Statistics
Tutorial 07

Philipp Scherer & Jens Wiederspohn

17.06.2020

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!**

Quick recap: independent dummy variables

- In regression, we usually assume data with interval-scale
  - Regression line depends on ratio of differences $\frac{\Delta y}{\Delta x}$
- With nominal data, we cannot interpret the differences
  - e.g. difference for CDU=1, SPD=2,Greens=3,...
- Incorporate binary data with dummy explanatory variables
  - Remember experiments? Treatment variable is a dummy
  - Interpret effect as difference between two groups (c.p.)

    → Intercept shift
- Same concept applies for ordinal data
  - Transform categories in dummies and one base group

    → Dummy variable trap!
  - Gives us a constant partial effect for each category

→ *How do we account for different slopes across groups?*

Differences between groups & dummies

- We can allow for intercept difference with a dummy
- What if the slopes also differ across groups?
  → Allow intercept and slope to differ with dummies & interactions
- We test existence of group difference with either:
  1. Include group dummy and all interactions (= unrestricted) and test their joint significance (against 'no effect')
     → Standard F test: restricted vs. unrestricted model
  2. Estimate two separate regressions for the dummy groups and compare with pooled model (H_0: no group difference) with $k+1$ restrictions to group models (SSR_1, SSR_2)
     - test whether the intercept and all slopes are the same across the two groups
     - Under H_0, error variances for the groups must be equal (After all, we assume that there is no group difference)

     → Chow statistic: $F = \frac{SSR_p - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \cdot \frac{n - 2(k+1)}{k+1}$

Dependent dummy variables

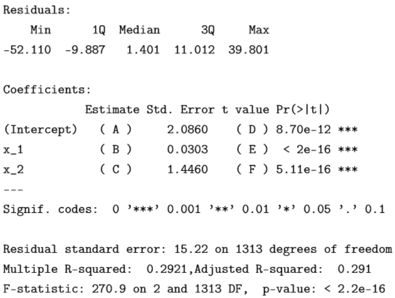
- With regression, we can model dependent dummy variables
  - Model does not give exact $y = 0$ or $y = 1$ (*'success'*)
  - Interpretation changes: What is probability that $y = 1$?

    → **Linear probability model** (LPM)
- In LPM, β_j[1] measures change in $Pr(y = 1|x)$ for Δx_j
  - Apart from that, linear model works as before
- Problematic if predicted $Pr(y = 1)$ is negative/greater than 1
  - Usually works well for values of y that are close to $\bar{y}$

[1] whether we use β_j or β_k does not matter too much

Exercise 1

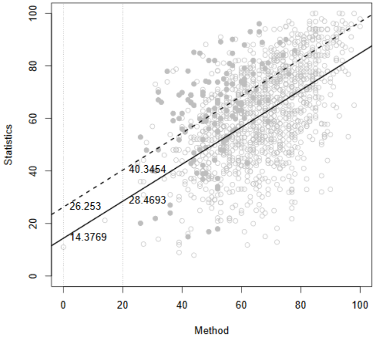
MLR Model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ with y representing the results of Statistics, x_1 those of Methods and x_2 as a binary variable with 1 for those who started to study in 2018 and 0 for the others



```
Residuals:
    Min      1Q  Median      3Q     Max
-52.110  -9.887   1.401  11.012  39.801

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   ( A )     2.0860    ( D ) 8.70e-12 ***
x_1           ( B )     0.0303    ( E )  < 2e-16 ***
x_2           ( C )     1.4460    ( F ) 5.11e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 15.22 on 1313 degrees of freedom
Multiple R-squared:  0.2921,Adjusted R-squared:  0.291
F-statistic: 270.9 on 2 and 1313 DF,  p-value: < 2.2e-16
```



Exercise 1a and b

How many students do you have in your dataset?

- $n = df + k + 1 = 1313 + 2 + 1 = 1316$

Compute the values for (A) - (F)

A Intercept of solid line → 14.3769

B Slope parameter (increase in y if x increases per one unit) $\rightarrow \frac{28.4693 - 14.3769}{20} = 0.7046$

C Difference between solid and dotted line → 11.8761

D $t = \frac{Estimate}{Std.Error} = \frac{14.3769}{2.086} = 6.89209$

E $t = \frac{Estimate}{Std.Error} = \frac{0.7046}{0.0303} = 23.25413$

F $t = \frac{Estimate}{Std.Error} = \frac{11.8761}{1.4460} = 8.213071$

Exercise 1c and d

What is the base group whose intercept corresponds to the overall intercept?

- Base group: $x_2 = 0 \rightarrow$ Students who did NOT start to study in 2018

You are now extending Model 1. Of which model can you NOT estimate the parameters?

C You add two binary variables for major: x_3 (1: political science as major; 0: others) and x_4 (1: other discipline than political science as major; 0: political science as major)

D You add three binary variables for major/minor: x_3 (1: political science as major; 0: others), x_4 (1: political science as minor; 0: others) and x_5 (1: political science as neither major nor minor; 0: others)

→ Danger of dummy variable trap!

Excercise 1e

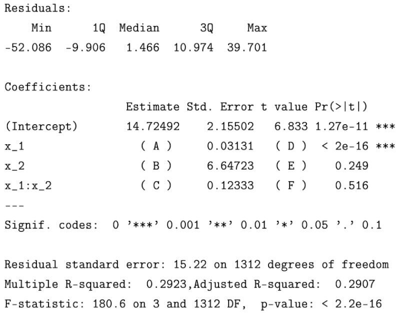
Why can you NOT estimate the parameters of the model(s) above? Name the assumption, which is violated

- No Perfect Collinearity
- x_4 and x_5 are perfect linear functions from x_3 which would lead to perfect collinearity in our model

→ Danger of dummy variable trap!

Exercise 1f

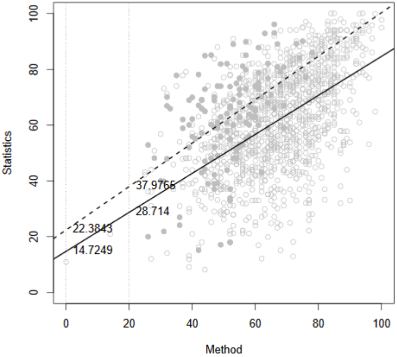
Extended Version of Model 1 with $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$, where $\beta_3 x_1 x_2$ represents the interaction effects between x_1 and x_2



```
Residuals:
    Min      1Q  Median      3Q     Max
-52.086  -9.906   1.466  10.974  39.701

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       14.72492    2.15502   6.833 1.27e-11 ***
x_1                 ( A )     0.03131   ( D )  < 2e-16 ***
x_2                 ( B )     6.64723   ( E )    0.249
x_1:x_2             ( C )     0.12333   ( F )    0.516
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 15.22 on 1312 degrees of freedom
Multiple R-squared:  0.2923,Adjusted R-squared:  0.2907
F-statistic: 180.6 on 3 and 1312 DF,  p-value: < 2.2e-16
```



Exercise 1f

Compute the values for (A) - (F)

A Slope parameter (increase in y if x increases per one unit)

$\rightarrow \frac{28.714-14.7249}{20} = 0.69945$

B Difference between solid and dotted line →7.65934

C Interaction effect (increase in difference between solid and dotted line if x increases per one unit

$\rightarrow \frac{(22.3843-14.7249)-(37.9766-28.714)}{20} = 0.08016$

D $t = \frac{Estimate}{Std.Error} = \frac{0.69945}{0.03131} = 22.33951$

E $t = \frac{Estimate}{Std.Error} = \frac{7.65934}{6.64723} = 1.15226$

F $t = \frac{Estimate}{Std.Error} = \frac{0.08016}{0.1233} = 0.6501217$

Exercise 1g and h

You are interested in the hypothesis that average results of “Statistics” exam are identical for those who started to study in 2018 and those who started earlier if they have the same result at of “Method” exam. For this hypothesis test, which parameter(s) of Model 2 should you focus on? If you need multiple parameters, name all of them

- β_1
- β_3

Which test do you need for the hypothesis test above?

- F-Test

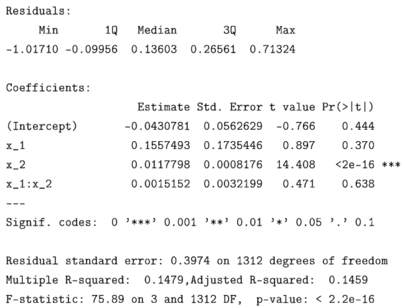
Exercise 1i

Above test is equivalent to test whether the regression line for those who started to study in 2018 and that for the other students are identical. Consequently, you do not have to estimate Model 2 to obtain the corresponding test statistic. What is the name of the test statistic, which you obtain in such a way?

- Chow statistic
- $F = \frac{SSR_p - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \cdot \frac{n - 2(k+1)}{k+1}$
- This particular F statistic is usually called the Chow statistic in econometrics.
- Because the Chow test is just an F test, it is only valid under homoskedasticity.
- In particular, under the null hypothesis, the error variances for the two groups must be equal. As usual, normality is not needed for asymptotic analysis.

Exercise 1j

You replaced the dependent variable of Model 2 with a binary variable whether the students passed the exam or not (1: pass and 0: fail). What is the name of this kind of linear regression model with a binary dependent variable?



```
Residuals:
     Min       1Q   Median       3Q      Max
-1.01710 -0.09956  0.13603  0.26561  0.71324

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.0430781  0.0562629  -0.766    0.444
x_1                0.1557493  0.1735446   0.897    0.370
x_2                0.0117798  0.0008176  14.408   <2e-16 ***
x_1:x_2            0.0015152  0.0032199   0.471    0.638
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 0.3974 on 1312 degrees of freedom
Multiple R-squared:  0.1479,Adjusted R-squared:  0.1459
F-statistic: 75.89 on 3 and 1312 DF,  p-value: < 2.2e-16
```

- linear probability model
- The multiple linear regression model with a binary dependent variable is called the linear probability model (LPM).
- In the LPM, β_j measures the change in the probability of success when x_j changes, holding other factors fixed.

Exercise 1k

The above model must violate one of the Gauss-Markov-Assumption. Which one is that?

- Homoskedasticity
- When y is a binary variable, its variance, conditional on x, is $Var(y|x) = p(x)[1 - p(x)]$, where $p(x)$ is shorthand for the probability of success: $p(x) = \beta_0 + \beta_1 x_1 + ... + \beta_k$.
- This means that, except in the case where the probability does not depend on any of the independent variables, there must be heteroskedasticity in a linear probability model.

Exercise 1l

Predict the probability for a student who started his study in 2008 and whose result of “Method”-exam was 93 points to pass the “Statistic”-exam.

- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$
- $\hat{y} = -0.0430781 + 0.1557493 * x_1 + 0.0117798 * x_2 + 0.0015152 * x_1 * x_2$
- $\hat{y} = -0.0430781 + 0.1557493 * 0 + 0.0117798 * 93 + 0.0015152 * 0 * 93$
- $\hat{y} = 1.052443$

Exercise 1m

We build another prediction variable which is 0 for those whose predicted probability is lower than 0.5 and 1 for the others. Below you will find the contingency table of this variable and the dependent variable. Calculate the percent correctly predicted.

		prediction	
		0	1
Dependent	0	57	265
variable	1	29	965

- $\frac{57+965}{57+29+265+965} * 100\%$
- $\frac{1022}{1316} * 100\%$
- 77.65957%