Advanced Topics in Applied Regression

Day 1: OLS basics & assumptions

Constantin Manuel Bosancianu

Doctoral School of Political Science
Central European University, Budapest
bosancianu@icloud.com

July 31, 2017

Welcome!

Thanks for taking the class!

Reasoning:

- ✓ Some of these topics you have to have in your toolbox, e.g. interactions or fixed effects;
- ✓ Others you need so as to avoid more complex procedures, e.g. clustered SEs, which, *in certain specific cases*, can obviate the need for MLM.

Setup

Lecture + lab, both of which will be highly interactive. Ask questions and bring examples from data sets or research projects that you are working on!

R syntax supplied by myself (usually the morning of the class).

We will make some *moderate* use of statistical notation.

Why notation?

You will encounter it in a lot of quantitative literature, so it’s good to get familiar with it early.

Some statistical topics you will have to learn on your own, so it’s best if you get used with the symbols which many books use.

It slows you down a bit in the short term, but makes things faster in the long term.

Glossary (I)

- ✓ X, Y: variables
- ✓ $\overline{X}$: mean of X
- ✓ σ_X^2: variance of X, computed as $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
- ✓ n: sample size

Glossary (II)

- $cov(X, Y)$: covariance between X and Y
  - computed as $\frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$
- r_{XY}: correlation between X and Y, if both are continuous variables[1]
  - computed as $\frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{\frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sigma_x \sigma_y}$

[1] Other types of correlations have different symbols, such as ρ.

Regression recap

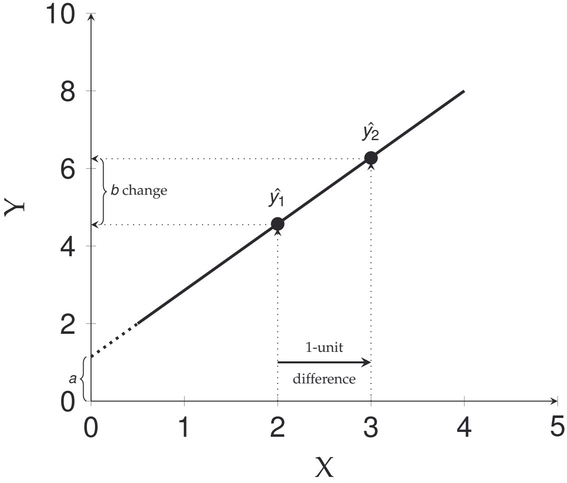
Standard multiple regression

$$Y = a + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k + e \tag{1}$$

Each individual in the sample has a different value on Y and $X_1, X_2, \ldots, X_k$.

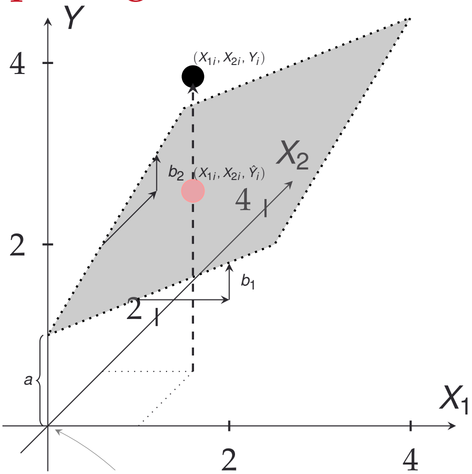
This means that the es (errors, residuals) will also be different for each individual.

Visualizing simple regression



Slope interpretation (I call the predicted value of Y for x_i as $\hat{y}_i$)

Visualizing multiple regression



(0,0,0)
origin

Model fit

Most common is R^2, interpreted as the share of the variance in Y explained by the influence of $X_1, X_2, \ldots X_k$.

$$Adjusted\ R^2 : \tilde{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1} \tag{2}$$

For the adjusted R^2 R uses the “Wherry Formula -1”.

$$Residual\ SE : \sigma_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n - k - 1}} \tag{3}$$

σ_e is an alternative measure of fit, interpreted as a sort of “average residual” (sadly, it’s often not reported).

Inference with regression

For simple regression:

$$V(b) = \frac{\sigma_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma_e^2}{(n-1)\sigma_x^2} \qquad (4)$$

- ✓ larger n means smaller $V(b)$;
- ✓ as σ_e^2 increases, so does $V(b)$;
- ✓ as $\sum_{i=1}^{n}(x_i - \bar{x})^2$ increases, $V(b)$ gets smaller.

Inference for multiple regression

$$V(b_j) = \underbrace{\frac{1}{1 - R_j^2}}_{\text{VIF}} \times \frac{\sigma_e^2}{\sum_{i=1}^{n}(x_j - \bar{x}_j)^2} \tag{5}$$

The second part is the same as for simple regression. The first part is called the *variance inflation factor* (VIF).

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

Introduction

Why assumptions

We can only “trust” the estimated a, bs 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.

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";[2]
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: $cov(X, e) = 0$.

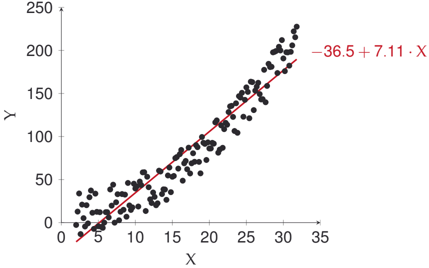
[2] The violation of this assumption if called "heteroskedasticity". Sometimes you encounter this term with a "c" instead of a "k".

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*.[3]

For each observation for a specific predictor, compute

$$e_i^{(k)} = e_i + b_k X_k \tag{6}$$

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

[3] In other texts you will encounter it as the *partial-residual* plot.

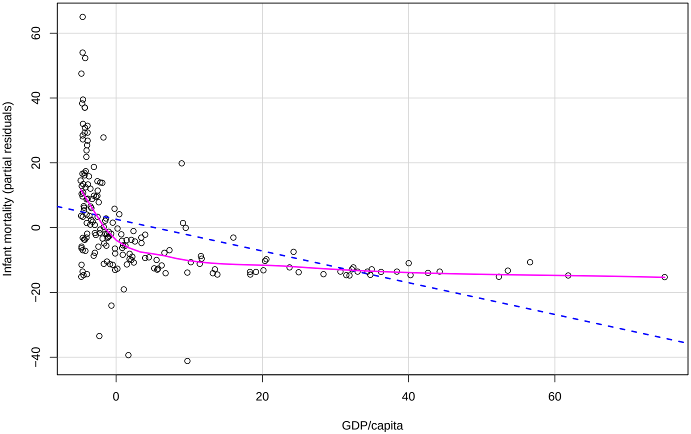
Example: infant mortality

Model for infant mortality

	DV: Infant mortality
(Intercept)	19.55^{***}
	(1.40)
GDP/capita (1,000s)	-0.49^{***}
	(0.08)
Sub-Saharan Africa (yes)	35.39^{***}
	(2.65)
R^2	0.63
Adj. R^2	0.63
Num. obs.	183

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$. GDP/capita rescaled by subtracting 5,000 USD.

Example: infant mortality



Component-plus-residual plot for infant mortality regression (solid line is a *lowess* fit).

Addressing linearity

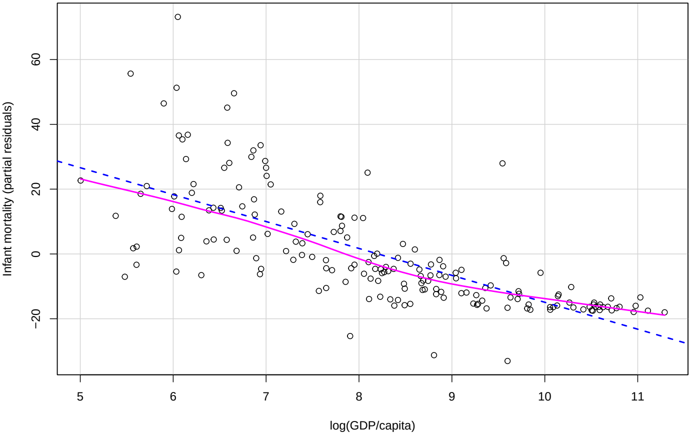
A standard solution is to transform one of the predictors; in our case, this is GDP/capita.

	Model 1	Model 2
(Intercept)	19.55***	87.86***
	(1.40)	(6.29)
GDP/capita (1,000s)	-0.49^{***}	
	(0.08)	
Sub-Saharan Africa	35.39***	24.53***
	(2.65)	(2.52)
log(GDP/capita)		-8.31^{***}
		(0.71)
R^2	0.63	0.74
Adj. R^2	0.63	0.74
Num. obs.	183	183

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$. GDP/capita used in its original form.

Comparison of models—problematic nonlinearity for infant mortality regression

Addressing linearity



Component-plus-residual plot for infant mortality regression, with changed specification.

Addressing linearity (cont.)

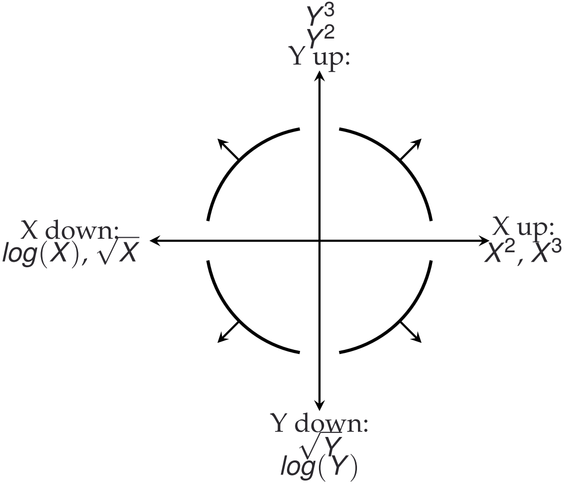
Another strategy would be to change the functional form of the model.

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[4], which would show how the slope of GDP/capita changes depending on …GDP/capita.

[4] More on this Wednesday.

Mosteller and Tukey's rules



Y^3

Y^2

Y up:

X down:

$log(X)$, $\sqrt{X}$

X up:

X^2, X^3

Y down:

$\sqrt{Y}$

$log(Y)$

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.

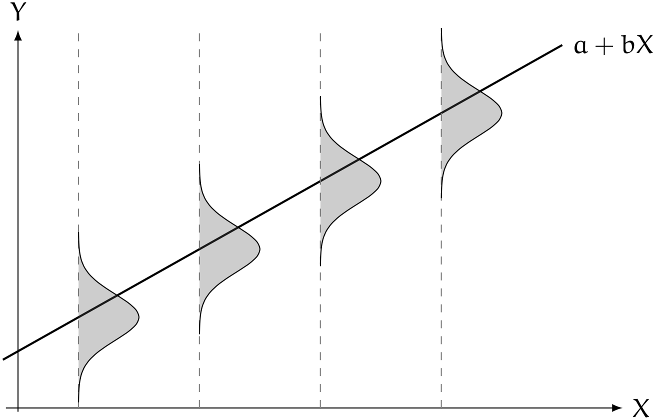
c_2: constant from which old values are subtracted so that smallest new value is 1.

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

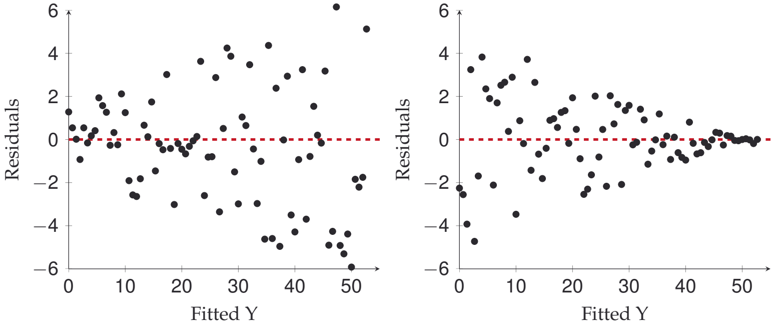
Homoskedasticity

Homoskedasticity

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

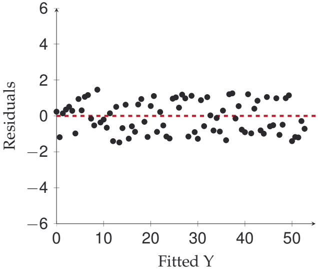


Heteroskedasticity



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

Diagnosing heteroskedasticity

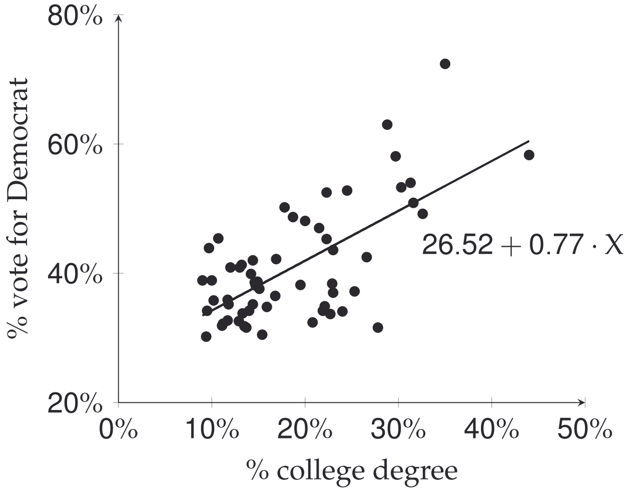


Is σ_e^2 constant?

- ✓ a plot of studentized residuals versus fitted values ($\hat{Y}$);[5]
- ✓ a plot of studentized residuals versus predictors (X_k).

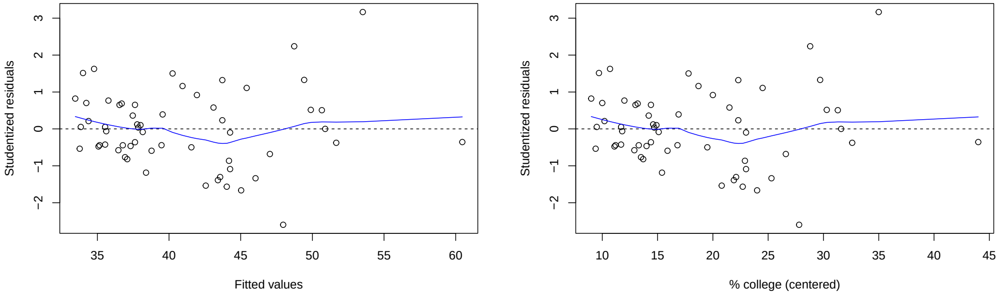
[5] Using Y would result in a tilted plot, which is a bit harder to interpret; $\hat{Y}$ and the studentized residuals are uncorrelated, though, so the plot will be "flat".

Example: California counties in 1992



OLS estimates: education and vote choice (CA 1992)

Example: California counties in 1992



No clear evidence of heteroskedasticity.

Take another case: average Boston house prices, at the neighborhood level. The goal is to understand what influences the price.

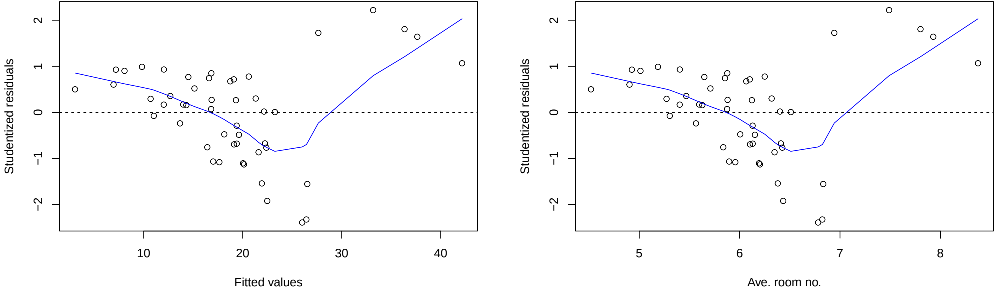
Example: Boston house prices

	DV: House price (ave.)
(Intercept)	-42.757^{***}
	(9.620)
Average num. rooms	10.139^{***}
	(1.568)
R^2	0.471
Adj. R^2	0.460
Num. obs.	49

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Predicting house price using number of rooms

Example: Boston house prices



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

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 doesn’t give us much.

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

Addressing heteroskedasticity (cont.)

Use background knowledge about the topic to find whether heteroskedasticity is due to an omitted variable.

In our case, I omitted a dummy variable and an interaction, since the data was collected from 2 different towns.[6]

Including these two predictors in the model results in a better distribution of the errors.

[6] And the slope is different in the two towns.

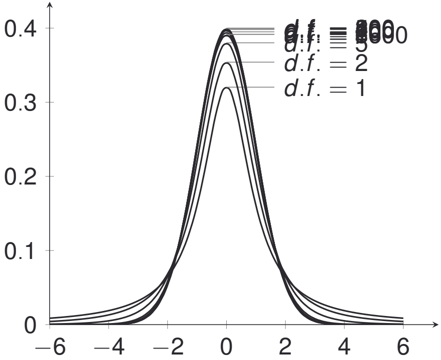
Addressing heteroskedasticity (cont.)

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

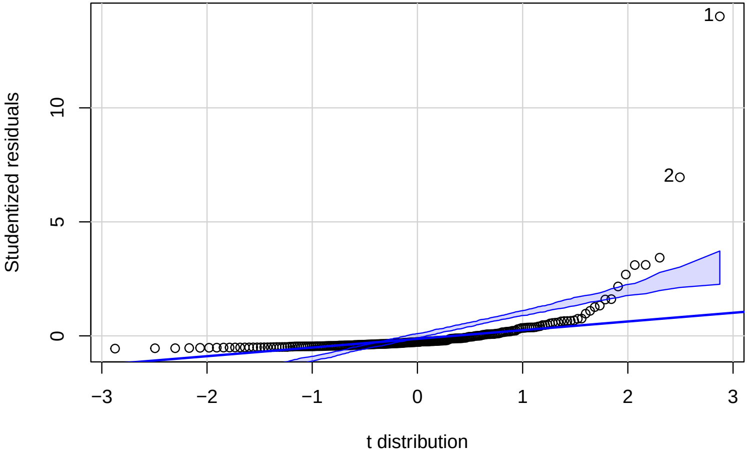
The standard tool is a quantile-comparison plot (Q-Q plot).

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.

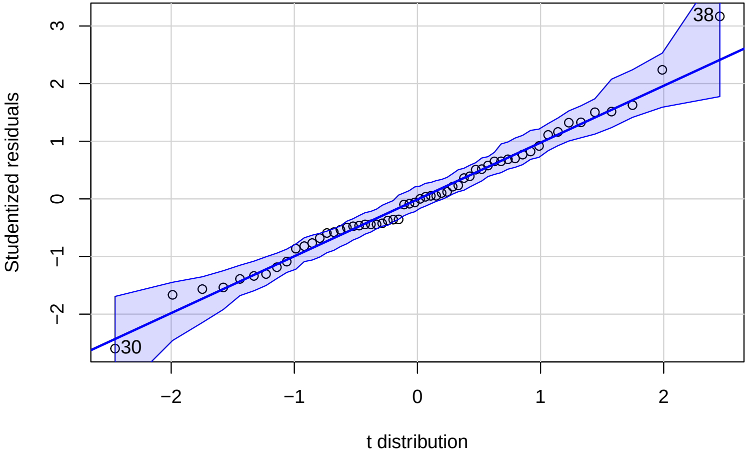
If our residuals are normally distributed, then the points ought to line up on a diagonal line in the graph.

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

Example: *Fortune*'s 1992 billionaires



Example: California counties in 1992



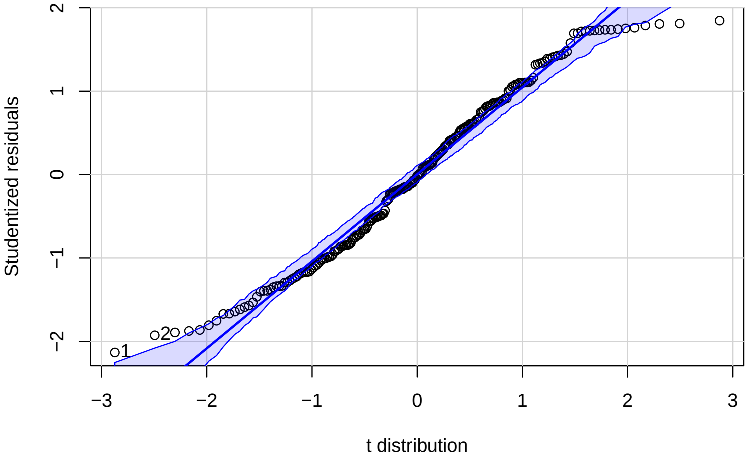
Addressing non-normality

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



Unusual and influential data

Outliers and high leverage cases

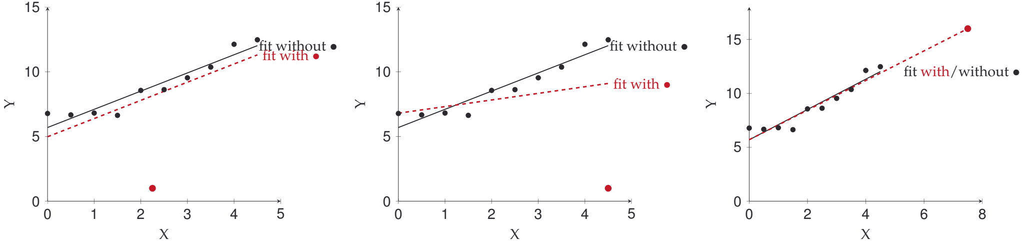
OLS estimates are easily influenced by outliers in the data.

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



Left panel: Outlier, but with low leverage. **Center panel**: Outlier, with high leverage. **Right panel**: High leverage, but not an outlier.

Influence on coefficients

$$\mathit{Influence} = \mathit{Leverage} \times \mathit{Discrepancy} \tag{7}$$

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

The case in the third panel is nevertheless problematic.

$$V(b) = \frac{\sigma_{\epsilon}^{2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{8}$$

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

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 variables that ought to be there.

Theory provides a list of variables.

Take an example. Although the full model is:

$$Y = a + b_{11}X_1 + b_{12}X_2 + \epsilon_1 \tag{9}$$

we can only test:

$$Y = a + b_{21}X_1 + \epsilon_2 \tag{10}$$

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 bs 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 ⇒ a and bs are unbiased, but SEs are larger, and R^2 is lower (Berry, 1993, p. 51).

Systematic error ⇒ 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.

$$D = \frac{\sum_{t=2}^{n}(\epsilon_t - \epsilon_{t-1})^2}{\sum_{t=1}^{n}\epsilon_t^2} \tag{11}$$

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:

$$VIF = \frac{1}{1 - R_j^2} \tag{12}$$

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!

References I

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

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.