Lab 3: Crabs

Graeson Gardner

April 23, 2018

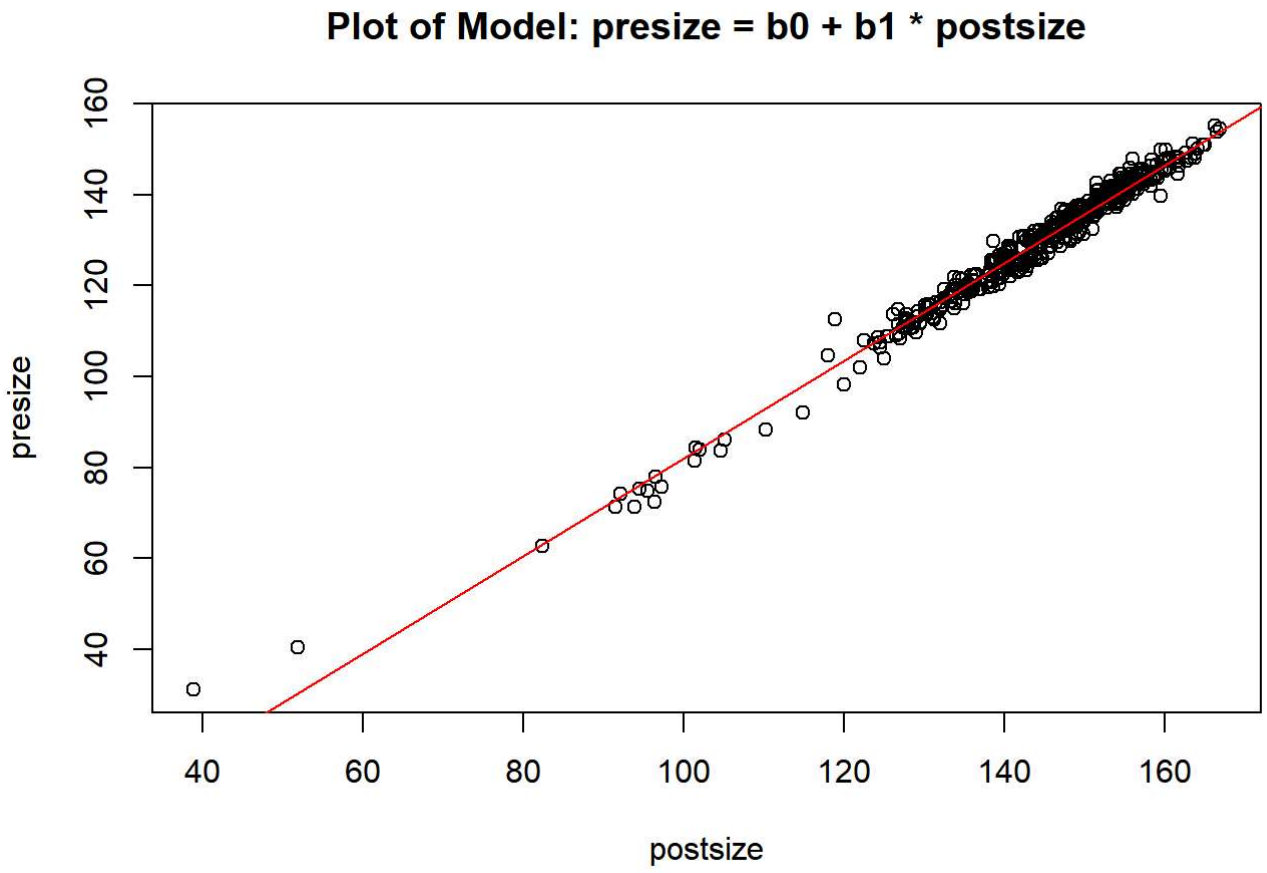
```
library(ggplot2)
setwd("C:\\Users\\ggard\\stat 135\\crabs")
molt = read.csv("crabmolt.csv")
pop = read.csv("crabpop.csv")

#PART 1
# *see part 5 for the graph I chose
# One variabe that stood out to me in particular was the difference variable included and especi
ally its relationship to the residuals. increased/decreased differences (or rather "outliers") i
n terms of difference did appear to have larger residuals, but the smallest difference was only
 the third smallest residual. The model measures expected change as a multiple of the post-valu
e, meaning change is expected to be proportionate to the size, the change for smaller values are
 not outliers in terms of difference, but rather their small size cause the change to be underes
timated. Perhaps we should include an interaction term for Crabs under 100 in size in our model.

#PART 2

#The regression done with R's buit in funtion. I will compare to the Beta's
# p, values, sd, and R^2 value I find "manually"

df = data.frame(molt$postsz, molt$presz)
names(df) = c("postsize","presize")
linr = lm(presize ~ postsize, df)
summary(linr)
```

```
##
## Call:
## lm(formula = presize ~ postsize, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1557 -1.3052  0.0564  1.3174 14.6750
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -25.21370    1.00089  -25.19   <2e-16 ***
## postsize      1.07316    0.00692  155.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.199 on 470 degrees of freedom
## Multiple R-squared:  0.9808, Adjusted R-squared:  0.9808
## F-statistic: 2.405e+04 on 1 and 470 DF,  p-value: < 2.2e-16
```

```
X = cbind(rep(1, times = length(molt$presz)), molt$postsz) # matrix of input with post size and
 vector of 1's
Y = molt$presz # vector of reponse with pre size
B = solve(t(X) %*% X) %*% t(X) %*% Y #solution to Beta
B # = -25.213703, 1.073162      These are nearly identical to the values lm() found.
```

```
##            [,1]
## [1,] -25.213703
## [2,]   1.073162
```

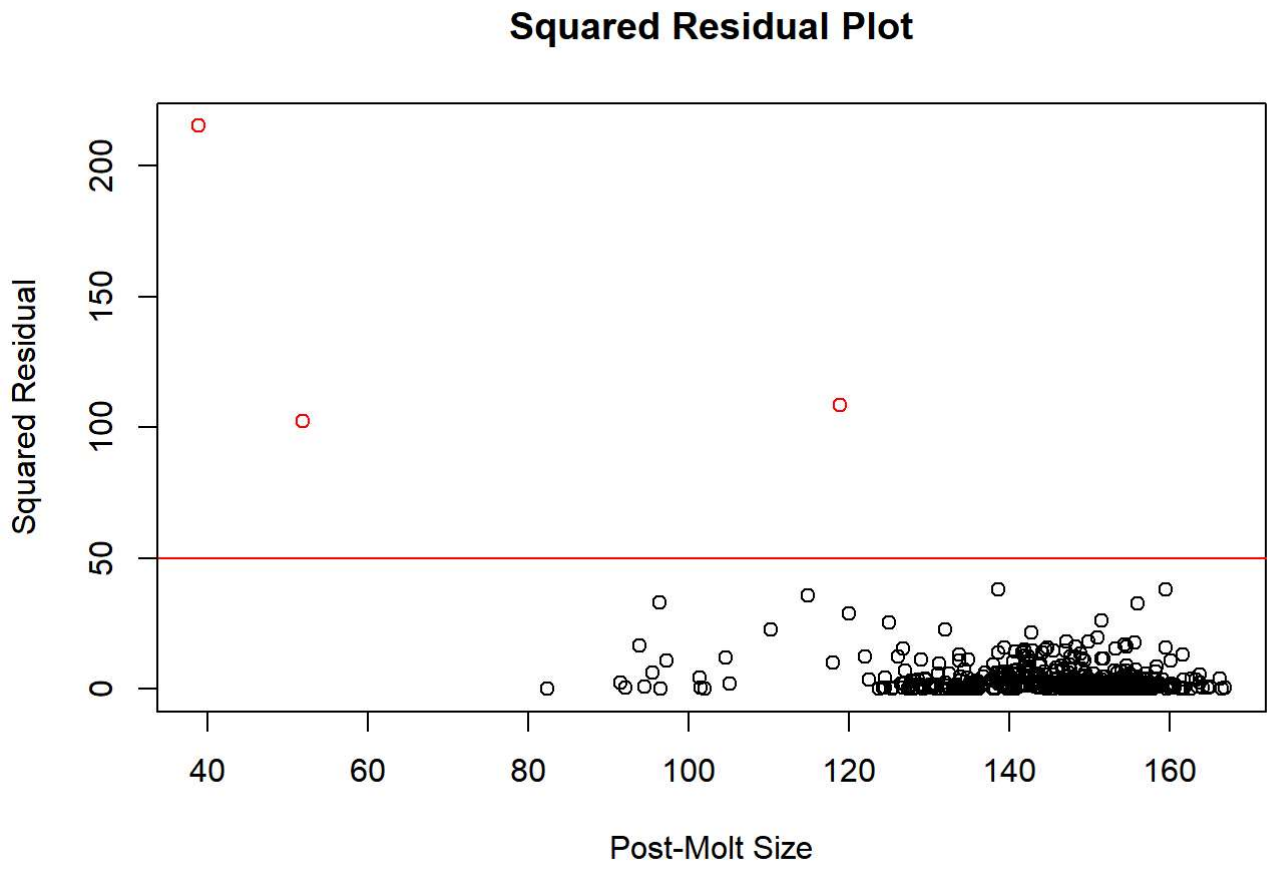
```
#PART 3
plot(df, main = "Plot of Model: presize = b0 + b1 * postsize")
abline(linr, col = "red")
```




```
# The model seems to do a good job at face value. The relationship from the scatter plot itslef
 seems very close to linear, and highly correlated.
```

```
#PART 4

# The R^2 value of .98 is very high and suggests the model predicts most (98%) variance in the
 "dependant" variable.
# Calculating R^2:
ST = (molt$presz - mean(molt$presz))^2 # sum ~= n*var(y)
SR = (molt$presz - (B[1] + B[2]*molt$postsz))^2
Rsq = 1 - sum(SR)/sum(ST) # = 0.9808326
```

```
#PART 5

resdf = data.frame(molt$postsz, SR)
resdf$color = "black"
resdf$color[SR > 50] = "red"
plot(resdf$molt.postsz, resdf$SR, col = resdf$color, main = "Squared Residual Plot", xlab = "Pos
t-Molt Size", ylab = "Squared Residual")
abline(h=50, col= "red")
```




```
molt[which(SR > 50),]
```

```
##     presz postsz  inc year lf
## 177  31.1   38.8  7.7   81  1
## 178  40.5   51.8 11.3   81  1
## 302 112.7  118.8  6.1   92  1
```

```
# I found 3 apparent major outliers that contributed the most to the SSR, each having a squared
 residual value over 50 (see plot). All outliers were lab caught, and two of them had the two sm
allest post molt values, meaning their difference from the model could be due to crab growth not
 following the same model (ex. growth may decellerate rapidly later in life compared to younger
 stages of growth).
(6.1-mean(molt$inc))/sd(molt$inc)
```

```
## [1] -3.512383
```

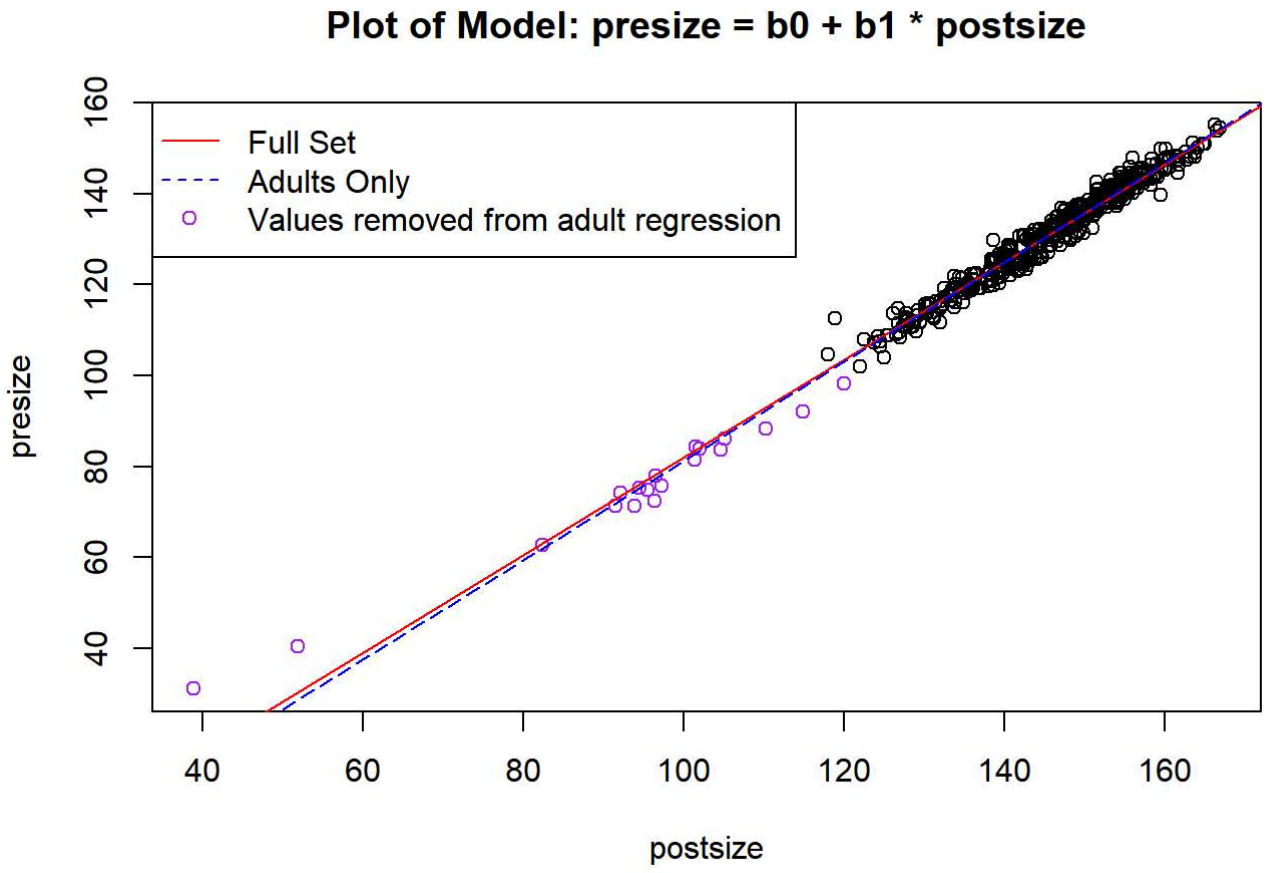
```
#The observation at 302 seems to have no apparent explanation, the change in size was unusually
 small (6.1, the smallest in the set). This is actually 3.5 SD's below the mean change in size.
 My conclusion is this there was likely measurement error or something exceptional about this cr
ab correlated with shell growth such as illness.

#plot(lm) gives some interesting graphs. The third, a plot of values vs residuals seemed the mos
t interesting and seemed to indicate similar information to my plot so I did not include it.


#PART 1
```

```
#PART 6

#The readout from lm() gives us a t-statistic of ~ 155.08 and a p-value of 2*10^(-16) which mean
s the value Beta 1 is highly signifigant; we reject the null that Beta1 = 0 with confidence 1 -
 2*10^(-16) ~ 99.99%

#Now, I will find the SD of Beta1 to replicate the lm() readout
e = molt$presz - (B[1] + B[2]*molt$postsz)
s2 = (t(e) %*% e)/(length(e) - 2) #error^2/(n-2)
seBeta = (s2[1]*solve(t(X) %*% X))^.5 # ~ 1.000889, 0.006919915      these are the same SD's as
 lm() returned.

# H0: Beta1 = 0 (99% confidence two sided)
Beta1 = B[2] = # 1.073162
seBeta1 = seBeta[2,2]
tstat = (Beta1 - 0)/seBeta1 # = 154.91, about .2 less than the t-stat lm found

#We reject at 99% confidence if this above 2.58, which it is.

# The following line returns the probability Beta1 is greater than its estimated value of 1.07 u
nder the null hypothesis, which turns out to be effectively zero.
1-pnorm(tstat) # ~0
```

```
## [1] 0.1586553
```

```
#So again, we can say with 99% (arguably 99.999999%) certainty  that there is a positive relatio
nship between pre and post molt size (under the assumptions our hypothesis test).
```

```
#PART 7

col2 = rep("black", length(df$postsize))
col2[df$presize <= 100] = "purple"
dfA = df[df$presize >= 100,] # 19 observations deleted
linA = lm(presize ~ postsize, dfA)
plot(df, main = "Plot of Model: presize = b0 + b1 * postsize", col = col2)
abline(linr, col = "red")
abline(linA, col = "blue", lty = "longdash")
legend("topleft",legend=c("Full Set", "Adults Only", "Values removed from adult regression" ),lt
y=c(1:2,NA), pch =c(NA,NA,1), col=c("red", "blue","purple"))
```




```
#PART 8

molted = data.frame(size  = pop[pop$shell == 1,]$size)
unmolted = data.frame(size = pop[pop$shell == 0,]$size)
#We take the set of crabs that molted this year and use our regeression to attempt to predict wh
at their size was prior to molting
moltpresize = data.frame(size = predict(linr, data =molted))

summary(moltpresize)
```

```
##       size
##  Min.   : 16.43
##  1st Qu.:122.86
##  Median :132.97
##  Mean   :129.21
##  3rd Qu.:139.44
##  Max.   :153.79
```

```
ggplot() +
  geom_histogram(data = moltpresize,aes(size, legend = "Distribution of Predicted Presize for Mo
lted Crabs"), fill = "blue",  alpha = .9) +
  geom_histogram(data = unmolted,aes(size,legend  = "Size of Crabs that have not Molted"), fill
 = "red", alpha = .5)+
  xlab("Count") +
ylab("Size")+ggtitle("Size Counts of Crabs in 1983")
```

```
## Warning: Ignoring unknown aesthetics: legend

## Warning: Ignoring unknown aesthetics: legend
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```