Statistics
Tutorial 06

Philipp Scherer & Jens Wiederspohn

10.06.2020

Disclaimer!

- **The content of the slides partly relies on material of Philipp Prinz, a former Statistics tutor. Like us, he's just a student. Therefore we provide no guarantee for the content of the slides or other data/information of the tutorial.**
- **Please note that the slides will not cover the entire lecture content. To pass the exam, it is still absolutely necessary to deal with the Wooldridge in detail!**

Effects of data scaling on OLS statistics

1. If dependent variable is changed:
   - Coefficients change $\rightarrow \beta^* = \frac{Cov(x,a*y)}{Var(x)} = a * \frac{Cov(x,y)}{Var(x)} = a * \beta$
     - One unit increase in x has another effect when y is changed
   - Standard errors of the coefficients change
     - When coefficients change, standard errors also need to change

     $\rightarrow se(\beta^*) = se\left(\frac{Cov(x,a*y)}{Var(x)}\right) = se\left(\frac{a*Cov(x,y)}{Var(x)}\right) = |a| * sd(\beta)$
   - R^2 remains unchanged
     - Scaling of data does not affect explanatory power of a variable
   - Residuals change, since they depend on y $(y_i - \hat{y}_i)$
   - t- and F-statistic are unchanged

     $\rightarrow t^* = \frac{\hat{\beta}^*}{se(\hat{\beta}^*)} = \frac{a*\hat{\beta}}{|a|*se(\hat{\beta})} = t$

     $\rightarrow F^* = \frac{(R^{2*}_{ur} - R^{2*}_{r})/q}{(1-(R^{2*}_{ur})/df_{ur}} = \frac{(R^{2}_{ur} - R^{2}_{r})/q}{(1-(R^{2}_{ur})/df_{ur}} = F$

Effects of data scaling on OLS statistics

2. If independent variable is changed:
   - Coefficients change $\rightarrow \beta^* = \frac{Cov(a*x,y)}{Var(a*x)} = \frac{a*Cov(x,y)}{a^2*Var(x)} = \frac{1}{a}\beta$
     - One unit increase in x has another implication for y
   - Standard errors of the coefficients also change
     $\rightarrow se(\beta^*) = se\left(\frac{Cov(a*x,y)}{Var(a*x)}\right) = se\left(\frac{a*Cov(x,y)}{a^2*Var(x)}\right) = \frac{1}{|a|}sd(\beta)$
   - R^2 remains unchanged
     - Scaling of data does not affect explanatory power of a variable
   - Standard errors of residuals do not change
     - Horizontal deviation is independent of changes on the x axis
   - t- and F-statistic are unchanged
     $\rightarrow t^* = \frac{\hat{\beta}^*}{se(\hat{\beta}^*)} = \frac{\hat{\beta}}{se(\hat{\beta})} = t$
     $\rightarrow F^* = \frac{(R^{2*}_{ur} - R^{2*}_{r})/q}{(1-(R^{2*}_{ur})/df_{ur}} = \frac{(R^{2}_{ur} - R^{2}_{r})/q}{(1-(R^{2}_{ur})/df_{ur}} = F$

Dummies

- Dummy = binary variable with values 0 and 1
  - East/west, unmarried/married, loss/victory
- Effect is difference (e.g. east and west) if all else is equal
  - Base group vs. reference group
    - Never use both groups as a dummy → dummy variable trap
    - Multicollinearity if all categories are included
  - Constant difference → intercept shift
  - Shifts entire regression line up/down (equal shift for each x)

Quadratic forms

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

- Simple way to capture diminishing marginal effects (if $\beta_2 < 0$)
  - Slope decreases as x increases (after maximum $x^* = \frac{\beta_1}{-2\beta_2}$)
  - Increasing/decreasing marginal effects depend on the βs
- How quickly is the slope changing? $\rightarrow$ derivative

Interaction effects

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u \\
&= \beta_0 + (\beta_1 + \beta_3 x_2) x_1 + \beta_2 x_2 + u
\end{aligned}
$$

- Explanatory variable depends on magnitude of another x
- Reparameterize model to get overall effect of x
  - Easiest to see if we use interaction with a dummy
    $\rightarrow$ Allows different slopes (effects) for both groups (see graph)
    $\rightarrow$ Since $x_2 = 1$ effect of $x_1 = \beta_1 + \beta_3$ (as opposed to β_1)

R^2

- $R^2 = 1 - \frac{SSR/n}{SST/n}$ $\rightarrow$ recall: $\frac{SST_y}{n} = Var(y)$
- R^2 = estimate of how much variation in y is explained by x_k
  - Small $R^2 \rightarrow$ error variance is large relative to variance of y
    - Much variation comes from unexplained factors
      $\rightarrow$ Complicates estimation of $\hat{\beta}$ and prediction
  - But: large error variance can be offset by a large sample size
    $\rightarrow$ Poor expl. power has nothing to do with unbiased estimation
- R^2 can be used to select variables for a model:
  - F statistic measures relative change in R^2 when x are added
  - Use multiple (not adjusted) R^2 for F statistic

Adjusted R^2

- $R^2_{adj} = 1 - \frac{SSR}{SST} \cdot \frac{n-1}{n-k-1} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$
- Adjusted R^2 imposes penalty for additional independent variables
  - Account for fact that SSR of a model with more explanatory variables never increases (maximum is no change)
  - Dependency on number of x (i.e. k) decreases R^2_{adj} if additional variables have no explanatory power
- R^2_{adj} uses unbiased estimator for population variances
  - Replace biased estimators for population variances:
    - $\rightarrow$ Use $\frac{SSR}{n-k-1}$ rather than $\frac{SSR}{n}$ to estimate $\hat{\sigma}^2$
    - $\rightarrow$ Replace $\frac{SST}{n}$ with $\frac{SST}{n-1}$ to estimate $\hat{\sigma}^2_y$
- Unfortunately, R^2_{adj} is not an unbiased estimator of ρ^2

Exercise 1

You regress the vote share of AfD (in %) in districts at the German federal election 2017 on GDP per capita (in thousand Euro). The unit of analysis is electoral districts ($n = 299$). We call this model Model 1 and its result is as follows:

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.065  -3.537  -1.050   2.247  20.542

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.90134    0.85240  21.001  < 2e-16 ***
gdp.tsd     -0.14599    0.02269  -6.434 4.96e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 5.126 on ( A ) degrees of freedom
Multiple R-squared:  ( B ),Adjusted R-squared:  ( C )
F-statistic:  41.4 on 1 and ( A ) DF,  p-value: 4.956e-10
```

Exercise 1a and b

Which value should be in (A) - (C) in the above output? As additional information, you can rely on the following information: SST: 8891.792 , SSE: 1087.865

A $df = n - k - 1 = 299 - 1 - 1 = 297$

B $R^2 = 1 - \frac{SSR/n}{SST/n} = 1 - \frac{7803.927/299}{8891.792/299} = 0.1223448$

C $R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k-1} = 1 - (1 - 0.1223448) * \frac{298}{297} = 0.1193898$

Suppose that you regress the same dependent variable on GDP per capita in Euro. Estimate the intercept and slope coefficient of the regression line. Do not round your result.

- Intercept remains the same $\rightarrow$ 17.90134
- $\beta_1^* = \frac{1}{a}\beta = \frac{1}{1000} * -0.14599 = -.00014599$

Exercise 1c and d

Suppose that you regress the vote share of AfD (not in %) on GDP per capita in thousand Euro. Estimate the intercept and slope coefficient of the regression line

- $\beta_0^* = a * \beta = 0.01 * 17.90134 = 0.1790134$
- $\beta_1^* = a * \beta = 0.01 * -0.14599 = -.0014599$

Suppose that you regress the vote share of AfD (not in %) on GDP per capita in Euro. Estimate the intercept and slope coefficient of the regression line

- $\beta_0^* = a * \beta = 0.01 * 17.90134 = 0.1790134$
- $\beta_1^* = a * \beta = 0.01 * \frac{1}{1000} * -0.14599 = -.0000014599$

Excercise 1e

Coming back to Model 1 (the vote share of AfD in % on GDP per capita in thousand Euro), we denote the dependent and independent variable by y and x, respectively. We modify the model by using the natural logarithm function and obtained the results as follows:

Model	dep.var	ind.var	$\hat{\beta}_0$	$\hat{\beta}_1$
2	y	$\ln(x)$	37.9898	-7.1901
3	$\ln(y)$	x	2.832911	-0.010269
4	$\ln(y)$	$\ln(x)$	4.18455	-0.48826

Exercise 1e and f

If GDP per capita increases by 1000 Euro, how much growth in % can you predict for the vote share of AfD?

- $log(y)$ and x: one unit increase in X increases Y by $\beta \cdot 100\%$
- $\hat{\beta}_1 = -0.010269 * 100\% = -1.0269\%$

If GDP per capita increases by 1%, how much growth in % can you predict for the vote share of AfD?

- $log(y)$ and $log(x)$: one % increase in X increases Y by $\beta\%$
  $\rightarrow \beta$ measures 'elasticity'
- -0.48826%

Exercise 1g

Calculate elasticity of the vote share of AfD in respect to GDP (in TSD)?

- $log(y)$ and $log(x)$: one % increase in X increases Y by β%
  $\rightarrow \beta$ measures 'elasticity'
- -0.48826%
- A one % increase stays a one % increase, independent of the unit of measurement of the independent variable.
- Changing the unit of measurement of the independent variable, when it appears in logarithmic form, does not affect the slope estimates.

Exercise 1h

Suppose that you add another variable to Model 1, whereby the new variable has zero covariance with GDP per capita in Euro. We call this model Model 5. The new variable has nonzero effect on the vote share of AfD, however its effect was not significant at 5%-level. Which following statement(s) is(are) true for Model 5?

B The standard error of GDP per capita in thousand Euro is smaller than that in Model 1. $\rightarrow$ *TRUE*

- Because we added one more explanatory variables to the equation, the error variance gets reduced. At the same time R_j^2 stays the same, because the new variable is uncorrelated with GDP per capita in Euro. This leads to an overall lower standard error.
- $se(\hat{\beta}_k) = \frac{\hat{\sigma}}{\sqrt{SST_k(1-R_k^2)}}$

Exercise 1i

You are now predicting the vote share of AfD (in %) in a district where GDP per capita is one thousand Euro. Calculate the point estimate of the prediction using Model 2

- $\hat{voteshare} = \hat{\beta}_0 + \hat{\beta}_1 \ln(gdp)$
- $\hat{voteshare} = 37.9898 - 7.1901 * \ln(gdp)$
- $\hat{voteshare} = 37.9898 - 7.1901 * \ln(1)$
- $\hat{voteshare} = 37.9898 - 7.1901 * 0$
- $\hat{voteshare} = 37.9898$

Exercise 1j

To obtain the uncertainty of the predicted value above as average value of AfD vote share for the corresponding district, which procedure do you need to conduct?

C You construct a new variable z by subtracting $\ln(1)$ from $\ln(x)$ and estimate a regression model $y = \beta_0 + \beta_1 z + u$. You use the standard error of β_0 as uncertainty measure.

Exercise 1k

You are now predicting the vote share of AfD (in %) in a district where GDP per capita is one thousand Euro. Calculate the point estimate of the prediction using Model 2

A larger.

- The standard error for the average value in the subpopulation is not the same as a standard error for a particular unit from the population. When calculating a standard error for an unknown outcome on y, we must account for another very important source of variation: the variance in the unobserved error, which measures our ignorance of the unobserved factors that affect y. This leads to an overall larger uncertainty.

Exercise 1l

Assume that Model 3 satisfies the CLM assumptions. If you rescale your dependent variable from $\ln(y)$ to y, what distribution would the residuals have?

C A skewed distribution.

- Strictly positive variables often have conditional distributions that are heteroskedastic or skewed; taking the log can mitigate, if not eliminate, both problems.
- But this also works the other way around. If we rescale our dependent variable, from a model that satisfies the CLM assumptions, from $\ln(y)$ to y the distribution of the residuals would be skewed.

Exercise 1m

You are now predicting the vote share of AfD (in %) in a district where GDP per capita is one thousand Euro. The point estimate of the prediction based on Model 3 is:

D None of A - C

- $\hat{y}$ has a larger value than 16.82123 since $\exp(\hat{log}(y)) = 16.821234$ and $\hat{y} = \hat{\alpha_0} \exp(\hat{log}(y))$, with $\hat{\alpha_0}$ being necessarily greater than one.
- $\hat{log}(y) = \hat{\beta}_0 + \hat{\beta}_1 x$
- $\hat{log}(y) = 2.832911 - 0.010269 * x$
- $\exp(\hat{log}(y)) = \exp(2.832911 - 0.010269 * x)$
- $\exp(\hat{log}(y)) = \exp(2.832911 - 0.010269 * 1)$
- $\exp(\hat{log}(y)) = 16.821234$