Lab 3

Adam Shen

September 30, 2020

Packages

Load Packages

```r
library(car)
library(nortest)
```

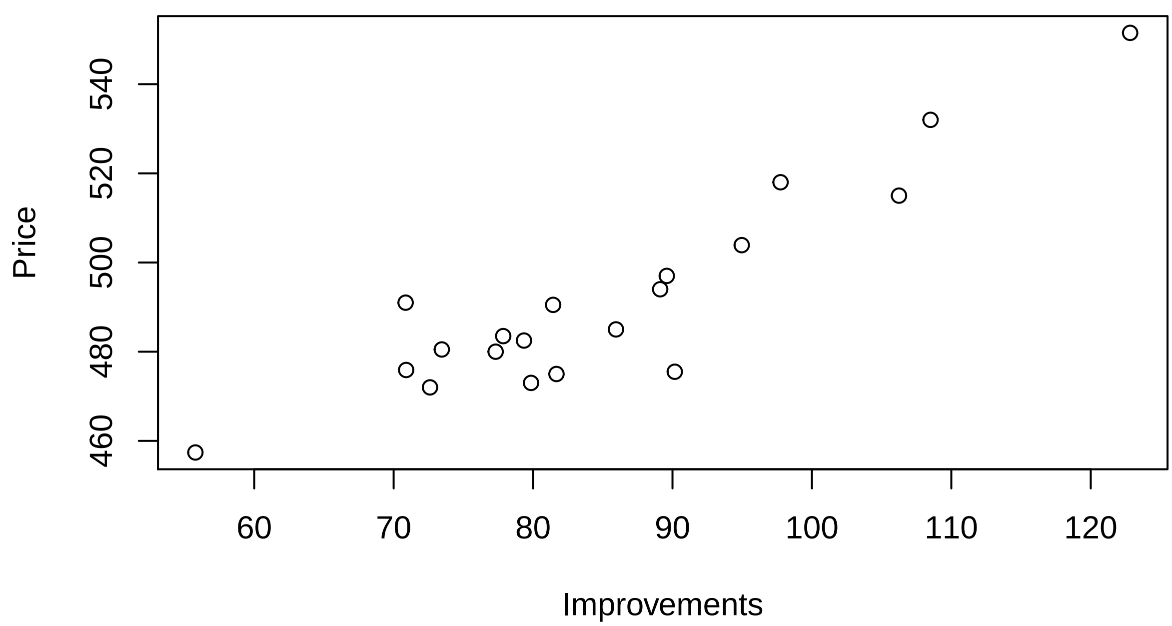
Housing data

Load the data

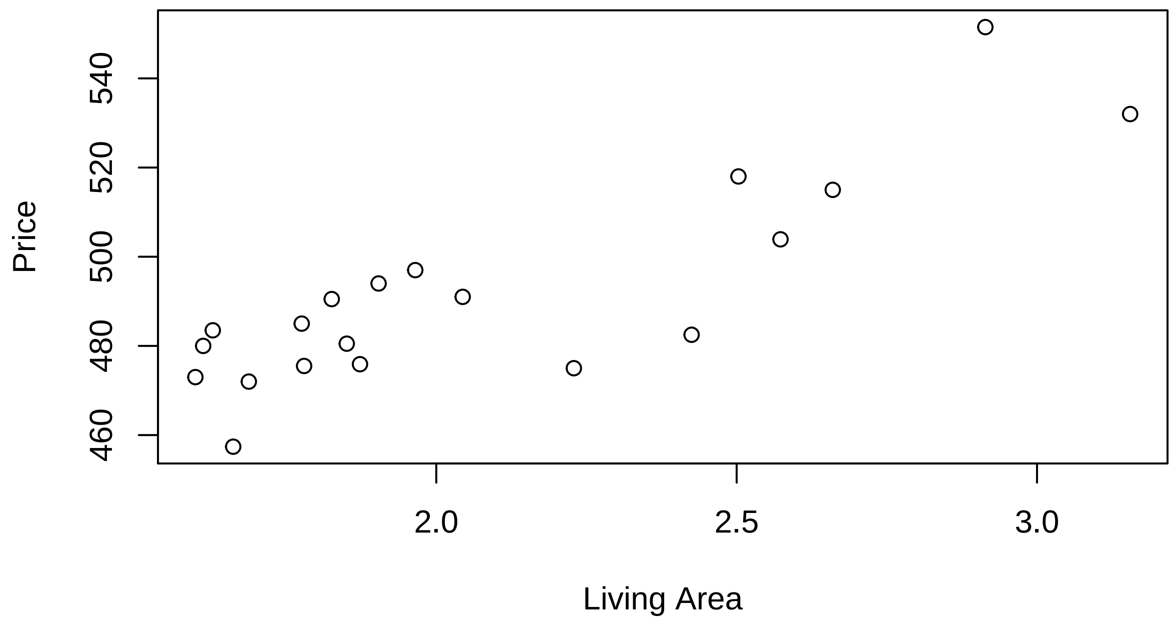
```r
homes <- read.table("./house.txt", header=TRUE)
```

Visualization

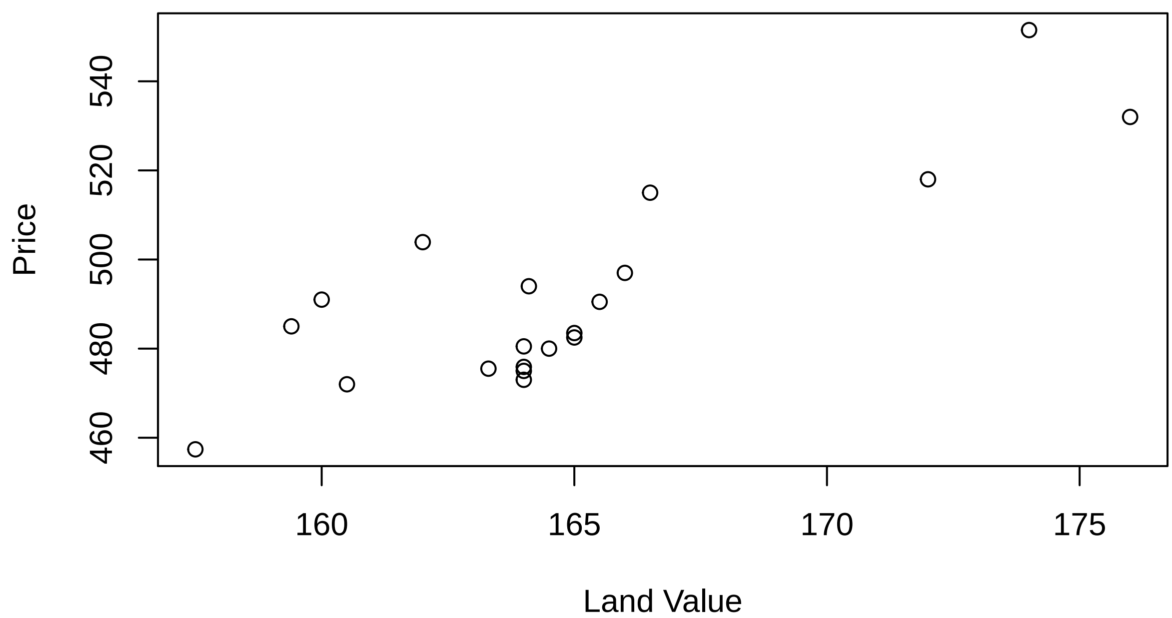
```r
with(homes, plot(x=Improve, y=Price, xlab="Improvements", ylab="Price"))
```



```r
with(homes, plot(x=Area, y=Price, xlab="Living Area", ylab="Price"))
```




```r
with(homes, plot(x=Land, y=Price, xlab="Land Value", ylab="Price"))
```



Fit a multiple linear regression model

```r
# Use `x=TRUE` to store the design matrix for later use
model <- lm(Price ~ Improve + Area + Land, x=TRUE, data=homes)
summary(model)
```

```
##
## Call:
## lm(formula = Price ~ Improve + Area + Land, data = homes, x = TRUE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.856  -2.897   1.797   2.783  16.246
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 229.5069    97.9279   2.344  0.03234 *
## Improve       0.7932     0.2232   3.553  0.00265 **
## Area         13.3934     6.6878   2.003  0.06246 .
## Land          1.0104     0.6735   1.500  0.15299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.979 on 16 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8763
## F-statistic: 45.88 on 3 and 16 DF,  p-value: 4.397e-08
```

Take note of this F-value and its associated p-value!

Model usefulness

```r
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## Improve    1 8202.5  8202.5 128.8261 4.589e-09 ***
## Area       1  418.5   418.5   6.5736   0.02081 *
## Land       1  143.3   143.3   2.2511   0.15299
## Residuals 16 1018.7    63.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You may recall from simple linear regression that the F-value and its associated p-value from `summary()` was the same as the one that resulted from `anova()`. We see here that the F-value and its associated p-value from `summary()` are not the ones displayed here! This ANOVA table is actually something else entirely, which you will learn in next week's set of lectures.

In order to compare our current model to the null model, we will need to actually construct the null model and run `anova()` with the null model and current model.

```r
null_model <- lm(Price ~ 1, data=homes)
anova(null_model, model)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ 1
## Model 2: Price ~ Improve + Area + Land
```

```
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     19 9783.2
## 2     16 1018.7  3    8764.4 45.884 4.397e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For multiple linear regression, the hypotheses associated with this output are:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{vs} \quad H_A : \text{At least one } \beta_i \text{ non-zero, } i = 1, 2, 3$$

Since the p-value is less than 0.05, we reject the null hypothesis and conclude that the model is useful.

Store fitted values and residuals for later

```
fits <- fitted(model)
res <- resid(model)
```

Computing the covariance matrix from scratch

```
X <- model$x
XTXinv <- solve(t(X) %*% X)
MSE <- sum(res^2)/model$df.residual
(Vhatb <- MSE*XTXinv)
```

```
##             (Intercept)     Improve         Area         Land
## (Intercept) 9589.875334  9.68984296 191.4994500 -65.58278131
## Improve        9.689843  0.04983480  -0.7521917  -0.07505678
## Area         191.499450 -0.75219172  44.7260869  -1.33751305
## Land         -65.582781 -0.07505678  -1.3375131   0.45353762
```

Computing the covariance matrix with built-in function

```
vcov(model)
```

```
##             (Intercept)     Improve         Area         Land
## (Intercept) 9589.875334  9.68984296 191.4994500 -65.58278131
## Improve        9.689843  0.04983480  -0.7521917  -0.07505678
## Area         191.499450 -0.75219172  44.7260869  -1.33751305
## Land         -65.582781 -0.07505678  -1.3375131   0.45353762
```

```
all.equal(Vhatb, vcov(model))
```

```
## [1] TRUE
```

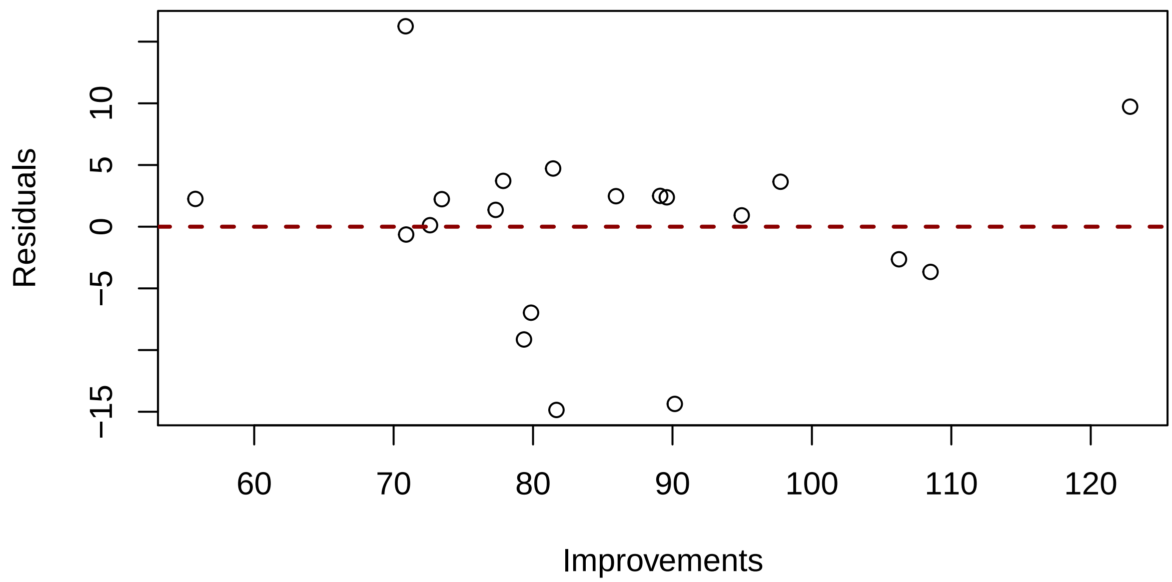
Note that we can get the standard errors of the regression coefficients by taking the square root of the diagonal of the covariance matrix.

```
sqrt(diag(vcov(model)))
```

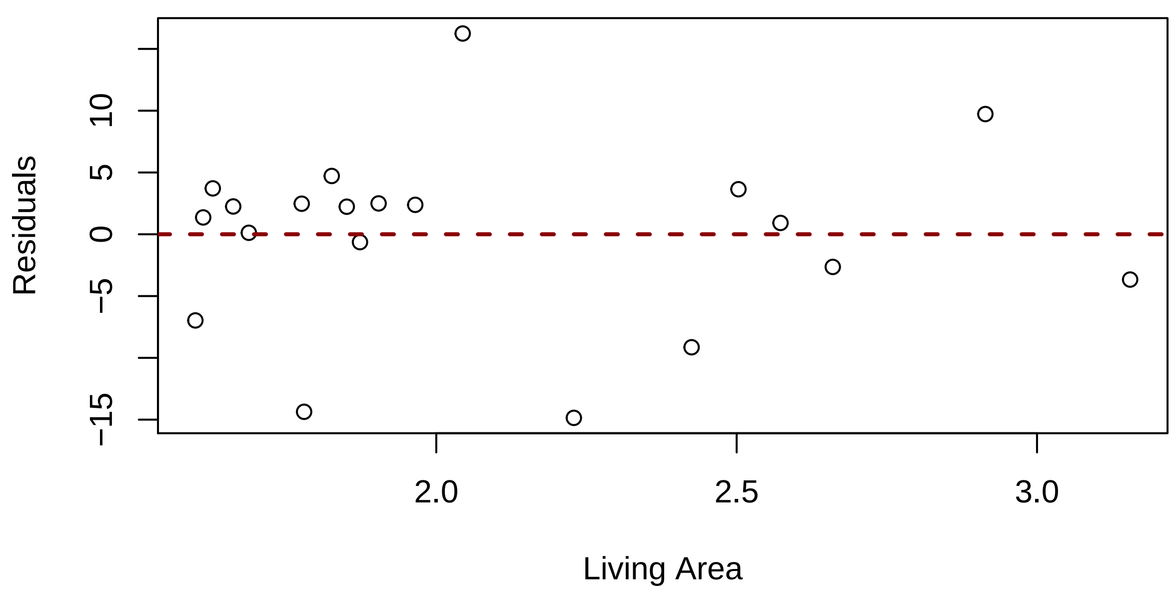
```
## (Intercept)     Improve        Area        Land
##  97.9279089   0.2232371   6.6877565   0.6734520
```

Diagnostics

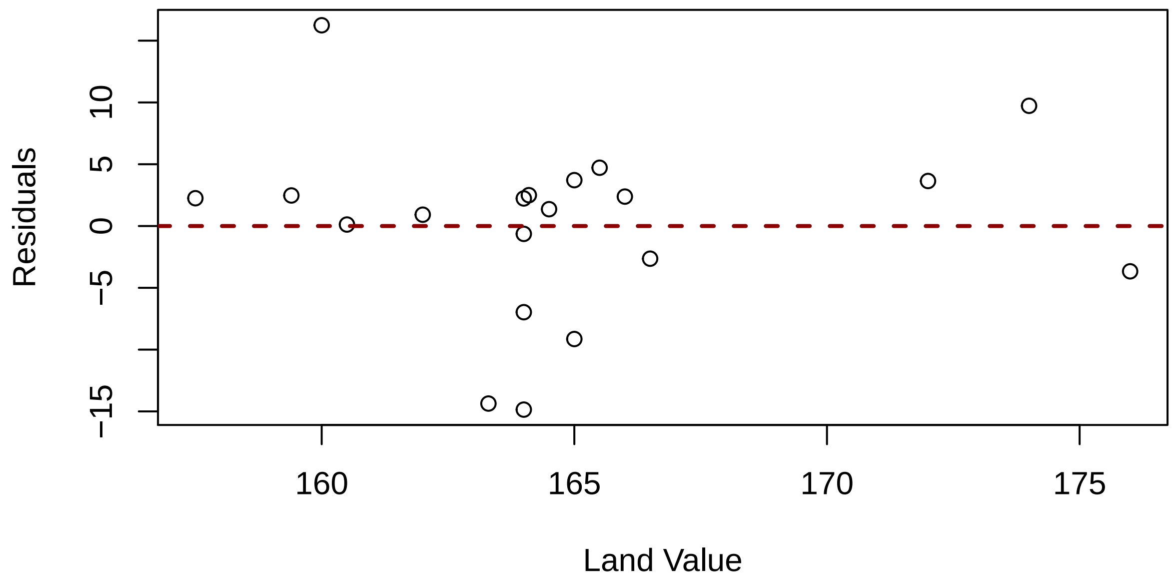
```
with(homes, plot(x=Improve, y=res, xlab="Improvements", ylab="Residuals"))
abline(h=0, col="darkred", lty=2, lwd=2)
```



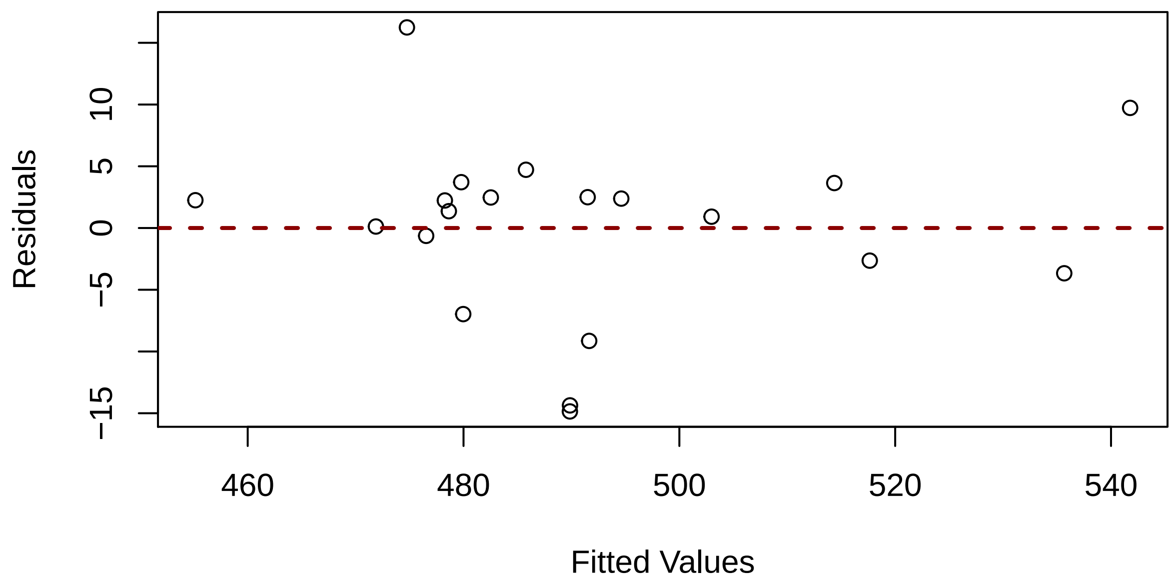

```
with(homes, plot(x=Area, y=res, xlab="Living Area", ylab="Residuals"))
abline(h=0, col="darkred", lty=2, lwd=2)
```



```r
with(homes, plot(x=Land, y=res, xlab="Land Value", ylab="Residuals"))
abline(h=0, col="darkred", lty=2, lwd=2)
```




```r
plot(x=fits, y=res, xlab="Fitted Values", ylab="Residuals")
abline(h=0, col="darkred", lty=2, lwd=2)
```



```r
fitsize <- factor(fits <= 485)
leveneTest(res, group=fitsize)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  1.3326 0.2634
##       18
```

```r
ncvTest(model)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.007676665, Df = 1, p = 0.93018
```

For both of the above tests, the corresponding hypotheses are:

$$H_0 : \text{The error term variance is constant}$$

$$H_A : \text{The error term variance is non-constant}$$

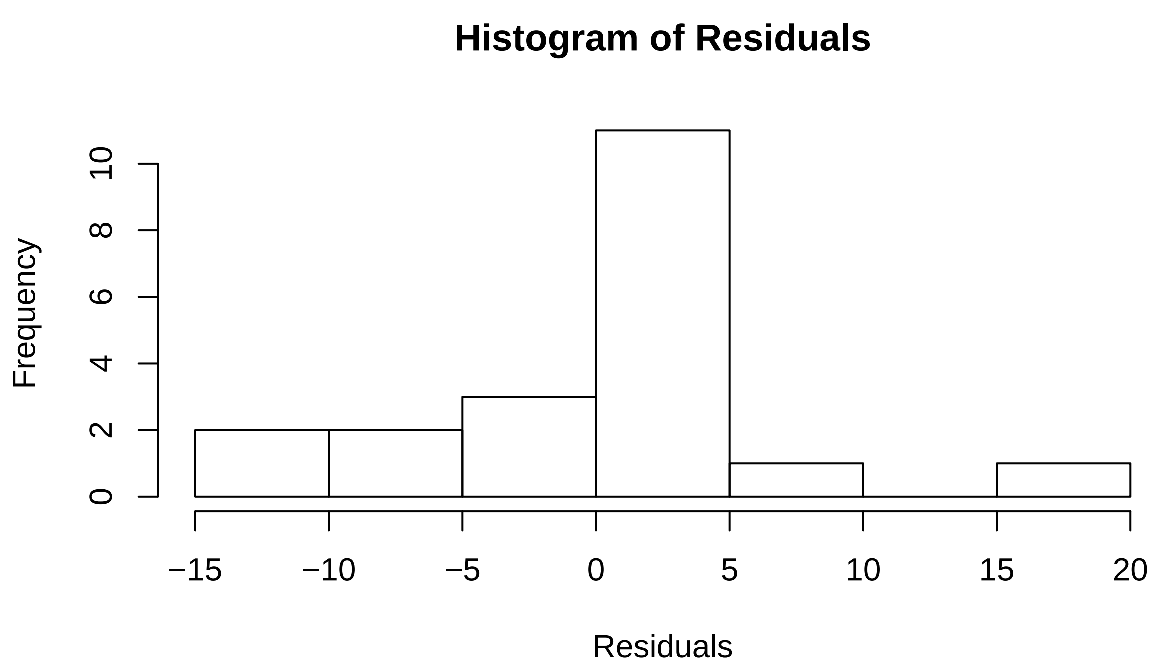
Using $\alpha = 0.10$ for both tests, we fail to reject the null hypothesis as the p-values are larger than 0.10. We conclude that there is insufficient evidence suggesting a non-constant error term variance.

Distribution of residuals

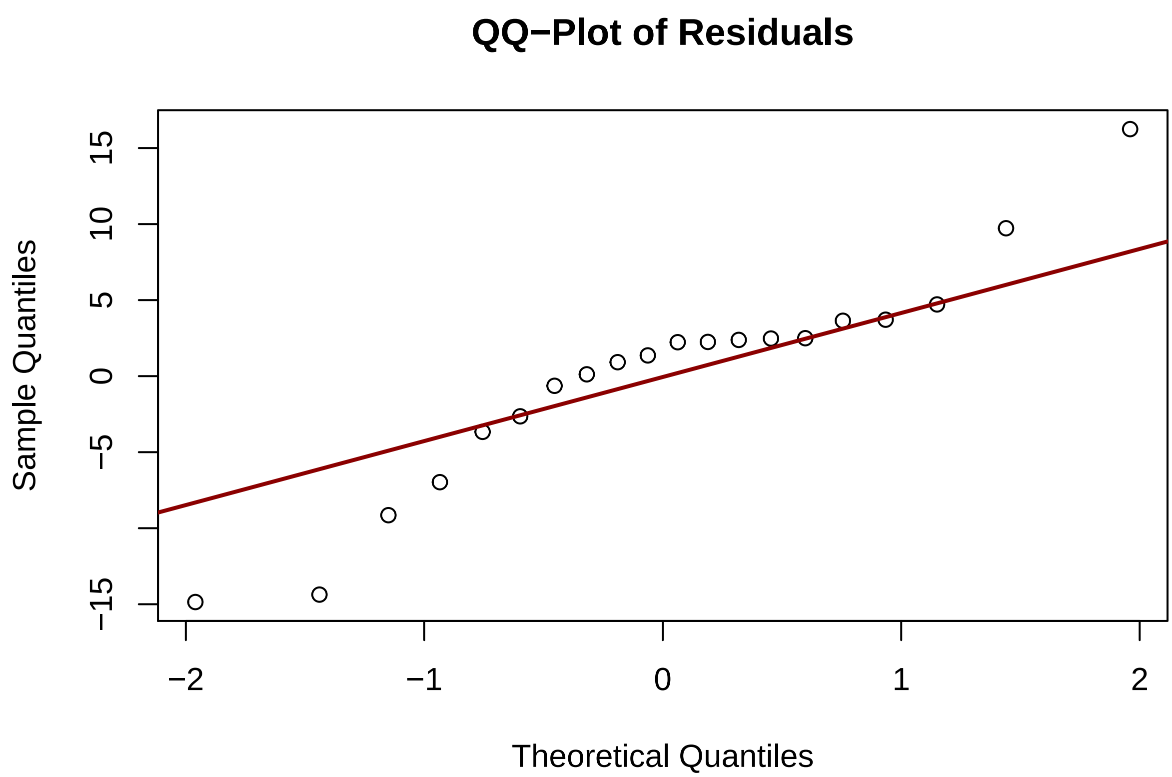
```r
stem(res)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   -1 | 54
##   -0 | 97431
##    0 | 01122222445
##    1 | 06
```

```r
hist(res, main="Histogram of Residuals", xlab="Residuals")
```



```r
qqnorm(res, main="QQ-Plot of Residuals")
qqline(res, col="darkred", lwd=2)
```



Possible deviation from normality assumption?

```r
ad.test(res)
```

```
##
##  Anderson-Darling normality test
##
## data:  res
## A = 0.775, p-value = 0.03654
```

```r
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.92562, p-value = 0.1271
```

```r
lillie.test(res)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  res
## D = 0.16541, p-value = 0.1615
```

The hypotheses associated with the above tests are:

$$H_0 : \text{The errors are normally distributed} \quad \text{vs} \quad H_A : \text{The errors are not normally distributed}$$

Using $\alpha = 0.10$, we reject the null hypothesis for the Anderson-Darling test and fail to reject the null hypothesis for the Shapiro-Wilk and Kolmogorov-Smirnov tests. We are probably nearing the violation of the normality assumption since the Shapiro-Wilk and Kolmogorov-Smirnov p-values are not far from 0.10 despite failing to reject the null hypothesis for these tests. Since two of our three tests are suggesting that the normality assumption is not implausible, we will go with this conclusion.

Making intervals

```r
predict(model, level=0.95, interval="confidence", se.fit=TRUE,
        newdata=data.frame(Improve=100, Area=2, Land=170))
```

```
## $fit
##        fit      lwr      upr
## 1 507.3829 498.4215 516.3443
##
## $se.fit
## [1] 4.22726
##
## $df
## [1] 16
##
## $residual.scale
## [1] 7.979439
```

```r
predict(model, level=0.95, interval="prediction", se.fit=TRUE,
        newdata=data.frame(Improve=100, Area=2, Land=170))
```

```
## $fit
##        fit      lwr      upr
## 1 507.3829 488.2401 526.5257
##
## $se.fit
## [1] 4.22726
##
## $df
## [1] 16
##
## $residual.scale
## [1] 7.979439
```