F5 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:

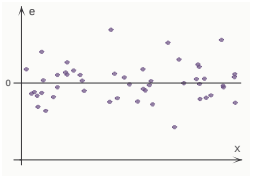
$$r_{(-i)} = \frac{e_i}{S_{(-i)}\sqrt{1-h_i}}$$

Observera! Olika namn på samma saker:

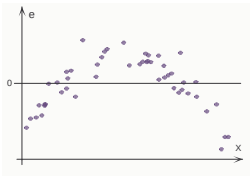
Kleinbaum et al		Minitab
residuals	$\longleftrightarrow$	residuals
standardized	$\longleftrightarrow$	*(finns ej, får man skapa själv)*
studentized	$\longleftrightarrow$	standardized
jackknife	$\longleftrightarrow$	studentized eller deleted t residuals

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

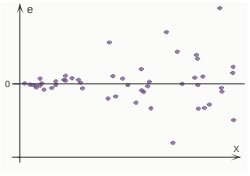
Plottas ofta mot de anpassade värdena $\hat{y}_i$ men även mot varje prediktorvariabel x för sig.



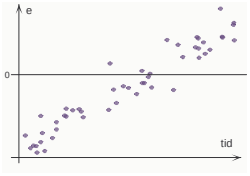
Uppfyller villkoren?



Mönster ska inte finnas!
Kvadratiskt term verkar här!

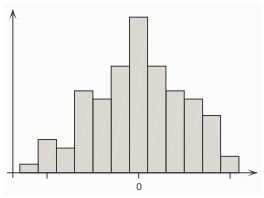


Ej lika varians
för alla X!

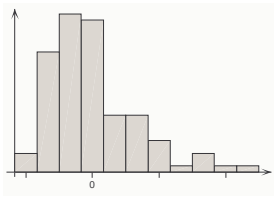


Tidsberoende!
(Minitab: Residuals v. Order)

Histogram över residualerna:



Normalfördelat?

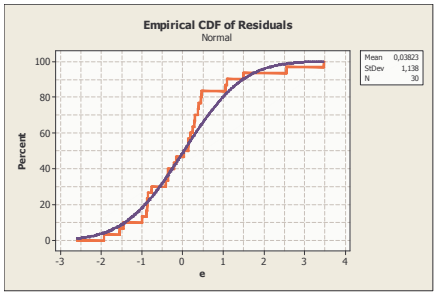


Normalfördelat?

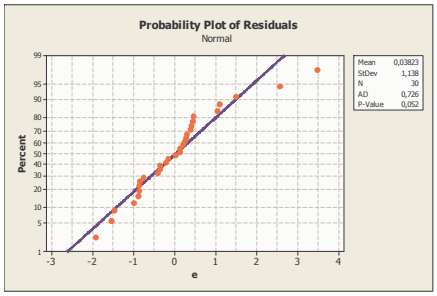
Plottar och histogram kan vara svåra att bedöma, medför ett subjektivt element.

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)

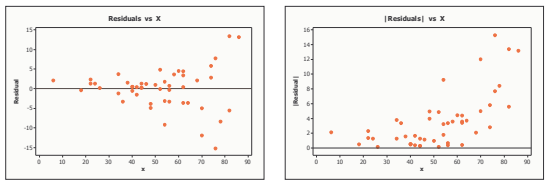
Finns några (se sid 227). Förutsätter (ibland) flera observationer per observerat x-värde.

Enkelt alternativ: studera *rangkorrelationen* mellan

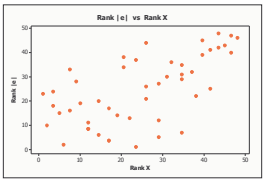
$$|e_i| \quad \text{och} \quad x_i$$

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



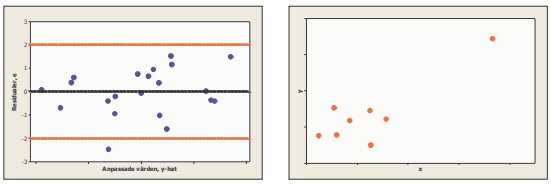
Rangkorrelationen blir i detta fall 0.624;

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:

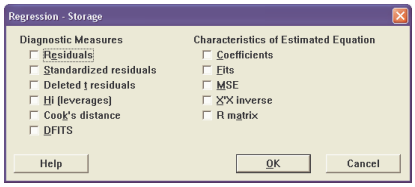
$$r_{(-i)} = \frac{e_i}{s^2_{(-i)}\sqrt{1-h_i}}$$

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)

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

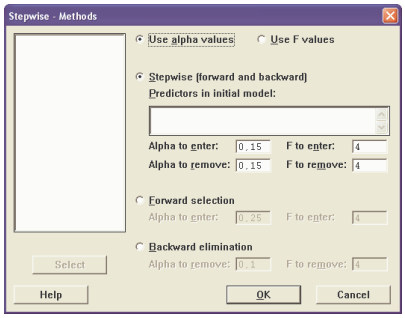
olika modeller att skatta och analysera!

- 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)

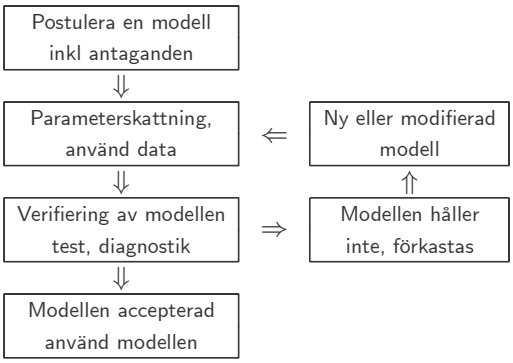
5. Validering, bedömning av modellens tillförlitlighet

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:



.