Exercise solutions 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?
      Yes, the very low p-value for the F-test for model significance (see the last row of the R output) and the very high value for R-squared suggest that this model may be relevant for making predictions and can explain almost all (96%) variability in product sales.

   b) Can we specify the estimates and approximate confidence intervals for each predictor?
      Considering that the t-values are computed by dividing the estimated coefficients by their respective standard error, basic algebra helps to get back the estimates by multiplying the t value and Std. Error column entries. For example, for the intercept:

$$\frac{\hat{\beta}_0}{\mathrm{SE}(\hat{\beta}_0)} = \text{t value, i.e. } \frac{\hat{\beta}_0}{37.7015} = 0.798;$$

      therefore, we have that

$$\hat{\beta}_0 = 0.798 * 37.7015 = 30.086.$$

      Similarly, the estimate for the coefficient of Novelty is 0.3469*5.139=1.783, the coefficient of Relevance is 0.3997*21.646=8.652, the coefficient of Advertisements is 0.3782*16.277=6.156.

The coefficient estimates and standard errors can be used to get the confidence intervals, and the simple formula with "2" instead of the 97.5% quantile of the t distribution can be used for sample sizes of approximately at least 30. In this case, the sample size is exactly 30, as we can see from the 26 degrees of freedom – recall that df = n-p-1 and here we have p=3 predictors, so 26 = n-3-1 gives that n=30. Therefore, the 95% confidence intervals are as follows:

$$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0) = 30.086 \pm 2 \cdot 37.7015, \text{ i.e. the interval } [-45.317, 105.489];$$
$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1) = 1.783 \pm 2 \cdot 0.3469, \text{ i.e. the interval } [1.089, 2.477];$$
$$\hat{\beta}_2 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_2) = 8.652 \pm 2 \cdot 0.3997, \text{ i.e. the interval } [7.853, 9.451];$$
$$\hat{\beta}_3 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_3) = 6.156 \pm 2 \cdot 0.3782, \text{ i.e. the interval } [5.400, 6.912].$$

c) At what advertisement budget could we expect to sell 1000 units of each product?

As we have computed the coefficient estimates in part b), we have the following equation for predicting sales values:

$$\widehat{\mathrm{Sales}} = 30.086 + 1.783 \times \mathrm{Novelty} + 8.652 \times \mathrm{Relevance} + 6.156 \times \mathrm{Advertisements}.$$

In part c), we know that we are aiming at a prediction of 1000 sold units, so we can write 1000 on the left side of the equation. Considering product A, the novelty and relevance values are given to us as Novelty = 90, Relevance = 20, so we can plug these numbers in the right side of the equation. Therefore, for product A, the prediction equation looks like this:

$$1000 = 30.086 + 1.783 \times 90 + 8.652 \times 20 + 6.156 \times \mathrm{Advertisements}.$$

In this equation, Advertisements is the only unknown, so we can solve it with basic algebra:

$$\mathrm{Advertisements} = \frac{1000 - 30.086 - 1.783 \times 90 - 8.652 \times 20}{6.156} = 103.385.$$

This gives the required Advertisements value for product A, in 1000 dollars.

For products B and C, we need to do exactly the same computation, except that we plug in different values for Novelty and Relevance. Therefore, the necessary advertisement budgets for those products, in 1000 dollars, are:

$$\mathrm{Advertisements} = \frac{1000 - 30.086 - 1.783 \times 30 - 8.652 \times 40}{6.156} = 92.651;$$
$$\mathrm{Advertisements} = \frac{1000 - 30.086 - 1.783 \times 70 - 8.652 \times 80}{6.156} = 24.849.$$

This helps us to see that product C requires the smallest financial investment while product A requires the largest financial investment to expect 1000 sold units.

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	0	0	1
Stockholm	0	0	0
Sydney	0	0	1
Shanghai	0	1	0
New York	1	0	0
London	0	0	1

Note that destinations in the EU have all-0 rows for the relevant dummy variables, because EU was chosen as the baseline level. Note also that since the end of last week, London is no longer in the EU, hence it has become an "Other" destination.

b) What does it mean if $X_{\text{Other}} = 0$?
It means that the destination is not in the Other category. Since we consider four possible levels of this variable, EU, US, China and Other, not being in Other means that the destination is in the EU, US or China.

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?
The coefficient of a dummy variable corresponding to a destination – like US – means the extra cost of transporting to the US compared to the baseline level, EU, assuming the same distance. Considering that we have advantageous trade rules within the EU, it is reasonable to assume that transporting to either US, China or other non-EU destinations would entail extra costs compared to transporting within the EU, so it is reasonable to expect that all coefficients will be positive.
Trying to guess which coefficient would be largest and smallest is way more difficult, because it depends on the rules and regulations of trade (e.g. taxes) between EU and China, respectively EU and US. Additionally, the "Other" category includes everything from Australia to Norway, so it is very difficult to know what to expect for that category and also to interpret the result.

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} ?
In that case, all comparisons are made against the baseline level of transporting to US. Recall that we expected that transporting to US would be more expensive than transporting to the EU. Therefore, compared to the US baseline level, EU transports should be cheaper, hence the coefficient of EU should be negative. To predict the sign or size of the other coefficients, more detailed knowledge would be needed about trade rules to China versus US.

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?
The suggested classification is probably too crude. For example, it would make sense to separate at least Norway, the UK and Switzerland from the Other category, because there are special rules in place with these countries.

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;

    Here we start with the all-variable model:

```
             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 ***
```

    Since all variables in this model have very small p-values, we stop here and do not remove any variable. We conclude that the model selected by backward selection is the three-variable model, with coefficients as specified in the output above.

    b) Forward selection;

    We start with the null model (step 0) and consider each one-variable model. We conclude that the model with school gives the highest R^2 value and school is a significant predictor in this model. Therefore, the model after the first update is this:

```
             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
```

    Now we consider all two-variable models in which one of the predictors is school. There are two such models, school+exper and school+sex. The one with school+sex gives a highest R^2 value and sex is a significant predictor in this model. Therefore, the model after the second update is this:

```
             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
```

    Finally, we consider all three-variable models that include school and sex as predictors. There is only one such model, namely the three-variable model in which exper is added to school and sex. In this model, exper is highly significant, so we add

it. There are no more variables to consider, so we stop and conclude that our final model using forward selection is this one:

```
             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 ***
```

In other words, the final models from backward and forward variable selection agree for this dataset.

c) Mixed selection;
In this case, we follow the same path as the forward algorithm, but in each step, we do not only check whether the newly added variable is significant but also whether other variables have become non-significant as a result of the addition. However, we see that in each model after each update, all variables are highly significant. Therefore, the mixed selection algorithm uses exactly those steps and results in exactly the same final model as the forward selection algorithm in part b).

d) Best subset selection.
In this case, AIC is given for each model – we can simply check all AIC values and choose the model with the lowest AIC (because AIC is a measure that gives low values for good models). Looking at all outputs, we see that the AIC of AIC = 16692.18 in the three-variable model is lower than any other AIC value. Therefore, even here, the best model is the one with three variables, exactly as in parts a)-c).

Are there any differences between the final models chosen by these algorithms? Why?
There are no differences. Looking at the three-variable model, we see that each variable has very low p-values, i.e. each variable has a significant relationship with the response even in the presence of all other variables. In this sense, it is not surprising that all algorithms preferred that model.

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.
The R code and output are 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
```

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

a) Check which of the intervals is wider;

The prediction interval is much wider, going from approximately 15 to 34, while the confidence interval is just the interval from 24 to 25.

b) Check whether the prediction point estimate is in the midpoint of both intervals;
The prediction point estimate is listed under "fit". One can check that indeed, it equals (14.8094+34.12476)/2 as well as (23.97308+24.96108)/2

c) Describe a correct interpretation of the confidence interval;
The confidence interval is a range that is very likely to contain the miles per gallon value for *an average car* with horsepower=98.

d) Describe a correct interpretation of the prediction interval.
The prediction interval is a range that is very likely to contain the miles per gallon value for *a specific car* with horsepower=98 (which car may or may not be close to an average car with this horsepower value – hence, it is much harder to predict its corresponding mpg range than for the average).

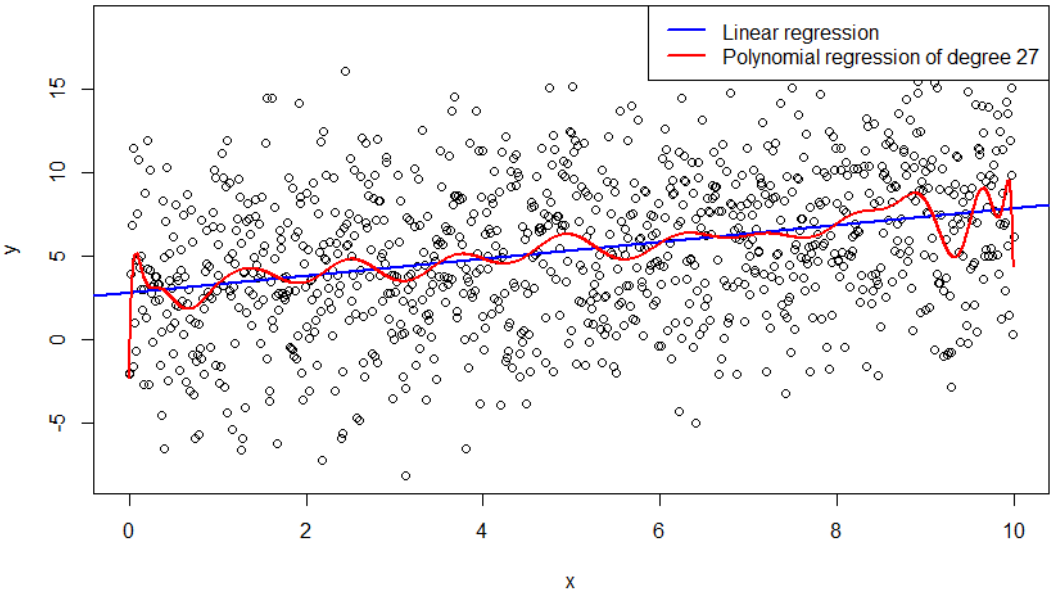
5. An analyst got a dataset containing about 1000 values of response y and corresponding values of explanatory variable x. The analyst decided to try both simple linear regression and polynomial regression including higher degrees of x to capture potential non-linear effects. The two resulting models are displayed overlaid on the scatterplot of x and y, see below.

a) In your opinion, which model is better?
Here, the question is a bit vague and can be understood in different ways. A very good answer pointed out that the linear regression model is easier to interpret, which is a big advantage. Another very good answer gave a correct assessment of parts b) and c) (see below) and concluded that the model that would give better predictions for new data should be preferred.

b) Which of the two models would give a better fit with the observations?
It is the polynomial regression of degree 27 (i.e. the red curve). That contains all polynomials up to degree 27 including the linear model (which has degree 1) and chooses the model that is closest to the observations. Therefore, this model gives at least as good fit on the observations as the linear model.

c) Which of the two models would give better predictions for new data?
It seems unlikely that new data would require all the wiggles that are produced by the red curve. It seems more like a case of overfitting, where the model tried to follow the random error in the training set, and we would expect that it would not do very well on a test set, i.e. on new data. Therefore, we expect that the linear model, i.e. the blue curve would give better predictions for new data.

d) Can any specific problems with one of the models be pointed out?
The red curve looks especially problematic (unnecessarily curvy) around the ends of the x-range, i.e. between 0 and 0.5, respectively between 9 and 10. This is often the case with too complex models, especially higher degree polynomials, that their behavior is bad around the ends of the prediction range.



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