ST221 Assessed coursework 2

Deadline: 02 May 2023, 1 pm

Please read these instructions carefully!

This assignment counts for **20%** of your final module mark. The maximum score for this coursework is 50 marks.

Your solutions must be produced using a word processor, R Markdown, LaTeX or similar and must be converted to a pdf file before submission. Please do not add your name on your submission to allow for anonymous marking. You may cut-and-paste R-output. Handwritten answers will not be awarded marks. Use a font size of 11pt or larger. Question sub-sections must be clearly labelled for ease of marking.

If you do not submit your solutions in a typed format, then this will not be accepted as a submission. You should convert your solutions into **one PDF file** to be submitted on the ST221 moodle page.

Please read Chapter 5 in the course guide which gives details around the procedures regarding coursework including applying for extensions and lateness penalties. Please ensure that you submit in good time before the deadline. Penalties will apply if work is submitted more than 1 minute after the deadline unless an extension or waiver is granted. Coursework is not eligible for mitigating circumstances due to the loss of work in progress. The penalty for late submission is 5% per 24 hour period encompassing a working day. However no submission will be accepted more than 5 working days after the original deadline unless there is a pre-approved extension extending past the cut off period.

If you have any queries about the coursework, please post them on the ST221 forum, but do not post any part of your solutions. You can also submit questions to the anonymous question form on moodle.

Please be aware that your work will be submitted to TurnItIn, a piece of plagiarism-detection software. Cases of suspected collusion or plagiarism will be followed up as outlined in Section 5.3 of the course guide. Note that detailed discussions of the assignment or comparisons of numerical/graphical results or computer code are **not permitted.**

Make sure to read questions carefully. If asked to produce a plot, then please include the plot in your report. Make sure it is of appropriate scale and the axes are clearly labelled. Include R code only if requested to do so.

Good luck with the assignment!

Download the file `houses.csv` from the ST221 moodle page and load it into `R`.[1] The original dataset consists of information about the value and various characteristics of a sample of houses in Saratoga County, USA, from 2006.

The variables (adapted from the original dataset) are:

- `price`: the value of the property (in 1000 US dollars);
- `livingArea`: the living area of the property (in square feet);
- `bathrooms`: the number of bathrooms;
  (a bathroom that has no shower or bathtub is counted as a half bathroom);
- `bedrooms`: the number of bedrooms;
- `fuel`: the type of fuel used for heating (gas, electric or oil).

We aim to develop a normal linear model that predicts the value of the property from its characteristics.

Question 1 - Pairwise relationships

[Total 20 marks]

(a) **[2 marks]** Create a new variable `log2livingArea` by transforming `livingArea` using a logarithm of base 2. Fit a simple linear regression of `price` on `log2livingArea`. Report the `R` model summary.

(b) **[2 marks]** Give a quantitative interpretation of the estimated coefficient for `log2livingArea`.

(c) **[3 marks]** Briefly explain why it is not sensible to give an interpretation of the estimated intercept for the model in (a).

(d) **[3 marks]** Implement `fuel` as a factor variable. Produce a graphical illustration of the relationship between `price` and `fuel` and give a brief description of the salient features in the plot.

(e) **[1 mark]** In the linear model defined by the `R` model formula `price ~ 0 + factor(fuel)`, what do the estimated coefficients correspond to?

(f) **[3 marks]** The variable `bathrooms` is discrete taking only a small number of values. Fit the two alternative linear models specified below:

$$M_0 : \texttt{price} \sim \texttt{bathrooms} \qquad \text{and} \qquad M_1 : \texttt{price} \sim 0 + \texttt{factor}(\texttt{bathrooms})$$

Produce a plot of the house prices (in US Dollars) predicted by model M_0 against bathrooms. Using a different colour and plotting symbol add to the plot the predicted house prices (in US Dollars) produced by model M_1. Make sure to include a legend.

[1] The data is adapted from the Saratoga Houses which is, for example, available from DASL, the Data and Story library.

(g) [**6 marks**] Perform a test for non-linearity for the simple linear regression of `price` on `bathrooms` using a significance level of 5%. Include the relevant R output and clearly state

- the model under the null hypothesis and the model under the alternative hypothesis, (you may use an `R` formula notation to specify the models);
- the test statistic and its distribution under the null hypothesis;
- the observed value of the test statistic;
- the p-value of the test;
- the outcome of the test and
- the conclusion that can be drawn from the outcome of the test.

Question 2 - Multiple regression

[Total 20 marks]

(a) [**1 mark**] Consider the `R` summary output below for a linear model with response variable `price` and explanatory variables: `fuel`, `bathrooms` and `log2livingArea`. Which category of `fuel` has been used as reference category?

```
Call:
lm(formula = price ~ fuel + bathrooms + log2livingArea, data = houses)

Residuals:
    Min      1Q  Median      3Q     Max
-53.432 -11.862   0.149  12.104  58.212

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -307.701     20.688 -14.873  < 2e-16 ***
fuelelectric     -5.861      1.703  -3.442 0.000608 ***
fueloil          -6.269      1.952  -3.212 0.001375 **
bathrooms        14.512      1.497   9.695  < 2e-16 ***
log2livingArea   44.932      2.058  21.833  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.34 on 763 degrees of freedom
Multiple R-squared:  0.6421,    Adjusted R-squared:  0.6403
F-statistic: 342.3 on 4 and 763 DF,  p-value: < 2.2e-16
```

(b) [**2 marks**] Give a quantitative interpretation of the coefficient reported for `fueloil`.

(c) [**1 mark**] Suppose a house has a living area of 1600 square feet, 2 bathrooms and gas fueled heating. Using the model in Question 2 (a) predict the price of the house in US Dollars.

(d) [**3 marks**] Below is the R model summary of a linear model fitted to the same data and using the same explanatory variables as the model in Question 2 (a). Explain why it is not a contradiction that the coefficient for electric fuel is significant in Question 2 (a) [at a 5% significance level] but not significant in the R output below.

```
Call:
lm(formula = price ~ fuel + bathrooms + log2livingArea, data = houses)

Residuals:
    Min      1Q  Median      3Q     Max
-53.432 -11.862   0.149  12.104  58.212

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -313.9700    20.6811 -15.181  < 2e-16 ***
fuelgas           6.2695     1.9521   3.212  0.00137 **
fuelelectric      0.4089     2.2500   0.182  0.85586
bathrooms        14.5120     1.4969   9.695  < 2e-16 ***
log2livingArea   44.9317     2.0580  21.833  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.34 on 763 degrees of freedom
Multiple R-squared:  0.6421,    Adjusted R-squared:  0.6403
F-statistic: 342.3 on 4 and 763 DF,  p-value: < 2.2e-16
```

(e) [**4 marks**] For the model in Question 2 (a), identify the datapoint with the highest influence and report its Cook's distance. Is this point influential due to being a regression outlier, due to having a high leverage or due to both, having a high leverage and being a regression outlier? Support your answer with appropriate evidence.

(f) [**6 marks**] Below are the results of a **sequential** ANOVA for the model in Question 2 (a), that is for the model

$$LM_2 : \texttt{price} \sim \texttt{log2livingArea} + \texttt{bathrooms} + \texttt{fuel}$$

```
Analysis of Variance Table

Response: price
                Df Sum Sq Mean Sq   F value    Pr(>F)
log2livingArea   1 369280  369280 1228.7391 < 2.2e-16 ***
bathrooms        1  36829   36829  122.5461 < 2.2e-16 ***
fuel             2   5343    2671    8.8885 0.0001528 ***
Residuals      763 229309     301
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Consider a test for the existence of regression for the model

$$LM_1 : \texttt{price} \sim \texttt{log2livingArea} + \texttt{bathrooms}$$

(Note the difference between model LM_1 and model LM_2.) Explain how to use the information provided in the sequential ANOVA above to produce an ANOVA table for the test of the existence of regression for the model LM_1. Write out this ANOVA table.

(g) [**3 marks**] An estate agent suggests that, all else being equal, the increase in the average value of a house when doubling the living area is dependent on the type of fuel used by its heating system. Perform an appropriate hypothesis test at a 5% significance level to examine whether there is evidence for this in the data set given here. Present the `R` output of the test and state the conclusion that can be drawn from the outcome of the test.

Question 3 - Adding bedrooms as a predictor variable

[**Total 10 marks**]

(a) [**3 marks**] Another suggestion by the estate agent is to add bedrooms as an additional predictor variable to the model. Produce boxplots of `price` grouped by `bedrooms`. Briefly comment on the plot.

(b) [**3 marks**] Fit a linear model with response variable `price` and explanatory variables: `fuel`, `bathrooms`, `log2livingArea` and `bedrooms`, where `bedrooms` is implemented as a quantitative predictor, not a factor. Report the `R` model summary output. Assuming a 5% significance value, what can you conclude about the explanatory variable `bedrooms` from the model summary?

(c) [**4 marks**] Consider the model in Question 3 (b). The variance inflation factor of the jth explanatory variable X_j is defined as $VIF_j = \frac{1}{1-R_j^2}$ where R_j^2 is the coefficient of determination from a regression of X_j on the other predictors. By fitting the appropriate linear models compute the VIF for `bedrooms` and for `log2livingArea`. Include your `R` code in your answer and comment on the results. Hint: the coefficient of determination can be extracted from a fitted model using the command `summary(model)$r.squared`, where `model` is the `R` object containing the fitted model.