

Linear Models in R

Jim Bentley

1 Sample Data

The following code creates a couple of sample data frames that we will use in our examples.

```
> sex = c(rep("Female",12),rep("Male",7))
> mass = c(36.1, 54.6, 48.5, 42.0, 50.6, 42.0, 40.3, 33.1, 42.4, 34.5,
+         51.1, 41.2, 51.9, 46.9, 62, 62.9, 47.4, 48.7, 51.9)
> rate = c(995, 1425, 1396, 1418, 1502, 1256, 1189, 913, 1124, 1052,
+         1347, 1204, 1867, 1439, 1792, 1666, 1362, 1614, 1460)
> bps5.4.9 = data.frame(sex, mass, rate)
> htwt = read.csv(
+ "http://bulldog2.redlands.edu/facultyfolder/jim_bentley/downloads/math111/htwt.csv")
> htwt$Group = factor(htwt$Group,levels=c(1,2),labels=c("Male","Female"))
> hospitals = read.csv(
+ "http://bulldog2.redlands.edu/facultyfolder/jim_bentley/downloads/math111/hospitals.
```

Note that the plus signs (+) at the beginning of the lines are there to indicate that R is reading from a new line. They should not be entered as part of the code.

We can now check to see if the data frames have been created by entering

```
> ls()

[1] "bps5.4.9" "hospitals" "htwt"      "mass"      "rate"      "sex"
```

Note that the listing also shows the individual variables that were used to create the data frame. These can be deleted by using `rm()`.

```
> rm("sex","mass","rate")
> ls()

[1] "bps5.4.9" "hospitals" "htwt"
```

2 Loading R Packages

```
> ## load a few required packages
> #install.packages("xtable")
> library(Hmisc)
```

```
> library(xtable)
> library(ggplot2)
> library(survival)
> library(Rcmdr)
```

3 Testing The Population Mean

3.1 The One Sample Test

A simple test for the population mean of the `Weight` variable in the `htwt` data can be obtained via the `t.test` function. To compute the one sample t-test of $H_0 : \mu = 145$ we enter:

```
> t.test(htwt$Weight, mu=145, alternative=two.sided,
+        conf.level=.95)
```

```
One Sample t-test
```

```
data: htwt$Weight
t = -0.56, df = 19, p-value = 0.582
alternative hypothesis: true mean is not equal to 145
95 percent confidence interval:
 119.4182 159.7818
sample estimates:
mean of x
 139.6
```

An equivalent test of $H_0 : \mu = 145$ may be carried out using a linear model via the `lm` function.

```
> summary(lm((Weight-145)~1, data=htwt))
```

```
Call:
```

```
lm(formula = (Weight - 145) ~ 1, data = htwt)
```

```
Residuals:
```

```
   Min       1Q   Median       3Q      Max
-57.60 -31.35 -16.10   27.15   88.40
```

```
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -5.400      9.642   -0.56   0.582
```

```
Residual standard error: 43.12 on 19 degrees of freedom
```

Notice that adding the coefficient from the model to the hypothesized mean gives the sample mean. That is $145 + (-5.4) = 139.6$. Note, too that the p-values computed by `t.test` and `lm` are the same ($p = 0.582$).

3.2 The Two Sample Test

A simple test to compare the male and female population means of the `Weight` variable in the `htwt` data can also be obtained via the `t.test` function. To compute the two sample t-test of $H_0 : \mu_m = \mu_f$ we enter:

```
> t.test(Weight~Group, alternative=two.sided, conf.level=.95,  
+       var.equal=TRUE, data=htwt)
```

Two Sample t-test

```
data: Weight by Group  
t = 1.4903, df = 18, p-value = 0.1534  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
-11.4713  67.4713  
sample estimates:  
mean in group Male mean in group Female  
          155          127
```

An equivalent test of $H_0 : \beta_1 = 0 = \mu_m - \mu_f$ may be carried out using a linear model via the `lm` function.

```
> summary(lm(Weight~Group, data=htwt))
```

Call:

```
lm(formula = Weight ~ Group, data = htwt)
```

Residuals:

```
    Min      1Q  Median      3Q      Max  
-68.00 -31.50  -6.50   31.25   73.00
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)   155.00      13.93   11.12 1.69e-09 ***  
GroupFemale   -28.00      18.79   -1.49  0.153
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

Residual standard error: 41.8 on 18 degrees of freedom

Multiple R-squared: 0.1098, Adjusted R-squared: 0.06039

F-statistic: 2.221 on 1 and 18 DF, p-value: 0.1534

Notice that intercept term (155) is the sample mean for the males. The sample mean for the females is the model evaluated for a female ($155 + (-28) = 127$). As in the one sample problem the p-values computed by `t.test` and `lm` are the same ($p = 0.153$).

3.3 Correcting for Height

It is fairly clear from graphing `Weight` as a function of `Height` that when modeling a person's weight we should correct for height. While this cannot be accomplished using a t-test, a linear model makes the correction fairly easy.

To test for $H_0 : \beta_1 = 0$ when controlling for `Height` using the model

$$\text{Weight} = \beta_0 + \beta_1 \text{Female} + \beta_2 \text{Height} + \epsilon$$

we compute

```
> summary(lm(Weight~1+Group+Height, data=htwt))
```

Call:

```
lm(formula = Weight ~ 1 + Group + Height, data = htwt)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-9.539 -6.022 -1.253  4.032 14.720
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -170.6997    13.8866  -12.292 6.96e-10 ***
GroupFemale   -1.5796     3.4779   -0.454  0.655
Height         5.0108     0.2103   23.826 1.68e-14 ***
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 7.334 on 17 degrees of freedom

Multiple R-squared: 0.9741, Adjusted R-squared: 0.9711

F-statistic: 319.9 on 2 and 17 DF, p-value: 3.239e-14

Notice that as before there does not appear to be a difference between females and males ($p = 0.655$). However, it is clear that `Height` is predictive of `Weight` ($p < 0.001$).

Interaction Terms At this point we may be convinced that no differences exist in the weights of our two groups. Clearly the means for this sample are not significantly different. A little more insight may be gained by including an interaction term.

We now fit the model

$$\text{Weight} = \beta_0 + \beta_1 \text{Female} + \beta_2 \text{Height} + \beta_3 \text{Female} * \text{Height} + \epsilon$$

```
> lm.htwt = lm(Weight~1+Group*Height, data=htwt)
```

```
> summary(lm.htwt)
```

Call:

```
lm(formula = Weight ~ 1 + Group * Height, data = htwt)
```

Residuals:

Min	1Q	Median	3Q	Max
-9.968	-3.413	-1.104	2.697	13.163

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-198.2609	16.6933	-11.877	2.39e-09	***
GroupFemale	54.4858	23.2997	2.338	0.0327	*
Height	5.4348	0.2547	21.340	3.51e-13	***
GroupFemale:Height	-0.9013	0.3713	-2.427	0.0274	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.463 on 16 degrees of freedom

Multiple R-squared: 0.9811, Adjusted R-squared: 0.9775

F-statistic: 276.6 on 3 and 16 DF, p-value: 5.425e-14

It is now clear that not only is height predictive of weight ($p < 0.0001$), more importantly, females and males put weight on differently. Since the interaction term is significant ($p = 0.0274$) this indicates that their slopes are different with the women putting on about one pound less per inch than the men.

Diagnostic plots can be generated by using the `plot` function on the `lm` object, `lm.htwt`. Figure 1 shows the four diagnostic plots that are the default. The analysis of variance table may also be generated.

```
> # Set up the page to take all four images
> par(mfrow=c(2,2))
> plot(lm.htwt)
> anova(lm.htwt)
```

Analysis of Variance Table

Response: Weight

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Group	1	3880.8	3880.8	92.9116	4.570e-08	***
Height	1	30535.6	30535.6	731.0636	8.778e-15	***
Group:Height	1	246.1	246.1	5.8921	0.02738	*
Residuals	16	668.3	41.8			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The question of mean differences is thus shown to be the wrong question. The investigator should have been looking to see if men and women put on an equivalent number of pounds for each inch difference in height. This is something that is not apparent when looking at t-tests.

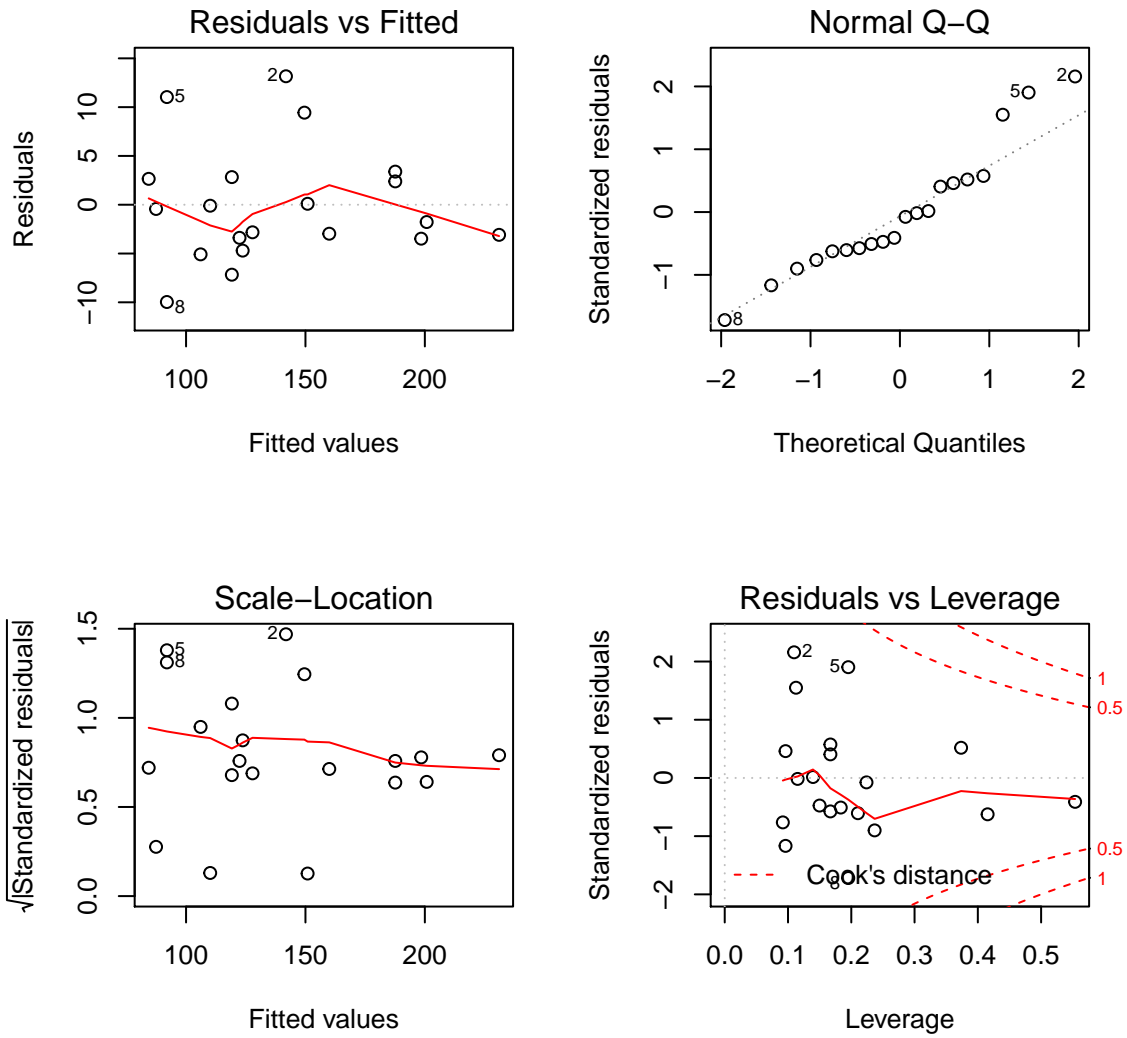


Figure 1: Default diagnostic plots for the full model fitted to the `htwt` data.