**Exercises for exercise class 7b in MMS075, Mar 4, 2020**

A couple of exercises are collected here that provide an idea about what kind of exam questions you can expect. Note: this does not mean that the other exercises in the exercise classes cannot be included in the exam, but rather that questions like those below are likely to be included.

Under the text of each exercise, it is shown which exercise class it was taken from and the number of the exercise; for example, the first one is exercise 2 from exercise class 1, hence it is denoted as (Ex cl 1, Ex 2). This allows you to check the solution on Canvas.

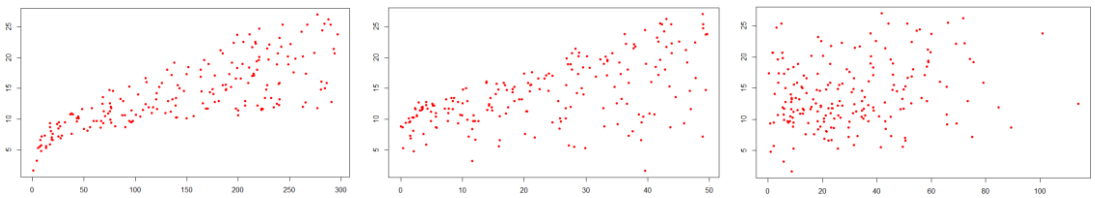
The current solution document adds a solution for exercise 6 only, which is a new exercise. The solutions to all other exercises can be found in the documents on Canvas – for example, the solution to the first exercise, marked as Ex cl 1, Ex 2, can be found in 'Exercise class 1 – exercises with solutions.docx'.

Furthermore, it is also indicated how many points would be assigned to the exercises and their distribution.

With these exercises and assigned points, the following grading limits would be used: pass (grade 3): 40% of total points; grade 4: 60%, grade 5: 80%.

No books or notes are allowed, but a formula sheet will be provided. You may use a Chalmers approved calculator.

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? Explain your answer to both questions!
(Ex cl 1, Ex 2, 2 points: one for correct answer, one for correct explanation)

2. Consider the multiple linear regression model with sales (in 1000 units) as response and the usual 3 predictor variables of TV advertisements, radio advertisements and newspaper advertisements as predictors. The R output of this model is given below.

```
Call:
lm(formula = AdData$sales ~ AdData$TV + AdData$radio + AdData$newspaper)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.938889   0.311908   9.422   <2e-16 ***
AdData$TV       0.045765   0.001395  32.809   <2e-16 ***
AdData$radio    0.188530   0.008611  21.893   <2e-16 ***
AdData$newspaper -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

a) Is there a relationship between at least one type of advertisement and sales? Where do you see this in the output above?
b) Formulate the interpretation of the following quantities:
   - 0.86 in the row of AdData$newspaper
   - <2e-16 in the row of AdData$radio
   - <2e-16 in the row of (Intercept)
c) Specify an approximate confidence interval for each coefficient.
d) The management decides to spend 115000$ on TV advertisements and 40000$ on radio advertisements. How many sold units does the multiple linear regression model predict for these values?

(Ex cl 2, Ex 3, 7 points: 1 for a), 2 each for b)-d))

3. The R outputs after the exercise descriptions 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.

   Briefly explain each step in the variable selection process!

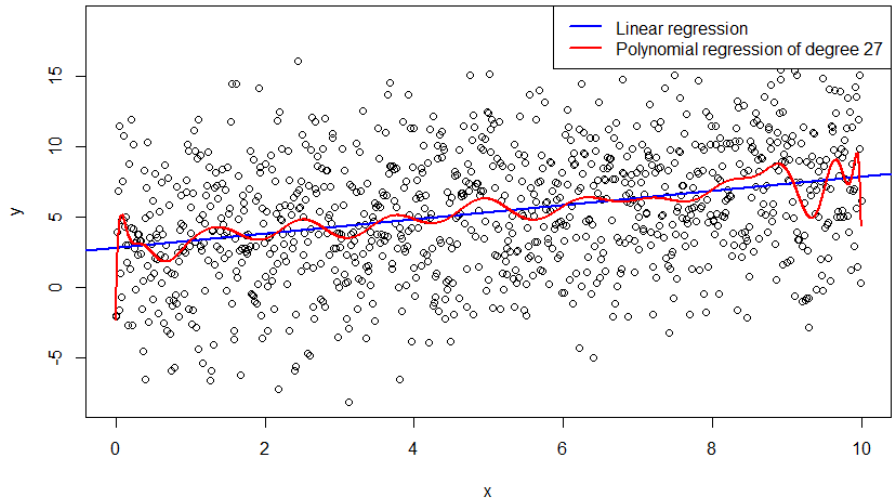
   (Ex cl 3, Ex 3, 6 points: 3 each for a)-b), of which 1 for correctly performing the algorithm and 2 for correct explanation of what you are doing)

4. 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.

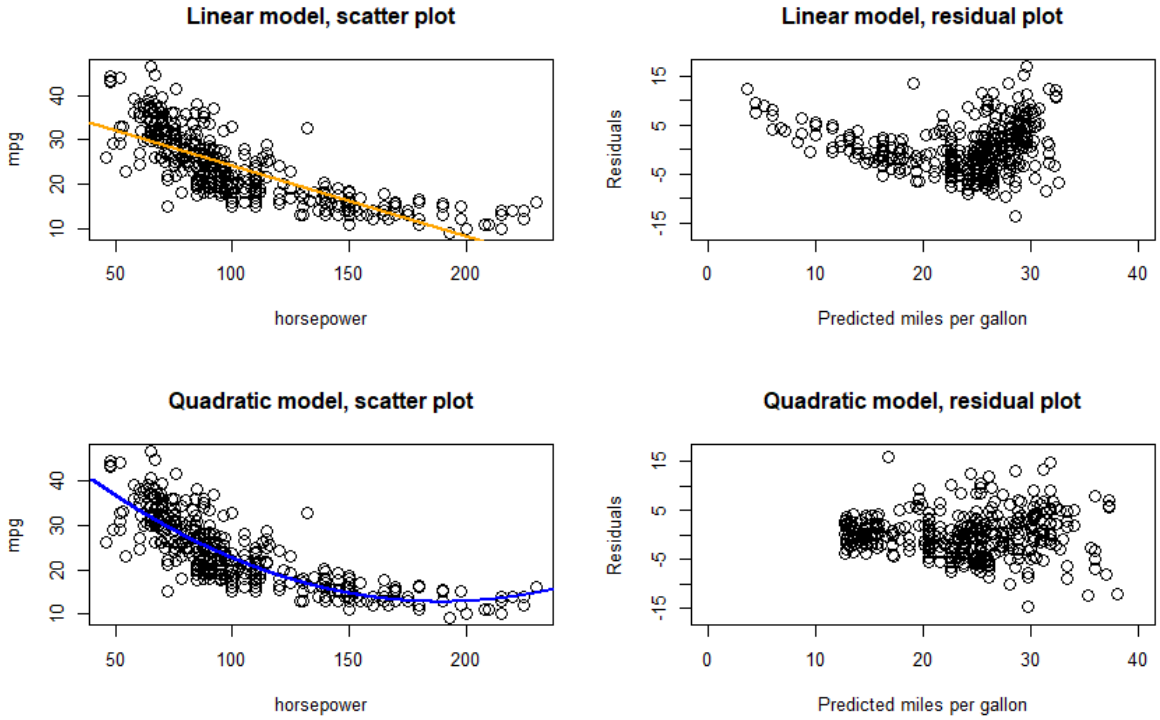
a) Which of the two models would give a better fit with the observations?
b) Which of the two models would give better predictions for new data?

Explain your answers to both questions!

(Ex cl 3, Ex 5, 4 points: 2 each for a)-b), of which 1 for correct answer, 1 for correct explanation)




5. Check the mpg vs horsepower graphs presented in the lecture:




Based on these graphs, address the following points:

a) Why are there several points in the residual plot of the linear model with x-coordinates between 0 and 10 and no such points at all in the residual plot of the quadratic model?

b) Mark the corresponding points on the scatter plots!

(Ex cl 4, Ex 3, 3 points: 2 for a), 1 for b))

6. Which of the following combinations is most likely to result in an influential point:
   a. Outlier and high leverage point
   b. Not outlier and not high leverage point
   c. Not outlier and high leverage point
   d. Outlier and not high leverage point?

   Explain your answer!

   (This is a new exercise, 3 points, 1 for correct answer and 2 for correct explanation.)

   The correct answer is a) – it is discussed in the lectures that outliers that also are high leverage points are often influential. You can also think about the residuals vs leverage plot in R – this plot shows areas marked by red dashed lines in the lower-right and upper-right corners, and a point is influential if it is in one of these marked corner areas. To be there, it needs to have a high leverage value (to be on the right end for the x-axis) and either be close to the top or close to the bottom of the graph, and that's exactly where outliers could be found (by having standardized/studentized residual values >3 or <-3).

7. Recall that the prediction model for Advertising example was as follows:

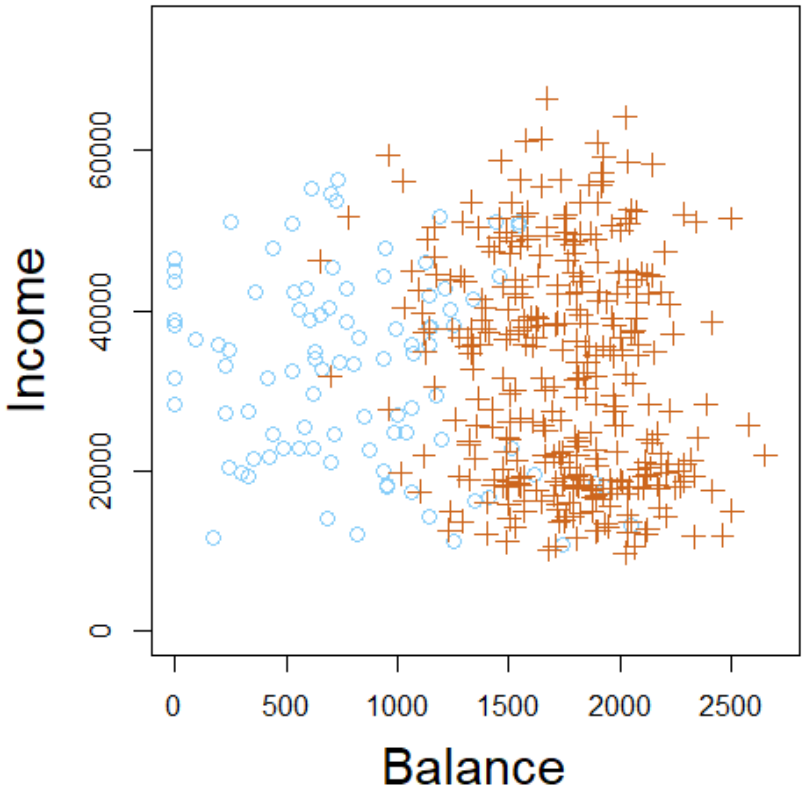
$$\widehat{\text{sales}} = 6.7502 + 0.0191 \times \text{TV} + 0.0289 \times \text{radio} + 0.0011 \times \text{TV} \times \text{radio}$$

   Predict the number of sold units using this model for the following advertisement budget distributions:
   a) TV budget: \$0, radio budget: \$100 000;
   b) TV budget: \$50 000, radio budget: \$50 000;

   Furthermore, estimate the effect of:
   c) \$1000 increase of TV advertisement on sold units if the radio budget is \$10 000.

   (Ex cl 4, Ex 2, 4 points: 1 each for a)-b), 2 for c))

8. The scatter plot below corresponds to the removal of data for 99% of non-defaulted individuals. The blue points represent the non-defaulted cases and the brown crosses represent the defaulted individuals. What predictions do you expect for the default probability at balance = \$500, balance = \$1000, balance = \$1500 and balance = \$2000, based on a logistic regression model that uses balance as a single predictor to predict default? Draw a function that you expect to be close to the estimated default probability curve!




(Ex cl 6a, Ex 2, 4 points: 3 for providing reasonable estimates with reasonable explanations for the given balance values, 1 for drawing a curve that has approximately correct shape)

9. The logistic regression model using that corresponds to the above scatter plot has the following R output:

```
Call:
glm(formula = default ~ balance, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2277   0.0489   0.1607   0.3726   2.2302

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.9439716  0.7432864  -7.997 1.28e-15 ***
balance      0.0054321  0.0005759   9.433  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 456.15  on 428  degrees of freedom
Residual deviance: 196.85  on 427  degrees of freedom
AIC: 200.85

Number of Fisher Scoring iterations: 6
```

Recall that in logistic regression, the estimated probability of a “case” for given values of the predictors can be computed as follows:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2+\cdots+\hat{\beta}_p X_p}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2+\cdots+\hat{\beta}_p X_p}}$$

Plug the coefficient estimates from the above R output into this formula and write down the resulting equation. Using this equation, make a prediction of the probability of default at the following balance values:

a) balance = $1000
b) balance = $2000.

(Ex cl 6a, Ex 3, 4 points: 2 for a) and b) each of which 1 is for using the correct formula and 1 is for getting the correct result)

10. An air conditioning company collects information about factors that affect the probability of installing central air conditioning. As a first step, they analyze the **Housing** dataset in R and consider the following variables as predictors: **price**, denoting the sale price of a house ($), **gashw** indicating whether the house uses gas for hot water heating (yes/no), and **stories** indicating the number of stories excluding basement (numerical value). Their analysis in R returns the following summary output:

```
Call:
glm(formula = airco ~ gashw + stories + price, family = "binomial",
    data = Housing)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7904  -0.7341  -0.5198   0.6765   3.1124

Coefficients:
                Estimate   Std. Error z value Pr(>|z|)
(Intercept) -4.275690040  0.387730993 -11.027  < 2e-16 ***
gashwyes    -3.689570472  1.086346957  -3.396 0.000683 ***
stories      0.294471594  0.130692028   2.253 0.024248 *
price        0.000043195  0.000005342   8.085  6.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 681.92  on 545  degrees of freedom
Residual deviance: 532.87  on 542  degrees of freedom
AIC: 540.87

Number of Fisher Scoring iterations: 6
```

**Using this output, do the following tasks:**

a) Specify the equations predicting the probability that a house with 2 stories has air conditioning, depending on its price and whether it uses gas for hot water heating!
b) Estimate the probability that a house with a sale price of $90000 and 2 stories that does not use gas for hot water heating has central air conditioning!

In addressing parts a) and b), remember that the equation for estimating the probability of a "case" in logistic regression was given in the previous exercise.

(Ex cl 6b, Ex 2, 4 points: 2 each for a)-b))

11. The **Mode** dataset in the **Ecdat** library in R contains data about travel modes: the estimated cost and time of car, carpool, bus or rail for different trips and the decision of which travel

mode is chosen for the trip. We want to understand how the decision depends on the various parameters, in particular, what influences people to choose the different alternatives instead of driving a car. Therefore, we fit a multinomial logistic regression model with "car" as reference level, see the R commands and the output below.

```
> Mode$TravelMode=relevel(Mode$choice,ref="car")
> multinom(TravelMode~.-choice,data=Mode)
# weights:  40 (27 variable)
initial  value 627.991346
iter  10 value 394.826983
iter  20 value 360.160469
iter  30 value 342.077240
final  value 342.073501
converged
Call:
multinom(formula = TravelMode ~ . - choice, data = Mode)

Coefficients:
        (Intercept)  cost.car cost.carpool   cost.bus  cost.rail
carpool   -4.106305 0.6360771   -0.4472983 0.04505779 -0.5501033
bus       -4.788587 0.8461170    0.2162670 0.01013198 -0.5276687
rail      -4.299980 0.8900743    0.2058304 0.56600590 -1.2752642

  time.car time.carpool     time.bus    time.rail
0.12366772  -0.06922734  0.007442305 -0.027732676
0.02285042   0.09461637 -0.107750616 -0.006393055
0.03639195   0.07297203 -0.018132194 -0.075282632
```

Based on this output, how are the following changes expected to affect the probability of choosing various alternatives:

a) Increasing the cost of car trips;
b) Decreasing the time of car trips;
c) Increasing the price of train tickets;

(Ex cl 6b, Ex 5, 5 points: 1 for a), 2 each for b)-c))

12. Consider a simple model A and a very flexible model B. Evaluate whether the following inequalities always hold:
    a. Training MSE for model B ≤ Training MSE for model A;
    b. Test MSE for model B ≤ Test MSE for model A.

    (Ex cl 7a, Ex 3, 4 points: 2 each for a)-b))

13. Feedback quiz (optional): Go to www.menti.com and use the code 36 48 46.

## 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