UNIVERSITY OF MANNHEIM
School of Social Sciences

Quantitative Methods in Political Science: Linear Regression Model and Diagnostics

Thomas Gschwend & Oliver Rittmann & Viktoriia Semenova

Week 8 - 27 October 2021

The Course

Roadmap

- Understand and model stochastic processes
- Understand statistical inference
- Implement it mathematically and learn how to estimate it
  - OLS
  - Maximum Likelihood
- Implement it using software
  - R
  - Basic programming skills

Overview: Week 8

The Classical Linear Model Assumptions

Regression Diagnostics

1. Misspecification
   - Monte Carlo Simulation
2. Measurement Error
3. Multicollinearity
4. Heteroskedasticity
5. Influential Observations

Quiz

Suppose you constructed a regression model. A reviewer makes the following suggestions to bolster your argument that you have a well-specified model. Do they make sense?

1. If your model is well specified then 95% of the observations fall within the 95% confidence intervals around the regression line.
2. Plot a scatterplot of the residuals by $\hat{Y}$ and draw lines at $+/-$ 1 SE of the predicted values. Roughly 2 out of 3 observations should fall between the lines.
3. When you also draw lines at $+/-$ 2 SE of the predicted values, you should see about 95% of your observations to fall between the lines.
4. You can also use a scatterplot of first-differences instead of the residuals to check whether your model is well specified.

The Classical Linear Model Assumptions

The classical linear model assumptions

1. Linearity in parameters. Our model can be written as:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

2. The zero conditional mean assumption requires that

$$E(\epsilon | x_1, x_2, ..., x_k) = 0$$

3. Error variance is constant given the data (homoskedasticity assumption), which is

$$Var(\epsilon | x_1, x_2, ..., x_k) = \sigma_\epsilon^2$$

4. Errors for any two observations are uncorrelated given the data (independence assumption). Hence,

$$Cov(\epsilon_i, \epsilon_j | x_i, x_j) = 0, i \neq j$$

5. X is non-stochastic. Observations on our independent variables are fixed in repeated samples. Hence, variability is due to stochastic component (i.e., in ϵ) and not due to measurement error in X.

6. Normality of errors

$$\epsilon_i \sim N(0, \sigma_\epsilon^2)$$

The classical linear model assumptions

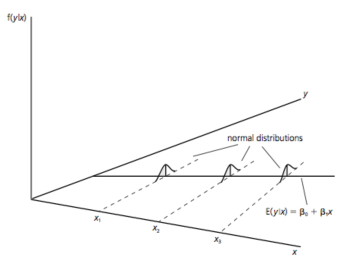
- The OLS estimator is *unbiased*, i.e., $E(\hat{\beta}_j) = \beta_j$, if the errors are independently distributed with zero expectation and constant variance (conditions 2,3, and 4).
- Under the full CLM assumptions (+ normality of errors), the OLS estimator has *minimum variance* among the unbiased estimators, i.e. it is *best linear unbiased estimator* (BLUE). This is the famous Gauss-Markov Theorem.
- A succinct way to summarize the assumptions of the CLM is:

$$Y_i = X_i\beta + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma_\epsilon^2)$$

Or, alternatively:

$$Y_i \sim N(\mu_i, \sigma_\epsilon^2) \quad \text{Stochastic component}$$
$$\mu_i = X_i\beta \quad \text{Systematic component}$$



Regression Diagnostics

1. Misspecification

- We already know what happens when you omit an important variable from the regression
- The other predictors will try to “make up the difference” - and will do so unless they are totally unrelated to the excluded variable
- But: how much are our estimates biased?
- The problem: data analysis does not help us because we don’t know whether we omitted an important variable or whether our model is correctly specified.
- We need to investigate the statistical properties of OLS when we assume *a priori* that the fitted regression model misrepresents the true data generating process.

1. Misspecification

- Suppose we know the true data generating process. In other words, we know the values of our parameters β_0, β_1, etc.
- Let the true data generating process be the following:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$\epsilon \sim N(0, \sigma_\epsilon^2)$$

- This is the standard multivariate linear model with normally distributed disturbances. Suppose we also know that the explanatory variables x_1 and x_2 are related in the following way (i.e., x_1 and x_2 are *correlated*):

$$x_2 = \delta x_1 + \nu$$

$$\nu \sim N(0, \sigma_\nu^2)$$

- Note that, for $\delta \neq 0$, as $\sigma_\nu^2 \to 0$, we'll approach exact linear dependency. As σ_ν^2 increases, the correlation between x_1 and x_2 becomes less severe.

1. Misspecification

- We want to compare the performance of the OLS estimators in two different regression models:

  1. Correctly specified multiple regression (true model)
  2. Misspecified regression model

- Model 1: correctly specified (true model)

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- Model 2: misspecified (x_2 is omitted)

$$Y = \beta_0^* + \beta_1^* x_1 + \epsilon$$

- In particular, we want to compare β_1 and β_1^* to determine the size and direction of the bias!

- It can be shown that:

$$E(\beta_1^*) = \beta_1 + \beta_2\hat{\delta}$$

where $\hat{\delta}$ is the regression coefficient from a bivariate regression of x_2 on x_1.

- The expected omitted variable bias is then:

$$E(\beta_1^*) - \beta_1 = \beta_2\hat{\delta}$$

- Thus, OLS is unbiased if
  1. $\beta_2 = 0$ (i.e. x_2 does not appear in true model), or if
  2. $\hat{\delta} = 0$ (i.e. if x_1 and x_2 are uncorrelated)

Monte Carlo Simulation

Rather than analytical, we can also use a computational approach, relying on *Monte Carlo simulations*, to examine the consequences of violating the classical linear model assumptions. You have done this in the lab already several times.

- A Monte Carlo simulation is a method for estimating the value of an unknown quantity using principles of inferential statistics, i.e. through drawing random samples from a (population) distribution
- The key is: we need to "know" (i.e., assume) the "true" values of the population parameters. We do this by defining the true data generating process (through a regression equation including draws from a normal distribution to represent the error term) and simulating observations for the independent and dependent variables.

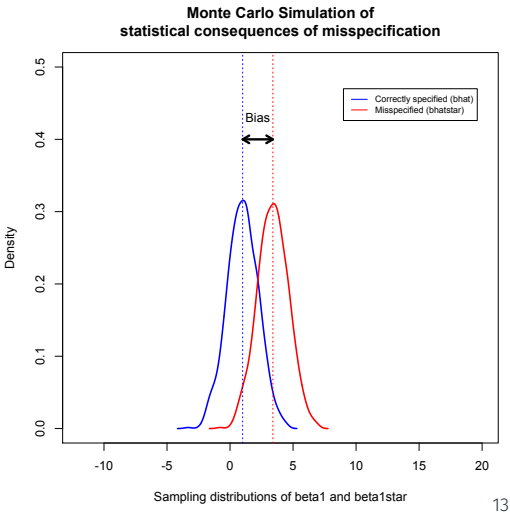
Monte Carlo Simulation

1. We generate “fake” (i.e., simulate) data, X, and assume that we know the true coefficients, $\boldsymbol{\beta}$ and, hence, the true data generating process (including ϵ). Of course, this is never true with real data.
2. Given the data, X, and the true coefficients, $\boldsymbol{\beta}$ and random draws from a normal distribution to represent the error term, we calculate the outcome variable Y from the assumed “true” linear model.
3. Given Y and X, we can run an OLS regression, and extract the estimated coefficients $\hat{\boldsymbol{\beta}}$.
4. If we repeat steps (2) and (3) often enough, e.g., 1,000 times, we generate a empirical sampling distribution consisting of 1,000 different regression coefficients.

- Note that this is very similar to what we have done last week, except that we now simulate based on the assumed “truth”, to see what happens when we violate assumptions.
- When generating QoI last week, we approximated the sampling distribution of the regression coefficients to account for estimation uncertainty and added random draws from the error distribution to account for fundamental uncertainty.

1. Misspecification

1. Set number of observations and “true” parameter values for $\beta_0, \beta_1, \beta_2, \delta$.
2. Set values for x_1 and generate values for x_2 as sum of δx_1 and a random draw from $N(0, \sigma_\nu^2)$.
3. Generate Y using the true data generating process including a random draw from $N(0, \sigma_\epsilon^2)$.
4. Run two regressions (correctly specified and misspecified model) and record $\hat{\beta}_1$ and $\hat{\beta}_1^*$
5. Repeat steps (3) and (4), say, 1,000 times, each time recording the coefficients
6. Plot and compare simulated sampling distributions for β_1 and β_1^*



2. Measurement Error

- Recall one of our classical linear model assumptions:

  X is non-stochastic. This means that the observations on our independent variables are fixed in repeated samples.

- This implies no measurement error in X

- Recall how we set up X in our Monte Carlo study:

  X was generated once and held fixed in each simulation. The only stochastic component was the generation of the error terms (disturbances).

- But: measurement error might exist. Simulate it as robustness test of your findings! E.g., how much measurement error does there have to be before coefficient turns insignificant? And is that realistic?

2. Measurement Error

- Our theories often refer to constructs that are difficult to observe directly
- Yet, we almost always assume a perfect measurement process.
- But most of our data in political science are poorly measured
- Examples
  - Job approval of prime ministers as measured in a poll
  - Degree of democracy (construction of “democracy scales” through an additive index, e.g. POLITY)
  - Ideology of voters (issue questions in surveys)
  - Ideology of political parties (coding of manifestos, or survey of “experts”)
  - Policy area of legislation (coding of legislation into a policy area can lead to misclassification)
  - Reported vs actual income

- Consider the following simple regression model:

$$Y = \beta_0 + \beta_1 X + u$$

- The problem is: we do not observe X, but X^* (measured with error e) instead:

$$X^* = X + e$$

$$e \sim N(0, \sigma_e^2)$$

- Since $X = X^* - e$, we are estimating in the regression:

$$Y = \beta_0 + \beta_1 (X^* - e) + u$$

re-arranging leads to:

$$Y = \beta_0 + \beta_1 X^* + (u - \beta_1 e)$$

2. Measurement Error

- We assumed that the measurement error *e* is uncorrelated with the *unobserved* explanatory variable *X*.

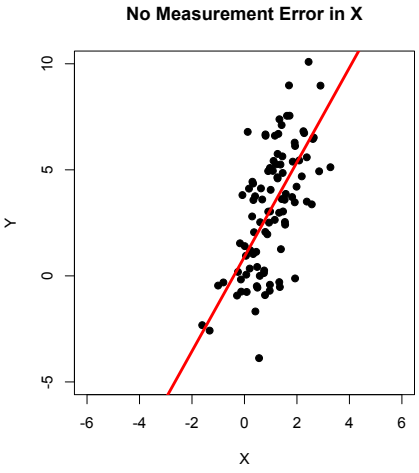
- This means that the measurement error *e* must be correlated with our *observed* explanatory variable *X**.

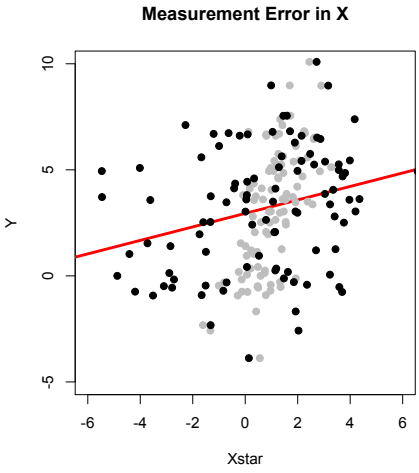
- Recall CLM assumption that X's are fixed and not correlated with the errors. The correlation between the observed explanatory variable and the error will cause biased estimates. When estimating via OLS, we get asymptotically:

$$\hat{\beta}_1 \to \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}\beta_1$$

2. Measurement Error

- This attenuation bias, which is introduced through measurement error, biases our coefficients towards zero.
- Hence, we underestimate the true effect.
- Why is this important?
- It is important because the estimated effect in a simple regression model might turn out to be statistically insignificant although the true effect is larger (and, thus, potentially significant).
- In the multivariate case, this problem gets more severe. The true effect can be under- or overestimated


No Measurement Error in X
Y
X
10
5
0
-5
-6
-4
-2
0
2
4
6

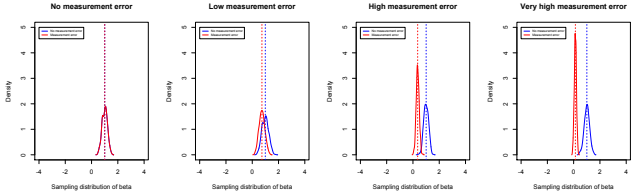

Measurement Error in X
Y
Xstar
10
5
0
-5
-6
-4
-2
0
2
4
6

2. Measurement Error

- We can use our Monte Carlo setup to simulate data and run regressions on X with measurement error added and on X without measurement error.
- With an attenuation bias in the asymptotics, given as

$$\hat{\beta}_1 \rightarrow \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}\beta_1,$$

  the size of the measurement error can be defined by varying σ_e^2.
- Expectation: sampling distribution of β_1 (for X with measurement error) is biased towards zero.
- Robustness test: Find out which level of measurement error does render our coefficient to be still significant.



- So far, measurement error was in the independent variable. But what happens if the error is in the dependent variable?
- Assume that we are modeling Y, but we only observe Y^* (measured with error u):

$$Y^* = Y + u$$

- Hence, the model we estimate is

$$Y + u = \beta_0 + \beta_1 X + e.$$

- Rearranging leads to:

$$Y = \beta_0 + \beta_1 X + (e - u)$$

- Does this cause bias?
  - As long as $E(u \mid X) = 0$, i.e., measurement error is uncorrelated with X, the OLS estimators are unbiased, but the variance is inflated (larger variances, larger standard errors).
  - Thus, measurement error in Y is no big deal!

3. Multicollinearity

- Multicollinearity refers to high correlation between two or more independent variables. It is not limited to pairwise correlation.
- But multicollinearity can lead to instability in the regression estimates - the same consequence that follows from having a small sample
- Recall the sampling variance of the OLS estimator in multiple regression:

$$Var(\hat{\beta}_j) = \frac{\sigma_\epsilon^2}{\sum (x_{ij} - \bar{x}_j)^2 (1 - R_j^2)}$$

σ_ϵ^2 = error variance (noise)
$\sum (x_{ij} - \bar{x}_j)^2$ = sample variation in x_j
R_j^2 = linear relationship among independent variables, i.e. the R^2 from a regression of x_j on all other independent variables.

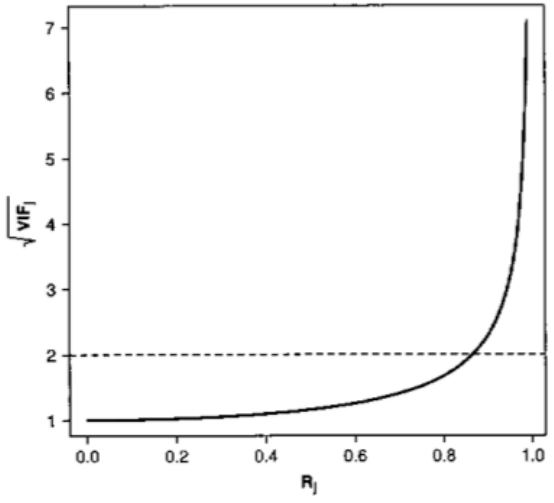
Thus, given $Var(\hat{\beta}_j) = \frac{\sigma_\epsilon^2}{\sum (x_{ij} - \bar{x}_j)^2 (1 - R_j^2)}$, we get:

- As $R_j^2 \to 1$, more sample variation in x_j can be explained by the other independent variables, and $Var(\hat{\beta}_j)$ increases.
- The consequence on the variance of the estimator makes intuitive sense: When the independent variables are strongly correlated, the data contain little information about the impact of, say, x_1 on Y holding $x_2, ..., x_k$ constant, because there is little variation in x_1 when $x_2, ..., x_k$ is fixed.
- Strong, but less than perfect, multicollinearity substantially increases the sampling variances of the OLS coefficients. Consequently, this also increases confidence intervals & standard errors of the coefficients. Thus, it also affects hypothesis testing.
- The good news is: even strong multicollinearity does not bias our coefficients!

- A correlation table gives you a first idea. Pairwise correlations cannot reveal highly collinear linear combinations, though.
- A good strategy to detect it is to calculate the variance-inflation factor for each independent variable, which is defined as

$$VIF_j = \frac{1}{(1 - R_j^2)},$$

  where R_j^2 is again the R^2 measure from a regression of x_j on the remaining explanatory variables.
- A problem with VIF is that we need to determine an arbitrary threshold above which we consider multicollinearity to be high.
- Sometimes, 10 is chosen as such as threshold value, but this does not necessarily imply that the standard error would be too large.
- It is unclear what “too large” means in this context.

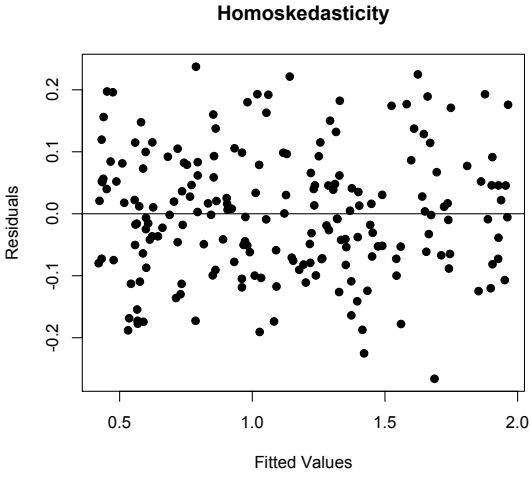

7
6
5
4
3
2
1
$\sqrt{VIF_j}$
0.0
0.2
0.4
0.6
0.8
1.0
R_j

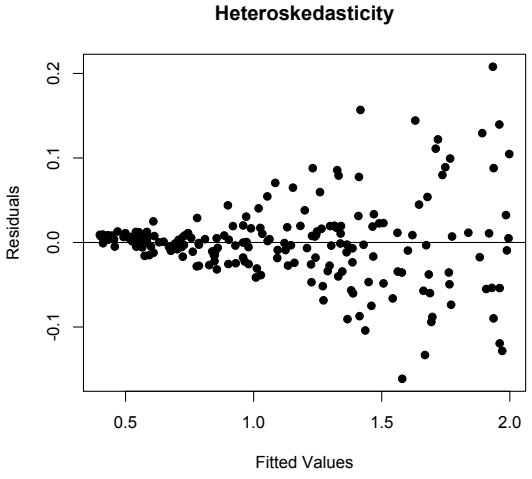
3. Multicollinearity

- There is no good way to reduce variances of unbiased estimators other than to collect more data (increase sample size!).
- Possibly simplify model, i.e. drop variables to reduce collinearity. Unfortunately, dropping relevant independent variables leads to omitted variable bias.
- Possibly combine variables (e.g. “participation index”, “legislator’s expertise”). This way we would no longer be trying to estimate the partial effect of each separate category (e.g. committee assignments, number of cosponsored bills, etc.). May seem reasonable if we are asking questions that are too subtle for the available data to answer with any precision.
- If we believe certain variables belong in the model (e.g., to infer causality), then there is not much we can do about it. Our theories should tell us what to include, but we may be able to use different operationalizations.
- If we expect to see an effect, but confidence intervals appear too wide, checking VIF is a good diagnostic tool.

4. Heteroskedasticity

- Recall the CLM assumption of constant error variance (homoskedasticity). It is necessary to justify the usual *t*-tests, *F*-tests, and confidence intervals for OLS estimation.
- But: It is common for the variance of the errors to increase or decrease with the level of the independent variable.
- Such a pattern is called heteroskedasticity. OLS standard errors are no longer valid for constructing confidence intervals or hypothesis testing.
- Heteroskedasticity can be detected in a plot of residuals against fitted values

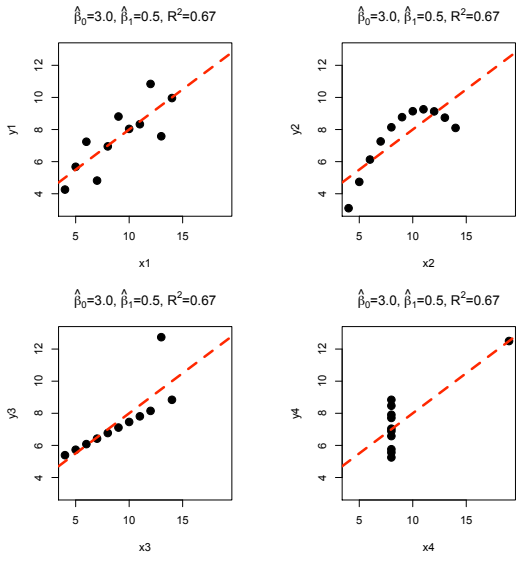

Homoskedasticity
Residuals
0.2
0.1
0.0
-0.1
-0.2
0.5
1.0
1.5
2.0
Fitted Values


Heteroskedasticity
Residuals
0.2
0.1
0.0
-0.1
0.5
1.0
1.5
2.0
Fitted Values

4. Heteroskedasticity

- The error variance might vary because of model *specification errors* (i.e., the model is not correctly specified).
  - Transform your independent variables and potentially the depend variable (using *log*'s, square-roots or square's). Plot residuals against the respective variables to find candidates for this.
  - or include interaction effects.
- Heteroskedasticity can arise as a result of the presence of *outliers* (more later today!)
- Instead of using OLS standard errors, we can use heteroskedasticity-robust (*Huber-White* or *robust*) standard errors for $\hat{\beta}$ as a statistical fix. More about this next semester.
- Another better alternative is to model heteroskedasticity explicitly. If you are interested in how to do this, register with the "Advanced Methods Course" (AQM) next semester.

5. Influential Observations



5. Influential Observations

- The above examples show that OLS estimation can be very sensitive to singletons.
- Therefore, it is important
  - to identify observations that may be dominating the estimates, and
  - to assess how results change when such dominating data points are dropped.
- Some terminology is needed here:
  - Outlier: An observation with a large residual; may indicate a sample peculiarity, a data entry error, or some other problems.
  - Leverage: An observation with an extreme value on an independent variable is said to have high leverage. Leverage is a measure of how far an independent variable deviates from its mean. These leverage points may have an effect on the estimate of regression coefficients.
  - Influence: An observation is said to be influential if removing the observation substantially changes the estimate of coefficients.

5. Influential Observations: Cook's Distance

- The Cook's distance is a measure of influence, which aggregates outlier and leverage properties.
- For each observation i Cook's distance is defined as

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j\setminus i})^2}{(k+1)\hat{\sigma}},$$

  where $\hat{y}_j$ is the predicted value for j with all data, $\hat{y}_{j\setminus i}$ is the predicted value for j when observation i is dropped, k is the number of independent variables, and $\hat{\sigma}$ is the estimated error variance.
- The measure of Cook's distance is based on the difference in predicted values with and without particular observations.

5. Influential Observations: Cook's Distance

- The rule of thumb is to be suspicious of observations with $D_i > 1$.
- Bollen and Jackman (1990) suggest to determine the threshold as $4/n$ where n is the sample size.
- So what to do?
  - Create a dummy in your dataset which is "1" when Cook's distance is too large.
  - Respecify model to account for outliers. Do they have something in common?
  - Potentially drop those observations from analysis (and acknowledge that). Do you still get the same results?