**Exercise solutions 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) 🡪 you want to explain the completion time as a function of the number of employees. Therefore, in the linear model, the input variable, X, is the number of employees and the output (or response), Y, is the completion time. The plot should look approximately like this:



1. Draw an approximate line that you think would fit the points;
2. Compute the least squares coefficient estimates and corresponding 95% confidence intervals;

See the separate document “Exercise class 1 - solution to Exercise 1, parts c and e.pdf”

1. Draw the least squares line (did you guess approximately right in advance?)

The plot should look approximately like this:



1. Compute the proportion of variability in completion time explained by the number of employees assigned to the project.

See the separate document “Exercise class 1 - solution to Exercise 1, parts c and e.pdf”

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!

The linear model is a simple model that is easy and quick to prepare and present. The points seem to be close to a line, and the computation shows that about 91% of the variation in completion time can be explained using the simple linear model with the number of employees as a predictor.

However, there could be other, more elaborate ways that describe the dependence of completion time on the number of employees based on more data than these four points, and they should be considered if a more elaborate estimate is needed. Also, the linear model gives meaningless results outside the range of the observed x values; for example, it gives that the project would take 100 days without any employees working on it and that having more than 100 employees on the projects allows completion in the past. (Generally, predictions should be made within the range of observed predictor values, i.e. with the number of employees between 10 and 90, but not outside that interval).

Overall, the linear model seems appropriate as a first approximation.

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?

The plots suggest a good fit in case of TV advertisements, where the points seem to be close to a line, and the worst fit for Newspaper, where no obvious linear relationship can be observed on the plot.

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!

The results are in line with the answer to 2: the R2 values indicating the proportion of variance in sales explained by the advertisement spending, shown as “Multiple R-squared” in the outputs, is highest in the first output, with TV as predictor, and lowest in the last one with Newspaper as predictor.

**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?

The slope values in the different models indicate the increase in sales in 1000 units as a result of additional $1000 investment in advertisements of the given type. Comparing the outputs shows that the model with radio advertisement spending as predictor has the highest slope of 0.203, indicating that additional $1000 on radio advertisements is expected to result in an increase of 203 units in sales. Since the corresponding numbers are 48 for TV and 55 for newspaper advertisements, it seems like investing all our advertisement budget in radio ads should be the best strategy (but we will see in later classes that this is not the full truth).

1. 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. Yes, the extremely small p-value (which we will discuss in Lecture 2) indicates that “no relationship” can be rejected.
2. The R2 value shows that there is a strong linear relationship between the predictor and the response.
3. The relationship is negative, as witnessed by the negative slope.
4. The predicted mpg for horsepower=98 can be computed as 39.936 - 98\*0.1578 = 24.4716. The interval estimates were not discussed in detail yet, and the necessary R output was not provided in the exercise, so the R code to get it will be given below:

> library(ISLR)

> attach(Auto)

> Ex8Model=lm(mpg~horsepower)

> newdata=data.frame(horsepower=98)

> predict(Ex8Model,newdata,interval="predict")

 fit lwr upr

1 24.46708 14.8094 34.12476

> predict(Ex8Model,newdata,interval="confidence")

 fit lwr upr

1 24.46708 23.97308 24.96108

(Note: if this code does not work for you, then you may need to install the ISLR package in RStudio under “Tools > Install Packages” in the menu.)

Therefore, the associated confidence interval is [23.97,24.96] and prediction interval is [14.8094 34.12476].

1. Group discussion: come up with at least one example related to logistics where simple linear regression could be used.

This will be addressed in Exercise class 2.

1. Exercise 6 in ISL (page 121).

