STAT 210
Applied Statistics and Data Analysis
Week 8 - Summary

Joaquin Ortega

King Abdullah University of Science and Technology

Video 27: Linear Regression I

Suppose we have a joint sample

$$(X_1, Y_1), (X_2, Y_2) \dots, (X_n, Y_n)$$

and we want to determine whether there exists a relationship between the two variables.

The simplest relation is a linear model such as

$$Y = \beta_0 + \beta_1 X. \tag{1}$$

In this model,

- Y is the **response** or dependent variable,

- X is a (continuous) **explanatory** or independent variable, also known as a **regressor**.

There are two **parameters** in the model, the slope β_1 and the intercept β_0.

Since we have a sample of values from both variables (X_i, Y_i), $i = 1, \ldots, n$, the model is usually written as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n. \tag{2}$$

This model is

- **simple**, because it has only one regressor or independent variable,
- **linear**, because it is linear in the parameters,

and is also linear on the variables because the predictor variable only appears raised to the power 1.

For instance,

$$Y = \beta_0 + \beta_1 X^2 + \beta_2 \log(X) + \varepsilon$$

is also a linear model, but

$$Y = \beta_0 + \frac{\beta_1 \log(X)}{1 + \beta_2 X^2} + \varepsilon$$

is not.

We assume that the ϵ_i are

- centered, $E[\epsilon_i] = 0$,
- have equal variance $Var(\epsilon_i) = \sigma^2, i = 1, \ldots, n$,
- have normal distribution, and
- are independent.

The expected value of Y given X is

$$\begin{aligned} E[Y|X] &= E[\beta_0 + \beta_1 X + \epsilon_i] \\ &= \beta_0 + \beta_1 E[X] + E[\epsilon_i] \\ &= \beta_0 + \beta_1 X. \end{aligned}$$

The distribution of Y **when X is known** is Gaussian with mean $\beta_0 + \beta_1 X$ and variance σ^2.

The slope β_1 represents the expected change in Y when X changes one unit.

When $\beta_1 = 0$, the response Y is independent of the explanatory variable X.

We want to estimate the parameters for the model from a sample of values $(x_1, y_1), \ldots, (x_n, y_n)$.

We can write the relation as a system of linear equations

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\
&\;\;\vdots \quad\;\; \vdots \quad\;\; \vdots \quad\;\; \vdots \\
y_n &= \beta_0 + \beta_1 x_n + \epsilon_n.
\end{aligned}
$$

In matrix notation this can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}; \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}; \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

$\mathbf{X}$ is known as the **design matrix**, while $\boldsymbol{\beta}$ is the vector of parameters.

Estimation

The model we want to fit is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where (x_i, y_i) are the observed values.

The errors are the differences between the observed values and the values that the model predicts:

$$\epsilon_i = y_i - \beta_0 - \beta_1 x_i.$$

We adopt the least-squares criterion for choosing the parameter values. We want

$$(\hat{\beta}_0, \hat{\beta}_1) = \text{argmin}_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{3}$$

The sum on the right of (3) is known as the **error sum of squares** SSE:

$$SSE = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

and in matrix notation

$$SSE = \epsilon' \epsilon = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta).$$

The fitted values are given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

and the residuals are given by

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

Derivation of the estimated values for the βs

We want to minimize SSE. For this, we take partial derivatives wrt the parameters and set them to zero:

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial SSE}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \qquad (4)$$

These are known as the **normal equations**.

The least squares estimators for the model parameters are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (5)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}. \qquad (6)$$

Information from the `lm` function in Example 1:

```r
summary(lm1)
```

```
##
## Call:
## lm(formula = FL ~ CL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86395 -0.51746 -0.02826  0.50456  1.77009
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.15316    0.23477   0.652    0.515
## CL           0.48060    0.00714  67.313   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.717 on 198 degrees of freedom
## Multiple R-squared:  0.9581, Adjusted R-squared:  0.9579
## F-statistic:  4531 on 1 and 198 DF,  p-value: < 2.2e-16
```

Residuals and Properties of the Regression Line

The difference between the observed values y_i, and the fitted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \ldots, n$ are the residuals:

$$\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

The model we have fitted is the one that minimizes the sum of squares of these residuals. Observe that

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$= n\bar{y} - n(\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 n\bar{x} = 0. \qquad (7)$$

We have already seen in (7) that $\sum_i \hat{\epsilon}_i = 0$. Other properties are:

❶ $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$.

❷ $\sum_{i=1}^{n} x_i \hat{\epsilon}_i = 0$.

❸ $\sum_{i=1}^{n} \hat{y}_i \hat{\epsilon}_i = 0$.

❹ The regression line always goes through $(\bar{x}, \bar{y})$.

Video 28: Simple Linear Regression II

Recall that the error sum of squares in matrix notation is given by

$$SSE = \epsilon'\epsilon = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta). \tag{8}$$

Multiplying out the terms in this expression we have

$$SSE = \mathbf{Y}'\mathbf{Y} - \beta'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta. \tag{9}$$

Since $\beta'\mathbf{X}'\mathbf{Y}$ is a scalar, it is equal to its transpose so

$$\beta'\mathbf{X}'\mathbf{Y} = (\beta'\mathbf{X}'\mathbf{Y})' = \mathbf{Y}'\mathbf{X}\beta$$

and we get

$$SSE = \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta. \tag{10}$$

The derivative of SSE is given by

$$\frac{\partial SSE}{\partial \beta} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta$$

Setting this expression equal to zero and solving for β we obtain

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{11}$$

which is the matrix version of the normal equations.

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \\
&= \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix} \\
&= \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i \\ -\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i + n\sum_{i=1}^{n} x_i y_i \end{pmatrix}
\end{aligned} \tag{12}$$

Sampling Distribution of $\hat{\beta}$.

We have assumed that $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and therefore

$$\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}).$$

Thus,

$$\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}). \tag{13}$$

Consequently, $\hat{\beta}$ is a linear unbiased estimator for β.

The unbiased estimator for the error variance σ^2 is

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2. \tag{14}$$

Hypothesis tests

A test statistic for $H_0 : \beta_i = \beta_{i,0}$ versus $H_1 : \beta_i \neq \beta_{i,0}$ can be obtained from the pivotal quantity

$$\frac{\hat{\beta}_i - \beta_{i,0}}{s(\hat{\beta}_i)} \tag{15}$$

for $i = 0, 1$, which has a t_{n-2} distribution under H_0. Here, $s(\hat{\beta}_i)$ is the standard error for $\hat{\beta}_i$.

For the test $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$, the function `summary` applied to a linear model object will provide

$$t_{obs} = \hat{\beta}_i / s(\hat{\beta})$$

and the corresponding p-value:

$$p - \text{value} = 2P(t_{n-2} \geq |t_{obs}|).$$

Let us review again the results for the initial model

```
summary(lm1)
```

```
##
## Call:
## lm(formula = FL ~ CL)
##
## Residuals:
##      Min      1Q   Median      3Q      Max
## -1.86395 -0.51746 -0.02826  0.50456  1.77009
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.15316    0.23477   0.652    0.515
## CL           0.48060    0.00714  67.313   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.717 on 198 degrees of freedom
## Multiple R-squared:  0.9581, Adjusted R-squared:  0.9579
## F-statistic:  4531 on 1 and 198 DF,  p-value: < 2.2e-16
```

Confidence Intervals

Using the pivotal quantity (15) we can get confidence intervals for the parameters of the model.

A $100(1 - \alpha)\%$ confidence interval for $\beta_i, i = 0, 1$ is

$$CI_{1-\alpha}(\beta_i) = \left(\hat{\beta}_i - t_{n-2,1-\alpha/2}s(\hat{\beta}_i), \hat{\beta}_i + t_{n-2,1-\alpha/2}s(\hat{\beta}_i) \right)$$

Video 29: Linear Regression 3
Confidence Bands and Anova

Confidence Bands for the Regression Line

Notation: $\mu_{Y|x} = E(Y|X = x)$

Recall that

$$E(Y|X) = \beta_0 + \beta_1 X$$

For $\mu_{Y|x}$, there are two sources of variability, $\hat{\beta}_0$, and $\hat{\beta}_1$.

The standard error (or empirical standard deviation) of $\mu_{Y|x}$ is

$$se_{\mu_{Y|x}} = \hat{\sigma}\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)^{1/2}. \qquad (16)$$

Observe that the standard error is a minimum when $x = \bar{x}$.

A confidence interval for the average value of Y at x at the $(1 - \alpha)$ level is given by

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x - t_{n-2,1-\alpha/2} se_{\mu_{Y|x}}, \hat{\beta}_0 + \hat{\beta}_1 x + t_{n-2,1-\alpha/2} se_{\mu_{Y|x}} \right)$$
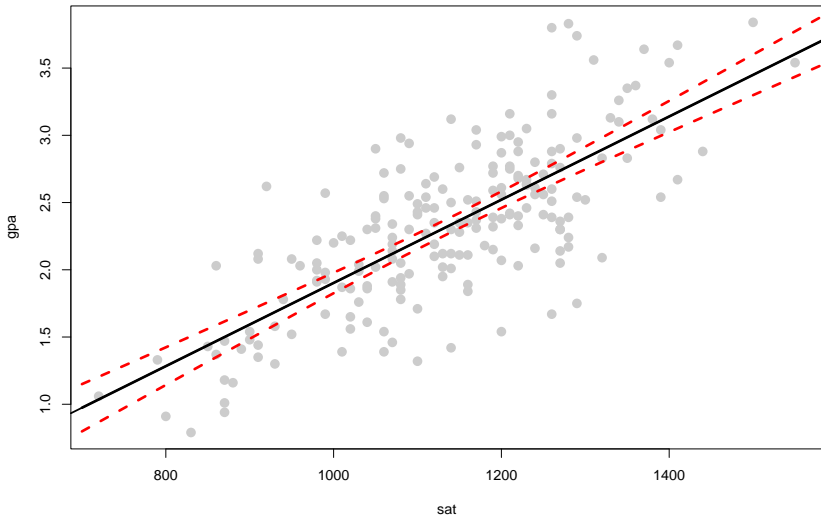
We can get these intervals using the function `predict`, which, when applied to an object of class `lm` and a data frame of x values, will give the values of the regression line at the x values, with the option of adding confidence intervals.

```
new.data <- data.frame(x=c(900,1100,1300))
predict(model,new.data,interval='c')
```

```
##        fit      lwr      upr
## 1 1.592779 1.486950 1.698609
## 2 2.211634 2.154371 2.268896
## 3 2.830488 2.746088 2.914887
```

This can be used to draw 'confidence bands' for the regression line in the SAT/GPA example.

```r
plot(sat, gpa)
modelA <- lm(gpa~sat, data = Grades)
abline(modelA)
new.sat <- data.frame(sat=seq(700,1600,
                              length.out = 15))
pc <- predict(modelA,new.sat, int='c')
matlines(new.sat$sat, pc, lty=c(1,2,2),
         lwd=rep(2,3),
         col=c('black','red','red'))
```

If we wanted to predict the value of *y* corresponding to a given value of *x* (instead of predicting the *average* value of *y* at *x*), we would expect a wider confidence band.

To avoid confusion, these are called **prediction** intervals.

Prediction intervals are wider because they take into account sampling variability due to the error term in the model.

Also, since the uncertainty in the estimation of the parameters is less important, their curvature is less pronounced.

The standard error for the predicted value $\hat{y}$ at x is given by

$$se_{\hat{y}|x} = \hat{\sigma}\left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i(x_i - \bar{x})^2}\right)^{1/2}.$$

A prediction interval for the value $\hat{y}$ at the point x and the $(1 - \alpha)$ level is given by

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x - t_{n-2,1-\alpha/2} se_{\hat{y}|x}, \hat{\beta}_0 + \hat{\beta}_1 x + t_{n-2,1-\alpha/2} se_{\hat{y}|x} \right)$$

The `predict` function also calculates prediction intervals.

```
new.data <- data.frame(x=c(900,1100,1300))
predict(model,new.data,interval='p')
```
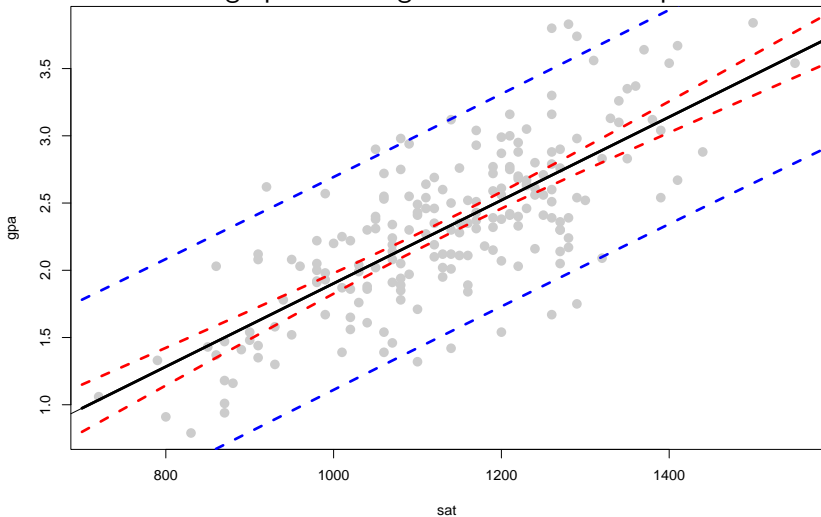
```
##        fit       lwr      upr
## 1 1.592779 0.7979927 2.387566
## 2 2.211634 1.4218455 3.001422
## 3 2.830488 2.0382696 3.622706
```

```
predict(model,new.data,interval='c')
```

```
##        fit      lwr      upr
## 1 1.592779 1.486950 1.698609
## 2 2.211634 2.154371 2.268896
## 3 2.830488 2.746088 2.914887
```

Let's now draw a graph including both bands for comparison.

```r
plot(sat, gpa)
modelA <- lm(gpa~sat, data = Grades)
abline(modelA)
new.sat <- data.frame(sat=seq(700,1600,
                             length.out = 15))
pc <- predict(modelA,new.sat, int='c')
matlines(new.sat$sat, pc, lty=c(1,2,2),lwd=rep(2,3),
        col=c('black','red','red'))
pp <- predict(modelA,new.sat, int='p')
matlines(new.sat$sat, pp, lty=c(1,2,2),lwd=rep(2,3),
        col=c('black','red','red'))
```

Analysis of Variance in Linear Regression

Anova is based on dividing the sums of squares and degrees of freedom associated with the response variable Y.

The difference $y_i - \bar{y}$ is divided into two parts:

1.- The deviation of y_i from the regression line: $y_i - \hat{y}_i$.

2.- The deviation of the fitted value $\hat{y}_i$ from the mean: $\hat{y}_i - \bar{y}$.

$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$$

Squaring this relation and summing up over i we get

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

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{17}$$

This relation is commonly expressed as

$$SST = SSE + SSR$$

where

- SST denotes the total sum of squares,
- SSE is the error or residual sum of squares and
- SSR is the regression sum of squares.

The terms $y_i - \bar{y}$ represent the distance from the observed values to the average, $y_i - \hat{y}_i$ is the distance between the observed and the fitted value and $\hat{y}_i - \bar{y}$ is the distance between the fitted value and the average observed value.

The degrees of freedom are similarly distributed.

There are $n - 1$ degrees of freedom associated with SST; one degree is lost since we need to estimate $\bar{y}$.

These degrees of freedom are divided into SSR and SSE.

SSE has $n - 2$ degrees of freedom; two are lost because we need to estimate parameters β_0 and β_1, to fit the regression line.

Finally, there are two degrees of freedom associated with the regression line, one for the slope and one for the intercept, but one is lost since $\sum_i (\hat{y}_i - \bar{y}) = 0$ by property 1, so that SSR has one degree of freedom.

Sums of squares divided by their degrees of freedom are known as **mean squares** and are denoted by MS, thus

$$MSE = \frac{SSE}{n-2}, \qquad \text{and} \qquad MSR = \frac{SSR}{1} = SSR.$$

We have assumed that the errors in the regression are centered normal with variance σ^2, and therefore $SSE/\sigma^2 \sim \chi^2_{n-2}$, this gives $E(SSE/\sigma^2) = n-2$ and

$$E(MSE) = E\left(\frac{SSE}{n-2}\right) = \sigma^2,$$

which means that MSE is an unbiased estimator of σ^2.

$$E(MSR) = E(\hat{\beta}_1^2) \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \left(Var(\hat{\beta}_1) + (E(\hat{\beta}_1))^2 \right) \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

When $\beta_1 = 0$, the mean of the sampling distribution of MSR is σ^2 and coincides with the mean of MSE.

If $\beta_1 = 0$, the variables SSR/σ^2 and SSE/σ^2 have a χ^2 distribution with 1 and $n-2$ degrees of freedom, and it is possible to show that they are independent.

In consequence,

$$\frac{MSR}{MSE} = \frac{\frac{SSR/\sigma^2}{1}}{\frac{SSE/\sigma^2}{n-2}} = \frac{\chi_1^2/1}{\chi_{n-2}^2/(n-2)} \sim F_{1,n-2}.$$

Therefore, to test $H_0 : \beta_1 = 0$ we use this statistic. If msR and msE are the observed values for the sums of squares then

$$F_{obs} = \frac{msR}{msE}$$

and large values of F_{obs} give evidence against the null hypothesis.

At a confidence level of $1 - \alpha$, the null hypothesis will be rejected if

$$F_{obs} \geq F_{1,n-2,1-\alpha}.$$

The usual way to sum up these results is through an Analysis of Variance (Anova) table.

Table 1: Anova table for example 1.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F_{obs}	p-value
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$F_{obs} = \frac{MSR}{MSE}$	$P(F_{1,n-2} \geq F_{obs})$
Error	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$		
Total	SST	$n-1$			

In R we get an anova table with the command `anova` acting on an object of class `lm`:

```
anova(lm1)

## Analysis of Variance Table
##
## Response: FL
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## CL          1 2329.45 2329.45  4531.1 < 2.2e-16 ***
## Residuals 198  101.79    0.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
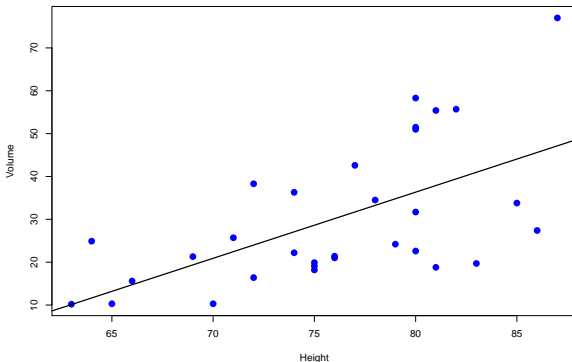
Example 3

The data set `trees` has data on girth, height, and volume of timber
in 31 felled black cherry trees. Girth is the diameter of the tree in
inches measured at 4 ft 6 in above the ground.

```
plot(Volume ~ Height, data=trees, cex = 1.5, pch = 16, col = 'blue')
lm4 <- lm(Volume ~ Height, data=trees)
abline(lm4, lwd = 2)
```

Example 3

```r
summary(lm4)
```

```
##
## Call:
## lm(formula = Volume ~ Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.274  -9.894  -2.894  12.068  29.852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.1236    29.2731  -2.976 0.005835 **
## Height        1.5433     0.3839   4.021 0.000378 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 29 degrees of freedom
## Multiple R-squared:  0.3579, Adjusted R-squared:  0.3358
## F-statistic: 16.16 on 1 and 29 DF,  p-value: 0.0003784
```

Example 3

```r
anova(lm4)
```

```
## Analysis of Variance Table
##
## Response: Volume
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Height      1 2901.2 2901.19  16.165 0.0003784 ***
## Residuals  29 5204.9  179.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Problem List 7

1