STA 563 PROJECT 1

SIMPLE LINEAR REGRESSION ANALYSIS
ON
HOUSING DATA

Submitted by:
Alisha Rajbhandari
Erika Rasnick

October 23, 2017

Introduction

Buying a house is one of the largest financial decisions most people will make in their lifetime. Therefore, understanding what factors influence the price of the house can give insight into this big decision. The Ames Housing dataset was compiled by Dean DeCock for use in data science education, and consists of many variables describing 1,460 individual houses in Ames, Iowa. For this project, we are using two of these variables – the property's sale price in dollars (SalePrice) is the response variable that we are trying to predict, and the size of the home's first floor in square feet (1stFlrSF) is the predictor variable. We believe that there is a linear relationship between the square footage of the first floor of a house and the house's selling price. That is, the larger the first floor, the higher the price of the home.

Exploratory Data Analysis

Upon first inspection of the data, we found that both the predictor and response variables are heavily skewed to the right (Figure 1). For a simple linear regression model, we assume the error terms are normally distributed with a mean of zero and variance of σ^2, and the expected value of the response variable is normal with a mean of $\beta_0+\beta_1X_i$ and variance σ^2. Because our response variable is not normally distributed, we transformed the sale price variable by taking the logarithm of each value. After the transformation, we can see that the response variable is more symmetric and appears to be more normally distributed (Figure 2).

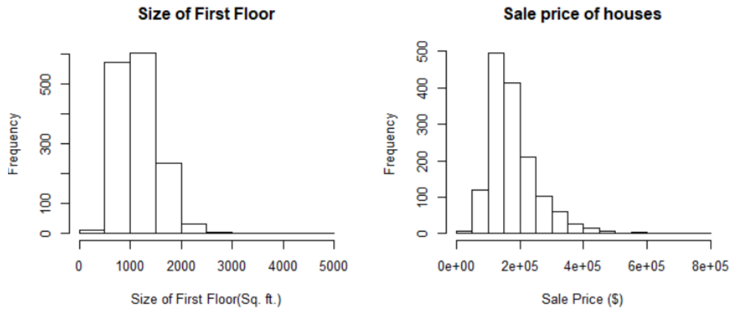


Figure 1. Histogram for size of first floor and sale price

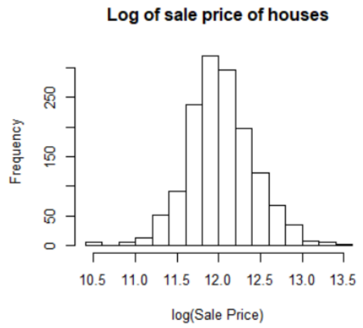


Figure 2. Histogram for log of sale price

This log transformation is further supported by a scatter plot of the residuals from the linear regression of sale price vs. size of the first floor (Figure 3). From this plot, we can see that the constant variance assumption would be violated. Hence, it would not be appropriate to fit the model without transforming the response variable. The output and graphs for this full analysis can be found in the appendix (Figures 4 and 5).

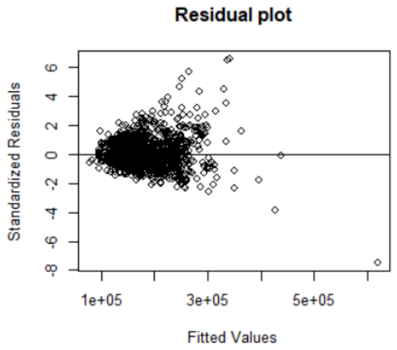


Figure 3. Residual plot

Analysis

We now perform a simple linear regression analysis with the log of the sale price as the response variable and the size of the first floor as the predictor. The simple linear regression model is of the form

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i\ , \qquad i = 1, 2, 3, \ldots, 1{,}460$$

where $\ln(Y_i)$ is the response variable (log of sale price of the house in dollars), X_i is the predictor (size of the first floor in square feet), and ϵ_i is the error term. We assume the error terms are independent and identically distributed as normal with mean zero and variance σ^2.

A simple linear regression analysis of the logarithm of sale price and the size of the first floor yields the following regression equation

$$\ln(\widehat{\text{Sale Price}}) = 11.31 + 0.00062 * \text{ Size of First Floor} .$$

This can be interpreted to indicate that for every one-square foot increase in the size of the first floor of the home, the geometric mean sale price of the house increases by a factor of 1.0006. Approximately 35.64% of the variability in the log of the sale price of houses is explained by this linear relationship with the size of the first floor as shown by the R-squared. We would like to test the following hypotheses:

H_0: $\beta_1 = 0$ (There is no linear relationship between sale price of house and size of first floor.)

H_1:$\beta_1 \neq 0$ (There is a linear relationship between sale price of house and size of first floor.)

Under null hypothesis, we have a t-statistic of 30.894 and a p-value of less than 2e-16, which is smaller than any reasonable significance level. Hence, we reject the null hypothesis and conclude that there is a linear relationship between the log of the sale prices of homes and the size of their first floor. The R output for this regression analysis can be found in the appendix (Figure 6).

Before the transformation of the sale price, the model form assumption and the constant variance assumptions were violated (Appendix, Figure 5), but the transformation remediated these violations as seen in Figure 7a and 7c in the appendix. However, there are still outliers and influential points, violating those two assumptions as well as the normality assumption (Appendix, Figure 7b and 7d).

Upon removal of the outliers, the simple linear regression analysis of the logarithm of sale price and the size of the first floor of the house changed very little. In this analysis, the geometric mean sale price of the house increased by a factor of 1.0007 for every one-square foot increase in the size of the first floor, and approximately 38.75% of the variability in the log of the sale price of the house is explained by the linear relationship with its first-floor size as shown by the R-squared. The t-statistic is now 30.29 and the p-value did not change. The R output for this regression analysis can be found in Figure 8.

Now the outliers and influential points assumptions are not violated (Figure 9), and the distribution of the errors is roughly normal (Figure 9d). The relationship between the two variables appears linear (Figure 10) and the variance of the residuals is mostly constant (Figure 9c). Although outliers cannot simply be removed from a dataset, temporarily removing them did not change our results significantly, so we can be confident in our results.

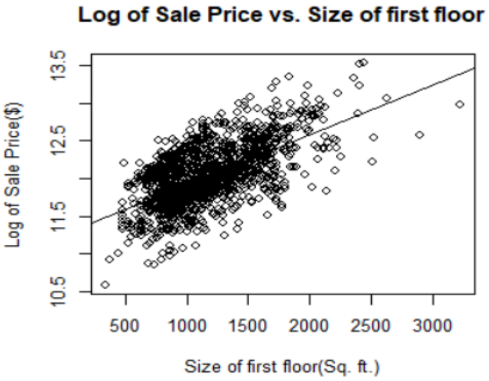
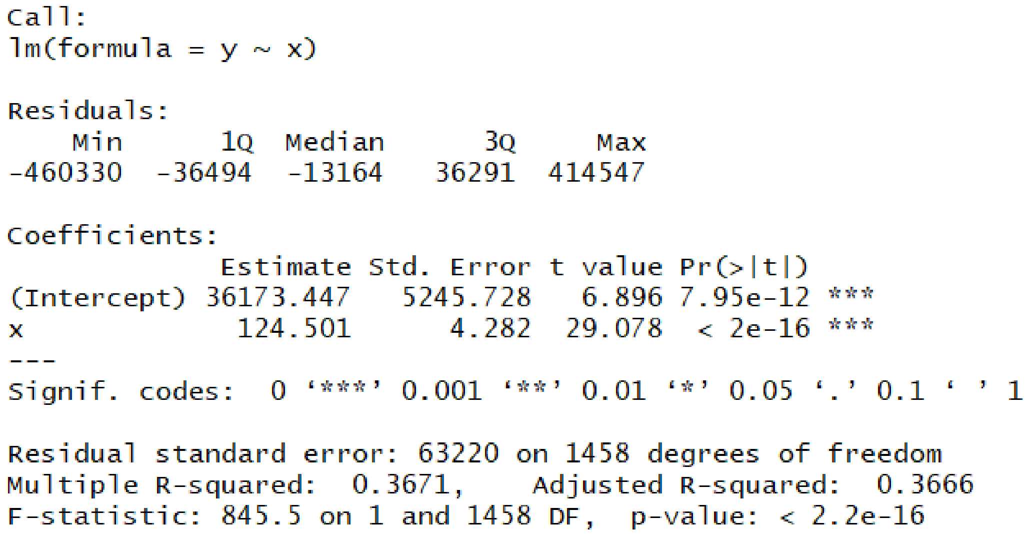


Figure 10. Scatterplot of sale price of house vs. size of first floor after transforming sale price and removing outliers

Conclusion

From this analysis, we conclude that there is a linear relationship between the log of a house's selling price and the size of its first floor. However, we could not accurately check the assumption that the errors are independent because we do not know how the data were collected. Although we found a significant relationship between these two variables, the coefficient of determination was not very large (.3875). This means that there is a significant amount of variation in the sale price of houses that is not explained by the size of the first floor. However, this is to be expected, as there are many other predictors that could influence the price of the home (e.g., number of bedrooms, number of bathrooms, year of sale, neighborhood, finished garages, air conditioning system, etc.).

Appendix



```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-460330  -36494  -13164   36291  414547 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 36173.447   5245.728   6.896 7.95e-12 ***
x             124.501      4.282  29.078  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63220 on 1458 degrees of freedom
Multiple R-squared:  0.3671,	Adjusted R-squared:  0.3666 
F-statistic: 845.5 on 1 and 1458 DF,  p-value: < 2.2e-16
```

Figure 4. R output for the regression of sale price and size of the house's first floor

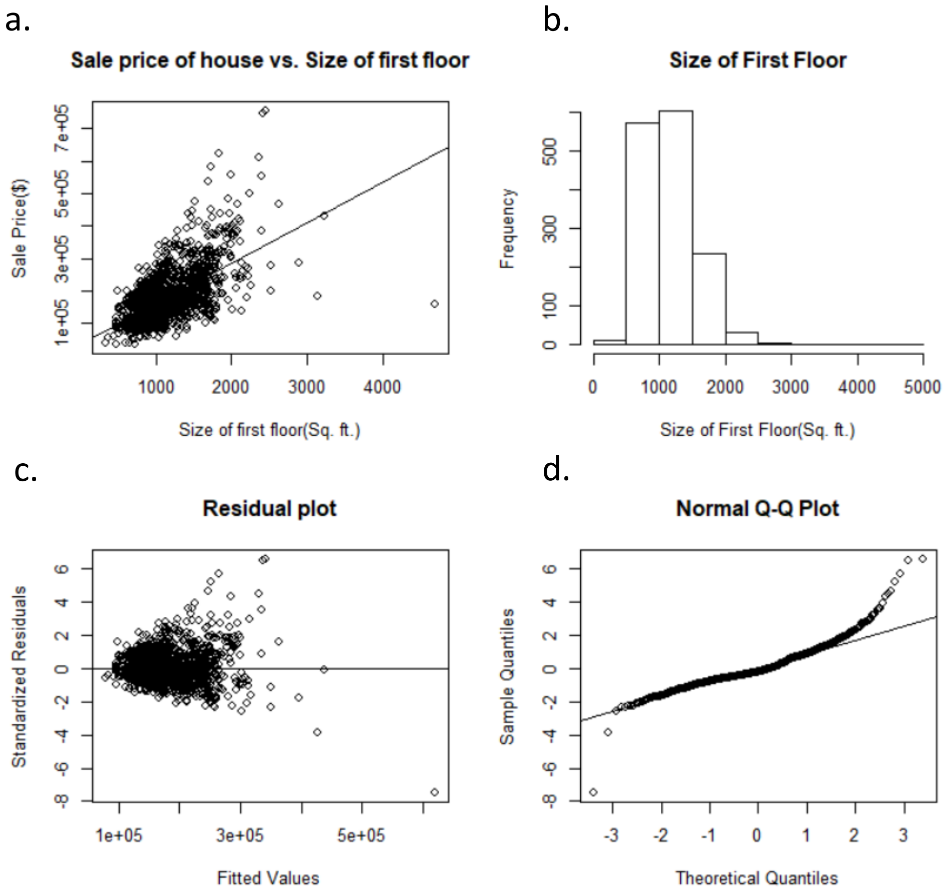
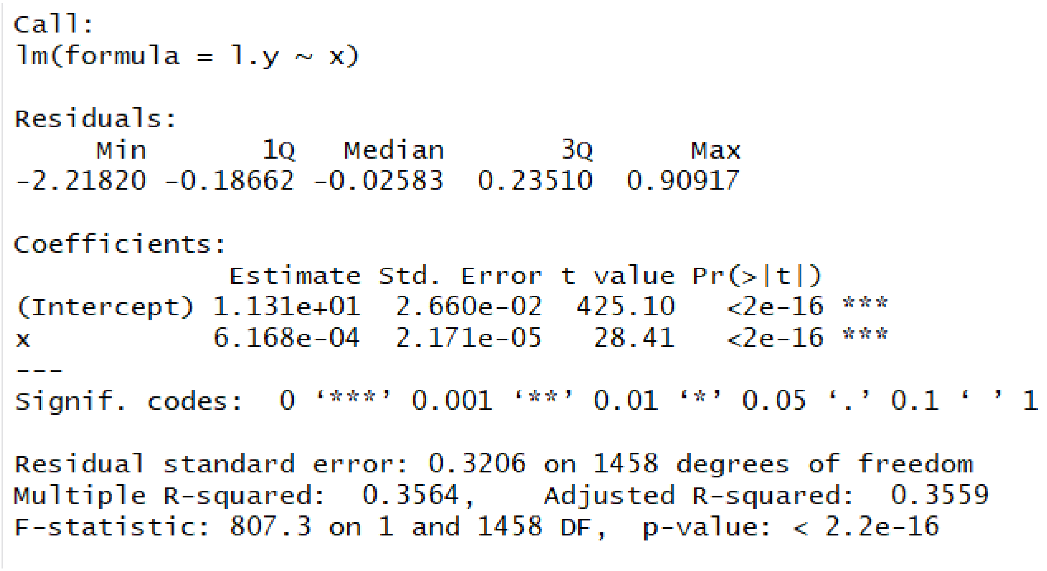


Figure 5. Plots for regression of sale price and size of house's first floor showing violations of assumptions



```
Call:
lm(formula = l.y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.21820 -0.18662 -0.02583  0.23510  0.90917 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.131e+01  2.660e-02  425.10   <2e-16 ***
x           6.168e-04  2.171e-05   28.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3206 on 1458 degrees of freedom
Multiple R-squared:  0.3564,    Adjusted R-squared:  0.3559 
F-statistic: 807.3 on 1 and 1458 DF,  p-value: < 2.2e-16
```

Figure 6. R output for the regression of log of sale price and size of the house’s first floor

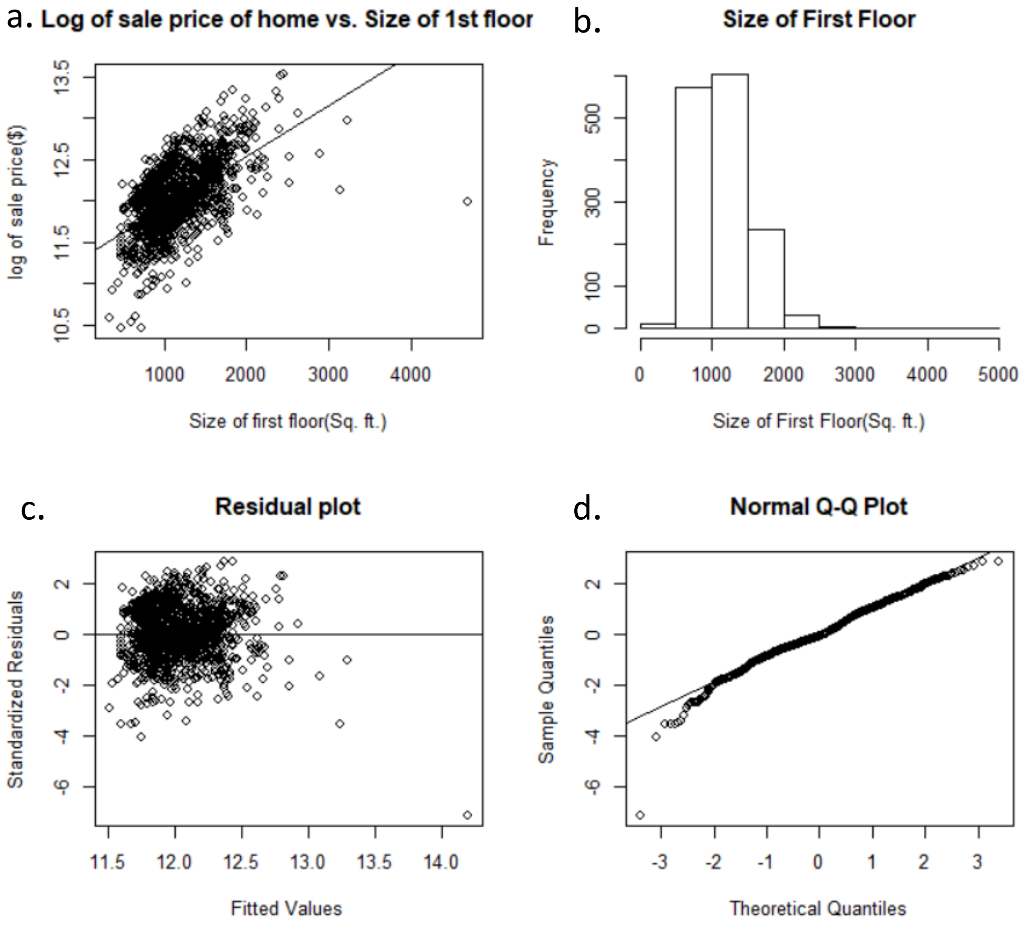
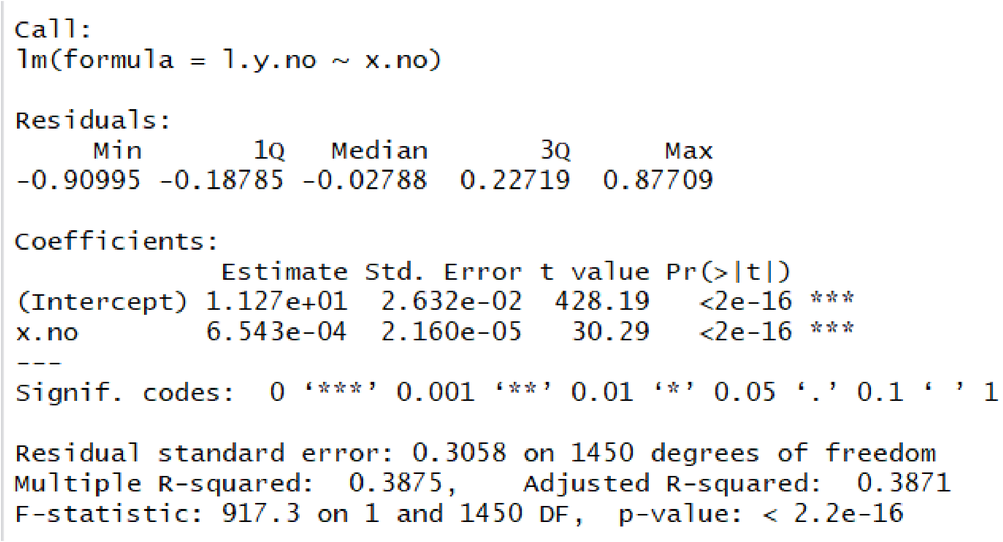


Figure 7. Plots for regression of log of sale price and size of house’s first floor that show violations of outliers and influential points assumptions



```
Call:
lm(formula = l.y.no ~ x.no)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.90995 -0.18785 -0.02788  0.22719  0.87709 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.127e+01  2.632e-02  428.19   <2e-16 ***
x.no        6.543e-04  2.160e-05   30.29   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3058 on 1450 degrees of freedom
Multiple R-squared:  0.3875,	Adjusted R-squared:  0.3871 
F-statistic: 917.3 on 1 and 1450 DF,  p-value: < 2.2e-16
```

Figure 8. R output for the regression of log of sale price and size of the house’s first floor with outliers removed

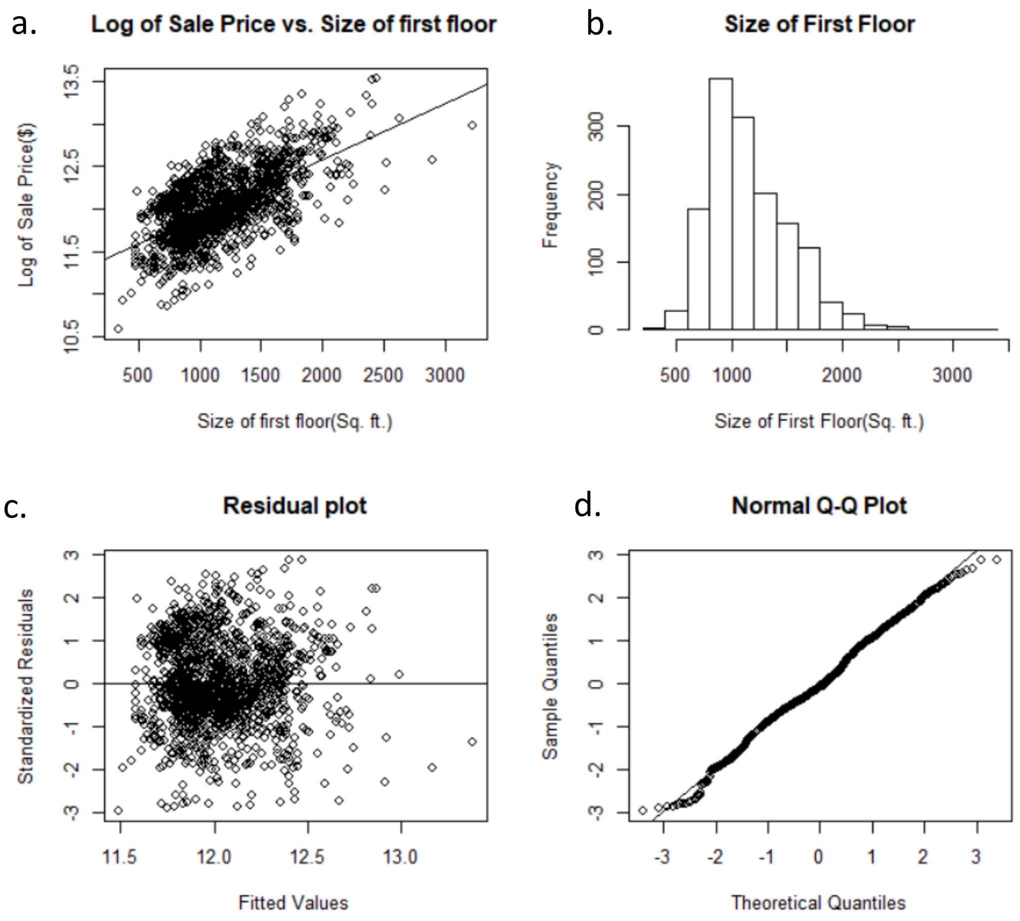


Figure 9. Plots for regression of log of sale price and size of home’s first floor that show no violations of assumptions

R code to generate the outputs

```
## Dataset - housing.csv

## Data preparation
house <- read.csv("D:/ALISHA/.MIAMI Classes/1st sem/Regression/Project1/housing.csv",
              header=TRUE)
x = house$X1stFlrSF
y = house$SalePrice

## make a 1x2 grid of plots (Figure 1)
par(mfrow=c(1,2))

## create a histogram
hist(x, xlab="Size of First Floor(Sq. ft.)", main="Size of First Floor")
hist(y, xlab="Sale Price ($)", main="Sale price of houses")

## Figure 2
par(mfrow=c(1,1))
hist(log(y), xlab="log(Sale Price)", main="Log of sale price of houses")

## Simple linear regression
house.slr <- lm(y~x)
summary(house.slr)                  ## Figure 4

## make a 2x2 grid of plots (Figure 5)
par(mfrow=c(2,2))
plot(x,y, xlab="Size of first floor(Sq. ft.)", ylab = "Sale Price($)",
    main = "Sale price of house vs. Size of first floor")
abline(house.slr)
hist(x, xlab="Size of First Floor(Sq. ft.)", main="Size of First Floor")
plot(house.slr$fitted,rstandard(house.slr),xlab="Fitted Values",          ## Figure 3 and 5c
    ylab="Standardized Residuals", main="Residual plot")
abline(h=0)
qqnorm(rstandard(house.slr))
qqline(rstandard(house.slr))

## log transforming y
l.y <- log(y)
log.slr <- lm(l.y~x)
summary(log.slr)                    ## Figure 6

## Figure 7
par(mfrow=c(2,2))
plot(x,l.y,xlab="Size of first floor(Sq. ft.)", ylab = "log of sale price($)",
    main = "Log of sale price of home vs. Size of 1st floor")
abline(log.slr)
hist(x, xlab="Size of First Floor(Sq. ft.)", main="Size of First Floor")
plot(log.slr$fitted,rstandard(log.slr),xlab="Fitted Values",
    ylab="Standardized Residuals", main="Residual plot")
abline(h=0)
```

```
qqnorm(rstandard(log.slr))
qqline(rstandard(log.slr))

par(mfrow=c(1,1))
plot(log.slr$fitted,rstandard(log.slr),xlab="Fitted Values",ylab="Standardized Residuals")
abline(h=-3)
abline(h=3)

## identify the outliers in the residual plot
identify(log.slr$fitted,rstandard(log.slr),xlab="Fitted Values",ylab="Standardized Residuals")
# [1]   31  411  496  524  813  917  969 1299

## log transforming y after removing more outliers
l.y.no <- log(y.no)
plot(x.no, l.y.no)
log.slr.no <- lm(l.y.no~x.no)
summary(log.slr.no)                    ## Figure 8

## Figure 9
par(mfrow=c(2,2))
plot(x.no,l.y.no, xlab="Size of first floor(Sq. ft.)", ylab = "Log of Sale Price($)",
    main = "Log of Sale price of vs. Size of first floor")
abline(log.slr.no)
hist(x.no, xlab="Size of First Floor(Sq. ft.)", main="Size of First Floor")
plot(log.slr.no$fitted,rstandard(log.slr.no),xlab="Fitted Values",
    ylab="Standardized Residuals", main="Residual plot")
abline(h=0)
qqnorm(rstandard(log.slr.no))
qqline(rstandard(log.slr.no))
```