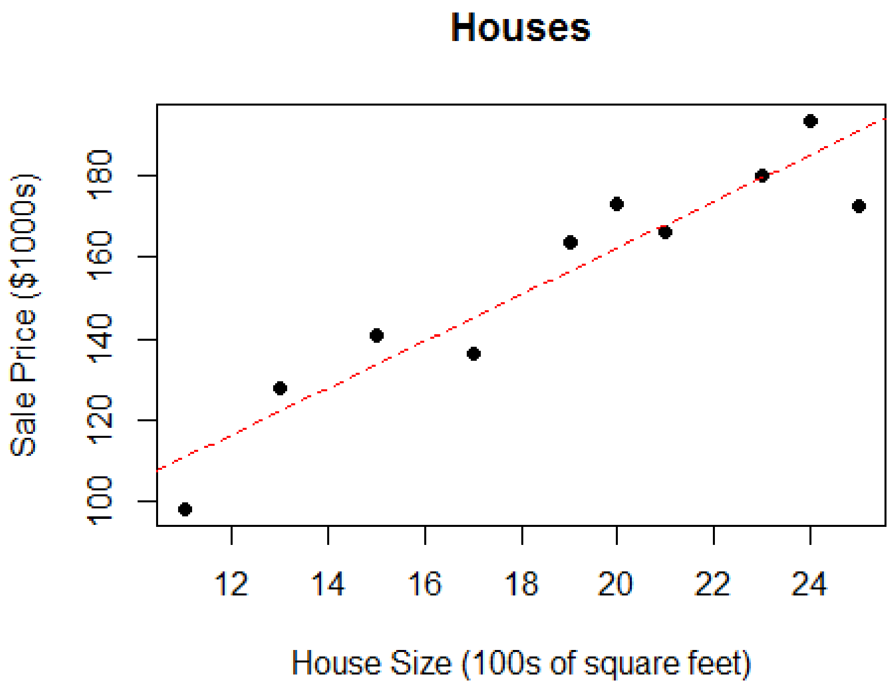
Regression_1Basics.R

Fitting Section

```
# A real estate agency collects data concerning
#   house sales prices ($1000s) and house sizes (100s of square feet).
#install readxl package first
library(readxl)
houses<-read_excel("Houses.xlsx", na="NA", col_names = TRUE)

# scatter plot w/ fitted linear regression line
plot(houses$Size, houses$Price, pch = 16, main = "Houses", xlab = "House Size
(100s of square feet)", ylab = "Sale Price ($1000s)")
abline(lm(houses$Price ~ houses$Size), lty=2, col="red")
```



```
# fit the model
linefit1 <- lm(houses$Price ~ houses$Size)
# linefit1 stores information that can be accessed
# to see an information summary of the fitted model
summary(linefit1)
```

```
##
## Call:
## lm(formula = houses$Price ~ houses$Size)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.032  -6.780   3.270   7.396  11.070
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  48.0244    14.4135   3.332   0.0104 *
## houses$Size   5.7003     0.7457   7.644 6.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.59 on 8 degrees of freedom
## Multiple R-squared:  0.8796, Adjusted R-squared:  0.8645
## F-statistic: 58.43 on 1 and 8 DF,  p-value: 6.05e-05
```

```
# there are also functions for seeing specific features, e.g.,
  # to see the coefficients Beta-hats
coefficients(linefit1)
```

```
## (Intercept) houses$Size
##   48.024405    5.700298
```

```
  # to see the coefficient of determination R-squared
summary(linefit1)$r.squared
```

```
## [1] 0.8795784
```

```
  # correlation (use positive value since beta-hat-1 > 0)
sqrt(summary(linefit1)$r.squared)
```

```
## [1] 0.9378584
```

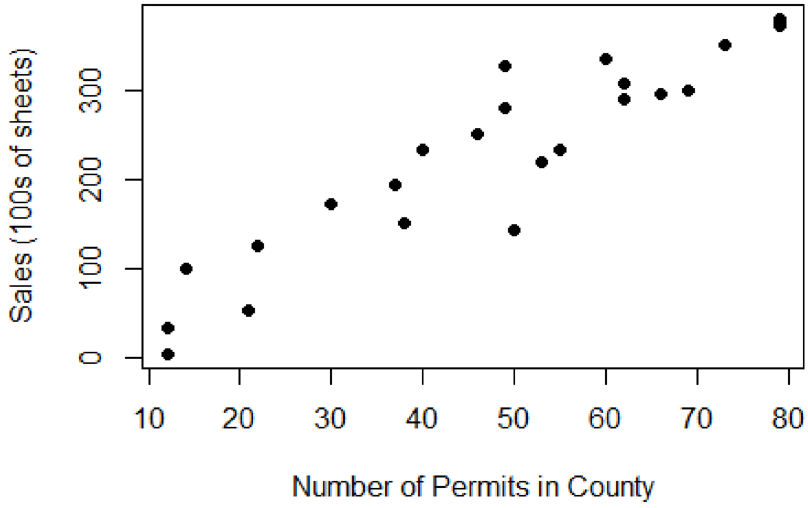
```
  # the observed residuals, epsilon-hats
resids <- residuals(linefit1)
resids
```

```
##          1          2          3          4          5          6
##   0.868750 -12.627679  11.069643  -8.429464   7.471131  -1.830655
##          7          8          9         10
##   8.668452   5.671726   7.169940 -18.031845
```

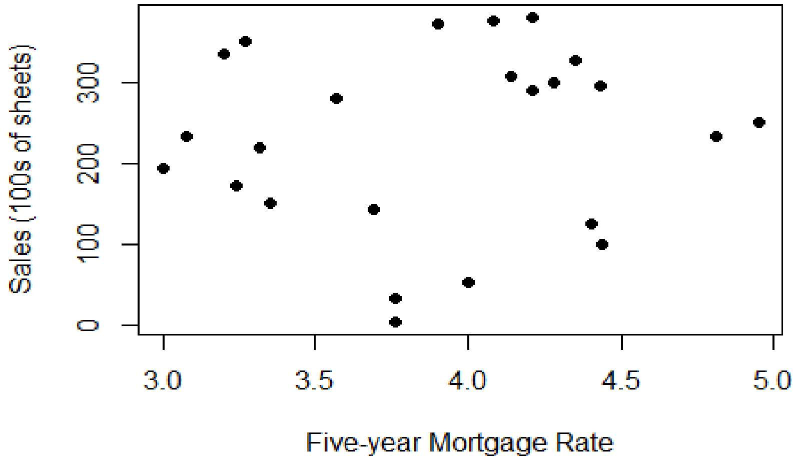
```
  # standard deviation of the residuals = sqrt(MSE)
summary(linefit1)$sigma
```

```
## [1] 10.58797
```

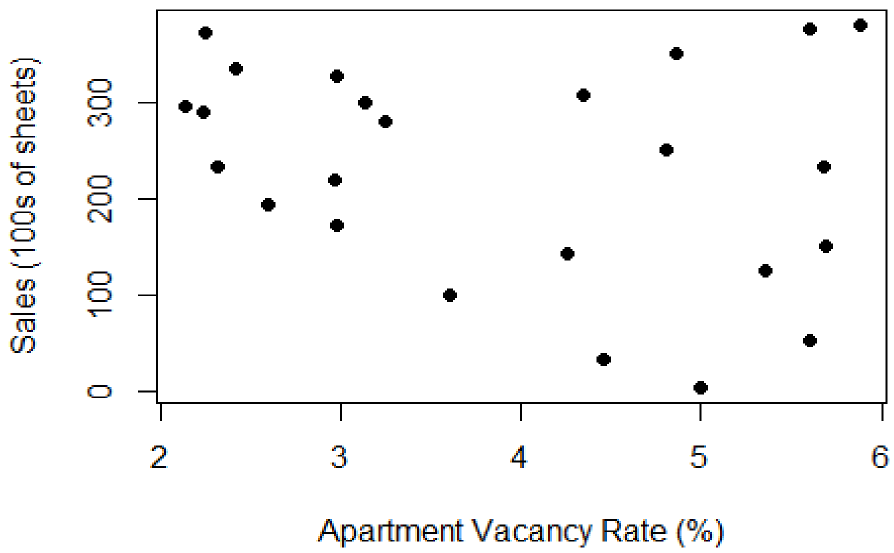
```
# Multiple regression
  # Data file: see file for documentation
drywall<-read_excel("Drywall.xlsx", na="NA", col_names = TRUE)
  # Shorthand to allow referring to dataframe columns without stating the dat
aframe name
attach(drywall)

  # scatter plots: Y vs. each X
plot(Permits, Sales, pch = 16, xlab = "Number of Permits in County", ylab = "
Sales (100s of sheets)")
```



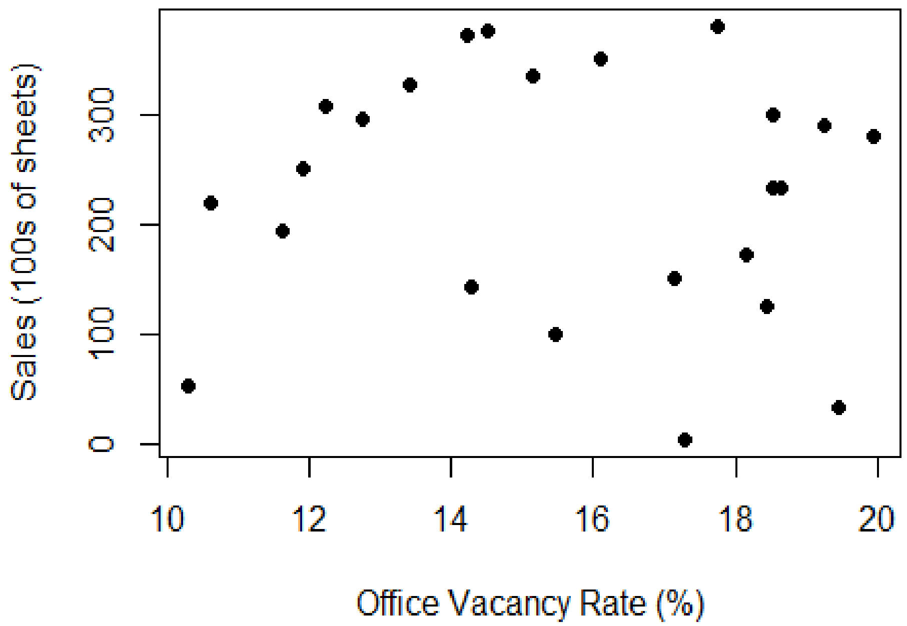
```
plot(Mortgage, Sales, pch = 16, xlab = "Five-year Mortgage Rate", ylab = "Sal
es (100s of sheets)")
```



```
plot(A_Vacancy, Sales, pch = 16, xlab = "Apartment Vacancy Rate (%)", ylab =
"Sales (100s of sheets)")
```



```
plot(O_Vacancy, Sales, pch = 16, xlab = "Office Vacancy Rate (%)", ylab = "Sa
les (100s of sheets)")
```



```
  # fit the model
linefit4 <- lm(Sales ~ Permits + Mortgage + A_Vacancy + O_Vacancy)

  # information summary of the fitted model
summary(linefit4)
```

```
## 
## Call:
## lm(formula = Sales ~ Permits + Mortgage + A_Vacancy + O_Vacancy)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -86.822 -25.351   9.409  22.602  78.391 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -43.873     83.213  -0.527    0.604    
## Permits        4.763      0.395  12.057 2.39e-10 ***
## Mortgage      16.988     15.159   1.121    0.276    
## A_Vacancy    -10.528      6.394  -1.646    0.116    
## O_Vacancy      1.308      2.791   0.469    0.645    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 40.13 on 19 degrees of freedom
## Multiple R-squared:  0.8935, Adjusted R-squared:  0.8711 
## F-statistic: 39.86 on 4 and 19 DF,  p-value: 5.448e-09
```

```
  # the observed residuals, epsilon-hats
residsMR <- residuals(linefit4)
residsMR
```

```
##          1          2          3          4          5          6 
##  78.390864  14.256410 -20.600557 -86.822107  22.822671 -27.580036 
##          7          8          9         10         11         12 
##  22.528809 -47.759222 -24.607992  27.740794  15.757952  26.559573 
##         13         14         15         16         17         18 
##  -4.549211  21.712180  -1.870581 -20.658747 -34.671535 -42.122439 
##         19         20         21         22         23         24 
##  44.388902  37.969736  20.533014 -42.999311  17.018283   4.562549
```

```
  # standard deviation of the residuals = sqrt(MSE)
summary(linefit4)$sigma
```

```
## [1] 40.13239
```

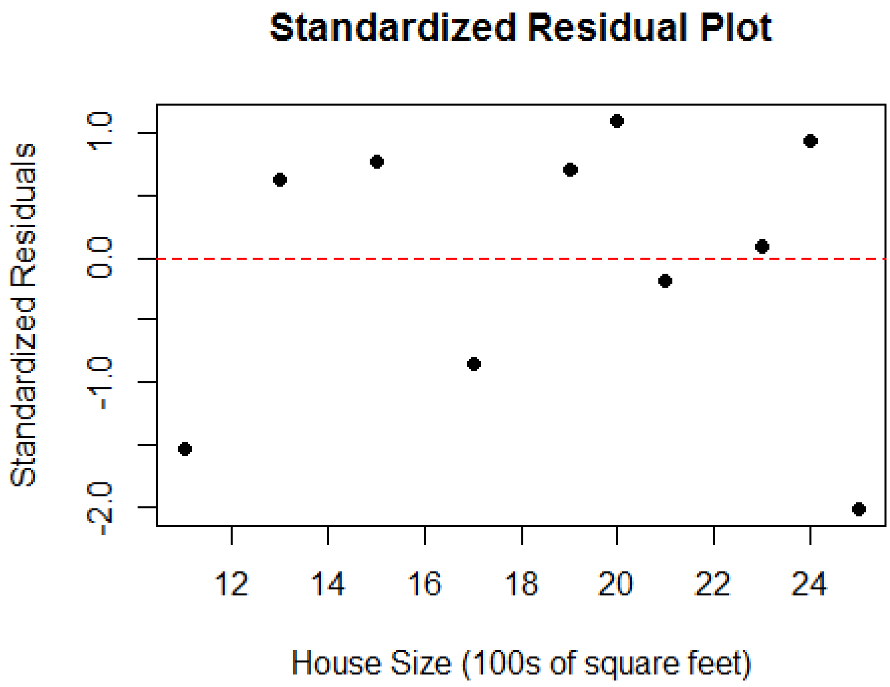
```
  # clean up
detach(drywall)
```

Inference Section

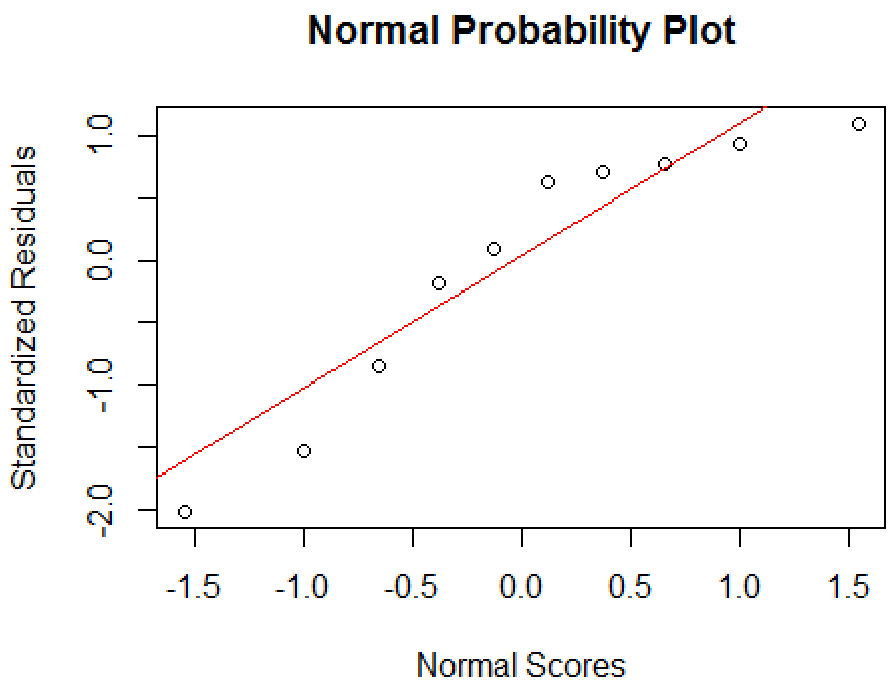
```
# A real estate agency collects data concerning
# house sales prices ($1000s) and house sizes (100s of square feet).
#install readxl package first
library(readxl)
houses<-read_excel("Houses.xlsx", na="NA", col_names = TRUE)

# fit the linear regression model
linefit1 <- lm(houses$Price ~ houses$Size)

  # confidence intervals for Beta-i
confint(linefit1, level = .90)
##                   5 %      95 %
## (Intercept) 21.221720 74.827090
## houses$Size  4.313622  7.086973

confint(linefit1, level = .75)
##                12.5 %   87.5 %
## (Intercept) 30.147018 65.90179
## houses$Size  4.775385  6.62521

  # two-tailed hypothesis tests for H0: Beta-i = 0
  # along with other summary information
summary(linefit1)

## Call:
## lm(formula = houses$Price ~ houses$Size)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.032  -6.780   3.270   7.396  11.070
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  48.0244    14.4135   3.332   0.0104 *
## houses$Size   5.7003     0.7457   7.644 6.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.59 on 8 degrees of freedom
## Multiple R-squared:  0.8796, Adjusted R-squared:  0.8645
## F-statistic: 58.43 on 1 and 8 DF,  p-value: 6.05e-05

  # ANOVA data for simple linear regression
anova(linefit1)

## Analysis of Variance Table
## Response: houses$Price
##             Df Sum Sq Mean Sq F value   Pr(>F)
## houses$Size  1 6550.7  6550.7  58.433 6.05e-05 ***
## Residuals    8  896.8   112.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Assumptions Section

```r
# A real estate agency collects data concerning
# house sales prices ($1000s) and house sizes (100s of square feet).
#install readxl package first
library(readxl)
houses<-read_excel("Houses.xlsx", na="NA", col_names = TRUE)

# fit the model
linefit1 <- lm(houses$Price ~ houses$Size)

# standardized residual plot
linefit1.stres <- rstandard(linefit1)
plot(houses$Size, linefit1.stres, pch = 16, main = "Standardized Residual Plot", xlab = "House Size (100s of square feet)", ylab = "Standardized Residuals")
abline(0,0, lty=2, col="red")
```



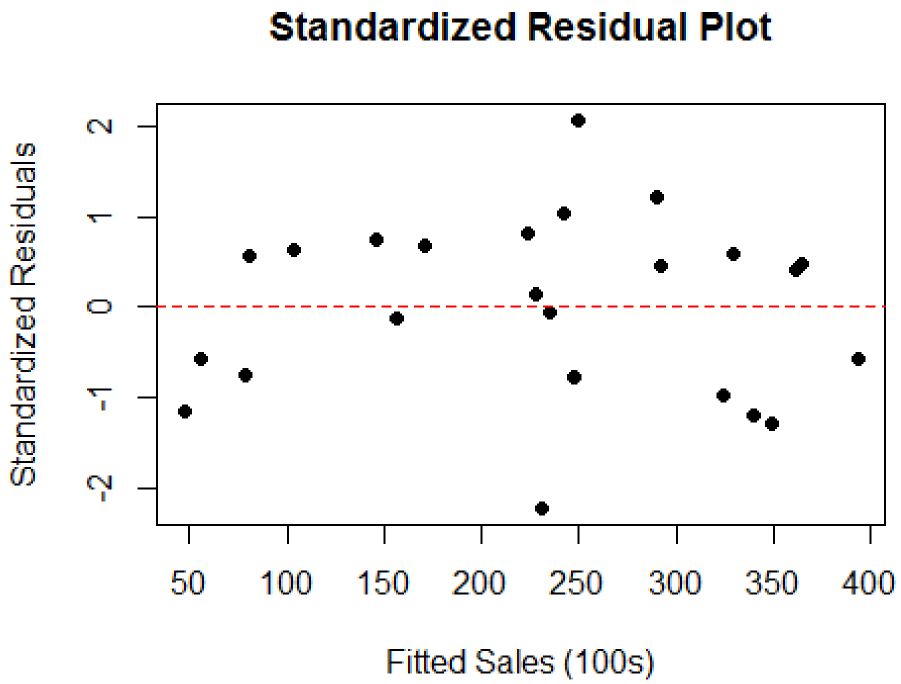
```
# normal probability plot
qqnorm(linefit1.stres, main = "Normal Probability Plot", xlab = "Normal Score
s", ylab = "Standardized Residuals")
qqline(linefit1.stres, col = "red")
```



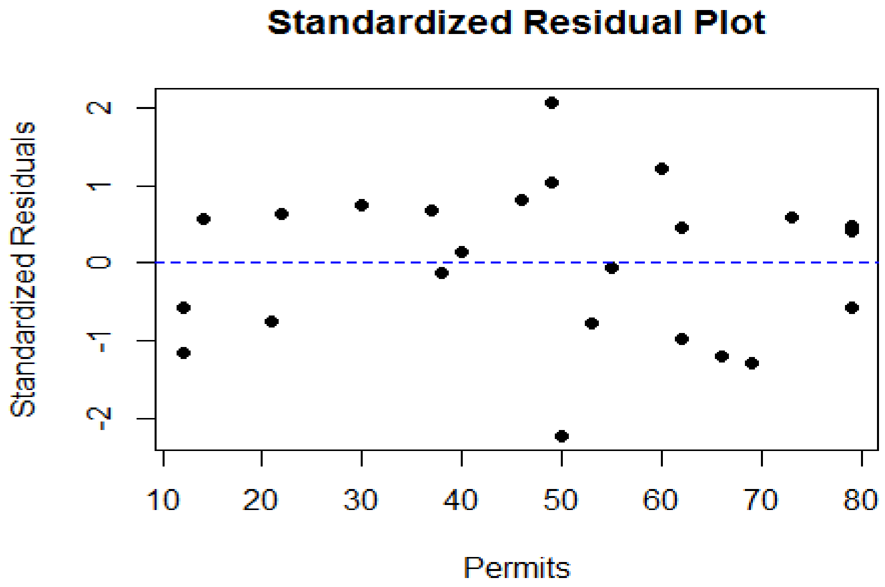
```
shapiro.test(linefit1.stres)

##
##  Shapiro-Wilk normality test
##
## data:  linefit1.stres
## W = 0.88246, p-value = 0.1393
```

```r
# Multiple regression
# Data file: see file for documentation
drywall<-read_excel("Drywall.xlsx", na="NA", col_names = TRUE)
attach(drywall)

# fit the model
linefit4 <- lm(Sales ~ Permits + Mortgage + A_Vacancy + O_Vacancy)
```

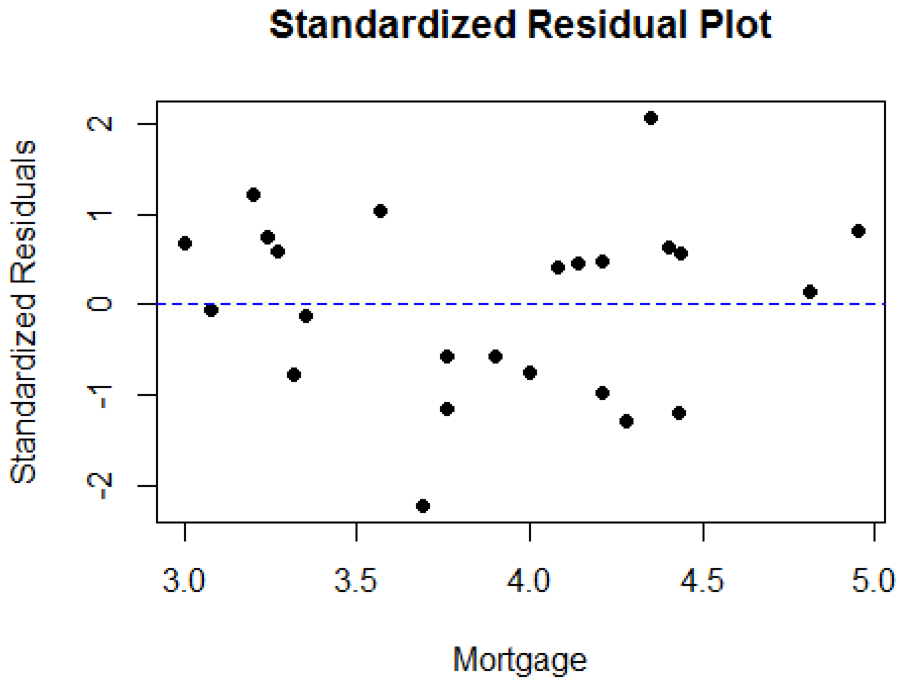
```r
# standardized residual plot - on fitted values
linefit4.stres <- rstandard(linefit4)
plot(linefit4$fitted.values, linefit4.stres, pch = 16, main = "Standardized Residual Plot", xlab = "Fitted Sales (100s)", ylab = "Standardized Residuals")
abline(0,0, lty=2, col="red")
```



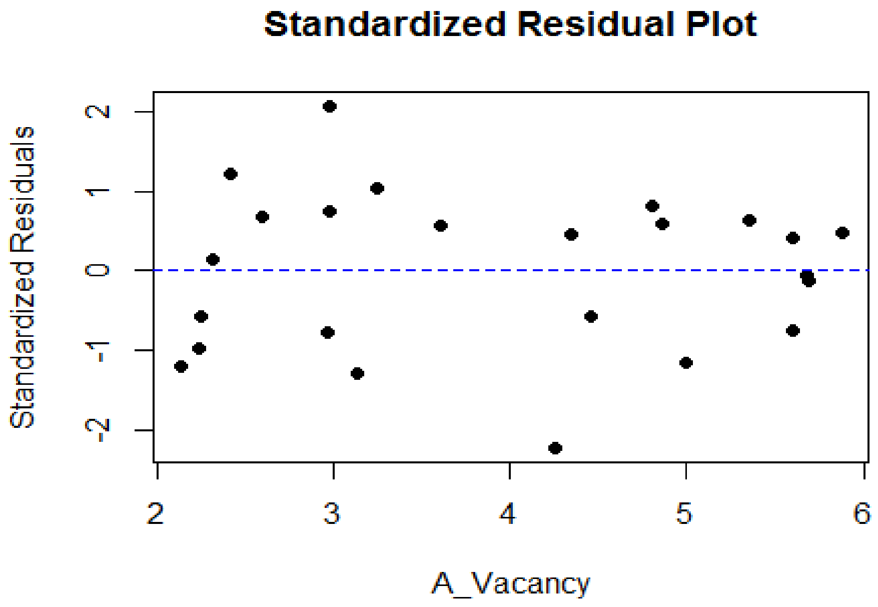
```
# standardized residual plot - on Permits
plot(Permits, linefit4.stres, pch = 16, main = "Standardized Residual Plot",
xlab = "Permits", ylab = "Standardized Residuals")
abline(0,0, lty=2, col="blue")
```



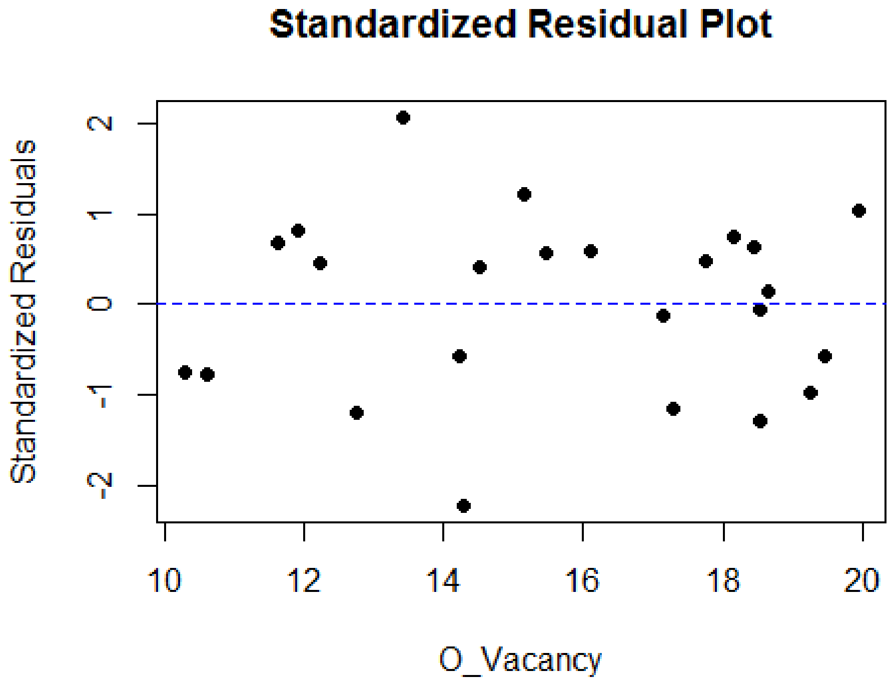
```
# standardized residual plot - on Mortgage
plot(Mortgage, linefit4.stres, pch = 16, main = "Standardized Residual Plot",
xlab = "Mortgage", ylab = "Standardized Residuals")
abline(0,0, lty=2, col="blue")
```



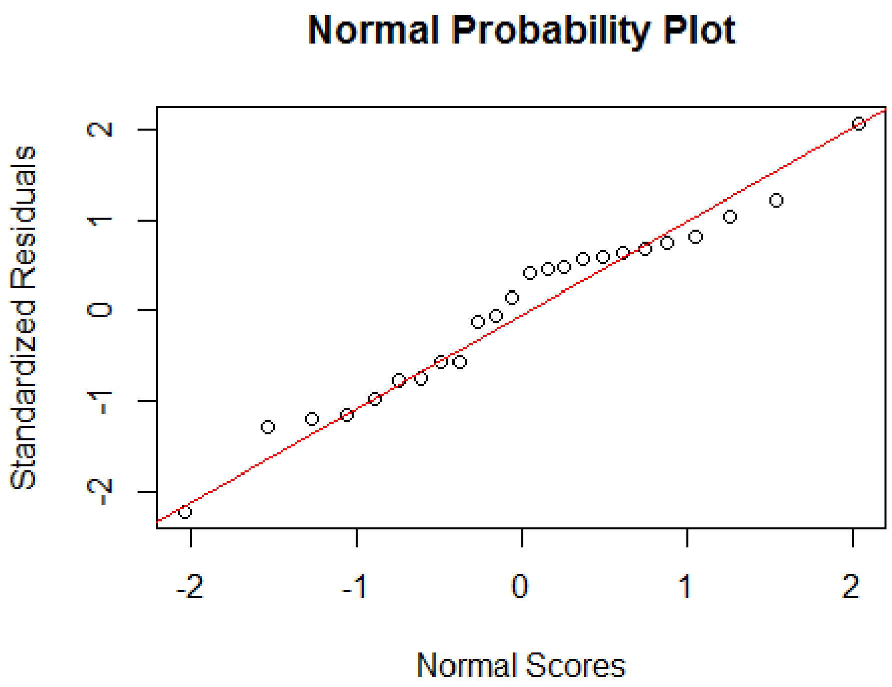
```r
# standardized residual plot - on A_Vacancy
plot(A_Vacancy, linefit4.stres, pch = 16, main = "Standardized Residual Plot"
, xlab = "A_Vacancy", ylab = "Standardized Residuals")
abline(0,0, lty=2, col="blue")
```



```r
# standardized residual plot - on O_Vacancy
plot(O_Vacancy, linefit4.stres, pch = 16, main = "Standardized Residual Plot"
, xlab = "O_Vacancy", ylab = "Standardized Residuals")
abline(0,0, lty=2, col="blue")
```



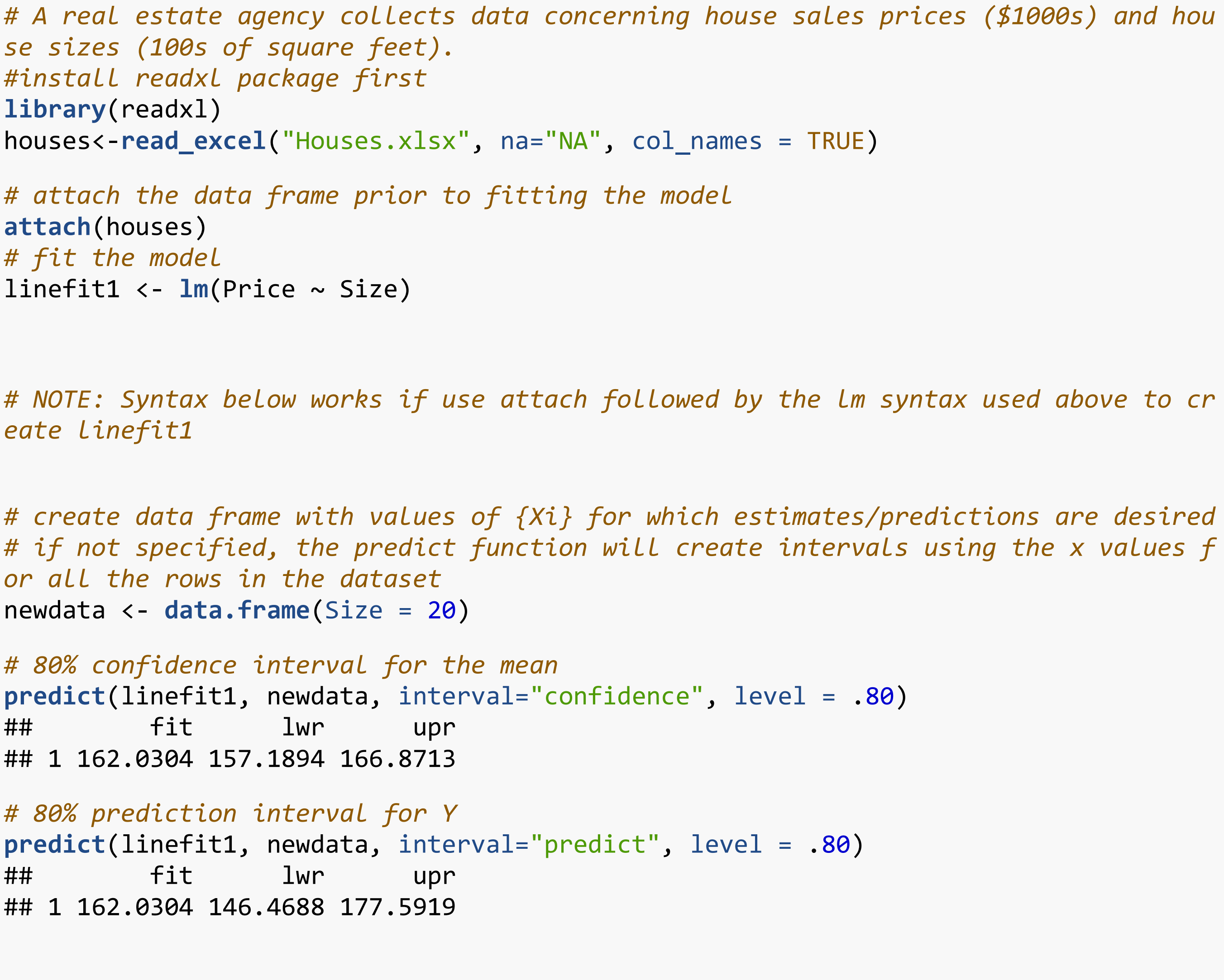
```
# normal probability plot
qqnorm(linefit4.stres, main = "Normal Probability Plot", xlab = "Normal Score
s", ylab = "Standardized Residuals")
qqline(linefit4.stres, col = "red")
```



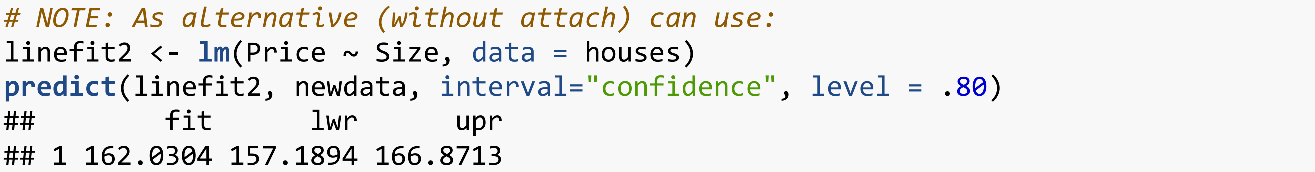
```
shapiro.test(linefit4.stres)

##
##  Shapiro-Wilk normality test
##
## data:  linefit4.stres
## W = 0.9687, p-value = 0.6351
```

```
detach(drywall)
```

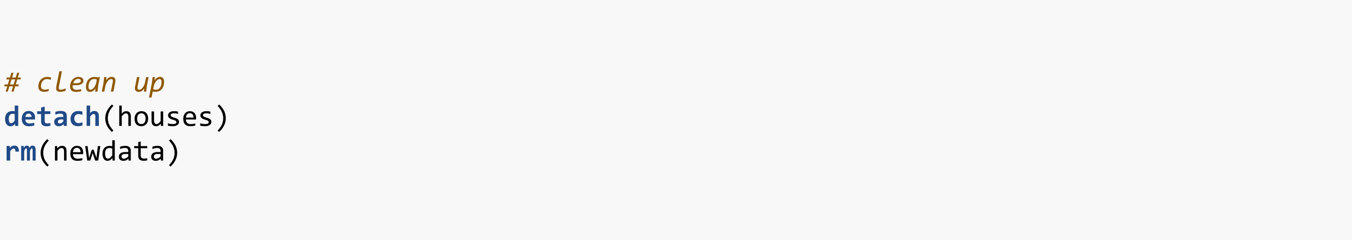
Using Section



```
# A real estate agency collects data concerning house sales prices ($1000s) and hou
se sizes (100s of square feet).
#install readxl package first
library(readxl)
houses<-read_excel("Houses.xlsx", na="NA", col_names = TRUE)

# attach the data frame prior to fitting the model
attach(houses)
# fit the model
linefit1 <- lm(Price ~ Size)



# NOTE: Syntax below works if use attach followed by the lm syntax used above to cr
eate linefit1


# create data frame with values of {Xi} for which estimates/predictions are desired
# if not specified, the predict function will create intervals using the x values f
or all the rows in the dataset
newdata <- data.frame(Size = 20)

# 80% confidence interval for the mean
predict(linefit1, newdata, interval="confidence", level = .80)
##        fit      lwr      upr
## 1 162.0304 157.1894 166.8713

# 80% prediction interval for Y
predict(linefit1, newdata, interval="predict", level = .80)
##        fit      lwr      upr
## 1 162.0304 146.4688 177.5919
```



```
# NOTE: As alternative (without attach) can use:
linefit2 <- lm(Price ~ Size, data = houses)
predict(linefit2, newdata, interval="confidence", level = .80)
##        fit      lwr      upr
## 1 162.0304 157.1894 166.8713
```



```
# clean up
detach(houses)
rm(newdata)
```

```r
# Multiple regression
  # Data file: see file for documentation
drywall<-read_excel("Drywall.xlsx", na="NA", col_names = TRUE)
attach(drywall)

  # fit the model
linefit4 <- lm(Sales ~ Permits + Mortgage + A_Vacancy + O_Vacancy)
  # create data frame with desired values of {Xi} for the inference
newdata <- data.frame(Permits = 250, Mortgage = 4, A_Vacancy = 3.5, O_Vacancy = 14)

  # 90% confidence interval for the mean
predict(linefit4, newdata, interval="confidence", level = .90)
##        fit      lwr      upr
## 1 1196.308 1058.854 1333.762

  # 90% prediction interval for Y
predict(linefit4, newdata, interval="predict", level = .90)
##        fit     lwr      upr
## 1 1196.308 1042.33 1350.285

  # clean up
detach(drywall)
```