SOS3003
Applied data analysis for social science
Lecture note 06-2009

Erling Berge
Department of sociology and political science
NTNU

Literature

- Regression criticism II
  Hamilton Ch 4 p109-137

Let us repeat some basics from last lecture:

Analyses of models are based on assumptions

- OLS is a simple technique of analysis with very good theoretical properties. But
- The good properties are based on certain assumptions
- If the assumptions do not hold the good properties evaporates
- Investigating the degree to which the assumptions hold is the most important part of the analysis

OLS-REGRESSION: assumptions

- I SPECIFICATION REQUIREMENT
  - The model is correctly specified
- II GAUSS-MARKOV REQUIREMENTS
  - (1) x is known, without stochastic variation
  - (2) Errors have an expected value of 0 for all i
  - (3) Errors have a constant variance for all i
  - (4) Errors are uncorrelated with each other

  (Ensures that the estimates are “BLUE”)
- III NORMALLY DISTRIBUTED ERROR TERM
  - Ensures that the tests are valid

Problems in regression analysis that cannot be tested

- If all relevant variables are included
- If x-variables have measurement errors
- If the expected value of the error is 0
- (This means that we are unable to check if the correlation between the error term and x-variables actually is 0 and is actually the same as the first point that we are unable to test if the model is correctly specified)

The most important problems in regression analysis that can be tested

- Non-linear relationships
- Non-constant error of the error term (heteroscedasticity)
- Autocorrelation for the error term
- Non-normal error terms

Heteroscedastisity

- Is present if the variance of the error term varies with the size of x-values
- Predicted y is an indicator of the size of x-values (hence scatter plot of residual against predicted y)
- Heteroscedasticity (non-constant variance of error term) can arise from
  - Measurement error (e.g. y more accurate the larger x is)
  - Outliers
  - The wrong functional form
  - If ε_i contain an important variable that varies with one or more x and y. The error term ε_i is not independent of the x-es. Hence the Gauss-Markov requirements 1 and 2 cannot be correct.

Indicators of heteroscedastisity

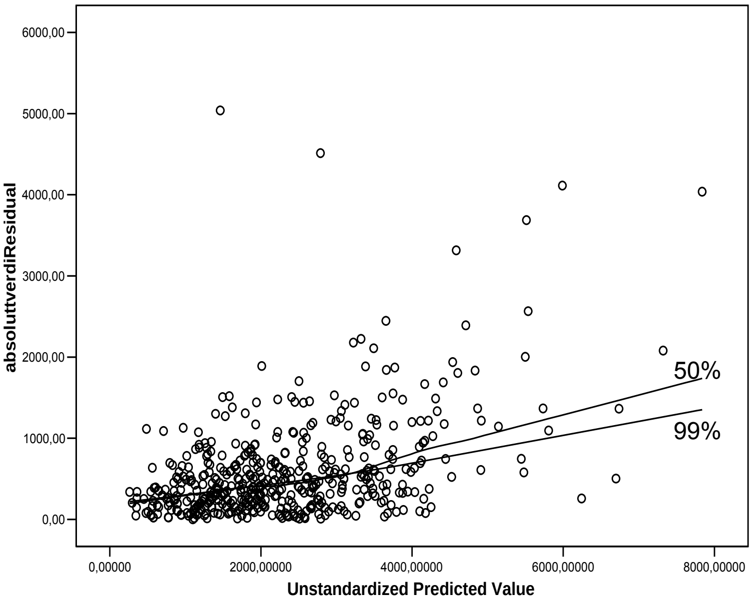
- Inspection of the scatter plot of residual against predicted value of y
- Band regression of the scatter plot

An interesting option here is:

- Locally weighted / ”sliding” regression on the central part of the sample

”Sliding” adapted line by means of locally weighted OLS regression

The procedure is called LOESS (see next slide)



A footnote: SPSS explains

Fit Lines

- In a fit line, the data points are fitted to a line that usually does not pass through all the data points. The fit line represents the trend of the data. Some fit lines are regression based. Others are based on iterative weighted least squares.
- Fit lines apply to scatter plots. You can create fit lines for all of the data values on a chart or for categories, depending on what you select when you create the fit line.

Loess

- Draws a fit line using iterative weighted least squares. At least 13 data points are needed. This method fits a specified percentage of the data points, with the default being 50%. In addition to changing the percentage, you can select a specific kernel function. The default kernel (probability function) works well for most data.

Autocorrelation

- Correlation among variable values on the same variable across different cases (e.g. between ε_i and ε_{i-1})
- Autocorrelation leads to larger variance and biased estimates of the standard error - similar to heteroscedasticity
- Autocorrelation is the result of a wrongly specified model
- Typically it is found in time series and geographically ordered cases. In a simple random sample from a population autocorrelation is improbable
- Tests (e.g. Durbin-Watson) is based on the sorting of the cases. Hence: hypotheses about autocorrelation need to specify the sorting order of the cases

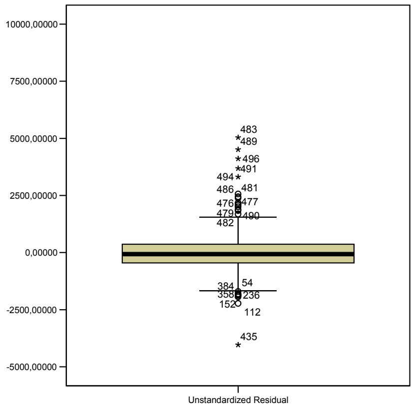
Non-normal residuals

- Imply that t- and F-tests cannot be used
- Since OLS estimates of parameters are easily affected by outliers, heavy tails in the distribution of the residual will indicate large variation in estimates from sample to sample
- We can test the assumption of normally distributed error term by inspecting the distribution of the residual, e.g. by inspecting
  - Histogram, box plot, or quantile-normal plot
  - There are also more formal tests (but not very useful) based on skewness and kurtosis

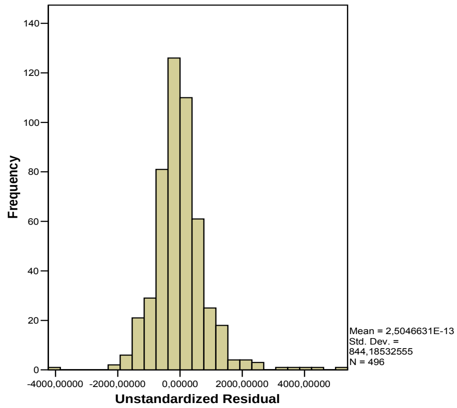
Diagram of the residual shows:

Heavy tails, many outliers, and weakly positively skewed distribution

BOX PLOT

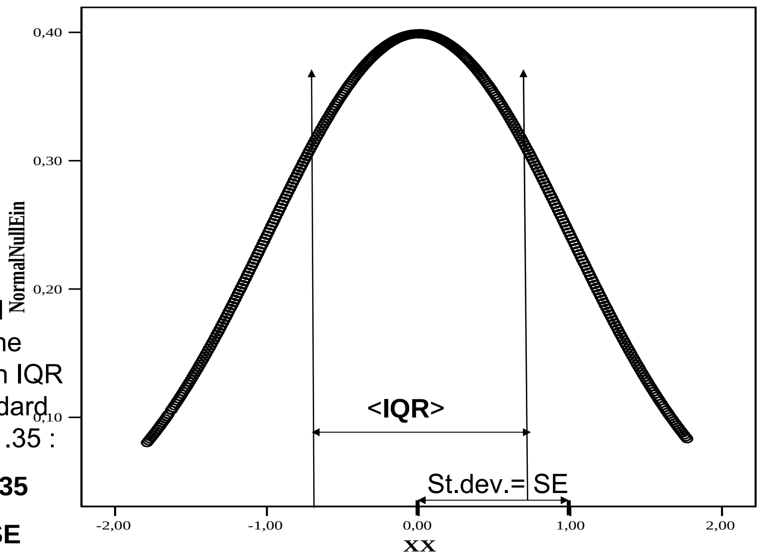


HISTOGRAM



Fall 2009 © Erling Berge 2009 13

Skewed distribution of the residual (1)



In the normal distribution the ratio between IQR and the standard deviation is 1.35 :

IQR/ SE = 1.35

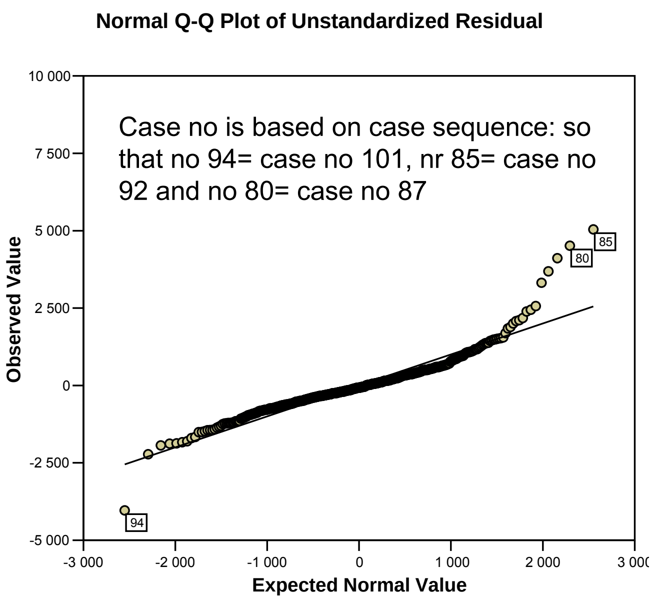
IQR/1.35 = SE

Fall 2009 © Erling Berge 2009 14

Skewed distribution of the residual (2)

- Since the average of the residuals (e_i) always equals 0, the distribution will be skewed if the median is unequal to 0
- It is known that for the normal distribution the standard deviation (or the standard error) equals approximately IQR/1.35
- If the distribution of the residual is symmetric we can compare SE_e to IQR/1.35. If
  - SE_e > IQR/1.35 the tails are heavier than the normal distribution
  - $SE_e \approx$ IQR/1.35 the tails are approximately equal to the normal distribution
  - SE_e < IQR/1.35 the tails are lighter than the normal distribution

Quantile-Normal plot of residual from regression in table 3.2 in Hamilton



Options if non-normality is found

- Test out if the right function has been used
- Test out if some important variable has been excluded
  - If the model cannot be improved substantially, we may try transforming the dependent variable to symmetry
- Test out if lack of normality is caused by outliers or influential cases
  - If there are outliers, transforming of the variable where the case is outlier may help

Influence (1)

- A case (or observation) has influence if the regression result changes when the case is excluded
- Some cases have unusually large influence because of
  - Unusually large y-value (outliers)
  - Unusually large value on an x-variable
  - Unusual combinations of variable values

Influence (2)

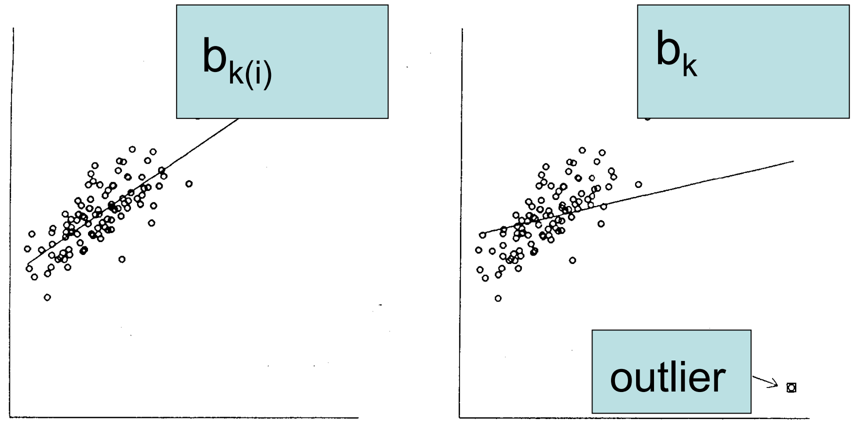
- We can see if a case has influence by comparing regressions with and without a particular case. One may for example
- Inspect the difference between b_k and $b_{k(i)}$ where case no i has been excluded in the estimation of the last coefficient
- This difference measured relative to the standard error of $b_{k(i)}$ is called $DFBETAS_{ik}$

$DFBETAS_{ik}$

$$DFBETAS_{ik} = \frac{b_k - b_{k(i)}}{\frac{s_{e(i)}}{\sqrt{RSS_k}}}$$

$s_{e(i)}$ is the standard deviation of the residual when case no i has been exclude from the analysis
RSS_k is Residual Sum of Squares from the regression of x_k on all other x-variables

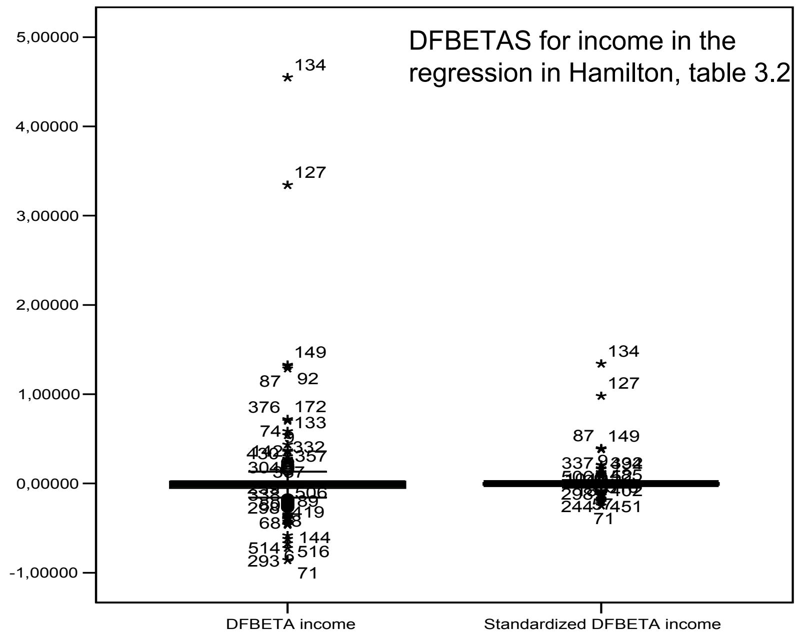
$DFBETAS_{ik}$:

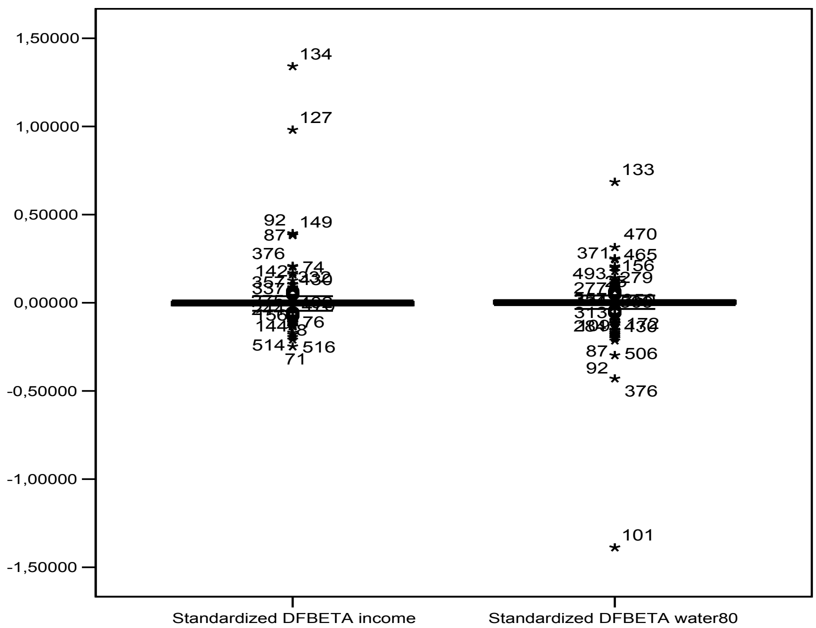


One case may make a lot of difference

What is a large DFBETAS?

- $DFBETAS_{ik}$ is calculated for every independent variable for every case. We do not want to inspect all values for it
- Three criteria for finding large values we need to inspect are
  - External scaling. $|DFBETAS_{ik}| > 2/ SQRT(n)$ $\sqrt{n}$
  - Internal scaling. Look for **severe outliers** in the box plot of $DFBETAS_{ik}$:
    $DFBETAS_{ik} < Q_1 - 3IQR$
    $Q_3 + 3IQR < DFBETAS_{ik}$
  - Gap in the distribution of $DFBETAS_{ik}$
- None of the $DFBETAS_{ik}$ needs to be problematic

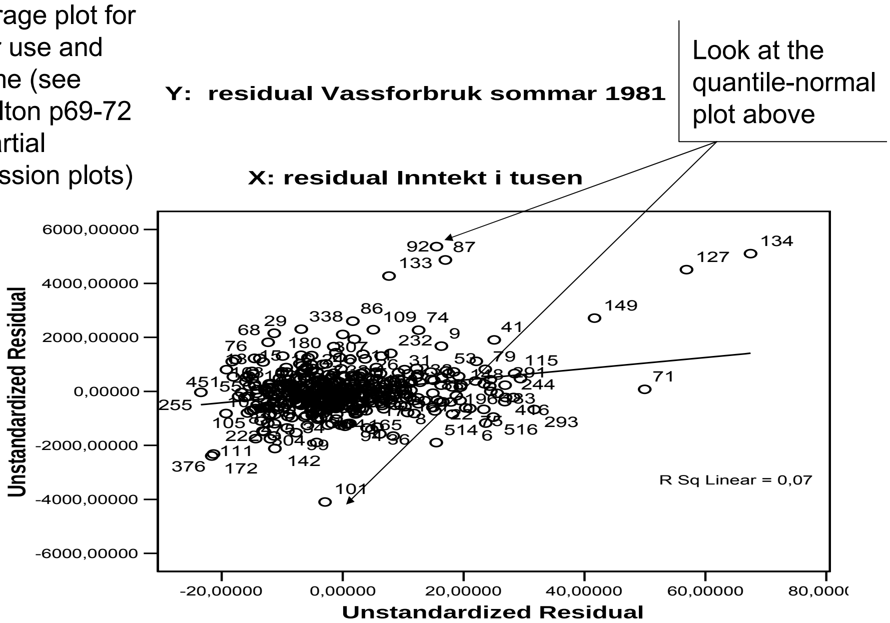

DFBETAS for income in the regression in Hamilton, table 3.2
5,00000
4,00000
3,00000
2,00000
1,00000
0,00000
-1,00000
134
127
149
92
87
376
172
133
144
516
514
293
71
134
127
87
149
71
DFBETA income
Standardized DFBETA income


1,50000
1,00000
0,50000
0,00000
-0,50000
-1,00000
-1,50000
134
127
92
149
87
376
514
516
71
133
470
371
465
156
493
279
87
506
92
376
101
Standardized DFBETA income
Standardized DFBETA water80

Sequence in the data set and case no is not the same. Case no is fixed. Variable values.

Sequence no	Case nr	water81	water80	water 79	educat	retire	peop 81	cpeop
91	98	1500	1300	1500	16	0	2	0
92	99	3500	6500	5100	14	0	6	0
93	100	1000	1000	2700	12	1	1	0
94	101	3800	12700	4800	20	0	5	0
95	102	4100	4500	2600	20	0	5	0
96	103	4200	5600	5400	16	0	5	-1
97	104	2400	2700	800	16	0	6	0
98	105	1600	2300	2200	14	0	4	0
99	107	2300	2300	3100	16	0	4	-2

Leverage plot for water use and income (see Hamilton p69-72 on partial regression plots)



Consequences of case with large influence

- If we discover case with large influence we should not necessarily remove them from the analysis
- Report results both with and without the cases
- Take a careful look at influential cases, maybe there are measurement errors
- When influential cases are outliers their influence can be reduced by transformation
- Use robust regression not so easily affected as OLS regression

Potential influence: leverage

- The potential for influence of a case from a particular combination of x-values is measured by the hat statistic h_i
- h_i varies from 1/n to 1. It has an average of K/n (K = # parameters)
- SPSS reports the centred h_i
  - i.e. $(h_i – K/n)$, we may call this for h^c_i

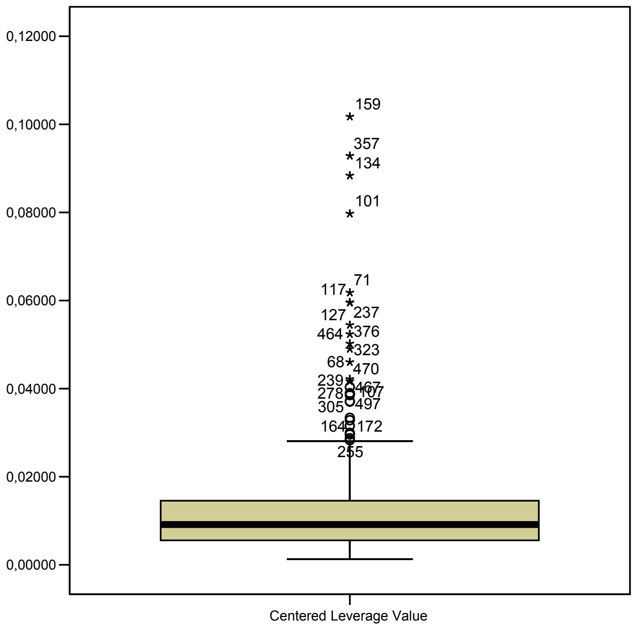
What is a large value of leverage?

- As for DFBETAS different criteria can be suggested. They all depend on the sample size n
  - If $h_i > 2K/n$ (or $h^c_i > K/n$) we find the ca 5% largest h_i ; alternatively
    - If max $(h_i) \leq 0.2$ there is no problem
    - If $0.2 \leq$ max $(h_i) \leq 0.5$ there is some risk for a problem
    - If $0.5 \leq$ max (h_i) probably there is a problem

Fall 2009 © Erling Berge 2009 29

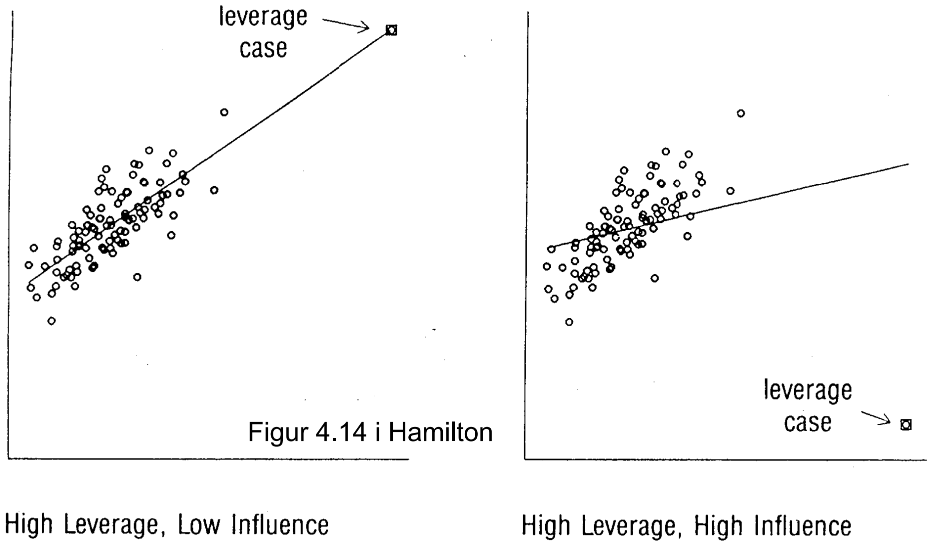
Centred leverage (h^c_i) from the regression in table 3.2 in Hamilton

Max av h^c_i er 0.102



Fall 2009 © Erling Berge 2009 30

The difference between influence and leverage



The leverage statistic is found in many other case statistics

- Variance of the i-th residual $\mathrm{var}[e_i] = s_e^2[1-h_i]$
- Standardized residual (*ZRESID in SPSS) $z_i = \dfrac{e_i}{s_e\sqrt{1-h_i}}$
- Studentized residual (*SRESID in SPSS) $t_i = \dfrac{e_i}{s_{e(i)}\sqrt{1-h_i}}$
- And remember that the standard deviation of the residual is $s_e = \sqrt{RSS/(n-K)}$

Total influence: Cook's D_i

- Cook's distance D_i measure influence on the model as a whole, not on a specific coefficient as $DFBETAS_{ik}$

$$D_i = \frac{z_i^2 h_i}{K\left(1-h_i\right)}$$

where z_i is the standardized residual

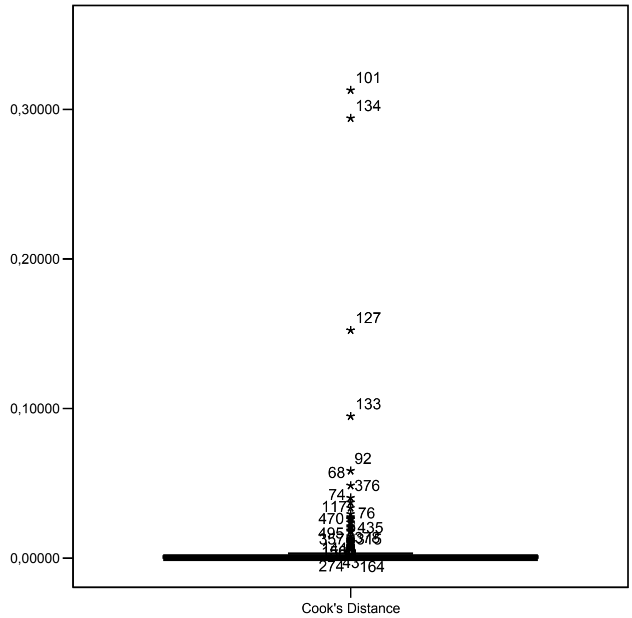
and h_i is the hat statistic (leverage)

What is a large D_i ?

- One might want to take a look at all
  - $D_i > 1$ or
  - $D_i > 4/n$ these are about the 5% largest D_i
- Even if a case has low D_i it may still be the case that it affects the size of single coefficients (it has a large $DFBETAS_{ik}$)

Cook's distance D_i from the regression in table 3.2 in Hamilton

Also see table 4.4 (p133) in Hamilton



Summarizing

What can be done with outliers and cases with large influence? We can

- Investigate if data are erroneous. If data are wrong the case can be removed from the analysis
- Investigate if transformation to symmetry helps
- Report two equations: with and without cases with unreasonable large influence
- Get more data

Multicollinearity

- Means very high intercorrelations among x-variables
- Check if parameter estimates are correlated
- Check if tolerance (the part of the variation of x that is not shared with other variables) is less than say 0.1. If so there may be a problem
- VIF = variance inflation factor = 1/tolerance
- If multicollinearity is caused by squaring of variables or interaction terms it should not be seen as problematic

Tolerance

- The amount of variation in a variable x_k unique to that variable is called the tolerance of the variable
- Let R^2_k be the coefficient of determination in the regression of x_k on all the rest of the x-variables. The other x-variables explain the proportion R^2_k of the variation in x_k.
- Then 1- R^2_k is the unique variation: tolerance= 1- R^2_k
- Perfect multicollinearity means that
  - $R^2_k = 1$ and tolerance = 0
- Low values of tolerance make regression results less precise (larger standard errors)

Variance Inflation Factor (VIF)

- The standard error of the regression coefficient b_k can be written

$$SE_{b_k} = \frac{s_e}{\sqrt{RSS_k}} = \frac{s_e}{\sqrt{\left(1-R_k^2\right)TSS_k}} = \sqrt{VIF}\,\frac{s_e}{\sqrt{TSS_k}}$$

- 1/tolerance = 1/(1-R^2_k) = VIF
- Other things being equal lower tolerance (larger VIF) for x_k will give higher standard error for b_k [SE increase with a factor equal to square root of VIF]

Indicators of multicollinearity

- The best indicators is tolerance or VIF (both are based on R^2_k)
- Other indicators are
  - Correlation among singe variables (not reliable)
  - Inclusion/ exclusion of single variables give large changes in the effect of other variables
  - Unexpected signs on the effects of some variable
  - Standardized regression coefficients larger than1 or less than -1
  - Correlation among parameter estimates

Tolerance and VIF from regression in table 3.2 in Hamilton

Dependent Variable: Summer 1981 Water Use	Unstandardized Coefficients		t	Sig.	**Collinearity Statistics**	
	B	Std. Error			**Tolerance**	**VIF**
(Constant)	242,220	206,864	1,171	,242		
Summer 1980 Water Use	,492	,026	18,671	,000	,675	1,482
Income in Thousands	20,967	3,464	6,053	,000	,712	1,404
Education in Years	-41,866	13,220	-3,167	,002	,873	1,145
head of house retired?	189,184	95,021	1,991	,047	,776	1,289
# of People Resident, 1981	248,197	28,725	8,641	,000	,643	1,555
Increase in # of People	96,454	80,519	1,198	,232	,957	1,045

What is low tolerance?

When $R^2_k > 0,9$ tolerance is < 0,1 and VIF > 10

Factor of multiplication for the standard error is the square root of VIF (ca 3.2 for $R^2_k = 0,9$)

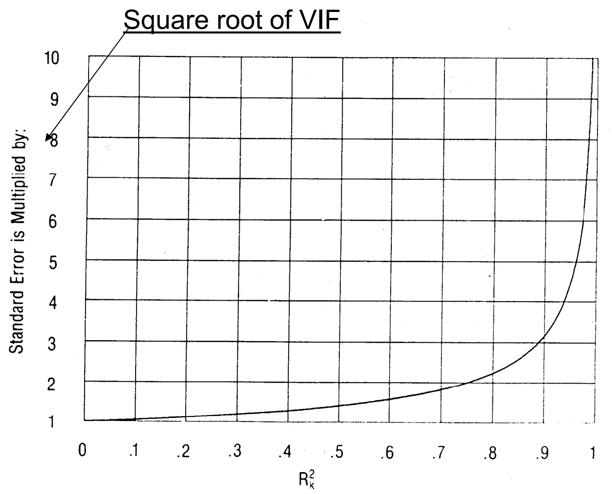


Figure 4.15 Effect of multicollinearity on standard errors (simplified).

When is multicollinearity a problem?

- It is not a problem if the reason is curvilinearity or interaction terms in the model. But in testing we need to take account of the fact that if VIF is high parameter estimates are imprecise (high standard errors). They are tested as a group by the F-test
- If the reason is that two variables measure the same concept one of them should be dropped, or they can be combined in an index
- It is a problem if we need estimates of the separate effects of two highly correlated variables (if a test of their joint effect is not sufficient)

Summarizing (1)

- When errors are independent and identically normally distributed OLS estimates are as good or better than other possible estimates
- But the assumptions are rarely satisfied completely, we have to test the degree to which they are satisfied
- Many problems can be corrected if we learn about them
- Check early on if curvilinearity, outliers or heteroscedasticity are problems (for example by use of scatter plots)

Summarizing (2)

- Do more exact investigations using residual/predicted Y plots and leverage plots
  - Curvilinearity (leverage plot, residual vs predicted Y plot)
  - Heteroscedasticity (leverage plot, [absolute value of residual] against predicted Y plot)
  - Non-normal residuals (quantile-normal plot, box-plot with analysis of median and IQR/1.35
  - Influence (check DFBETAS and Cook's D)
  - When we do not find serious problems we can have more confidence in our conclusions