Stat 302
Statistical Software and Its Applications
Regression

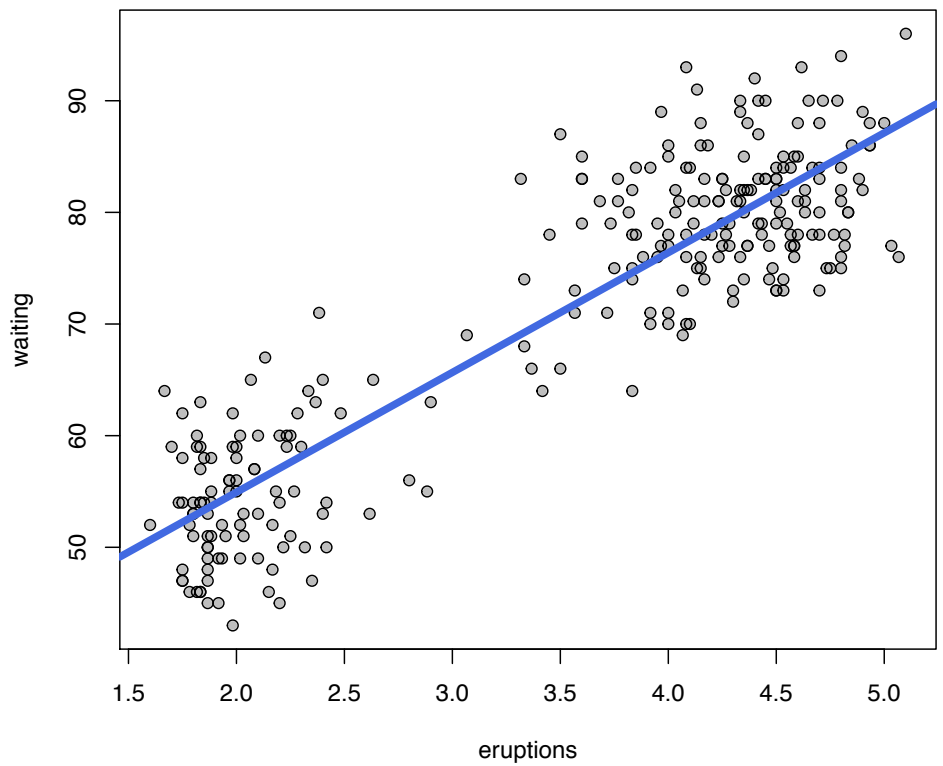
Yen-Chi Chen

Department of Statistics, University of Washington

Autumn 2016

An example: old faithful dataset

```
> plot(faithful, pch=20, col="gray")
> points(faithful)
> fit <- lm(waiting~eruptions, data = faithful)
> abline(fit, lwd=5, col="royalblue")
```



`lm()`: fitting variables in a dataset

```
> fit <- lm(waiting~eruptions, data = faithful)
> fit

Call:
lm(formula = waiting ~ eruptions, data = faithful)

Coefficients:
(Intercept)    eruptions
      33.47        10.73
```

- `lm()` is a function for linear regression.
- Here we apply it to fit variable `waiting` as Y and `eruptions` as X using the data `faithful`.
- Without `data = faithful`, we will get an error.

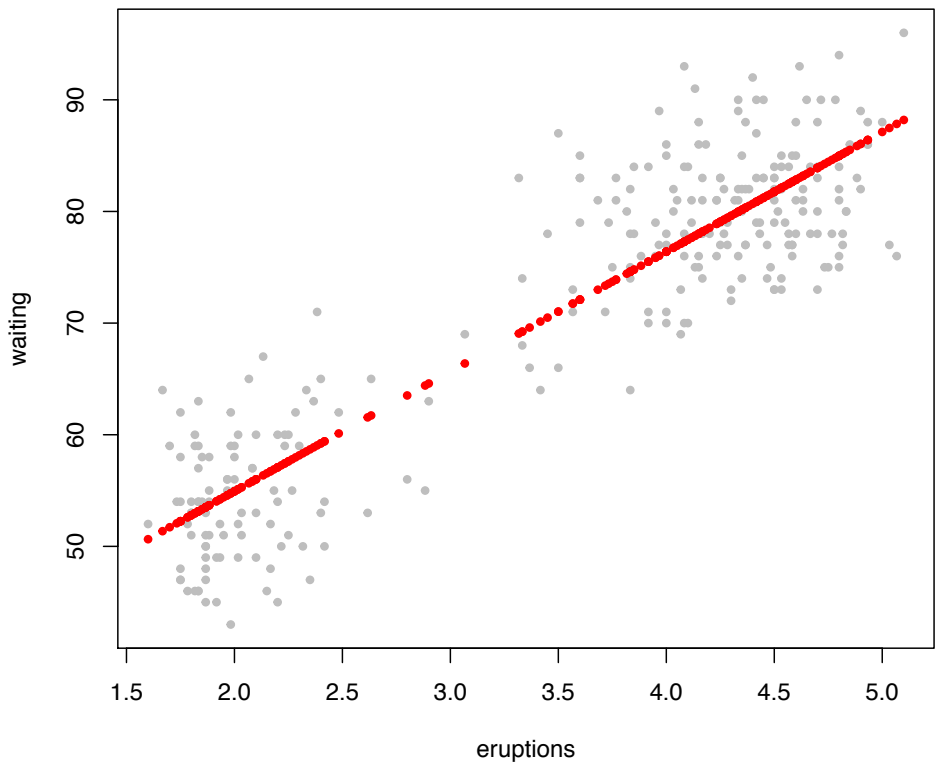
names(): checking components in an object

```
> names(fit)
 [1] "coefficients"  "residuals"      "effects"
 [4] "rank"          "fitted.values"  "assign"
 [7] "qr"            "df.residual"    "xlevels"
[10] "call"          "terms"          "model"
```

- `fit$coefficients`: fitted coefficients.
- `fit$residuals`: the residual of each observation under the fit.
- `fit$fitted.values`: the fitted value for each observation.
- More: check `help(lm)`

Fitted value at each x= eruptions

```
> plot(faithful, pch=20, col="gray")
> points(x=faithful$eruptions, y=fit$fitted.values,
+        col="red", pch=20)
```



Fitting a Line by Least Squares

Fitting a straight line model to $y = \texttt{waiting}$ and $x = \texttt{eruptions}$, i.e.,

$$y_i = \alpha + \beta x_i + \epsilon_i \quad \text{with } \epsilon_i \sim i.i.d.\ \mathcal{N}(0, \sigma^2)$$

where ϵ_i captures measurement error or the extent to which a line does not fit the data at the i^{th} data point.
α and β are found by the method of least squares, minimizing

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

over α and β. The solutions are

$$\hat{\beta} = \frac{SXY}{SXX} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad \text{and} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

with fitted values $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i = \bar{y} + \hat{\beta}(x_i - \bar{x})$ and $\bar{\hat{y}} = \bar{y}$.

Sum of Squares Decomposition

$$\begin{aligned} SYY = \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ &= RSS + SS_{reg} \end{aligned}$$

SS_{reg} = sum of squares of fitted values due to the regression.

RSS = residual sum of squares in addition to the regression.

Multiple Correlation Coefficient R^2

The multiple correlation coefficient or coefficient of determination

$$R^2 = \frac{SYY - RSS}{SYY} = 1 - \frac{RSS/(n-1)}{SYY/(n-1)} = \frac{SS_{reg}}{SYY}$$

$=$ proportion of Y variability explained by the regression on X.

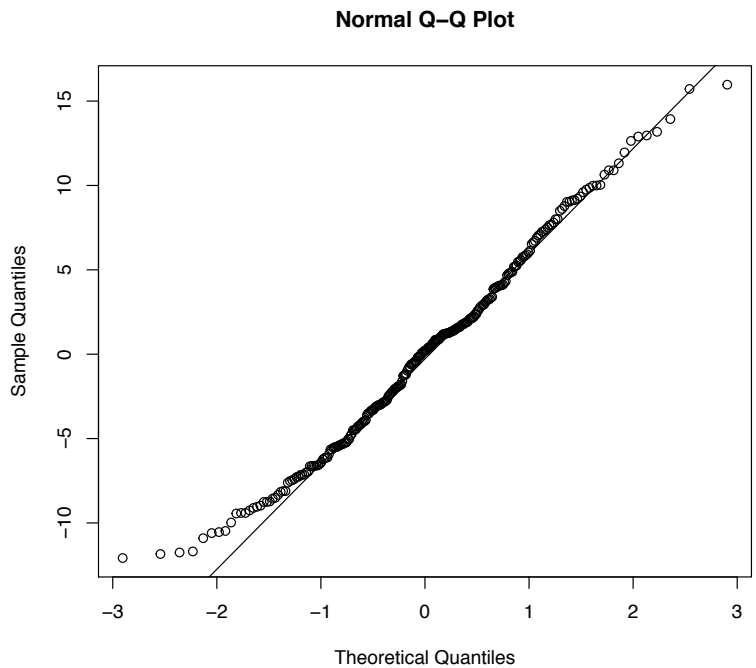
Note that R^2 is just the square of the Pearson correlation between x's and y's, i.e., $r = SXY/\sqrt{SXX \cdot SYY}$.

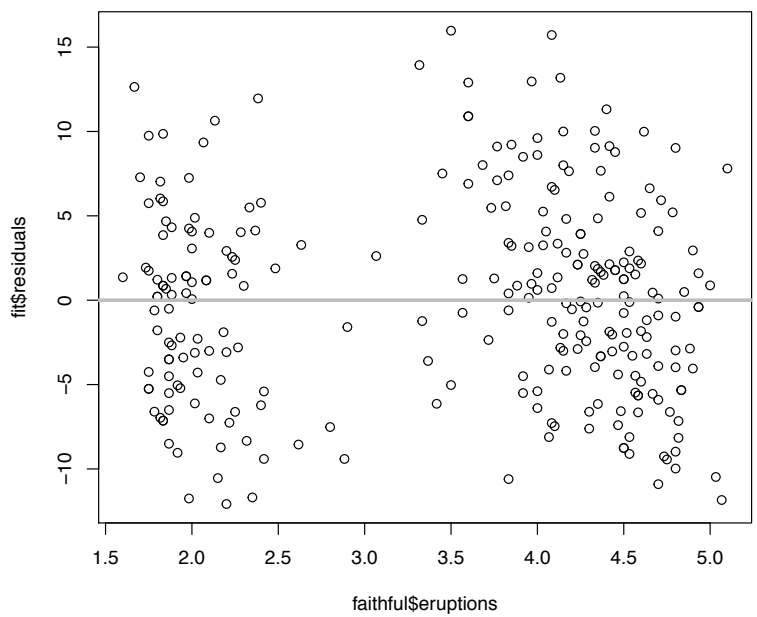
The adjusted $R^2 = \bar{R}^2$ uses $n-2$ in the RSS denominator above, i.e.,

$$\bar{R}^2 = 1 - \frac{RSS/(n-2)}{SYY/(n-1)}$$

Model checking for the linear regression

```
> qqnorm(fit$residuals)
> qqline(fit$residuals)
>   # looks like a normal
>
> plot(x=faithful$eruptions, y=fit$residuals)
> abline(h=0, lwd=3, col="gray")
>   # the residual is indepdent of covariate
```





summary(): summary of the fit

```
> summary(fit)
Call:
lm(formula = waiting ~ eruptions, data = faithful)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0796  -4.4831   0.2122   3.9246  15.9719

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.4744     1.1549   28.98   <2e-16 ***
eruptions    10.7296     0.3148   34.09   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.914 on 270 degrees of freedom
Multiple R-squared:  0.8115,    Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

$coefficients: features of regression coefficients

```
> summary(fit)$coefficients
             Estimate Std. Error  t value      Pr(>|t|)
(Intercept) 33.47440  1.1548735 28.98534  7.136015e-85
eruptions   10.72964  0.3147534 34.08904 8.129959e-100
```

- The 90% confidence interval (CI) for a parameter θ is (asymptotically)

$$[\hat{\theta} - z_{0.95} \cdot \hat{\sigma}_\theta, \hat{\theta} + z_{0.95} \cdot \hat{\sigma}_\theta],$$

  where $\hat{\theta}$ is the estimator to θ and $\hat{\sigma}_\theta$ is the estimated error.

- The 90% CI for the intercept is $[31.57480, 35.37399]$:

```
> summary(fit)$coefficients[1,1]+
+   c(-1,1)*qnorm(0.95)*summary(fit)$coefficients[1,2]
[1] 31.57480 35.37399
```

- The 90% CI for the slope is $[10.21192, 11.24736]$:

```
> summary(fit)$coefficients[2,1]+
+   c(-1,1)*qnorm(0.95)*summary(fit)$coefficients[2,2]
[1] 10.21192 11.24736
```

Hypothesis test and p-value

- If we want to compute p-values for some hypothesis, what should we do?
- For instance, assume we want to compute the p-value for testing the slope β

$$H_0 : \beta = 10.$$

- A test statistics is

$$\frac{|\hat{\beta} - 10|}{\hat{\sigma}_\beta},$$

where $\hat{\sigma}_\beta$ is the error for the estimator $\hat{\beta}$.

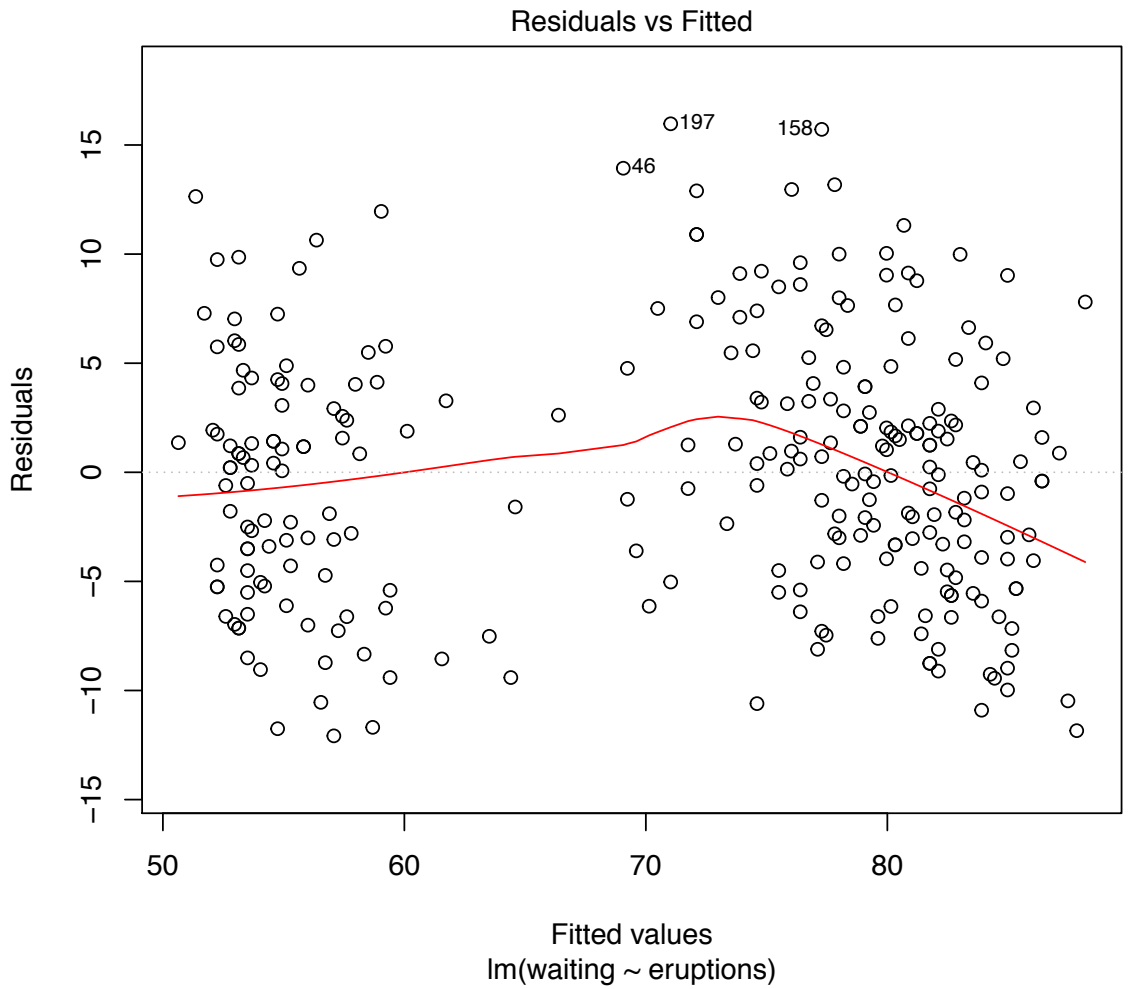
- In R, we will write

```
> test_stats <-
+   abs(summary(fit)$coefficients[2,1]-10)/
+   summary(fit)$coefficients[2,2]
> (1-pnorm(test_stats))*2
[1] 0.0204419
```

- So the p-value is 0.0204419.

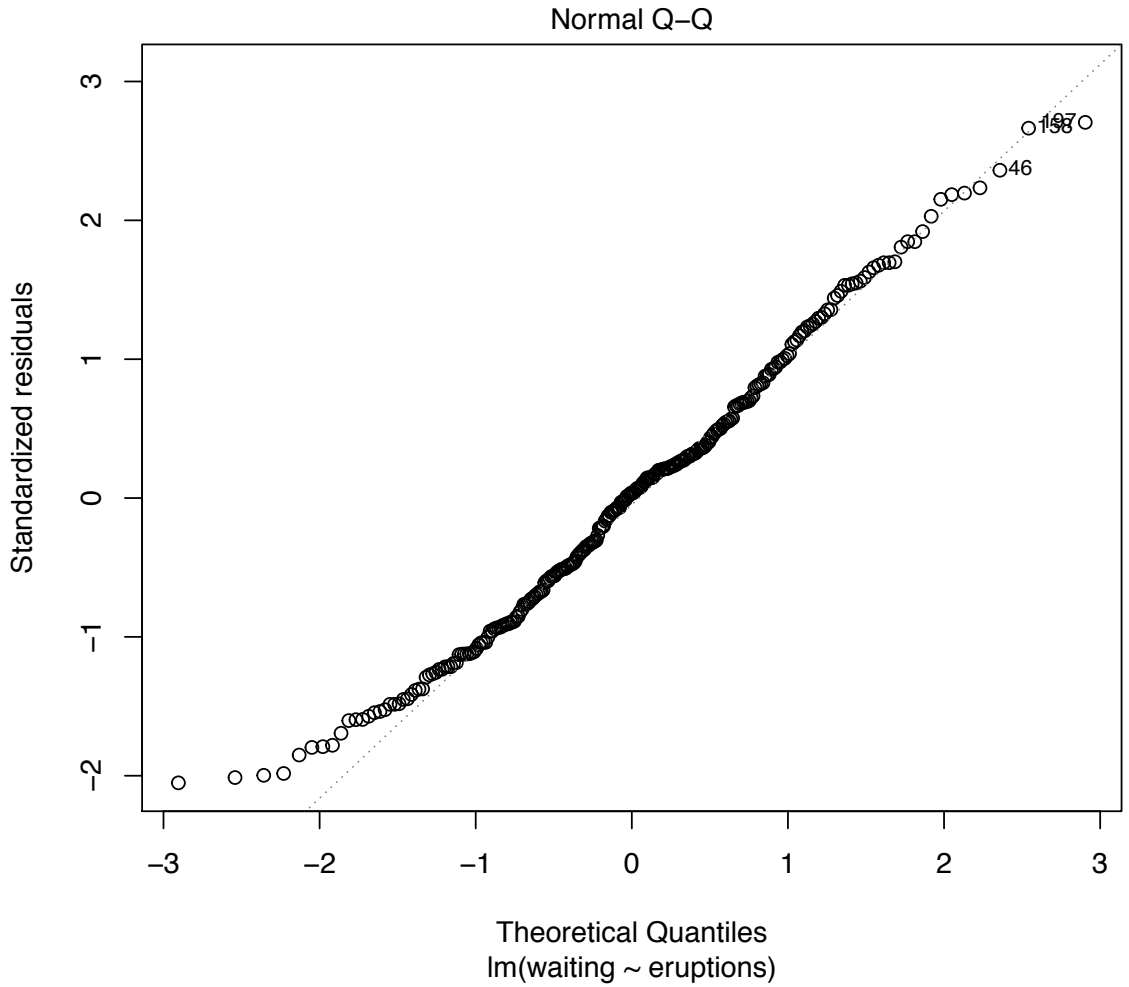
plot(): show diagnostic plots for the fit – 1

```
> plot(fit)
```



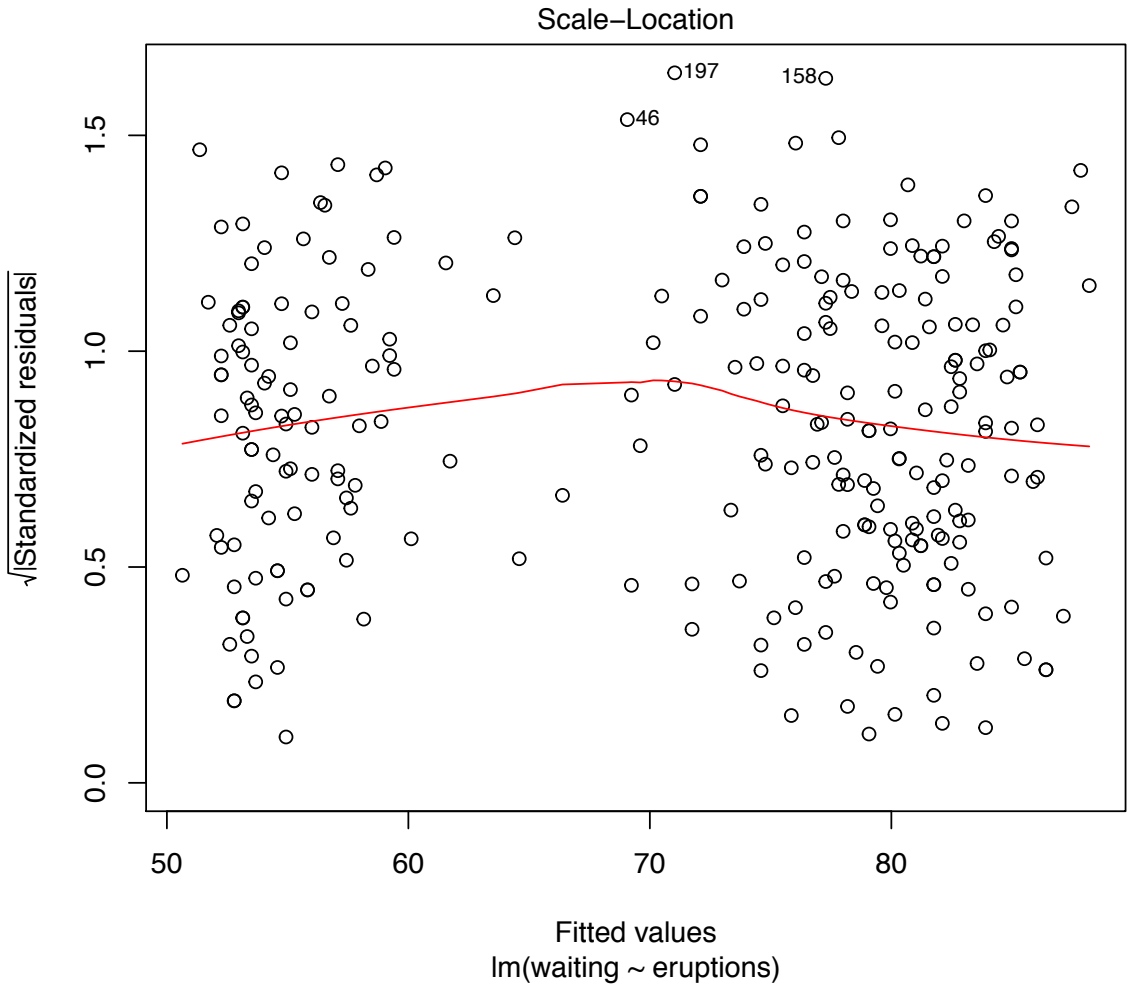
plot(): show diagnostic plots for the fit – 2

```
> plot(fit)
```



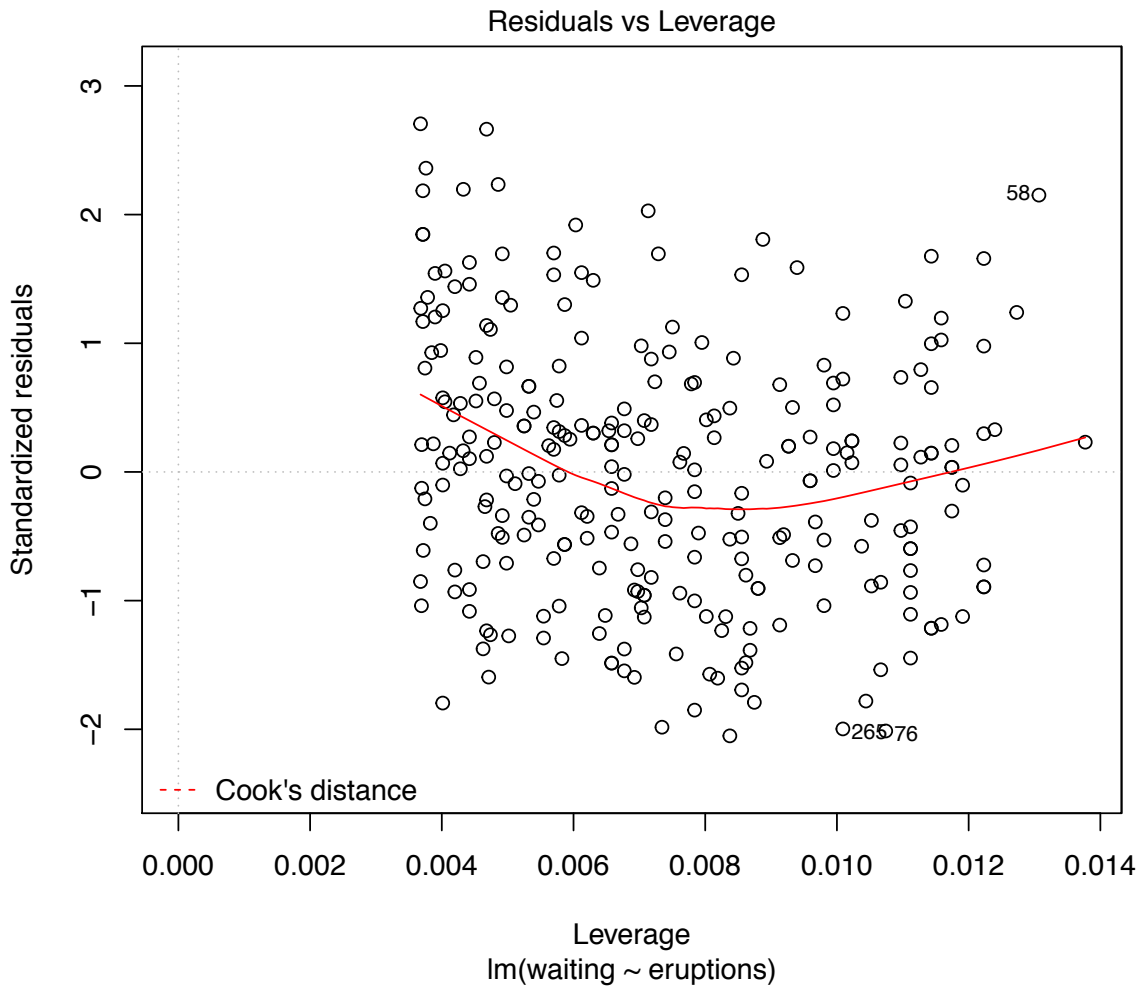
plot(): show diagnostic plots for the fit – 3

```
> plot(fit)
```



plot(): show diagnostic plots for the fit – 4

```
> plot(fit)
```



Fitting of two vectors – 1

```
> fit1 <- lm(faithful$waiting~faithful$eruptions)
>
> fit1

Call:
lm(formula = faithful$waiting ~ faithful$eruptions)

Coefficients:
       (Intercept)  faithful$eruptions
             33.47                10.73
```

$\longrightarrow$ the slope and the intercept are same as the `fit`.

Fitting of two vectors – 2

```
> x0 <- faithful$eruptions
> y0 <- faithful$waiting
> fit2 <- lm(y0~x0)
>
> fit2

Call:
lm(formula = y0 ~ x0)

Coefficients:
(Intercept)           x0
      33.47        10.73
```

$\longrightarrow$ again, the slope and the intercept are the same as the `fit`.

Reported Measurements

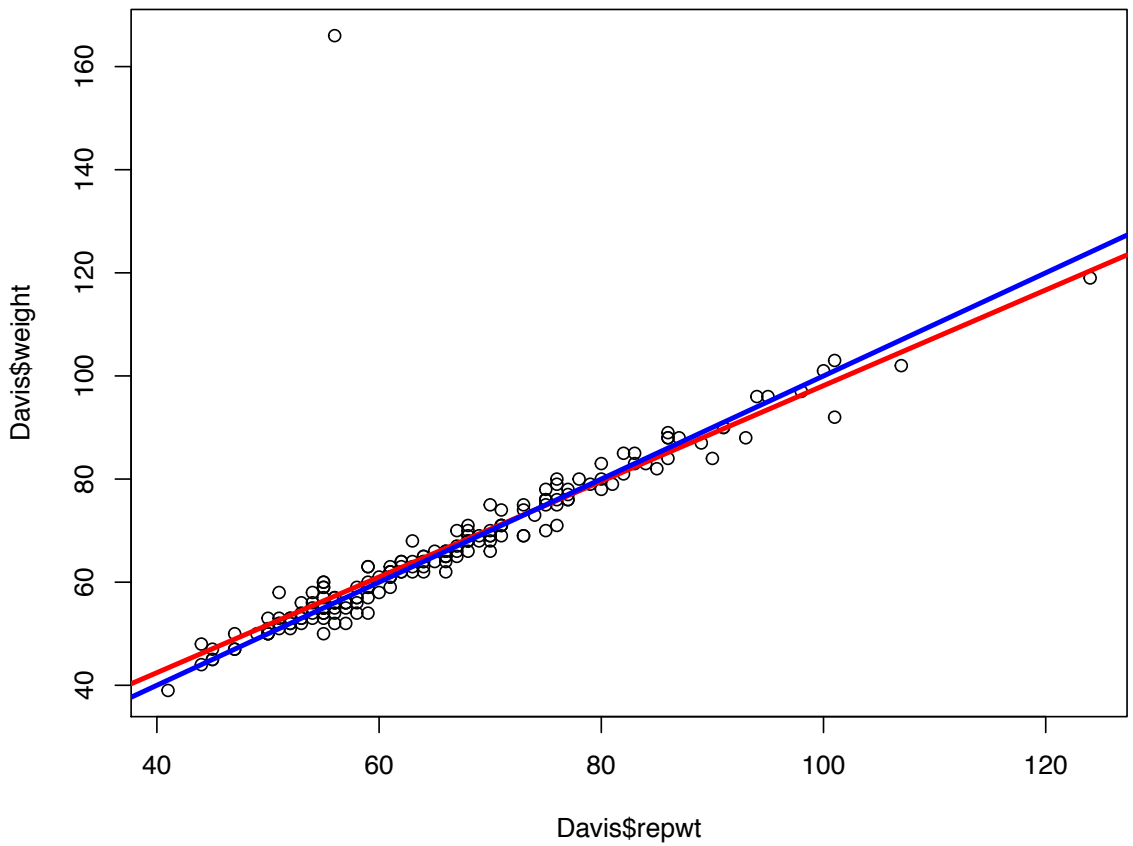
- The data set is `Davis` from the `car` package Companion for Applied Regression by John Fox.
- The idea is to check the reliability of self report.
- `install.packages("car")` followed by `library(car)` per R session.

```
> library(car)
> head(Davis)
  sex weight height repwt repht
1   M     77    182    77   180
2   F     58    161    51   159
3   F     53    161    54   158
4   M     68    177    70   175
5   F     59    157    59   155
6   M     76    170    76   165
```

```
> Davis.model <- lm(weight~repwt, data=Davis)
> summary(Davis.model)

Call:
lm(formula = weight ~ repwt, data = Davis)

Residuals:
    Min      1Q  Median      3Q     Max
 -7.048  -1.868  -0.728   0.601 108.705

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.3363     3.0369   1.757   0.0806 .
repwt         0.9278     0.0453  20.484   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.419 on 181 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared:  0.6986,    Adjusted R-squared:  0.697
F-statistic: 419.6 on 1 and 181 DF,  p-value: < 2.2e-16
```

Data Plot

```
> plot(Davis$repwt,Davis$weight)
> abline(Davis.model, col="red", lwd=3)
> abline(a=0,b=1, col="blue", lwd=3)
```



Finding the outlier

```
> which(Davis.model$residuals==
+         max(Davis.model$residuals))
12
12
>   # think about why there are two 12?
> which(Davis$weight>160)
[1] 12
>
> Davis[12,]
   sex weight height repwt repht
12   F    166     57    56   163
```

Outlier Treatment

```
> Davis[12,]
   sex weight height repwt repht
12   F    166     57    56   163
```

- it appears weight and height were interchanged for case 12
- can either interchange these 2 values and run regression again
- or omit case 12 and run regression again
- or update the regression

```
> Davis.model2 <- update(Davis.model, subset=-12)
> summary(Davis.model2)
```

```
> summary(Davis.model2)

Call:
lm(formula = weight ~ repwt, data = Davis, subset = -12)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5296 -1.1010 -0.1322  1.1287  6.3891

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.73380    0.81479   3.355 0.000967 ***
repwt        0.95837    0.01214  78.926  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.254 on 180 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared:  0.9719,    Adjusted R-squared:  0.9718
F-statistic:  6229 on 1 and 180 DF,  p-value: < 2.2e-16
```

Multiple Regression

- We use the `babies` data set of the package `UsingR`.
- First `install.packages("UsingR")`
- For each new R session execute `library(UsingR)`
- Examine that data frame `babies`. Here we focus on

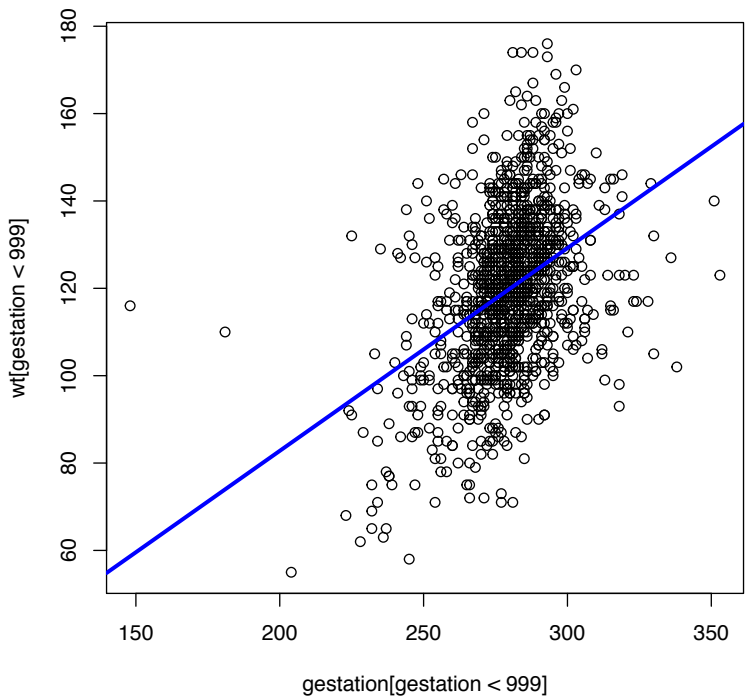
```
wt # baby's weight at birth (ounces)
gestation # pregnancy length (days), 999 = unknown
age # mother at end of pregnancy (yrs), 99 = unkown
ht # mother's height (inches), 99 = unknown
wt1 # mother's prepreg. weight (lbs), 999 = unknown
dage # father's age (yrs), 99 = unknown
dht # father's height (inches), 99 = unknown
dwt # father's weight (lbs), 999 = unknown
```

attach(): import the variables from a dataset

```
> head(babies$wt)
[1] 120 113 128 123 108 136
> head(wt)
Error in head(wt) : object 'wt' not found
> attach(babies)
> head(wt)
[1] 120 113 128 123 108 136
```

Simple Linear Regression

```
> wtgest.reg <- lm(wt ~ gestation,
+                  subset = gestation < 999)
>   # no need to say babies$XXX
> plot(gestation[gestation < 999],
+      wt[gestation<999])
> abline(wtgest.reg,col="blue", lwd=3)
```



Simple Linear Regression: Summary

```
> summary(wtgest.reg)

Call:
lm(formula = wt ~ gestation, subset = gestation < 999)

Residuals:
    Min      1Q  Median      3Q     Max
-49.394 -11.125   0.071  10.106  57.353

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.06418    8.32220  -1.209    0.227
gestation     0.46426    0.02974  15.609   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.66 on 1221 degrees of freedom
Multiple R-squared:  0.1663,    Adjusted R-squared:  0.1657
F-statistic: 243.6 on 1 and 1221 DF,  p-value: < 2.2e-16
```

Multiple Regression

```
> wtreg <- lm(wt ~ gestation + age + ht +
+    wt1 + dage + dht + dwt, data = babies,
+    subset = gestation < 999 & age < 99 & ht < 99 &
+    wt1 < 999 & dage < 99 & dht < 99 & dwt < 999)
```

Here we fit the following model, again by least squares

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i \quad \text{with } \epsilon_i \sim i.i.d.\ \mathcal{N}(0, \sigma^2)$$

where $y_i \sim$ `wt`$_i$, $x_{i1} \sim$ `gestation`$_i$, ..., $x_{ip} \sim$ `dwt`$_i$, and $p = 7$.

Note that here the cases used are specified by a logic vector,
only cases with `TRUE` enter the analysis.
Compare with `subset = -12` (omitting case 12),
as in the previous usage.

Multiple Regression (Summary)

```
> summary(wtreg)

Call:
lm(formula = wt ~ gestation + age +
    ht + wt1 + dage + dht + dwt,
    data = babies, subset = gestation < 999 &
    age < 99 & ht < 99 & wt1 < 999 &
    dage < 99 & dht < 99 & dwt < 999)

Residuals:
    Min      1Q  Median      3Q     Max
-48.675 -10.528   0.403  10.123  54.960
```

Multiple Regression (Summary Cont.)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -101.90746   23.29181  -4.375 1.40e-05 ***
gestation      0.45031    0.03909  11.520  < 2e-16 ***
age            0.13497    0.18811   0.717   0.4733
ht             1.22304    0.28517   4.289 2.05e-05 ***
wt1            0.03078    0.03427   0.898   0.3693
dage           0.06032    0.16551   0.364   0.7157
dht           -0.07833    0.27056  -0.290   0.7723
dwt            0.07831    0.03307   2.368   0.0182 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.43 on 693 degrees of freedom
Multiple R-squared:  0.2117,    Adjusted R-squared:  0.2038
F-statistic: 26.59 on 7 and 693 DF,  p-value: < 2.2e-16
```

Multiple Regression Decomposition and R^2

Via matrix algebra one gets again the following decomposition

$$SYY = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = RSS + SS_{reg}$$

where the fitted values are

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}$$

The multiple correlation coefficient is

$$R^2 = \frac{SYY - RSS}{SYY} = \frac{SS_{reg}}{SYY} = 1 - \frac{RSS}{SYY}$$

$$\text{with ``adjusted } R^2\text{''} = \bar{R}^2 = 1 - \frac{RSS/(n-p-1)}{SYY/(n-1)}$$

$p+1$ is the number of fitted parameters $\beta_0, \ldots, \beta_p$.

Comments on Multiple Regression Output Summary

- The `Std. Error` is the estimated standard deviation of the respective estimate.
- The `t value` is the ratio `t = Estimate/Std. Error`.
- `Pr( > |t|)` represents the two-sided p-value for the observed value `t`, when testing the hypothesis $H_j : \beta_j = 0$.
- The `F statistic` tests the hypothesis of no regression effect at all, i.e., $H_0 : \beta_1 = \ldots = \beta_p = 0$.
- It appears that the intercept $\beta_0 \neq 0$ and that `gestation` and `ht` are significant predictor variables, i.e., $\beta_1 \neq 0$ and $\beta_3 \neq 0$.
- The "significance" of `dwt` should be viewed with caution, in the context of multiple tests performed.

Model Update

We delete the non-significant variables from previous analysis

```
> wtreg0 <- update(wtreg,.~.-age-wt1-dage-dht)
> summary(wtreg0)

Call:
lm(formula = wt ~ gestation + ht + dwt,
    data = babies, subset = gestation <
    999 & age < 99 & ht < 99 & wt1 < 999 &
    dage < 99 & dht < 99 & dwt < 999)

Residuals:
    Min      1Q  Median      3Q     Max
-48.230 -10.482   0.394  10.379  56.108
```

Note the modified model call.

Model Update (Output Continued)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -102.46196   18.91277  -5.418 8.33e-08 ***
gestation      0.44803    0.03893  11.509  < 2e-16 ***
ht             1.31132    0.25456   5.151 3.37e-07 ***
dwt            0.07553    0.02839   2.660  0.00799 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.45 on 697 degrees of freedom
Multiple R-squared:  0.2056,    Adjusted R-squared:  0.2021
F-statistic: 60.12 on 3 and 697 DF,  p-value: < 2.2e-16
```

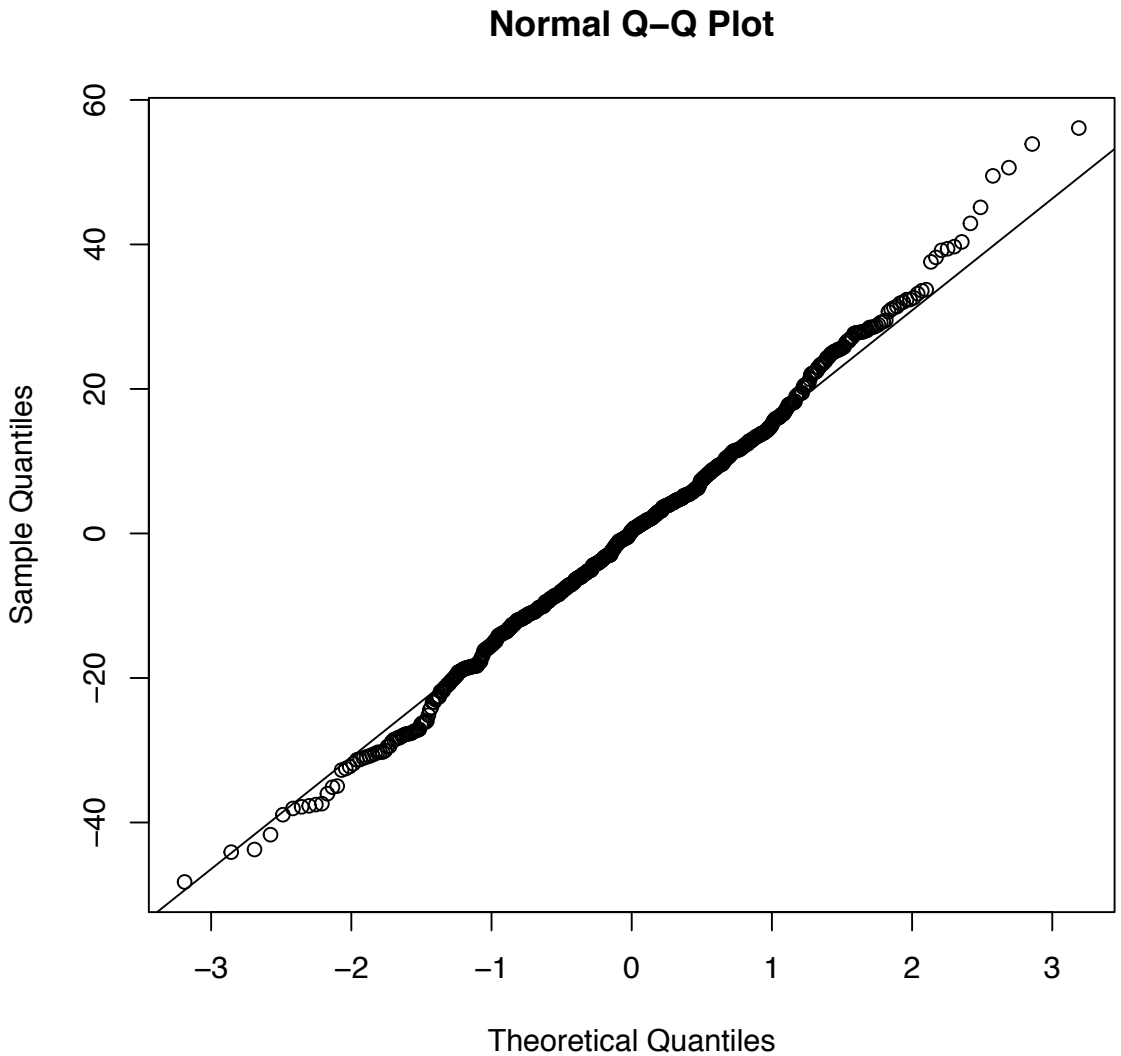
Confidence Intervals

- Confidence intervals for each parameter can be computed as:

```
> summary(wtreg0)$coefficients[,1] +
+   summary(wtreg0)$coefficients[,2]%*%
+   matrix(c(-1,1),nrow=1)*qnorm(0.95)
              [,1]         [,2]
[1,] -133.57070586 -71.3532147
[2,]    0.38399878   0.5120685
[3,]    0.89260453   1.7300402
[4,]    0.02882768   0.1222286
```

Examining Normality of Residuals

```
> qqnorm(wtreg0$resid)
> qqline(wtreg0$resid)
```



In-class Exercises

- Go back to the `Davis` example.
- Find the 95% confidence intervals for both the slope and the intercept for fitting $Y =$`weight` versus $X =$`repwt`.
- Test the following hypothesis:

$$H_0 : \text{intercept} = 5.$$

  What is the p-value? Can you reject H_0?
- Now remove the 12-th observation and redo all the analysis again. Can you reject the H_0?