Linear Regression with R/Stata

Day 4: Regression Assumptions

Constantin Manuel Bosancianu

February 28, 2019

Wissenschaftszentrum Berlin
Institutions and Political Inequality
bosancianu@icloud.com

Preamble

Recap from yesterday

- ✓ Categorical predictors are easily handled by OLS, under the form of dummy indicators. The most important part is always being aware of the reference category;
- ✓ Sampling variability depends on: (1) sample size, (2) spread of errors, (3) variance of the predictor;
- ✓ Large sampling variance generally results in a low t value $\Rightarrow$ lack of statistical significance;

Example: US union density 1982

	DV: Public sector union dens.
(Intercept)	33.21***
	(2.52)
Public sector coverage (yes)	7.47**
	(2.78)
Right-to-work law (yes)	−9.82**
	(2.83)
R^2	0.37
Adj. R^2	0.34
Num. obs.	50

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$. Standard errors in brackets.
DV (outcome) is union density in US states (0-100 range).

One more attempt at interpretation

Introduction

Why assumptions?

We can only “trust” the estimated *a*, *b*s and SEs if the data follows certain specifications.

Without these, we can’t be sure that the population effects are the same as the estimated sample effects.

Coefficients suffer from bias, and standard errors from inefficiency.

MIA (most important assumptions)

The residuals:

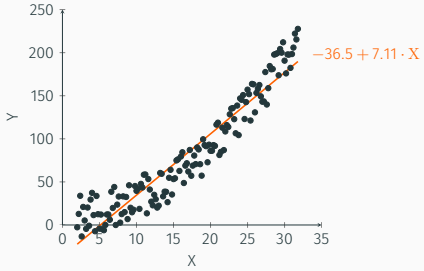
1. Average of the es is 0 along the length of Xs: $E(e|x_i) = 0$;
2. Variance is constant along the length of Xs: $V(e|x_i) = \sigma_e^2$. This is also called the assumption of "homoskedasticity"; when it does not hold, we are presented with "heteroskedasticity";
3. Errors are normally distributed: $e_i \sim \mathcal{N}(0, \sigma_e^2)$;
4. Errors are independent from each other: $cov(e_i, e_j) = 0$, for any $i \neq j$;
5. Predictors are measured without error, and are independent of the errors: $cov(X, e) = 0$.

Linearity

Linearity assumption

Two understandings:

- ✓ The bivariate relationship between X and Y is linear;
- ✓ The mean of e_i is 0 along the length of X.



Nonlinear relationship

The two are equivalent.

Diagnosing linearity

Plotting Xs against Y can detect some cases, but can miss some others for multiple regression.

The standard way is the *component-plus-residual plot* (in some texts, called the *partial-residual* plot).

For each observation for a specific predictor, compute

$$e_i^{(k)} = e_i + b_k X_k \tag{1}$$

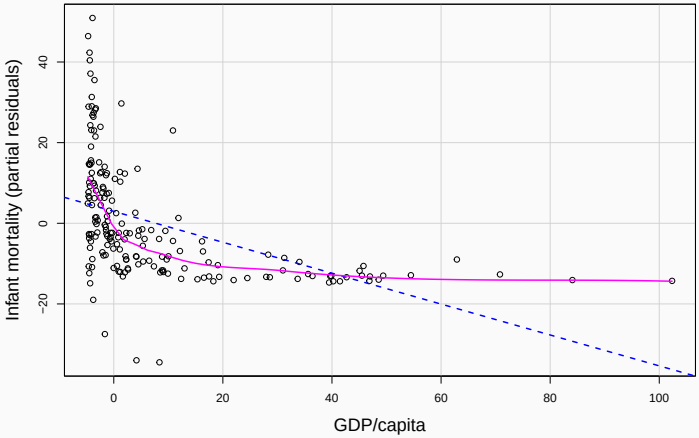
This is called the *partial residual*. Plotting $e_i^{(k)}$ against X_k should reveal any nonlinearity.

Example: infant mortality

	DV: Infant mortality
(Intercept)	19.34***
	(1.28)
GDP/capita (1,000s)	−0.38***
	(0.06)
Sub-Saharan Africa (yes)	30.70***
	(2.32)
R^2	0.63
Adj. R^2	0.63
Num. obs.	187

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$. Standard errors in brackets.
GDP/capita rescaled by subtracting 5,000 USD.
DV (outcome) is no. of deaths per 1,000 live births.

Example: infant mortality



Component-plus-residual plot (solid line is a *lowess* fit).

Addressing linearity

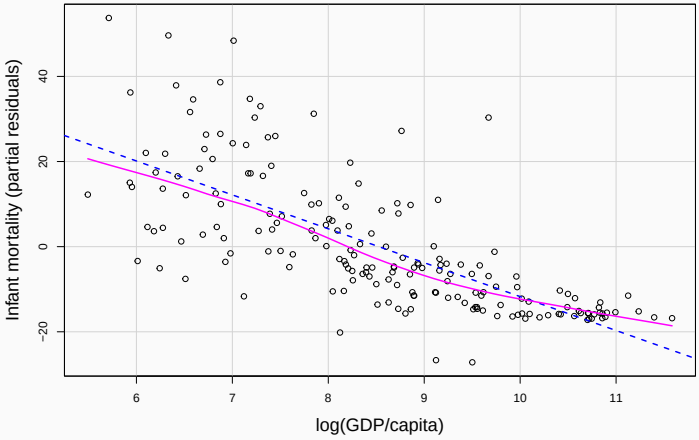
A standard solution is to transform one of the predictors; in our case, this is GDP/capita.

	Untransformed GDP	Transformed GDP
(Intercept)	19.34 (1.28)***	86.83 (6.19)***
GDP/capita (1,000s)	−0.38 (0.06)***	
Sub-Saharan Africa	30.70 (2.32)***	20.87 (2.25)***
log(GDP/capita)		−7.96 (0.68)***
R^2	0.63	0.74
Adj. R^2	0.63	0.74
Num. obs.	187	187

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$. Standard errors in brackets.
GDP/capita rescaled by subtracting 5,000 USD.
DV (outcome) is no. of deaths per 1,000 live births.

Problematic nonlinearity for infant mortality regression

Addressing linearity



Component-plus-residual plot (GDP is transformed).

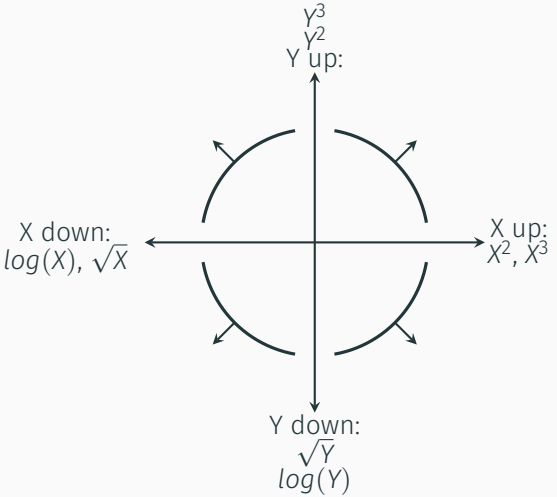
Addressing linearity (cont.)

Another strategy would be to change the functional form of the model.

In our case, this would mean adding both GDP/capita and a squared GDP/capita (the latter is called a *quadratic term*).

The squared version can be considered a multiplicative interaction, which would show how the slope of GDP/capita changes depending on ...GDP/capita.

Mosteller and Tukey's rules



Mosteller and Tukey's (1977, p. 84) set of rules for transformations.

Transformations for univariate distributions

Positive skew

Moderate: $NEW = \sqrt{OLD}$

Substantive: $NEW = \log_{10} OLD$

Substantive (with 0s): $NEW = \log_{10}(OLD + c_1)$

Negative skew

Moderate: $NEW = \sqrt{c_2 - OLD}$

Substantive: $NEW = \log_{10}(c_2 - OLD)$

c_1: constant added so that smallest value is 1.

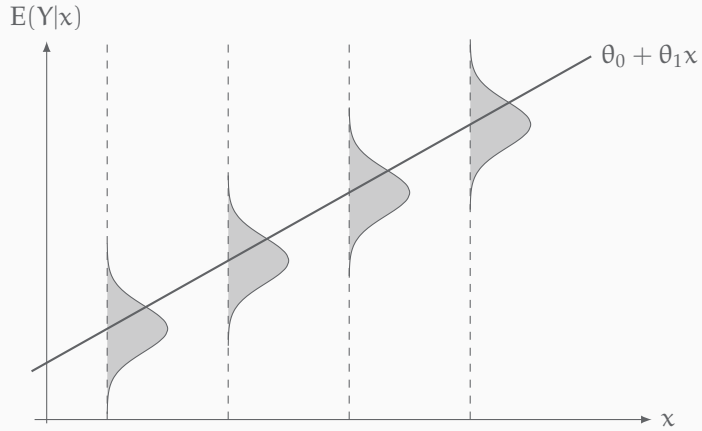
c_2: constant from which old values are subtracted so that smallest new value is 1.

Naturally, transformations also imply a change in interpretation of the transformed variable.

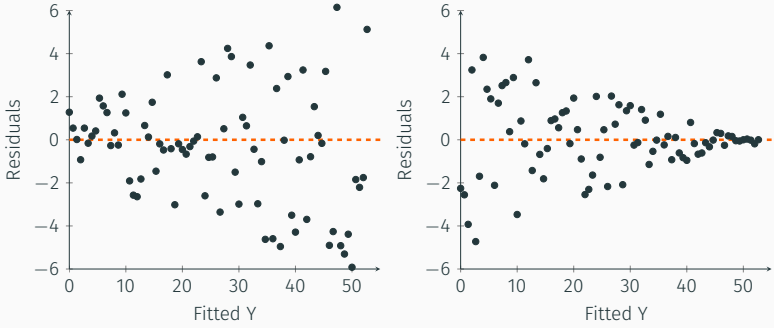
Homoskedasticity

Homoskedasticity

The spread of e_i should be constant along the length of $\hat{Y}$.

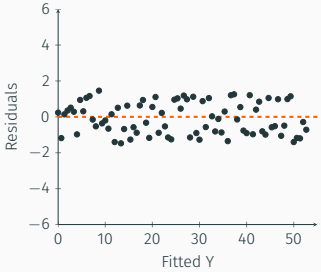


Heteroskedasticity



a and *b*s are unbiased, but their SEs are imprecise, which means significance tests are affected.

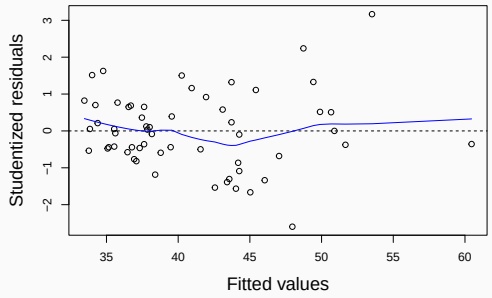
Diagnosing heteroskedasticity

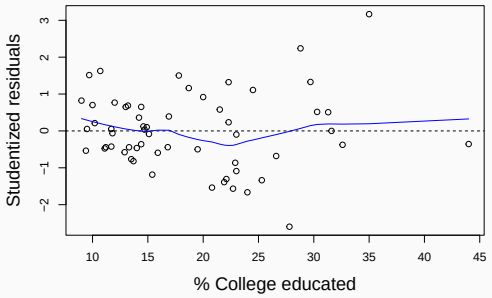


Is σ_e^2 constant?

- ✓ a plot of studentized residuals versus fitted values ($\hat{Y}$);
- ✓ a plot of studentized residuals versus predictors (X_k).

Example: California 1992

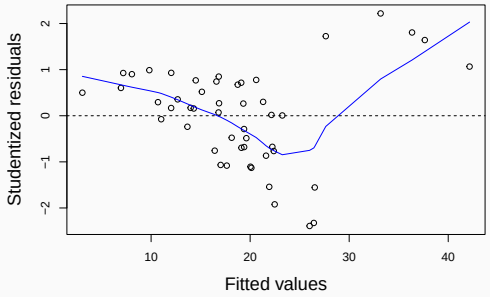


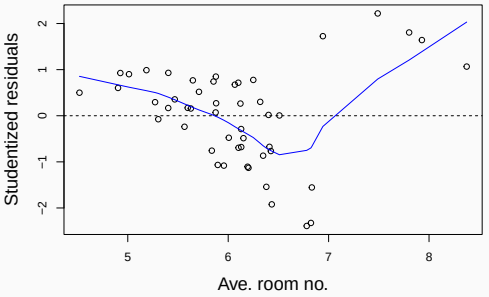


No clear evidence of heteroskedasticity.

Take the case of Boston house prices, but with the data at the neighborhood level, and from only 2 towns: Cambridge and Roxbury.

Example: Cambridge and Roxbury house prices





Clear heteroskedasticity: the variance in the middle of the plot is considerably larger than at the left edge.

Breusch–Pagan test

Can be used for a summary diagnosis.

It uses a form of standardized squared residuals, which are then regressed on a set of predictors to produce fitted residuals. These predictors can be the original *X*s, but can also include other variables.

The test value, computed with the fitted residuals, has a χ^2 distribution with k degrees of freedom (k is the # of predictors from the second regression).

H_0 for the test is that the data is homoskedastic.

Breusch–Pagan test

```
studentized Breusch-Pagan test

data:  model1
BP = 14.768, df = 1, p-value = 0.0001216
```

For our example with Cambridge and Roxbury, the test value is 14.768, with 1 degree of freedom.

p-value is $1.2 \times 10^{-4} \Rightarrow$ the test is statistically significant.

H_0 of homoskedasticity is rejected $\Rightarrow$ data is heteroskedastic.

Addressing heteroskedasticity

None of the solutions are particularly simple.

The first is *Weighted Least Squares*—the quantities $\frac{1}{e_i}$ are used as weights to re-estimate the model.

Observations with large e_i are down-weighted in this setup.

Addressing heteroskedasticity (cont.)

Since the SEs are the problem, we can do a correction on the SEs.

“Huber–White standard errors”, “robust standard errors”, “sandwich estimator” (Huber, 1967; White, 1980).

If the heteroskedasticity is caused by omitted variables in the model specification, though, then the Huber–White correction does not give us much.

In this case, Huber–White SEs provide accurate estimates of uncertainty for wrong estimates of effect (Freedman, 2006).

Addressing heteroskedasticity (cont.)

Is it due to an omitted variable?

In our case, I omitted a dummy variable and an interaction (data was collected from 2 different towns, and the slope is different in the 2 towns).

Including these two predictors improves things.

```
^^Istudentized Breusch-Pagan test

data:  model2
BP = 2.4962, df = 3, p-value = 0.476
```

H_0 of homoskedasticity cannot be rejected anymore.

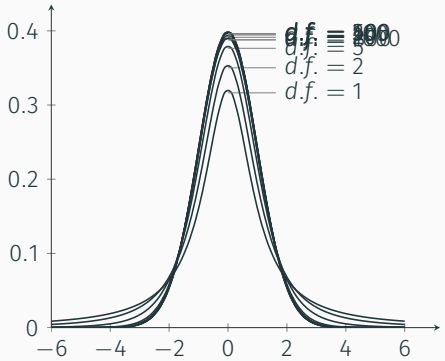
Addressing heteroskedasticity (cont.)

You ought not be deterred by small differences in residual variances.

Results are problematic only when σ_e at its highest level is about 2 or 3 times as large as σ_e at its lowest level (Fox, 2008).

Normality of errors

Normality



t distributions with varying degrees of freedom: 1, 2, 5, 20, 50, 100, 200, 400, 800, 1,600. Practically, the t distribution with 1,600 degrees of freedom can be considered a normal distribution.

In case errors are not normal, the SEs are affected.

Diagnosing normality

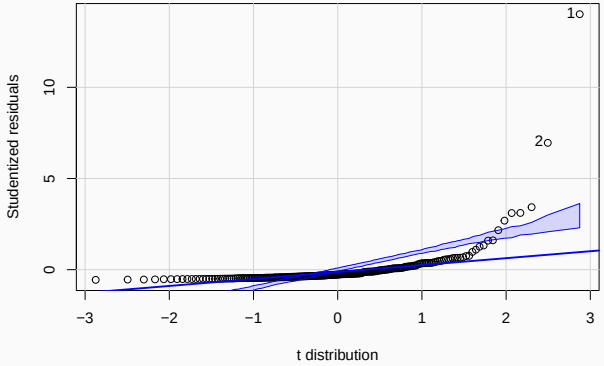
The standard tool is a quantile-comparison plot (Q–Q plot).

Logic: plot on horizontal axis where we would expect an observation to be, based on the normal distribution, and on the vertical where the observation actually is.

If our residuals are normally distributed, then the points ought to line up on a diagonal line in the graph.

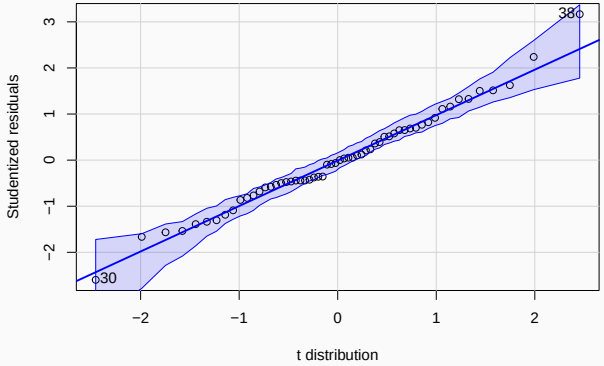
Useful to examine in particular the behavior of residuals at the tails of the distribution.

Example: *Fortune*'s 1992 billionaires



Non-normal errors in model of wealth for billionaires in 1992 ($N = 233$).
Specification: $Wealth = a + b_1 Age + e$.

Example: California 1992



Normal errors in model of Democratic vote % in California, 1992. Specification: $Vote = a + b_1 Education + e$.

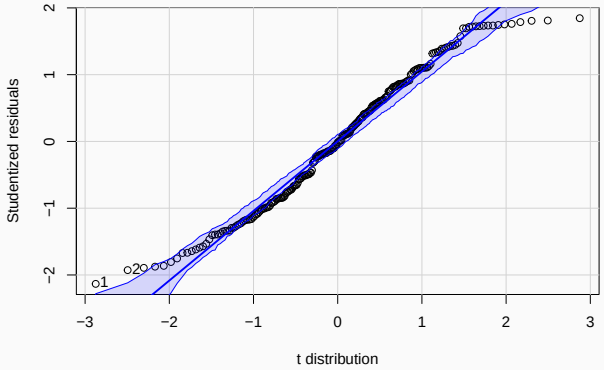
Addressing non-normality

A frequent cause for non-normal errors is non-normal predictors $\Rightarrow$ data transformations.

In our case, the “culprit” is wealth, which has a severe positive skew: most billionaires have between 1 and 3 billion USD, while the richest person in the world then had 37 billion USD.

The inverse transformation might work in this case, $\frac{1}{wealth}$, making the outcome into an index of “poverty”.

Example: *Fortune*'s 1992 billionaires



Normal errors in respecified model of wealth for billionaires in 1992 ($N = 233$).

Unusual and influential data

Outliers and high leverage cases

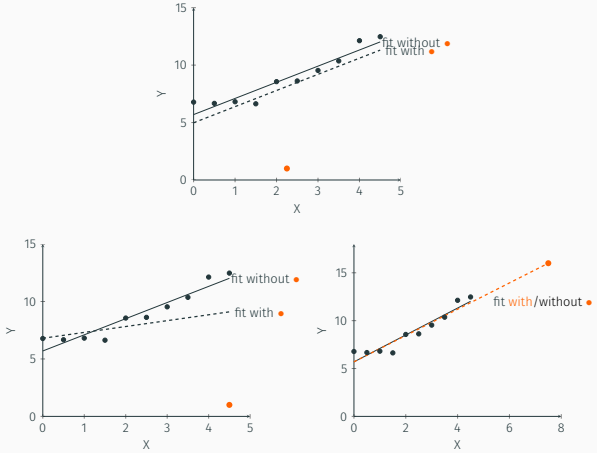
OLS estimates are easily influenced by outliers in the data.

Outlier: a case which, *given its value for X*, has an unusual value for *Y*.

(high) Leverage: a case with a value for *X* that is far away from the mean of *X*.

These two characteristics sometimes coincide, but not always.

Examples



Top panel: Outlier, but with low leverage. **Bottom left panel**: Outlier, with high leverage. **Bottom right panel**: High leverage, but not an outlier.

Influence on coefficients

$$Influence = Leverage \times Discrepancy \tag{2}$$

The case in the second panel has high influence (on the regression slope).

The case in the third panel is nevertheless problematic.

$$V(b) = \frac{\sigma_\epsilon^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3}$$

The sampling variance is “artificially” reduced in such cases.

Assessing leverage

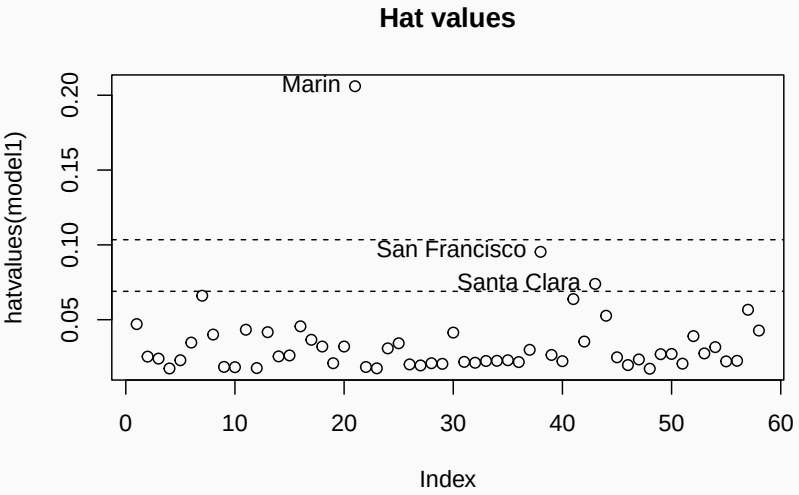
"Hat-values" are used.

It's possible to express every $\hat{Y}_i$ as a weighted sum of Y_i.

$$\hat{Y}_j = h_{1j}Y_1 + h_{2j}Y_2 + \cdots + h_{jj}Y_j + \cdots + h_{nj}Y_n \tag{4}$$

Any observation that has a h_{ij} larger than $2 \times \bar{h}$ or $3 \times \bar{h}$, should be considered a high leverage case.

Example: California 1992



Detecting outliers

Studentized residuals:

$$E_i^* = \frac{e_i}{S_{E(-i)}\sqrt{1-h_i}} \tag{5}$$

Computed from:

- ✓ OLS residuals, e_i;
- ✓ standard error of the regression, $S_E = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-k-1}}$;
- ✓ hat values, h_i.

Detecting outliers

Instead of the regression SE, S_E, we use the regression S_E without the i observation, $S_{E(-1)}$.

This makes the top and bottom part of Equation 5 independent of each other $\Rightarrow E_i^*$ has a t distribution ($n - k - 2$ degrees of freedom).

We're interested in the maximum value of the studentized residual, E_{max}^*.[1]

Example: California 1992

```
outlierTest(model1)

No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferroni p
38 3.166318           0.002518      0.14605
```

Observation 38 = San Francisco.

The *Bonferroni-adjusted p*-value suggests that it's not unusual to see a residual of such magnitude in a sample of 58 observations.

Assessing influence

A large number of quantities have been proposed.

$DFBETA_{ij}$: a distance between the OLS estimate with and without a particular observation i in the sample.

A derivate measure for influence is $DFBETAS_{ij}$, which simply standardizes the D_{ij}.

The problem with both is that each measure can be computed for each observation and each predictor.

Cook (1977) introduces a distance measure based on the *standardized* residuals, which applies only to observations: Cook's D.

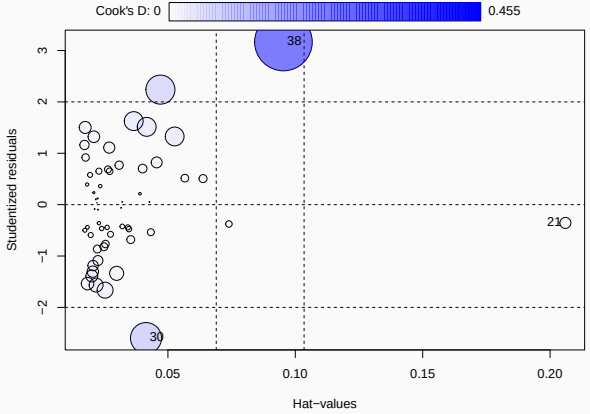
Cook's D

$$D_i = \underbrace{\frac{E_i'^2}{k+1}}_{\text{discrepancy}} \times \underbrace{\frac{h_i}{1-h_i}}_{\text{leverage}} \tag{6}$$

k is the number of predictors in the model, and h_i is the hat-value for observation i. $E_i'^2$ represent the squared standardized residuals^2, where

$$E_i' = \frac{\epsilon_i}{S_E\sqrt{1-h_i}} \tag{7}$$

Bubble plot



"Bubble plot" of hat-values versus studentized residuals (the area of the circle is proportional to Cook's *D*).

Additional assumptions

Other assumptions

Encountered frequently (Berry, 1993), but without a clear solution:

- ✓ some depend on the researcher's background knowledge, and not on a statistical fix, e.g. measurement error;
- ✓ for others the solutions are not straightforward, and sometimes depend on learning more advanced procedures, e.g. collinearity.

No specification error

The assumption requires that the estimated model is *complete*: excludes variables that ought not to be there, and includes all relevant variables.

Theory provides a list of variables.

Take an example. Although the full model is:

$$Y = a + b_{11}X_1 + b_{12}X_2 + \epsilon_1 \tag{8}$$

we can only test:

$$Y = a + b_{21}X_1 + \epsilon_2 \tag{9}$$

Effects of mis-specification

Further assume that X_1 and X_2 are weakly correlated.

- ✓ X_2 excluded from the model
- ✓ X_2 is correlated with X_1
- ✓ X_2 has a partial effect on Y

The effect of X_2 is now part of $X_1 \Rightarrow b_{21} \neq b_{11}$.

Diagnosis

There is a test: Ramsey's RESET (Regression Equation Specification Error Test).

This is limited, as it only refers to functional specification, and tests for any omitted non-linear predictors.

Ultimately, it's down to knowing the theory and having the right data available.

No measurement error (in the predictors)

Measurement error in the outcome can be accommodated in OLS.

e_i have a non-normal distribution, but *a* and *b*s are still BLUE (*Best Linear Unbiased Estimators*).

BLUE requires only the assumption of linearity, homoskedasticity, and error independence.

Measurement in the predictors, though, impacts the estimates.

No measurement error (in the predictors)

Measurement error in the predictors tends to bias coefficients downward (they are smaller than they should be).

Ultimately, probably all indicators have some measurement error to them.

Two aspects are important:

- ✓ the size of the error;
- ✓ whether it's random or systematic.

No measurement error (in the predictors)

Random error $\Rightarrow$ a and bs are unbiased, but SEs are larger, and R^2 is lower (Berry, 1993, p. 51).

Systematic error $\Rightarrow$ even the a and bs are biased.

No magic bullet: put a lot of time in concept operationalization.

No autocorrelation

Particularly salient in time-series analysis.

e_t tends to correlate with e_{t+1} because some phenomena exhibit slow change and multi-year trends (e.g. unemployment, GDP/capita).

A test is available: the Durbin-Watson test.

$$D = \frac{\sum_{t=2}^{n} (\epsilon_t - \epsilon_{t-1})^2}{\sum_{t=1}^{n} \epsilon_t^2} \tag{10}$$

No autocorrelation

For a large n, $D \approx 2(1 - \rho)$, where ρ is the correlation coefficient between ϵ_t and ϵ_{t+1}.

A $D \approx 2$ suggests that $\rho \approx 0$, which is ideal.

The limits of D are 0, when $\rho = 1$, and 4, when $\rho = -1$.

No (perfect) collinearity

The formula for sampling variance of a particular predictor, X_j, in multiple regression had the VIF:

$$VIF = \frac{1}{1 - R_j^2} \tag{11}$$

R_j^2 is the model fit from a regression of X_j on all the other predictors in the model.

The higher the correlation between X_j and another predictor, the higher the $R_j^2 \Rightarrow$ high VIF $\Rightarrow$ high sampling variance.

Large SEs means that there isn't enough (independent) information to properly estimate b.

Solutions: high collinearity

Yet again, no magic bullet:

- ✓ create an index, if it's theoretically plausible;
- ✓ drop a variable from the model, and risk mis-specification error;
- ✓ collect more data, to estimate b with more precision;
- ✓ ridge regression: accept a bit of bias in your coefficients, for a larger gain in efficiency.

Thank you for the kind attention!

Belsley, D. A., Kuh, E., & Welsch, R. E. (2004). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Hoboken, NJ: Wiley-Interscience.

Berry, W. D. (1993). *Understanding Regression Assumptions*. Thousand Oaks, CA: Sage Publications.

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*, *19*(1), 15–18.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models* (2nd ed.). Thousand Oaks, CA: Sage.

Freedman, D. A. (2006). On The So-Called "Huber Sandwich Estimator" and "Robust Standard Errors". *The American Statistician*, *60*(4), 299–302.

Huber, P. J. (1967). The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 221–233). Berkeley, CA: University of California Press.

Mosteller, F., & Tukey, J. W. (1977). *Data Analysis and Regression. A Second Course in Statistics*. Reading, MA: Addison-Wesley.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, *48*(4), 817–838.