Erling Berge
POL3507 IMPLEMENTERING OG EVALUERING AV OFFENTLEG POLITIKK

Regresjonsanalyse

Ref.: L. B. Mohr 1995 Chapter 5 and 6

Literature

- Breen, Richard 1996 "Regression Models. Censored, Sample Selected, or Truncated Data", Sage University Paper: QASS 111, London, Sage
- Hamilton, Lawrence C. 1992 "Regression with graphics", Belmont, Duxbury, Kap. 1-7
- Hardy, Melissa A. 1992 "Regression with dummy variables" Sage University Paper: QASS 93, London, Sage,
- Mohr, Lawrence B. 1995 "Impact Analysis for Program Evaluation", Sage, London
- Winship, Christopher, and Robert D. Mare 1992 «Models for sample selection bias», Annual Review of Sociology, 18:327-350
- Winship, Chrisopher, and Stephen L. Morgan 1999 "The Estimation of Causal Effects from Observational Data", Annual Review of Sociology Vol 25: 659-707

Regression towards the mean

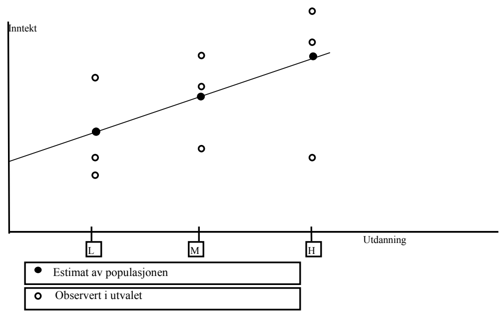
- Occurs in the posttest Y_2 because sample is selected based on high or low values of pretest Y_1
- In general: selection based on the dependent variable gives biased results like this
- $Y_1 = \alpha + \beta_{11}X1_1 + \beta_{12}X2_1 + ... + \varepsilon_1$
- Y_1 is large if X1 and/ or X2 and/ or the error term ε are large
- In measuring Y_2 we can control changes in X1 and X2 but not in ε
- This is a serious problem only if ε comprises important (unmeasured) causal variables (causing a correlation between Y og ε)

Bivariat Regresjon: Modell for populasjon

- $Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$
- i=1,...,n n = # case i populasjonen
- Y og X må definerast eintydig, og Y må ha målenivå **intervallskala** i ordinær regresjon

Bivariat Regresjon: Modell for utval

- $Y_i = b_0 + b_1 x_{1i} + e_i$
- i=1,...,n n = # case i utvalet
- Y og X må definerast eintydig, og Y må ha målenivå **intervallskala eller høvestalskala** (målevariabel) i ordinær regresjon



Oppsummering

- I bivariat regresjon kan ein seie at OLS-metoden freistar finne den beste LINJA eller KURVA som passar til eit to-dimensjonalt spreiingsmønster
- Scatter-plott og residualanalyse er hjelpemiddel for å diagnostisere problem i regresjonen
- Transformsjon er eit **generelt hjelpemiddel** mot fleire typar problem, som t.d.:
  - Kurvelinearitet
  - Heteroskedastisitet
  - Ikkje-normalitet i residualfordelinga
  - Case med stor innverknad
- Regresjon med transformerte variablar er alltid kurvelineær. Vi tolkar resultatet letast ved hjelp av grafar

Multippel regresjon: modell (1)

- Målet med multippel regresjon er å finne nettoeffekten av ein variabel, kontrollert for variasjonen i alle dei andre
- Sett K = talet på parametrar i modellen (dvs. K-1 er talet på variablar). Da kan (populasjons) modellen skrivast
- $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_{K-1} x_{i,K-1} + \varepsilon_i$

Multippel regresjon: modell (2)

- Dette kan skrivast $y_i = E[y_i] + \varepsilon_i$,

  dette tyder at
- $E[y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_{K-1} x_{i,K-1}$

  $E[y_i]$ les vi som forventa verdi av y_i

Multippel regresjon: modell (3)

- Vi finn OLS estimata av modellen som dei b-verdiane i
  $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + ... + b_{K-1} x_{i,K-1}$
  ($\hat{y}_i$ les vi som estimert eller ”predikert” verdi av y_i)
  som minimerer kvadratsummen av residualane

$$RSS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} e_i^2$$

Interaksjonseffektar i regresjon

- Ein interaksjonseffekt mellom x og w kan inkluderast i ein regresjonsmodell ved å ta inn ein hjelpevariabel lik produktet av dei to, dvs. Hjelpevariabel H=x*w
- $y_i = b_0 + b_1{*}x_i + b_2{*}w_i + b_3{*}H_i + e_i$
- $y_i = b_0 + b_1{*}x_i + b_2{*}w_i + b_3{*}x_i{*}w_i + e_i$

Example from Hamilton(p85-91)

Define

- y = natural logarithm of Chlorid consentration
- x = dept of well (1=deep, 0=shallow)
- w = natural logarithm of distance to road
- xw = interaction term between dept and distance to road (product x*w). Then
- $\hat{y}_i = b_0 + b_1 x_i + b_2 w_i + b_3 x_i w_i$

First take a look at the simple models without interaction

Figures 3.3 and 3.4 (Hamilton p85-86)

Figure 3.3 is based on

Dependent Variable: ln[ChlorideConcentra]	B	Std. Error	Beta	t	Sig.
(Constant)	3.775	.429		8.801	.000
x= Deep (DEEP OR SHALLOW WELL?)	-.706	.477	-.205	-1.479	.145

Figure 3.4 is based on

Dependent Variable: ln[ChlorideConcentra]	B	Std. Error	Beta	t	Sig.
(Constant)	4.210	.961		4.381	.000
w= ln[DistanceFromRoad]	-.091	.180	-.071	-.506	.615
x= Deep (DEEP OR SHALLOW WELL?)	-.697	.481	-.202	-1.449	.154

Figure 3.3

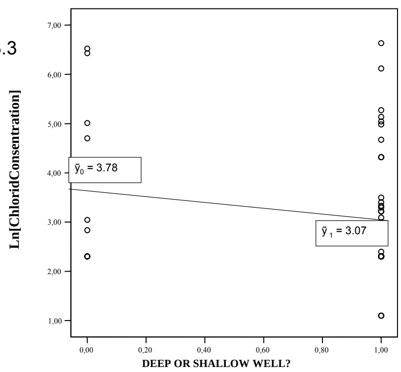
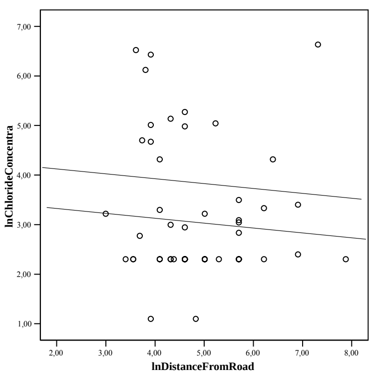


Figure 3.4



Figures 3.5 and 3.6 (Hamilton p89-91)

Figure 3.5 is based on

Dependent Variable: ln[ChlorideConcentra]	B	Std. Error	Beta	t	Sig.
(Constant)	3.666	.905		4.050	.000
w= ln[DistanceFromRoad]	-.029	.202	-.022	-.144	.886
w*x= ln[DistFromRoad]*Deep	-.081	.099	-.128	-.819	.417

Figure 3.6 is based on

Also see Table 3.4 in Hamilton p90 Dependent Variable: ln[ChlorideConcentra]	B	Std. Error	Beta	t	Sig.
(Constant)	9.073	1.879		4.828	.000
w= ln[DistanceFromRoad]	-1.109	.384	-.862	-2.886	.006
x= Deep (DEEP OR SHALLOW WELL?)	-6.717	2.095	-1.948	-3.207	.002
w*x= ln[DistFromRoad]*Deep	1.256	.427	1.979	2.942	.005

Figure 3.5

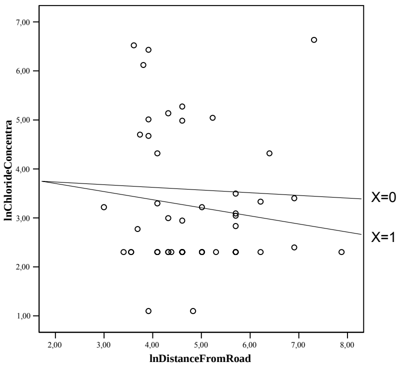
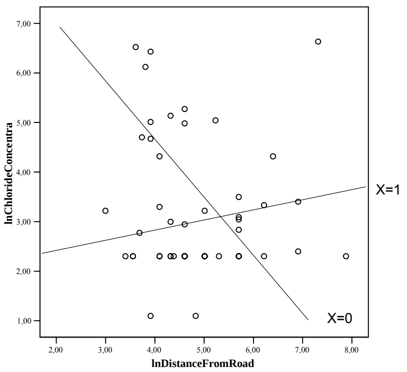


Figure 3.6



Adequasy ratio

- Adequacy = size of impact relative to size of problem, ie the proportion of the problem eleminated
- A= 1- R/C
- R = observed outcome of treatment
- C = ”counterfactual”: that is the outcome without the treament effect
- A = β_T/Y_{0E}

Comparisons

- Comparisons of gain scores are suspect
- Comparisons of proportional gain scores are suspect:
- Get individual level data and do regressions!

Test of significance

- Statistic
- Sampling distribution
- Model
  - $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_{K-1} x_{i,K-1} + \varepsilon_i$
- Testing β_i to determine confidence intervals
  - The statement

  $b_k - t_\alpha(SE_{b_k}) \le \beta_k \le b_k + t_\alpha(SE_{b_k})$

  is true with a probability of $1 - \alpha$

Footnote on t-test

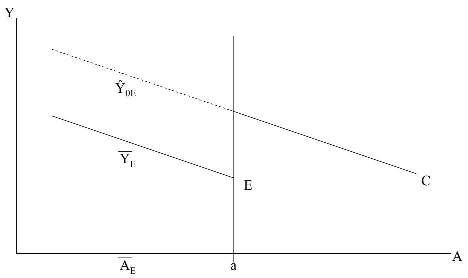
- Skilnaden mellom observert koeffisient (b_k) og uobserverte koeffisient (β_k) standardisert med standardavviket til den observerte koeffisienten (SE_{b_k}) vil normalt vere svært nær null dersom den observerte b_k ligg nær populasjonsverdien. Dette tyder at dersom vi i formelen
- $t = (b_k - \beta_k)/ SE_{b_k}$ set inn H_0: $\beta_k = 0$ og finn at "t" er liten vil vi tru at populasjonsverdien β_k eigentleg er lik 0. Kor stor "t" må vere for at vi skal slutte å tru at $\beta_k = 0$ kan vi finne ut frå kunnskap om samplingfordlingane til b_k og SE_{b_k}

Ways of testing

- Testing in causal models
- Testing in experimental designs: random assignment to treament group
- Testing of a descriptive statistic in surveys
- Assessing size of impact: What is large? Which is largest? Is it significanly different from zero?

Regression discontinuity design

- Assignment to treatmeent is not random and not autonomous, but controlled
- Selection is determined by the size of some measured quantitative variable A such that one group comprises those scoring below A=a and the other group scores above
- Assumes A measured without error and the relationship of Y and A is known



Why does assignment work?

- $Y = \beta_0 + \beta_A A + \beta_T T + u(X_2, u')$
- $X_2 = \beta_2 A + u'$
- $A \equiv \beta_2 A + (1- \beta_2)A$
- $u \equiv u' + (u-u')$
- $Y = \beta_0 +\beta_A [\beta_2 A +(1-\beta_2)A] + \beta_T T +u' +(u-u')$
- $Y = \beta_0 +\beta_A \beta_2 A +(\beta_A A +u') -\beta_A \beta_2 A +\beta_T T +(u-u')$
- Confounding impact of X_2 is removed from β_T

Threats to validity

- Random measurement error in A is not a problem
- Fuzzy cutpoints
  - Loss of testpossibilities and distortion of effect estimate
- Non-random mesurement error in A will often jeopardise the functional reltionship between A and X_2
- The functional form of the A and X_2 relationship is critical, but in most cases relations other than linear (or logistic for dichotomous A) would seem farfetched or unrealistic, particularly if pretest data are available as controls

Modelling selection

- In ordinary regression the remedy to problems of selection bias is to estimate a model of the selection process
- The regression discontinuity design is doing the same to an extreme degree: determining the selection process unambigously