Exercises for exercise class 3 in MMS075, Feb 5, 2020

1. We first consider the last exercise from the previous class, with an extra hint at the bottom of the text:
   An analyst has used a multiple linear regression model to predict the sales of products in development using the novelty value of the product, its relevance to the market (both on a scale of 0-100 with large values corresponding to more novel and more relevant products) and the advertisement costs in 1000$. However, under serious time pressure while preparing a report summarizing the results, the analyst forgets to copy-paste all relevant information to the report. The resulting table looks like this:

Parameters	Std. Error	t value	Pr(>|t|)
(Intercept)	37.7015	0.798	0.432
Novelty	0.3469	5.139	2.33e-05***
Relevance	0.3997	21.646	< 2e-16***
Advertisements	0.3782	16.277	3.75e-15***

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

Residual standard error: 53.08 on 26 degrees of freedom
Multiple R-squared:  0.9648, Adjusted R-squared:  0.9607
F-statistic: 237.5 on 3 and 26 DF,  p-value: < 2.2e-16
```

   Fed up with the work conditions, the analyst resigns soon after preparing the report and starts a trip around the world without leaving any contact information. However, the management immediately needs to decide the advertisement strategy for three new products, product ‘A’ with Novelty = 90, Relevance = 20, product ‘B’ with Novelty = 30, Relevance = 40 and product ‘C’ with Novelty = 70, Relevance = 80.
   a) Does it make sense to use the multiple linear regression model to decide the advertisement budget?
   b) Can we specify the estimates and approximate confidence intervals for each predictor?
   c) At what advertisement budget could we expect to sell 1000 units of each product?

   Answer as many questions as you can, based on the presented information!

   Hint: think about the way that t-values are computed from coefficient estimates and standard errors. The corresponding formula will allow you to reconstruct the estimated coefficients from the t-value and standard error columns. Once you have those available, this exercise will be similar to the exercises in the previous class.

2. As in the lecture, model how the cost of transporting goods from Sweden depends on distance by linear regression and consider also a destination variable with four possible levels: EU, US, China, Other.

   a) Fill the value taken by the three dummy variables representing the destinations with EU as baseline for some specific destinations given in the table.

Destination	X_{US}	X_{China}	X_{Other}
Cape Town			
Stockholm			
Sydney			
Shanghai			
New York			
London			

b) What does it mean if $X_{\text{Other}} = 0$?

c) Without computing the coefficients or software use, what kind of results do you expect to get? Which coefficients will be positive and which ones will be negative? Which coefficient do you expect to be largest and the smallest?

d) How do your answers to part c) change if we re-define the model using US as baseline and defining dummy variables X_{China}, X_{Other} and X_{EU} ?

e) Is the classification of destinations with the four levels specified above the best way to model the dependence? If not, how could it be refined?

3. The R outputs after the exercise descriptions (see next page) belong to an analysis of all possible combinations of predictors being used for explaining wages based on education, experience and sex, based on data from the United States from 1976 to 1982 (from the R library called Ecdat). The format of the variables is as follows:
   - exper: experience in years;
   - sex: a factor with levels (male,female);
   - school: years of schooling;
   - wage: wage (in 1980 \$) per hour.

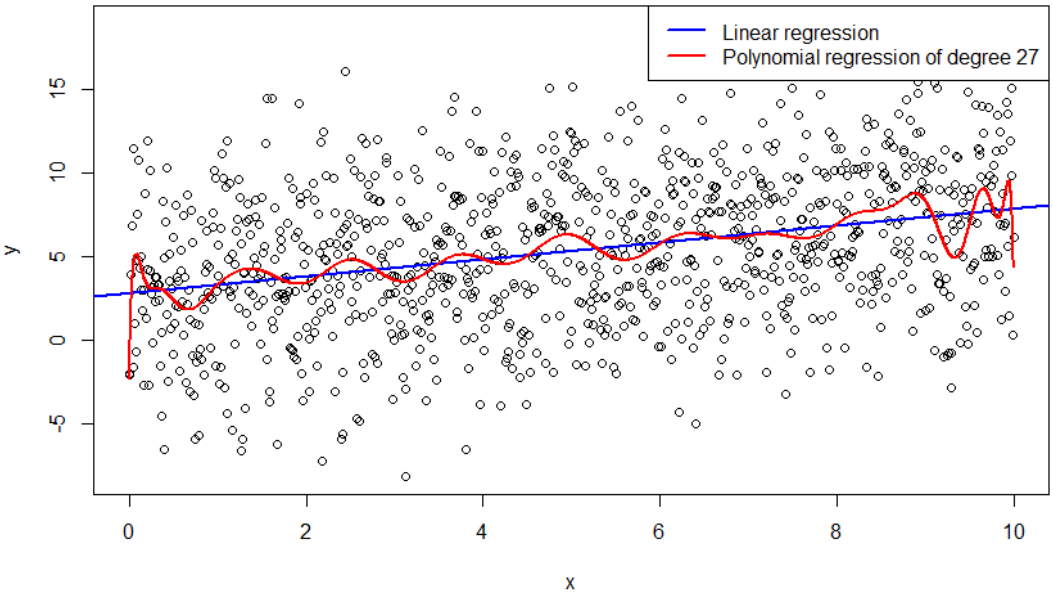
   We would like to understand the most relevant way of modelling. Therefore, perform the following data selection procedures based on the R outputs:
   a) Backward selection;
   b) Forward selection;
   c) Mixed selection;
   d) Best subset selection.

   Are there any differences between the final models chosen by these algorithms? Why?

4. In the solution document provided to the exercises in Exercise class 1 (i.e. Exercise class 1 - exercises with solutions.docx), confidence intervals and prediction intervals are provided for Exercise 8 in ISL (page 121), part a) - iv.
   a) Check which of the intervals is wider;
   b) Check whether the prediction point estimate is in the midpoint of both intervals;
   c) Describe a correct interpretation of the confidence interval;
   d) Describe a correct interpretation of the prediction interval.

5. An analyst got a dataset containing about 1000 values of response y and corresponding values of explanatory variable x. The analyst decided to try both simple linear regression and polynomial regression including higher degrees of x to capture potential non-linear effects. The two resulting models are displayed overlaid on the scatterplot of x and y, see below.
   a) In your opinion, which model is better?
   b) Which of the two models would give a better fit with the observations?
   c) Which of the two models would give better predictions for new data?

d) Can any specific problems with one of the models be pointed out?



6. Feedback quiz (optional): Go to www.menti.com and use the code 14 71 24.

R outputs to use for Exercise 3:

Null model:

```
Call:
lm(formula = wage ~ 1)

Residuals:
   Min     1Q Median     3Q    Max
-5.681 -2.136 -0.552  1.547 34.051

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.75759    0.05696   101.1   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.269 on 3293 degrees of freedom

AIC = 17154.72
```

One-predictor models:

```
Call:
lm(formula = wage ~ exper)

Residuals:
   Min     1Q Median     3Q    Max
-6.110 -2.153 -0.560  1.479 34.275

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.16777    0.20775  24.875  < 2e-16 ***
exper        0.07333    0.02484   2.952  0.00318 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.265 on 3292 degrees of freedom
Multiple R-squared:  0.00264,   Adjusted R-squared:  0.002337
F-statistic: 8.714 on 1 and 3292 DF,  p-value: 0.003181
```

AIC = 17148.02

```
Call:
lm(formula = wage ~ sex)

Residuals:
   Min     1Q Median     3Q    Max
-6.160 -2.102 -0.554  1.487 33.496

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.14692    0.08122   63.37   <2e-16 ***
sexmale      1.16610    0.11224   10.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.217 on 3292 degrees of freedom
Multiple R-squared:  0.03175,   Adjusted R-squared:  0.03145
F-statistic: 107.9 on 1 and 3292 DF,  p-value: < 2.2e-16
```

AIC = 17050.46

```
Call:
lm(formula = wage ~ school)

Residuals:
   Min     1Q Median     3Q    Max
-6.744 -2.024 -0.482  1.443 34.403

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.72251    0.38739  -1.865   0.0623 .
school       0.55716    0.03298  16.896   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.137 on 3292 degrees of freedom
Multiple R-squared:  0.0798,    Adjusted R-squared:  0.07952
F-statistic: 285.5 on 1 and 3292 DF,  p-value: < 2.2e-16
```

AIC = 16882.77

Two-variable models:

```
Call:
lm(formula = wage ~ exper + sex)

Residuals:
   Min     1Q Median     3Q    Max
-6.397 -2.113 -0.550  1.462 33.633

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.82930    0.20737  23.288   <2e-16 ***
exper        0.04108    0.02468   1.665   0.0961 .
sexmale      1.14169    0.11317  10.089   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.216 on 3291 degrees of freedom
Multiple R-squared:  0.03256,   Adjusted R-squared:  0.03197
F-statistic: 55.38 on 2 and 3291 DF,  p-value: < 2.2e-16
```

AIC = 17049.68

```
Call:
lm(formula = wage ~ exper + school)

Residuals:
   Min     1Q Median     3Q    Max
-6.879 -1.989 -0.518  1.393 34.908

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.47668    0.47000  -5.270 1.46e-07 ***
exper        0.15726    0.02417   6.507 8.86e-11 ***
school       0.59923    0.03340  17.940  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.117 on 3291 degrees of freedom
Multiple R-squared:  0.09149,   Adjusted R-squared:  0.09094
F-statistic: 165.7 on 2 and 3291 DF,  p-value: < 2.2e-16
```

AIC = 16842.67

```
Call:
lm(formula = wage ~ sex + school)

Residuals:
   Min     1Q Median     3Q    Max
-7.584 -1.970 -0.423  1.465 33.765

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.04584    0.39105  -5.232 1.79e-07 ***
sexmale      1.40621    0.10746  13.086  < 2e-16 ***
school       0.60763    0.03238  18.763  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.058 on 3291 degrees of freedom
Multiple R-squared:  0.1253,    Adjusted R-squared:  0.1248
F-statistic: 235.7 on 2 and 3291 DF,  p-value: < 2.2e-16
```

AIC = 16717.69

Three-variable model

```
Call:
lm(formula = wage ~ exper + sex + school)

Residuals:
   Min     1Q Median     3Q    Max
-7.654 -1.967 -0.457  1.444 34.194

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.38002    0.46498  -7.269 4.50e-13 ***
exper        0.12483    0.02376   5.253 1.59e-07 ***
sexmale      1.34437    0.10768  12.485  < 2e-16 ***
school       0.63880    0.03280  19.478  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.046 on 3290 degrees of freedom
Multiple R-squared:  0.1326,    Adjusted R-squared:  0.1318
F-statistic: 167.6 on 3 and 3290 DF,  p-value: < 2.2e-16
```

AIC = 16692.18