GLM in R

Oliver and Thurber

Data Entry

The following code creates a sample data frame that we will use in our examples.

```
# Import the height and weight data
htwt = read.csv("http://facweb1.redlands.edu/fac/jim_bentley/downloads/math111/htwt.csv")
# The Group variable is actually categorical and not numeric
htwt$Group = factor(htwt$Group,levels=c(1,2),labels=c("Male","Female"))
head(htwt)
```

```
##   Height Weight  Group
## 1     64    159   Male
## 2     63    155 Female
## 3     67    157 Female
## 4     60    125   Male
## 5     52    103 Female
## 6     58    122 Female
```

Testing The Population Mean

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 H0: mu=145 we enter:

```
t.test(htwt$Weight, mu=145, alternative='two.sided', conf.level=.95)
```

```
##
##  One Sample t-test
##
## data:  htwt$Weight
## t = -0.56003, 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 H0: 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).

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 H0: 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 between group Male and group Female is not equal to
## 95 percent confidence interval:
##  -11.4713  67.4713
## sample estimates:
##   mean in group Male mean in group Female
##                  155                  127
```

An equivalent test of H0: beta1 = 0 = mu.m-mu.f may be carried out using a linear model via the **lm** function.

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

```
##
## 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
```

```
plot(htwt$Group,htwt$Weight)
points(c(1,2),by(htwt[,2],htwt[,3],mean),type="b")
```
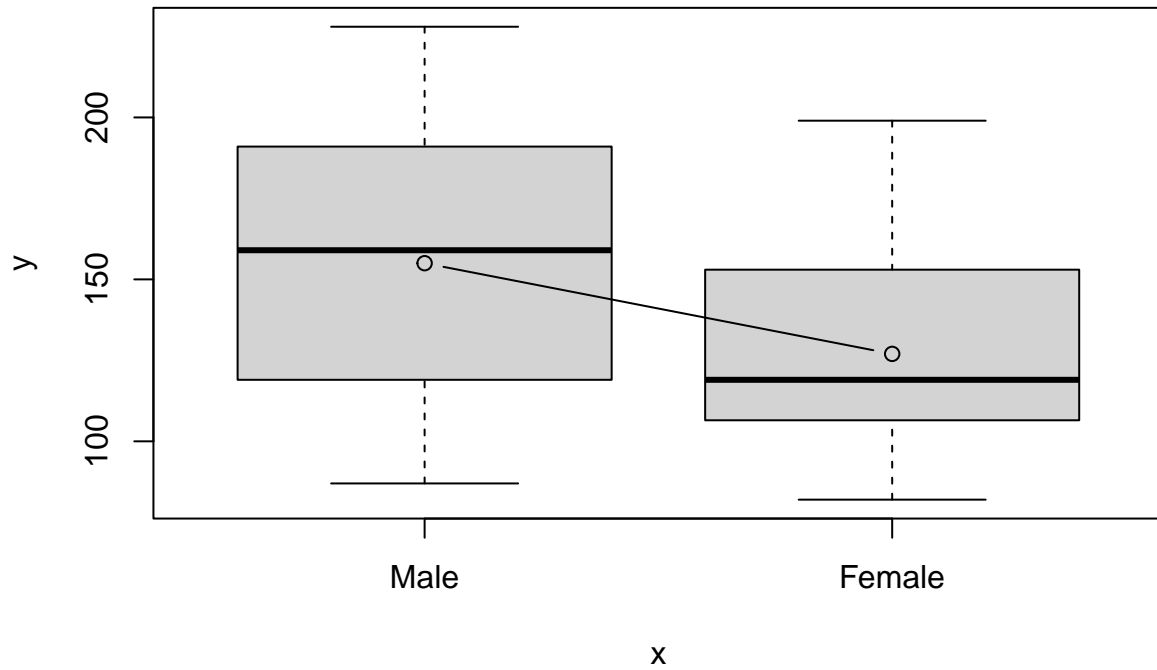


```
par(mfrow=c(2,2))
plot(htwt.lm.Group)
```
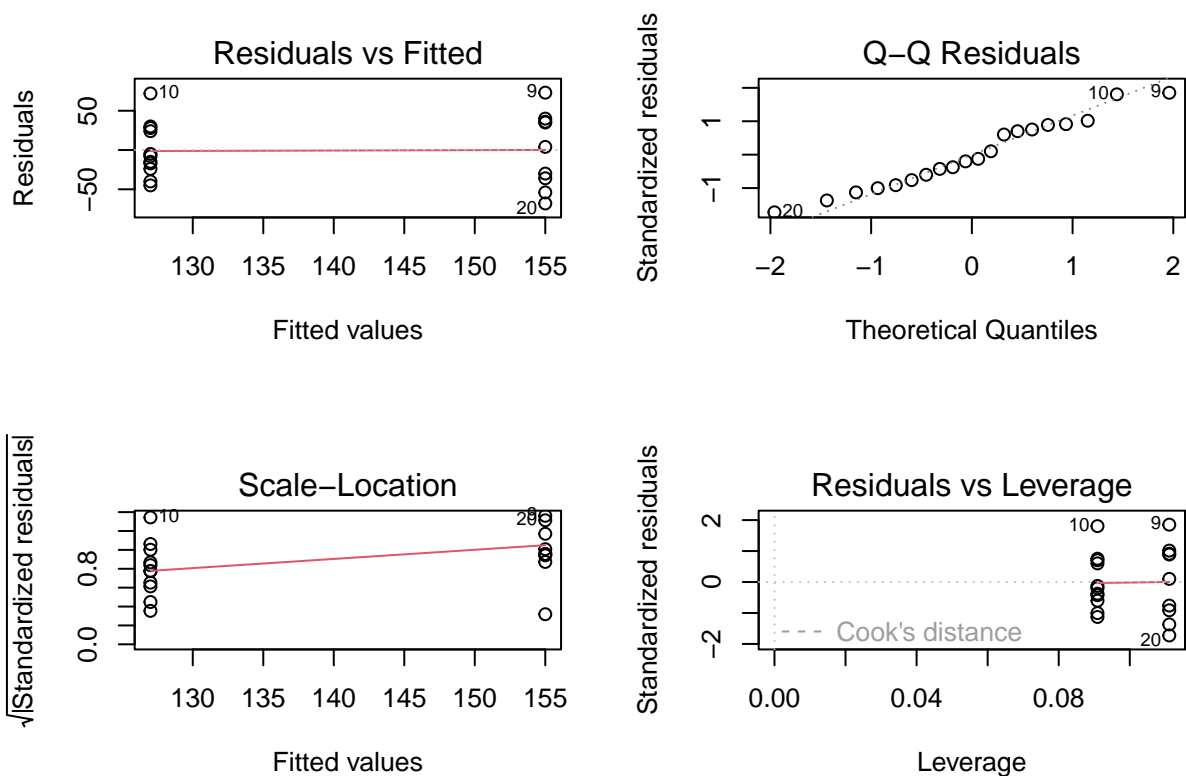
```r
par(mfrow=c(1,1))
```

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).

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 H0: beta.1 = 0 when controlling for **Height** using the model Weight = beta.0 + beta.1 Female + beta.2 Height + epsilon we compute

```r
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
```

```r
plot(htwt$Height,htwt$Weight,pch=as.numeric(htwt$Group),xlab="Height",ylab="Weight")
htwt.lm.HG=lm(Weight~Height+Group,data=htwt)
abline(coef(htwt.lm.HG)%*%cbind(c(1,0,0),c(0,1,0)),lty=1)
abline(coef(htwt.lm.HG)%*%cbind(c(1,0,1),c(0,1,0)),lty=2)
legend(55,200,legend=c("Male","Female"),pch=c(1,2),lty=c(1,2))
```
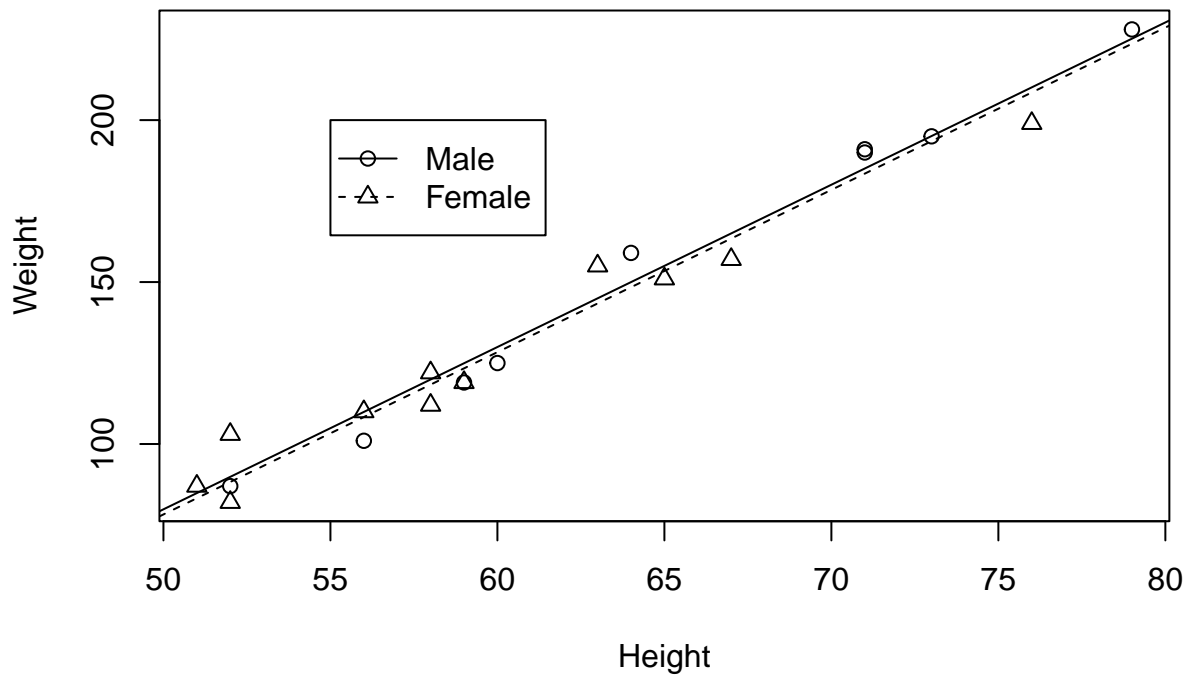

Notice that as before there does note appear to be a difference between females and males (p=0.655). However, it is clear that **Height** is predictive of **Weight** (p<0.001).

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

Weight = beta.0 + beta.1 Female + beta.2 Height + beta.3 Female x Height + epsilon

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

```
##
## 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
```
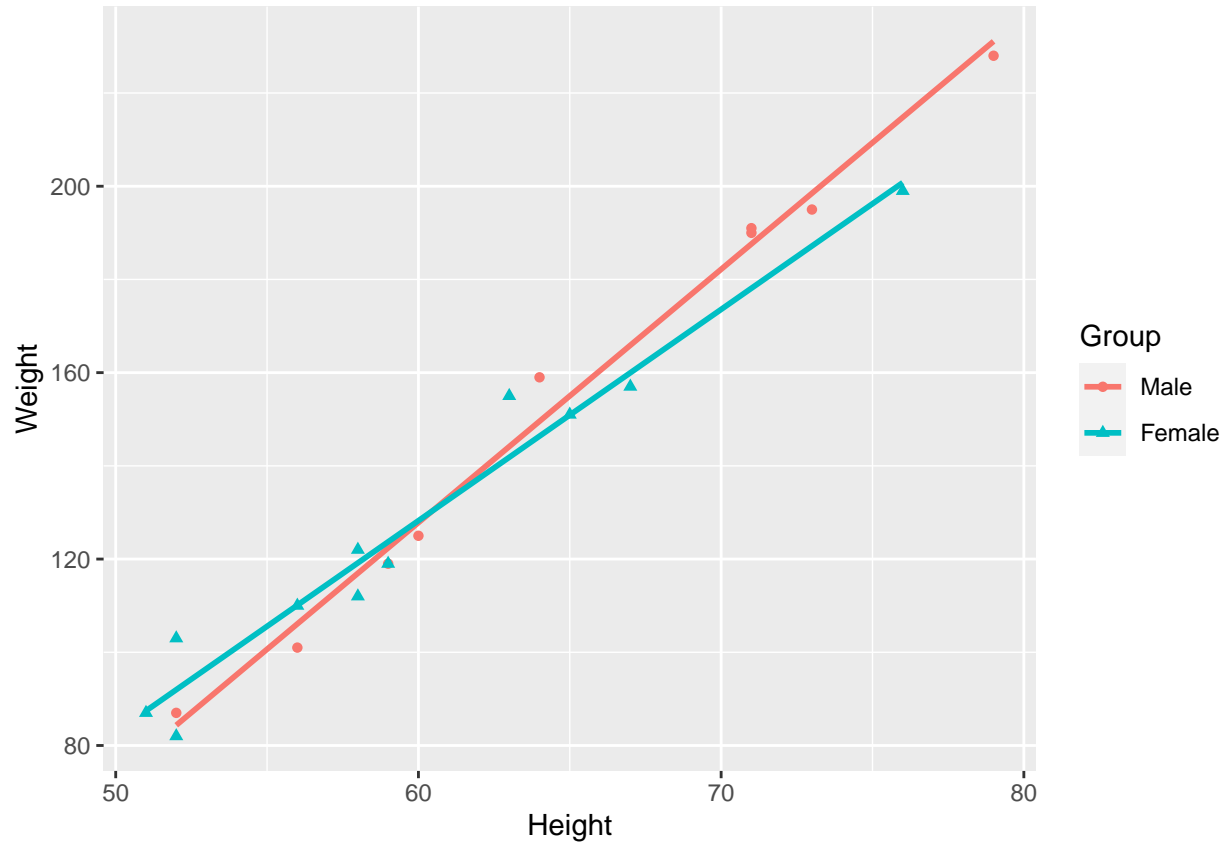
```
# Using lattice
p_load(lattice)
xyplot(Weight~Height, group=Group,data=htwt,type=c("p","r"), pch=c(1,3), col=c(5,6), key=list(text=li
```
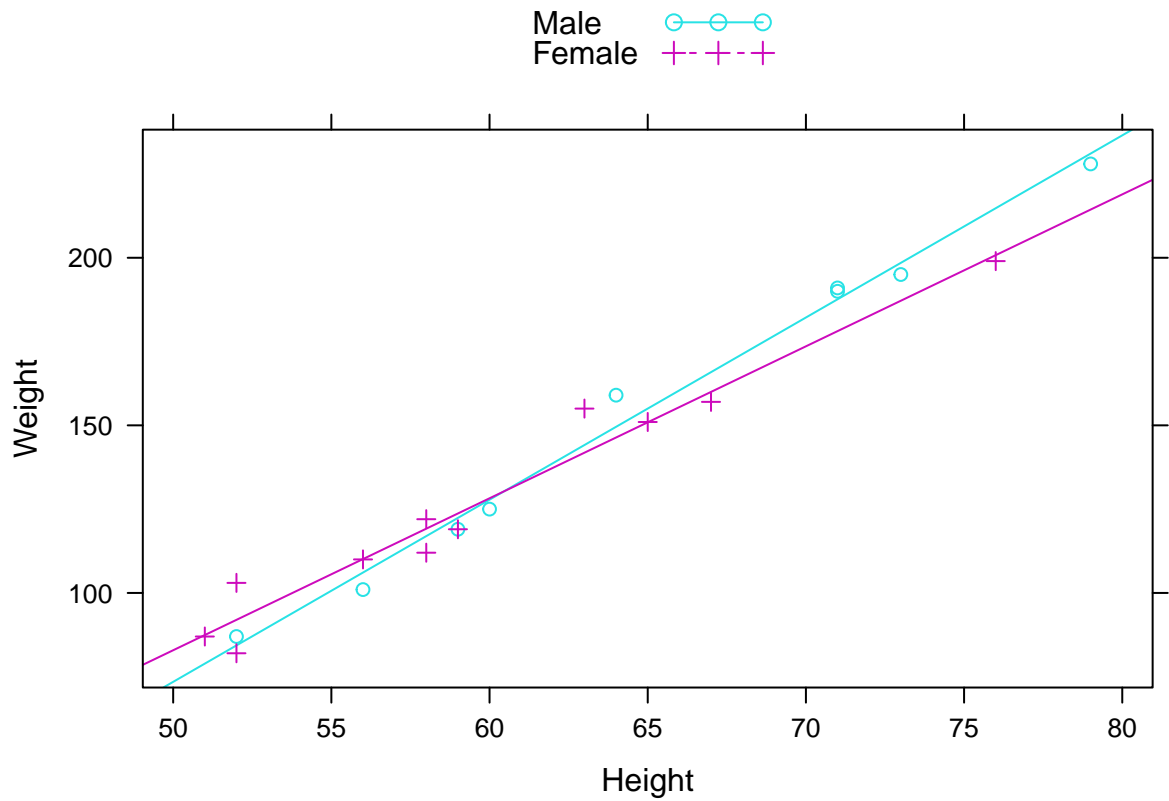
```
# Using ggplot
p_load(ggplot2)
ggplot(htwt, aes(x=Height, y=Weight, color=Group, shape=Group))+geom_point()+geom_smooth(method=lm, se
```

`geom_smooth()` using formula = 'y ~ x'

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, **htwt.lm.GbyH**. The figure shows the four diagnostic plots that are the default. An analysis of variance table may also be generated.

```
# Set up the page to take all four images
par(mfrow=c(2,2))
plot(htwt.lm.GbyH)
```
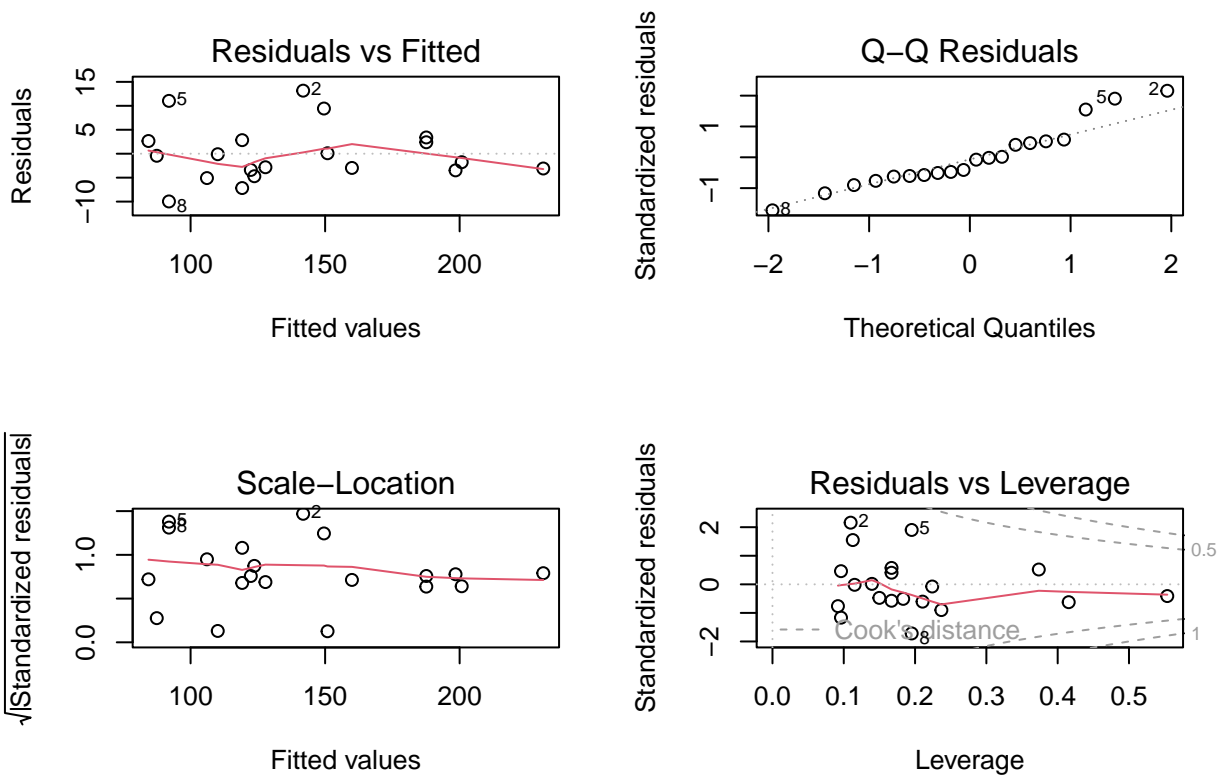
```
par(mfrow=c(1,1))
anova(htwt.lm.GbyH)
```

```
## 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.