F21 Regressionsanalys, diagnostik och modellval

Kap 12 Modelldiagnostik, forts

- Residualanalys: Residualerna skattar ju slumptermerna varpå modellantagandena vilar. Hur kan vi verifiera att antagandena är uppfyllda?
- Outlieranalys: Extrema observationer som antingen uppvisar extremt stora residualer eller har ett starkt inflytande på skattningarna, eller både och.
- Kollinearitet: Hur samvarierar prediktorvariablerna? Är detta ett problem?
- Sakkunskap: Vad är det jag analyserar? Hur har data samlats in? Vad är troliga värden?

Residualanalys

Vanliga residaler:

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Standardiserade residualer:

$$z_i = \frac{e_i}{S_e}$$

Studentiserade residualer:

$$r_i = \frac{e_i}{S_e\sqrt{1-h_i}} = \frac{z_i}{\sqrt{1-h_i}}$$

Jackknife residualer:

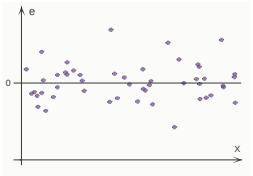
$$r_{(-i)} = \frac{e_i}{S_{(-i)}\sqrt{1-h_i}}$$

Observera! Olika namn på samma saker:

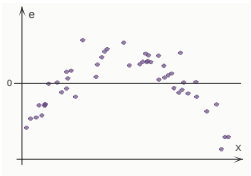
Kleinbaum et al		Minitab
residuals	$\longleftrightarrow$	residuals
standardized	$\longleftrightarrow$	*(finns ej, får man skapa själv)*
studentized	$\longleftrightarrow$	standardized
jackknife	$\longleftrightarrow$	studentized eller deleted t residuals

Indikationer på om modellantagandena håller eller inte får vi framförallt genom att studera plottar över residualerna.

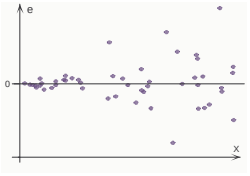
Plottas ofta mot de anpassade värdena $\hat{y}_i$ men även mot varje prediktorvariabel x för sig.



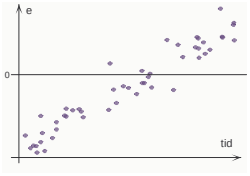
Uppfyller villkoren?



Mönster ska inte finnas!
Kvadratiskt term verkar här!

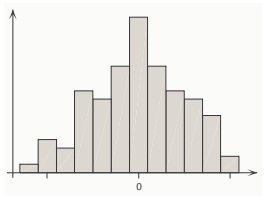


Ej lika varians för alla X!

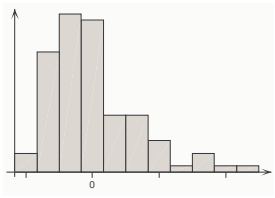


Tidsberoende!
(Minitab: Residuals v. Order)

Histogram över residualerna:



Normalfördelat?

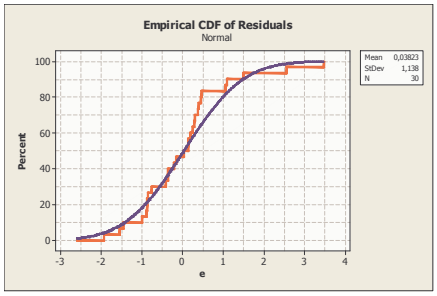


Normalfördelat?

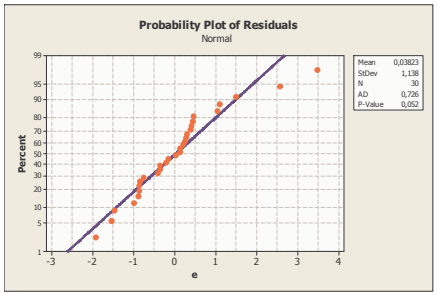
Plottar och histogram kan vara svåra att bedöma, medför ett subjektivt element.

Alternativ är kvantitativa mått och test.

Den empiriska fördelningsfunktionen med en normalfördelning. Är avvikelsen stor?



Probability plot. Är avvikelsen stor? "Korvar" den sig runt linjen?



Normalfördelningstest

Tre tillgängliga i Mintab.

- Anderson-Darling - baserat på den empiriska fördelningsfunktionen.

- Smirnov-Kolmogorov - baserat på χ^2-test.

- Ryan-Joiner (Shapiro-Wilks) - baserat på korellationer.

Testen har som nollhypotes H_0 att observationerna är normalfördelade. Förkasta H_0 vid låga p-värden.

Notera dock att de kan vara väldigt känsliga för outliers, särskilt vid små stickprov.

Test för tidsberoende

Runs-test, ett icke-parametriskt test, inga modelantaganden görs (se Hogg&Tanis sid 585-587)

Låt e_i beteckna residualerna i den ordning de är arrangerade efter (ex tid).

Beteckna med A om $e_i > 0$ och u om $e_i < 0$ och skriv ned sekvensen av alla A och u. Räkna sedan antalet "runs" dvs grupper av A resp u :

$$u\ u\ u\ u\ \ A\ \ u\ u\ \ A\ A\ A\ A\ A$$

4 runs här . Ser inte slumpmässigt ut! Fler negativa residualer (u) i början av serien.

$$A\ u\ A\ u\ A\ u\ A\ u\ A\ u\ A\ u$$

12 runs här. Ser det mer slumpmässigt ut?

För att det ska fungera i Minitab (som använder en normalapproximering) rekommenderas ett minimum av 10 av vardera A och b. Mintab-kommando: RUNS C1.

Autokorrelation, studera sambandet mellan närliggande residualer i termer av korrelation. Kan testas med Durbin-Watson test (finns i Minitab, Regression > Options)

C1	C2
e_1	$*$
e_2	e_1
e_3	e_2
$\vdots$	$\vdots$
e_{n-1}	e_{n-2}
e_n	e_{n-1}

i Minitab: LAG C1 C2
CORR C1 C2

Detta kommer att användas i större omfattning på momentet Tidsserieanalys.

Test för lika varians (homoskedasticitet)

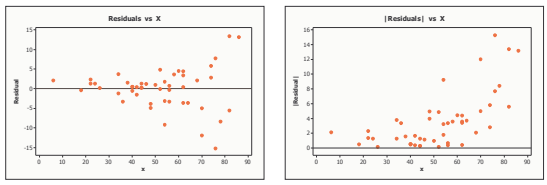
Finns några (se sid 227). Förutsätter (ibland) flera observationer per observerat x-värde.

Enkelt alternativ: studera *rangkorrelationen* mellan

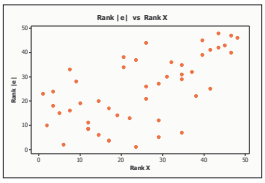
$$|e_i| \quad \text{och} \quad x_i$$

dvs *Spearmans rangkorrelationstest* från grundkursen. Om denna är större än ett kritiskt värde (från en tabell) förkastas nollhypotesen att rangkorrelationen är 0, vilket motsvarar idén att storleken på en residual inte beror på x.

Exempel) Vi får följande:



och plottar vi sedan *rangerna* för residualerna mot *rangerna* för X får vi



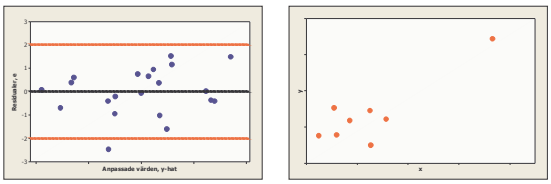
Rangkorrelationen blir i detta fall 0.624;

detta är större än det kritiska värdet 0.375723 ($\alpha = 1\%$, $n = 48$);

alltså finns det samband mellan storleken på e_i och storleken på x_i, dvs det är inte homoskedastiskt.

Outlieranalys

Outlier = extrem observation som ligger "långt utanför det område där de flesta övriga observationerna befinner sig.



Stor residual? Stort inflytande?

Outlier; antingen i prediktorrummet eller som stor residual eller både och. När är det tillräckligt långt bort resp för stort, när ska man reagera?

Man vill mäta

leverage och inflytande (influence)

Tre mått (i Kleinbaum et al.)

-Leveragemått h_i
-Cooksavståndsmått d_i
-Jackkniferesidualer $r_{(-i)}$

Leverage måttet h_i

är ett avståndsmått i *prediktorrummet*, mäter hur långt bort den i:te prediktorpunkten befinner sig från centrum, dvs avståndet mellan

$$(x_{i1}, x_{i2}, \ldots, x_{ik}) \quad \text{och} \quad (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k)$$

jämför med figur ovan. Typiskt beräknas h_i mha dator, kräver matrisalgebra.

Vid enkel linjär regression (med en prediktor) kan det visas att

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)\, s_x^2}$$

Egenskaper allmänt med k prediktorer:

$\frac{1}{n} \leqslant h_i \leqslant 1$ (om man har ett intercept i modellen)

$$\sum h_i = k + 1 \quad \Leftrightarrow \quad \bar{h} = \frac{k+1}{n}$$

Om h_i ligger nära ett innebär detta att den i:te observationen har tvingat in modellen nästan genom punkten (x_i, y_i). Approximativt ska ett histogram över h_i'na påminna om en χ^2-fördelning.

Tumregel: se upp för observationer där

$$h_i > \frac{2\,(k+1)}{n}$$

Cook's avstånd d_i

är ett inflytandemått. Hur mycket påverkas skattningen av regressionskoefficienterna om den i:te observationen tas med/tas bort?

Om prediktorerna har medelvärde $= 0$, samma varians och är okorrelerade gäller att d_i är proportionell mot

$$\sum_{j=0}^{k} \left(\hat{\beta}_j - \hat{\beta}_{j(-i)}\right)^2$$

där $\hat{\beta}_j$ är skattningen med alla observationer och $\hat{\beta}_{j(-i)}$ är skattningen med den i:te borttagen.

Allmänt gäller att

$$d_i = \frac{e_i^2 h_i}{(k+1)\, s_e^2\, (1-h_i)^2} = \frac{r_i^2}{(k+1)} \frac{h_i}{(1-h_i)} > 0$$

dvs d_i är positivt men ej begränsat uppåt. d_i kan alltså vara stor pga att observationen är extrem i prediktorrummet eller för att den får en stor studentiserad residual. Approximativt är

$$d_i \sim F\,(k, n-k-1)$$

Tumregel (möjligen otillförlitlig): se upp för observationer där $d_i > 1.0$ (se sid 232).

Tabell A-10 i Kleinbaum et al ger kritiska värden för den största av samtliga observerade d_i, dvs $\max d_i$, för n observationer och k prediktorer (egentligen $d_i \times$ antalet frihetsgrader).

Om man observerar $\max d_i$ som är större än tabellvärdet $\Longrightarrow$ reagera!

Jackkniferesidualer

e_i : resdual, avstånd mellan observerat och predicerat

$s^2_{(-i)}$: jackknife-skattning av σ^2_ε

h_i : leverage, mått på avståndet mellan den i:te prediktorpunkten och centrum

Jackknife-residualen sammanför dessa tre mått:

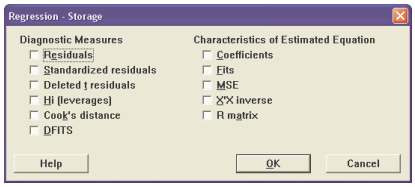
$$r_{(-i)} = \frac{e_i}{s^2_{(-i)}\sqrt{1-h_i}}$$

Testa om $r_{(-i)}$ är signifikant skilt från noll, t-fördelat med $n - p - 2$ frihetsgrader.

Obs! Vi kan testa n stycken h_i, d_i och $r_{(-i)}$, men detta kräver justering av signifikansnivån, se sidorna 229-233 och s.k. "Bonferroni-korrektion".

- Residualanalys och olika inflytande mått är indikatorer på besvärliga/konstiga observationer.
- Vi kommer alltid att observera en största residual, leverage och Cook's mått osv.
- Att blint förlita sig på kvantitativa mått är inte att rekommendera. Vi kan alltid indentifiera outliers och eventuellt ta bort dessa från analysen, men med största försiktighet!
- Vad man ska fråga sig är varför en given observation är extrem. Föreligger det mätfel? Datainsamlingsproblem? Är de observerade värdena omöjliga, otroliga eller bara "lagom" extrema?
- Man måste veta något om bakgrunden och förutsättningarna för analysen!

I Minitab: Som en övning kan ni, via Storage-knappen välja att spara de olika måtten genom att klicka i resp box.



Sedan kan ni undersöka dessa genom tex att skapa histogram, DESC-kommandot, osv.

I enkel linjär regression kan ni se något intressant om ni plotttar h_i mot prediktorvariabeln X. Vad beror mönstret på?

Multikollinearitet

Samband mellan prediktorerna $X_1, X_2, \ldots, X_k$

Ett exempel med två prediktorer,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

Det kan visas att skattningarna kan skrivas som

$$\hat{\beta}_j = c_j \left(\frac{1}{1 - r^2_{x_1 x_2}} \right), \quad j = 1, 2$$

där c_j är en konstant som beror på data och $r_{x_1 x_2}$ är korrelationen mellan X_1 och X_2.

Antag att $x_{i1} = x_{i2}$. Då är $r_{x_1 x_2} = 1$ och om vi skattar regressionsmodellen får vi

$$\hat{\beta}_j = c_j \left(\frac{1}{0} \right) = ?$$

Koefficienterna $\hat{\beta}_1$, $\hat{\beta}_2$ och $\hat{\beta}_0$ kan ej bestämmas entydigt! Dessutom är variansskattningarna för $\hat{\beta}_j$ proportionella mot inflationsfaktorn $1/\left(1 - r^2_{x_1 x_2}\right)$ och osäkerheten i skattningarna ökar!

Det uppstår problem om en prediktor kan skrivas som en linjärkombination av de övriga, tex

$$X_1 = \gamma_0 + \gamma_2 X_2 + \ldots + \gamma_k X_k$$

Att kvantifiera och bedöma kollinearitet rör endast *prediktorerna*, ej responsvariabeln Y.

Antag att man har k stycken prediktorer i modellen. Skatta alla k modeller som kan formuleras med den j:te prediktorn som responsvariabel och de övriga $k-1$ som prediktorer:

$$\begin{aligned}
1 \ &: \quad X_1 = \alpha_{10} + \alpha_{12} X_2 + \ldots + \alpha_{1k} X_k \\
2 \ &: \quad X_2 = \alpha_{20} + \alpha_{21} X_1 + \alpha_{23} X_3 \ldots + \alpha_{2k} X_k \\
&\ \ \vdots \\
k \ &: \quad X_k = \alpha_{k0} + \alpha_{k1} X_1 + \ldots + \alpha_{k(k-1)} X_{k-1}
\end{aligned}$$

För varje sådan modell får vi ett R^2-värde,betecknat R^2_j avseende prediktor j. Vad mäter R^2_j?

Variation Inflation Factor (VIF)

$$VIF_j = \frac{1}{1 - R^2_j}, \quad j = 1, 2, \ldots, k$$

Om $R_j^2 \to 1$ så gäller att $VIF_j \to \infty$. Alternativt definieras *toleransen*

$$\frac{1}{VIF_j} = 1 - R_j^2$$

Tumregel: se upp om $VIF_j > 10.0$

I Minitab: via Options-knappen, klicka för boxen "Variance inflation factors" och studera sedan utskriften.

(Läs även diskussionen om VIF_0, dvs för interceptet, sid 242-243, men skippa fortsättningen sid 243-245.)

Problem med multikollinearitet kan ofta uppstå i samband med Polynomregression (kap13) och modeller med Samspelstermer (kap 11).

En metod som kan reducera multikollinearitet är att *centrera* observationerna runt respektive medelvärde, dvs för varje prediktor ($j = 1, \ldots, k$) och observation ($i = 1, \ldots, n$) beräknas

$$W_{ij} = X_{ij} - \bar{X}_j$$

som får ersätta de ursprungliga observerade värdena.

Exempel med polynomregression på sidorna 239-240. Modellen

$$\mu_{Y|X} = \beta_0 + \beta_1 X + \beta_2 X^2$$

kan ge problem om X och X^2 är starkt linjärt korrelerade i data. Centrera istället enligt ovan och skatta modellen

$$\begin{aligned} \mu_{Y|X} &= \beta_0^* + \beta_1^* W + \beta_2^* W^2 \\ &= \beta_0^* + \beta_1^* \left(X - \bar{X}\right) + \beta_2^* \left(X - \bar{X}\right)^2 \end{aligned}$$

Exempel med samspelstermer, se datorövning 3-4.

Skippa sidorna 246-252 dock ej avsnitt 12-9.

Kap 16 Att Välja Bästa Modell

1. Definiera en maximal modell
   - alla tänkbara prediktorer: $X_1, X_2, \ldots$
   - potenser av prediktorer: $X_1^2, X_1^3, \ldots$
   - ev. transformationer: ln X_1,$1/X_2, \ldots$
   - samspelstermer: $X_1X_2, X_1X_3, \ldots$
   - kontrollvariabler och deras ev potenser, transformationer, samspelstermer...

Problem med överanpassning (overfitting), dvs att ta med sådant om inte ingår i den "sanna" modellen medför inte bias. Däremot ev. problem med multikollinearitet. Även möjliga problem med antalet frihetsgrader ($df = n - q - 1$). Om $df = 0$ så blir $R^2 = 1$ även om det i populationen är så att $R^2 = 0$.

Tumregel i boken: minst 5 observationer per prediktor (även potenser, samspelstermer etc).

Även tolkningsproblem med alltför stora modeller (kom ihåg Occam's razor).

2. Definiera kriterium för val av modell

Modellvalskriterium som kan beräknas för varje modell vi analyserar / skattar dvs ett index att jämföra modeller med varandra.

Fyra behandlas i boken: R_p^2, F_p, MSE_p och C_p

Utgå ifrån en jämförelse mellan två modeller, en full (maximal) och en mindre (reducerad):

$$\underbrace{\underbrace{\beta_0 + \beta_1 X_1 + .. + \beta_p X_p}_{\text{reducerad modell, } p \text{ pred}} + \beta_{p+1} X_{p+1} + .. + \beta_k X_k}_{\text{maximal modell, } k \text{ st prediktorer}}$$

- R_k^2 jämfört med R_p^2, andel förklarad variation
  - tenderar att överskatta sant R^2
  - fler prediktorer $\Rightarrow$ högre R^2
  - R_k^2 alltid större än R_p^2
  - alternativ: R_{adj}^2

- F_p dvs F-kvoten från multipelt partiellt F-test (ekv. (16.4))

$$F_p = \frac{[SSE\,(\text{reducerad}) - SSE\,(\text{full})]\,/\,(k-p)}{SSE\,(\text{full})\,/\,(n-k-1)}$$

- MSE_p dvs skattningen

$$\hat{\sigma}^2_{Y|X} = \hat{\sigma}^2_{\varepsilon} = \frac{SSE\,(\text{reducerad})}{n-p-1}$$

Vi vill ha en modell med liten slumptermsvarians. Hur ska man jämföra modeller?

- Mallow's C_p definieras (ekv. (16.6))

$$C_p = \frac{SSE\,(\text{reducerad})}{MSE\,(\text{full})} - (n - 2\,(p+1))$$

Om $MSE\,(\text{reducerad}) \approx MSE\,(\text{full})$ så blir

$$C_p \approx p + 1$$

Om $C_p > p+1$ så finns det förmodligen utrymme för fler prediktorer, om $C_p < p+1$ har ni förmodligen överanpassat modellen.

- Kriterier baserade på informationsteori, tex
  - Akaike Information Criterium

$$AIC = n \log MSE + 2p$$

  - Schwartz Bayesian Criterium

$$SBC = BIC = n \log MSE + p \log n$$

  - straffar för "stora" modeller och belönar stora minskningar i osäkerhet (residualvarianser)
  - när man jämför värden är det bättre med det mindre

3. Definiera en strategi för val av variabler

- Jämför alla möjliga modeller

  Många prediktorer $\Rightarrow$ ännu fler modeller

  Ex) p prediktorer ger (med bara huvudeffekter)

$$\binom{p}{0} + \binom{p}{1} + \binom{p}{2} + \ldots + \binom{p}{p} = 2^p$$

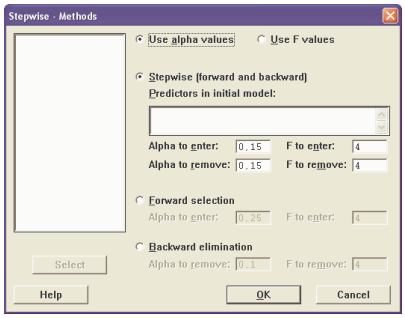
olika modeller att skatta och analysera!

- Reducera en full modell
  1. Skatta full modell
  2. Vilken prediktor kan tas bort utan att det försvinner en stor andel förklarad variation? Kriterium: F-test.
  3. Ta bort denna
  4. Skatta reducerad modell, gå till 2.

- Bygga upp från tom modell
  1. Skatta "tom" modell
  2. Givet modellen, vilken enskild prediktor bör läggas till? Kriterium: F-test.
  3. Lägg till denna
  4. Skatta utökad modell, gå till 2.

- Stegvis regression (Minitab: Stepwise)
  - variant av de två oföregående, välj själva om ni vill börja uppifrån eller nerifrån
  - i varje steg analyseras om en prediktor man tidigare slängde bort borde tas med igen alternativt en som tidigare togs med nu bör slängas bort
  - i Minitab anges kriterier för att lägga till resp ta bort

Minitab dialogfönster för kriterier (Methods-knappen) med stegvis regression:



4. Genomför analysen

När man väl har bestämt sig för en modell ska en mer fullständig analys genomföras (residualanalys, outlieranalys enligt tidigare)

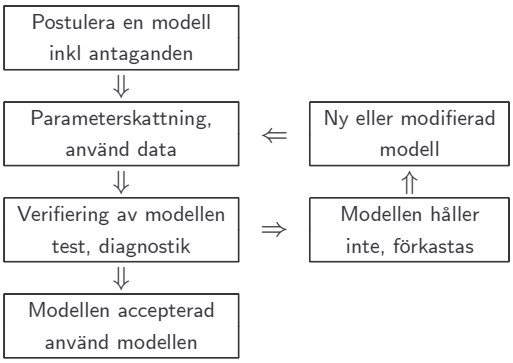
5. Validering, bedömning av modellens tillförlitlighet

Om modellen lyckas predicera nya observationer "bra" så är det en tillförlitlig (reliable) modell. Hur ska man bedöma detta innan man får nya observationer?

Split-sample metoden:

- Dela upp materialet i två (eller flera) grupper
- Identifiera modellen och skatta den genom att använda en grupp
- Predicera nya Y-värden, dvs Y, med prediktorvärdena från den andra gruppen.
- Beräkna prediktionsfelet dvs $Y - \hat{Y}$ och dess varians

Kom ihåg, arbetsgången är ofta en *iterativ* process:



.

F22 Regressionsanalys forts och Logistisk regression

Kap 13 Polynomregression

Emellanåt upptäcker man samband som är kvadratiska, kubiska osv.

Allmänt: polynom av k:te ordningen

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k + \varepsilon$$

där (som tidigare)

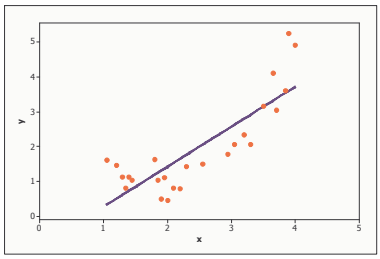
$$\varepsilon \overset{iid}{\sim} N\left(0, \sigma_\varepsilon^2\right)$$

tillsammans med övriga antaganden (vilka är dessa?). Ett antagande måste dock ändras. Vilket?

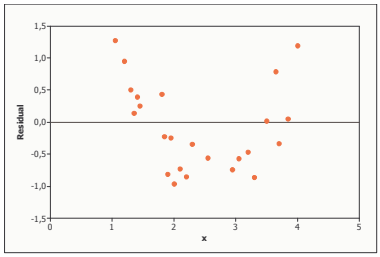
Modellen uttryckt i termer av betingat väntevärde

$$\mu_{Y|X} = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k$$

Antag att man börjar med enkel linjär regression.



Typisk indikation på att man gjort fel ska man se i mönstret som uppstår i residualplotten:

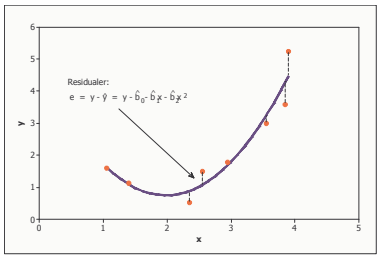


Alltså anpassa istället en kvadratisk modell till data!

MK-skattningen går till precis som förut, vi minimerar nu

$$Q = \sum \left(y_i - \beta_0 + \beta_1 x_i + \beta_2 x_i^2\right)^2$$

Vi vill alltså minimera summan av de kvadrerade avstånden till linjen som nu är en kvadratisk linje.



Betrakta x^2 som en extra prediktorvariabel precis som vanligt i multipel linjär regression.

I Minitab: säg att Y ligger i C1 och X i C2.

Kommando:

MTB> LET C3=C2**2

Anger sedan C2 och C3 som prediktorer i dialogrutan.

Resultatet blir som vanligt en ANOVA-tablå:

```
Analysis of Variance
Y = 4,22 - 3,47 X + 0,902 X2

Predictor      Coef   StDev      T      P
Constant     4,2201  0,6931   6,09  0,000
X           -3,4708  0,6010  -5,78  0,000
X2           0,9022  0,1163   7,76  0,000

S = 0,401725  R-Sq = 91,7%  R-Sq(adj) = 90,9%


Analysis of Variance
Source           DF      SS      MS       F
Regression        2  39,127  19,564  121,23
Residual Error   22   3,550   0,161
Total            24  42,678

Analysis of Variance
Source  DF  Seq SS
X        1  29,408
X2       1   9,720
```

Inferensen går till precis som förut:

1. $MSR \;/\; MSE \sim F(2, n-3)$, overall test för modellen som helhet

2. $SSR(X) \;/\; MSE(X) \sim F(1, n-2)$, test för enbart X (först in, enkel linjär modell)

3. $\frac{SSR(X^2|X)}{MSE(X,X^2)} \sim F(1, n-3)$, partiellt F-test för X^2 givet X

4. $\frac{SSR(X,X^2|Z)/2}{MSE(Z,X,X^2)} \sim F(2, n-4)$, multipelt partiellt F-test för X och X^2 givet något Z.

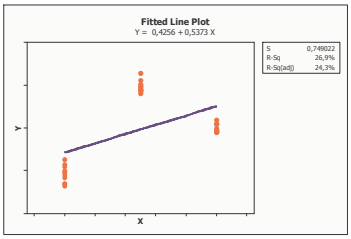
5. och motsvarande ekvivalenta t-test...(sist-in-test)

Lack of Fit test

Beskrivs på sid 290-292. Ofta i samband med just polynomregression.

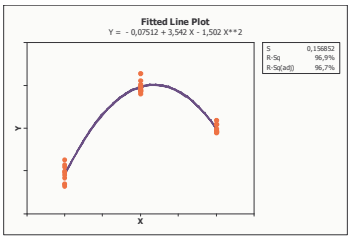
I Minitab: i Regressionsfönstret, klicka på Options-knappen och under "Lack of Fit Tests" välj antingen "Pure Error" om ni har replikat dvs flera observationer per X-värde eller "Data subsetting" om ni inte har det.

Ex) Anpassning till en linjär modell verkar inte vara rätt här:



Resultatet i Minitab:
Possible curvature in variable X (P-Value=0,000)
Possible lack of fit at outer X-values (P-Value=0,000)
Overall lack of fit test is significant at P=0,000

Anpassning till en kvadratisk modell verkar bättre



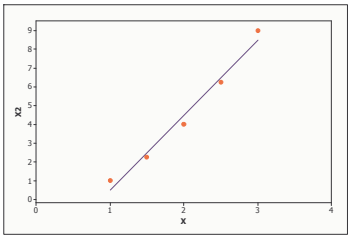
I Minitab:
No evidence of lack of fit (P >= 0,1).

Vad man ska tänka på är att det kan bli problem med multikollinearitet!

Ex) Antag att fem värden på X observeras enligt nedan. Insättning av en kvadratisk term i modellen ger

X	1.0	1.5	2.0	2.5	3.0
X^2	1.00	2.25	4.00	6.25	9.00

$Corr\left(X, X^2\right) = 0.989 \quad \Rightarrow \quad VIF = \frac{1}{1-0.989} = 45.7$



Hur ska man lösa detta?

Boken ger en lång beskrivning av sk *ortogonala polynom* som kan användas för att skapa nya prediktorer med en mängd trevliga egenskaper men det är alldeles för mycket för den här kursen!

Enklare metod som fungerar alldeles utmärkt med kvadratiska samband är *centrering runt medelvärdet*, dvs skapa följande nya prediktorer,

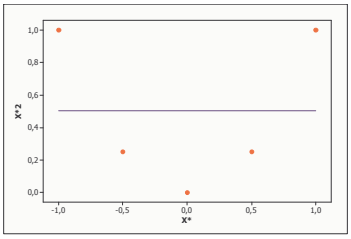
$$X_1^* = \left(X - \bar{X}\right) \quad \text{och} \quad X_2^* = \left(X - \bar{X}\right)^2$$

Effekten av denna enkla transformering kan illustreras med exempelt ovan:

X	1.0	1.5	2.0	2.5	3.0
$X_1^* = \left(X - \bar{X}\right)$	−1.0	−0.5	0.0	0.5	1.0
$X_2^* = \left(X - \bar{X}\right)^2$	1.0	0.25	0.0	0.25	1.0

$$Corr\left(X_1^*, X_2^*\right) = 0 \quad \Rightarrow \quad VIF = \frac{1}{1-0} = 1$$

Ny plott



Använd X_1^* och X_2^* i skattningen av modellen och vi får

$$\hat{Y} = \hat{\beta}_0^* + \hat{\beta}_1^*\left(X - \bar{X}\right) + \hat{\beta}_2^*\left(X - \bar{X}\right)^2$$

vilket kan skrivas om enligt

$$\hat{Y} = \underbrace{\left(\hat{\beta}_0^* - \hat{\beta}_1^*\bar{X} + \hat{\beta}_2^*\bar{X}^2\right)}_{=\hat{\beta}_0} + \underbrace{\left(\hat{\beta}_1^* - 2\hat{\beta}_2^*\bar{X}\right)}_{=\hat{\beta}_1} X + \underbrace{\hat{\beta}_2^*}_{=\hat{\beta}_2} X^2$$

dvs de skattningar man skulle få utan att centrera!

Detta är ett enkelt sätt att bli av med ett "självförvållat" problem! Svårare däremot att bli av med multikollinearitetsproblem vid polynomregression av högre ordning, särskilt med udda potenser (X^3, X^5), se sid 299.

Kap 14 Dummyvariabler

Hittills har vi haft kontinuerliga prediktorer X.

Ofta har man kategoriska prediktorer (nominal- el ordinalskala).

Ex)

- X = man eller kvinna
- X = sockerpiller el Medicin A el Medicin B
- X = kallt, ljummet, varmt
- X = Saab, Audi, Volvo, Toyota, ...

För att hantera detta används dummyvariabler!

Annat namn: indikatorvariabler.

En dummyvariabel antar typiskt värdena 0 eller 1 och konstrueras så att man kan identifiera vilken kategori det är frågan om.

Ex) X = kön

Alt. 1:

$$X = \begin{cases} 0 & \text{man} \\ 1 & \text{kvinna} \end{cases}$$

Modellen skrivs

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

vilket är ekvivalent med

$$Y = \begin{cases} \beta_0 + \varepsilon & \text{man} \\ \beta_0 + \beta_1 + \varepsilon & \text{kvinna} \end{cases}$$

Tolkningen är

- β_0 är väntevärdet i Y för män
- β_1 är förväntade skillnaden mellan kvinnor och män

Alt. 2:

$$X = \begin{cases} -1 & \text{man} \\ 1 & \text{kvinna} \end{cases}$$

Modellen skrivs

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

vilket är ekvivalent med

$$Y = \begin{cases} \beta_0 - \beta_1 + \varepsilon & \text{man} \\ \beta_0 + \beta_1 + \varepsilon & \text{kvinna} \end{cases}$$

Tolkningen är

- β_1 är halva förväntade skillnaden i Y mellan kvinnor och män
- β_0 är väntevärdet i Y för båda tillsammans förutsatt att män och kvinnor är exakt lika många

Om man har en kategorisk prediktor X med k olika kategorier behövs $(k-1)$ st dummyvariabler.

Det fungerar inte att utöka med ytterligare en nivå för dummyvariablen, tex

$$X = \begin{cases} 0 & \text{socker} \\ 1 & \text{medicin A} \\ 2 & \text{medicin B} \end{cases}$$

En ökning i X från 1 till 2 blir meningslös! Vi vet väl inte om skillnaden mellan socker och med.B är två gånger skillnaden mellan socker och med.A?

Ej ekvidistans! Kanske inte ens en ökning!

Inför ytterligare en dummyvariabel:

$$X_1 = \begin{cases} 0 & \text{ej medicin A} \\ 1 & \text{medicin A} \end{cases} \qquad X_2 = \begin{cases} 0 & \text{ej medicin B} \\ 1 & \text{medicin B} \end{cases}$$

Tre möjligheter

$$(X_1, X_2) = \begin{cases} (0,0) & \text{socker} \\ (0,1) & \text{medicin A} \\ (1,0) & \text{medicin B} \end{cases}$$

Vi kan identifiera vilken behandling det är frågan om.

Modellen

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

ger

$$Y = \begin{cases} \beta_0 + \varepsilon & \text{socker} \\ \beta_0 + \beta_1 + \varepsilon & \text{medicin A} \\ \beta_0 + \beta_2 + \varepsilon & \text{medicin B} \end{cases}$$

Tolkning av koefficienter:

β_0 : väntevärdet i Y efter behandling med socker

β_1 : förväntad skillnad i Y mellan medicin A och socker

β_2 : förväntad skillnad i Y mellan medicin B och socker

Bakgrunden ovan är ett kontrollerat experiment där man får ett av tre möjliga preparat.

Om man kan kombinera preparat, dvs om både medicin A och B kan ges samtidigt, kan det finnas anledning till att ta med samspelseffekter:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12}\,(X_1 X_2) + \varepsilon$$

ger

$$Y = \begin{cases} \beta_0 + \varepsilon & \text{socker} \\ \beta_0 + \beta_1 + \varepsilon & \text{medicin A} \\ \beta_0 + \beta_2 + \varepsilon & \text{medicin B} \\ \beta_0 + \beta_1 + \beta_2 + \beta_{12} + \varepsilon & \text{medicin A \& B} \end{cases}$$

Antag att vi har en prediktor X med två nivåer, $X = A$ eller B. Vi skattar modellen, dvs parametrarna β_0 och β_1.

Inferens:

$$H_0 : \beta_1 = 0 \quad \text{mot} \quad H_1 : \beta_1 \neq 0$$

Testvariabeln är som tidigare

$$T = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} \sim t\,(n - p - 1)$$

Detta test är ekvivalent med testet att tjämföra väntevärden i två populationer

$$H_0 : \mu_A = \mu_A \quad \text{mot} \quad H_1 : \mu_A \neq \mu_B$$

med antagande om lika varians.

Två pop.modell	Regr.modell
$Y_A \sim N\left(\mu_A, \sigma^2\right)$	$Y_A = \beta_0 + \varepsilon$
$Y_B \sim N\left(\mu_B, \sigma^2\right)$	$Y_B = \beta_0 + \beta_1 + \varepsilon$

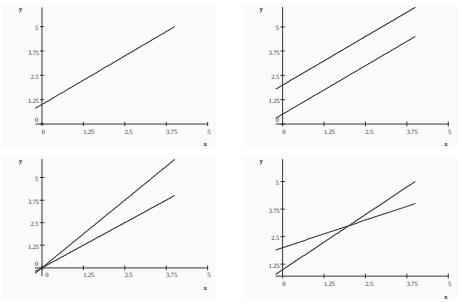
där $\varepsilon \sim N\left(0, \sigma^2\right)$. Ekvivalensen är uppenbar genom att inse att

$$\mu_A = \beta_0 \quad \text{resp} \quad \mu_B = \beta_0 + \beta_1$$

Man kan givetvis utöka modellen och ta med "vanliga" kontinuerliga prediktorer.

Vad som är intressant är frågor av typen:

Är det samma lutning på regressionslinjen i de olika grupperna? Är det samma intercept?



Kom ihåg att använda en multiplikativ samspelsterm!

Metod I (sid 327 avsnitt 14-8)

Två *separata* modellskattningar för olika grupper (sid322 avsnitt 14-7):

$$\begin{aligned} Y_A &= \beta_{0A} + \beta_{1A} X + \varepsilon \\ Y_B &= \beta_{0B} + \beta_{1B} X + \varepsilon \end{aligned}$$

dvs gör två regressionsanalyser, en för varje grupp och jämför resultaten mellan grupperna/modellerna.

Kräver en del ytterligare beräkningar, tex poolade variansskattningar (se sid 323) eftersom ett av antagandena är att slumptermsvariansen är lika mellan grupperna.

Krångligare kan man tycka men ibland har man inte tillgång till rådata, endast färdiga resultat som publicerats.

Ex) En undersökning i Spanien finns redovisad i en artikel och man upprepar undersökningen i Sverige. Finns det skillnader? Vilka?

Metod II

En modellskattning med allt-i-ett (sid 327 avsnitt 14-8):

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_{12} XZ + \varepsilon$$

där Z är en dummyvariabel som anger grupptillhörighet. Detta ger

$$\begin{aligned} A &: \quad Y = \beta_0 + \beta_1 X + \varepsilon \\ B &: \quad Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_{12})\, X + \varepsilon \end{aligned}$$

För att avgöra vilken av de fyra situationerna som gäller kör man på med den vanliga inferensen, dvs testa

$$H_0 : \beta_2 = 0 \quad \text{mot} \quad H_1 : \beta_2 \neq 0$$

resp

$$H_0 : \beta_{12} = 0 \quad \text{mot} \quad H_1 : \beta_{12} \neq 0$$

Detta alternativ verkar enklare och vettigare än alternativ I, peta in allt och gör en stor allt-i-ett analys. Men som sagt...

Boken redogör för varje tänkbart test som kan göras för respektive Metod I och II.

- test för lika intercept
- test för lika lutning
- test för att regressionslinjerna sammanfaller

Det ser ut som mycket att lära in...

Testen är precis som förut, inget principiellt nytt jämfört med multipel linjär regression!

Rekommendation: Läs igenom avsnitt 14-1 - 14-10 utan att uppehålla er för länge, skippa avsnitt 14-11 - 14-13.

Kap 23 + KD's häfte, Logistisk regression

Exempel på icke-linjär regressionsmodell.

(1) Börja med e*nkel linjär regression* där Y är en kontinuerlig variabel.

Obetingat

$$\begin{aligned} Y &\sim Distr\left(\mu_Y\ ,\ \sigma_Y^2\right) \\ &: \quad E(Y) = \mu_Y, \quad Var(Y) = \sigma_Y^2 \end{aligned}$$

där $Distr$ står för någon fördelning med ändligt väntevärde och ändlig positiv varians.

Betingat (modell):

$$\begin{aligned} Y \ \mid \ X &\sim N\left(\beta_0 + \beta_1 X\ ,\ \sigma_\varepsilon^2\right) \\ &: \quad E(Y \mid X) = \beta_0 + \beta_1 X \\ &: \quad Var(Y \mid X) = \sigma_\varepsilon^2 < \sigma_Y^2 \end{aligned}$$

Det betingade väntevärdet bestäms av X och dessutom kan vi förklara/predicera Y mycket bättre.

(2) Tänk om Y ej är kontinuerlig utan en dikotom variabel, dvs kan anta värdena 0 el 1.

Obetingat

$$\begin{aligned} Y &\sim Bernoulli(p_Y) \\ &: \quad E(Y) = p_Y \\ &: \quad Var(Y) = p_Y(1 - p_Y) \end{aligned}$$

där p_Y är sannolikheten för att observera $Y = 1$.

Kan man förklara att man i vissa lägen tycks observera "etta" oftare än "nolla"?

Dvs att sannolikheten för lyckat utfall kanske kan förklaras med någon prediktor?

Hur ska modellen i så fall formuleras?

- Allmänt är det betingade väntevärden för Y givet X som modelleras.
- Med enkel linjär regression i (1) ovan har man
  $$f(x) \ = \ \beta_0 + \beta_1 x \ = \ \mu_{Y|X=x}$$
  och det inses att $\quad -\infty < \ \mu_{Y|X} \ < \ \infty$.
- Problem i (2) ty väntevärdet är ju en *sannolikhet* så vi behöver en funktion $f(x) = \mu_{Y|X} = p_{Y|X}$ sådan att
  $$0 \ < \ p_{Y|X} \ < \ 1$$
  En linjär funktion uppfyller inte detta krav!
- Däremot kan ju i princip vår prediktor tillåtas anta vilka värden som helst.
- Önskemål: hitta en funktion $f(x)$ med obegränsad definitionsmängd
  $$-\infty \ < \ x \ < \ \infty$$
  men begränsad värdemängd
  $$0 \ < \ f(x) \ < \ 1$$

Ett av de vanligaste valen är den logistiska funktionen som definieras

$$f(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} = p_{Y|X}$$

I Kleinbaum et al skrivs detta som (sid 657)

$$f(x) = \frac{1}{1+\exp(-\alpha-\beta x)}$$

men uttrycken är identiska.

Om $\beta > 0$ så gäller att

$$\begin{aligned} x &\to -\infty \quad \Rightarrow \quad f(x) \to 0 \\ x &\to \infty \quad \Rightarrow \quad f(x) \to 1 \end{aligned}$$

och om $\beta < 0$ så gäller det omvända.

En intressant punkt som man kan se i rapportering är det värde på X när sannolikheten för $Y = 1$ är precis 0.5. Denna punkt kallas LD50 vilket står för *lethal dosage* 50% och inträffar alltid när

$$x = -\alpha/\beta$$

Forts (2)

Betingat

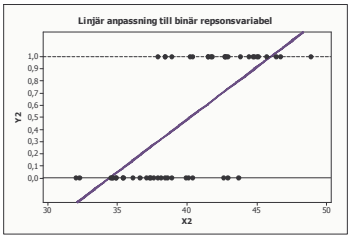
$$\begin{aligned} Y \mid X &\sim Bernoulli\left(p_{Y|X}\right) \\ : E(Y) &= p_{Y|X} \\ : Var(Y) &= p_{Y|X}\left(1-p_{Y|X}\right) \end{aligned}$$

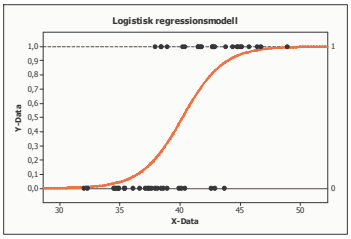
dvs det betingade väntevärdet för Y givet X tolkas / definieras som en sannolikhet

$$p_{Y|X} = P(Y=1 \mid X=x)$$

Den andra möjliga händelsen, $Y = 0$, har en betingad sannolikhet

$$\begin{aligned} P(Y=0 \mid X=x) &= 1 - P(Y=1 \mid X=x) \\ &= 1 - \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} \\ &= \frac{1}{1+\exp(\alpha+\beta x)} \end{aligned}$$





Odds, Logodds och Oddskvoter

Oddset för en händelse definieras som kvoten mellan sannolikheten för händelsen och sannolikheten för komplementet:

$$Odds(A) = \frac{P(A)}{P(A')} = \frac{P(A)}{1-P(A)}$$

Med logistisk regression får man det betingade oddset för händelsen $Y = 1$ givet X

$$\begin{aligned} Odds(Y=1 \mid X) &= \frac{P(Y=1 \mid X)}{P(Y=0 \mid X)} = \frac{\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}}{\frac{1}{1+\exp(\alpha+\beta x)}} \\ &= \exp(\alpha+\beta x) \end{aligned}$$

Oddset för händelsen $Y = 0$ givet X blir på samma sätt

$$\begin{aligned} Odds(Y=0 \mid X) &= \frac{P(Y=0 \mid X)}{P(Y=1 \mid X)} \\ &= \frac{1}{\exp(\alpha+\beta x)} \\ &= (\exp(\alpha+\beta x))^{-1} \\ &= \exp(-\alpha-\beta x) \end{aligned}$$

Logoddset för en händelse är helt enkelt den naturliga logaritm av oddset:

$$\begin{aligned} LogOdds\,(Y = 1 \mid X) &= \ln\left(Odds\,(Y = 1 \mid X)\right) \\ &= \ln\left(\exp\left(\alpha + \beta x\right)\right) \\ &= \alpha + \beta x \end{aligned}$$

Nu kan vi tolka regressionskoefficienter:

- α är *logoddset* för $Y = 1$ då $X = 0$
- om x ändras med 1 enhet så förändras *logoddset* för $Y = 1$ med β

Kom ihåg att responsvariabeln Y nu är något annat än vad vi har haft tidigare.

I (1) betraktas Y som en observerbar företeelse som kan mätas och sedan jämföras med den skattade $\hat{Y}$ direkt. I princip skulle vi kunna observera

$$y_i = \hat{Y} \quad \Leftrightarrow \quad e_i = 0$$

De betingade väntevärdena $\hat{\mu}_{Y|X=x}$ ligger i utfallsrummet för Y.

I (2) och den logistiska modellen är Y förvisso också observerbar men kan bara anta värdena 0 eller 1. De skattade $\hat{Y}$ ska nu kanske hellre betecknas med $\hat{p}_{Y|X=x}$ och tolkas som betingade sannolikheter (för $Y = 1$). Vi kommer aldrig att observera något fall där

$$Y = \hat{p}_{Y=1|X=x}$$

ty betingade sannolikheterna $\hat{p}_{Y|X=x}$ antar typiskt värden som ligger *mellan* 0 och 1.

En residual måste sålunda tolkas annorlunda:

$$\begin{aligned} y_i = 1 \quad \Rightarrow \qquad e_i = \; & y_i - \hat{p}_{Y=1|X=x_i} \\ & = 1 - \hat{p}_{Y=1|X=x_i} \\ & = \hat{p}_{Y=0|X=x_i} \end{aligned}$$

dvs sannolikheten för $Y = 0$ givet $X = x$.

$$\begin{aligned} y_i = 0 \quad \Rightarrow \qquad e_i = \; & y_i - \hat{p}_{Y=1|X=x_i} \\ & = 0 - \hat{p}_{Y=1|X=x_i} \\ & = -\hat{p}_{Y=1|X=x_i} \end{aligned}$$

dvs minus sannolikheten för $Y = 1$ givet $X = x$.

Ytterligare en tolkning av regressionskoefficienten β

Oddskvoter beskrivs kanske enklast med ett exempel från boken (sid 659–660).

Anta en logistisk modell med tre prediktorer

$$\begin{aligned} X_1 &= \text{röker / röker ej} \\ X_2 &= \text{ålder} \\ X_3 &= \text{vit / svart} \end{aligned}$$

Man får

$$LogOdds = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

och vill sedan jämföra rökare/icke-rökare bland 45-åriga svarta.

$$\begin{aligned} Oddsvkot &= OR \\ &= \frac{Odds\,(Y = 1 \mid \text{45 år, svart, rökare})}{Odds\,(Y = 1 \mid \text{45 år, svart, icke-rökare})} \\ &= \frac{\exp\left(\alpha + \beta_1 \cdot 1 + \beta_2 \cdot 45 + \beta_3 \cdot 1\right)}{\exp\left(\alpha + \beta_1 \cdot 0 + \beta_2 \cdot 45 + \beta_3 \cdot 1\right)} \\ &= \exp\left(\beta_1\right) \end{aligned}$$

Alltså om vi ökar X_1 med 1 enhet (allt annat konstant) så är den *relativa förändringen i oddset*, för händelsen $Y = 1$, lika med $\exp(\beta_1)$.

Oddskvoter används typiskt för jämförelser mellan olika grupper med avseende på risker för tex sjukdomar.

Ex) Hur mycket större är risken för en 45-årig vit rökare jämfört med en 20-årig svart icke-rökare?

$$\begin{aligned} OR &= \frac{\exp(\alpha + \beta_1 \cdot 1 + \beta_2 \cdot 45 + \beta_3 \cdot 0)}{\exp(\alpha + \beta_1 \cdot 0 + \beta_2 \cdot 20 + \beta_3 \cdot 1)} \\ &= \exp(\beta_1 + 25\beta_2 - \beta_3) \end{aligned}$$

Odds kan vara svåra att tolka i början. Men vet man sannolikheten för en händelse så vet man också oddset:

$$Odds(A) = \frac{P(A)}{1 - P(A)}$$

Och vet man oddset så vet man också sannolikheten

$$P(A) = \frac{Odds(A)}{1 + Odds(A)}$$

Odds här är inte samma sak som odds i spel som definieras av hur mycket pengar man kommer att få om man satsar rätt. Eller

$$SpelOdds(A) = \frac{1}{Odds(A)} + 1 = \frac{P(A')}{P(A)} + 1$$

där $P(A)$ är andelen i pengar som har satsats på att A vinner, vilket kan tolkas som en sorts sannolikhet.

Inget sagt ännu om skattning av en logistisk regressionsmodell.

Typiskt maximeras en likelihoodfunktion, se tex eq. (23.14) sid 673 och eq. (23.17) sid 676.

Detta kräver numeriska metoder dvs inget vi frivilligt gör för hand! Använd ett färdigt datorprogram!

.

.