Midterm on the 13th of November

WS 2020/2021, 11:00 – 13:00

This exam should take you at most 105 minutes. Please allocate enough time to upload your solutions on Learn before 13:00! Please upload one single pdf file only.
Good luck!

1 Case Study 1 – Simple Linear Regression

In this case study we examine the influence of per capita income on per capita expenditure on public schools in all states of the US. To this end we examine a dataset with 50 observations (note that Wisconsin is missing, but Washington DC is included) from 1979, with the columns `Expenditure`, which contains the per capita expenditure on public schools, and `Income`, which contains per capita income.

1.1

We consider the following linear model:

$$Y = \beta_0 + X\beta_1 + u, \quad \mathbb{E}(u|X) = 0. \tag{1}$$

The response variable Y is the per capita expenditure on schools (in dollar per capita), the covariate X is the per capita income (in dollar per capita). The results from this regression and potentially relevant plots can be found on page 8.

What is the OLS estimate of the regression coefficient associated with the explanatory variable `Income`? Report the value of the estimate as well as the interpretation of this parameter. The interpretation should specifically mention the appropriate unit of change.

1.2

What is the sign of the correlation between the variables X `(Income)` and Y `(Expenditure)`? Provide a justification.

1.3

Do the assumptions that $\mathbb{E}(u|X) = 0$ and $\mathbb{V}(u|X) = \sigma^2$ (homoskedasticity) seem to be fulfilled for Model 1? Provide justifications.

1.4

What is the expected change in per capita expenditure on public schools given
a) an increase in per capita income of 1000 dollars and
b) a decrease in per capita income of 2300 dollars?

1.5

We now consider the following log-log model:

$$\log(Y) = \xi_0 + \xi_1 \log(X) + \tilde{u}, \quad \mathbb{E}(\tilde{u}|X) = 0, \tag{2}$$

where Y and X are the same as in question 1.1. Similarly, you can find the results from this regression and potentially relevant plots on page 9.

Specify the model this implies for Y, i.e., provide an equation of the form $Y = \dots$.

1.6

What is the OLS estimate of the regression coefficient associated with the explanatory variable `sales` in the log-log model (2)? Report the value of the estimate as well as the interpretation of this parameter. The interpretation should specifically mention the appropriate unit of change.

1.7

Do the assumptions that $\mathbb{E}(\tilde{u}|X) = 0$ and $\mathbb{V}(\tilde{u}|X) = \sigma^2$ (homoskedasticity) seem to be fulfilled for Model 2? Provide justifications.

1.8

Which of the two models (1) or (2) would you prefer and why?

1.9

How does this regression analysis inform you from a social policy angle? Mention possible measures to address the issues raised.

Midterm on the 13th of November

WS 2020/2021, 11:00 – 13:00

2 Case Study 2 – Multiple Regression

In the following case study, we examine what factors contribute to CEO's salary. The data set contains information on 177 chief executive officers for the year 1990. The dependent variable `salary` records the salary in 1000 USD per unit. Furthermore, the explanatory variables are:

- `mktval`: market value of the compnay (1 unit = 1 million USD)
- `sales`: sales of the company (1 unit = 1 million USD)
- `profits`: profits of the company (1 unit = 1 million USD)
- `ceoten`: CEO tenure in the company (1 unit = 1 year)

These variables will be referred to as Y_{salary} and $X_{\text{mktval}}, X_{\text{sales}}, X_{\text{profits}}$ and X_{ceoten} in models we consider with `salary` as dependent variable and other variables as potential explanatory variables to be included in the model. Furthermore, a set of explanatory variables will be collectively denoted as X whenever appropriate.

Answer the following questions based on the tables found on page 10.

2.1

We fit the multiple regression model

$$\log(Y_{\text{salary}}) = \beta_0 + \beta_1 \log(X_{\text{mktval}}) + \beta_2 \log(X_{\text{sales}}) + u, \quad \mathbb{E}(u|X) = 0$$

where $X = (X_{\text{mktval}}, X_{\text{sales}})$, which we refer to as Model 1.

What is the OLS estimate of the regression coefficient associated with the explanatory variable `sales`? Report the value of the estimate as well as the interpretation of this parameter. The interpretation should specifically mention the appropriate unit of change.

2.2

What is the OLS estimate of the intercept? Provide an interpretation in the current context!

2.3

i) How would you incorporate the variable `profits` in to Model 1? Does it make sense to add it as a log variable?

ii) Now we fit a new model:

$$\log(Y_{\text{salary}}) = \beta_0 + \beta_1 \log(X_{\text{mktval}}) + \beta_2 \log(X_{\text{sales}}) + \beta_3 X_{\text{profits}} + u, \quad \mathbb{E}(u|X) = 0,$$

where $X = (X_{\text{mktval}}, X_{\text{sales}}, X_{\text{profits}})$, which we refer to as Model 2.

What is the OLS estimate for the regression coefficient associated with the explanatory variable `profits`? Report the value of the estimate as well as the interpretation of this parameter. The interpretation should specifically mention the appropriate unit of change.

2.4

Do you think adding `profits` to Model 1 makes sense? To answer this question, first consider the following questions and provide answers:

- Compare the R-squared values for Model 1 and Model 2 we considered previously. Which one is higher?
- Is there a high multicolinearity between the variables `mktval` and `profits`? (Please check the provided correlation values on the table found on page 10.) If so, how would adding the variable `profits` into Model 1 impact the standard error of the OLS estimator of the variable `mktval`?

Finally, provide an answer to the original question whether you deem adding `profits` to Model 1 sensible.

2.5

Now we fit a new model:

$$\log(Y_{\text{salary}}) = \beta_0 + \beta_1 \log(X_{\text{mktval}}) + \beta_2 \log(X_{\text{sales}}) + \beta_3 X_{\text{ceoten}} + u, \quad \mathbb{E}(u|X) = 0,$$

where $X = (X_{\text{mktval}}, X_{\text{sales}}, X_{\text{ceoten}})$, which we refer to as Model 3.
Compare R-squared values for Model 1 and Model 3. Which model would you pick according to this criteria? Indicate your choice and justify your answer.

2.6

Using Model 3, find the average difference between the log-salary of a CEO with 20 years of CEO tenure and the log salary of a CEO with 15 years of CEO tenure.

2.7

Using Model 3, calculate the estimated conditional mean salary of a CEO in a company with a market value of 100 million USD, sales worth 200 million USD who spent their career as a CEO in that company for 5 years.

2.8

Calculate the (unbiased) estimate of the error variance, $(\hat{\sigma}^2)$, implied by Model 3 using the SSR reported in the table on page 10.

2.9

When considering factors that affect the salary of a CEO, one can expect that not only their tenure as a CEO there, but also the number of years they have spent in other positions in the same company prior to becoming a CEO influences their salary.

For this reason, your instructor suggested you to add this varaiable in addition to another variable representing total number of years spent in the company to Model 3. Would you follow your instructor's suggestion and include these two new variables as additional predictors? Provide an answer to this suggestion including your reasoning.

3 True or false?

True or false? For true statements, say that they are true and give a brief explanation of why they are true. For false statements, say that they are false, and provide a correct statement or an explanation as to why they are false.

3.1

Consider the simple linear regression model

$$Y = \beta_0 + \beta_1 X + u,$$

where $\mathbb{E}(u|X) = 0$. Suppose further that $\beta_0 = -1$, $\beta_1 = -3$ and $\mathbb{E}(X) = 2$.

1. If X increases by 2 units, Y decreases exactly by 6 units.
2. The expected value of Y is -1.

3.2

Consider the simple linear regression model

$$Y = \beta_0 + \beta_1 X + u,$$

where $\mathbb{E}(u|X) = 0$, $\mathbb{V}(u|X) = \sigma^2 > 0$. Suppose you have a random sample (y_i, x_i), $i = 1, \ldots, N$, from this population. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the OLS estimators of β_0 and β_1.

1. The larger the sample variation $s_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$ of the regressor, the smaller is the variance of the estimator $\hat{\beta}_1$.
2. The OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are always uncorrelated.

3.3

Consider the multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u,$$

where $\mathbb{E}(u|X_1, X_2, X_3) = 0$, $\mathbb{V}(u|X_1, X_2, X_3) = \sigma^2 > 0$. Let $(y_i, x_{1,i}, x_{2,i}, x_{3,i})$, $i = 1, \ldots, N$, be a random sample.
Let $\hat{\boldsymbol{\beta}}$ be the OLS estimator for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$.

1. It holds that $y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \hat{\beta}_3 x_{3,i}) = 0$ for all $i = 1, \ldots, N$.
2. It holds that $\hat{\beta}_j = \beta_j$ for all $j = 0, 1, 2, 3$.
3. Consider data from several local elections in different cities with three possible parties A, B and C. Suppose Y is the total number of votes, X_1 indicates the proportion of votes for party A, and similarly, X_2 and X_3 indicate the proportion for party B and C, respectively.
   Then the OLS estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ can be computed as

   $$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \qquad \text{where} \quad \boldsymbol{y} = (y_1, y_2, \ldots, y_N)'.$$

3.4

Consider the regression model

$$Y = cX_1^{\beta_1} X_2^{\beta_2} e^u,$$

where $Y > 0,\ c > 0,\ \mathbb{E}(u|X_1, X_2) = 0$, $\mathbb{V}(u|X_1, X_2) = \sigma^2 > 0$.

1. The conditional expectation $\mathbb{E}(\log(Y)|X_1, X_2)$ is a linear function in β_1 and β_2.

2. If X_2 increases by 1% while X_1 is held fixed, we expect Y to increase by β_2 units.

3.5

Suppose you fit a simple linear regression model, relating weight (dependent variable), Y, in kg to height (explanatory variable), X, in cm. Your sample contained $N = 1000$ adults whose heights range from 150cm to 200cm. The fitted model takes the form

$$Y = X - 100 + u,$$

where u has mean 0 given X.
The fitted model can be used by paediatritians to determine whether their patients are obese or not.

Midterm on the 13th of November

WS 2020/2021, 11:00 – 13:00

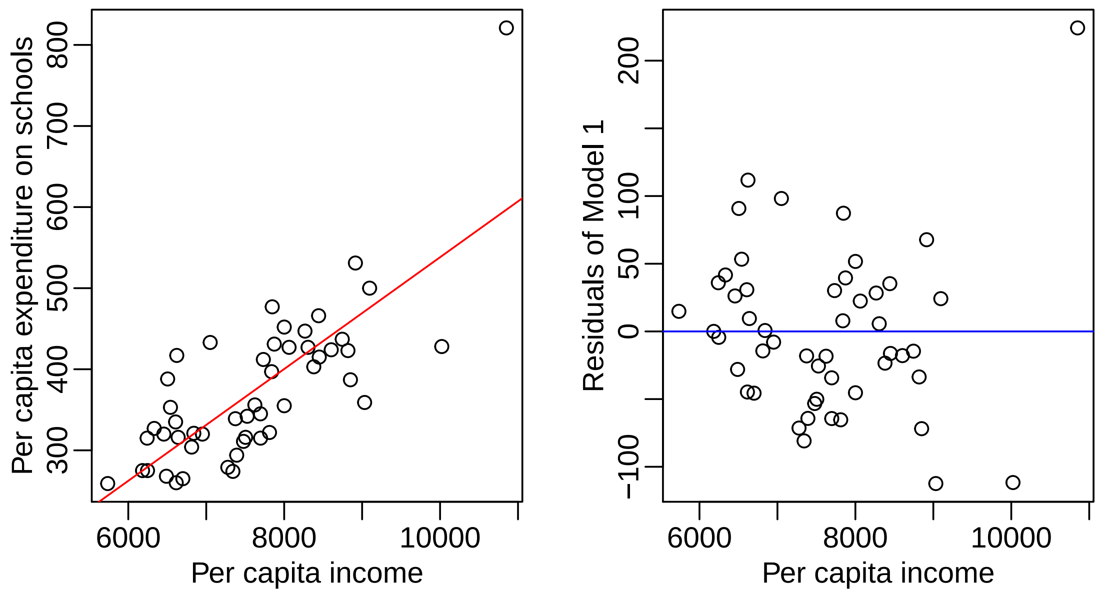

Figure 1: Plots belonging to Model 1 from Case Study 1

	Model 1
(Intercept)	−151.27 (64.12)
Income	0.07 (0.01)
SSR	181015.217(df = 48)
R^2	0.59
Adj. R^2	0.58
Num. obs.	50

. Standard errors reported in parentheses

Table 1: Estimation results for Model 1

Midterm on the 13th of November

WS 2020/2021, 11:00 – 13:00

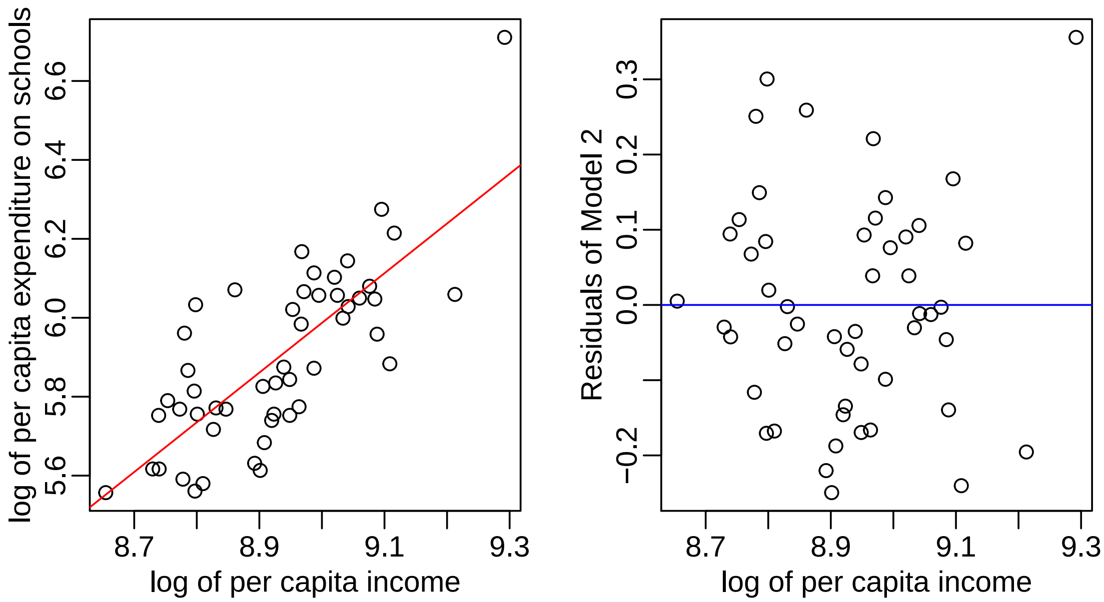

Figure 2: Plots belonging to Model 2 from Case Study 1

	Model 2
(Intercept)	−5.35 (1.38)
log(Income)	1.26 (0.15)
SSR	1.017(df = 48)
R^2	0.58
Adj. R^2	0.57
Num. obs.	50

. Standard errors reported in parentheses

Table 2: Estimation results for Model 2

	Model 1	Model 2	Model 3
(Intercept)	4.62092 (0.25441)	4.68692 (0.37973)	4.50379 (0.25723)
log(mktval)	0.10671 (0.05012)	0.09753 (0.06369)	0.10924 (0.04959)
log(sales)	0.16213 (0.03967)	0.16137 (0.03991)	0.16285 (0.03924)
profits		0.00004 (0.00015)	
ceoten			0.01171 (0.00533)
SSR	45.31(df = 174)	45.295(df = 173)	44.079(df = 173)
R^2	0.29911	0.29934	0.31815
Adj. R^2	0.29106	0.28719	0.30633
Num. obs.	177	177	177

. Standard errors reported in parentheses

Table 3: Statistical models

Table 4: Correlation matrix of coefficient estimates

	salary	mktval	sales	profits	ceoten
salary	1.0000000	0.4063071	0.3802239	0.3939276	0.1429477
mktval	0.4063071	1.0000000	0.7546616	0.9181280	0.0066094
sales	0.3802239	0.7546616	1.0000000	0.7982872	-0.0677147
profits	0.3939276	0.9181280	0.7982872	1.0000000	-0.0216068
ceoten	0.1429477	0.0066094	-0.0677147	-0.0216068	1.0000000