ST663 - Semester 1 - Assignment 5 - Solutions

Sean O'Riogain (18145426)

11 December 2018

```
knitr::opts_chunk$set(echo = TRUE)
getwd()
```

```
## [1] "C:/Users/oriogain/Dropbox/Maynooth/Statistical Methods/Semester 1 - Assignment 5"
```

```
suppressMessages(library(tidyverse))
```

Question 1

Consider a multiple regression model which predicts the calories in breakfast cereals from sodium, potassium and sugar content in grams.

Examine the regression results given below.

```
library(knitr)
df<-data.frame(x=c("(Intercept)","sodium","potass","sugars"),
  matrix(c(83.0469,0.0572,-0.0193,2.3876,
           5.1984,0.0215,0.0251,0.4066,
           15.98,2.67,-0.77,5.87,
           0.0000,0.0094,0.4441,0.0000),nrow = 4))
names(df)<-c("","Estimate","Std. Error","t value","Pr(>|t|")
kable(df)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	83.0469	5.1984	15.98	0.0000
sodium	0.0572	0.0215	2.67	0.0094
potass	-0.0193	0.0251	-0.77	0.4441
sugars	2.3876	0.4066	5.87	0.0000

```
df<-data.frame(c("Residual standard error: 15.6 on 73 degrees of freedom",
                 "Multiple R-squared: 0.3844, Adjusted R-squared: 0.3591",
                 "F-statistic: 15.2 on 3 and 73 DF, p-value: 8.868e-08"))
names(df)<-""
kable(df)
```

Residual standard error: 15.6 on 73 degrees of freedom
Multiple R-squared: 0.3844, Adjusted R-squared: 0.3591
F-statistic: 15.2 on 3 and 73 DF, p-value: 8.868e-08
One of the interpretations below is correct. Which is it?

Explain what is wrong with the others.

a. Each extra gram of sugar increases calories by 2.39.

This is not correct because it would only be valid if the values of the the other predictor (independent) variables (i.e. sodium and potassium) were fixed.

b. Every extra gram of sodium is associated with a 0.0572 increase in average calories, for cereals with a given potassium and sugar content.

This is the correct interpretation of the coefficient (slope) value for a predictor (independent) variable in multiple linear regression.

c. Every extra calorie means the potassium content drops by 0.0193g.

This is not correct because it is treating calories as a predictor (independent) variable when, for this regression, it is the response (dependent) variable.

d. The model does not fit because R2 is only 38.4%.

The R2 value provides some indication on the linear 'strength' of the model (indicating some degree of weakness in this case) but it does not, in itself, invalidate the model.

It merely indicates that 38.4% of the variation in the value of the response variable (Calories) is explainable by the model in question.

Question 2

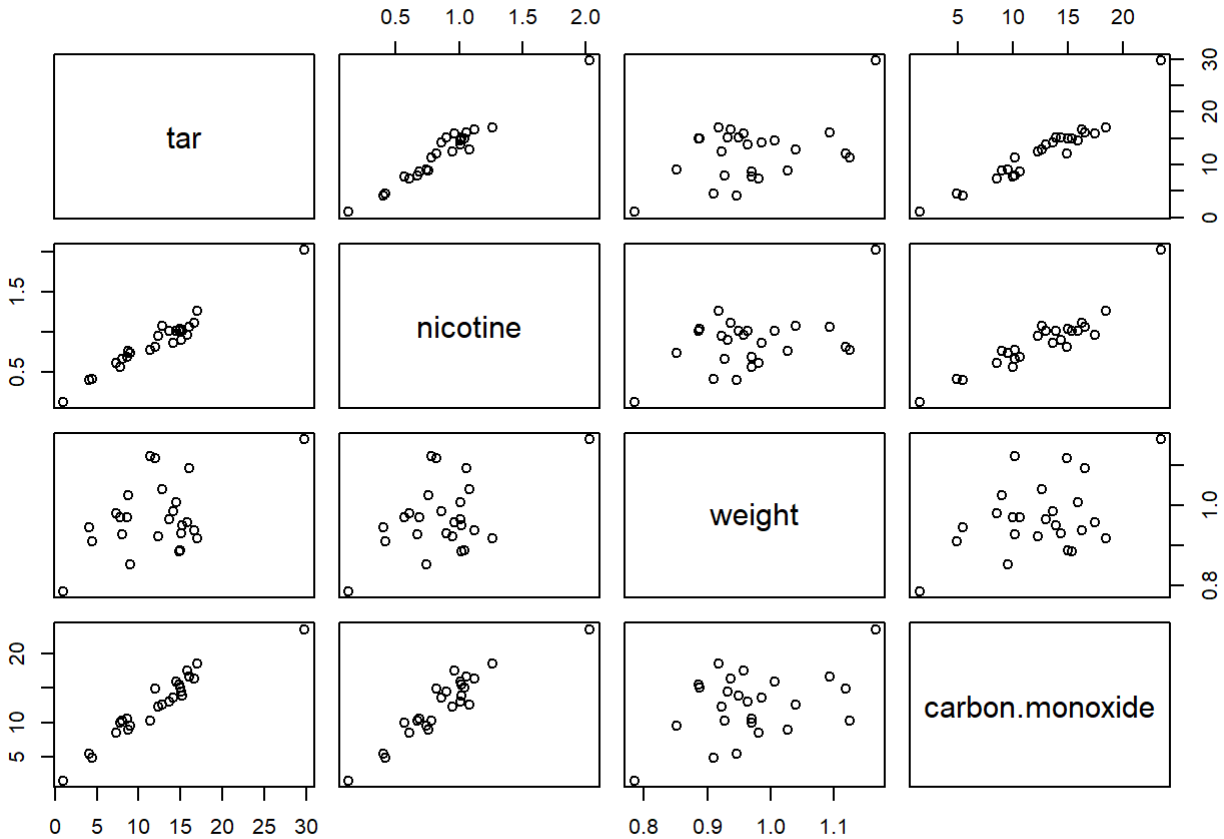
2. The FDA recorded data on tar, nicotine, weight and carbon monoxide for 25 cigarette brands. The data is in file Cigarette.csv.

```
d<-read.csv("Cigarette.csv")
str(d)
```

```
## 'data.frame':    25 obs. of  5 variables:
##  $ brand          : Factor w/ 25 levels "Alpine","Benson&Hedges",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ tar            : num  14.1 16 29.8 8 4.1 15 8.8 12.4 16.6 14.9 ...
##  $ nicotine       : num  0.86 1.06 2.03 0.67 0.4 1.04 0.76 0.95 1.12 1.02 ...
##  $ weight         : num  0.985 1.094 1.165 0.928 0.946 ...
##  $ carbon.monoxide: num  13.6 16.6 23.5 10.2 5.4 15 9 12.3 16.3 15.4 ...
```

a. Read in the data. Use pairs to make a plot of the 4 variables.

```
pairs(d[,2:5])
```



Comment on the correlation between the variables.

Visually, in the plots above, there appears be a linear relationship (correlation) between tar and nicotene, tar and carbon monoxide, and nicotene and carbon monoxide, and, possibly, between weight and carbon monoxide.

Which of the three predictors has the hightest correlation with carbon.monoxide?

```
cor(d[,2:5])
```

```
##                     tar  nicotine    weight carbon.monoxide
## tar             1.0000000 0.9766076 0.4907654       0.9574853
## nicotine        0.9766076 1.0000000 0.5001827       0.9259473
## weight          0.4907654 0.5001827 1.0000000       0.4639592
## carbon.monoxide 0.9574853 0.9259473 0.4639592       1.0000000
```

As shown in the carbon.monoxide column above, tar and nicotine have the highest (strongest) positive correlations (in decending order).

The lowest?

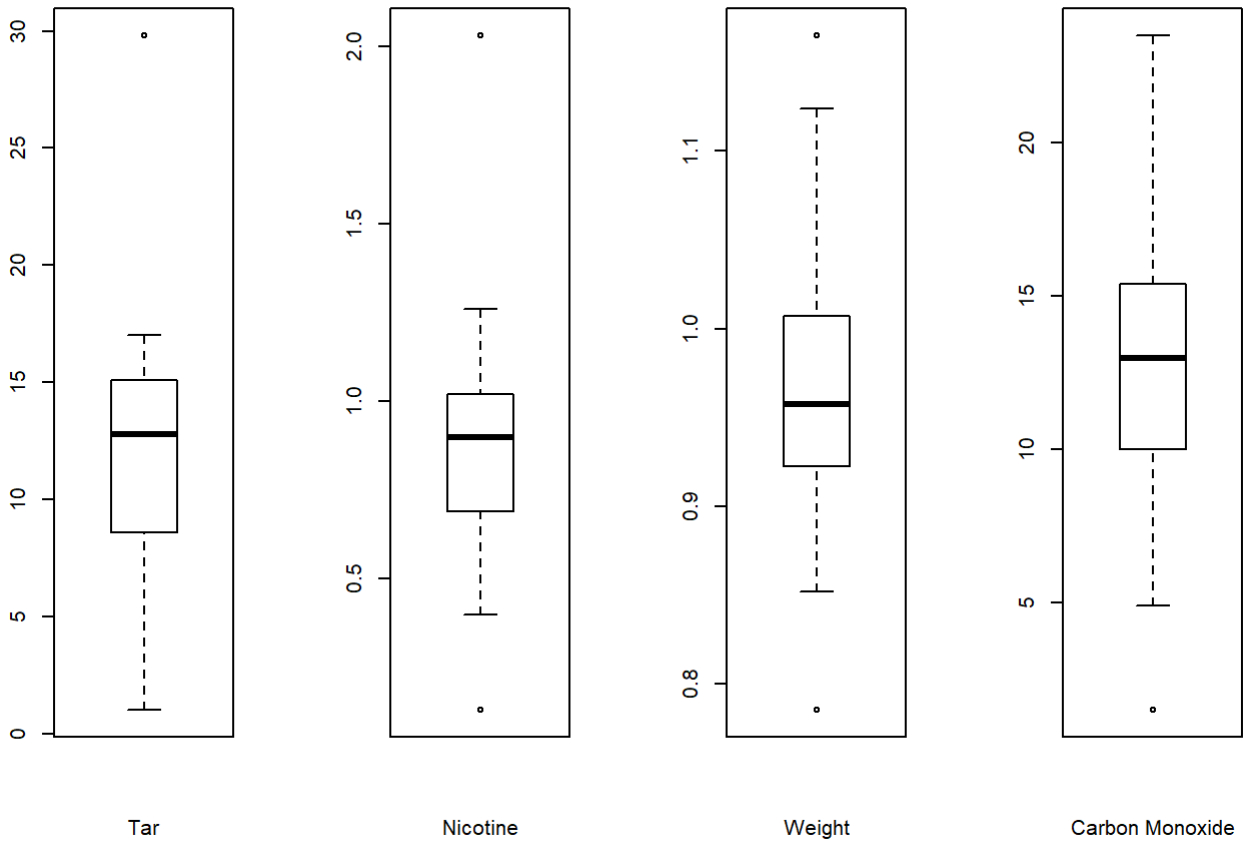
That column also shows that Weight has the lowest (weakest) correlation.

Are there any outliers evident?

In the pairs plot above, there is evidence that there is at least one outlier for all 3 predictors (tar, nicotine and weight).

The following boxplots support those observations.

```
par(mfrow=c(1,4))
boxplot(d$tar,xlab="Tar")
boxplot(d$nicotine,xlab="Nicotine")
boxplot(d$weight,xlab="Weight")
boxplot(d$carbon.monoxide,xlab="Carbon Monoxide")
```



b. Fit the regression model relating carbon.monoxide to tar, nicotine and weight.

```
f<-lm(data=d,carbon.monoxide~tar+nicotine+weight)

summary(f)
```

```
##
## Call:
## lm(formula = carbon.monoxide ~ tar + nicotine + weight, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89261 -0.78269  0.00428  0.92891  2.45082
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2022     3.4618   0.925 0.365464
## tar           0.9626     0.2422   3.974 0.000692 ***
## nicotine     -2.6317     3.9006  -0.675 0.507234
## weight       -0.1305     3.8853  -0.034 0.973527
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.446 on 21 degrees of freedom
## Multiple R-squared:  0.9186, Adjusted R-squared:  0.907
## F-statistic: 78.98 on 3 and 21 DF,  p-value: 1.329e-11
```

Write down the regression equation relating carbon monoxide to tar, nicotine and weight.

$$y_{carbon.monoxide} = 3.20 + 0.96 * tar - 2.63 * nicotine - 0.13 * weight$$

c. Interpret β_1.

A single unit increase in the Tar value will result in an increase of 0.96 in Carbon Monoxide value, where the Nicotine and Weight values are fixed.

Find a 95% confidence interval for β_1.

```
ci<-confint(f,"tar");ci
```

```
##         2.5 %   97.5 %
## tar 0.4587991 1.466349
```

Interpret this interval carefully in words.

A single unit increase in the Tar value will result in a change in the range of 0.46 and 1.47 in the Carbon Monoxide value (with 95% confidence), assuming that both the Nicotine and Weight values are fixed.

d. Interpret β_2.

A single unit increase in the Nicotine value will result in a decrease of 2.63 in the Carbon Monoxide value, where the Tar and Weight values are fixed.

Test H0 : $\beta_2 = 0$ versus Ha : $\beta_2 \neq 0$.

```
summary(f)
```

```
##
## Call:
## lm(formula = carbon.monoxide ~ tar + nicotine + weight, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89261 -0.78269  0.00428  0.92891  2.45082
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2022     3.4618   0.925 0.365464
## tar           0.9626     0.2422   3.974 0.000692 ***
## nicotine     -2.6317     3.9006  -0.675 0.507234
## weight       -0.1305     3.8853  -0.034 0.973527
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.446 on 21 degrees of freedom
## Multiple R-squared:  0.9186, Adjusted R-squared:  0.907
## F-statistic: 78.98 on 3 and 21 DF,  p-value: 1.329e-11
```

As the p-value for Nicotine is approximately 0.51 and, therefore, greater than 0.05, we cannot reject the Null hyppthesis and must conclude (with 95% confidence) that Nicotine does not signicantly influence the Carbon Monoxide value.

e. Interpret β_3.

A single unit increase in the Weight value will result in a decrease of 0.13 in the Carbon Moonoxide value, where the Tar and Nicotine values are fixed..

Test H0 : $\beta_3 = 0$ versus Ha : $\beta_3 \neq 0$.

As the p-value for Weight is 0.97 and, therefore, is greater than 0.05, we cannot reject the Null hyppthesis and must conclude (with 95% confidence) that Weight does not significantly influence the Carbon Monoxide value.

f. Which of the three predictors is most important in explaining the response?

Because its p-value of approximately 0.0007 is (much) lower than 0.05, Tar is the most important predictor.

g. What is the estimate of σ in the fit?

As per the summary of the results of the lm function above, 1.446 (Residual standard error) is the estimate of sigma for this model.

What is R2?

As per the summary of the results of the lm function above, 0.9186 (Multiple R-Squared) is the R2 estimate for this model.

h. Use the model to estimate the mean carbon.monoxide for brands with tar=10, nicotine=1 and weight = .9.

```
predict(f,data.frame(tar=10,nicotine=1,weight=0.9))
```

```
##        1
## 10.07883
```

The result of the predict function above indicates that the mean Carbon Monoxide value for such brands would be 10.08, approximately.

Find the associated confidence interval.

```
predict(f,data.frame(tar=10,nicotine=1,weight=0.9),interval="confidence")
```

```
##        fit      lwr      upr
## 1 10.07883 7.794905 12.36276
```

Interpret this interval carefully in words.

The latest results of the predict function indicate (with a 95% level of confidence) that the mean Carbon Monoxide reading for brands with the specified Tar, Nicotine and Weight values, would be in the 7.79 to 12.36 range, approximately.

i. Use the model to predict the carbon.monoxide for a brand with tar=10, nicotine=1 and weight = .9.

Find the associated prediction interval.

```
predict(f,data.frame(tar=10,nicotine=1,weight=0.9),interval="prediction")
```

```
##        fit      lwr     upr
## 1 10.07883 6.303164 13.8545
```

Interpret this interval carefully in words.

The latest results of the predict function indicate (with a 95% level of confidence) that the Carbon Monoxide reading for brands with the specified Tar, Nicotine and Weight values, would be in the 6.3 to 13.85 range, approximately.

j. Fit the reduced model with tar as the single predictor.

```
f_tar<-lm(carbon.monoxide ~ tar, data=d)

summary(f_tar)
```

```
##
## Call:
## lm(formula = carbon.monoxide ~ tar, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1124 -0.7167 -0.3754  1.0091  2.5450
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.74328    0.67521   4.063 0.000481 ***
## tar          0.80098    0.05032  15.918 6.55e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.397 on 23 degrees of freedom
## Multiple R-squared:  0.9168, Adjusted R-squared:  0.9132
## F-statistic: 253.4 on 1 and 23 DF,  p-value: 6.552e-14
```

Use anova to compare this reduced model with the model fit in part (c).

```
anova(f_tar,f)
```

```
## Analysis of Variance Table
##
## Model 1: carbon.monoxide ~ tar
## Model 2: carbon.monoxide ~ tar + nicotine + weight
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     23 44.869
## 2     21 43.893  2    0.9765 0.2336 0.7937
```

State the hypothesis being tested and give your conclusions carefully.

$H_0 : \beta_{nicotine} = \beta_{weight} = 0$ H_A **: At least one of** $\beta_{nicotine}, \beta_{weight} \neq 0$

As the p-value of 0.7937 achieved above is greater than 0.05, we cannot reject the Null hypothesis above, and conclude (with a 95% level of confidence) that we can omit the Nicotine and Weight predictors from the model.

Question 3

The file Real estate.csv has data on recent home sales in the home towns of students in a large U.S. Statistics class. In this question, you will look at how Price (in dollars) relates to living area (square feet) and number of bedrooms. Later you will also look at the location (urban, suburban or rural).

Read in the data with

```
house <- read.csv("Real_estate.csv")
house1 <- house[150:250,c(1:3, 8)]
house1$location.type <- factor(house1$location.type)

str(house1)
```

```
## 'data.frame':    101 obs. of  4 variables:
##  $ Price        : int  339000 339999 228500 289000 230000 299911 244500 1100000 599900 394900 ...
##  $ Living.area  : int  2875 3320 2028 2250 2212 1844 1923 4501 1740 1289 ...
##  $ bedrooms     : int  5 4 4 4 4 3 3 5 3 2 ...
##  $ location.type: Factor w/ 3 levels "r","s","u": 1 1 2 2 2 2 2 2 2 2 ...
```

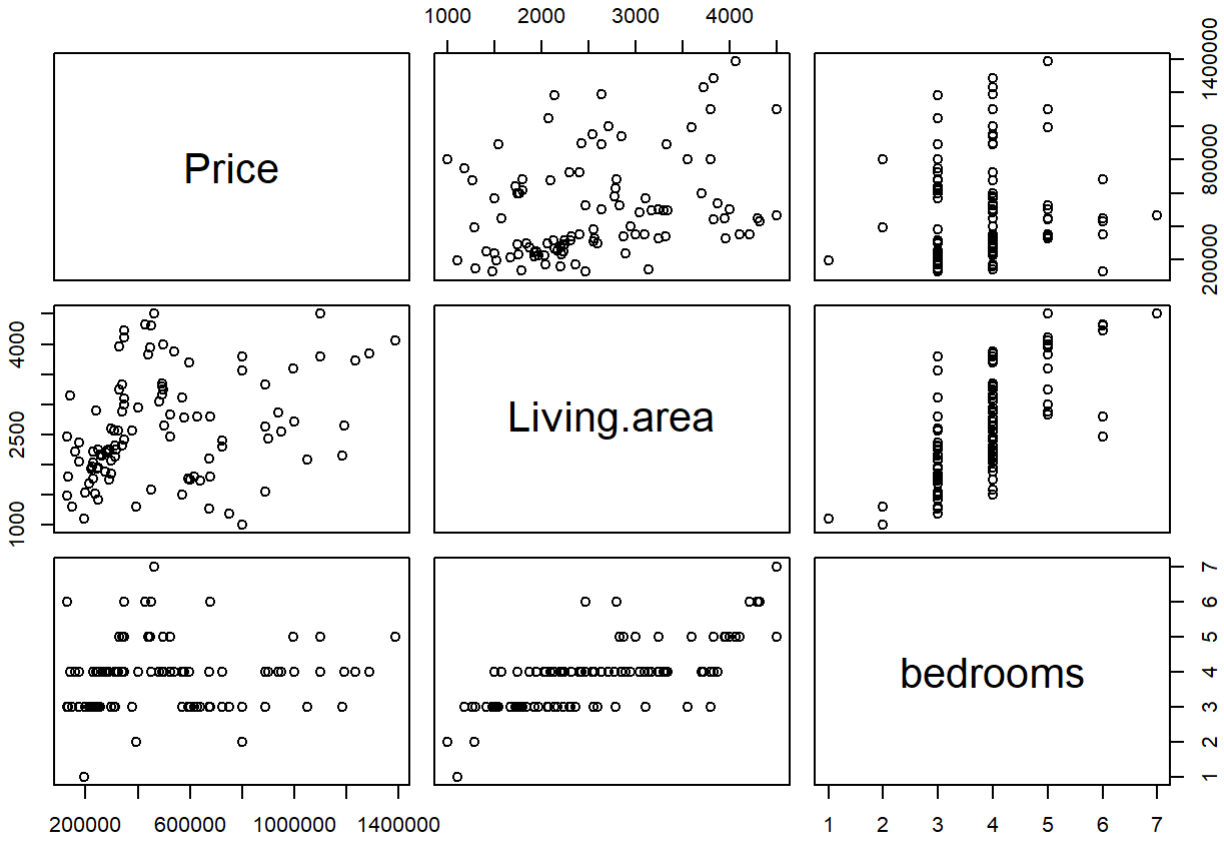
```
head(house1)
```

```
##      Price Living.area bedrooms location.type
## 150 339000        2875        5             r
## 151 339999        3320        4             r
## 152 228500        2028        4             s
## 153 289000        2250        4             s
## 154 230000        2212        4             s
## 155 299911        1844        3             s
```

As the dataset is large, use the specified subset only, ie the dataset house1.

a. Use pairs to make a plot of the 3 variables Price, Living.area and bedrooms.

```
pairs(house1[,1:3])
```



Comment on the correlation between the variables.

```
cor(house1[,1:3])
```

```
##                  Price Living.area   bedrooms
## Price       1.00000000   0.3098038 0.07481714
## Living.area 0.30980376   1.0000000 0.70343050
## bedrooms    0.07481714   0.7034305 1.00000000
```

The results of the cor function above indicate that Living Area has a weakly positive correlation with Price, while Bedrooms has a negligible (positive) correlation with Price.

Which of the predictors has the highest correlation with Price?

Living Area has the higher (weakly positive) correlation with Price.

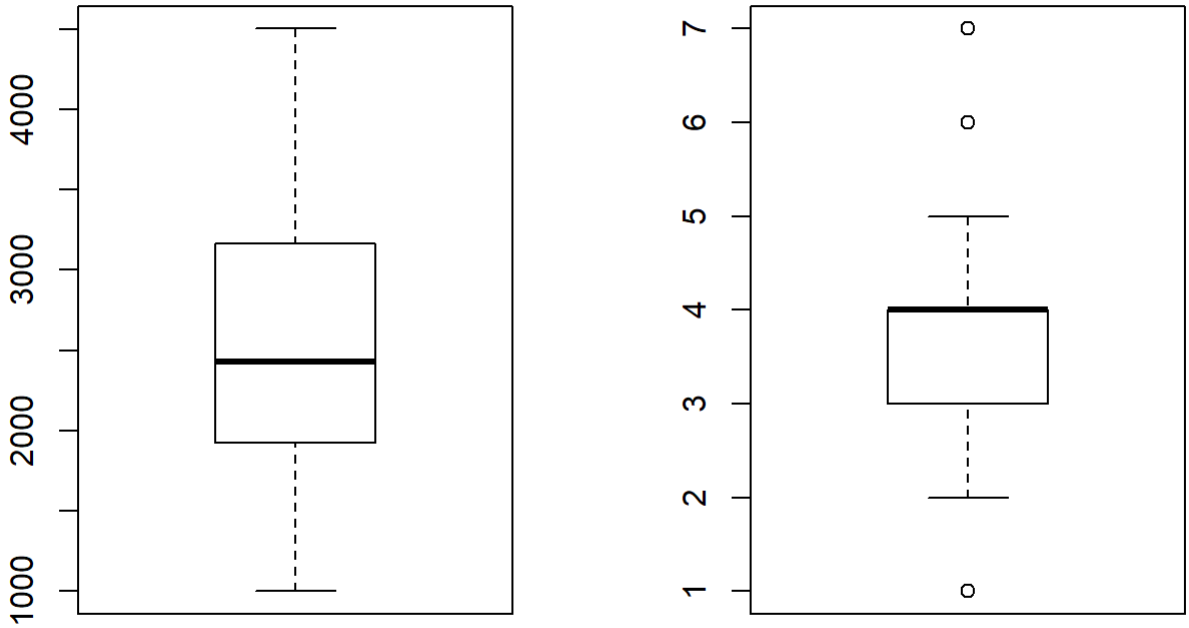
The lowest?

Bedrooms has the lower (negligibly positive) correlation with Price.

Are there any outliers evident?

**The following boxplots reveal that the Bedrooms predictor has 3 data points that are in the outlier category.

```
par(mfrow=c(1,2))
boxplot(x=house1$Living.area,y=house1$Price,xlab="Living Area")
boxplot(x=house1$bedrooms,y=house1$Price,xlab="Bedrooms")
```



b. Fit the regression model relating Price to Living.area and bedrooms. Call this fit f1.

```
f1<-lm(Price ~ Living.area + bedrooms, data=house1)

summary(f1)
```

```
##
## Call:
## lm(formula = Price ~ Living.area + bedrooms, data = house1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -444142 -193078 -111736  196949  730824
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 393569.62  117390.20   3.353 0.001139 **
## Living.area    175.46      45.52   3.855 0.000207 ***
## bedrooms    -89690.08   41813.72  -2.145 0.034425 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 282200 on 98 degrees of freedom
## Multiple R-squared:  0.1365, Adjusted R-squared:  0.1189
## F-statistic: 7.747 on 2 and 98 DF,  p-value: 0.0007523
```

Write down the regression equation.

$$y_{Price} = 393569.62 + 175.46 * Living.area - 89690.08 * bedrooms$$

c. Interpret β_1.

A single unit increase in Living Area value will result in a 175.46 increase in the Price value, assuming that the Bedrooms value is fixed.

Find a 95% confidence interval for β_1.

```
confint(f1,"Living.area")
```

```
##                2.5 %   97.5 %
## Living.area 85.12842 265.7853
```

Interpret this interval carefully in words.

A single unit increase in the Living Area value will result in a change in the Price value in the range of 85.13 to 265.793 (with 95% confidence), assuming that Bedroom value is fixed.

d. Interpret β_2. Find a 95% confidence interval for β_2.

```
confint(f1,"bedrooms")
```

```
##             2.5 %    97.5 %
## bedrooms -172668 -6712.124
```

Interpret this interval carefully in words.

A single unit increase in the Bedrooms value will result in a decrease in the Price value in the range of 6712 to 172,668 (with 95% confidence), assuming that Living Area value is fixed.

Comment: The Bedrooms variable does not appear to be a good predictor because of its extremely wide CI and because it indicates the Price decrease as the number of Bedrooms increase.

e. What is the estimate of σ in the fit?

The sigma estimate is 282,200 (the Residual Standard Error figure in the summary of the results of the lm function call for this fit above).

What is R^2?

0.1365 is the R2value for this model (the Multiple R-Squared value in the summary of the results of the lm function call for this fit above).

f. Use your fit to predict the Price of a house with Living.area=2800 and bedrooms=3.

```
predict(f1,data.frame(Living.area=2800,bedrooms=3))
```

```
##        1
## 615778.6
```

As shown above, the current model predicts a Price value of 615,778 for the specified house type.

Find the associated prediction interval.

```
predict(f1,data.frame(Living.area=2800,bedrooms=3),interval="prediction")
```

```
##        fit     lwr     upr
## 1 615778.6 46855.9 1184701
```

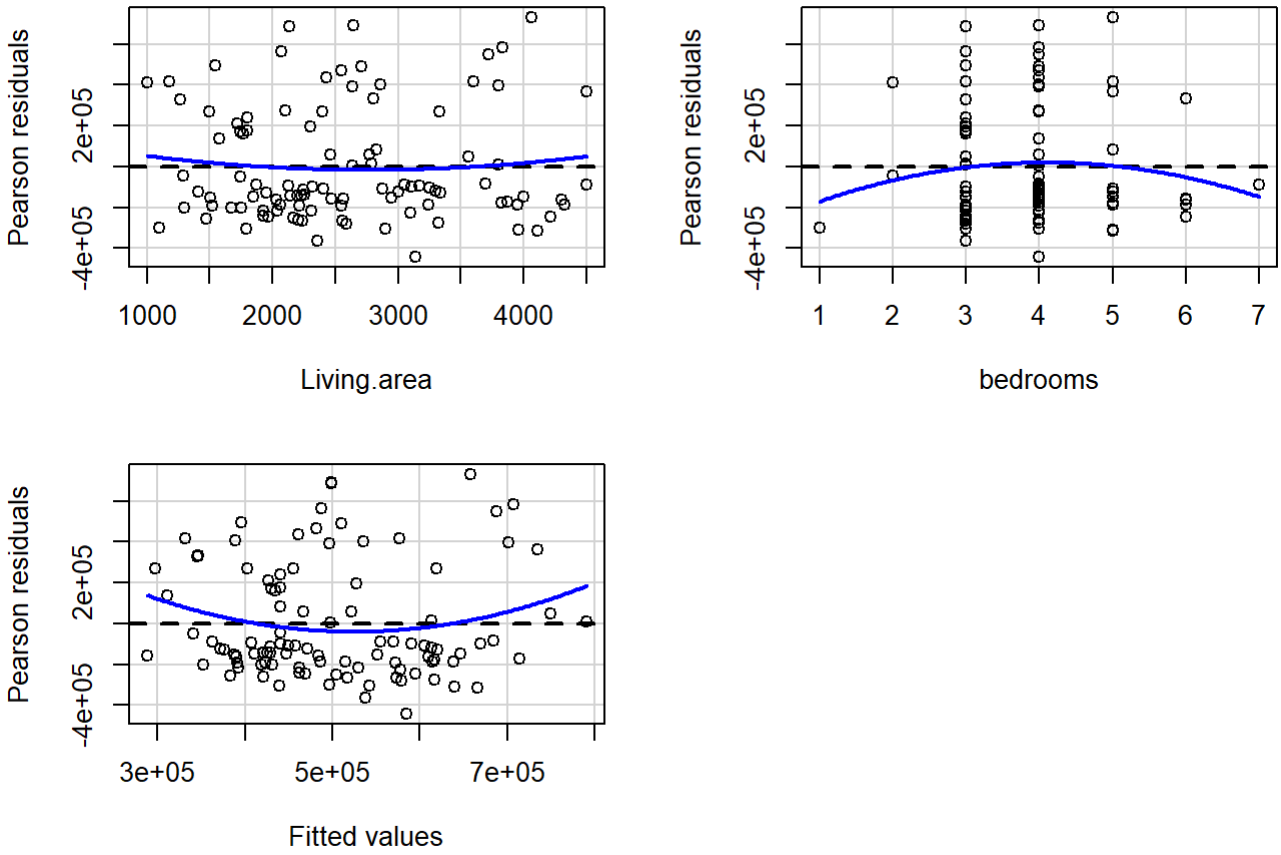
Interpret this interval carefully in words.

The latest results of the predict function indicate (with a 95% level of confidence) that the Pricevalue for the specified type of house would be in the 46,865 to 1,1184,701 range, approximately.

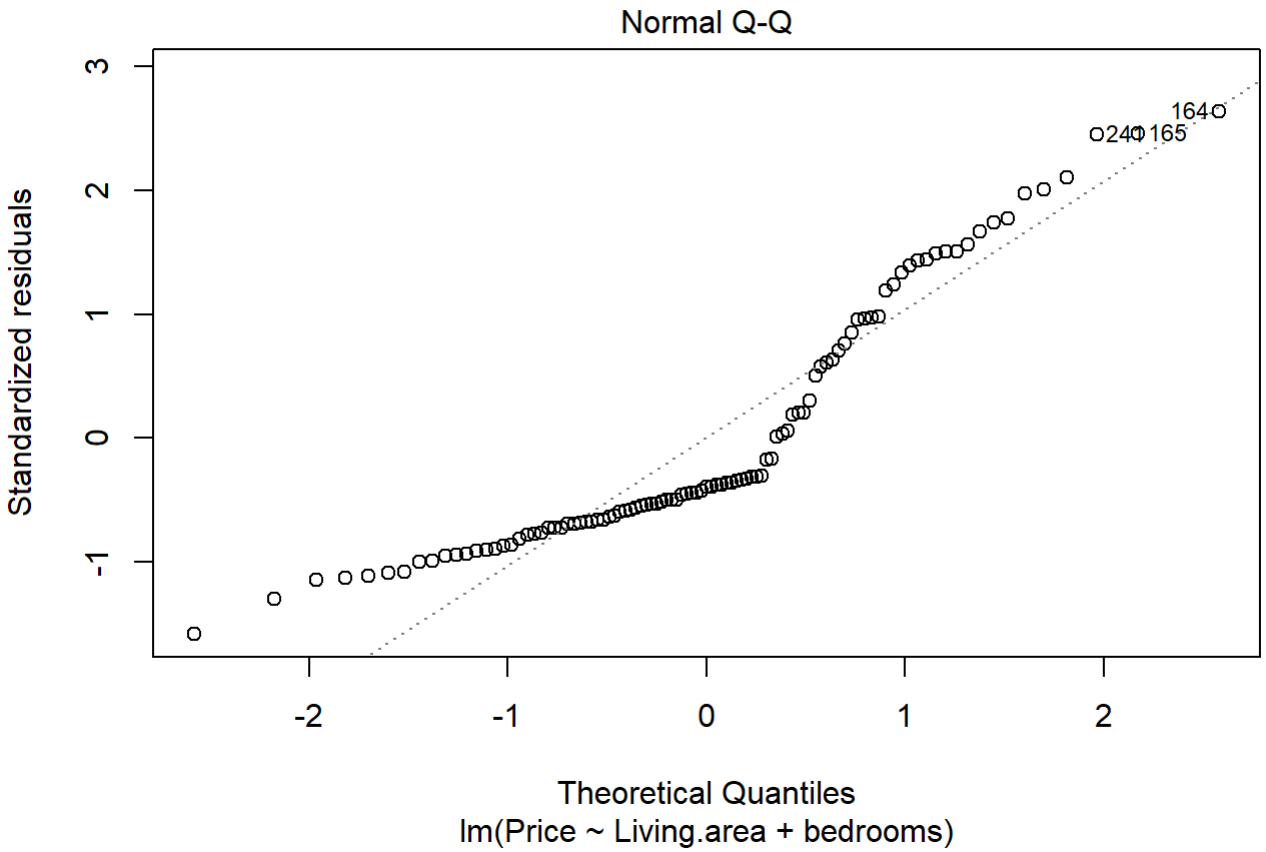
Refer to my previous comment in relation to relying on Bedrooms as a predictor in this model.

g. Assess the model assumptions. (See the code at the end of question 4).

```
suppressMessages(library(car))
residualPlots(f1, tests=F)
```




```
plot(f1,which=2)
```



The above plots clearly indicate that two of the main assumptions of linear regression (i.e. that the residuals are normally distributed and with constant variance) are not satisfied by this model. Living Area appears to do better on the constant variance front.

Question 4

For the house1 data of the previous question we will look at how the Price varies with location.type.

```
levels(house1$location.type)
```

```
## [1] "r" "s" "u"
```

```
levels(house1$location.type)<-c("Rural","Suburban","Urban")

levels(house1$location.type)
```

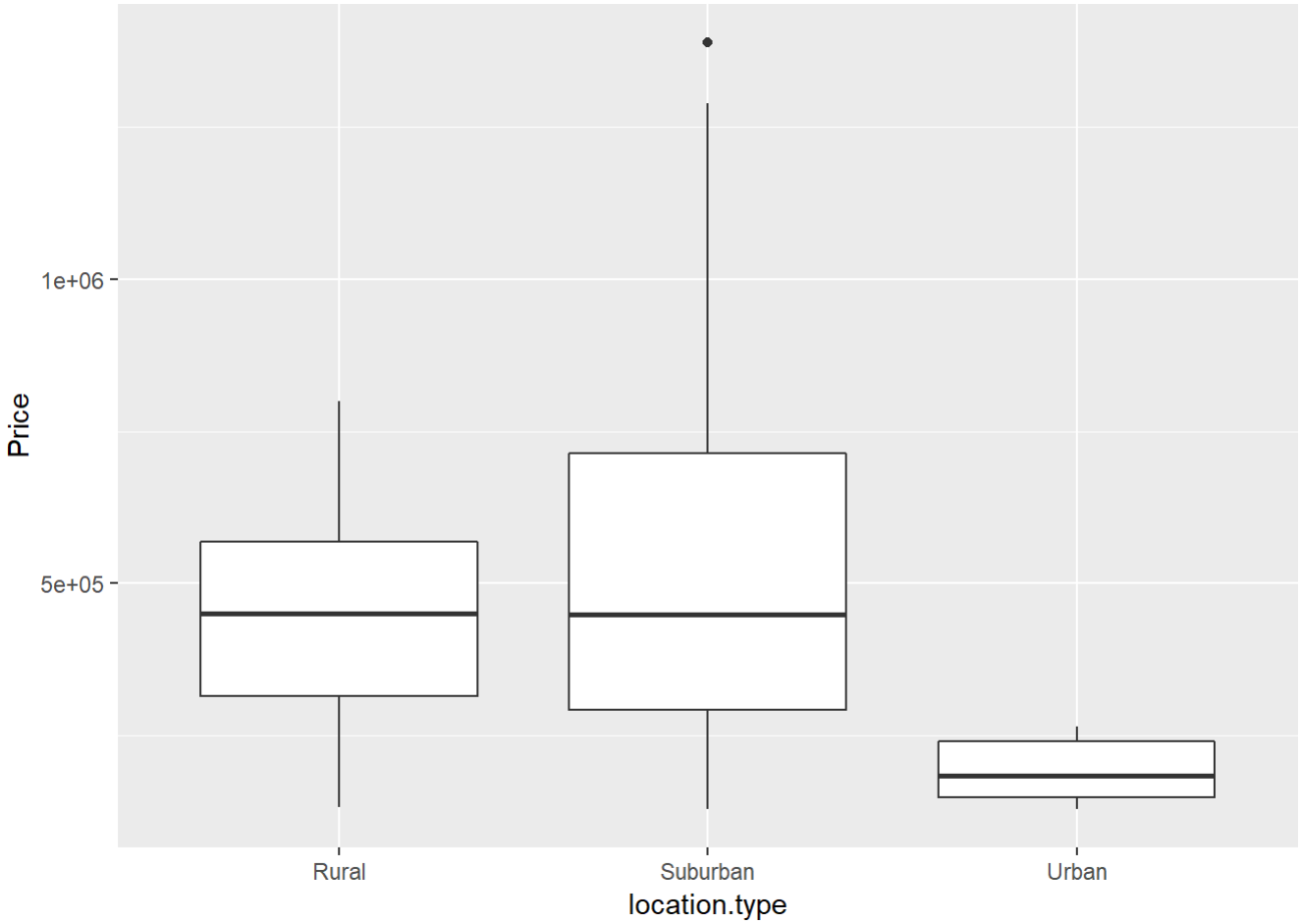
```
## [1] "Rural"    "Suburban" "Urban"
```

```
f2<-lm(Price ~ location.type, data=house1)

summary(f2)
```

```
##
## Call:
## lm(formula = Price ~ location.type, data = house1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -409872 -224872  -52133  139228  849228
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             445782      70576   6.316 7.88e-09 ***
## location.typeSuburban    93990      77888   1.207   0.2304
## location.typeUrban     -252749     138180  -1.829   0.0704 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 291000 on 98 degrees of freedom
## Multiple R-squared:  0.08163,    Adjusted R-squared:  0.06289
## F-statistic: 4.355 on 2 and 98 DF,  p-value: 0.01541
```

a. Draw boxplots for the Price for each group.

```
levels(house1$location.type)<-c("Rural","Suburban","Urban")

ggplot(data=house1,aes(x=location.type,y=Price)) + geom_boxplot()
```



Comment on the differences between the groups.

1. It is suprising that house prices in Urban areas have a significantly tighter spread and lower price range than those in both Rural and Suburban areas. 2. It is also surprising that Suburban and Rural areas have roughly the same median price and that Suburban prices have a greater spread than those in Rural areas. 3. The relative lack of outliers is also remarkable. 4. The previous comments indicate the need to review the spread of data across the location type for the selected sample of 101 houses (see below). <>

```
group_by(house1,location.type) %>%
  summarise(n())
```

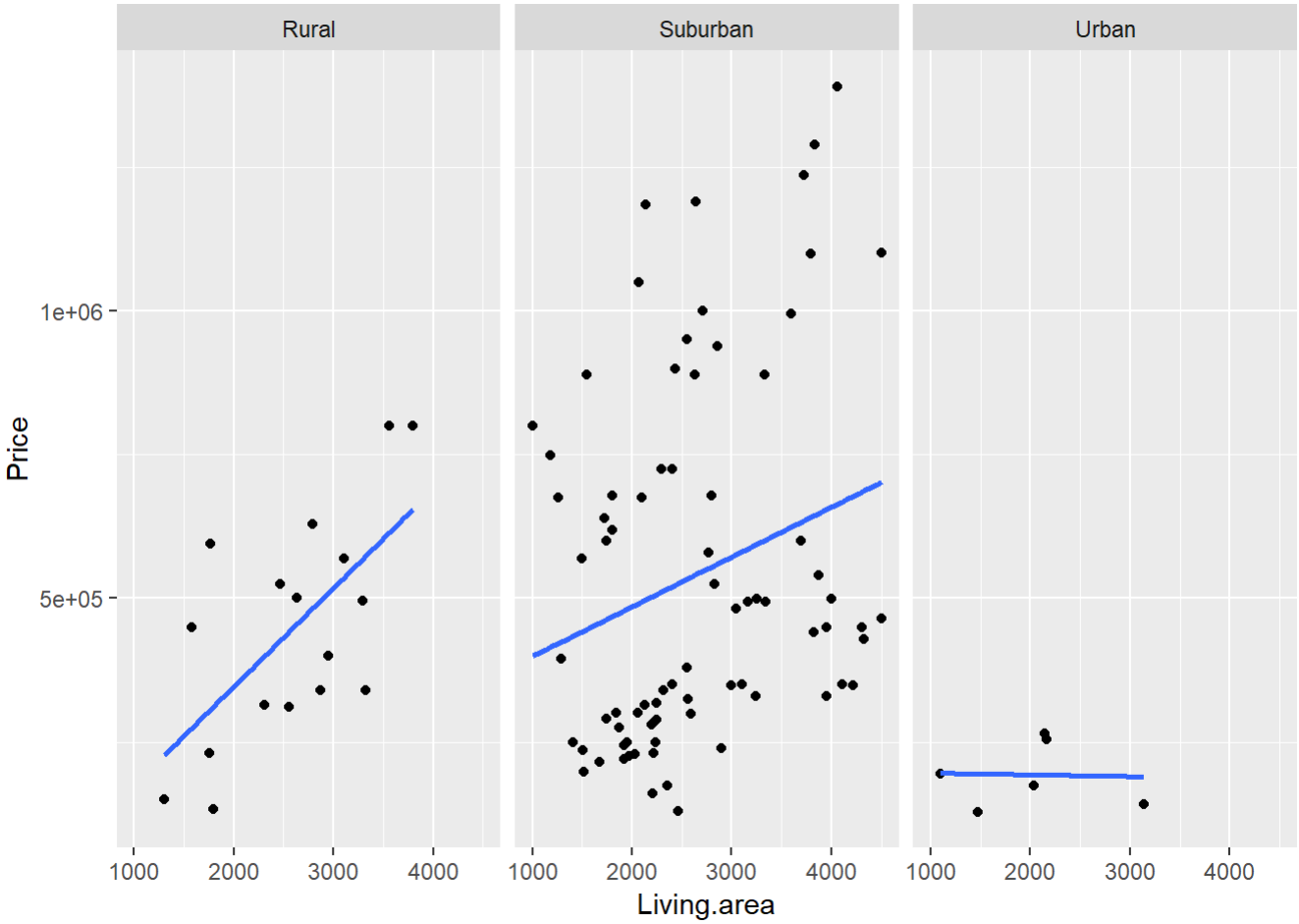
```
## # A tibble: 3 x 2
##   location.type `n()`
##   <fct>         <int>
## 1 Rural            17
## 2 Suburban         78
## 3 Urban             6
```

The above breakdown of data points by Location Type may well explain the comments in relation to the features of the boxplots above. In particular, the relatively size of the sample of Urban prices and, to a leser extent, of the Rural prices may be skewing the overall results.

Does it look like the variability is roughly constant across the 3 groups?

The following scatter plots show a lack of constant variance accross all 3 house locations - although the paucity of Urban data points doesn't lend itself to pattern recognition.

```
ggplot(house1,aes(Living.area,Price),color=location.type) + geom_point() + geom_smooth(method=lm,se=F)
 +facet_wrap(~location.type)
```



b. Fit a model relating Price to location.type. Call your fit f2. Use the anova function on the fit.

```
f2<-lm(Price ~ location.type, data=house1)

summary(f2)
```

```
##
## Call:
## lm(formula = Price ~ location.type, data = house1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -409872 -224872  -52133  139228  849228
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             445782      70576   6.316 7.88e-09 ***
## location.typeSuburban    93990      77888   1.207   0.2304
## location.typeUrban     -252749     138180  -1.829   0.0704 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 291000 on 98 degrees of freedom
## Multiple R-squared:  0.08163,    Adjusted R-squared:  0.06289
## F-statistic: 4.355 on 2 and 98 DF,  p-value: 0.01541
```

```
anova(f2)
```

```
## Analysis of Variance Table
##
## Response: Price
##               Df     Sum Sq    Mean Sq F value  Pr(>F)
## location.type  2 7.3759e+11 3.6880e+11  4.3553 0.01541 *
## Residuals     98 8.2983e+12 8.4677e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What conclusions can you draw from this about location.type?

As the Anova results give a p-value (0.01541) for Location Type that is less than 0.05, we can conclude that Location Type is significant when explaining variation in Price.

c. What is the estimate of σ in the fit?

291000 is the sigma (Residual standard error) for this model.

What is R^2?

0.08163 is the R2 (Multiple R-squared) value for this model.

Compare these values to those of fit f1.

The sigma value for f2 is greater than that of f1 and the R2 value for f2 is less than than for f1.

This (esp. the R2 value) indicate that f1 does a better job of explaining variation in price - although neither of them are impressive.

d. Fit the model f3 <- lm(Price ~ Living.area+bedrooms+location.type, data=house1)

Use anova to compare f1 and f3.

```
f3 <- lm(Price ~ Living.area+bedrooms+location.type, data=house1)

summary(f3)
```

```
##
## Call:
## lm(formula = Price ~ Living.area + bedrooms + location.type,
##     data = house1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -430315 -195524 -121763  176702  710506
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             386255.2   122602.4   3.150  0.00217 **
## Living.area                182.5       44.4   4.110 8.34e-05 ***
## bedrooms                -118492.1    41667.1  -2.844  0.00545 **
## location.typeSuburban    143018.4    75002.6   1.907  0.05953 .
## location.typeUrban      -184791.7   130090.2  -1.420  0.15871
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 271100 on 96 degrees of freedom
## Multiple R-squared:  0.2191, Adjusted R-squared:  0.1865
## F-statistic: 6.733 on 4 and 96 DF,  p-value: 8.061e-05
```

```
anova(f1,f3)
```

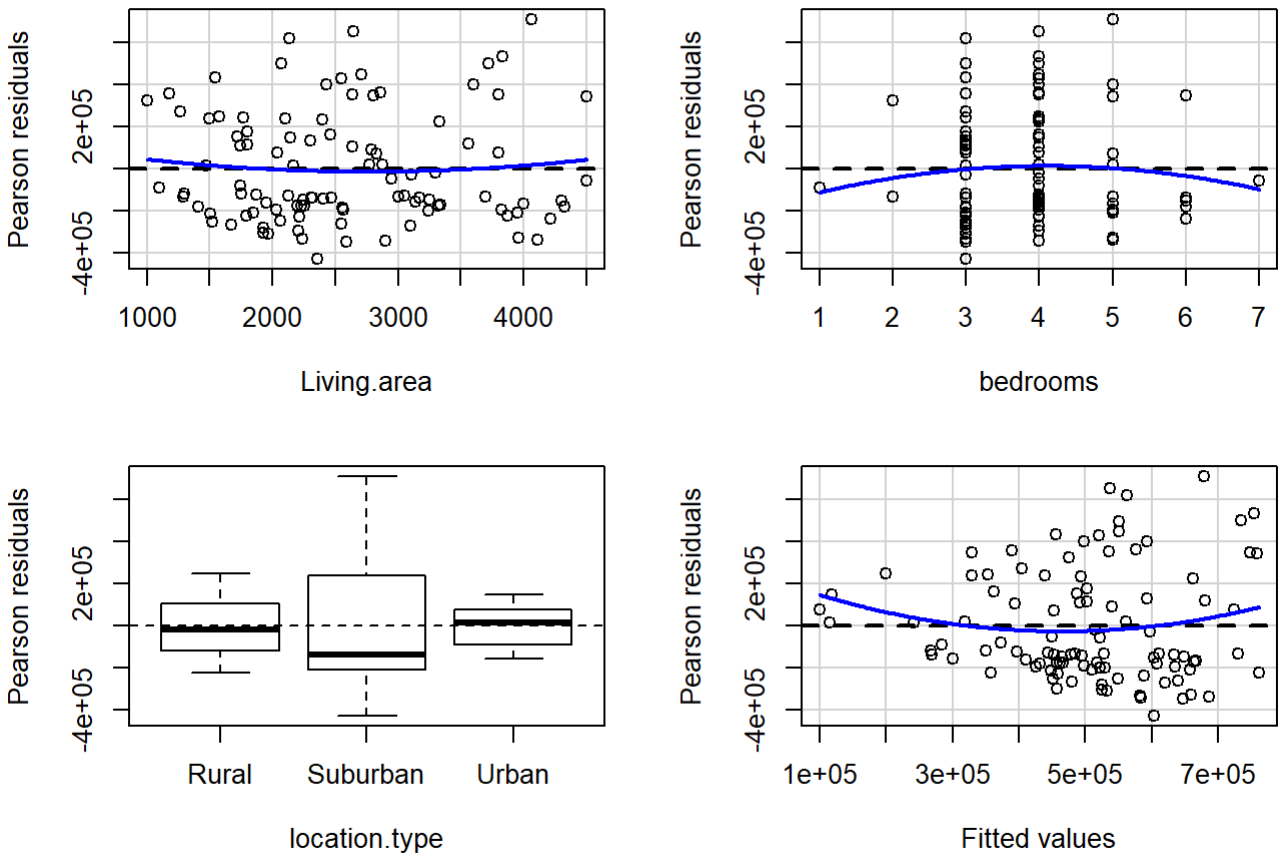
```
## Analysis of Variance Table
##
## Model 1: Price ~ Living.area + bedrooms
## Model 2: Price ~ Living.area + bedrooms + location.type
##   Res.Df        RSS Df  Sum of Sq      F   Pr(>F)
## 1     98 7.8023e+12
## 2     96 7.0563e+12  2 7.4604e+11 5.0749 0.008034 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What do the results of the anova tell you?

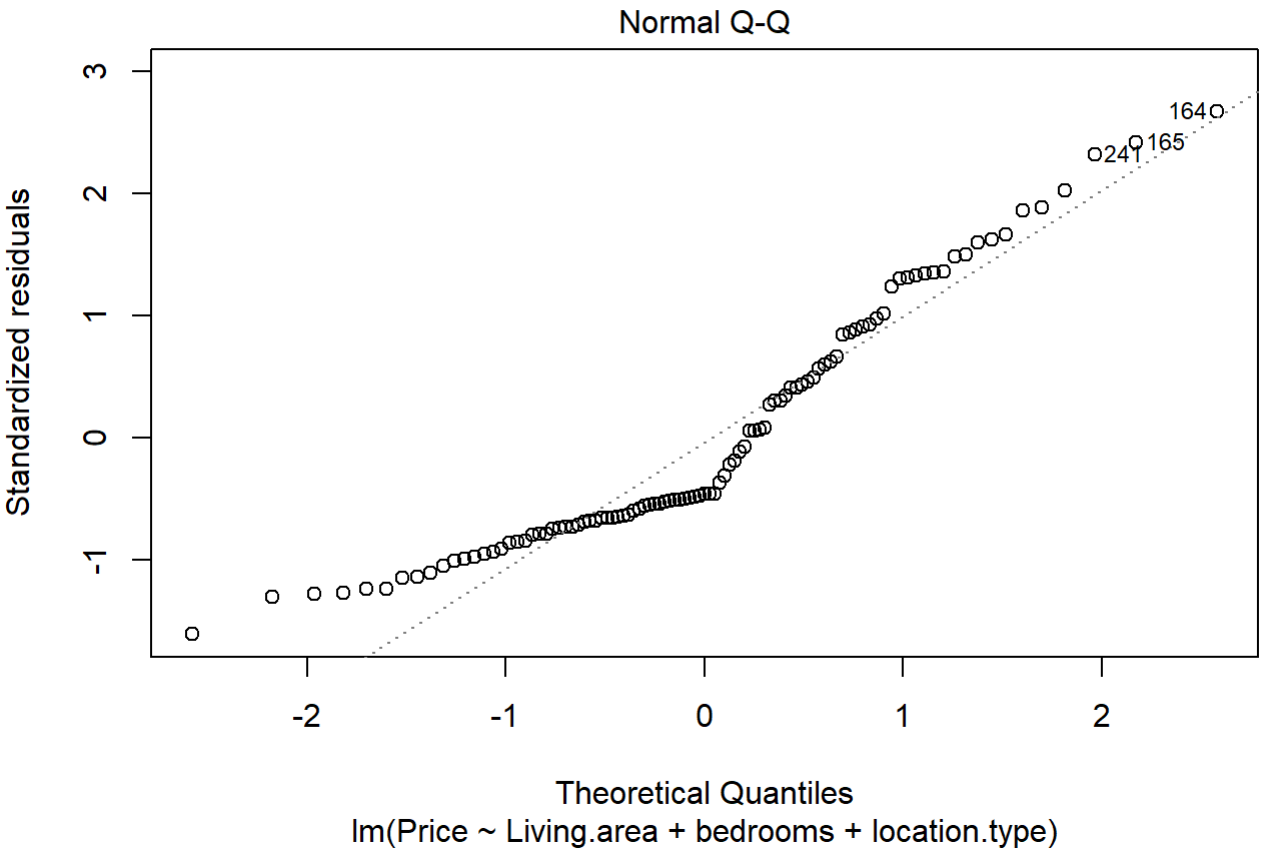
As the p-value (0.008034) for the extended model is less than 0.05, we can conclude that the addition of Location Type to the f1 model can a significant contribution to explaining variation in Price.

e. Assess the model assumptions for the fit f3 based on the plots produced by the given code.

```
library(car)
residualPlots(f3, tests=F)
```




```
plot(f3,which=2)
```


Normal Q-Q
Standardized residuals
3
2
1
0
-1
164
165
241
-2
-1
0
1
2
Theoretical Quantiles
lm(Price ~ Living.area + bedrooms + location.type)