Regression and Residuals

Oliver and JuJu

## 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

## Esitmating Weight While 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.

A linear model makes the correction fairly easy.

 p\_load(lattice)
 xyplot(Weight~Height, data=htwt)



To fit the model **Weight** = a + b **Height** we can use the “simple” equations given in class or we can use the function ***lm***.

 r = cor(htwt$Weight,htwt$Height)
 b = r \* sqrt(var(htwt$Weight))/sqrt(var(htwt$Height))
 b

## [1] 5.041217

 a = mean(htwt$Weight) - b \* mean(htwt$Height)
 a

## [1] -173.4596

 htwt.lmw = lm(Weight ~ 1 + Height, data=htwt)
 summary(htwt.lmw)

##
## Call:
## lm(formula = Weight ~ 1 + Height, data = htwt)
##
## Residuals:
## Min 1Q Median 3Q Max
## -10.6729 -5.4001 -0.6165 3.9014 14.3163
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) -173.4596 12.2080 -14.21 3.18e-11 \*\*\*
## Height 5.0412 0.1949 25.87 1.09e-15 \*\*\*
## ---
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.171 on 18 degrees of freedom
## Multiple R-squared: 0.9738, Adjusted R-squared: 0.9723
## F-statistic: 669.1 on 1 and 18 DF, p-value: 1.09e-15

 print(c(a, b))

## [1] -173.459595 5.041217

 coef(htwt.lmw)

## (Intercept) Height
## -173.459595 5.041217

The estimated reqression line is seen to be **Weight** = -173.4595952 + 5.0412173 **Height**.

The regression line can be plotted using standard R functions.

 # Using base R graphics
 plot(htwt$Height, htwt$Weight, xlab="Height", ylab="Weight")
 abline(htwt.lmw)



 # Using lattice graphics
 p\_load(lattice)
 xyplot(Weight ~ Height, data=htwt, type=c("p","r"))



 # Using ggplot2
 p\_load(ggplot2)
 ggplot(data = htwt, aes(x = Height, y = Weight)) + geom\_point() + geom\_smooth(method = "lm", se=FALSE, color="lightblue", formula = y ~ 1 + x)



## Residual Analysis

We need to find the predicted and residual values for our data. Remember that the residual is **residual = observed - predicted**.

 # Using definitions
 obsserved = htwt$Weight
 predicted = a + b \* htwt$Height
 htwt.lmw.resid = obsserved - predicted
 htwt.lmw.resid

## [1] 9.8216871 10.8629044 -7.3019648 -4.0134436 14.3162949 3.0689910
## [7] -7.8485744 -6.6837051 3.2034274 -10.6729207 0.4507313 1.1514256
## [13] 6.5331659 -3.2195302 -4.9722263 -4.9722263 -6.9310090 3.3575122
## [19] 5.5331659 -1.6837051

 # Using internal functions
 res = as.numeric(residuals(htwt.lmw))
 res

## [1] 9.8216871 10.8629044 -7.3019648 -4.0134436 14.3162949 3.0689910
## [7] -7.8485744 -6.6837051 3.2034274 -10.6729207 0.4507313 1.1514256
## [13] 6.5331659 -3.2195302 -4.9722263 -4.9722263 -6.9310090 3.3575122
## [19] 5.5331659 -1.6837051

When fitting a regression model, it is always a good idea to check the residuals to make certain that underlying model assumptions are satisfied. A first step is to plot the residuals against the explanatory variables.

 plot(htwt$Height, res, ylab="Residual")
 abline(h=0)



Since no obvious patterns exist in the residuals, we can be fairly sure that we have not missed some non-linear pattern in the data. Since the vertical spread of the residuals seems to be the same across the x values, we can assume that the variability is constant.

If the data were collected using random selection and we wish to test to see if the slope is different from zero, the residuals must be normally distributed.

 hist(res, xlab="Residual")



 qqnorm(res)
 qqline(res, lty=2)



Neither the histogram nor the normal quantile plot suggest any reason to question the normality of the residuals. Statistical testing appears to be valid in this situation.

R generates a number of plots automatically.

 plot(htwt.lmw)

