Statistics
Tutorial 04

Philipp Scherer & Jens Wiederspohn

23.05.2020

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!**

Normality assumption

- **Normality assumption**: errors u are normally distributed
  - $u \sim \mathcal{N}(0, \sigma^2)$ MLR.6
  - $u =$ sum of random variables → CLT → normality
  - Before, we assumed $E(u|x) = 0$ and $Var(u|x) = \sigma^2$
    - MLR.6 requires MLR.4 and MLR.5 to hold!
  - Classical linear model = all GM + normality assumption
- Y is a linear function of u, therefore it is normally distributed
  - Linear function of normally distributed variable is also normal
  - $y|x \sim \mathcal{N}(\hat{y}, \sigma^2)$

Normality assumption

- $y|x_1 \sim \mathcal{N}(\hat{\beta}_0 + \hat{\beta}_1 x_1, \sigma^2)$
  - Given each value of x, y follows a normal distribution
    - Note that there is only one independent variable in this graph!
  - Most likely value = expected value = predicted value $\hat{y}$
  - Range of possible y depends on σ^2

Normality of $\hat{\beta}$

- $\hat{\beta}$ is a linear function of u, therefore it is normally distributed
    - $\hat{\beta} = \beta + \sum \frac{\hat{r}_i}{SSR} u_i$
- $\hat{\beta}_j \sim \mathcal{N}\left(E(\hat{\beta}_j), var(\hat{\beta}_j)\right) \rightarrow \mathcal{N}\left(\beta_j, \frac{\sigma^2}{SST_j(1-R_j^2)}\right)$
    - Note that we still use the unknown σ^2 here
- Like any normal random variable, we can standardize $\hat{\beta}_j$
    - Substract mean and divide by standard deviation
    - $Z = \frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim \mathcal{N}(0, 1)$

Hypothesis tests

- True β_j are unknown, but we can still hypothesize about them
  - $\beta = 0, 1, 2, ...$ → true effect in population is 0,1,2,...
  - $\beta \neq 0$ → true effect is different from zero → x has an effect
- Hypotheses can be tested with statistical inference
  - Requires knowledge about distribution of tested parameter
  - Requires test statistic that gives us error probabilities

t test

- Distribution of $\hat{\beta}$ depends on σ^2, which is unknown
  - $\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \frac{\sigma^2}{SST_j(1-R_j^2)}\right)$
    $\rightarrow$ Replace σ^2 by $\hat{\sigma}^2$
- Using $\hat{\sigma}^2$, $\hat{\beta}_j$ follows a t distribution with $df = n - k - 1$
  - $t = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$
  - Most often, we test $\beta = 0$ so that $t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$
- t measures how far (in se) $\hat{\beta}_j$ is from β_j
  - How likely is it to get $\hat{\beta}_j$ under the null that $\beta_j = x$?
  - E.g. if $\hat{\beta}_j$ is $\pm 2se$ away from $\beta_j = 0$, can we reject $H_0 : \beta_j = 0$?

One sided test

Hypothesis testing:

1. Decide on alternative hypothesis H_1 (and null hypothesis H_0)
   - H_0 is opposite of H_1 (statement includes equality)
   - One sided test: is β greater/lower than some value?
     → Often 0 to check whether effect is positive/negative
     - $H_1 : \beta_j > 0$ or $H_1 : \beta_j < 0$
     - $H_0 : \beta_j \leq 0$ or $H_0 : \beta_j \geq 0$
2. Decide on a significance level (usually $\alpha = 0.05$)
   - Max. probability of rejecting H_0 when it is in fact true
3. Compute test statistic t with sample data and β_j
4. Reject H_0 if t-statistic exceeds critical value(from df & α)
   - Always check if test statistic and theorized effect match
   - Careful! We cannot reject H_0 if $c = 1.96$ and and $t = -2.4$

One sided test in *R*

- p-value $= Pr(T \leq t)$, not $Pr(|T| \leq |t|)$ as in R
  → R always reports p-value for two-sided test!
- For *Statistics*, cut reported p-value in half to get one-sided one

Two sided test

- Two sided test allows test about equality of $\hat{\beta}$ and true value
  - $H_1 : \beta_j \neq x$
  - $H_0 : \beta_j = x$
- We usually test whether effect differs from 0 ($H_0 : \hat{\beta}_j = 0$)
  - Again, H_0 includes equality
- Reject H_0 if $|t| > c$ (we no longer care for direction)
  - If H_0 is rejected, x_j is statistically significant at α percent lvl
  - Analogous to critical values (c), we can argue with p-values
    - If $p \leq \alpha$, we can reject H_0
    - p-value $= Pr(|T| \leq |t|)$, where T $=$ rv, t $=$ test statistic
    - Larger t-value → lower p-value

Two sided test

- For two sided test, α is symmetrically divided into two parts
    - Instead of one, we have two rejection regions (greater/lesser)

Significance and power

- Check whether variable is statistically significant
    - If it is, check whether it is practically relevant
    - A significant effect of 0.02 is not necessarily relevant
- But: Do not place too much emphasis on significance
    - We might end up with significant results by accident
      → check statistical power!
    - Statistical power is the probability of a hypothesis test to find an effect if there is an effect to be found
        - Affected by effect size and sample size
        - Check EGAP Power Calculator
          → Useful for papers!

Confidence interval

- Confidence interval (CI) for population parameter β_j
  - $CI(\hat{\beta}) = \hat{\beta}_j \pm c * se(\hat{\beta}_j)$
    - c = critical value based on $(1 - \frac{\alpha}{2})$ percentile and df=n-k-1
    - Replace sd() by se() since we use t-distribution
    - For $df > 120$, we can use standard normal percentiles (Wool.)
- 95% CI = true value is in 95% of constructed CIs
- If CI covers 0, coefficient is not significant at α percent level
  - We cannot exclude possibility that $\beta_j = 0$ (usually H_0)
- *What happens to CI under OVB or heteroskedasticity?*

Testing combinations of parameters

- How can we test combinations of parameters?
  - Is degree in math as good as a physics one? → $H_0 : \beta_1 = \beta_2$
- Standard procedure, but slightly more difficult:
  - Rewrite H_0 as difference: $\beta_1 - \beta_2 = 0$
    → Is difference between coefficients significant?
  - Compute t statistic as usual: $t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{se(\hat{\beta}_1 - \hat{\beta}_2)}$
- New hurdle: what is $se(\hat{\beta}_1 - \hat{\beta}_2)$?
  - Requires statistical programs or matrix algebra
  - Sometimes its easier to just reformulate the model

F-test

- So far, we only tested one restriction at a time
  - What happens if we want to test more restrictions at once?
  - e.g. tutorials and motivation have effect on final grade
    $\rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
  - $H_0 : \beta_1 = \beta_2 = 0$
    H_1 : at least one parameter is $\neq 0$ ($= H_0$ is not true)
- Under H_0, we have restricted model: $y = \beta_0 + u$
  → *How do we predict y in the restricted model?*
  → Which model is better?
- F-test allows to test multiple linear restrictions **jointly**
  - Compare restricted to unrestricted model
  - Test decision is based on how much SSR increases if restrictions are imposed (fewer variables → larger SSR)
    - Is increased SSR large enough to reject restricted model (H_0)?
      → F-statistic ≈ measure of ratio of SSRs

→ *How do we test overall significance of model?*

F-test

$$F = \underbrace{\frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}}_{\hat{\sigma}^2} \sim F_{q,n-k-1}$$

- Shape of F-distribution depends on df_r and df_{ur}
- 'ur' = unrestricted, 'r' = restricted (→ restrictions from H_0)
- $df_{ur} = n - k - 1 =$ denominator df
- q $= df_r - df_{ur} =$ "numerator df"
  - q corresponds to number of restrictions
  - if n=1000 in prev. example: $q = (1000 - 1) - (1000 - 3) = 2$ → restricted: k=0, unrestricted: k=2
- Since $SSR_r \geq SSR_{ur}$ and $SSR \geq 0$, it always holds that $F \geq 0$
  - F cannot be negative!

Hypothesis testing with the F-test

- Hypothesis testing works as usual:
  1. Define H_0 and H_1
  2. Set level of significance α
  3. Compute test statistic ($\rightarrow$ F)
     - MLR.1 to MLR.6 need to hold
  4. Compare F to critical value c (given q, df_{ur} and α)
     - For each level α we have a table for values of df_r and df_{ur}
  5. Decide whether H_0 can be rejected
     - Reject H_0 if $F > c$ or if $Pr(\mathcal{F} > F) \leq \alpha$
     - If H_0 is rejected, tested expl. variables are **jointly significant**

Hypothesis testing with the F-test

FIGURE 4.7 **The 5% critical value and rejection region in an $F_{3,60}$ distribution.**

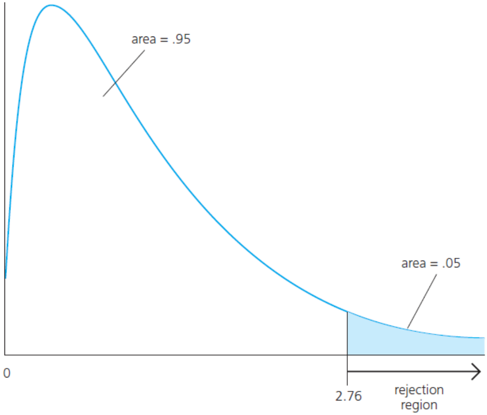

Figure 1: Wooldrige Figure 4.7

F-test

- We can express SSR in terms of R^2
  - $R^2 = 1 - \frac{SSR}{SST} \rightarrow SSR = SST(1 - R^2)$
    - $SSR_r = SST(1 - R_r^2)$
    - $SSR_{ur} = SST(1 - R_{ur}^2)$

$$
\begin{aligned}
F &= \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \\
&= \frac{\left(SST(1 - R_r^2) - SST(1 - R_{ur}^2)\right)/q}{SST(1 - R_{ur}^2)/(n-k-1)} \\
&= \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/df_{ur}}
\end{aligned}
$$

→ 'ur' and 'r' flipped because of minus sign outside the brackets

Overall significance of regression

- When joint exclusion of all independent variables is tested
- H_0: None of the independent variables help to explain y
  $\rightarrow$ all $\beta_j = 0$
- Since $R_r^2 = 0$ when there are no independent variables:

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

- The above F-statistic is reported in R automatically

Exercise 1a and b

Which classical linear model assumption does not belong to the Gauss-Markov-Assumptions?

- Normality (of the population error)
- errors u are normally distributed $\rightarrow u \sim \mathcal{N}(0, \sigma^2)$

Why is the above unique assumption of the classical linear model relevant for multiple regression Analysis?

A Without the assumption, the OLS estimators are biased

B Without the assumption, the OLS estimators are not BLUE

C **With this assumption, we can know the shape of the distribution of $\hat{\beta}$**

$\rightarrow \hat{\beta}$ is a linear function of u, therefore it is normally distributed!

Exercise 1c and d

Suppose you know exactly about the variance of the population error in your classical linear model. Which distribution does $\hat{\beta}$ have?

A **Normal distribution**

→ In cases where σ^2 is known: Z

Suppose you do not know exactly about the variance of the population error in your classical linear model and estimate it by using the variance of the residuals. Which distribution does $\hat{\beta}$ have?

B **t-distribution**

→ In cases where we estimate σ^2 with $\hat{\sigma}_2$: t

Exercise 2a

Complete the tables above by filling parentheses (A) to (G).

- A : 2.042
  - $t_{vegetarian} = \frac{\hat{\beta}_{vegetarian}}{se(\hat{\beta}_{vegetarian})}$
  - $t_{vegetarian} = \frac{1.1096}{0.5434}$
  - $t_{vegetarian} = 2.042$
- B : 0.3115
  - $se(\hat{\beta}_{beer}) = \frac{\hat{\beta}_{beer}}{t_{beer}}$
  - $se(\hat{\beta}_{beer}) = \frac{-0.6429}{-2.064}$
  - $se(\hat{\beta}_{beer}) = 0.3115$
- C : 1.246336
  - $\hat{\beta}_{organic} = t_{organic} * se(\hat{\beta}_{organic})$
  - $\hat{\beta}_{organic} = 3.745 * 0.3328$
  - $\hat{\beta}_{organic} = 1.246336$

Exercise 2a

Complete the tables above by filling parentheses (A) to (G).

- D : 226
  - $df = n - k - 1$
  - $df = 235 - 8 - 1$
  - $df = 226$
- E : 8
  - $k = 8$
  - There are $k = 8$ restrictions (numerator degrees of freedom) when we are determining the overall significance of the regression, because we have $k = 8$ independent variables.

Exercise 2a

Complete the tables above by filling parentheses (A) to (G).

- F : 230
  - $df = n - k - 1$
  - $df = 235 - 4 - 1$
  - $df = 230$
- G : 4
  - $k = 4$
  - There are $k = 4$ restrictions (numerator degrees of freedom) when we are determining the overall significance of the regression, because we have $k = 4$ independent variables.

Exercise 2b

You are interested in the hypothesis “Ceteris paribus, a citizen’s evaluation of the Green party depends on whether s/he is occasional vegetarian.”. Given the result of Model 1, can you reject the corresponding null hypothesis at the 5% level? Answer with yes or no.

- No
  - We can not reject the corresponding null hypothesis, because the p-value of an two sided test is 0.062047 and therefore bigger than 0.05.

Exercise 2c

You are interested in the hypothesis “Ceteris paribus, occasional vegetarians evaluate more positively the Green party than the others.”. Given the result of Model 1, can you reject the corresponding null hypothesis at the 5% level? Answer with yes or no.

- Yes
  - We can reject the corresponding null hypothesis, because the p-value of an one sided test is 0.0310235 (the p-value of the two sided divided by 2) and therefore smaller than 0.05.

Exercise 2d

Construct the 95%-confidence interval of the effect of occasional vegetarian based on the result of Model 1. You can use the table at the bottom of this exercise sheet.

- lower bound: -0.0333 , upper bound: 1.3449
  - $CI(\hat{\beta}) = \hat{\beta}_j \pm c * se(\hat{\beta}_j)$
  - $\underline{\beta_j} = \hat{\beta}_j - c * se(\hat{\beta}_j)$
  - $\overline{\beta_j} = \hat{\beta}_j + c * se(\hat{\beta}_j)$
  - $\underline{\beta_j} = 0.6558 - 1.97052 * 0.3497$
  - $\overline{\beta_j} = 0.6558 + 1.97052 * 0.3497$
  - $\underline{\beta_j} = -0.0333$
  - $\overline{\beta_j} = 1.3449$

Exercise 2e and f

If you compare Model 1 with Model 2, which model is the restricted one?

- Model 2
  - The restricted model always has fewer parameters than the unrestricted model.

To compare Model 1 with Model 2, calculate the F value.

- $F = 2.31329$
  - $F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/df_{ur}}$
  - $F = \frac{(0.2282 - 0.1966)/4}{(1 - 0.2282)/226}$
  - $F = 2.31329$

Exercise 2g

With which distribution do you compare the F value calculated above? Give the name of the distribution. If the distribution has degrees of freedom, you also have to specify them.

- F distribution with df = 4 and 226
  - It can be shown that, under H_0 (and assuming the CLM assumptions hold) F is distributed as an F random variable with $(q, n-k-1)$ degrees of freedom. We write this as $F \sim F_{q,n-k-1}$.
  - $q =$ numerator degrees of freedom $= df_r - df_{ur} = 4$
  - $n - k - 1 =$ denominator degrees of freedom $= df_{ur} = 226$

Exercise 2h and i

Suppose that you test whether the following null-hypothesis can be rejected: “None of the independent variables in Model 2 has an effect on the dependent variable”. Can you reject the null-hypothesis based on the result above? Answer with yes or no.

- Yes
  - We can reject the null hypothesis, because the p-value for the F test for the overall significance of the regression is 0.000000000276 and therefore way smaller than 0.05.

Which test statistic do you rely on for the last task? Give the name of the test statistic and its value.

- F statistic
- $F = 14.07$