1st online quiz (until Mon May 15, 14:15 = 2:15 pm / 1 point): Some simple sums [graded May 21]

Started on	Saturday, 13 May 2023, 10:12 PM
State	Finished
Completed on	Saturday, 13 May 2023, 10:22 PM
Time taken	9 mins 40 secs
Grade	**1.00** out of 1.00 (**100**%)

Question 1

Correct

Mark 0.60 out of 0.60

Flag question

Calculate the following sums:

a. $\sum_{j=1}^{4} j^3 =$ 100 ✓

b. $\sum_{i=4}^{8} (i+3) =$ 45 ✓

c. $\sum_{k=1}^{3} (k-4)^2 =$ 14 ✓

You should have obtained:

a. 100
b. 45
c. 14

2nd online quiz (until Friday, May 26 - 23:59 = 11:59 pm / 3.6 points): Simple Linear Regression [graded May 29]

Started on	Thursday, 25 May 2023, 10:36 PM
State	Finished
Completed on	Thursday, 25 May 2023, 10:45 PM
Time taken	8 mins 59 secs
Grade	**3.60** out of 3.60 (**100**%)

Question 1

Correct

Mark 1.80 out of 1.80

Flag question

Simple linear regression:

You are supposed to analyse a dataset in R/RStudio with two variables:

X explanatory variable
Y dependent variable

Your tasks (please do save all your R code to upload it with your solutions):

- Download the dataset Simple-32209.csv (click to download). Note, that this is an English .csv file.
- Estimate a simple linear regression to explain Y by X.
- Draw a scatterplot and add the estimated regression line.
- Using R, determine the values for a. to f. (see below) and enter them in the provided fields.
- Your R code should run in R/RStudio without errors and calculate all required terms.
- Finally save your R code into an R script (or R Markdown file) and upload it, see below.

For the following entries: Do not round too much! If relevant, give numerical values with at least 4 decimal places.

a. Sample size n = 42 ✓

b. Estimated intercept = 0.1005218 ✓

c. Estimated slope = -1.4509552 ✓

d. Coefficient of determination = 0.401859 ✓

e. Residual sum of squares = 41.55212 ✓

f. For $x = 0.9$ predict $\hat{y}$ = -1.205338 ✓

This is the `summary` for your linear regression:

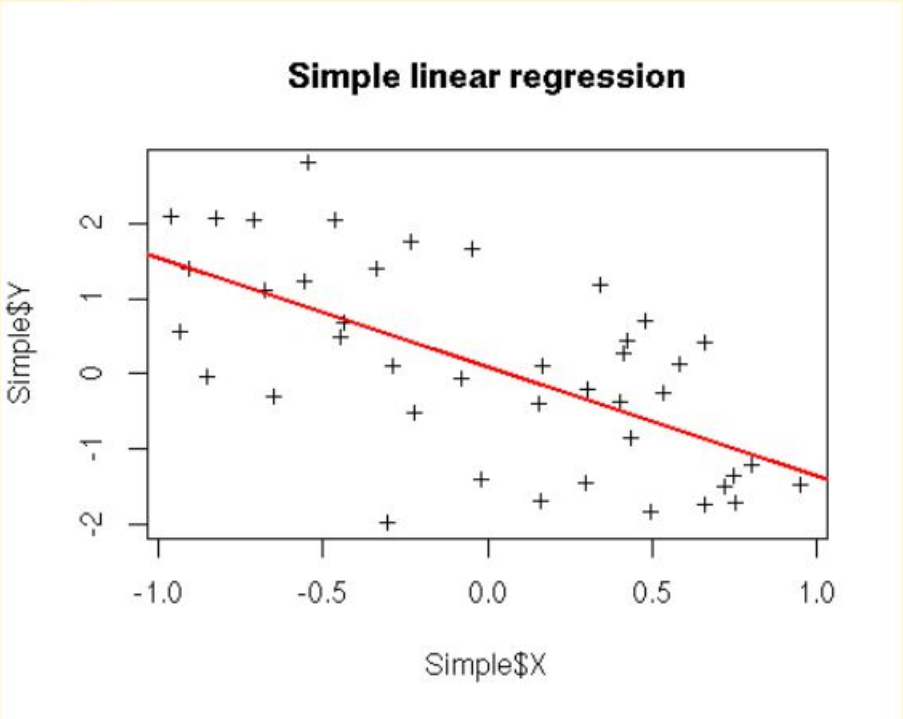
```
Call:
lm(formula = Y ~ X, data = Simple)

Residuals:
     Min       1Q   Median       3Q      Max
-2.51200 -0.67192 -0.04618  0.78923  1.91086

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1005     0.1573   0.639    0.526
X            -1.4510     0.2799  -5.184 6.58e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.019 on 40 degrees of freedom
Multiple R-squared:  0.4019,    Adjusted R-squared:  0.3869
F-statistic: 26.87 on 1 and 40 DF,  p-value: 6.576e-06
```

Your graph should look similar to:



You should have obtained the values:

a. n = 42
b. $\widehat{\beta}_0$ = 0.1005
c. $\widehat{\beta}_1$ = -1.451
d. R^2 = 0.4019
e. RSS = 41.5521
f. For $x = 0.9$ we predict $\hat{y}$ = -1.2053

3rd online quiz (until Monday, June 19 - 14:15 = 2:15 pm / 2.4 points): Summary of a LM [graded July 3]

Started on	Thursday, 15 June 2023, 2:57 PM
State	Finished
Completed on	Thursday, 15 June 2023, 3:11 PM
Time taken	14 mins 30 secs
Grade	**2.40** out of 2.40 (**100**%)

Question 1

Correct

Mark 1.20 out of 1.20

Flag question

Summary of a linear regession

A linear regression model was estimated using R and its summary output is:

```
Call:
lm(formula = Y ~ X + I(X^2) + I(X^3) + I(X^4), data = Data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.17116 -0.63542 -0.04809  0.49752  1.98267

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1439     0.2083  10.290 3.60e-13 ***
X            -1.3938     0.5204  -2.678   0.0104 *
I(X^2)        7.4769     1.3728   5.446 2.32e-06 ***
I(X^3)       -3.9092     0.7864  -4.971 1.11e-05 ***
I(X^4)       -2.6577     1.5426  -1.723   0.0921 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8144 on 43 degrees of freedom
Multiple R-squared:  0.9258,    Adjusted R-squared:  0.9189
F-statistic: 134.1 on 4 and 43 DF,  p-value: < 2.2e-16
```

Your tasks

- Based on this summary output, find answers to the questions a.-d. below.
- In addition, write into the field below how you obtained your solutions (explain briefly or give an R formula).

In addition, enter the values for a.-c. right here. (If relevant, give the numerical values with at least 4 decimal places.)

a. What is the value of the sample size?
b. How many of the coefficients (incl. intercept) are significantly different from zero at the level 5%?
c. What is the predicted value of the regression function for $X = 2.5$?
d. Explain the value 4 in the last line of the summary. (How is it related to the data or the model?)

a. 48 ✓

b. 4 ✓

c. -119.5076 ✓

You should have obtained:

a. 48
b. 4
c. -119.5076

Question 2

Complete

Mark 1.20 out of 1.20

Flag question

Please give short explanations for your solutions to a.-d. here:

a) Given, n-p-1 =43, and p = 4. So, n = 43 + p + 1 = 43 + 4 + 1 = 48.

b) From Pr(>|t|) column, we can see that 4 values are less than 0.05 or 5%.

c) If X = 2.5, then Y_hat = 2.1439 + (2.5 * -1.3938) + (7.4769 * (2.5^2)) + (-3.9092 * (2.5^3)) + (-2.6577 * (2.5^4)) = -119.5076

d) 4 or "p" represents the number of parameters estimated in the model. In other words, p is the number of coefficients without considering the intercept.

Comment:
all ok

4th online quiz (until Friday, July 7 - 23:59 = 11:59 pm / 5 points): Multiple Linear Regression [graded July 10]

Started on	Thursday, 29 June 2023, 3:06 PM
State	Finished
Completed on	Thursday, 6 July 2023, 3:47 PM
Time taken	7 days
Grade	**4.20** out of 5.00 (**84**%)

Question 1

Correct

Mark 2.50 out of 2.50

Flag question

Linear regression for Munich rent data:

You are asked to analyse a subset of the Munich rent data. The sample consists of appartments, which are characterized by:

`netrent` (net rent per month, in Euro),
`space` (living space, in square metres),
`rooms` (no. of rooms),
`year` (year of construction),
`kitchen` (upscale kitchen: yes/no).

Your tasks (please do save all your R code to upload it with your solutions):

- Download the dataset Rent-17807.csv (click to download).
- Estimate the four linear regression models `m1` - `m4` described below.
- Find (using R) the answers for a. - e. and enter them below.
- Within your R code do also provide solutions to the additional questions f. - h. (just add comments to your R code).
- Finally, save your R code (into an .R or .Rmd file). Upload it together with the plot file you obtained in task h. (you find the upload area at the end of this page).

Models to estimate

Model `m1` : Explain `netrent` by `space`.
Model `m2` : Explain `netrent` by `space` and `space^2`.
Model `m3` : Explain `netrent` by `rooms` and `year` and `space` .
Model `m4` : Explain `netrent` by `space` and `space^2` and `year` and `kitchen`.

Additional questions f.-h.

f. For models `m1` and `m2` do a scatterplot of the data and draw the estimated regression curves. Save the plot into a file (to upload it).

g. Using model `m4` - what difference in net rent do you expect between an appartment with and without upscale kitchen?

h. Compare the four models using F tests. Which F tests could be performed? What are the results? Mention for each F test the model you would decide for.

Questions a.-e.

Give the following answers right here. For b. and c. give the numerical values with at least 3 decimal places.

a. The sample size equals: 99 ✓

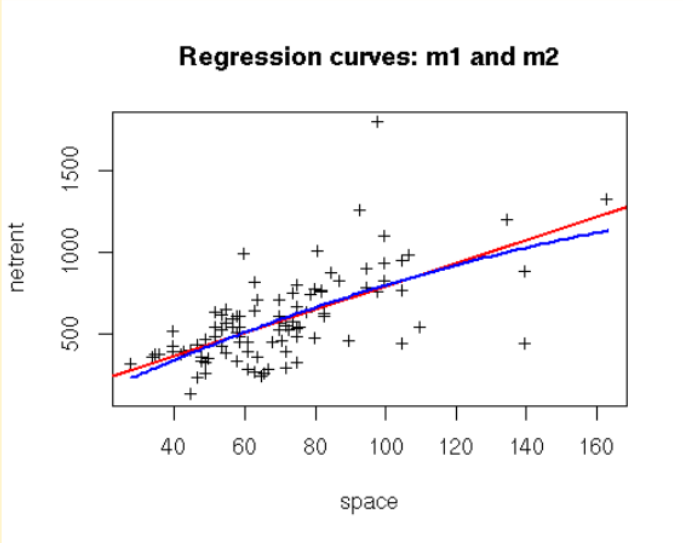
b. Enter $\widehat{\sigma}$ for model `m3`: 183.8335 ✓

c. Using model `m4`, predict the net rent for an appartment of 60 square metres that was built in 1975 and has *no* upscale kitchen: 539.9545 ✓

d. Which of the four models has the smallest AIC value? Answer by entering `m1`, `m2`, `m3` or `m4`: m3 ✓

e. The coefficient for rooms in model m3 is significant at 5%. ✓

Your graph for for the regression curves of `m1` and `m2` should look similar to:



These are the summary outputs for models `m3` and `m4`:

```
Call:
lm(formula = netrent ~ rooms + year + space, data = Rent)

Residuals:
    Min      1Q  Median      3Q     Max
-536.15 -104.65  -43.39  119.17  850.19

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4103.767   1453.219  -2.824 0.005780 **
rooms        -142.249     38.551  -3.690 0.000374 ***
year            2.170      0.739   2.937 0.004163 **
space          11.662      1.427   8.174 1.29e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 183.8 on 95 degrees of freedom
Multiple R-squared:  0.5273,    Adjusted R-squared:  0.5124
F-statistic: 35.33 on 3 and 95 DF,  p-value: 2.002e-15

-------------------------------------------------


Call:
lm(formula = netrent ~ space + I(space^2) + year + kitchen, data = Rent)

Residuals:
    Min      1Q  Median      3Q     Max
-441.14 -108.50    5.18  108.88  767.66

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.765e+03  1.538e+03  -2.448  0.01620 *
space        1.082e+01  3.652e+00   2.963  0.00386 **
I(space^2)  -2.579e-02  2.145e-02  -1.202  0.23230
year         1.898e+00  7.891e-01   2.405  0.01811 *
kitchenyes   1.895e+02  7.919e+01   2.394  0.01867 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 191.5 on 94 degrees of freedom
Multiple R-squared:  0.4925,    Adjusted R-squared:  0.4709
F-statistic:  22.8 on 4 and 94 DF,  p-value: 3.47e-13
```

The four AIC values are: 1337.161, 1338.366, 1319.245, 1328.29

You should have obtained:

a. 99
b. 183.834
c. 539.954
d. m3
e. false / false / true: The coefficient for `rooms` in model `m3` is significant at 5%. / false

5th online quiz (until Wednesday, July 19 at 23:59 = 11:59 pm / 3 points): Summary of a GLM [graded July 21]

Started on	Saturday, 15 July 2023, 1:23 PM
State	Finished
Completed on	Saturday, 15 July 2023, 1:34 PM
Time taken	11 mins 28 secs
Grade	**3.00** out of 3.00 (**100**%)

Question 1

Correct

Mark 1.75 out of 1.75

Flag question

Summary of a GLM

A generalized linear model `m1` was estimated using R and its summary output is:

```
Call:
glm(formula = Y ~ X + I(X^2) + I(X^3) + I(X^4), family = binomial(link = "logit"),
    data = Daten)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0958     0.4094   2.677  0.00743 **
X            -2.6634     0.9396  -2.834  0.00459 **
I(X^2)       -3.0258     2.6686  -1.134  0.25685
I(X^3)        4.1322     1.4754   2.801  0.00510 **
I(X^4)        3.8514     3.0385   1.268  0.20496
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 142.91  on 112  degrees of freedom
Residual deviance: 131.91  on 108  degrees of freedom
AIC: 141.91

Number of Fisher Scoring iterations: 4
```

This model was also compared to a simpler model `m0` for the same data using the χ^2 test performed by `anova(m0,m1,test="Chisq")`:

```
Analysis of Deviance Table

Model 1: Y ~ X + I(X^2) + I(X^3)
Model 2: Y ~ X + I(X^2) + I(X^3) + I(X^4)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       109     133.56
2       108     131.91  1   1.6502   0.1989
```

Your tasks

- Based on this summary output, find answers to the following questions (a.-e.).
- In addition, write into the field below how you obtained your solutions for a.-e. (describe briefly or give an R formula).

Give the numerical values for c. with 3 decimal places and for d. with at least 2 decimal places:

a. What is value of the sample size?
b. Using the summary of `m1`: How many of the coefficients (incl. intercept) are significantly different from zero at the level 1% ?
c. Using the summary of `m1`: Determine the predicted probability of $Y = 1$ given $X = 0.9$.
d. Using the summary of `m1`: What is the value of the log-likelihood for the estimated model?
e. Considering the χ^2 test: Which model would you choose given a significance level of 1% ? (Answer by entering `m0` or `m1`.)

a. 113 ✓

b. 3 ✓

c. 0.8565737 ✓

d. -65.955 ✓

e. m0 ✓

You should have obtained:

a. 113
b. 3
c. 0.857
d. -65.96
e. m0

Question 2

Complete

Mark 1.25 out of 1.25

Flag question

Please give short explanations for your solutions to a.-e. here:

a) df = n-p-1 | here, p=4, and df = 108 | so, n = 113

b) Based on these p-values, we can conclude that three coefficients (including the intercept) are significantly different from zero at the 1% significance level.

c) Y1 <- b0 + (b1 * X) + (b2 * (X^2)) + (b3 * (X^3)) + (b4 * (X^4)); Y1

P1 <- 1 / (1 + exp(-Y1)); P1

d) AIC <- 141.91; P <- 5; loglik <- ((2 * P) - AIC) / 2

loglik # -65.955

e) In the output, the p-value for the chi-square test is given as "0.1989". Since this p-value is greater than the significance level of 1%, we fail to reject the null hypothesis and conclude that there is no significant difference between the two models.

Therefore, based on the chi-square test and the given significance level, we would choose the simpler model, m0, over the more complex model, m1.