ECONOMETRICS - Übung I

#Ü1

1.1.: • show: $Var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

→ per definition: $Var(X) = \mathbb{E}(X-\mu)^2$ ↳ $\mu = \mathbb{E}(X)$

$\Rightarrow \mathbb{E}\,Var(X) = \mathbb{E}(X^2 - 2\mu X + \mu^2) =$

$= \mathbb{E}(X^2) - 2\mu^2 + \mu^2 =$

($2\mu \cdot \mathbb{E}(X)$, $\mathbb{E}(a) = a$)

$\mathbb{E}(hX+a) = h\,\mathbb{E}(X) + a$

$= \mathbb{E}(X^2) - \mathbb{E}(X)^2$

$\mathbb{E}(a) = a$	(1.)
$\mathbb{E}(hX+a) = h\,\mathbb{E}(X) + a$	(2)
$\mathbb{E}(h_1 g(X) + h_2 h(X)) =$	
$= h_1 \mathbb{E}(g(X)) + h_2 \mathbb{E}(h(X))$	(3)

↑ calculation rules for $\mathbb{E}(X)$

1.2.: • $X \sim$ Bernoulli(p)

$f(x) = p^x (1-p)^{1-x}$ for $x \in \{0,1\}$

- $\mathbb{E}(X) = \sum_i (x_i \cdot f(x_i)) = \sum_{i=0}^{1} (x \cdot p^x (1-p)^{1-x}) =$

$= 0 \cdot (1-p) + 1 \cdot p \cdot 1 = p$

- $Var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = p - p^2 = p(1-p)$

↓ expl.: $X \in \{0,1\} \Rightarrow$

$\Rightarrow 0^2 = 0;\ 1^2 = 1$

1.6.: • to show: $\mathbb{E}(XY) = \mathbb{E}(X)\cdot\mathbb{E}(Y)$

→ independence $\Rightarrow p(x_i, y) = p_x(x)\cdot p_y(y)$

$\mathbb{E}(X) = \sum_i x_i f(x_i)$

$\Rightarrow \underline{\mathbb{E}(XY)} = \sum_{i\in X}\sum_{i\in Y} x_i y_i p(x_i y_i) =$

$= \sum_{i\in X}\sum_{i\in Y} x_i y_i p_x(x_i) p_y(y_i) = \sum_{i\in X}(x_i p(x_i))\cdot$

$\cdot \sum_{i\in Y}(y_i p(y_i)) = \underline{\mathbb{E}(X)\cdot\mathbb{E}(Y)}$

$\Rightarrow \underline{\mathrm{cov}(X,Y)} = \mathbb{E}[(X-\mu_x)(Y-\mu_Y)] =$

$= \mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))] = \mathbb{E}[XY -$

$- X\mathbb{E}(Y) - Y\mathbb{E}(X) + \mathbb{E}(X,Y)] =$

$= \mathbb{E}(XY) - \mathbb{E}(Y)\mathbb{E}(X) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X,Y) =$

$= \underline{0}$

1.8.: (a) $A = \begin{pmatrix}1&2\\3&4\end{pmatrix}$; $B = \begin{pmatrix}4&5\\6&2\end{pmatrix}$

$AB = \begin{pmatrix}1\cdot4+2\cdot6 & 1\cdot5+2\cdot2\\ 3\cdot4+4\cdot6 & 3\cdot5+4\cdot2\end{pmatrix} = \begin{pmatrix}16&19\\36&43\end{pmatrix}$

$BA = \begin{pmatrix}4\cdot1+5\cdot3 & 4\cdot2+5\cdot4\\ 6\cdot1+2\cdot3 & 6\cdot2+2\cdot4\end{pmatrix} = \begin{pmatrix}19&28\\12&20\end{pmatrix}$ $\neq$

→ to show: $(AB)^t = B^t A^t$

(b) unbiased $\Rightarrow \mathbb{E}(\hat{\theta}_n) = \theta_n$

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathbb{E}\left(\sum X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{n}{n}\mathbb{E}(X_i) = \mathbb{E}(X)$$

→ E-Mail Tutor: tobias.grossauer@tuwien.ac.at

→ *-A Exercises not relevant for exams

1.7.: $$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \underbrace{\frac{1}{n}\sum x_i \bar{y}}_{\bar{x}\bar{y}} - \underbrace{\frac{1}{n}\sum \bar{x} y_i}_{\bar{x}\bar{y}} + \frac{1}{n}\sum \bar{x}\bar{y} =$$

$\frac{1}{n}\sum x_i = \bar{x}$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} y_i(x_i - \bar{x})$$

$\bar{x}\bar{y} = \frac{1}{n}\sum y_i \bar{x}$

105.628 Econometrics for Business Informatics 2017S

Exercise Sheet 2

Due April 11, 2018.

2.1 The following table contains the ACT (*American college testing*) scores and the GPA (*grade point average*) for $n = 8$ college students. The GPA is based on a four-point scale (4 being the best) and has been rounded to one digit after the decimal. The perfect ACT score is a 36.

Student	GPA	ACT
1	2.8	21
2	3.4	24
3	3.0	26
4	3.5	27
5	3.6	29
6	3.0	25
7	2.7	25
8	3.7	30

(a) Estimate the relationship between GPA as a dependent variable and ACT as an explanatory variable using simple least-squares regression

(b) and produce a plot of your results. Comment on the "direction" of the relationship.

(c) What is the expected increase in GPA if the ACT score was increased by one unit?

2.2 Continuation of Problem **2.1**:

(a) Compute the fitted values and residuals for each observation, and verify that the residuals (approximately) sum to zero.

(b) What is the predicted value of GPA when ACT = 20?

(c) How much variation in GPA for these 8 students is explained by ACT?

2.3 For given data points (x_i, y_i) for $i = 1, \ldots, n$ and $x_i, y_i \in \mathbb{R}$ where neither all x_i nor all y_i are equal to the same value for all $i = 1, \ldots, n$, show that

$$S_{x\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(\hat{y}_i - \bar{\hat{y}}) = 0.$$

2.4 For the same setup as in Problem **2.3**, show that

$$S_{yy} = S_{\hat{y}\hat{y}} + S_{\hat{u}\hat{u}}.$$

Problems **2.5** - **2.8** will be used in Chapter 2.

ECO - Exercise II

2.1 (a)(b) → see R code

(c) increase ACT ⟶ increase GPA (how much? ⇒ β_2)

↑ unit

$$\underline{\beta_2} = \frac{s_{xy}}{s_{xx}} = \underline{0,102}$$

$$\underline{s_{xx}} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - \bar{x}^2\right) = \frac{56,875}{8} = \underline{7,109}$$

$$\underline{s_{xy}} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i y_i - \bar{x}\bar{y}\right) = \frac{5,813}{8} = \underline{0,727}$$

2.2 (a) $\underline{\beta_1} = \bar{y} - \beta_2 \bar{x} = \underline{0,573}$

n	$\hat{y}_i = \beta_1 + x_i \cdot \beta_2$	$u_i = y_i - \hat{y}_i$
1	2,714	0,086
2	3,021	0,329
3	3,225	-0,225
4	3,324	0,123
5	3,531	0,068
6	3,123	-0,123
7	3,123	-0,423
8	3,634	0,066

(b) $\hat{y} = 0,573 + 0,102 \cdot 20 = 2,612$

(c) $R^2 = 58\%$ → see R → summary(lma)

2.3 $s_{x\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(\hat{y}_i - \bar{\hat{y}}) = 0$ → to show

drop one derivation: $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(\hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \hat{u}_i)$

$= \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x} y_i - x_i \hat{u}_i + \bar{x}\hat{u}_i)$???

$\hat{y}_i = \beta_1 + \beta_2 x$ ein und dann multiplizieren → see sheet !!

105.628 Econometrics for Business Informatics 2018S

Exercise Sheet 3

Due April 25, 2018.

In most of the exercise problems below, the following definitions are relevant: Similarly as in the simple linear model, we define also in the multiple linear model the *fitted values* $\hat{y}_i = x_i'\hat{\beta}$ for $i = 1, \ldots, n$, so that the vector of $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)'$ satisfies $\hat{y} = X\hat{\beta}$ and the *residuals* $\hat{u}_i = y_i - \hat{y}_i$ for $i = 1, \ldots, n$, so that $\hat{u} = (\hat{u}_1, \ldots, \hat{u}_n)'$ is given by $\hat{u} = y - \hat{y}$.

3.1 For the *simple* linear regression with data points (x_i, y_i) for $i = 1, \ldots, n$, show how the least-squares estimates, the fitted values, the residuals and the coefficient of determination R^2 change when the explanatory variable is changed by a factor of c with $c > 0$, so that the new data points are given by (cx_i, y_i) for $i = 1, \ldots, n$.

3.2 Consider the data set `wagedata.csv` with the variables wage (average hourly earnings), education and experience (in years) of $n = 526$ individuals.

Regress the variable `wage` on a constant as well as the variables `educ` and `exper`. This can (e.g) be done in `R` through the following commands.

```
wagedata = read.table("wagedata.csv")
lm(wage ~ educ + exper, data = wagedata)
```

Verify that the residual vector $\hat{u}$ is orthogonal to the columns of X as well as to the fitted data vector $\hat{y}$.

3.3 Continuation of Problem **3.2**. Also include the variable $\texttt{exper}^2$ into the regression. This can be done by adding `I(exper^2)` as an explanatory variable in the function `lm`.

(a) Compare the value of R^2 to the previous fit. Do the same also when using `log(wage)` as a dependent variable. Which model shows the best fit? Interpret the coefficient for `educ` in the two models with different dependent variables.

(b) Draw a plot the curve $\hat{\beta}_{\texttt{exper}}\,\texttt{exper} + \hat{\beta}_{\texttt{exper}^2}\,\texttt{exper}^2$ as a function of `exper`, once for the fit with `wage` and once for the fit with `log(wage)` as a dependent variable (meaning that the corresponding coefficients $\hat{\beta}_{\texttt{exper}}$ and $\hat{\beta}_{\texttt{exper}^2}$ are taken from those fitted models). Comment.

3.4 Derive a formula for the least-squares estimator $\hat{\beta} \in \mathbb{R}$ in a homogeneous linear regression model (a model with no intercept) with just one explanatory variable, i.e. a model with regressor matrix

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n.$$

3.5 Using the data set from Problem **3.2**, regress `log(wage)` on `educ` only without using an intercept. In `R`, this can be done with the command

```
lm(log(wage) ~ educ - 1, data = wagedata)
```

where -1 indicates that no constant should be included. Verify the formula for the estimator derived in Problem **3.4** and also verify that in this case, `R` outputs the non-centered coefficient of determination $\bar{R}^2$

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n y_i^2}.$$

3.6 Again using the data set from Problem **3.2**, now regress `log(wage)` on `educ` including an intercept. Compute the non-centered coefficient of determination $\bar{R}^2$ as well as the centered coefficient of determination R^2 (the latter one can be read off the `R` output). Verify in this example that

$$R^2 \leq \bar{R}^2.$$

3.7 Derive the formulas for the least squares estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ in the simple linear model from the general formula in the multiple linear model, meaning that for

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

show that

$$\hat{\beta} = (X'X)^{-1}X'y = \begin{pmatrix} \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x} \\ \frac{S_{xy}}{S_{xx}} \end{pmatrix}$$

3.8 Repeat the definitions of positive and negative (semi-)definiteness of symmetric square matrices (4 definitions).

3.9* Show that a matrix of the form $A'A$ where A is a matrix of arbitrary dimension is always positive semi-definite.

3.10* Show that the non-centered coefficient of determination

$$\bar{R}^2 = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2} = \frac{\|\hat{y}\|^2}{\|y\|^2} = 1 - \frac{\|\hat{u}\|^2}{\|y\|^2}$$

satisfies $R^2 \leq \bar{R}^2$ (where R^2 is the centered coefficient of determination from class) in case the corresponding regression contains a constant.

$E(\hat{\beta}) = \beta$

$\hat{y}_i = x_i \hat{\beta}$ | $\cdot c$ $\quad$ $\hat{y}_i = (cx_i)\hat{\beta} =$

$= \begin{pmatrix} cx_{i1} \\ cx_{i2} \\ cx_{ik} \end{pmatrix} \hat{\beta} =$

$= \begin{pmatrix} cx_{i1}\hat{\beta}_1 \\ cx_{ik}\hat{\beta}_k \end{pmatrix} =$

$\hat{u}_i = y_i - \hat{y}_i = \quad \hat{u}_i = y_i - (cx_i)\hat{\beta} = c \cdot (\cdot)$

$= y_i - x_i \hat{\beta}$

$\hat{u} = y - \begin{pmatrix} cx_{11}\, cx_{21} \cdots cx_{k1} \\ cx_{12} \\ \vdots \\ cx_{1n} \cdots cx_{kn} \end{pmatrix} \hat{\beta}$

$\hat{\beta} = (X'X)^{-1} X'y \quad \hat{\beta} = (c^2 X'X)^{-1}?$

$cX'y = c\sum x_{i1} y_{i1} + \cdots$

$X^{n \times k} \quad X^{k \times n}$

$XX^{n \times n}$

$\begin{pmatrix} x_1 \cdots x_k \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} x_1 \cdots x_n \\ \vdots \\ x_k \end{pmatrix} \begin{pmatrix} x_1 \cdots x_k \\ \vdots \\ x_n \end{pmatrix}$

$n \cdot \sum x_i^2 - \left(\sum x_i\right)^2$

$n \cdot \sum x_i^2 - n^2 \bar{x}^2$

$n \cdot \left(\sum x_i^2 - n\bar{x}^2\right)$

$n \cdot \sum (x_i - \bar{x})^2$

$n^2 \cdot S_{xx}$

→ Erklärung (Regel):

$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \left(\frac{1}{ad - bc}\right) \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

→ anwenden!

$X^T y = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$

$\beta = \frac{1}{n^2 S_{xx}} \begin{pmatrix} \sum x_i^2 \cdot \sum y_i - \sum x_i \sum x_i y_i \\ -\sum x_i \sum y_i + n \sum x_i y_i \end{pmatrix} \begin{matrix} b_1 \\ b_2 \end{matrix}$

$b_2 = \frac{-\sum x_i \sum y_i + n\left(\sum (x_i - \bar{x})(y_i - \bar{y}) + \sum x_i \bar{y}\right)}{n^2 S_{xx}}$

$= \frac{-n^2 \bar{x}\bar{y} + n^2 \bar{x}\bar{y} + n \sum (x_i - \bar{x})(y_i - \bar{y})}{n S_{xx}}$

$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$

$b_1 = \left(\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i\right) \cdot \frac{1}{\dots} =$

$= \left[n \cdot \bar{x}^2 n\bar{y} - \sum x_i \left(\sum (x_i - \bar{x})(y_i - \bar{y}) + \sum x_i \bar{y}\right)\right] \cdot \frac{1}{\dots}$

$= \left[n^2 \bar{x}^2 \bar{y} - n\bar{x} \cdot n\bar{x}\bar{y} - n \sum (x_i - \bar{x})(y_i - \bar{y})\right] \cdot \frac{1}{\dots}$

$= \left[n^2 \bar{x}^2 \bar{y} - n^2 \bar{x}^2 \bar{y}\right] \cdot \frac{1}{\dots} = \bar{y} - b_2 \bar{x} =$

$= \left[n\bar{y}\left(n\bar{x}^2 - n\bar{x}^2\right)\right] \cdot \dots = n\bar{y}\left(n \frac{1}{n} \sum x_i^2 - n \frac{1}{n} \sum x_i^2\right) \cdot \dots$

$= n^2 \bar{y} \cdot \left(\frac{1}{n} \sum (x_i - \bar{x})^2\right) = \bar{y} - b_2 \bar{x}$

(2.1) Simple Linear Regression:

$$\beta_{2,\text{[illegible]}} = \frac{S_{xxy}}{S_{xx}} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

$$\beta_{2,\text{alle}} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

ECO - Regression for Multiple

(2D) $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \Rightarrow \hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i \Rightarrow \hat{u}_i = y_i - \hat{y}_i$

$\hat{\beta}_2 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$

(XD) $y_i = \beta_1 x_{i1} + \dots + \beta_k x_{ik}$

$n \times 1 \quad n \times k \quad k \times 1$

$y = X \beta$

$\hat{\beta} = (X'X)^{-1} X' y$

$\hat{u} = y - X\hat{\beta}$

↳ Implications of the normal equations

(2D) $\sum \hat{u} = 0$

$\bar{y} = \bar{\hat{y}}$

$\sum x \hat{u} = 0$

$\sum \hat{y} \hat{u} = 0$

$\sum y_i y_i = \sum \hat{y}_i \hat{y}_i + \sum \hat{u}_i \hat{u}_i$

(XD) precaution: regression contains an intercept

$\bar{\hat{u}} = 0$

$\bar{y} = \bar{\hat{y}}$

$TSS = ESS + RSS$

↳ R^2 - estimator

simple ratio: $R^2 = \frac{ESS}{TSS}$ → it not only shows the correlation

↳ XD - case $R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum \hat{u}_i^2}{\sum (y_i - \bar{y})^2}$

↳ if regression contains an intercept

also, $R^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2}$

105.628 Econometrics for Business Informatics 2018S

Exercise Sheet 4

Due May 16, 2018.

4.1 Explain what a *ceteris paribus* interpretation is (for the coefficients in a multiple linear regression model).

4.2 The median starting salary for new law school graduates is determined by

$$\log(salary_i) \approx \beta_1 + \beta_2\, GPA_i + \beta_3\, LSAT_i + \beta_4 \log(libvol_i) + \beta_5 \log(cost_i) + \beta_6\, rank_i,$$

where $LSAT_i$ is the median LSAT score for the graduating class, GPA_i is the median college GPA for the class, $libvol_i$ is the number of volumes in the law school library, $cost_i$ is the annual cost of attending law school, $rank_i$ is a law school ranking (with $rank_i = 1$ being the best) and $i = 1, \ldots, n$ varies over different law schools.

(a) Explain why we expect $\beta_6 \leq 0$. What signs do you expect for the other slope parameters (slope parameters are all parameters that are not the intercept)? Justify your answers.

(b) Estimate the above regression equation using the data set `lawschool.xls`. What is the predicted *ceteris paribus* difference in salary for schools with a median GPA different by one point? (Report your answer as a percentage.)

(c) Would you say it is better to attend a higher ranked law school? How much is a difference in ranking of 20 worth in terms of predicted starting salary?

4.3 Consider the data set `bwght.csv` and the following description. *A problem of interest to health officials (and others) is to determine the effects of smoking during pregnancy on infant health. One measure of infant health is birth weight; a birth weight that is too low can put an infant at risk for contracting various illnesses. Since factors other than cigarette smoking that affect birth weight are likely to be correlated with smoking, we should take those factors into account. For example, higher income generally results in access to better prenatal care, as well as better nutrition for the mother.* An equation that recognizes this is

$$bwght_i \approx \beta_1 + \beta_2\, cigs_i + \beta_3\, faminc_i$$

(a) Estimate the above equation. Are the signs of the coefficients what you would expect? Comment.

(b) If you convert the variable *bwght* from ounces to grams, how will the estimated coefficients change?

(c) If you do not include the variable *faminc* into the model, how does the model fit change? Other than R^2, which quantity could you use to assess the difference in fits?

4.4 The following equation describes the median housing price in a community in terms of amount of pollution (*nox* for nitrous oxide) and the average number of rooms in houses in the community (*rooms*):

$$\log(price_i) \approx \beta_1 + \beta_2 \log(nox_i) + \beta_3\, rooms_i.$$

What are the probable signs of β_2 and β_3? What is the interpretation of β_2? Explain.

105.628 Econometrics for Business Informatics 2018S

Exercise Sheet 5

Due May 30, 2018.

5.1 Try to motivate $\hat{\sigma}^2 = \frac{\text{RSS}}{n-k}$ as an estimator for σ^2. (You do not need to explain why there is a factor $\frac{1}{n-k}$ instead of $\frac{1}{n}$.)

5.2 We consider a classical linear regression model (CLM) $y = X\beta + u$ and the LS estimator $\hat{\beta}$ in this model. Assume that we are given a "future" observation for the explanatory variables (i.e., a new row for the matrix X) $x_{f.} \in \mathbb{R}^k$ (f stands for "future"). The "predictor" $x'_{f.}\hat{\beta}$ is an estimator for the expected future observation $x'_{f.}\beta$. Show that this predictor is unbiased and compute its variance $\text{Var}(x'_{f.}\hat{\beta})$.

5.3 Assume that we are given the following linear regression model

$$y_i = \beta_1 + \beta_x x_i + \beta_z z_i + u_i \quad \text{for } = 1, \ldots, n$$

under assumptions A1)-A4) and that in the estimation process, we do not use the explanatory variable z_i. (This might be due to various reasons – for instance, we may not know that the variable z_i matters or we may not have data available.) This means that we instead consider the model

$$y_i = \beta_1 + \beta_x x_i + v_i \quad \text{for } = 1, \ldots, n,$$

which does not satisfy the $\mathbb{E}(v_i) = 0$ anymore. Omitting z_i implies that for estimating β_x, we use the formula from the simple model

$$\tilde{\beta}_x = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Show that the bias of $\tilde{\beta}_x$ as an estimator for β_x, $\text{Bias}(\tilde{\beta}_x) = \mathbb{E}(\tilde{\beta}_x) - \beta_x$ is given by

$$\frac{S_{xz}}{S_{xx}}\beta_z.$$

HINT: Summary.

5.4 Consider again the data set `bwght.csv` and the equation

$$bwght_i = \beta_1 + \beta_2 cigs_i + \beta_3 faminc_i + u_i.$$

(a) Compute the LS estimator $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)'$ for $\beta = (\beta_1, \beta_2, \beta_3)'$.

(b) Compute $\hat{\sigma}^2$ through the formula from class and compute $\hat{\sigma}$. Which quantity in your program output corresponds to this?

(c) Compute the 3×3 matrix $\hat{\sigma}^2(X'X)^{-1}$, an estimator for the variance-covariance matrix of $\hat{\beta}$. Use this to come up with estimates for the standard errors of the 3 coefficient estimates. Which quantities in your program output corresponds to this?

105.628 Econometrics for Business Informatics 2017S

Exercise Sheet 6

Due June 13, 2018.

6.1 Consider the data set `wage_full.csv` with hourly wages and regress `log(wage)` on `educ`, `exper` and `tenure`. Test individually if each coefficient is equal to 0 and compute p-values. Explain which quantities in your program output correspond to this test. Comment on the "significance" of the coefficients.

6.2 Explain the idea and procedure of a one-sided t-test.

6.3 Use the same data set as in Problem **6.1**, but now regress `log(wage)` on `educ`, `exper`, `exper`^{2} and `tenure`. Test whether an additional year of education will result in an increase of more than 5% in expected income. Carefully specify both H_0 and H_1 and compute the corresponding test statistic. Carry out the test at the 5%-level by giving the appropriate critical value. Also compute the corresponding p-value.

6.4 In a single plot, plot the density function of a t-distribution for different degrees of freedom, as well as the density function of a standard normal distribution. What can you observe for increasing degrees of freedom?

6.5 Explain what kind of general hypothesis can by tested through an $\mathcal{F}$-test. Give the formula of the corresponding test statistic and its distribution under the null hypothesis.

6.6 Formulate the hypothesis $H_0 : \beta_2 = \cdots = \beta_k$ in the form $H_0 : R\beta = r$, that is, specify the corresponding matrix $R \in \mathbb{R}^{q \times k}$ and $r \in \mathbb{R}^q$. What is q here?

6.7 Same setup as in Problem **6.3**. Compute the quantity

$$F = \frac{R^2}{1 - R^2} \frac{n - k}{q}$$

where R^2 is the centered coefficient of determination and $q = k - 1$. Find this quantity also in your `R` output.

6.8 Same setup as in Problem **6.3**. Based on Problems **6.5** and **6.6**, compute the F-statistic for $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ through the general formula, as well as the corresponding p-value. Which quantities in your program output describe this test?