STAT202 Assignment 3: Multiple linear regression II

Due on 18 August 12 noon

This assignment deals with the analysis of the kc_house_data.csv[1] data set, which contains data for over 26 thousand house sales in King County WA, USA. The response variable is house **price** (in USD), for which we have twenty predictors: a sale ID (**id**), date of sale (**date**), number of bedrooms (**bedrooms**), number of bathrooms (**bathrooms**), living area in foot^2 (**sqft_living**), lot area in foot^2 (**sqft_lot**), number of floors (**floors**), is in waterfront (**waterfront**, 0/1), quality of view (**view**, 0-4), condition score (**condition**), quality of building standards (**grade**), built area above ground (**sqft_above**), basement area (**sqft_basement**), year of construction (**yr_built**), year renovated (**yr_renovated**, 0 for no renovated), zipcode (**zipcode**), latitude (**lat**), longitude (**long**), and two other functions of size (**sqft_living15**) and (**sqft_lot15**).

1. As always, create a new folder for your work and also a new RStudio project (using the 'existing folder' option) in that folder. Download the kc_house_data.csv file in your folder

2. You will use a subset of the data, which depends on your student ID. Load the tidyverse functions, read the kc_house_data.csv dataset and choose a subset of 20,000 houses. Assuming that you name the original dataset houses, name your sample `my_houses` and then drop the following variables from the dataset: id, date, lat, long, zipcode, sqft_above, sqft_living15 and sqft_lot15. For the last part we use the select function (if we use a negative sign the variable is dropped/deleted):

```
my_houses <- my_houses %>%
  select(-id, -date, -lat, -long, -zipcode, sqft_above,
         -sqft_living15, -sqft_lot15)
```

3. Estimate the correlations between all variables (using `my_houses %>% cor()`) and, based on the correlations, choose three variables to predict price. Using `ggpairs` (from the `GGally` package[2]) create a scatterplot matrix with your three predictors and price. Explain in 50 words the relationships you observe in that plot.

4. Fit that model (call it `m1`) and write down the coefficients, the adjusted-r2 and residual standard error.

5. Now use the leaps package to fit all regression subsets. Plot the results of regression subsets, and explain which predictors are contained in the best model.

```
library(leaps)
all_mods <- regsubsets(model with all predictors, data = my_houses)
plot(all_mods, scale = 'Cp')
```

[1] This dataset is available in Kaggle https://www.kaggle.com/harlfoxem/housesalesprediction

[2] We showed the syntax in lab 2 and also in lecture 9.

6. Now fit the best model from the previous step (call it `m2`) and compare its adjusted-r2 and residual standard error with `m1`. Discuss in 50 words the similarities and differences between the results of the 2 models.

7. Create diagnostic plots for the residuals of the full model; use `plot(m2)`. Check the model for assumptions for the residuals. Explain in no more than 70 words if there is anything unusual or wrong.

8. Obtain the predicted quality (95% prediction and 95% confidence intervals) for new houses with the following characteristics:

```
bedrooms bathrooms sqft_living sqft_lot floors waterfront  view
    3       1.5         1200    15606      1           0      0
    4       2.5         1920     8562      2           0      0

condition grade sqft_basement yr_built yr_renovated
     3     7              0     1985             0
     4     7              0     1994             0
```

You will need to create a tibble and use the predict() function for this. Have a look at slides 5 and 6 in lecture 10 to get an idea of how it works.

9. Create a Word file with your graphs, code and answers and submit it in Learn.