Statistics
Recap Questions

Philipp Scherer & Jens Wiederspohn

25.06.2020

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!**

Sorry guys :-/

I will not give any written solutions to the questions below as I believe that there is a difference between learning and memorizing. Learning is psychologically defined as *"any relatively permanent change in behavior brought by* **experience or practice**" and I do share that view.

In order for you to succeed in the exam, it is of importance that you at least thought about some of the problems in order to understand the concepts of regression analysis. Nonetheless, you will find all answers to the problems in the Wooldrige and the lecture scripts. Do not shy away from the R markdown (html for starters) either!

Best of luck!
Philipp

Appendix A

- Suppose that the return from a firm's stock goes from 15% to 18%. The majority shareholder claims that "*the stock return only increased by 3%*", while the CEO claims that "*the return on the firm's stock increased by 20%*". Reconcile their disagreement.
- Suppose the success of politicians is related to campaign funding as $success = 17 + .09\sqrt{funds}$. How does this relationship compare to a linear one?
- It is proposed that group size and presentation grades are related as follows: $grade = 2.8 - 0.6size + 0.1size^2$. What is the ideal group size? Interpret the intercept. Any issues? Is this relationship deterministic?

Appendix B

- Why are results of an IQ test considered a random variable?
- Assume that a newspaper correctly predicts elections with $p = 0.7$. What is the probability that its edition on 26 May 2019 correctly predicts all elections the upcoming day? What is the probability that at least 2 predictions are correct?
- X is the proportion of unemployed youth in a European country. It is described by $F(x) = 3x^2 - 2x^3$ for $0 \leq x \leq 1$. What is the probability that youth unemployment is larger than 0.4?
- Suppose that the average PolVer student starts his/her professional career with 35,000 euros. The standard deviation is 8.5. What would be the mean and variance in dollars if 1 Dollar = 0.9 Euro?

Chapter 2^+

- Assume we regress income on education. What factors are likely contained in the error term? Does $Cov(u, x) = 0$ hold?
- Do you think the above regression yields causal estimates?
- You suppose GDP has an effect on the quality of national football teams. Why would reg. estimates be biased? How?
- Under what conditions would an estimate be unbiased?
- How would you model the relationship between housing prices and proximity to highway systems?
- Suppose u in the above regression is $u = \sqrt{distance} * e$ with $Var(e) = \sigma_e^2$. Does homoskedasticity assumption still hold?
- Suppose you estimate a model ($k = 1$) where the intercept is 0. Is $\tilde{\beta}_1$ still an unbiased estimator without an intercept?
- How is $\hat{y}$ predicted in models without any x_k?

Chapter 2^+

- How would you identify the causal effect of tutorials on final grades in an ideal world?
- Why can regression models help us in our imperfect world?
- How would β_1 change if we formulated x as percentage (i.e. $\frac{attended}{total}$)?
- What would happen to β_0 if we changed y from points to grade?

$\rightarrow$ For log transformations, see Table 2.3 (p. 44) in Wooldrige

Gauss-Markov assumptions

Recall the Gauss-Markov assumptions

→ What do they **imply**?

→ Why do we **need** each? What if they do not hold?

1. Model is linear in its parameters β_k
2. We have a random sample of n observations
3. No perfect collinearity
4. Zero conditional mean $\rightarrow E(u|x_1, x_2, ..., x_k) = 0$
5. Homoskedasticity $\rightarrow Var(u|x_1, x2, ..., x_k) = \sigma^2$

Chapter 3

- We assess the impact of hours of sleep on health. Any problems with $health = \beta_0 + \beta_1 \cdot sleep + \beta_2 \cdot awake + u$?
- Under what conditions is an OLS estimator biased?
  1) Heteroskedasticity
  2) $corr(x_1, x_2) = 0.97$
  3) We omit x_a with $corr(x_a, y) = 0.93$ and $corr(x_a, x_b) = 0$
  3) We omit x_b with $corr(x_b, y) = 0.48$ and $corr(x_a, x_b) = 0.26$
- If *catholic* is omitted from a regression of $vote_{CDU}$ on *married*, what is the direction of the bias?
- Which factors determine the standard error of a coefficient?
- Compare a simple regression model ($y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$) and a multiple regression model ($y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$).
  - If $corr(x_1, x_2) = 0.86$, is $se(\hat{\beta}_1)$ larger or smaller than $se(\tilde{\beta}_1)$?
  - If x_1 and x_2 are vote shares of the democratic and republican parties in the USA, why could $\tilde{\beta}_1$ be a better estimator?

Chapter 4

- You want to test the effect of education on income. State the null hypothesis in terms of the effect size.
- If β_1 is 0.1, how does $log(income)$ approximately change if we increase *schooling* by one year?
- If we test whether conservatives are more fond of security measures than liberals, what regression would you run and how would you interpret the R-output?
- Suppose the CI of β_{beer} is $CI_{0.95} = [-2.03, 0.08]$. What do you conclude about the effect of beer on points in the exam?
- Assume you test $H_0 : \beta_1 - 2\beta_2 = 5$. State the formula for the t-statistic. What is the easiest way to compute it?
- Suppose you want to test the overall significance of a model where u is strictly increasing in x. What do you do?

  $\rightarrow$ Check 4.1 and MLR.6

Inference and tests

- How is the 95% confidence interval interpreted?
- State which z-values you would use if you wanted to estimate a CI with a level of confidence of 90%.
- Why do we perform t-tests? How do you interpret the t-statistic?
- How are t-statistic and p-value related?
- What is tested with the F-test in the standard R output? Why can we not always use this value if we are interested in the F-statistic?
- How are the degrees of freedom in the t- and F-test determined, respectively?

Chapter 5

- What is the difference between consistency and unbiasedness?
- Under what conditions do consistency and unbiasedness hold, respectively?
- What is the effect of increasing the sample size on an inconsistent OLS estimator?
- Under what conditions is the estimator $\hat{\beta}_2$ not inconsistent, given that we have $Corr(x_1, u) \neq 0$?
- What are the classical linear model (CLM) assumptions?
- Is it possible to perform t-tests if the normality assumption is violated? State why (not).
- Let $z_i = g(x_i), \forall i$ (= for all i). Under what conditions (name two) is an estimator $\tilde{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})u_i}{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})x_i}$ consistent?

Chapter 6

- How do we interpret β in models where the dependent variable is $log(y)$?
- Assume a regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2$ yields decreasing marginal effects. For what value x_1 do you get the maximum value of y?
- When would you want to use interaction effects?
- Why is over controlling problematic for our inferences?
- Which regressors should we include in our model? Why?
- State why the prediction interval is larger if we predict individuals values rather than group averages.

Chapter 7

- How is non-metric (below interval scale) data ideally accounted for in regression models? Why?
- How exactly is ordinal data incorporated into regression models?
- State why we want to avoid the dummy variable trap and what the consequences on the intercept are if we do.
- How do we interpret interactions with a dummy, conceptually and graphically?
- What is tested with the Chow test?
- When do we use a linear probability model? What are potential downsides of the LPM? Name alternative models that solve the issue(s).
- How would you test whether there is discrimination against women when it comes to promotions in an office?

Chapter 8

- What is the effect of heteroskedasticity on OLS estimators?
- Name two common tests for heteroskedasticity in our data.
- How can heteroskedasticity be tackled in regression analysis?
- What information needs to be known to compute weighted least squares with GLS? Which estimator would you use if that information is unknown (and therefore estimated)?
- Speaking about heteroskedasticity, why does the LPM yield heteroskedastic errors by construction? When can we (not) use WLS to get asymptotically efficient estimators?
- Can you imagine why the estimator is not BLUE, even though the regression model accounts for heteroskedasticity?