**Exercises for exercise class 1 in MMS075, Jan 22, 2020**

1. A (hypothetical) very large company called Maintain-IT is responsible for a project task that needs to be repeated every year. They want to determine how the number of employees assigned to the project affects the completion time. An analyst at Maintain-IT decides to use simple linear regression to model this dependence, based on the following observations:

|  |  |  |
| --- | --- | --- |
| Year | Employees assigned to the project | Completion time (days) |
| 1 | 70 | 20 |
| 2 | 30 | 60 |
| 3 | 10 | 100 |
| 4 | 90 | 20 |

Do the following steps using the above data:

1. Plot (x,y) values (draw by hand);
2. Draw an approximate line that you think would fit the points;
3. Compute the least squares coefficient estimates and corresponding 95% confidence intervals;
4. Draw the least squares line (did you guess approximately right in advance?)
5. Compute the proportion of variability in completion time explained by the number of employees assigned to the project.

Based on all these steps, how would you interpret the results of the linear regression model? Do you think that the decision taken by the analyst of using a linear regression model was justified? Try to find arguments for and against this modelling!

1. In the context of the advertising example in ISL, the three plots below show sales in 1000 units as a function of 1000 dollars invested.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| TV | Radio | Newspaper |

Where do you expect linear modelling to give the best fit to the data points? Where do you expect it to give the worst fit?

1. For the same data sets as in exercise 2, we get three separate outputs from linear modelling, see below. Is this in line with your answer to question 2? Explain how you can check this!

**Output 1:**

Call:

lm(formula = sales ~ TV)

Residuals:

 Min 1Q Median 3Q Max

-8.3860 -1.9545 -0.1913 2.0671 7.2124

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.032594 0.457843 15.36 <2e-16 \*\*\*

TV 0.047537 0.002691 17.67 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.259 on 198 degrees of freedom

Multiple R-squared: 0.6119, Adjusted R-squared: 0.6099

F-statistic: 312.1 on 1 and 198 DF, p-value: < 2.2e-16

**Output 2:**

Call:

lm(formula = sales ~ radio)

Residuals:

 Min 1Q Median 3Q Max

-15.7305 -2.1324 0.7707 2.7775 8.1810

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.31164 0.56290 16.542 <2e-16 \*\*\*

radio 0.20250 0.02041 9.921 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.275 on 198 degrees of freedom

Multiple R-squared: 0.332, Adjusted R-squared: 0.3287

F-statistic: 98.42 on 1 and 198 DF, p-value: < 2.2e-16

**Output 3:**

Call:

lm(formula = sales ~ newspaper)

Residuals:

 Min 1Q Median 3Q Max

-11.2272 -3.3873 -0.8392 3.5059 12.7751

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 12.35141 0.62142 19.88 < 2e-16 \*\*\*

newspaper 0.05469 0.01658 3.30 0.00115 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.092 on 198 degrees of freedom

Multiple R-squared: 0.05212, Adjusted R-squared: 0.04733

F-statistic: 10.89 on 1 and 198 DF, p-value: 0.001148

1. Formulate an interpretation of the linear modelling results for each of the three advertisement forms considered, based on the outputs presented in exercise 3. Compare the results to each other. What does this comparison suggest?
2. Exercise 8, part a) in ISL (page 121); see computer outputs below:

Call:

lm(formula = Auto$mpg ~ Auto$horsepower)

Residuals:

 Min 1Q Median 3Q Max

-13.5710 -3.2592 -0.3435 2.7630 16.9240

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 39.935861 0.717499 55.66 <2e-16 \*\*\*

Auto$horsepower -0.157845 0.006446 -24.49 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.906 on 390 degrees of freedom

Multiple R-squared: 0.6059, Adjusted R-squared: 0.6049

F-statistic: 599.7 on 1 and 390 DF, p-value: < 2.2e-16

 2.5 % 97.5 %

(Intercept) 38.525212 41.3465103

Auto$horsepower -0.170517 -0.1451725

1. Group discussion: come up with at least one example related to logistics where simple linear regression could be used.
2. Exercise 6 in ISL (page 121).
3. Feedback quiz (optional): Go to [www.menti.com](http://www.menti.com) and use the code 31 25 43.