ST221 Assignment 2
Student Number: 2106983

# Question 1

a)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-66.253 -12.828  -0.071  13.796  63.267 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -428.885     19.398  -22.11   <2e-16 ***
log2livingArea   58.632      1.816   32.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.83 on 766 degrees of freedom
Multiple R-squared:  0.5763,	Adjusted R-squared:  0.5758 
F-statistic:  1042 on 1 and 766 DF,  p-value: < 2.2e-16
```

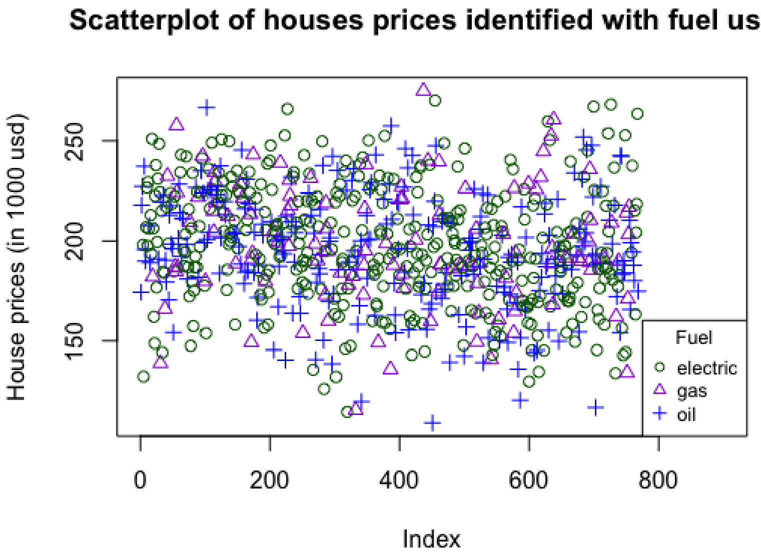
b)

As the slope for the model is 58.632 for each increase in square foot by a factor of 2, the average price of a house (in 1000 usd) is expected to increase by 58.632.

c)

The intercept is -428.885 which says for a house with a living area of 2 square feet the price is expected to be -428,885 usd which is clearly nonsense as you can’t have a house with a negative price.

d)

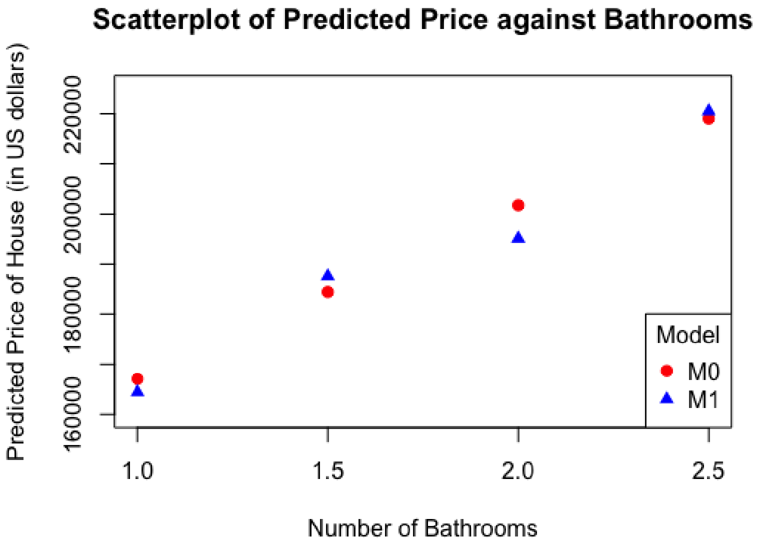


This scatterplot shows all the indexed house prices identified with fuel they use which is done by colour and symbol. There are too many data points to see a clear pattern between the 2 variables so more investigation would need to be done.

e)

The expected price of a house given which fuel it uses for heating.

f)



g)

The model under the null hypothesis is M_0 as defined in f) or lm(price ~ bathrooms) in R notation and the model under the alternative hypothesis is M_1 again as defined in f) or lm(price ~ 0 + factor(bathrooms)) in R notation.

The test statistic is $F = \frac{\frac{D_0 - D_1}{m-2}}{\frac{D_1}{n-m}}$ where D_0, D_1 are the deviances of M_0,M_1 (which are 385275.1 and 374856, see below) respectively, m is the number of distinct numbers of bathrooms in the data, and n is the total number of observations (which are 4 and 768 respectively in this case, see below). Under the null hypothesis this has an F distribution with $m - 2$ and $n - m$ degrees of freedom or more specifically 2 and 764. Equivalently, $F \sim F_{\{m-2,n-m\}} = F_{\{2,764\}}$

The observed value of the test statistic is ((385275.1-374856)/2)/(374856/764) which is 10.618 and the p-value is defined to be $P(F \geq F_{\{obs\}}) = P(F_{\{2,764\}} \geq 10.618)$ = 0.0000283 (see below) which is far less than 0.05 so at a 5% significance level we have strong evidence to reject the null hypothesis and we conclude that a more complex model than the straight line one is needed.

```
> deviance(M0)
[1] 385275.1

> deviance(M1)
[1] 374856
```

```
> length(unique(houses$bathrooms))
[1] 4
```

```
> length(houses$price)
[1] 768
```

```
> pf(10.618,2,764,lower.tail=FALSE)
[1] 2.828693e-05
```

# Question 2

a)

gas

b)

The expected price of a house using oil as fuel is 6.269 1000 usd (or 6269 usd) less than a house that has the same number of bathrooms and same living area but uses gas as fuel.

c)

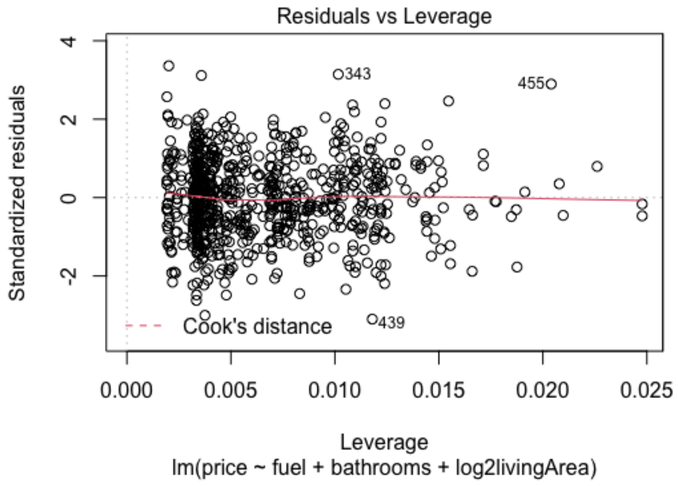
Predicted price in 1000 usd is -307.701 + (log2(1600)*44.932) + (2*14.512) = 199.523, so 199,523 usd

d)

The reference category has changed from gas to oil and as seen in 2a) the estimated coefficients for oil and electric are quite similar so hard to tell if there's really a difference.

e)

The datapoint with the highest influence has index 455 and it's cook's distance is 0.03494 (see below). As seen in the residuals vs leverage plot below this data point has one of the highest standardised residual values and also one of the highest leverage values therefore this point is influential due to both, being a regression outlier and having a high leverage.



```
> which.max((cooks.distance(slr.houses3)))
455
455

> cooks.distance(slr.houses3)[455]
    455
0.03493618
```

f)

The information provided gives us the regression sum of squares added by each variable in the formula for LM_2. As we are considering the existence of regression for the model LM_1 (which does not include fuel as a predictor variable) then the regression sum of squares contributed by the variable fuel is added to the residual sum of squares and the regression sum of squares is the sum from the other 2 variables. The total sum of squares remains the same and that enables us to produce the ANOVA table below. The regression degrees of freedom is number of parameters, of which the model has 3, minus 1 then the residual degrees of freedom is the number of observations minus the number of parameters, (768-3). Mean square is sum of squares divided by degrees of freedom and then the observed F statistic is the Regression mean square over the residual mean square as found in the notes.

SOURCE	DF	SS	MS	F
Regression	2	406109	203054.5	661.987
Residual	765	234652	306.735	
Total	767	640761		

g)

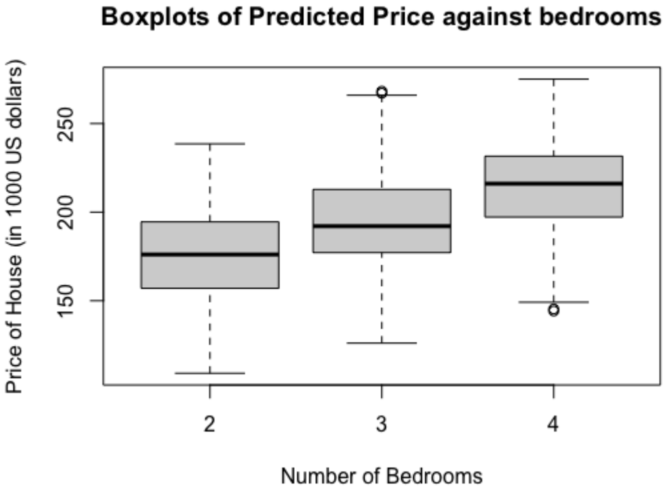
The null hypothesis is that price can be sufficiently modelled by LM_1 and the alternative hypothesis is that LM_2 is a significant improvement (aka adding fuel as a predictor variable is a significant improvement). If we look at the 3rd row of the sequential ANOVA table the p value is equivalent to the p value in our hypothesis test as the 3rd row corresponds to H_0 vs H_1. The value is 0.0001528 so at a 5% significance level we have strong evidence to reject the null hypothesis and therefore conclude that including fuel as a predictor variable is a significant improvement to the model. In R you can define the 2 models and use the anova function on them to produce the anova table that demonstrates the same result.

```
> anova(LM1,LM2)
Analysis of Variance Table

Model 1: price ~ log2livingArea + bathrooms
Model 2: price ~ log2livingArea + bathrooms + fuel
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    765 234651
2    763 229309  2    5342.6 8.8885 0.0001528 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Question 3

a)



There is a clear pattern in the plot that as the number of bedrooms increases the price of a house increases. It’s clear as the quartiles (inc. median) and the extremities (min and max) all increase as number of bedrooms increases.

b)

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

Residuals:
   Min      1Q  Median      3Q     Max
-54.445 -11.890  -0.146  11.764  57.550

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -330.99056   22.96310 -14.414  < 2e-16 ***
fuelgas          6.20059    1.71286   3.620 0.000314 ***
fueloil          0.08875    2.26726   0.039 0.968786
bathrooms       14.50238    1.49519   9.699  < 2e-16 ***
log2livingArea  47.13841    2.44635  19.269  < 2e-16 ***
bedrooms        -2.04002    1.22614  -1.664 0.096569 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.32 on 762 degrees of freedom
Multiple R-squared:  0.6434,	Adjusted R-squared:  0.6411
F-statistic:   275 on 5 and 762 DF,  p-value: < 2.2e-16
```

The p value of the estimated coefficient for bedrooms (0.0966) is not significant (as it is greater than 0.05) therefore including bedrooms as an explanatory variable is not a significant improvement on the model.

c)

```
> VIFlm1 = lm(bedrooms ~ bathrooms + log2livingArea, data=houses)
> VIFlm2 = lm(log2livingArea ~ bathrooms + bedrooms, data=houses)
> CoD1 = summary(VIFlm1)$r.squared
> CoD2= summary(VIFlm2)$r.squared
> VIF1 = 1/(1-CoD1)
> VIF1
[1] 1.653273
> VIF2 = 1/(1-CoD2)
> VIF2
[1] 2.139509
```

So the VIF of bedrooms is 1.653 and VIF of log2living Area is 2.140. Neither are particularly large and they are both certainly less than 10 so there is no concerns about those variables being collinear with the other predictors.