Statistics
Tutorial 02

Philipp Scherer & Jens Wiederspohn

06.05.2020

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!**

Normal distribution (Still Appendix B)

- Probability density function with mean μ and variance σ^2
  - **Standard** normal distribution has $\mu = 0$ and $\sigma^2 = 1$
- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Bell shaped, symmetric curve that is centered around μ
  - Many "normal" and fewer extreme observations
  - E.g. height, test results, errors in measurements
- Great importance because of **Central Limit Theorem**:
  - As n gets larger, the distribution of sample means approximates a normal distribution (certain assumptions needed)
    $\rightarrow$ irrespective of the population distribution shape!
- Normal random variables:
  - $X \sim \mathcal{N}(\mu, \sigma^2)$
  - X is distributed normally with expected value μ and variance σ^2

Standard normal distribution

- Recall: $\mu = 0$ and $\sigma^2 = 1 \rightarrow \sigma = 1$
- We can use standard normal distribution to compute the probability of any event that involves a standard normal random variable
  - Some quantile values of the standard normal distribution:
    1. 68% of data fall within ± 1 standard deviation of the mean
    2. 95% of data fall within ± 2 standard deviations of the mean
    3. 99.7% of data fall within ± 3 standard deviations of the mean
  - Standard normal cdf Φ gives $Pr(Z \leq z) \rightarrow$ integral up to z
- Use z-transformation (= standardization) to transform any normal distribution into a standard normal distribution
  - $Z = \frac{X - \mu}{\sigma}$

Recap: Standardization

- Suppose you want to compare the intelligence of two people, but they took different IQ tests
  - Amy, test A, 43 points (μ=27; SE = 7)
  - Bert, test B, 90 points (μ=53; SE = 16.8)
- With standardization, we can compare the results since both test results are tranformed into the same statistic
  $\rightarrow$ **Distance from the mean in standard deviations**
  - $z_A = \frac{43-27}{7} = 2.29$
    $z_B = \frac{90-53}{16.8} = 2.20$
- Amy and Bert did very well, but Amys result is still better
  - Given df=29, we find that Amy's result is in the top 1.5% of the results while Bert's is in "only" in the top 1.8%
    - $Pr(Z \leq 2.29) = 0.985$
    - $Pr(Z \leq 2.20) = 0.982$

Recap: Standardization

- Standardization expresses distance in standard deviations
  $\rightarrow$ **How many standard deviations is Z from the mean?**
- Can be used to calculate probabilities
  - Probability that Z is smaller than z? $\rightarrow \Phi(z)$
  - Probability that Z is larger than 1.96? $\rightarrow 1 - \Phi(1.96) = 0.025$

Examples normal distribution and probability

- Suppose $X \sim \mathcal{N}(5, 4)$ and calculate $P(X \leq 6)$
  - Standardize X to use values of standard normal distribution
  - $Z = \frac{X-\mu}{\sigma} \rightarrow Z = \frac{X-5}{2} \rightarrow Pr(Z \leq z)$
- $P(X \leq 6) = P\left(\frac{X-5}{\sqrt{4}} \leq \frac{6-5}{\sqrt{4}}\right) = P\left(\frac{X-5}{2} \leq \frac{1}{2}\right)$
  $\rightarrow P(Z \leq 0.5) = \Phi(0.5) = 0.6915$
- Suppose $X \sim \mathcal{N}(3, 4)$ and calculate $P(2 \leq X \leq 4)$
- $P(2 \leq X \leq 4) = P\left(\frac{2-3}{2} \leq Z \leq \frac{4-3}{2}\right) = P\left(-0.5 \leq Z \leq 0.5\right)$
  $= \Phi(0.5) - \Phi(-0.5) = 0.6915 - 0.3085 = 0.383$

Definition of Simple Regression Model

- Simple linear regression model is defined by:
  $y = \beta_0 + \beta_1 x + u$
  - "linear" = "linear in parameters" $\rightarrow x_1^2$ is okay, β_1^2 is not
  - Y = Dependent, explained variable
  - X = Independent, explanatory variable
  - β_0 = Constant = value of Y if all x_i are 0
  - β_1 = Slope parameter, effect
  - u = Error term = all factors besides X that affect Y
    $\rightarrow$ Captures all unobserved variables
- Independent variable X explains dependent variable Y
  $\rightarrow$ How does Y change **on average** if we change X?
- We can determine the linear effect of X on Y if $\Delta u = 0$
  - $\Delta y = \beta_1 \Delta x$ if $\Delta u = 0$
  - $\Delta u = 0$ = if x is increased by 1, u does not change $\rightarrow$ c.p.!
  - $\frac{\delta y}{\delta x} = \beta_1 \rightarrow$ if we increase x by 1, y changes by β_1
    - β_0 also drops out, derivative of a constant is 0

Unobserved factors

- Ceteris paribus assumption is often quite unrealistic
  - If we do not observe u, how can we say that it is fixed?
    $\rightarrow$ Look into section on causality from methods
- Restrictions on relationship of x and u needed to estimate β
  1. If we have an intercept: $E(u) = 0$
  2. $E(u|x) = E(u) = 0 \rightarrow$ average u does not depend on x
     - E.g. average ability is same for all education levels $\rightarrow$ sensible?
  3. If $E(u|x) = 0$, then $E(y|x) = \beta_0 + \beta_1 x$
     - One-unit increase in x changes the expected value of y by β_1
     - For given value x, distribution of y is centered about $E(y|x)$

Systematic and stochastic parts of regression analysis

- Full regression equation can be broken into two parts:
  1. Systematic part $E(y|x) = \hat{y} = \beta_0 + \beta_1 x$
     - Boils down to conditional expectation of y = regression line
     - Part of y that is explained by x via regression model
  2. Stochastic part u
     - Unsystematic part of y
     - Cannot be explained by x

Regression for samples of the population

- $y = \beta_0 + \beta_1 x + u \rightarrow y_i = \beta_0 + \beta_1 x_i + u_i$
  - Distinct values for y, x and u for each unit i in the sample
  - Graphically: scatterplot where x and y are plotted for each unit
- Given the data, we estimate parameters that best fit them

ID	y_i	x_i	u_i
1	5	1	na
2	1	4	na
3	2	3	na
4	4	2	na
5	3	3	na

  - Explain relationship between x and y based on data
  - Fit regression line to data points in scatterplot
  - On average, regressed values should be close to actual data
    $\rightarrow$ Minimize distance between predicted and empirical values!

How to get the estimates?

- First, we assume:
  1. $E(u) = E(y - \beta_0 - \beta_1 x) = 0$
     - On average, all unobserved factors are 0 $\rightarrow$ no impact on E(Y)
  2. $Cov(x, u) = E(xu) - 0 = E(xu) = E\left(x(y - \beta_0 - \beta_1 x)\right) = 0$
     - In the population, u is uncorrelated with x
- If $E(u) = 0$, then $E(y) = \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$
  - We do not know β and estimate $\hat{\beta}$ with sample data instead!
  - $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ $\rightarrow$ we get $\bar{x}$ and $\bar{y}$ from the data, but what is $\hat{\beta}_1$?
- $Cov(x, u) = \frac{1}{n} \sum_{i=1}^{n} \left(x_i(y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 \bar{x}) \right)$
  - Note: we replaced $\hat{\beta}_0$ with $\bar{y} - \hat{\beta}_1 \bar{x}$
  - Use $\sum_{i=1}^{n} x_i(y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 \bar{x}) = 0$ and rearrange it to get:

    $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)}$

Ordinary least squares (OLS)

- Alternative way to get the same parameter estimates
- "Least squares" $\rightarrow$ **minimize sum of squared residuals**
  - Residual = empirical values (y) - predicted values ($\hat{y}$)
  - Minimize residuals to get best regression line through the data
  - Squared sum since sum of $y_i - \bar{y}$ is 0 by construction: $\sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n}(y_i) - n\bar{y} = \sum_{i=1}^{n}(y_i) - \sum_{i=1}^{n}(y_i) = 0$
- Sum of squared residuals $= \sum_{i=1}^{n}(\hat{u}_i)^2 = \sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- Find minimal sum = optimization problem $\rightarrow$ derivative
  - With $f'(x) = 0$ and $f''(x) > 0$ we can identify smallest value

Derivation of OLS estimates (optional)

- $SSR = \sum_{i=1}^{n}(\tilde{u}_i)^2 = \sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2$
  - If we are very formal - which we are often not - we need to use $\tilde{\beta}_k$ until we have derived the OLS estimator $\hat{\beta}_k$.
- First order condition for $\hat{\beta}_0$ (= first derivative, chain rule)
  $\frac{\delta \sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2}{\delta \tilde{\beta}_0} = 2\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) * (-1) = 0$
  $\rightarrow \sum(y_i) - n * \hat{\beta}_0 - \hat{\beta}_1 \sum(x_i) = 0$
  $\rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- First order condition for $\hat{\beta}_1$
  $\frac{\delta \sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2}{\delta \tilde{\beta}_1} = 2\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) * x_i = 0$
  $\rightarrow \sum(y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i) * x_i = 0$
  $\rightarrow \sum(y_i - \bar{y}) * x_i = \hat{\beta}_1 \sum(x_i - \bar{x}) * x_i$
  $\rightarrow \hat{\beta}_1 = \frac{\sum(y_i - \bar{y})}{\sum(x_i - \bar{x})} = \frac{\sum(x_i - \bar{x})\sum(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)}$

Intermezzo

- Do not confuse error term and residuals!
  1. Error term
     - $u_i = y_i - \beta_0 + \beta_1 x_i$
     - Unobserved variables other than x that affect y
     - Usually unknown, since they are unobserved (wow)
  2. Residual
     - $\hat{u}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$
     - Difference between our model predictions and our data
     - Can be easily computed as difference between y and $\hat{y}$
- Wooldrige uses u for both, but "ϵ" is sometimes used for the error term to distinguish it from the residual u

Properties of OLS estimator

1. $\sum_{i=1}^{n} \hat{u}_i = 0$
   - Sum of all residuals is 0
   - On average, positive & negative residuals cancel each other out
2. $\sum_{i=1}^{n} x_i \hat{u}_i = 0$
   - Sample covariance between regressors x_k and residuals is 0
   - Residuals do not get larger/smaller if x becomes larger/smaller
3. $(\bar{x}, \bar{y})$ is always on the regression line
   - Plug $\bar{x}$ into $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ to get $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

From OLS, it follows:

- $y_i = \hat{y}_i + \hat{u}_i$
- $\bar{y} = \bar{\hat{y}} + \bar{\hat{u}} = \bar{\hat{y}} + \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i = \bar{\hat{y}} + 0 = \bar{\hat{y}}$
  - Mean of y = mean of our regression model

Them sums

- $y_i = \hat{y}_i + \hat{u}_i$
- $Var(y_i) = Var(\hat{y}_i + \hat{u}_i) = Var(\hat{y}_i) + Var(\hat{u}_i) + \underbrace{2Cov(\hat{y}_i, \hat{u}_i)}_{2Cov(\hat{\beta}_0 + \hat{\beta}_1 x, \hat{u}_i) = 0}$

  $\rightarrow Var(y_i) = Var(\hat{y}_i) + Var(\hat{u}_i)$

$\rightarrow SST + SSE + SSR$

- SST = **T**otal sum of squares $\rightarrow$ total sample variation in y_i
  - $SST = \sum(y_i - \bar{y})^2$
- SSE = **E**xplained sum of squares $\rightarrow$ sample variation in $\hat{y}_i$
  - Use $\bar{\hat{y}}_i = \bar{y}_i$ to get $SSE = \sum(\hat{y}_i - \bar{y})^2$
- SSR = **R**esidual sum of squares
  - $SSR = \sum(\hat{u}_i)^2$

$\rightarrow$ total variation = explained variation + unexplained variation

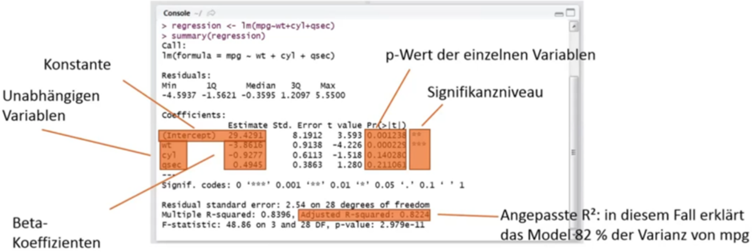
Gauss-Markov assumptions

1. The model is linear in its parameters β_k
2. We have a random sample
   - Possible violations: times series data, unrepresentative samples
3. There is variation in X
   - No perfect collinearity: x_i are not all the same value
   - What would happen to $\hat{\beta}_1$ if they were?
4. Zero conditional mean
   - $E(u_i) = 0$ for all values of X $\rightarrow$ implies $Cov(X, u) = 0$
   - Is it violated for $wage = \beta_0 + \beta_1 education + u$?
5. Model uncertainty equal for any $x_i \rightarrow Var(u_i) = \sigma_u^2$
   - "Homoskedasticity", equal (=constant) error variance

Recap: How to read R Outputs

- `lm(mpg ~ wt + cyl + qsec)` = linear regression model
  mpg (DV) regressed on three IVs (wt, cyl, qsec)

Console:



- $\beta_0 = 29.4291$
- $\beta_1 = -3.8616, \beta_2 = -0.9277, \beta_3 = 0.4945$
- Furthermore, keep an eye on:
  - $Pr(>|t|) \rightarrow$ P-Value (Recap Methods!)
  - Adjusted R-squared (will be covered in next weeks slides)

Exercise 1a and 1b

Suppose that you have the exam results of method and statistics lectures of students who enrolled between Winter Term 2008 - Summer Term 2019. Denote the method results by x and the statistics results y. We model the relationship between x and y by using the following equation:

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

What is the name of this model?

- Only one independent variable $\rightarrow$ Simple regression model

Which component in the equation above is this model's dependent variable?

- Dependent variable = "Outcome variable" $\rightarrow$ y

Exercise 1c and d

You make up groups of students with the same result of the method exam. If the expected value of the error term is not zero for all groups, what assumption of the model above is violated?

- Zero Conditional mean
- $E(u_i) = 0$ for all values of X

Given $Cov(Y, X) = 134.2$ **and** $Var(X) = 215.56$**, estimate** β_0 **and** β_1 **of the model above by using OLS**

- $\hat{\beta}_1 = \frac{Cov(Y,X)}{Var(X)} = \frac{134.2}{215.56} = 0.6226$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 61.9103 - 0.6226 * 65.6793 = 21.0184$

Exercise 1e

Denote your estimate in the last task by $\hat{\beta}_0$ and $\hat{\beta}_1$. What is the name of the following function?: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

- "hat" means, that we only estimate our parameters y, β_0, β_1, based on our sample.
- Sample regression function

Exercise 2a

If the number of total infected persons increases by a constant percentage, which model is more appropriate?

- Model 2 is more appropriate, because here the dependent variable "number of infected persons" appraise in logarithmic form.
- Due to that β_1 multipied by 100 in model 2 represents the percentage change in infected persons for one more day.

Exercise 2b

Calculate the constant percentage increase in the above task based on the results above

- The constant percentage increase is 24.2940%
  - Here you actually just had to find the estimated value for β_1 in the R output of model 2 and multiply it by 100 to get the constant percentage increase.

Model 2

```
Call:
lm(formula = log.infected ~ days)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9265 -0.1780  0.1728  0.3652  0.6051

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.683587   0.176276   9.551 3.73e-11 ***
days        0.242940   0.008308  29.241  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5178 on 34 degrees of freedom
Multiple R-squared:  0.9618,Adjusted R-squared:  0.9606
F-statistic:   855 on 1 and 34 DF,  p-value: < 2.2e-16
```

Exercise 2c

If the number of total infected persons increases by a constant number, calculate the constant number based on the results above.

- The constant number is 339.69
- Here we want to know the estimated value for β_1 in the situation that the total infected persons increases by a constant number.
  - Again you had not to calculate the result, you just had to find the estimated value for β_1 in the R output of model 1.

Model 1

```
Call:
lm(formula = infected.feb.mar ~ days)

Residuals:
    Min      1Q  Median      3Q     Max
-2761.8 -1768.8  -447.6  1667.9  4350.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3266.66     715.48  -4.566 6.24e-05 ***
days          339.69      33.72  10.073 9.67e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2102 on 34 degrees of freedom
Multiple R-squared:  0.749,Adjusted R-squared:  0.7416
F-statistic: 101.5 on 1 and 34 DF,  p-value: 9.673e-12
```

Exercise 2d and e

Suppose that you have y in how many persons have been infected in thousand instead of just the number of infected persons (e.g. 1.345 instead of 1345). Estimate β_1 of Model 1 by using OLS.

- The new estimated value for β_1 is 0.33969
  - OLS estimates change in entirely expected ways when the units of measurement of the dependent or independent variables change.
  - We obtain the new slope by simply dividing the estimate of β_1 in the R output of model 1 by 1,000.

In the same situation of the last task, estimate γ_1 of Model 2 by using OLS.

- The estimated value of γ_1 does not change due to that and is therefore still 0.242940
  - Because the change to logarithmic form of y in model 2 approximates a proportionate change, it makes sense that nothing happens to the slope.

Exercise 2f

Calculate the predicted number of infected persons at the 5th day (x=5) based on Model 1.

- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- $\hat{y} = -3266.66 + 339.69x$
- $\hat{y} = -3266.66 + 339.69 * 5$
- $\hat{y} = -1568.21$
- To calculate the predicted number just form the OLS regression function with the values for intercept and slope estimates provided by the R output of model 1 and include $x = 5$.

Exercise 2g

Calculate the predicted number of infected persons at the 5th day (x=5) based on Model 2.

- $\widehat{log}(y) = \hat{\gamma}_0 + \hat{\gamma}_1 x$
- $\widehat{log}(y) = 1.683587 + 0.242940x$
- $\hat{y} = \exp(1.683587 + 0.242940 * x)$
- $\hat{y} = \exp(1.683587 + 0.242940 * 5)$
- $\hat{y} = 18.14303971$
- Form the OLS regression function for $log(y)$ with the values for intercept and slope estimates provided by the R output of model 2.
- By exponentiating this function and including $x = 5$ we can calculate the predicted number of infected persons.

Exercise 3a and b

What assumption do you need to obtain the second equation, while it is not necessary to obtain the first equation?

- The homoskedasticity assumption
  - With the homoskedasticity assumption in place this equation for the variance of $\hat{\beta}_1$ holds.
  - The homoskedasticity assumption states that the error u has the same variance given any value of the explanatory variable.

What does SST stand for?

- Total sum of squares
  - SST is a measure of the total sample variation in the y_i
  - It measures how spread out the y_i are in the sample
  - If we divide SST by $n-1$, we obtain the sample variance of y

Exercise 3c

What does correspond to the error of $\hat{\beta}_1$?

- C: $\frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{SST_x}$
- The estimator $\hat{\beta}_1$ equals the population slope, β_1, plus this term that is a linear combination in the errors and corresponds to the error of $\hat{\beta}_1$
- Conditional on the values of x_i, the randomness in $\hat{\beta}_1$ is due entirely to the errors in the sample.
- The fact that these errors are generally different from zero is what causes $\hat{\beta}_1$ to differ from β_1.

Exercise 3d and e

If $\hat{\beta}_1$ is unbiased, is u_i always zero?

- No, if $\hat{\beta}_1$ is unbiased $E(u_i|x_i) = 0$, but not u_i.

If $\hat{\beta}_1$ is unbiased, is $\frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{SST_x}$ always zero?

- No, only the expected value of $\frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{SST_x}$ is always zero.To be more correct the expected value of u_i conditioning on the values of x_i in our sample is zero and therefore the whole term is zero.

Exercise 3f, g and h

If the zero conditional mean assumption holds, which two components in the first equation are uncorrelated?

- If the zero conditional mean assumption holds x and u are uncorrelated.

What is σ^2?

- A: σ^2 is the variance of errors

Is $\frac{1}{n}\sum_{i=1}^{n} u_i^2$ an unbiased estimator of σ^2?

- Yes, but unfortunately, this is not a true estimator, because we do not observe the errors u_i.
- But, we do have estimates of the u_i, namely, the OLS residuals $\hat{u}_i$.

Exercise 3i

Is $\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2$ an unbiased estimator of σ^2?

- No, even it is a true estimator, because it gives a computable rule for any sample of data on x and y, this estimator turns out to be biased (although for large n the bias is small).
- The estimator is biased essentially because it does not account for two restrictions that must be satisfied by the OLS residuals. These restrictions are given by the two OLS first order conditions $\sum_{i=1}^{n}\hat{u}_i = 0$ and $\sum_{i=1}^{n} x_i\hat{u}_i = 0$.
- Because it is easy to compute an unbiased estimator, we use that instead.
- The unbiased estimator of σ^2 that we will use makes a degrees of freedom adjustment: $\hat{\sigma}^2 = \frac{1}{(n-2)}\sum_{i=1}^{n}\hat{u}_i^2$