CHALMERS | GÖTEBORGS UNIVERSITET

Ofullständig justering vid regressionsanalys

Examensarbete för kandidatexamen i matematik vid Göteborgs universitet
Kandidatarbete inom civilingenjörsutbildningen vid Chalmers

Fredrik Sangberg
Henrik Imberg
Love Carlson
Tobias Mikaelsson

Institutionen för matematiska vetenskaper
Chalmers tekniska högskola
Göteborgs universitet
Göteborg 2014

Ofullständig justering vid regressionsanalys

Examensarbete för kandidatexamen i matematisk statistik inom matematikprogrammet vid Göteborgs universitet

Fredrik Sangberg Henrik Imberg Tobias Mikaelsson

Kandidatarbete i matematik inom civilingenjörsprogrammet Bioteknik vid Chalmers

Love Carlson

Handledare: Staffan Nilsson
Examinator: Maria Roginskaya

Institutionen för matematiska vetenskaper
Chalmers tekniska högskola
Göteborgs universitet
Göteborg 2014

Sammanfattning

Ofta rapporterar medier om underliga samband som upptäckts i olika studier och undersökningar. Man hänvisar då till statistiska metoder och säger att ett signifikant samband påvisats. Dessvärre är dessa samband inte alltid att lita på även om studien är genomförd med goda statistiska metoder, speciellt om det finns osäkerhet i data. Mätfel påverkar resultaten av de metoder som används, vilket kan leda till att ett signifikant samband mellan en utfallsvariabel och en icke-orsakande variabel påvisas, trots att hänsyn tagits till de sanna orsakande faktorerna. I rapporten studeras detta fenomen djupare inom regressionsanalys med hjälp av simuleringar. I regressionsmodeller beskrivs en utfallsvariabel som en funktion av ett antal förklarande variabler. För att undersöka hur de förklarande variablerna påverkar utfallsvariabeln skattas effekterna av dessa. Simuleringar som genomförs syftar till att illustrera hur mätfel och bristande modellering påverkar skattningar och signifikans av de förklarande variablernas effekter på utfallsvariabeln. De visar att även små mätfel leder till försvårade möjligheter att dra korrekta statistiska slutsater. En teoretisk studie i hur osäkerhet i data påverkar de skattade effekterna i linjära regressionsmodeller genomförs. De teoretiska resultaten styrker resultaten av simuleringarna och visar att mätfel ger skattningar som inte är väntevärdesriktiga, dvs. som systematiskt avviker från de sanna effekterna.

Abstract

It is not unusual for media to report strange and outrageous findings, citing studies and surveys. Usually there's a reference to statistical methods indicating a significant association. Even though the study was carried out using good statistical methods these findings are not always trustworthy, a lot of uncertainties in the data should be a large source for concern. Errors in measurements can have large effect on the results of statistical methods. These errors can lead to a significant association between a response variable and a non causal predictor, even if the effect is adjusted for confounded variables. This report strives to shed light on this phenomenon in regression analysis by the use of simulations. It furthermore delves into the basic theory of regression analysis. In regression models, a response variable is described as a function of a number of predictor variables. The predictor variable's influence on the response is analyzed by studying their estimated effect. In this report simulations are used to illustrate how errors in measurement as well as inadequate models affect the estimation and significance of the predictor variables. It is shown that even small errors in measurement lead to difficulty making correct inferences. A theoretic study is carried out and shows that uncertainty in data affects the estimated effects in linear regression models, resulting in biased estimations.

Innehåll

Förord

Denna rapport syftar till att återge resultaten av ett kandidatprojekt inom matematisk statistik som genomfördes av Fredrik Sangberg, Henrik Imberg, Love Carlson och Tobias Mikaelsson. Gruppen har fört dagbok och tidslogg under kursens gång. I dagboken har gruppen sammanfattat projektets utveckling vecka för vecka, och i tidsloggen har en individuell tidrapportering förts.

Hela gruppen har tillsammans skrivit och sammanställt projektrapporten, och alla har mer eller mindre varit delaktiga i samtliga delar av projektet. Fredrik och Henrik har haft ansvar för programmering och den fördjupning i teorin som ges i kapitel 4. En översiktlig beskrivning av gruppmedlemmarnas bidrag till rapporten ges i tabellen nedan.

Avsnitt	**Huvudansvar**
Sammanfattning, Abstract	Alla
1 Inledning	Alla
2 Teori	Alla
3 Simuleringar	Alla
4 Osäkerhet i variabler	Fredrik, Henrik
5 Diskussion	Tobias, Henrik
Appendix A - Teoretiska härledningar	Fredrik
Appendix B - Alternativa regressionsmetoder	Henrik

Avslutningsvis vill vi tacka vår handledare Staffan Nilsson samt handledare på Avdelningen för Fackspråk och Kommunikation för all hjälp under kandidatprojektet och rapportskrivandet.

1 Inledning

Vid regressionsanalys försöker man att bygga upp en modell som beskriver en utfallsvariabels variation som en funktion av ett antal förklarande variabler. Utfallsvariabeln kallas även för responsvariabel och de förklarande variablerna för prediktorer.

När man ägnar sig åt statistiska studier där linjära regressionsmodeller används är justering en viktig men problematisk fråga. Antag att man i en modell inkluderar endast en variabel för att beskriva en utfallsvariabel. Med linjär regression kan man påvisa association mellan dessa variabler, men detta behöver inte visa på ett orsakssamband. Variabeln i modellen kan i själva verket vara relaterad till en annan variabel som orsakar utfallet.

Låt säga att man vill påvisa ett samband mellan konsumtion av alkohol och förekomst av lungcancer. Här kan antal liter konsumerad öl fungera som prediktor medan förekomst av lungcancer är utfallsvariabel. Vi känner till att rökning ökar risken för lungcancer och att många som röker också dricker mer än vanligt. Det är därför troligt att en modell som endast inkluderar alkohol och lungcancer kommer att visa på ett tydligt samband. För att justera bort den effekt som alkoholkonsumtion har på lungcancer enbart genom dess samvariation med rökning utökas modellen till två prediktorer, rökning och alkoholkonsumtion. Det är detta som kallas justering, nämligen att i en modell inkludera ytterligare variabler relaterade till både respons och den studerade prediktorn. Om alkohol inte har någon verklig inverkan på lungcancer borde detta synas i den nya modellen och man kan därmed undvika att dra den felaktiga slutsatsen att alkohol medför lungcancer. Justering är därför viktigt vid studier av orsakssamband. Att finna de sanna orssakssambanden kan i praktiken vara svårt, speciellt när det råder stor osäkerhet i observationer av den orsakande variabeln.

Brus i data kan vara av varierande karaktär och ha olika ursprung. Den mest uppenbara orsaken är mätfel i instrument. Det är möjligt att mätningar varierar i tid, för olika platser eller när olika personer utför ett försök, vilket leder till brus i data. Vid studier där deltagande individer får svara på frågeformulär uppstår en annan typ av brus. Det kan handla om svårigheter att svara korrekt på vissa frågor, vilket ger tendenser till överdrift eller underdrift. Om det ställs frågor av känslig karaktär kan felaktiga svar ges med avsikt. Slutligen kan brus uppstå av att en latent variabel, vilken är omöjlig att observera, ersätts av en så kallad proxy-variabel som påminner om den sanna variabeln. Exempelvis används BMI som proxy-variabel för fetma och BNP för livskvalitet inom ett land. I själva verket existerar osäkerhet i data vid alla typer studier och det är möjligt att olika variabler observeras med varierande precision. Det är därför av intresse att undersöka hur statistiska metoder påverkas av brus i data.

Syftet är att studera olika typer av regressionsmodeller då det finns osäkerhet i data eller i de modeller som ansätts, samt hur det påverkar prediktorernas skattade effekter och signifikans.

Rapporten begränsas till att endast diskutera additiva linjära och logistiska regressionsmodeller. I en additiv modell antar man att det inte finns någon samverkan mellan prediktorer. Effekten av en prediktor på utfallsvariabeln är alltså oberoende av övriga prediktorer. I en linjär modell beror responsvariabeln linjärt av prediktorerna. Det är givetvis vanligt förekommande i praktiken icke-linjära samband mellan variabler, samt att prediktorer samverkar. Författarna har ändå valt av avgränsa rapporten från icke-linjära modeller. Detta eftersom linjära modeller är av mer grundläggande karaktär och det är naturligt att i en första studie fokusera på dessa. Vidare är linjära modeller vanligt förekommande i praktiken, och med hjälp av variabeltransformationer kan mer komplicerade samband ofta beskrivas med linjära termer.

Metoder som har använts är litteraturstudier och simuleringar. Med hjälp av läroböcker och artiklar har olika regressionsmodeller studerats, där fokus varit på enkel och multipel linjär regression samt logistisk regression. Artiklar som diskuterar mer specifika ämnen inom regression har använts för att studera hur mätfel i data påverkar linjära modeller.

Med hjälp av simuleringar studerades situationer då det fanns osäkerhet i data. Simulering och teori samverkar och flera av de resultat som gavs med simuleringar styrktes av teoretiska resultat. I dessa fall fyller simuleringarna ändå en funktion eftersom de illustrerar

teorin och på så vis underlättar förståelsen. Även det som är omöjligt att räkna på analytiskt går ofta att studera med hjälp av simuleringar. Därför har simuleringar en central roll i rapporten. En mer detaljerad beskrivning av metod vid simuleringar ges i kapitel 3. Simuleringar genomfördes med R, ett programspråk utvecklat för statistik och dataanalys. Språket valdes på grund av dess breda användning inom statistik och rika mängd av funktioner anpassade för regressionsanalys.

2 Teori - en översikt i linjär regression

I detta kapitel diskuteras grunderna i linjär regression översiktligt. Fokus ligger inte på matematiska härledningar av de formler som presenteras utan på att lyfta fram de begrepp som är nödvändiga för rapportens kommande delar. Läsaren antas vara bekant med grundläggande sannolikhetsteori och inferens. För en mer detaljerad genomgång hänvisas till litteratur inom ämnet och referenser ges i slutet av varje delkapitel.

Linjär regression handlar om att undersöka samband mellan en responsvariabel y och en eller flera andra variabler, så kallade prediktorer. En enkel regressionmodell med endast en prediktor x_1 ges enligt

$$y = \beta_0 + \beta_1 x_1 + \epsilon_y \tag{2.0.1}$$

där ϵ_y är ett slumpartat brus, det vill säga variationen i y som inte kan förklaras av prediktorn x_1. En modell med endast en prediktor kallas även för en univariat regressionsmodell.

Utifrån stickprov där x och y observerats vill man anpassa en linje som så bra som möjligt beskriver sambandet mellan dem. Flera olika metoder kan användas till detta, men vanligast är minstakvadratmetoden som går ut på att minimera följande summa

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{1i}))^2 \tag{2.0.2}$$

genom att välja β_0 och β_1 på ett optimalt sätt. Här betecknar y_i och x_{1i} observationer av y respektive x_1 och n betecknar antalet observationer. Skattningarna av koefficienterna skrivs $\hat{\beta}_0$ och $\hat{\beta}_1$ och den skattade regressionslinjen ges av:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 \tag{2.0.3}$$

$\hat{\beta}_0$ beskriver linjens skärning med y-axeln och $\hat{\beta}_1$ beskriver linjens lutning. Till varje observation y_i svarar en punkt $\hat{y}_i$ på linjen. Avståndet mellan observerat värde y_i och predikterat värde $\hat{y}_i$ kallas residual och betecknas e_i:

$$e_i = y_i - \hat{y}_i \tag{2.0.4}$$

Ekvivalent med att minimera (2.0.2) är att minimera kvadratsumman av residualerna:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 \tag{2.0.5}$$

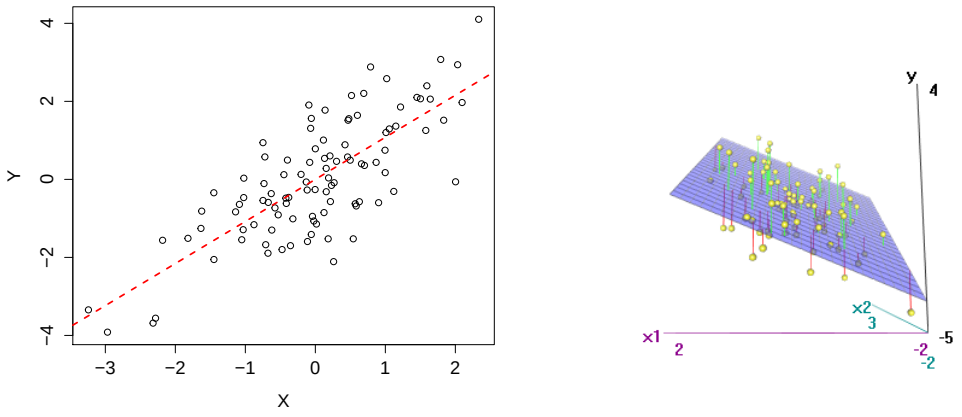
Då modellen innehåller två prediktorer vill man istället minimera residualkvadratsumman ifrån ett plan:

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}))^2 \tag{2.0.6}$$

Exempel på linjär regession med en respektive två prediktorer ges i figur 1 för att illustrera metoden.

Med p variabler $x_1, \ldots, x_p$ fås skattningarna av de $p+1$ regressionskoefficienterna $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ genom att minimera kvadratsumman av residualerna ifrån ett p-dimensionellt hyperplan

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j x_{ji}))^2. \tag{2.0.7}$$



Figur 1: Exemepel på linjär regression. Till vänster: y mot x med skattad regressionslinje. Till höger: y mot två prediktorer x_1 och x_2 med skattat regressionsplan. Här visas även residualerna med gröna och röda linjer. Punkter ovanför planet har gröna residualer och punkter under planet har röda.

2.1 Enkel linjär regression

Utgångspunkten i detta avsnitt är en enkel linjär regressionsmodell:

$$y = \beta_0 + \beta_1 x + \epsilon_y \tag{2.1.1}$$

y antas här vara kontinuerlig medan x kan vara antingen kontinuerlig eller kategorisk. Kategorisk innebär att variabeln antar diskreta värden. Det kan ofta röra sig om olika grupper i en population, exempelvis rökare/icke-rökare, man/kvinna eller indelningar i olika ålderskategorier. En variabel kallas binär eller dikotom om den endast antar två värden vilka ofta betecknas 0 och 1. Modeller då y är binär diskuteras i avsnitt 2.3.

Vid linjär regressionsanalys är det viktigt att y verkligen beror linjärt av x. Om ett icke-linjärt samband verkar gälla är det ibland möjligt transformera någon variabel och på så vis få ett linjärt samband. Man kan då utföra linjär regressionsanalys på de transformerade variablerna. I annat fall kan icke-linjära regressionsmodeller användas, då man exemepelvis kan inkludera x-termer av högre grad. Teorin kring detta är ett stort ämne i sig och kommer inte att diskuteras vidare.

Vid formulering av teorin för linjär regression betraktar man ofta observationerna x_i av x som fixa. Samtliga resultat i detta avsnitt håller dock även i fallet då prediktorn betraktas som en stokastisk variabel x under antagandet att $E[Y|X] = \beta_0 + \beta_1 X$.

Bruset i modellen antas vara oberoende och likafördelat för samtliga observationer x_i av x. Mer specifikt antas $\epsilon_y \sim N(0, \sigma_y)$, detta på grund av de många fördelar som normalfördelningsantagandet ger vid hypotestest. Vid regressionsanalys är det viktigt att noga undersöka om dessa antaganden verkar rimliga. Om så inte är fallet kan variabeltransformationer i vissa fall leda till att ovanstående antaganden är uppfyllda, varpå linjär regression kan användas. Om problemen kvarstår kan generaliserade linjära regerssionmodeller användas. Logistisk regression är ett exemepel på en generaliserad modell och diskuteras i kapitel 2.3.

Låt oss nu anta att y beror linjärt av x och att ovanstående antaganden gällande bruset i modellen är uppfylda. Då kan minsta-kvadratmetoden användas för att skatta regressionskoefficienterna, och skattningarna för β_0 och β_1 ges av

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} \tag{2.1.2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.1.3}$$

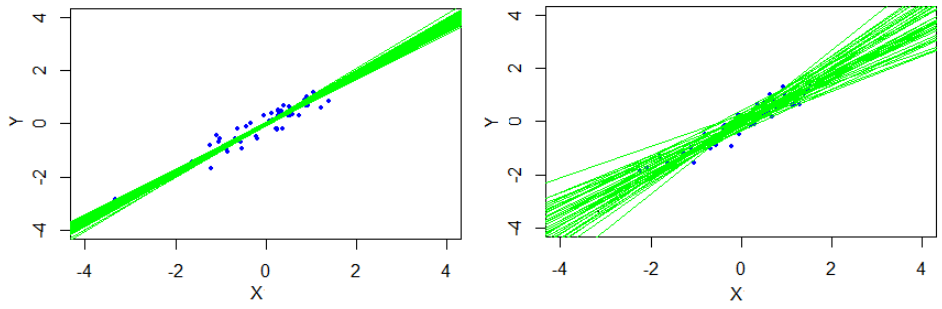
där $\bar{x}$ och $\bar{y}$ betecknar medelvärden av observationerna i x respektive y. Metoden ger väntevärdesriktiga skattningar, dvs. $E[\hat{\beta}_0] = \beta_0$, $E[\hat{\beta}_1] = \beta_1$. I själva verket kommer β_0 inte att

vara av stort intresse i denna rapport eftersom β_0 inte bidrar med information om förhållandet mellan prediktor och respons annat än med en förskjutning i y-led. Följande diskussion kommer därför i första hand att behandla skattningen av β_1. Om prediktorn x är kontinuerlig tolkas $\hat{\beta}_1$ som den förväntade förändringen i responsen y vid en enhetsförändring i x. Om x är dikotom är tolkningen istället den förväntade skillnaden i y mellan de grupper som x representerar. Om x är diskret med mer än två kategorier används ofta en av kategorierna som referens. För var och en av de övriga kategorierna skattas en regressionskoefficient. Om x har k kategorier skattas alltså $k-1$ regressionsparametrar $\beta_1, \ldots \beta_{k-1}$ som beskriver skillnaden mellan de olika kategorierna och referenskategorin.

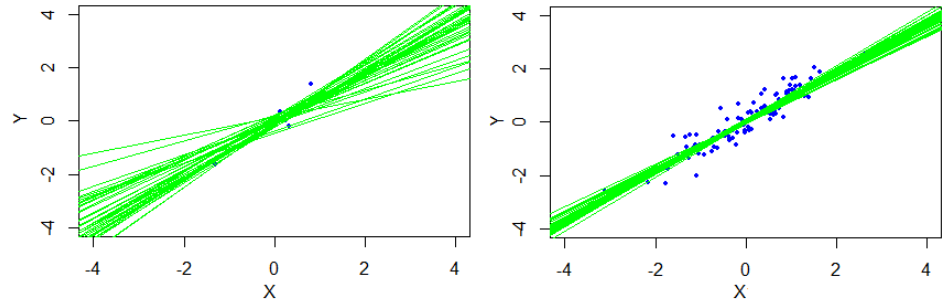
Variansen för den skattade lutningsparametern $\hat{\beta}_1$ ges av

$$Var[\hat{\beta}_1] = \frac{\sigma_y^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma_y^2}{(n-1)\sigma_x^2} \tag{2.1.4}$$

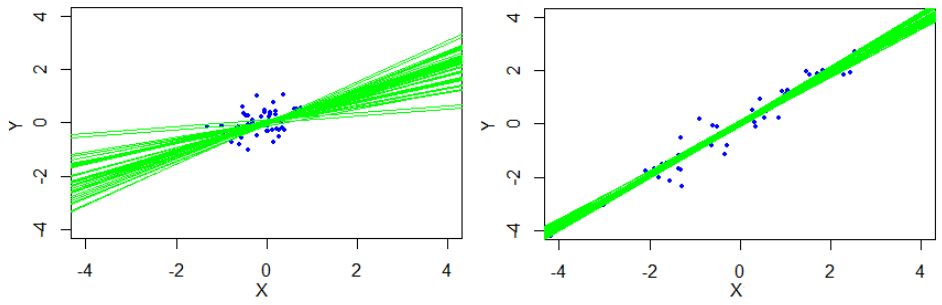
där σ_x^2 betecknar variansen av x. Stabila skattningar med liten varians fås alltså vid stor spridning i observationer, stora stickprov och litet brus i modellen. Detta illustreras i figur 2-4 där ett antal linjer anpassats efter olika data där σ_y^2, σ_x^2 och stickprovsstorleken n har varierats.



Figur 2: Variation i $\hat{\beta}_1$ vid störningar i data. n=50 och $\sigma_x^2 = 1$ i båda fallen. Till vänster: $\sigma_y^2 = 0.1$. Till höger: $\sigma_y^2 = 1$. Regressionslinjen visar en större variation för olika stickprov då modellbruset är stort.



Figur 3: Variation i $\hat{\beta}_1$ vid störningar i data. $\sigma_x^2 = 1$ och $\sigma_y^2 = 0.1$. Till vänster: n=10. Till höger: n=100. Regressionslinjen visar en större variation för olika stickprov då stickprovsstorleken är liten.



Figur 4: Variation i $\hat{\beta}_1$ vid störningar i data. n=50 och $\sigma_y^2 = 0.1$ i båda fallen. Till vänster: $\sigma_x^2 = 0.1$. Till höger: $\sigma_x^2 = 1$. Regressionslinjen visar en större variation för olika stickprov då spridningen i x är liten.

Det är ofta av intresse att specificera hur bra en modell förklarar relationen mellan variabler. Ett mått på detta är 'R-squared':

$$R^2 = \frac{SS_R}{SS_T} = \frac{SS_T - SS_E}{SS_T} = 1 - \frac{SS_E}{SS_T} \tag{2.1.5}$$

vilken tolkas som hur stor andel variationen i y som förklaras av x. Här betecknar

$$SS_T = \sum (y_i - \bar{y})^2, \qquad SS_E = \sum (y_i - \hat{y}_i)^2, \quad SS_R = \sum (\hat{y}_i - \bar{y})^2. \tag{2.1.6}$$

Summationsindex i utelämnas i utrycken ovan men det är underförstått att $i = 1, \ldots n$ där n är antalet observationer.

SS_T, 'total sum of squares', ger variationen kring medelvärdet av observationerna av y. SS_E, 'error sum of squares', ger variationen kring den skattade regressionlinjen. SS_R, 'regression sum of squares', ger de skattade värdenas variation kring $\bar{y}$.

Om prediktorn beskriver utfallsvariabeln väl kommer residualerna att vara små, så att $SS_E \approx 0$ och därmed $R^2 \approx 1$. På samma sätt gäller i en modell där sambandet mellan prediktor och respons är svagt att $SS_E \approx SS_T$ och därmed $R^2 \approx 0$. I denna situation med endast en prediktor är R^2 samma som kvadraten av korrelationen mellan prediktor och respons.

Eftersom SS_E beskriver variationen kring den skattade regressionslinjen används denna för att skatta variansen för bruset i modellen

$$\hat{\sigma}_y^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} = \frac{SS_E}{n-2}$$

där $(n-2)$ i nämnaren är lika med 'antalet observationer' minus 'antalet skattade parametrar'.

2.1.1 Inferens

Antag att man vill testa nollhypotesen $H_0 : \beta_1 = 0$ mot alternativet $H_1 : \beta_1 \neq 0$. Att β_1 är signifikant skild från noll vid signifikansnivå α innebär att det med $(1-\alpha)\%$ konfidensgrad finns ett linjärt samband mellan x och y. Observera att detta inte nödvändigtvis innebär att x orsakar y. Fallet kan lika gärna vara det motsatta, att y orsakar x, eller så kan både x och y vara relaterade till en tredje variabel vilket ger upphov till ett signifikant samband. Om antagandet att $\epsilon_y \sim N(0, \sigma_y)$ stämmer fås följande teststatistika:

$$T_{obs} = \frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \underset{H_0:\beta_1=0}{\sim} t_{n-2} \tag{2.1.7}$$

Standardfelet, dvs. den skattande standardavvikelsen, för $\hat{\beta}_1$ ges av:

$$SE[\hat{\beta}_1] = \sqrt{\frac{\hat{\sigma}_y^2}{\sum (x_i - \bar{x})^2}} = \sqrt{\frac{\hat{\sigma}_y^2}{(n-1)\hat{\sigma}_x^2}} \tag{2.1.8}$$

Antalet frihetsgrader, $n-2$, kommer från det faktum att det är två parametrar som skattas i modellen, nämligen $\hat{\beta}_0$ och $\hat{\beta}_1$. Utifrån detta kan man räkna ut p-värdet för den observerade teststatistikan, $P(|T| > |T_{obs}| \mid H_0)$, dvs sannolikheten för en lika eller mer extrem observation under förutsättning att H_0 är sann. H_0 förkastas om detta p-värde är mindre än en vald signifikansnivå α som typiskt väljs till 0.05. En mer ingående introduktion till linjära regressionsmodeller med en prediktor ges av Alm och Britton [1].

2.1.2 Exempel

Flera olika statistiska datorprogram som kan användas för att utföra linjär regression och inferens för regressionskoefficienterna. I detta arbete har programspråket R använts och nedanför ges ett exempel på hur informationen kan se ut. Motsvarande data och estimerad regressionslinje visas i figur 5.

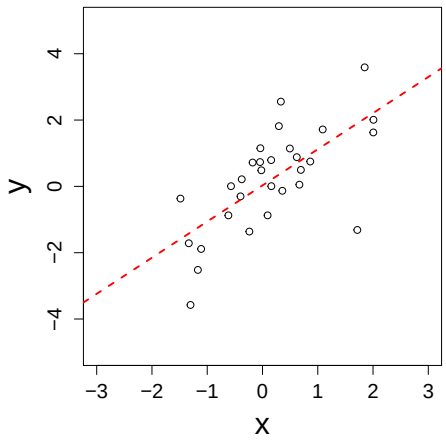
```
Call:
lm(formula =~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2135 -0.5797 -0.0476  0.5963  2.1643

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02941    0.21061   0.140     0.89
x            1.08987    0.21933   4.969 3.02e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.139 on 28 degrees of freedom
Multiple R-squared:  0.4686,Adjusted R-squared:  0.4496
F-statistic: 24.69 on 1 and 28 DF,  p-value: 3.017e-05
```

I sammanfattningen från R ges bland annat skattningen av lutningen ($\hat{\beta}_1 = 1.08987$) och dess tillhörande standardavfel ($SE[\hat{\beta}_1] = 0.21933$) och test-statistika ($T = 4.969$). I detta exempel är lutningen signifikant skild från noll med p-värde $3.02 * 10^{-5} < 0.05$. Data genererades från en modell med $\beta_0 = 0$ och $\beta_1 = 1$ och vi ser att båda skattningarna stämmer väl med de sanna värdena, samt att den sanna nollhypotesen att $\beta_0 = 0$ inte kan förkastas. $R^2 = 0.4686$, vilket innebär att 47% av variationen i y kan förklaras av x enligt modellen. Från sammanfattningen får man även ut residualernas standardfel och antalet frihetsgrader. Tillsammans med R^2 ger de en F-statistika med tillhörande p-värde, som testar nollhypotesen att ingen regressionskoefficent är skild från noll mot att någon koefficient har lutning skild från noll. Detta diskuteras mer i kapitel 2.2. I fallet då bara en variabel ingår i modellen blir p-värdet för T-statistikan och F-statistikan samma.



Figur 5: Scatterplot med anpassad regressionslinje efter data från exemplet på föregående sida.

2.2 Multipel linjär regression

En enkel linjär regressionsmodell kan utvecklas till att inkludera flera prediktorer, såväl kontinuerliga som kategoriska. Antag att man med $p-1$ prediktorer[1] vill beskriva en utfallsvariabel y enligt en linjär modell

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \epsilon_y \tag{2.2.1}$$

Låt $\mathbf{y} = \{y_1, \ldots, y_n\}^T$ vara en n x 1-vektor med observationer av utfallsvariabeln y, $\beta = \{\beta_0, \ldots, \beta_{p-1}\}^T$ en p x 1-vektor med regressionsparametrar och $\epsilon_y = \{\epsilon_{y1}, \ldots, \epsilon_{yn}\}^T$ en n x 1-vektor med normalfördelade feltermer, $\epsilon_{yi} \sim N(0, \sigma_y)$. Låt vidare $\mathbf{X}$ vara den så kallade designmatrisen given av:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{p-1,1} \\ 1 & x_{1,2} & \cdots & x_{p-1,2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \cdots & x_{p-1,n} \end{pmatrix} \tag{2.2.2}$$

Detta är alltså en n x p-matris bestående observationer. Det första indexet för elementen i matrisen betecknar variabel och det andra betecknar observation. Den första kolonnen motsvarar β_0 och i kolonn den j+1 återfinns observationer av x_j. Med ovanstående beteckningar kan den multipla linjära regressionsmodellen skrivas:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{2.2.3}$$

För att minimera kvadratsumman av residualerna används normalekvationerna

$$\mathbf{X^T X}\beta = \mathbf{X^T y} \tag{2.2.4}$$

vilket leder till minstakvadratskattningen för vektorn β

$$\hat{\beta} = (\mathbf{X^T X})^{-1} \mathbf{X^T y} \tag{2.2.5}$$

förutsatt att $\det(\mathbf{X^T X}) \neq 0$. Prediktorerna i modellen får alltså ej vara linjärt beroende. Matrisen $(\mathbf{X^T X})$ beskriver relationen mellan prediktorerna, dvs. deras varians och kovarianserna mellan dem. Om en prediktor x_j är oberoende av de övriga kommer dess skattade lutningskoefficient enbart att bero av observationerna av x_j, medan korrelerade prediktorer

[1] $p-1$ eftersom det ger p regressionskoefficienter då β_0 inkluderas.

kommer att påverka varandra. Kraftig korrelation medför att $\det(\mathbf{X^T X}) \approx 0$, så invertering av $\mathbf{X^T X}$ blir då en instabil operation och skattningarna blir känsliga för störningar i data.

$\hat{\beta}_j$ beskriver nu den förväntade förändringen i y vid en enhetsförändring i x_j, då övriga variabler hålls konstanta. Skattningen av vektorn β är väntevärdesriktig, dvs. $\mathbf{E}[\hat{\beta}] = \beta$, och kovariansmatrisen för regressionskoefficienterna ges av

$$\sigma^2(\mathbf{X^T X})^{-\mathbf{1}} \tag{2.2.6}$$

med varianser för koefficienterna på diagonalen. Statistikan för att testa om skattningen av en regressionskoefficient β_j är skild från noll ges av

$$T_{obs} = \frac{\hat{\beta}_j - \beta_j}{SE[\hat{\beta}_j]} \underset{H_0:\beta_j=0}{\sim} t_{n-p} \tag{2.2.7}$$

där

$$SE[\hat{\beta}_j] = \sqrt{\hat{\sigma}^2 (X^T X)^{-1}_{(j,j)}} = \sqrt{\frac{\hat{\sigma}_y^2}{(1 - R^2_{x_j})(n-1)\hat{\sigma}^2_{x_j}}} \tag{2.2.8}$$

$\hat{\sigma}^2_{x_j}$ betecknar variansen för x_j och $R^2_{x_j}$ är 'R-squared' när regression utförs med x_j som utfallsvariabel och övriga prediktorer som förklarande variabler. Termen

$$VIF(x_j) = \frac{1}{1 - R^2_{x_j}} \tag{2.2.9}$$

kallas för 'variance inflation factor' och förklarar hur variansen för en regressionsparameter påverkas av övriga variabler. Skattningar av regressionskoefficienter för prediktorer som är starkt korrelerade kommer alltså ha stor varians eftersom $R^2_{X_i} \approx 1$, vilket medför att VIF blir stor.

För att testa nollhypotesen $H_0 : \beta_1 = \ldots = \beta_{p-1} = 0$ mot alternativet $H_1 : \beta_j \neq 0$ för något β_j, $j = 1, \ldots p-1$ används ett F-test med test-statistika:

$$F_{obs} = \frac{SS_T - SS_E/(p-1)}{SS_E/(n-p)} \underset{H_0}{\sim} F_{p-1,n-p} \tag{2.2.10}$$

För regressionsmodeller med korrelerade variabler är det möjligt att F-statistikan blir signifikant skild från noll även om ingen av t-statistikorna är signifikant. Detta beror på att modellen förklarar utfallsvariabeln, vilket F-statistikan visar, men att det är omöjligt att avgöra vilken av prediktorerna som orsakar utfallet.

I avsnitt 2.1 introducerades R^2 som ett mått på hur 'bra' en modell förklarar responsvariabeln. Då fler prediktorer adderas till modellen kommer R^2 att förbättras, även om dessa är orelaterade till y. Detta eftersom anpassningen efter data alltid blir minst lika bra om en ny variabel inkluderas. För att kompensera för antalet variabler definieras

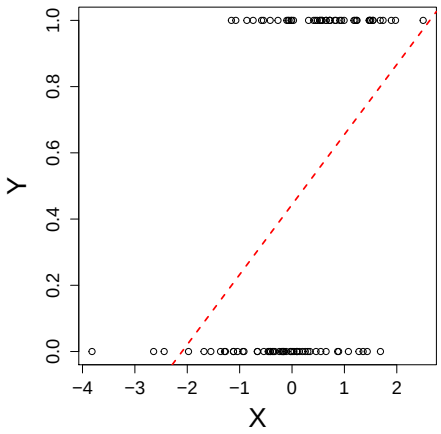
$$R^2_{adj} = 1 - \frac{SS_E}{SS_T}\left(\frac{n-1}{n-p}\right). \tag{2.2.11}$$

som alltså liknar R^2 men straffar då variabler som inte tillför information inkluderas i modellen. För en mer djupgående genomgång av multipel linjär regression hänvisas till Rawlings m. fl. [2].

2.3 Logistisk regression

I detta avsnitt antas y vara dikotom och kan exempelvis stå för förekomst av en viss sjukdom hos en individ. Prediktorn x kan vara antingen kontinuerlig eller kategorisk och är en faktor vars inverkan på y skall undersökas. Situationen då x är kategorisk med k kategorier kan illustreras med en 2 x k-tabell och sambandet mellan x och y och testas vanligtvis med ett χ^2-test.

Vid modellering då y är dikotom vill man dels kunna uttala sig om utfallet i y givet x med så stor säkerhet som möjligt, men även att uppskatta sannolikheten för y som en funktion av x. En möjlighet är givetvis att utföra vanlig linjär regression enligt en modell $y = \beta_0 + \beta_1 x + \epsilon$, för något brus ϵ, och anpassa en linje efter data enligt minstakvadratmetoden. Ett sådant exempel illustreras i figur 6.



Figur 6: Ett exempel på när linjär regression används trots att utfallsvariabeln är binär. Regressionslinjen kan inte användas för att beskriva sannolikheten att $y = 1$ eftersom linjen antar värden utanför intervallet [0,1] för |x|>2.

.

Utifrån modellen i figur 6 går det inte att uttala sig om sannolikheten för y eftersom den skattade regressionslinjen antar värden utanför [0,1]. Ytterligare ett problem med modellen är inferens för regressionskoefficienterna. I detta fall verkar det svårt att uttala sig om huruvida x och y är relaterade då utfallet i y inte visar på tydliga variationer för olika x. Det verkar dock som att sannolikheten för $y = 1$ tenderar att minska för små x. Med ett t-test från vanlig regression är $\hat{\beta}_1$ signifikant skild från noll. Problemet är att samtliga hypotestest som diskuterats tidigare bygger på antagandet att bruset är likafördelat för alla observationer med $\epsilon_y \sim N(0, \sigma_y)$ vilket inte håller i denna situation eftersom avvikelserna mellan linje och data är beroende av x.

Det man i allmänhet är intresserad av när y är dikotom är att kunna uttala sig om sannolikheten att $y = 1$ givet x. Man vill alltså hitta en funktion som för samtliga x returnerar värden mellan noll och ett. Med linjär regression kan i princip vilka värden som helst ges, varför logistisk regession används istället. Vi skall börja med att introducera den logistiska funktionen

$$p(z) = \frac{1}{1 + e^{-z}}, \quad z \in \mathbb{R} \tag{2.3.1}$$

vilken antar värden i (0,1) och alltså uppfyller kravet som beskrivits ovan. Genom att vidare låta

$$z = \beta_0 + \beta_1 x \tag{2.3.2}$$

fås

$$p(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}. \tag{2.3.3}$$

Här betecknar $p(x)$ sannolikheten att $y = 1$ givet x, och z kallas den linjära prediktorn. Formel (2.3.3) beskriver den logistiska regressionsmodellen för en prediktor. Även i logistiska regressionsmodeller kan flera prediktorer ingå genom att inkludera fler termer i den linjära prediktorn:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \tag{2.3.4}$$

En formulering ekvivalent med (2.3.3) är

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x \tag{2.3.5}$$

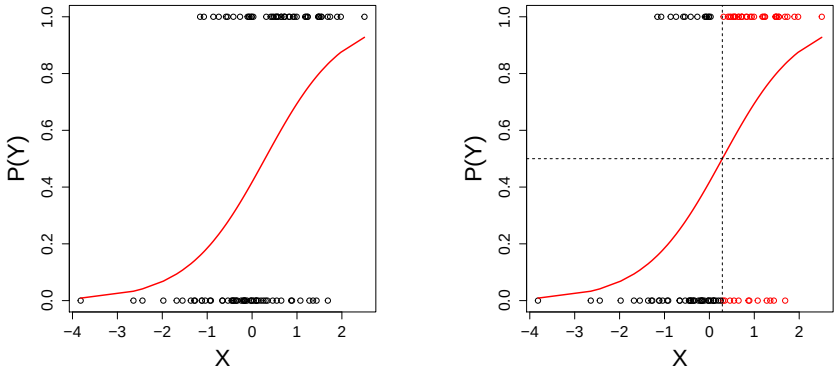
där vänsterledet beror linjärt av x.

En logistisk regressionsmodell anpassad efter data illustreras i figur 7. Vid prediktion klassificerar man observationer efter ett valt tröskelvärde. Detta kan exempelvis göras genom

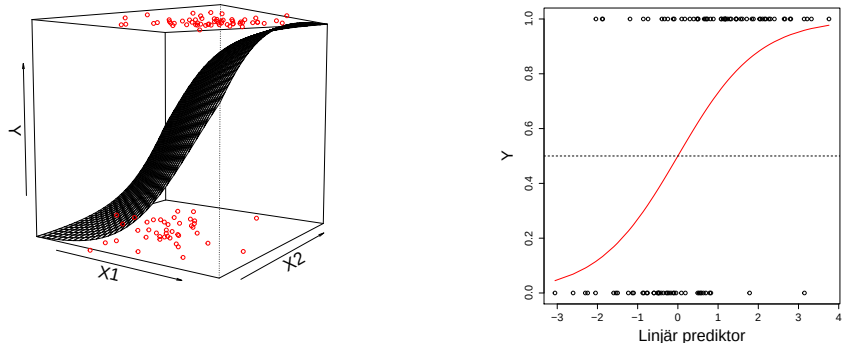
att låta observationer med predikterad sannolikhet $P(y) \geq 0.5$ tilldelas värdet 1 och övriga värdet 0. Observationer i figur 7 som har tilldelats värdet 1 har i den högra bilden markerats med rött. Det är svårare att ge en intuitiv tolkning av skattningen av β_1 än vid linjär regression, men om $\hat{\beta}_1 > 0$ förväntas sannolikheten att $y = 1$ öka för en positiv förändring i x.

En skattad regressionsyta för en modell med två prediktorer visas i figur 8. En alternativ framställning är att plotta den logistiska funktionen mot den linjära prediktorn $z = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ enligt uttryck (2.3.3), vilket visas i samma figur.

Det finns ingen explicit formel för att skatta parametrarna i logistiska regressionsmodeller, utan skattningar fås vanligtvis numeriskt med maximum likelihood-metoden. Även metoder för inferens är annorlunda. Att diskutera dessa mer ingående faller dock inte inom ramen för rapporten utan hänvisas till litteratur som behandlar generaliserade linjära regressionsmodeller, se exempelvis Kleinbaum och Klein [3].



Figur 7: Logistisk regression. Vänster: Logistiska funktionen ges av den s-formade kurvan, vilken beskriver sannolikheten att $y = 1$ som funktion av x. Höger: Observationer med $P(Y) > 0.5$ har tilldelats värdet $y = 1$ och markeras med rött. Övriga har predikterat värde $y = 0$.



Figur 8: Logistisk regression med två prediktorer. Vänster: Skattad yta för två prediktorer i 3 dimensioner. Höger: Logistiska funktionen mot den linjära prediktorn $z = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$.

2.4 Justering för samvarierande variabler

Låt oss betrakta en utfallsvariabel y och en prediktor x_2 med lutningskoefficient β_2. Beteckningarna är valda för att situationen skall stämma överens med kommande kapitel. Antag att det inte finns något orsakssamband mellan dessa variabler, dvs. $\beta_2 = 0$. P-värdet för $\hat{\beta}_2$ visar sannolikheten för en minst lika extrem observation av β_2 under nollhypotesen att $\beta_2 = 0$. Trots att det inte finns något orsakssamband mellan x_2 och y är det möjligt att skattningen blir signifikant. Man riskerar då ledas till slutsatsen att x_2 har en inverkan på responsvariabeln y. Det kan dock finnas andra omständigheter som leder till att $\hat{\beta}_2$ blir signifikant. Låt oss anta att x_2 samvarierar med en annan variabel x_1, som i sin tur orsakar y. Det existerar då ett indirekt samband mellan x_2 och y, vilket leder till att $\hat{\beta}_2$ blir signifikant när regression utförs med enbart x_2 som förklarande variabel. I sådana situationer är det viktigt att ta hänsyn till samvariationen mellan x_1 och x_2, vilket åstadkoms med justering.

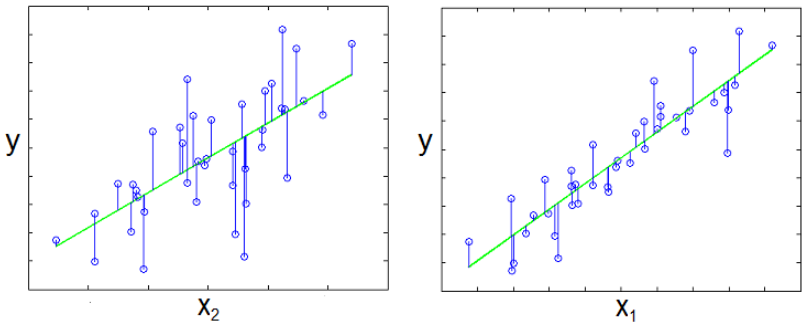
Kort beskrivet så innebär justering att inkludera ytterligare variabler i en modell. Man säger sig då ha justerat för dessa variabler. Istället för att skatta en linje $y = \beta_0 + \beta_2 x_2$ inkluderas även x_1 i modellen, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, vilket ger ett skattat regressionsplan. När parametrarna har estimerats uppstår fyra olika fall.

1. **Bara $\mathbf{x_1}$ blir signifikant.** Eftersom $\beta_2 = 0$ i exemplet ovan innebär detta att justeringen har lyckats och nollhypotesen att x_2 saknar inverkan på y kan inte förkastas.

2. **Bara $\mathbf{x_2}$ blir signifikant.** Om det tidigare var känt att x_1 orsakar y bör man förhålla sig kritiskt till sådana resultat, eftersom de antyder att det är x_2 och inte x_1 som orsakar y.

3. **Båda variablerna blir signifikanta.** Detta antyder att det finns ett samband mellan båda prediktorerna och utfallsvariabeln, och att sambandet mellan x_2 och y inte enbart kan förklaras av samvariationen mellan x_1 och x_2.

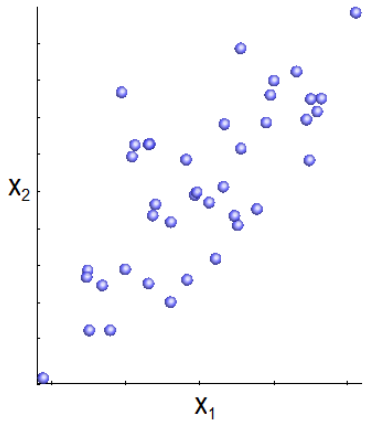
4. **Ingen av variablerna blir signifikant.** I dessa fall kommer F-statistikan troligtvis att vara signifikant, vilket visar att det finns ett signikant samband mellan prediktorer och respons, men det är omöjligt att avgöra vilken av prediktorerna som orsakar y.



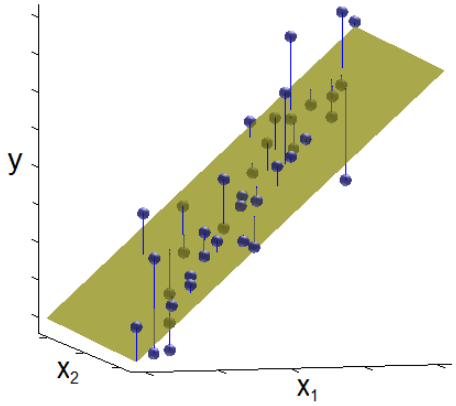
Figur 9: Vänster: Regression med y mot enbart x_2. Höger: Regression med y mot enbart x_1. Både x_1 och x_2 verkar visa på ett signifikant samband med y när de studeras var och en för sig.

Ett exempel på data som illustrerar detta visas i figur 9 - 11 och skall nu diskuteras. Som vi ser i figurerna ovan verkar det som om både x_1 och x_2 har en signifikant inverkan på responsvariabeln y eftersom de skattade parametrarna $\hat{\beta}_1$ och $\hat{\beta}_2$ är signifikant skilda från noll i regressionsmodeller som endast inkluderar en variabel åt gången. Vi låter som exempel y stå för förekomst av lungcancer i en population, x_1 och x_2 för konsumtion av tobak respektive alkohol. Enligt den vänstra figuren ser följande samband ut att vara bevisat: Ju mer alkohol du konsumerar, desto större risk löper du att drabbas av lungcancer. Den högra figuren visar det vedertagna sambandet mellan rökning och lungcancer. Det finns dock en dold problematik i det hela, nämligen att rökning och alkoholkonsumtion är korrelerade.

Att x_1 och x_2 är korrelerade visas i figur 10. Det skulle alltså kunna vara sambandet mellan konsumtion av alkohol och tobak som gör att alkohol verkar ha en inverkan på lungcancer. Detta underskös genom att studera en modell där båda variablerna ingår och skatta ett regressionsplan, vilket illustreras i figur 11. I och med detta justerar man alltså för rökning när man studerar sambandet mellan alkohol och lungcancer.



Figur 10: x_1 och x_2 är tydligt korrelerade.



Figur 11: Planet $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, dvs den justerade modellen

Genom att betrakta planets lutning på bilden kan vi förstå vad som hänt. Den justerade modellen ger ett plan som verkar öka längs med x_1-axeln. Förändring i x_2 leder däremot inte till märkvärd förändring i y-led. Justeringen har i detta fall lett till att x_2 inte längre är signifikant och har alltså lyckats.

Det bör påpekas att ovanstående exempel med alkoholkonsumtion, rökning och lungcancer inte är hämtat från någon genomförd studie. Data har simulerats utifrån hur det skulle kunna se ut i en verklig situation.

3 Simuleringar

Simuleringar har använts för att identifiera problem vid regressionsanalys då det finns mätfel och osäkerhet i data. Detta eftersom slutna analytiska lösningar till de problem som uppstår saknas för exempelvis logistisk regression. För linjär regression visar det sig vara möjligt att beskriva resultaten analytiskt, men de simuleringar som genomförs ger ändå en förståelse för problemen och illustrerar teorin. I följande avsnitt ges en genomgång av tillvägagångssätt i simuleringar samt diskussioner kring resultat av simuleringar. Endast modeller med två prediktorer, x_1 och x_2, studeras. x_1 betecknar genom hela kapitlet den kausala variabeln, det vill säga den variabel som har en sann inverkan på utfallsvariabeln y. x_2 har ingen inverkan på y men är korrelerad med x_1. x_2 samvarierar alltså med x_1, och denna samvariation ger upphov till en skenbar effekt av att ett orsakssamband skulle råda mellan x_2 och y. Både x_1 och x_2 antas mätas med viss osäkerhet och de observerade variablerna betecknas $\tilde{x}_1$ respektive $\tilde{x}_2$.

För att göra simuleringar mer realistiska och jämförbara mellan olika situationer väljs stickprovsstorlek och brus i modell på lämpligt sätt. Nämligen så att β_1 blir signifikant i 90% av fallen i då endast $\tilde{x}_1$ ingår som prediktor i modellen, och då det inte finns något mätfel i x_1. Man säger då att styrkan för testet att $\beta_1 \neq 0$ är 90%, där styrkan för ett test definieras som P(Förkasta $H_0|H_0$ falsk). Styrkan beskriver alltså hur bra ett test är på att förkasta en falsk nollhypotes, och en styrka nära 1 är önskvärt. Hög styrka fås genom en god design för ett experiment eller en studie. Det är vanligt att anpassa stickprovsstorlek i studier så att önskad styrka fås, då ökad stickprovsstorlek ger ökad styrka. I praktiken är det ofta dyrt med stora undersökningar och det finns ekonomiska och praktiska begräsningar till hur stora stickprov som kan fås, och en styrka mellan 0.8 och 0.9 är vanligt. En styrka på 0.9 vid simuleringar är alltså vald för att ge realistiska resultat. Genom att vidare utgå från samma styrka vid samtliga simuleringar blir resultaten jämförbara. I simuleringarna gäller att nollhypotesen $H_0 : \beta_1 = 0$, i modellen $y = \beta_0 + \beta_1 x_1$, är falsk. Om styrkan för testet är 0.9 förväntas alltså H_0 förkastas och β_1 bli signifikant i 90% av fallen. Teori kring hur parametrar skall väljas för att ge önskad styrka ges i appendix A.

Båda de univariata modellerna där y förklaras av antingen $\tilde{x}_1$ eller $\tilde{x}_2$, samt den justerade modellen där y förklaras av både $\tilde{x}_1$ och $\tilde{x}_2$ studeras. Syftet med simuleringarna är att upptäcka när $\tilde{x}_2$ visar på ett signifikant samband med y som inte kan förklaras av dess samvariation med $\tilde{x}_1$. Detta är alltså de fall då $\tilde{x}_2$ är signifikant även efter justering för $\tilde{x}_1$, vilket vi kallar för ofullständig justering. Antalet fall med ofullständig justering förväntas öka då brusets storlek ökar i observationer av x_1. Detta eftersom $\tilde{x}_2$ då riskerar att förklara y bättre än $\tilde{x}_1$. Mätfelet i x_1 kommer därför att varieras medan mätfelet i x_2 kommer att hållas fixt. Kvoten mellan dessa mätfel kommer att utgöra x-axeln i de figurer som presenteras.

3.1 Studerade modeller

De modeller som används vid simuleringar skall nu presenteras. Sambanden mellan variabler och de modeller som ansätts är valda för att på ett enkelt sätt kunna åskådliggöra resultaten. Mer komplicerade samband och modeller skulle givetvis kunna användas, men dessa enkla modeller räcker för att belysa effekterna av osäkerhet i data på resultaten av regressionsanalys.

Genomgående i kapitel 3 antas y orsakas av x_1 enligt

$$y = \beta_0 + \beta_1 x_1 + \epsilon_y, \quad \epsilon_y \sim N(0, \sigma_y) \tag{3.1.1}$$

där ϵ_y betecknar variation i y till följd av okända faktorer och individuell variation. Modell (3.1.1) visar hur det verkliga förhållandet mellan de studerade variablerna ser ut. Det råder inget verkligt orsakssamband mellan x_2 och y, men kunskap om hur dessa variabler förhåller sig är okänt.

Antag att man i en studie försöker påvisa ett samband mellan x_2 och y enligt en linjär modell:

$$y = \beta_0 + \beta_2 x_2 \tag{3.1.2}$$

Modell (3.1.2) försöker beskriva ett samband som inte finns. Antag att det visar sig att $\hat{\beta}_2$ är signifikant skild från noll. Antag vidare att det är känt att x_1 har en inverkan på y samt

att x_1 och x_2 samvarierar. För att försäkra sig om att det signifikanta sambandet mellan x_2 och y inte beror på samvariationen med x_1 studeras även en modell som inkluderar båda prediktorerna:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{3.1.3}$$

I denna modell borde β_2 inte bli signifikant eftersom $\beta_2 = 0$ i (3.1.1). Som kommande simuleringar visar är så inte alltid fallet, speciellt inte om det finns mätfel i data. Modell (3.1.3) kallas för 'den justerade modellen' eftersom man här har justerat för det samband som råder mellan x_2 och y via samvariationen med x_1.

Då x_1 och x_2 mäts med brus observeras två variabler, $\tilde{x}_1$ och $\tilde{x}_2$, vilka kan skrivas

$$\tilde{x}_1 = x_1 + \epsilon_1 \tag{3.1.4}$$

$$\tilde{x}_2 = x_2 + \epsilon_2 \tag{3.1.5}$$

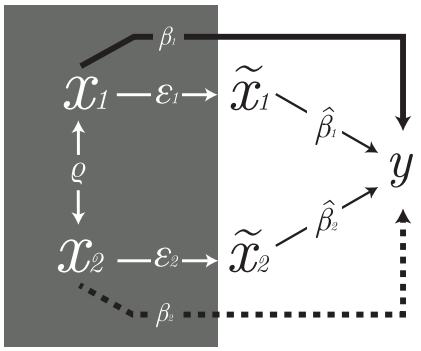
där ϵ_1 och ϵ_2 betecknar mätfelen i respektive variabel. Vidare ersätts modell (3.1.1) och (3.1.2) med

$$y = \beta_0 + \beta_1 \tilde{x}_1 \tag{3.1.6}$$

och

$$y = \beta_0 + \beta_2 \tilde{x}_2 \tag{3.1.7}$$

I figur 12 visas förhållanden mellan de variabler som studeras. x_1 orsakar y med effekt β_1. x_1 och x_2 samvarierar, vilket gör att x_2 får en skenbar inverkan på y med effekt β_2. I själva verket är $\beta_2 = 0$, dvs. x_2 har ingen inverkan på y. Prediktorerna mäts med brus och två relaterade variabler $\tilde{x}_1$ och $\tilde{x}_2$ observeras, från vilka effekterna av de sanna variablerna estimeras.



Figur 12: Samband mellan variabler vid simulering. Effekterna av variablerna i den grå zonen skattas med de observerade variablerna $\tilde{x}_1$ pch $\tilde{x}_2$.

3.2 Simulering med linjär regression

I detta delkapitel studeras modeller där samtliga variabler är kontinuerliga. x_1 och x_2 genererades från en bivariat normalfördelning

$$(x_1, x_2) \sim N(\mu_{x_1}, \mu_{x_2}, \sigma_{x_1}, \sigma_{x_2}, \rho) \tag{3.2.1}$$

där μ_{x_1} och μ_{x_2} betecknar de teoretiska medelvärdena för x_1 respektive x_2, σ_{x_1} och σ_{x_2} deras standardavvikelser och ρ korrelationen mellan variablerna. Detta kan också uttryckas som

$$x_2 = \alpha_0 + \alpha_1 x_1 + \delta \tag{3.2.2}$$

för några koefficienter α_0, α_1 och något normalfördelat brus δ. En bivariat normalfördelning valdes eftersom förhållandet mellan x_1 och x_2 då blir linjärt, och y kan skrivas som en linjär funktion av en eller båda prediktorerna. Då x_1 och x_2 samvarierar linjärt säger man att de är korrelerade, men även kvadratisk samvariation diskuteras i kapitel 3.4.

Även mätfelen genererades från en normalfördelning och de observerade variablerna kan skrivas

$$\tilde{x}_1 = x_1 + \epsilon_1, \quad \epsilon_1 \sim N(0, \sigma_1) \tag{3.2.3}$$

$$\tilde{x}_2 = x_2 + \epsilon_2, \quad \epsilon_2 \sim N(0, \sigma_2) \tag{3.2.4}$$

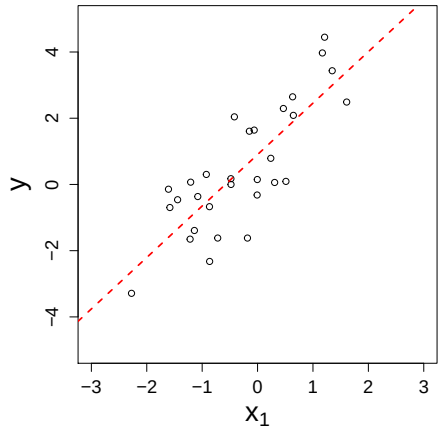
där σ_1 och σ_2 ger storleken på mätfelen i respektive variabel. Brusen har alltså väntevärde noll och genereras från en normalfördelnig, vilken är symmetrisk och har en stor del av massan koncentrerad nära väntevärdet. Ett normalfördelat brus är därför lämpligt för att beskriva brus där man i genomsnitt mäter data korrekt och det finns risk för fel med både positivt och negativt tecken, samt där extrema fel är ovanliga. Vilket stämmer väl med många typer av mätfel, men det är givetvis även möjligt med brus av annan karaktär.

Tabell 1 visar värden på parametrar vid de simuleringar som genomfördes. Stickprovsstorlek, β_1, brus i modellen samt spridning i x_1 valdes så att styrkan för testet att $\beta_1 \neq 0$ modell (3.1.1) blev 90%. Den konstanta termen β_0 sattes lika med noll eftersom den inte har någon inverkan på starkt sambandet mellan prediktor och respons är. Vidare valdes $\mu_{x_1} = \mu_{x_2} = 0$ och $\sigma_{x_2} = \sigma_{x_1} = 1$, vilket alltid är möjligt att åstadkomma i praktiken genom att standardisera data. Simuleringarna genomfördes med olika korrelation mellan x_1 och x_2, båda med en tydlig positiv samvariation. Tecknet på korrelationen spelar i själva verket inte någon roll, eftersom motsatt tecken på korrelationen kan fås genom att byta tecken på den ena variabeln. Slutligen varierades σ_1 men σ_2 var konstant i varje simulering. Simuleringar med olika värden för ρ och σ_2 genomfördes.

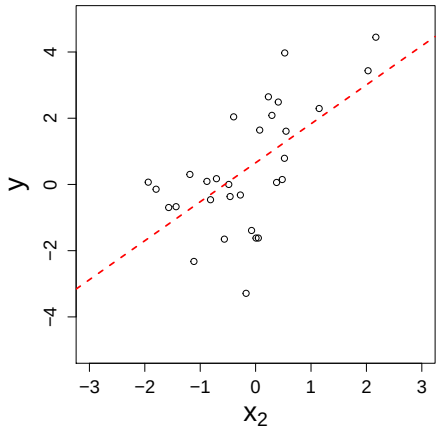
Tabell 1: Värden för parametrar vid simulering. σ_1 varierades mellan 0 och 1.

n	β_0	β_1	σ_y	μ_{x_1}	μ_{x_2}	σ_{x_1}	σ_{x_2}	ρ	σ_2	σ_1
30	0	1	1.5	0	0	1	1	0.6 och 0.8	0.2 och 0.5	0 - 1

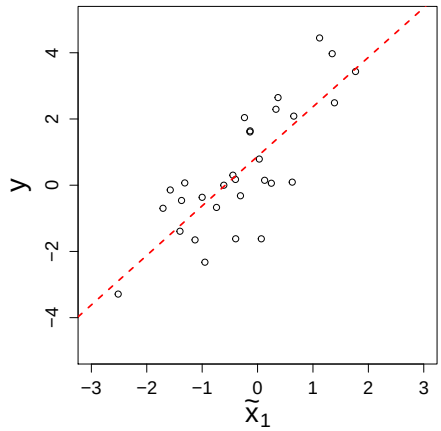
I figur 13 visas relationen mellan prediktorer och respons för ett stickprov med parametrar valda enligt tabell 1 med $\rho = 0.8$. Det syns ett tydligt samband mellan x_1 och y i figur 13a, men relationen mellan x_2 och y i figur 13b är något svagare. Detta eftersom x_2 egentligen inte har någon inverkan på y men ändå visar viss association på grund av dess samvariation med x_1. Sambandet mellan den sanna variabeln x_1 och den observerade variabeln $\tilde{x}_1$ för samma stickprov visas i figur 13d. Mätfelen är relativt små och relationen mellan x_1 och y beskrivs väl av $\tilde{x}_1$, vilket ses eftersom figur 13a och 13c inte visar några stora skillnader. I figur 13e visas relationen mellan $\tilde{x}_1$ och y för ett större mätfel i x_1. Lutningen för regressionslinjen har avtagit kraftigt jämfört med situationen med ett litet mätfel. Detta eftersom brus i observationer ger en systematisk underskattning av lutningen, vilket diskuteras vidare i kapitel 4 och appendix A. Det framgår av variationen kring linjen i figur 13f att det råder stor osäkerhet vid mätning av x_1.



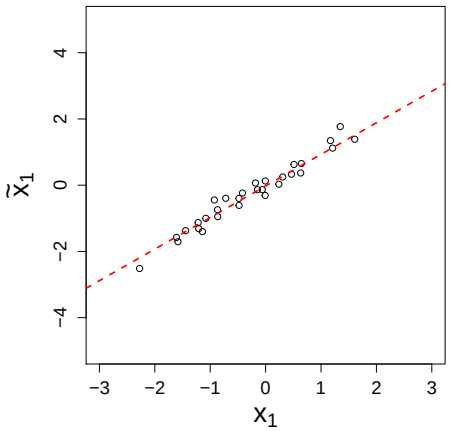
(a) Utfallsvariabel y mot orsakande variabel x_1.



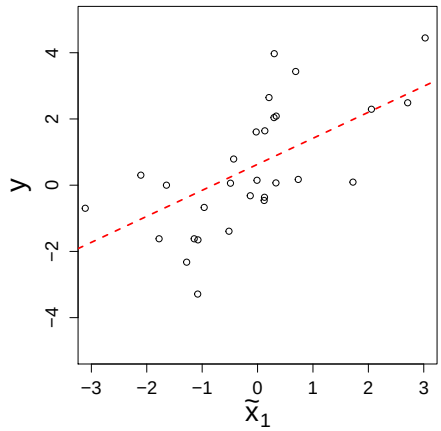
(b) y mot icke-orsakande variabel x_2.



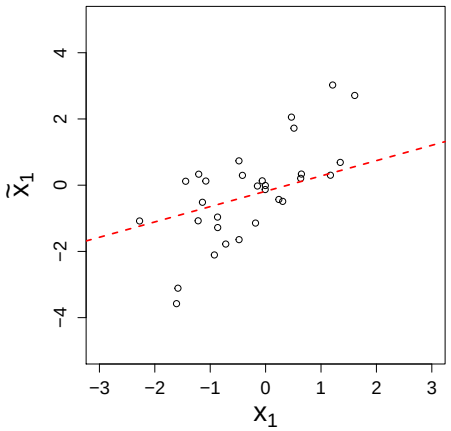
(c) y mot observerad variabel $\tilde{x}_1$. Varians av brus ges av $\sigma_1 = 0.2$.



(d) Observerad variabel $\tilde{x}_1$ mot sann variabel x_1. Varians av brus ges av $\sigma_1 = 0.2$.



(e) y mot $\tilde{x}_1$ med ökad varians av brus, $\sigma_1 = 1$.



(f) $\tilde{x}_1$ mot x_1 med ökad varians av brus, $\sigma_1 = 1$.

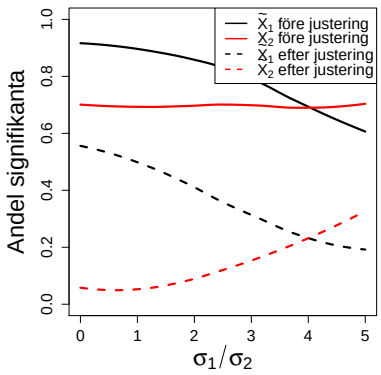
Figur 13: Utfallsvariabel y mot prediktorer och observationer av dessa med brus. Sambandet mellan prediktor och respons blir mindre tydligt då bruset ökar.

Med situationen som förklarats ovan som utgångspunkt simulerades ett stort antal stickprov då bruset i observationerna av den orsakande variabeln x_1 varierades. Övriga parametrar inklusive brus i den icke-orsakande variabeln x_2 hölls konstanta. Resultatet av simuleringarna visas i figur 14. Skalan på x-axeln ges av förhållandet mellan standardavvikelserna i mätfelen för x_1 och x_2. I bilderna till vänster visas hur stor andel av de simulerade stickproven $\tilde{x}_1$ blev signifikant före justering (svart heldragen linje) och efter justering (svart streckad linje), samt andelen då $\tilde{x}_2$ blev signifikant före justering (röd heldragen linje) och efter justering (röd streckad linje). I bilderna till höger visas andel signifikanta efter justering enligt de fyra fall som diskuterades i kapitel 2.4. Här visas alltså andel simulerade stickprov då endast $\tilde{x}_1$ (svart), endast $\tilde{x}_2$ (röd), båda (grön) eller ingen (blå) blev signifikant i den justerade modellen.

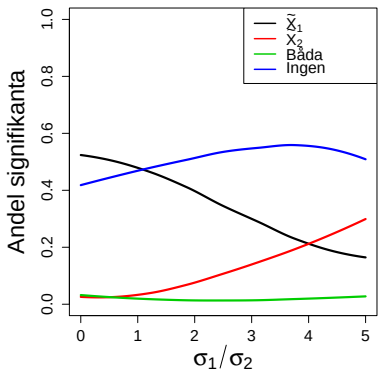
I de fall som visas i figur 14a och 14b är x_1 och x_2 starkt korrelerade och mätfelet i x_2 relativt litet. Figur 14a visar andel signifikanta före och efter justering för båda variablerna. Vid $\sigma_1/\sigma_2 = 0$ mäts x_1 med full precision. Den heldragna svarta linjen visar då att $\tilde{x}_1$ är signifikant i 90% av fallen, vilket stämmer med vald styrka. Typexempel för ett stickprov i denna punkt visades i figur 13a och 13b. Linjen avtar sedan, till en följd av ökande brus i x_1. Den streckade svarta linjen visar hur ofta $\tilde{x}_1$ blir signifikant efter justering. Linjen går betydligt lägre än den heldragna, vilket beror på att den starka korrelationen mellan x_1 och x_2 gör det svårt att avgöra vilken av prediktorerna som bäst förklarar y.

De röda linjerna visar hur ofta $\tilde{x}_2$ blir signifikant. Den heldragna röda linjen är nästan konstant eftersom inga faktorer som påverkar $\tilde{x}_2$ varieras. Att linjen inte är helt konstant beror på slumpmässig variation i de stickptov som genererats. Vid små mätfel lyckas justeringen väl och $\tilde{x}_2$ blir bara signifikant ca 5% av fallen, vilket är förväntat för den sanna nollhypotesen att $\beta_2 = 0$ eftersom den valda signifikansnivån $\alpha = 0.05$. I takt med att bruset i x_1 ökar blir $\tilde{x}_2$ oftare signifikant. Då $\sigma_1/\sigma_2 = 2.5$ är andelen ca 10%, vilket är dubbelt så många som vid situationen utan mätfel i x_1. Ännu större brus i x_1 resulterar i att $\tilde{x}_2$ blir signifikant oftare än $\tilde{x}_1$, och därmed ger intryck av att vara den orsakande variabeln.

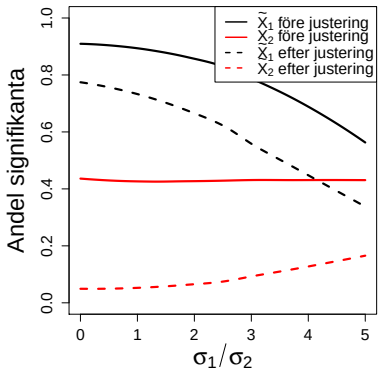
Figur 14b visar resultat av samma simulering efter justering. Då $\sigma_1/\sigma_2 = 0$ resulterar justeringen oftast i att endast $\tilde{x}_1$ (svart) eller ingen av variablerna (blå) blir signifikanta. I fallen då ingen blir signifikant gäller att F-statistikan, som testar om variablerna tillsammans förklarar y, är signifikant. Stickprov då bara $\tilde{x}_2$ (röd) eller båda variablerna (grön) blir signifikanta är sällsynta. För de gröna och blåa linjerna sker inga stora förändringar då bruset i x_1 ökar. Liksom i figur 14a framgår att sambandet mellan $\tilde{x}_1$ och y blir allt svagare då bruset ökar, vilket leder till att $\tilde{x}_2$ oftare blir signifikant.



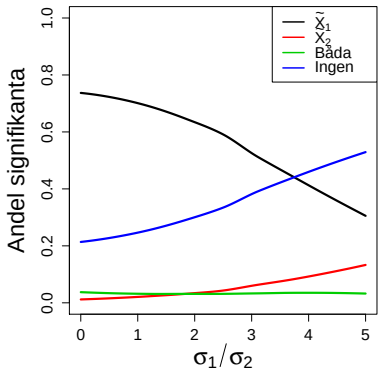
(a) Stark korrelation mellan x_1 och x_2: $\rho = 0.8$. Litet mätfel i x_2: $\sigma_2 = 0.2$.



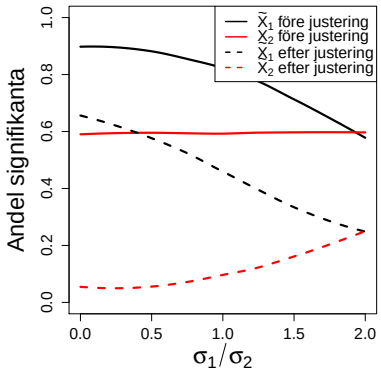
(b) Stark korrelation mellan x_1 och x_2: $\rho = 0.8$. Litet mätfel i x_2: $\sigma_2 = 0.2$.



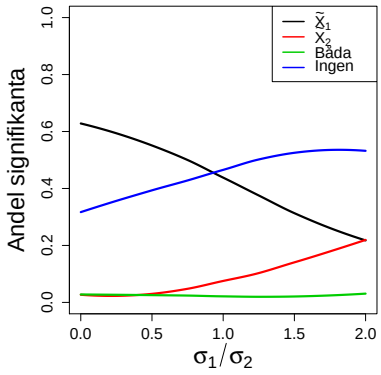
(c) Svagare korrelation mellan x_1 och x_2: $\rho = 0.6$. Litet mätfel i x_2: $\sigma_2 = 0.2$.



(d) Svagare korrelation mellan x_1 och x_2: $\rho = 0.6$. Litet mätfel i x_2: $\sigma_2 = 0.2$.



(e) Stark korrelation mellan x_1 och x_2: $\rho = 0.8$. Stort mätfel i x_2: $\sigma_2 = 0.5$.



(f) Stark korrelation mellan x_1 och x_2: $\rho = 0.8$. Stort mätfel i x_2: $\sigma_2 = 0.5$.

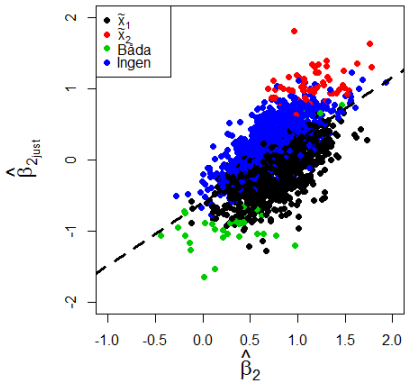
Figur 14: Resultat av simulering. Till vänster: andel fall $\tilde{x}_1$ (svart) respektive $\tilde{x}_2$ (röd) blev signifikanta före (heldragna) och efter justering (streckade). Till höger: andel signifikanta efter justering för fyra olika fall: endast $\tilde{x}_1$ (svart), endast $\tilde{x}_2$ (röd), båda (grön) eller ingen (blå) av variablerna blev signifikanta. Simuleringar genomfördes med olika korrelation mellan prediktorer och olika brus i x_2.

Även situationer då sambandet mellan prediktorerna var svagare, samt då mätfelet i x_2 var större, simulerades. Resultaten av simuleringarna illustreras i 14c-14f. Observera att de svarta heldragna linjerna bilderna till vänster ser likadana ut i samtliga fall. Detta eftersom endast faktorer som påverkar $\tilde{x}_2$ har varierats mellan simuleringarna. Märk även att skalan på x-axeln i figur 14e och 14f skiljer sig från de övriga. Mätfelet i x_1 har varierats på samma sätt som tidigare, men mätfelet i x_2 är större.

I figur 14c och 14d visas resultaten för $\rho = 0.6$, dvs. för ett svagare samband mellan x_1 och x_2. Ett svagare samband mellan prediktorerna leder till att även sambandet mellan $\tilde{x}_2$ och y blir svagare. I bilderna syns det genom att $\tilde{x}_1$ oftare blir signifikant efter justering, och att $\tilde{x}_2$ mer sällan blir signifikant. Risken att dra den felaktiga slutsatsen att x_2 har en inverkan på y är därför liten. Situationer med svårigheter att avgöra vilken prediktor som bäst förklarar y är sällsynta vid små mätfel. Denna typ av problem ökar dock då osäkerheten i x_1 blir större.

Simuleringar med ett stort mätfel i x_2 visas i figur 14e och 14f. Situationen påminner om det ursprungliga fallet, men det ökade bruset i den icke-orsakande variabeln gör att risken för felaktiga slutsatser är något mindre.

Ett annat sätt att illustrera resultaten ges i figur 15. Här visas förhållandet mellan skattningarna av β_2 före och efter justering. Varje punkt representerar ett stickprov från de simuleringar som genomfördes och punkterna markeras med olika färg i enlighet med bilderna till höger i figur 14. En regressionslinje har anpassats efter de punkter som genererats, vars lutning är något mindre än ett. Skattningen av β_2 kommer alltså i regel att vara mindre efter justering än i den modell som endast inkluderar $\tilde{x}_2$. För punkter under regressionslinjen har β_2 skattats lägre efter justering än vad det genomsnittliga sambandet visar, medan punkter ovanför linjen skattats högre än förväntat. I det fall som visas är prediktorerna starkt korrelerade med lika stora mätfel. Majoriteten av fallen visar att $\tilde{x}_1$ blir signifikant, eller att ingen blir det. Den skattade regressionslinjen utgör en tydlig gräns mellan de båda fallen. Situationer då endast $\tilde{x}_2$ blir signifikant uppstår när $\hat{\beta}_2$ är stor efter justering, vilket kan inträffa även om skattningen innan justering inte var så stor. När båda variablerna blir signifikanta är $\hat{\beta}_2 < 0$ i den justerade modellen. Eftersom korrelationen mellan x_1 och y samt mellan x_1 och x_2 är positiv är det orimligt att x_2 enbart via samvariationen med x_1 har en negativ inverkan på y. Därför leder stora negativa skattningar av β_2 ofta till signifikans i båda variablerna. I själva verket hör detta samman med estimeringsproblem som uppstår då korrelationen mellan prediktorerna är stark, vilket gör att skattningarna blir instabila och till och med kan variera i tecken.



Figur 15: Skattningar av β_2 före och efter justering för stark korrelation mellan prediktorer och litet brus i x_2: $\rho = 0.8$, $\sigma_1 = \sigma_2 = 0.2$. Varje punkt representerar ett simulerat stickprov markerad med färg enligt fyra olika fall: endast $\tilde{x}_1$ (svart), endast $\tilde{x}_2$ (röd), båda (grön) eller ingen (blå) av variablerna blev signifikanta. Streckad linje visar regression utförd med $\hat{\beta}_{2_{just}}$ mot $\hat{\beta}_2$. De stickprov där den skattade lutningskoefficienten avtar kraftigt i storlek efter justering har justeringen lyckats väl.

3.3 Simulering med logistisk regression

Föregående delkapitel behandlades modeller med kontinuerlig utfallsvariabel. Vi skall nu studera motsvarande problem för logistiska regressionsmodeller, vilka används då y är binär. I de modeller som presenterades i 3.1 ersätts y med

$$log(\frac{p}{1-p}) \tag{3.3.1}$$

där p betecknar sannolikheten att $y = 1$. För en given observation av x_1 genereras y från en Bernoulli-fördelning där sannolikheten är beoende av x_1 enligt:

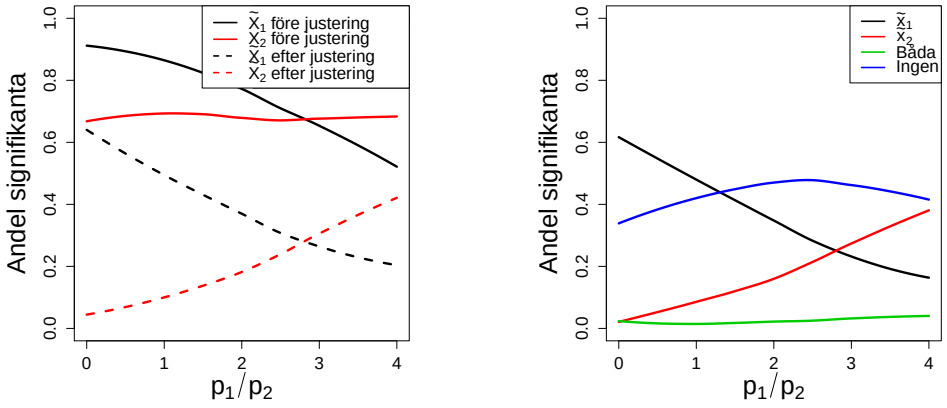
$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1)}} \tag{3.3.2}$$

Det är möjligt att två observationer med samma värde i x_1 får olika värden i y, vilket på ett naturligt sätt ger upphov till slumpmässiga avvikelser från modellen.

För att skapa ytterligare kontrast till de simuleringar som genomförts tidigare i detta kapitel så antas även prediktorerna vara binära. I en population skulle dessa exempelvis kunna stå för icke-rökare/rökare, man/kvinna etc. Här är x_1 och x_2 är relaterade, vilket innebär att om x_1 är känd, så går det med viss sannolikhet att uttala sig om utfallet i x_2. Det finns dock vissa avvikelser mellan variablerna, och hur stor sannolikheten att förutsäga den ena givet den andra avgör hur starkt relaterade variablerna är. Mätfelen i prediktorerna uppstår i denna situation av att vissa observationer klassificeras i fel kategori. Istället för σ_1 och σ_2 används här p_1 respektive p_2 för att beteckna sannolikheten att en observation av x_1 eller x_2 felklassificeras. Även i detta avsnitt skrivs de observerade variablerna $\tilde{x}_1$ och $\tilde{x}_2$. Stickprovsstorlek, β_1 samt proportionerna i grupperna i x_1 valdes så att styrkan för testet $\beta_1 \neq 0$ blev 90% i den univariata modellen med endast x_1, då denna variabel observerades utan brus. Avvikelsen mellan x_1 och x_2 sattes till 10%, vilket svarar mot korrelation 0.8. Sannolikheten för felklassificering i x_2, betecknad p_2, valdes till 0.05 och sannolikheten att felklassificera x_1 varierade i simuleringen.

Tabell 2: Värden för parametrar vid simulering. p_1 varierades mellan 0 och 0.2.

n	β_0	β_1	ρ	p_2	p_1
200	0	1	0.8	0.05	0 - 0.2



Figur 16: Simulering med logistisk regression. Till vänster: andelen signifikanta x_1 och x_2 före och efter justering. Till höger: Andel signifikanta efter justering för fyra olika fall. Resultaten påminner om de som presenterats för linjär regression.

Resultatet av simuleringen visas i figur 16. x-axeln utgör här förhållandet mellan antalet felklassificerade x_1 och antalet felklassificerade x_2. Den svarta linjen i bilden till vänster ser ut

som i figur 14. De statistiska slutsater som kan dras om relationen mellan $\tilde{x}_1$ och y stämmer alltså med tidigare simuleringar, vilket styrker jämförbarheten mellan dem. Även här är risken för felaktiga statistiska slutsatser små vid litet brus. Ökad felklassificering leder dock ganska snabbt till svårigheter att avgöra vilken variabel som bäst förklarar utfallsvariabeln, samt att $\tilde{x}_2$ ofta blir signifikant.

Simuleringen bygger på en logistisk regressionsmodell där samtliga variabler är dikotoma. Motsvarande skulle även kunna genomföras för kontinuerliga variabler. Det händer även att man har mätningar från kontinuerliga variabler men väljer att kategorisera observationer. Man kan exempelvis välja att klassificera individer efter olika viktklasser enligt 70-80kg, 80-90kg etc. Här är det möjligt att finheten i de kategorier som används bidrar till att de statistiska slutsatserna påverkas. Om x_1 och x_2 är kontinuerliga och korrelerade och x_1 delas in i två grupper, men x_2 delas in i fler kategorier, finns risk att den flexibilitet som fler kategorier medför, gör att x_2 blir signifikant istället för x_1. Även sådana fall simulerades, vilket gav liknande resultat som i figur 16 och presenteras därför ej mer ingående.

3.4 Simulering då modell och data ej överensstämmer

Simuleringar i avsnitt 3.2 och 3.3 belyser problem då prediktorer observeras med mätfel. I detta avsnitt visas att liknande problem kan uppstå även om samtliga variabler mäts med stor noggrannhet. Antag att y orsakas av x_1 enligt

$$y = \beta_0 + \beta_1 x_1 + \gamma x_1^2 + \epsilon_y \tag{3.4.1}$$

där γ bestämmer storleken på den kvadratiska termen. Antag vidare att

$$x_2 = \alpha_0 + \alpha_1 x_1^2 + \delta \tag{3.4.2}$$

för några koefficienter α_0, α_1 och brus $\delta \sim N(0, \sigma_\delta)$. Liksom tidigare undersöks sambandet mellan x_2 och y. Den modell som ansätts vid justering med x_1 är linjär i båda variablerna:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_y \tag{3.4.3}$$

Det finns enligt (3.4.1) en kvadratisk term i relationen mellan x_1 och y vilken inte är inkluderad i den studerade modellen men som fångas upp av x_2. Om det är uppenbart att y inte endast beror linjärt av x_1 bör man givetvis genomföra lämplig variabeltransformation eller inkludera x_1-termer av högre grad. Detta är dock inte alltid lätt i praktiken och icke-linjära förhållanden kan vara svåra att skilja från brus och slumpmässigt avvikande observationer.

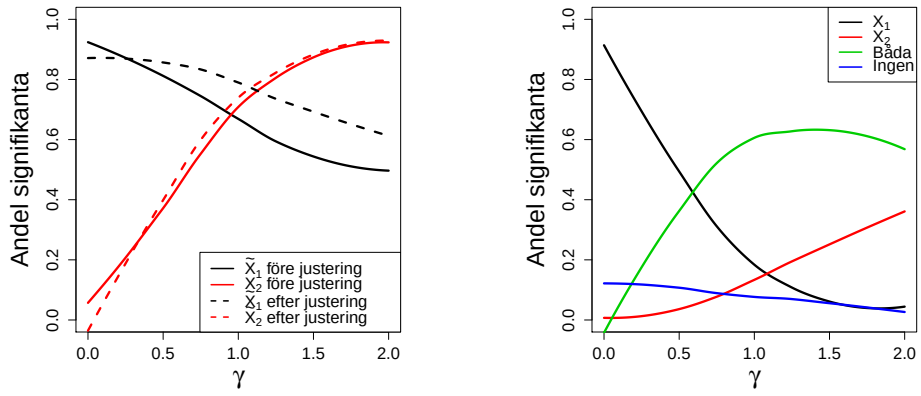
Tabell 3 visar värden på parametrar för denna simulering. Stickprovsstorlek, β_1, spridning i x_1 och brus i modellen valdes så att ett signifikant samband mellan x_1 och y identifierades i 90% av fallen då $\gamma = 0$.

Tabell 3: Värden för parametrar vid simulering. γ varierades mellan 0 och 2.

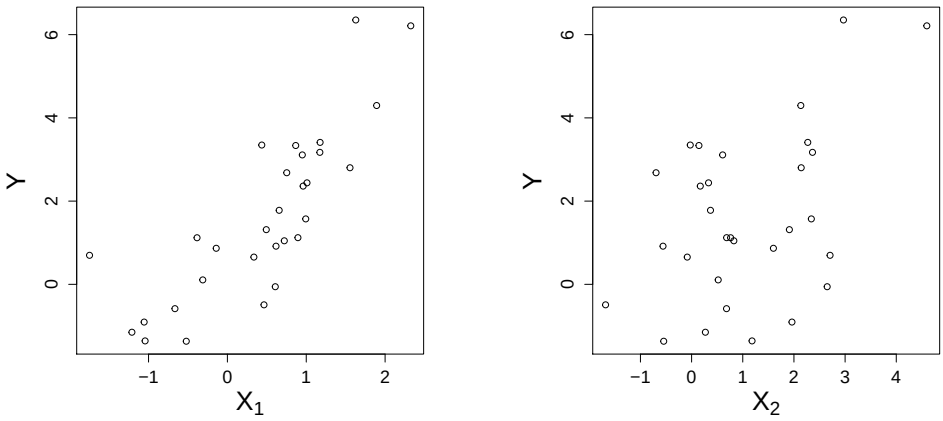
n	β_0	β_1	α_0	α_1	σ_y	σ_δ	σ_{x_1}	γ
30	0	1	0	1	1	1	1	0 - 2

Figur 17 visar resultatet för simuleringar vid situationen som presenterades ovan. Även denna gång gäller att andelen stickprov då x_1 är signifikant i den univariata modellen stämmer väl med tidigare simuleringar. I övrigt ser man att vissa skillnader gentemot tidigare simuleringar blir tydliga. En avvikelse i den vänstra figuren är att både x_1 och x_2 oftare blir signifikanta efter justering än före justering. Detta beror på att variablerna tillsammans förklarar y bättre än vad de gör var för sig. En annan skillnad är att x_2 ökar snabbt i andel signifikanta, både före och efter justering, då den kvadratiska termen i modell (3.4.1) ökar i storlek. I den högra bilden framgår det även att det är en stor andel där båda variablerna visar ett signifikant samband med y. Detta beror på att den linjära x_1-termen och den kvadratiska x_2-termen förklarar olika delar av y. Det behöver dock inte vara självklart att ett kvadratiskt förhållande mellan x_1 och y existerar även om γ är stor, vilket figur 18 visar. Slutligen blir x_2 oftare signifikant än x_1 då $\gamma > 1$, dvs. då den kvadratiska termen är större än den linjära. x_2 har alltså ingen inverkan på y och båda prediktorerna observeras utan brus, men bristerna i

den modell som ansätts gör att x_2 verkar ha en inverkan på y. I verkligheten råder ofta mer komplicerade samband och simuleringen belyser vikten av att ansätta en modell som på ett korrekt sätt beskriver aktuell data.



Figur 17: Till vänster: andelen signifikanta x_1 och x_2 före och efter justering. Till höger: Andel signifikanta efter justering. Båda variablerna blir ofta signifikanta även efter justering, eftersom den linjära x_1-termen och den kvadratiska x_2-termen förklarar olika delar av y.



Figur 18: Utfallsvariabel y mot prediktorer. $y = \beta_1 x_1 + \gamma x_1^2$ för $\beta_1 = \gamma = 1$. y ser ut att bero linjärt av både x_1 och x_2. Det är därmed inte självklart att en kvadratisk x_1-term borde ingå i modellen.

4 Osäkerhet i variabler - teori för skattningar vid mätfel i prediktorer

I kapitel 3 simulerades situationer där problem med att skatta och dra slutsatser om koefficienter i olika regressionsmodeller blev uppenbara. Detta beror i själva verket på att grundläggande antaganden för minstakvadratmetoden inte håller då det finns mätfel i data. Orsakerna till detta skall här beskrivas teoretiskt.

4.1 Mätfel i utfallsvariabeln

Antag först att utfallsvariabeln y observeras med mätfel δ_y så att

$$\tilde{y} = y + \delta_y \tag{4.1.1}$$

där mätfelet $\delta_y \sim N(0, \sigma_\delta)$ är likafördelat och oberoende av y för olika observationer. Antag vidare att man vill beskriva y med en prediktor x_1 enligt

$$y = \beta_0 + \beta_1 x_1 + \epsilon_y \tag{4.1.2}$$

för $\epsilon_y \sim N(0, \sigma_y)$. Eftersom man inte har tillgång till y, utan den observerade variablen $\tilde{y}$, ges den studerade modellen av:

$$\tilde{y} = \beta_0 + \beta_1 x + \epsilon_y \quad \Leftrightarrow \quad y = \beta_0 + \beta_1 x_1 + \epsilon_y - \delta_y = \beta_0 + \beta_1 x_1 + \gamma_y \tag{4.1.3}$$

Eftersom både bruset i modellen och mätfelet i y är oberoende av x_1 är även det nya bruset γ_y oberoende av x_1. Dessutom är $\gamma_y \sim N(0, \sigma_\gamma)$ eftersom summan av två normalfördelade stokastiska variabler med väntevärde noll är normalfördelad med väntevärde noll. Mätfelet vägs alltså in i bruset i modellen och antaganden för minstakvadratmetoden är uppfyllda. Detta innebär att mätfel i utfallsvariabeln inte orsakar några problem om dessa har väntevärde 0 och är oberoende av varandra samt av x_1.

4.2 Mätfel i prediktor vid enkel linjär regression

Antag att vi har en enkel linjär regressionsmodell

$$y = \beta_0 + \beta_1 x_1 + \epsilon_y \tag{4.2.1}$$

med $\epsilon_y \sim N(0, \sigma_y)$ och att x_1 observeras med mätfel

$$\tilde{x}_1 = x_1 + \epsilon_1 \tag{4.2.2}$$

där $\tilde{x}_1$ betecknar den observerade variabeln och $\epsilon_1 \sim N(0, \sigma_1)$. Genom att lösa ut x_1 i uttryck (4.2.2) och substituera i modell (4.2.1) fås följande modell där y förklaras av $\tilde{x}_1$:

$$y = \beta_0 + \beta_1 \tilde{x}_1 + \epsilon_y - \beta_1 \epsilon_1 \tag{4.2.3}$$

där $\epsilon_y - \beta_1 \epsilon_1$ är det nya modellbruset. Minstakvadratmetoden ger skattningar av β_1 som inte är väntevärdesriktiga

$$E[\hat{\beta}_1] = \beta_1 \frac{\sigma_{x_1}^2}{\sigma_{x_1}^2 + \sigma_1^2} = \beta_1 \frac{1}{1 + \sigma_1^2/\sigma_{x_1}^2} \tag{4.2.4}$$

vilket beror på att bruset inte är oberoende av $\tilde{x}_1$:

$$Cov[\tilde{x}_1, \epsilon_y - \beta_1 \epsilon_1] = -\beta_1 \sigma_1^2 \tag{4.2.5}$$

Det visar sig även vara omöjligt att skatta regressionskoefficienterna med maximum-likelihoodmetoden utan att göra ytterligare antaganden [4]. I vissa fall kan man dock bortse från de problem som orsakas av mätfel i prediktorer, exempelvis om mätfelen är små i relation till spridningen i data eller om modellen skall användas till prediktion och mätfelen förväntas vara av samma storleksordning vid framtida observationer. Mer detaljerade härledningar till uttrycken ovan ges i appendix A.

4.3 Mätfel i prediktorer vid multipel linjär regression

Vi betraktar den multipla linjära regressionsmodellen

$$\mathbf{y} = \mathbf{X}\beta + \epsilon_{\mathbf{y}} \tag{4.3.1}$$

där $\mathbf{y}, \beta$ och ϵ är vektorer och $\mathbf{X}$ designmatrisen med observationer av de ingående prediktorerna.

I fallet med en prediktor visade vi att väntevärdet av $\hat{\beta}_1$ avtar mot noll då det finns brus i mätning av x_1. Med flera prediktorer är det dock svårt att ge en generell regel då väntevärdesfelet beror på korrelationen mellan prediktorerna och storleken på de ingående mätfelen samt hur de är korrelerade.

Låt matrisen $\mathbf{E}$ beteckna mätfelen i prediktorerna. Element $\mathbf{E}_{ij}$ ger mätfelet för observation i av prediktor[2] $j+1$. Låt $\tilde{\mathbf{X}}$ beteckna designmatrisen för de observerade variablerna enligt:

$$\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{E} \tag{4.3.2}$$

Med minstakvadratmetoden skattas alltså

$$\tilde{\beta} = (\tilde{\mathbf{X}}^{\mathbf{T}}\tilde{\mathbf{X}})^{-\mathbf{1}}\tilde{\mathbf{X}}\mathbf{y} \tag{4.3.3}$$

om det finns mätfel i data, men i själva verket är man intresserad av

$$\hat{\beta} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}\mathbf{y}. \tag{4.3.4}$$

Antag nu att mätfelen i varje variabel är oberoende av varandra, att mätfelen för olika variabler är oberoende och att mätfelen är oberoende av $\epsilon_{\mathbf{y}}$, samt att de har medelvärde 0 och varians σ_i^2, för $i = 1, \ldots, p-1$ där p-1 är antalet prediktorer. Då ges väntevärdet för β av

$$E[\hat{\beta}] = \beta - (n-p)(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{S}\beta \tag{4.3.5}$$

där

$$\mathbf{S} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{p-1}^2 \end{pmatrix} \tag{4.3.6}$$

och skattningen av β-vektorn är alltså inte väntevärdesriktig om det finns mätfel i prediktorer. $\mathbf{S}_{1,1} = 0$ svarar mot β_0 som inte påverkas av mätfel. Matrisen $\mathbf{S}$ är sällan känd men kan approximeras genom att uppskatta brusen i variablerna, vilka i vissa kan fall fås genom upprepade observationer av samma individ. Genom att vidare approximera $\mathbf{X}^{\mathbf{T}}\mathbf{X} \approx \tilde{\mathbf{X}}^{\mathbf{T}}\tilde{\mathbf{X}}$ kan en uppskattning av väntevärdesfelet fås. Den vanliga skattningen för variansen av $\hat{\beta}$

$$Var[\hat{\beta}] \approx \hat{\sigma}_y^2(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}} \tag{4.3.7}$$

stämmer väl om mätfelen inte är alltför stora. Även här approximeras $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ med $\tilde{\mathbf{X}}^{\mathbf{T}}\tilde{\mathbf{X}}$. Härledningar till ovanstående uttryck ges av Hodges [5].

I praktiken är det dock inte alltid möjligt att få en uppskattning av mätfelen i variablerna. Det kan vara kostsamt och ibland omöjligt att skatta brus genom upprepade observationer. Situationer där en latent variabel ersätts med en proxy-variabel försvårar det hela ytterligare eftersom felet inte beror på mätfel utan på osäkerhet i relationen mellan variablerna. Det gäller även att approximationen $\mathbf{X}^{\mathbf{T}}\mathbf{X} \approx \tilde{\mathbf{X}}^{\mathbf{T}}\tilde{\mathbf{X}}$ stämmer dåligt för stora brus, vilket ger dåliga uppskattningar av väntevärdesfelet.

Under förutsättning att inga variabler observeras med mätfel ges den multivariata normalfördelningen för skattningarna av regressionskoefficienterna av

$$\hat{\beta} \sim N(\beta, \mathbf{\Sigma}). \tag{4.3.8}$$

[2] $j+1$ eftersom den första kolonnen svarar mot β_0 och mätfelen i den första prediktorn står i andra kolonnen.

medan mätfel i prediktorer enligt ovanstående diskussion ger följande fördelning:

$$\hat{\beta} \approx N\left(\beta - (n-p)(\tilde{\mathbf{X}}^{\mathbf{T}}\tilde{\mathbf{X}})^{-1}\mathbf{S}\beta, \tilde{\mathbf{\Sigma}}\right) \tag{4.3.9}$$

där kovariansmatriserna estimeras med

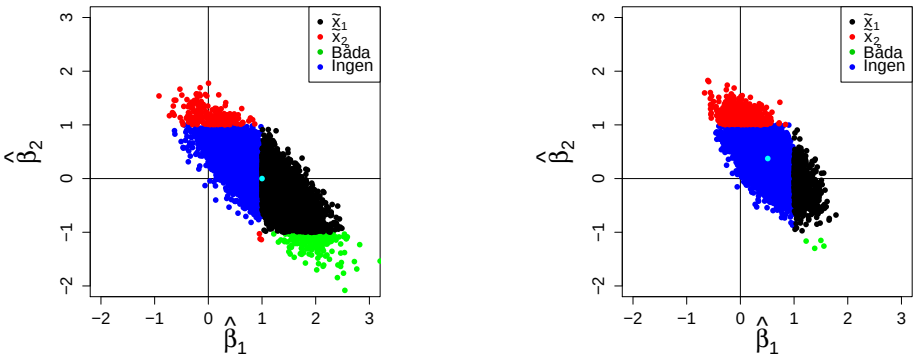
$$\hat{\mathbf{\Sigma}} = \sigma_y^2(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1} \tag{4.3.10}$$

respektive

$$\hat{\tilde{\mathbf{\Sigma}}} = \sigma_y^2(\tilde{\mathbf{X}}^{\mathbf{T}}\tilde{\mathbf{X}})^{-1}. \tag{4.3.11}$$

Fördelningarna för $\hat{\beta}$ i (4.3.8) och (4.3.9) illustreras i figur 19. De observerade variablerna betecknas som tidigare $\tilde{x}_1$ och $\tilde{x}_2$. Även här är x_1 den orsakande variabeln och x_2 den icke-orsakande men samvarierande variabeln. För fördelningen med brus gäller att $\sigma_1 = 0.6, \sigma_2 = 0.2$. Punkterna markeras med olika färger i enlighet med figurer i tidigare kapitel. De turkosa punkterna markerar medelvärden av β-vektorerna. Båda fördelningarna visar en tydlig negativ korrelation mellan $\hat{\beta}_1$ och $\hat{\beta}_2$. För stora skattningar av β_1 är alltså $\hat{\beta}_2$ liten eller har motsatt tecken, och vice versa. I fördelningen utan mätfel dominerar dem gångerna då $\tilde{x}_1$ är signifikant, samt då ingen är signifikant. Ganska vanligt förekommande är dock fall då $\tilde{x}_2$ är signifikant.

Det syns en tydlig skillnad i väntevärdena i de båda fördelningarna. I fördelningen med mätfel underskattas β_1, och β_2 överskattas. Det framgår även att signifikansen av skattningarna skiljer sig åt. Med mätfel är det med sällsynt att endast $\tilde{x}_1$ är signifikant eller att båda är signifikanta. Fallen där ingen är signifikant, samt där endast x_2 är signifikant har ökat markant. I den ideala situationen utan mätfel är det möjligt att felaktiga slutsatser dras, men situationen är tydligt försämrad då det finns mätfel.



Figur 19: Fördelning av $[\hat{\beta}_1, \hat{\beta}_2]$ utan brus enligt (4.3.8) till vänster och med brus enligt (4.3.9) till höger. Även i fördelningen utan brus i prediktorer händer det att $\tilde{x}_2$ blir signifikant efter justering. Situationen förvärras dock då det finns brus i x_1, eftersom skattningarna av regressionskoefficienterna ej längre är väntevärdesriktiga. Detta leder till fler fall där $\tilde{x}_2$ blir signifikant.

4.4 Metod för att hantera mätfel i prediktorer

Eftersom väntevärdesfelet för skattningarna av regressionskoefficienterna ges av

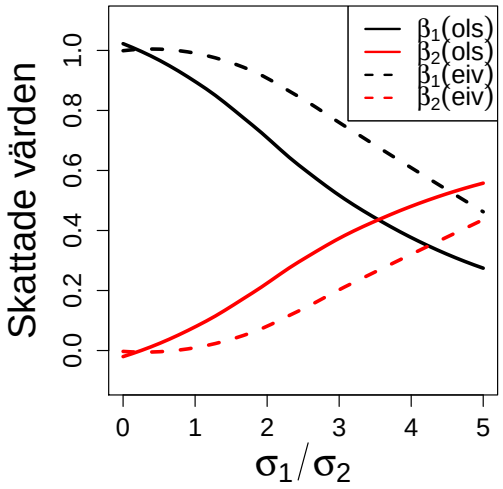
$$-(n-p)(\mathbf{X^T X})^{-1}\mathbf{S}\beta \tag{4.4.1}$$

är det möjligt att få väntevärdesriktiga skattningar då $\mathbf{S}$ är känd. Genom att addera ovanstående till de skattningar som fås med minsta kvadratmetoden erhålls följande:

$$\hat{\beta}^{(eiv)} = \hat{\beta}^{(ols)} + (n-p)(\mathbf{X^T X})^{-1}\mathbf{S}\hat{\beta}^{(ols)} \tag{4.4.2}$$

givet en observerad designmatris $\mathbf{X}$. I praktiken approximeras $\mathbf{X^T X}$ med $\tilde{\mathbf{X}}^\mathbf{T}\tilde{\mathbf{X}}$, vilken beskriver relationen mellan prediktorerna. OLS står för 'ordinary least squares' och används för att beteckna minstakvadratskattningen. EIV står för 'errors in variables', vilket är ett samlingsnamn för teori om regression med mätfel i prediktorer. Skattningen som ges i (4.4.2) kallas fortsättningsvis för EIV-skattningen av β eller EIV-metoden. Flera andra metoder finns, som tar hänsyn till mätfel i prediktorer vid skattning av regressionskoefficienterna. Några sådana diskuteras för regressionsmodeller med en prediktor i appendix B.

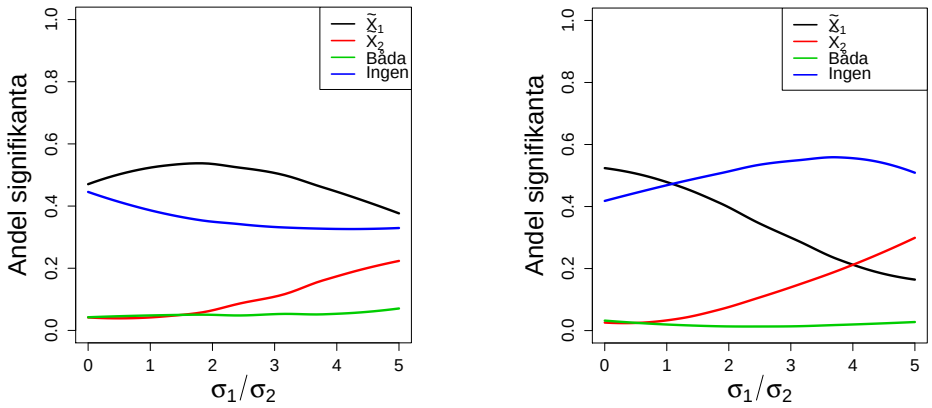
En jämförelse mellan EIV-metoden och minstakvadratmetoden gjordes med hjälp av simuleringar och illustreras i figur 20 och 21. Situationen är som i kapitel 3.2, med $\rho = 0.8$ och $\sigma_2 = 0.2$. Brusen i prediktorerna antogs vara kända. De streckade linjerna i figur 20 visar genomsnittliga EIV-skattningar av $\hat{\beta}_1$ och $\hat{\beta}_2$ vid simulering. Heldragna linjer visar minstakvadratskattningarna. Redan vid små mätfel uppstår väntevärdesfel med minstakvadratmetoden, medan EIV-metoden ger goda skattningar. Då mätfelet ökar ger även EIV skattningar som inte är väntevärdesriktiga, vilket beror på att approximationen $(\tilde{\mathbf{X}}^\mathbf{T}\tilde{\mathbf{X}}) \approx (\mathbf{X^T X})$ stämmer dåligt.



Figur 20: Skattningar av β_1 och β_2 med minstakvadratmetoden respektive EIV. Med EIV fås skattningar som är väntevärdesriktiga för små brus. För stora brus är EIV-skattningarna inte väntevärdesriktiga, men metoden ger mindre väntevärdesfel än minstakvadratmetoden.

Signifikans av de skattade effekterna efter justering illustreras i figur 21. Till vänster i figuren visas andelen signifikanta när EIV-metoden använts. Till höger ges motsvarande resultat för minstakvadratmetoden. Det framgår att EIV-metoden ger ett bättre resultat med mindre risk för felaktiga slutsatser.

Om man har god kännedom om storleken på brusen i prediktorerna, eller har möjlighet att uppskatta dem, är således (4.4.2) en mycket bra metod att använda sig av. Den ger då bättre resultat än minstakvadratmetoden, eftersom hänsyn tas till brusens inverkan på skattningarna och väntevärdesfelet är mindre. Minstakvadratmetoden lämpar sig bättre då ingen eller liten kännedom om bruset i mätningar existerar.



Figur 21: Justering med EIV (vänster) och minstakvadratmetoden (höger) för stark korrelation mellan prediktorer och litet brus i x_2: $\rho = 0.8$, $\sigma_2 = 0.2$ Vid jämförelse mellan bilderna framgår att EIV ger betydligt bättre resultat än minstakvadratmetoden med avseende på antal lyckade justeringar (svart linje).

5 Diskussion

Rapportens syfte var att belysa hur mätfel i prediktorer påverkar statistiska resultat vid olika typer av regressionsmodeller. Situationer med två samvarierande prediktorer studerades med simuleringar, där endast den ena hade en sann inverkan på utfallsvariabeln.

Först gjordes simuleringar med varierande brus i den orsakande prediktorn. Även korrelation mellan prediktorer och brus i observation av dessa varierades för att kunna dra mer generella slutsatser. Både linjär regression med kontinuerliga variabler samt logistisk regression med dikotoma variabler studerades. Då den orsakande variabeln mättes med full precision lyckades justeringen väl. Ett fåtal gånger blev den icke-orsakande variabeln signifikant, vilket genom slump är möjligt även för en sann nollhypotes. Stark korrelation mellan prediktorer gjorde det svårt att avgöra vilken som bäst förklarar utfallsvariabeln. Vidare bidrog brus i den orsakande variabeln till att sambandet med responsvariabeln blev svagare. Detta ledde i sin tur till att den icke-orsakande variabeln oftare blev signifikant. Även för relativt små mätfel i den orsakande variabeln ökade andelen fall med ofullständig justering. I extrema fall hände det att variabeln som inte hade någon inverkan på utfallsvariabeln oftare blev signifikant än den orsakande variabeln. Vid stark korrelation mellan prediktorer och stort brus i den variabel som har en inverkan på responsvariabeln, bör man vara uppmärksam på att justeringen kan vara ofullständig.

Därefter simulerades situationer där den modell som ansattes mellan orsakande variabel och utfallsvariabel stämde dåligt överrens med data. Bristerna i modellen fångades i detta fall upp av en prediktor utan sann inverkan på utfallsvariabeln. Här var det vanligt att båda prediktorerna blev signifikanta eftersom de förklarade olika delar av utfallsvariabeln. Simuleringen understryker vikten av att alltid kontrollera att ansatt modell på ett korrekt sätt beskriver förhållanden mellan prediktorer och respons.

Resultaten från simuleringarna motiverades sedan med teoretiska resonemang. Det visades att minstakvadratmetoden ger skattningar som inte är väntevärdesriktiga om det finns brus i prediktorer. Orsaken till detta är att modellbruset inte är oberoende av den observerade variabeln, vilket är ett grundläggande antagande för minstakvadratmetoden. I modeller med en prediktor leder det till att skattade effekter underskattar de sanna effekterna. I modeller med flera prediktorer visar det sig vara svårt ge generella förklaringar till hur skattningarna påverkas, då såväl korrelation mellan prediktorer samt storlek på mätfel och relationen mellan dessa spelar in. Man kan dock bortse från mätfel om de är små i relation till spridning i data. Mätfel spelar inte heller stor roll om modellen endast är avsedd för prediktion och mätfelen antas vara av samma storlek vid framtida observationer. Motsvarande teoretiska resonemang är svåra att genomföra för logistiska regressionsmodeller, men simuleringar visade att liknande

problem uppstår.

Slutligen visades att det är möjligt att förbättra de statistiska resultaten något vid god kännedom om bruset i prediktorer. Metoden bygger på kännedom om hur brus påverkar skattningarna av prediktorernas effekter. Det visade sig att skattningar med mindre väntevärdesfel gick att åstadkomma, vilket minskade risken för felaktiga statistiska slutsatser.

Både teori och simuleringar som presenterats bygger på antaganden och förenklingar av verkligheten. Endast ett fåtal fall simuleras, vilka knappast kan sägas ge en representativ bild av verkligheten. Inadekvata modellansatser och brus i variabler behandlades separat, men i tillämpningar är det inte orimligt att båda problemen existerar samtidigt. Vad gäller teori för skattningar vid brus i prediktorer antas exempelvis att brusen är oberoende för olika observationer samt att de är oberoende för olika variabler. I tillämpningar är det möjligt att mer komplicerade samband mellan dessa råder, vilket ytterligare försvårar för statistisk analys. Resultatet begränsas därmed till de något enklare fall som studerats, men liknande problem förväntas uppstå även för mer komplicerade modeller.

En fördjupning i hur osäkerhet i data påverkar icke-linjära och generaliserade linjära regressionsmodeller vore ett intressant ämne för framtida studier. Analytiska resultat i likhet med de som presenterades i kapitel 4 är troligtvis svåra att åstadkomma, då de mer komplicerade modellerna saknar explicita lösningar. Simuleringar kan användas för att studera de problem som uppstår, och brus i data leder troligtvis till att sanna effekter underskattas samt att de variabler med bäst precision gynnas.

Mätfel och brus i variabler är vanligt förekommande i många fall där regressionsanalys används. De statistiska resultaten påverkas av detta, men med god kännedom om problematiken är det möjligt att minimera riskerna för felaktiga statistiska slutsatser. Det är dock långt ifrån alla gånger som det är möjligt att uppskatta och hantera brus enligt de metoder som presenterats, och inte heller de ger perfekta resultat. Om det finns osäkerhet i data bör man därför vara aktsam på att det finns risk för ofullständig justering.

A Teoretiska Härledningar

I följande avsnitt diskuteras metoder som använts i avsnitt 3 och 4 med matematiska härledningar.

A.1 Styrka av test

Styrka av test definieras i kapitel 3. Här ges en en teoretisk bakgrund till hur olika parametrar valts vid simuleringar för att ge önskad styrka. Styrkan av ett test definieras som sannolikheten att förkasta en falsk nollhypotes, som således bör förkastas. En hög styrka för ett test är eftersträvansvärt, eftersom man inte vill godta en falsk hypotes för sanning.

Fortsättningsvis är det styrkan av testet att $\beta_1 \neq 0$ i den univariata modellen $y = \beta_0 + \beta_1 x_1$ som avses och nollhypotesen $H_0 : \beta_1 = 0$ är alltså falsk. När simuleringar utförts har en styrka på 0.9 eftersträvats som utgångspunkt i fallet då inget brus adderats till x_1. Detta för att få jämförbara resultat mellan olika typer av simuleringar. När bruset i x_1 sedan ökat i storlek så minskar styrkan av testet. Standardavfelet för β_1 ges av

$$SE[\hat{\beta}_1] = \sqrt{\frac{\sigma_y^2}{(n-1)\tilde{\sigma}_{x_1}^2}} = \sqrt{\frac{\sigma_y^2}{(n-1)(\sigma_{x_1}^2 + \sigma_1^2)}} \tag{A.1.1}$$

då det finns mätfel i data, där $\tilde{\sigma}_{x_1}^2$ betecknar variansen i den observerade variabeln $\tilde{x}_1$. Enligt nollhypotesen är väntevärdet

$$E[\hat{\beta}_1] = 0 \tag{A.1.2}$$

Utifrån detta kan man beräkna följande teststatistika:

$$T_{obs} = \frac{\hat{\beta}_1 - 0}{SE[\hat{\beta}_1]} \sim t_{n-2} \tag{A.1.3}$$

Låt $T_{\alpha/2}$ vara sådan att $P(|T| > T_{\alpha/2}) = 1 - \alpha$ får för någon signifikansnivå α. Då kan vi skriva sannolikheten att förkasta H_0 när den är sann som:

$$P\left(\frac{\hat{\beta}_1 - 0}{SE[\hat{\beta}_1]} \geq T_{\alpha/2}\right) \tag{A.1.4}$$

Men eftersom nollhypotesen är falsk, det vill säga $\beta_1 \neq 0$, kan vi subtrahera $\beta_1 / SE[\hat{\beta}_1]$ och få sannolikheten att förkasta H_0 när den är falsk enligt

$$P\left(\frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \geq T_{\alpha/2} - \frac{\beta_1}{SE[\hat{\beta}_1]}\right) = 1 - F_{t(n-2)}\left(T_{\alpha/2} - \frac{\beta_1}{SE[\hat{\beta}_1]}\right) = F_{t(n-2)}\left(\frac{\beta_1}{SE[\hat{\beta}_1]} - T_{\alpha/2}\right) \tag{A.1.5}$$

där $F_{t(n-2)}$ betecknar fördelningsfunktionen för en t-fördelning med n-2 frihetsgrader. Ju större arumentet till $F_{t(n-2)}$ är, dvs. uttrycket innanför parentesen, desto närmare 1 blir styrkan. Efter omskrivningar förväntat värde för β_1 vid mätfel, samt uttrycket för SE[$\hat{\beta}_1$] i (5.0.3) fås.

$$\frac{\beta_1}{SE[\hat{\beta}_1]} = \beta_1 \frac{\sigma_{x_1}^2}{(\sigma_{x_1}^2 + \sigma_1^2)} \sqrt{\frac{(n-1)(\sigma_{x_1}^2 + \sigma_1^2)}{\sigma_y^2}} \tag{A.1.6}$$

När värdet på detta uttryck blir större så ökar också styrkan, och när det blir mindre så minskar styrkan för testet. Intuitivt kan man dock utläsa att om bruset i y eller x_1 ökar, så avtar styrkan. En stor utspridning i x_1 och ett större stickprov bidrar å andra sidan till ökad styrka. Eftersom vi i utgångspunkt för våra simuleringar vill ha en viss styrka då bruset i x_1 är obefintligt, $\sigma_1 = 0$, samt utgår från att $\beta_1 = 1$ så kan vi förenkla uttrycket:

$$\beta_1 \frac{\sigma_{x_1}^2}{(\sigma_{x_1}^2 + \sigma_1^2)} \sqrt{\frac{(n-1)(\sigma_{x_1}^2 + \sigma_1^2)}{\sigma_y^2}} = \frac{\sqrt{(n-1)}\sigma_{x_1}}{\sigma_y} \tag{A.1.7}$$

Detta innebär att ändast tre parametrar behöver varieras för att få önskad styrka. Dessa valdes så att

$$F_{t(n-2)}(\frac{\beta_1}{SE(\hat{\beta}_1)} - T_{\alpha/2}) = 0.9 \tag{A.1.8}$$

dvs ungefär då:

$$\frac{\sqrt{(n-1)\sigma_{x_1}^2}}{\sigma_y} - T_{\alpha/2} \approx 1.3 \tag{A.1.9}$$

Här fås 1.3 genom normalapproximation. 10% av massan ligger till höger om värdet 1.2816 standardavvikelser men eftersom vi har en t-fördelning så blir det ett ungefärligt värde kring 1.3 beroende på antal frihetsgrader. Styrkan är alltså 0.9 om

$$\frac{\sqrt{(n-1)}\sigma_{x_1}}{\sigma_y} \approx 1.3 + Z_{\alpha/2} \tag{A.1.10}$$

A.2 Egenskaper för skattningar vid mätfel i prediktor

Som tidigare diskuteras ger minstakvadratmetoden väntevärdesriktiga skattningar för regressionskoefficienterna. Detta bygger på att obervationerna av prediktorerna betraktas som fixa, och situationen blir annorlunda då det finns mätfel i observationerna. Nedan beskrivs olika egenskaper för skattningen av β_1 då det finns brus i x_1. Här låter vi $\hat{x}_{1i}$ beteckna observation i av $\tilde{x}_1$ och $\bar{\tilde{x}}_1$ medelvärdet av observationerna.

Väntevärde av $\hat{\beta}_1$

Väntevärdet av $\hat{\beta}_1$ ges av

$$E[\hat{\beta}_1] = \frac{Cov[\tilde{x}_1, y]}{Var[\tilde{x_1}]} \tag{A.2.1}$$

och eftersom observerade $\tilde{x}_1 = x_1 + \epsilon_1$, där $\epsilon_1 \sim N(0, \sigma_1)$, ger det att

$$\frac{Cov[\tilde{x}_1, y]}{Var[\tilde{x}_1]} = \frac{Cov[x_1 + \epsilon_1, y]}{Var[x_1 + \epsilon_1]} \tag{A.2.2}$$

Eftersom y är okorrelerad med bruset för x_1 så blir bara kovariansen mellan x_1 och y kvar i täljaren. Nämnaren kan skrivas som summan av varianserna, eftersom x_1 och ϵ_1 är oberoende:

$$\frac{Cov[x_1 + \epsilon_1, y]}{Var[x_1 + \epsilon_1]} = \frac{Cov[x_1, y]}{Var[x_1] + Var[\epsilon_1]} \tag{A.2.3}$$

Vidare är y en linjär funktion av $x_1 + \epsilon_y$, där $y = \beta_0 + \beta_1 x_1 + \epsilon_y$. Genom att sätta in detta uttrycket ovan fås:

$$\frac{Cov[x_1, \beta_0 + \beta_1 x_1 + \epsilon_y]}{\sigma_{x_1}^2 + \sigma_1^2} \tag{A.2.4}$$

Och då x_1 är okorrelerad med modellbruset ϵ_y, och β_0 är en konstant, återstår följande uttryck:

$$\frac{Cov[x_1, \beta_1 x_1]}{\sigma_{x_1}^2 + \sigma_1^2} = \frac{\beta_1 \sigma_{x_1}^2}{\sigma_{x_1}^2 + \sigma_1^2} \tag{A.2.5}$$

Följande förväntade värde på β_1, som en funktion av storleken på bruset i x_1, har alltså härletts ovan:

$$E[\hat{\beta}_1] = \beta_1 \frac{\sigma_{x_1}^2}{\sigma_{x_1}^2 + \sigma_1^2} \tag{A.2.6}$$

Variansen av $\hat{\beta}_1$

Variansen för $\hat{\beta}_1$ kan vi skriva på följande sätt:

$$Var[\hat{\beta}_1] = \frac{\sigma_y^2}{(n-1)Var[\tilde{x}_1]} = \frac{\sigma_y^2}{(n-1)(\sigma_{x_1}^2 + \sigma_1^2)} \tag{A.2.7}$$

Där vi alltså ser att en hög varians i bruset på y ger en stor varians i $\hat{\beta}_1$, det vill säga en stor osäkerhet skattningar av β_1. Större stickprovsstorlek, såväl som större utspridning i observationer av x_1 bidrar till att variansen blir mindre, helt i enlighet med situationen utan mätfel.

Kovarians mellan $\tilde{x}_1$ och brus i modell

I kapitel 4.2 visade vi att mätfel i x_1 leder till en modell

$$y = \beta_0 + \beta_1\tilde{x}_1 + \epsilon_y - \beta_1\epsilon_1 \tag{A.2.8}$$

där $\tilde{x}_1$ står för den variabel som observeras och ϵ_1 mätfelet i x_1. Vi skall här visa att $\tilde{x}_1$ och modellbruset $\epsilon_y - \beta_1\epsilon_1$ inte är oberoende, vilket innebär att villkoren för att minsta kvadratmetoden skall ge väntevärdesriktiga skattningar inte är uppfyllda. Eftersom $\tilde{x}_1 = x_1 + \epsilon_1$ har vi att

$$Cov[\tilde{x}_1, \epsilon_y - \beta_1\epsilon_1] = Cov[x_1 + \epsilon_1, \epsilon_y - \beta_1\epsilon_1] \tag{A.2.9}$$

vilket enligt räkneregler för av kovarians av summor av stokastiska variabler ger:

$$Cov[x_1 + \epsilon_1, \epsilon_y - \beta_1\epsilon_1] = Cov[x_1, \epsilon_y] + Cov[x_1, \beta_1\epsilon_1] - Cov[\epsilon_1, \epsilon_y] - Cov[\epsilon_1, \beta_1\epsilon_1] \tag{A.2.10}$$

Vidare gäller enligt antagande att x_1, ϵ_y och ϵ_1 är oberoende och därmed är $Cov[x_1, \epsilon_y] = Cov[x_1, \beta_1\epsilon_1] = Cov[\epsilon_1, \epsilon_y] = 0$. Detta innebär att

$$Cov[x_1 + \epsilon_1, \epsilon_y - \beta_1\epsilon_1] = -Cov[\epsilon_1, \beta_1\epsilon_1] = -\beta_1\sigma_1^2 \tag{A.2.11}$$

och sammanfattningsvis gäller alltså att

$$Cov[\tilde{x}_1, \epsilon_y - \beta_1\epsilon_1] = -\beta_1\sigma_1^2. \tag{A.2.12}$$

B Metoder för att hantera mätfel i en prediktor

Situationen är som i kapitel 4.2 med en enkel linjär regressionsmodell där y förklaras av en prediktor x_1. Det finns mätfel i x_1 och den observerade variabeln betecknas $\tilde{x}_1$. Vi låter som tidigare $\sigma^2_{x_1}$ beteckna varians i x_1, σ_1^2 varians i mätfel och σ_y^2 varians av brus i modell.

B.1 Skattningar vid antaganden om kända brus

Om man har tillgång till skattningar av $\sigma^2_{x_1}$, σ_1^2 och σ_y^2 eller förhållanden mellan dessa uppstår olika fall där det är möjligt att åstadkomma goda skattningar av β_1. Givet $\hat{\beta}_1$ fås sedan

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{\tilde{x}}_1 \tag{B.1.1}$$

där $\bar{y}$ och $\bar{\tilde{x}}_1$ betecknar medelvärden för observationer av y respektive $\tilde{x}_1$. För att underlätta framställningen införs följande beteckningar:

$$s_{yy} = \sum (y_i - \bar{y})^2 \tag{B.1.2}$$

$$s_{xx} = \sum (\tilde{x}_{1i} - \bar{\tilde{x}}_{1i})^2 \tag{B.1.3}$$

$$s_{xy} = \sum (\tilde{x}_{1i} - \bar{\tilde{x}}_{1i})(y_i - \bar{y}) \tag{B.1.4}$$

där summering sker över observationer av y och $\tilde{x}_1$. I följande tre fall är det möjligt att få väntevärdesriktiga skattningar av β_1:

I: Antag att variansen, σ_1^2, av mätfelet i x_1 är känt:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx} - n\sigma_1^2} \tag{B.1.5}$$

II: Antag att 'error-ratio', $\lambda = \frac{\sigma_y^2}{\sigma_1^2}$, är känd:

$$\hat{\beta}_1 = \frac{s_{yy} - \lambda s_{xx} \sqrt{(s_{yy} - \lambda s_{xx})^2 + 4\lambda s_{xy}^2}}{2 s_{xy}} \tag{B.1.6}$$

III: Antag att 'reliability-ratio', $\frac{\sigma^2_{x_1}}{\sigma^2_{x_1} + \sigma_1^2}$, är känd. Då följer direkt av (4.2.4):

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} \left(\frac{\sigma^2_{x_1} + \sigma_1^2}{\sigma^2_{x_1}} \right) \tag{B.1.7}$$

eftersom minsta kvadratskattningen av β_1 är

$$\hat{\beta}_1 = \frac{\sum (\tilde{x}_{1i} - \bar{\tilde{x}}_1)(y_i - \bar{y})}{\sum (\tilde{x}_{1i} - \bar{\tilde{x}}_1)^2} = \frac{s_{xy}}{s_{xx}} \tag{B.1.8}$$

Uppskattningar av σ_1^2 kan ibland fås genom upprepade observationer av samma individ, vilket kan möjliggöra användning av metod I. I vissa tillämpningar kan 'error-ratio' eller 'reliability-ratio', eller approximationer av dessa, finnas till hands. Det är då möjligt att använda metod II eller III. Ovanstående är beskrivs mer ingående av Mandansky [6].

Ett fjärde alternativ är att göra upprepade mätningar för varje observation och beräkna medelvärdet av dessa. Medelvärdet förväntas ligga nära det sanna värdet och minsta kvadratmetoden kan då användas med goda resultat. I praktiken kan detta dock vara kostsamt och är ofta omöjligt att genomföra. I situationer där en latent variabel ersätts med en proxy-variabel hjälper det inte att upprepa mätningar då felet beror på att det finns brus i relationen mellan variablerna.

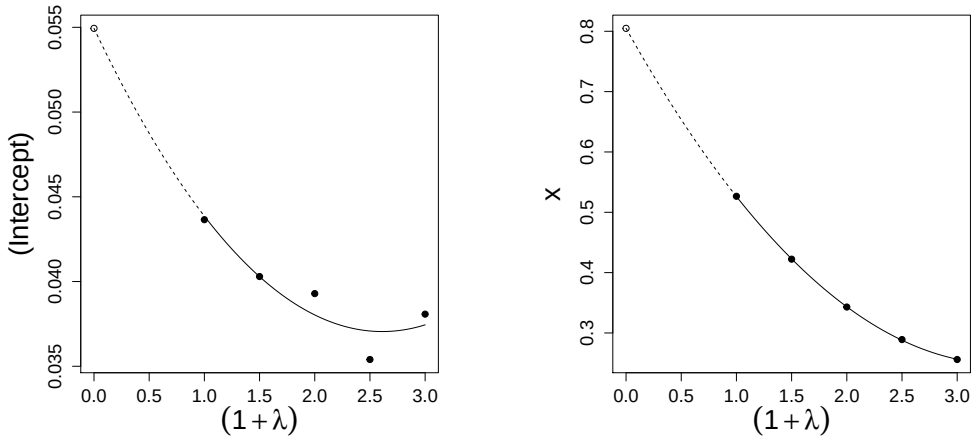
Ytterligare ett alternativ är att använda SIMEX[3], vilken i likhet med fall I kräver god kännedom om storleken av variansen mätfelet. Enligt (4.2.4) ger minsta-kvadratmetoden en väntevärdesriktig skattning av:

$$\beta_1 \frac{\sigma_x^2}{\sigma_x^2 + \sigma_1^2} \tag{B.1.9}$$

Med SIMEX simuleras ny data från observerad data med viktat mätfel $(1+\eta)\sigma_1^2$. För dessa nya observationer kommer minsta-kvadratmetoden att ge väntevädesriktiga skattningar av:

$$\beta_\eta = \beta_1 \frac{\sigma_x^2}{\sigma_x^2 + (1+\eta)\sigma_1^2} \tag{B.1.10}$$

$\eta = 0$ ger skattningen för ursprungsdata. Då η ökar avtar $\hat{\beta}_\eta$ och från de skattningar av som fås vid olika vikter erhålls en slutgiltig skattning av β_1 genom extrapolation till $\eta = -1$. Det är även möjligt att uppskatta varians för skattningarna, vilket gör det möjligt att genomföra hypotestest. Det krävs dock att kännedomen om mätfelen är god för att metoden skall ge goda resultat. Rawlings m. fl. ger en mer detaljerad förklaring till SIMEX-metoden [2]. Punkterna i figur 22 visar skattningarna av β_0 och β_1 då η varierar. Punkten 1 på x-axeln svarar mot $\lambda = 0$ vilket ger minstakvadratlösningen för ursprunglig data. De svarta heldragna linjerna har anpassats efter dessa punkter. Den streckade linjen visar extrapolation till punkten 0 på x-axeln, vilken svarar mot $\lambda = -1$. Om man följer de streckade linjerna återfinns SIMEX-estimatet för regressionskoefficienterna längst upp till vänster i figurerna. I detta exempel är $\hat{\beta}_0^{simex} = 0.055$ och $\hat{\beta}_1^{simex} = 0.81$. De sanna värdena för koefficienterna är $\beta_0 = 0$ och $\beta_1 = 1$. Skattningen av β_1 har förbättrats avsevärt med SIMEX jämfört med minstakvadratmetoden, viket gav $\hat{\beta}_1 = 0.53$.



Figur 22: SIMEX iterationer vid skattning av β_0 (vänster) och β_1 (höger). De svarta punkterna markerar skattningar för simulerad data med ökande brus. En kvadratisk kurva har anpassats efter data och genom extrapolation till $(1+\lambda) = 0$ fås SIMEX-estimatet.

B.2 Skattningar utan antaganden om kända brus

Vi skall nu ge en sammanfattning av metoder som inte kräver ytterligare antaganden. För flera av dessa saknas enkla metoder för inferens men bootstrap kan användas för att beräkna konfidensinterval och genomföra hypotestest.

Instrumentella variabler (IV)

Idén med denna metod bygger på att observera en variabel relaterad till x_1 som inte påverkas av samma mätfel som x_1, ett så kallat instrument. Utifrån denna kan man sedan skatta

[3]SIMulation EXtrapolation

β_1. Denna metod lämpar sig dock inte vid justering eftersom syftet med de modeller som studeras är att undersöka orsakssamband mellan x_1 och y, och variabler relaterade till båda dessa används för justering. Instrumenten till x_1 borde alltså ingå i den justerade modellen och kan därför inte användas som instrument.

Total least squares regression (TLS) & Ortogonal regression (OR)

I minstakvadratmetoden minimeras residualerna i y-led eftersom man tänker sig att x_1 är fix och variation endast finns i y-led. Då x_1 observeras med mätfel finns det brus i båda riktningar och således borde även residualer i x-led minimeras, vilket TLS gör. Skattningar för regressionkoefficienterna fås genom att hitta den linje som minimerar:

$$\sum (y_i - \hat{y}_i)^2 + \sum (\tilde{x}_{1i} - \hat{\tilde{x}}_{1i})^2 \tag{B.2.1}$$

där $(\tilde{x}_{1i} - \hat{\tilde{x}}_{1i})$ betecknar residualer i x-led. Detta är det samma som att hitta den linje som minimerar det ortogonala avståndet mellan punkter och linje och metoden kallas därför även för ortogonal regression (OR). Under antagandet att $\lambda = \frac{\sigma_y^2}{\sigma_1^2} = 1$ ger (B.1.6) samma lösning som TLS/OR, som alltså är väntevärdesriktig. Denna metod har dock den stora nackdelen att vara skalningsberoende. Detta eftersom en skalning med en faktor k i exempelvis y medför att man vill minimera

$$\sum (ky_i - k\hat{y}_i)^2 + \sum (\tilde{x}_{1i} - \hat{\tilde{x}}_{1i})^2 = k^2 \sum (y_i - \hat{y}_i)^2 + \sum (\tilde{x}_{1i} - \hat{\tilde{x}}_{1i})^2 \tag{B.2.2}$$

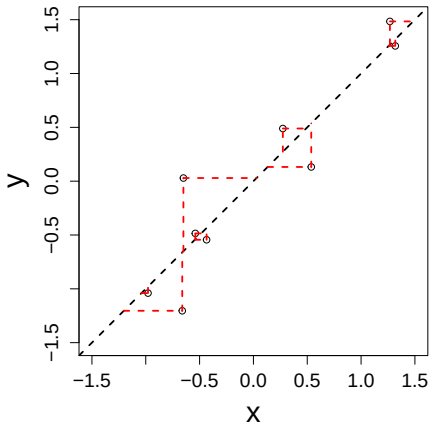
och vikten för residualerna i y-led har alltså ändrats med en faktor k^2. TLS ger alltså olika skattningar beroende på om man mäter längd i meter eller kilometer.

Geometric mean regression (GMR)

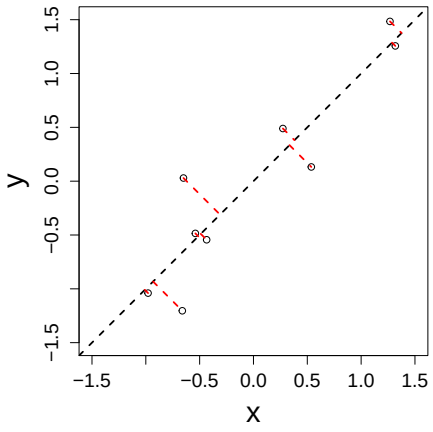
Denna metod påminner om TLS då hänsyn tas till både horisontella och vertikala residualer. Med denna metod skattas regressionlinjen med den linje som minimerar arean av trianglarna som uppstår mellan linjen och de vertikala och horisontella residualerna. Skattningen för lutningen ges av

$$\hat{\beta}_1 = \pm \sqrt{\frac{s_{yy}}{s_{xx}}} \tag{B.2.3}$$

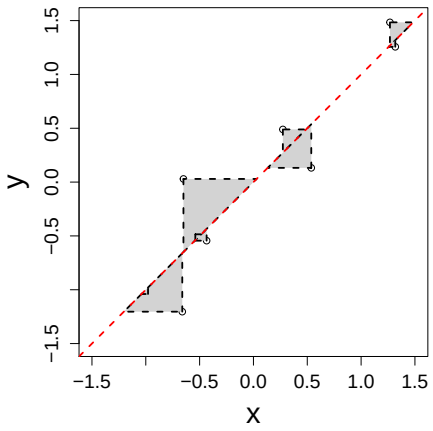
vilket även är det geometriska medelvärdet av lutningarna som fås då regression utförs med y mot $\tilde{x}_1$ samt $\tilde{x}_1$ mot y. Tecknet för skattningen är samma som tecknet för s_{xy}. Jämfört med TLS har GMR den stora fördelen av vara oberoende av skalning. Linjen är även symmetrisk; samma linje fås om $\tilde{x}_1$ betraktas som utfallsvariabel och y som prediktor, vilket inte är fallet för den linje som fås med minsta kvadratmetoden. Under antagandet att $\lambda = \frac{\sigma_y^2}{\sigma_1^2} = \frac{s_{yy}}{s_{xx}}$ ger (B.1.6) samma lösning som GMR och är då väntevärdesriktig. Nackdelen med denna metod är att den ofta visar på ett tydligt linjärt förhållande även om inget samband mellan x_1 och y existerar. Skattningen jämför nämligen bara spridningen i x-led och y-led. Om spridningen i båda variablerna är lika stora kommer $s_{yy} \approx s_{xx}$ och därmed $\hat{\beta}_1 \approx \pm 1$, oavsett hur relationen mellan variablerna ser ut. GMR lämpar sig därför inte för att undersöka orsakssamband. I figur 23 illustreras TLS, OR och GMR.



(a) 'Total Least Squares' (TLS).
Residualer i båda riktningar minimeras.



(b) 'Orthogonal Regression' (OR).
Residualer vinkelräta mot linjen minimeras.



(c) 'Geometric Mean Regression' (GMR).
Arean av trianglar som uppstår mellan residualer
och linje minimeras.

Figur 23: Alternativa metoder för att skatta β_1 vid linjär regression. De streckade linjerna är enbart exempel och ej nödvändigtvis optimala för någon av metoderna.

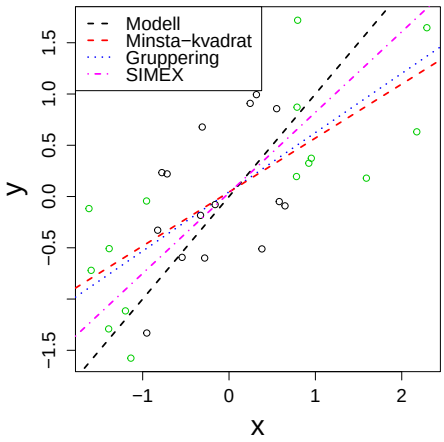
Grupperade variabler (GV)

I denna metod sorteras obervationerna efter $\tilde{x}_1$ och delas in i tre grupper. För grupp 1 och 3 beräknas medelvärden $(\bar{\tilde{x}}_1, \bar{y}_1)$ och $(\bar{\tilde{x}}_3, \bar{y}_3)$. Skattningen för β_1 ges av

$$\hat{\beta}_1 = \frac{\bar{y}_3 - \bar{y}_1}{\bar{\tilde{x}}_3 - \bar{\tilde{x}}_1} \tag{B.2.4}$$

dvs. lutningen mellan mittpunkterna i grupp 1 och 3. Observera att gruppen i mitten inte används för att skatta β_1. Vanligt är att dela in observationerna i tre lika stora grupper eller enligt proportioner 1:2:1. En förutsättning för att denna metod skall ge goda resultat är ordning efter $\tilde{x}_1$ skall vara samma som ordning efter x_1, vilket gäller vid små mätfel. Om så är fallet är grupperna oberoende av ϵ_1 och metoden ger bra skattningar. De metoder som presenterats i detta avsnitt är hämtade från Sprent och Dolby [7] samt Gillard [8].

I figur 24 visas skattade regressionslinjer för några av de metoder som diskuterats. Både minstakvadratmetoden och metoden med grupperade variabler underskattar lutningen ganska kraftigt som en följd av mätfel i x_1. Linjen som skattades med SIMEX stämmer däremot väl med den sanna modellen.



Figur 24: y mot $\tilde{x}_1$ samt skattade regressionslinjer för minstakvadratmetoden, grupperade variabler och SIMEX. Obervationer som används vid gruppering har markerats med grönt. SIMEX ger en regressionslinje som stämmer väl med den sanna linjen, men kräver också god kännedom om storleken på bruset i prediktorn. Både GV och minstakvadratmetoden underskattar linjens lutning, till följd av brus i x.

Metoderna som presenterats ovan har visat sig lämpliga inom vissa tillämpningar och bygger på en god idé om det finns mätfel i data eftersom hänsyn tas till residualer i båda riktingar. De ger dock i allmänhet inte väntevärdesriktiga skattningar och saknar alltså teoretiska garantier. I de situationer som TLS och GMR ger väntevärdesriktiga skattningar är det också möjligt att använda någon av (B.1.5), (B.1.6) eller (B.1.7). Då de senare gäller i mer allmäna situationer är de ofta att föredra.

Referenser

[1] Sven Erick Alm, Tom Britton. Stokastik: Sannolikhetsteori och Statistikteori med Tillämpningar. Stockholm: Liber; 2008:421-450.

[2] John O. Rawlings, Sastry G. Pantula, David A. Dickey. Applied Regression Analysis: A Research Tool, 2nd edition. New York: Springer-Verlag; 1998.

[3] David G.Kleinbaum, Mitchel Klein. Logistic Regression: A Self-Learning Text, 2nd edition. New York: Springer; 2002.

[4] Douglas C. Montgomery, Elizabeth A. Peck och G. Geoffrey Vining. Introduction to Linear Regression Analysis, 4th edition. New Jersey: John Wiley & Sons; 2006:475-510.

[5] S. D. Hodges och P. G. Moore. Data Uncertainties and Least Squares Regression. Journal of the Royal Statistical Society, Series C (Applied Statistics) Vol. 21, No. 2 (1972), pp. 185-195.

[6] Albert Mandansky, The Fitting of Straight Lines When both Variables are Subject to Error, Journal of the American Statistical Association, Vol. 54 No. 285 (Mar., 1959) pp 173-205.

[7] P. Sprent och G. R. Dolby, Query: The Geometric Mean Functional Relationship, Biometrics Vol. 36 No 3 (Sep., 1980) pp. 547-550.

[8] Jonathan Gillard, An Overview of Linear Structural Models in Errors in Variables Regression, Revstat Statistical Journal, Vol. 8 No. 1 (June 2010) pp 57-80.