Chapter 7 Question 11

Rahul Sirasao

3/5/2017

At https://archive.ics.uci.edu/ml/datasets/Abalone, you will find a dataset of measurements by W. J. Nash, T. L. Sellers, S. R. Talbot, A. J. Cawthorn and W. B. Ford, made in 1992. These are a variety of measurements of blacklip abalone (Haliotis rubra; delicious by repute) of various ages and genders.

a. Build a linear regression predicting the age from the measurements, ig- noring gender. Plot the residual against the fitted values.

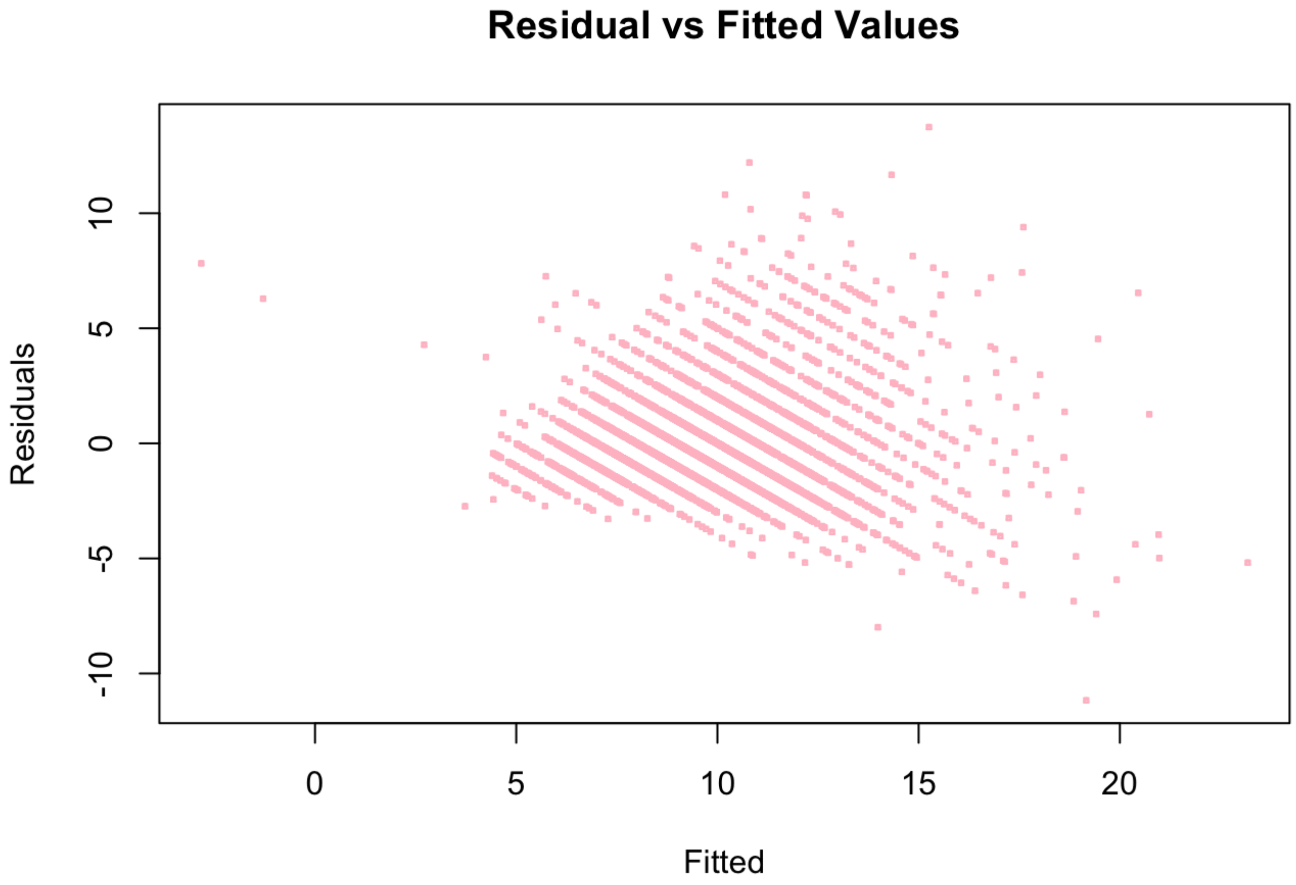
```
abalone_data <- read.csv("~/Desktop/abalone.data.csv", header = FALSE)
#View(abalone_data)

#In R, the lm(), or “linear model,” function can be used to create a simple regress
ion model                Since we are trying to predict age based on measurements (wh
ile ignoring gender), it is important that we only consider columns B to H (which c
orrespond to measurments length, diameter, height, and multiple weights) as well as
column I which gives us the age. We can assume here than a bigger, meatier plant wi
ll be older than a smaller one (however, that is for our graph to portray). Our pre
dictors would be columns B to H.

size_to_age_model = lm(abalone_data$V9 ~ abalone_data$V2 + abalone_data$V3 + abalo
ne_data$V4 + abalone_data$V5 + abalone_data$V6 + abalone_data$V7 + abalone_data$V8
, data = abalone_data)
anova(size_to_age_model) #Let's just check out the summary
```

```
## Analysis of Variance Table
##
## Response: abalone_data$V9
##                   Df  Sum Sq Mean Sq   F value    Pr(>F)
## abalone_data$V2    1 13454.5 13454.5 2735.4106 < 2.2e-16 ***
## abalone_data$V3    1  1059.0  1059.0  215.2985 < 2.2e-16 ***
## abalone_data$V4    1   920.7   920.7  187.1888 < 2.2e-16 ***
## abalone_data$V5    1     0.6     0.6    0.1163    0.7331
## abalone_data$V6    1  6632.6  6632.6 1348.4652 < 2.2e-16 ***
## abalone_data$V7    1   557.3   557.3  113.3068 < 2.2e-16 ***
## abalone_data$V8    1   280.0   280.0   56.9213 5.536e-14 ***
## Residuals       4169 20505.9     4.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
abalone.res = resid(size_to_age_model) #Let us get the residuals
abalone.predict = predict(size_to_age_model, data.frame(abalone_data[c(1:7)]))
plot(abalone.predict, abalone.res,pch = 14, cex = .3, col = "pink",xlab="Fitted",
ylab="Residuals", main="Residual vs Fitted Values") #Plot
```



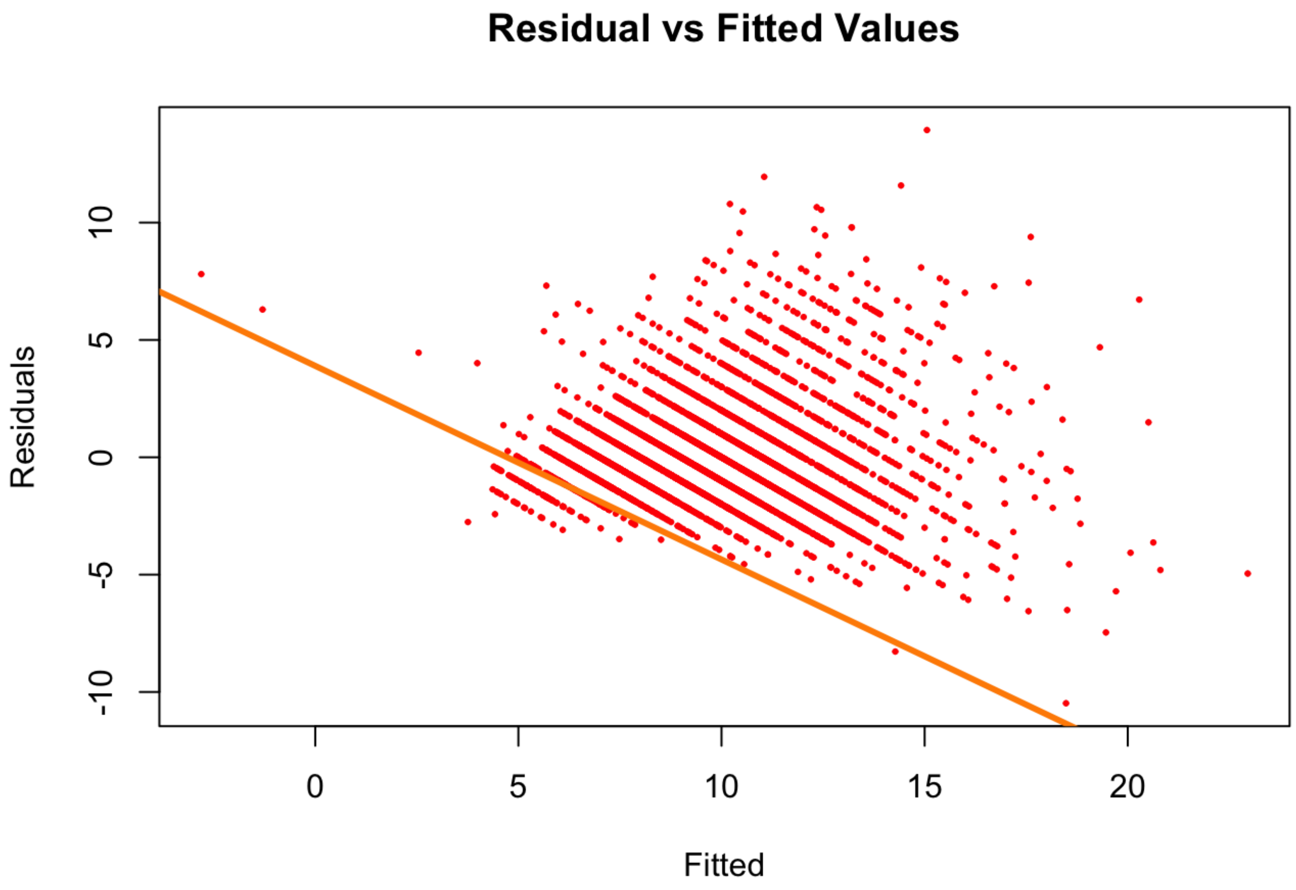
b. Build a linear regression predicting the age from the measurements, including gender. There are three levels for gender; I'm not sure whether this has to do with abalone biology or difficulty in determining gender. You can represent gender numerically by choosing 1 for one level, 0 for another, and -1 for the third. Plot the residual against the fitted values.

```
abalone_data <- read.csv("~/Desktop/abalone.data.csv", header = FALSE)
abalone_data$V1 <- factor(abalone_data$V1) ##Using factor to create the categorical gender as numeric
size_to_age_model_gen = lm(abalone_data$V9 ~ abalone_data$V1 + abalone_data$V2 + abalone_data$V3 + abalone_data$V4 + abalone_data$V5 + abalone_data$V6 + abalone_data$V7 + abalone_data$V8, data = abalone_data)
anova(size_to_age_model_gen) #Let's just check out the summary
```

```
## Analysis of Variance Table
##
## Response: abalone_data$V9
##                     Df  Sum Sq Mean Sq   F value    Pr(>F)
## abalone_data$V1     2  8381.1  4190.6  870.4591 < 2.2e-16 ***
## abalone_data$V2     1  6143.0  6143.0 1276.0098 < 2.2e-16 ***
## abalone_data$V3     1   777.7   777.7  161.5528 < 2.2e-16 ***
## abalone_data$V4     1   750.7   750.7  155.9349 < 2.2e-16 ***
## abalone_data$V5     1     2.7     2.7    0.5621    0.4534
## abalone_data$V6     1  6380.3  6380.3 1325.3191 < 2.2e-16 ***
## abalone_data$V7     1   623.5   623.5  129.5179 < 2.2e-16 ***
## abalone_data$V8     1   290.8   290.8   60.4094 9.639e-15 ***
## Residuals        4167 20060.7     4.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
abalone.res = resid(size_to_age_model_gen) #Resids
abalone.predict = predict(size_to_age_model_gen, data.frame(abalone_data[c(0:7)]))
#Include the first (index[0])
plot(abalone.predict, abalone.res,pch = 10, cex = .3, col = "red",xlab="Fitted", y
lab="Residuals", main="Residual vs Fitted Values")
abline(size_to_age_model_gen, lwd = 3, col = "darkorange") #Let us check out the fi
tted line to the scatterplot. It looks good and is consistent to the trend!
```

```
## Warning in abline(size_to_age_model_gen, lwd = 3, col = "darkorange"): only
## using the first two of 10 regression coefficients
```



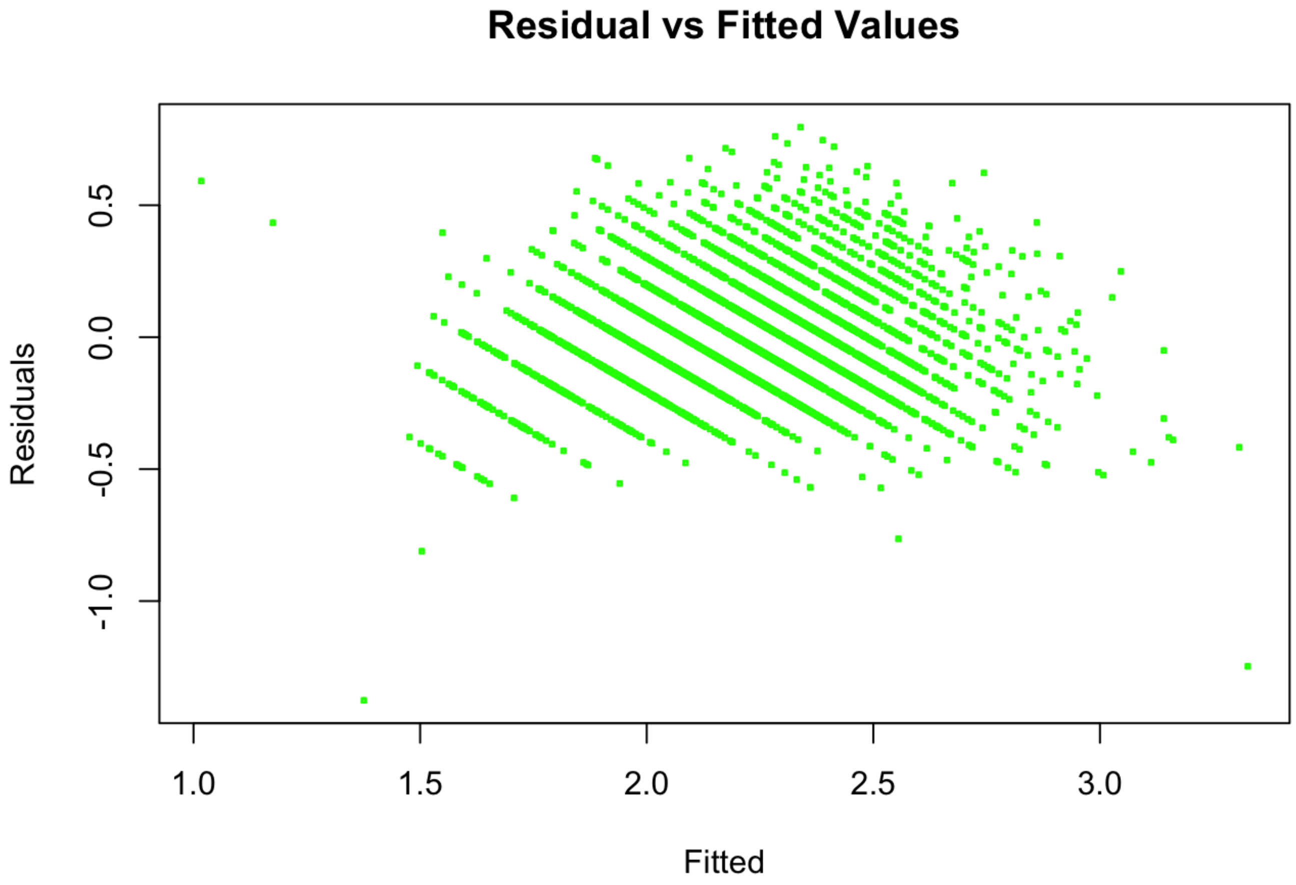
c. Now build a linear regression predicting the log of age from the measurements, ignoring gender. Plot the residual against the fitted values.

```
#We can use part a again as we are ignoring gender
abalone_data <- read.csv("~/Desktop/abalone.data.csv", header = FALSE)

size_to_age_model_log = lm(log(abalone_data$V9) ~ abalone_data$V2 + abalone_data$V
3 + abalone_data$V4 + abalone_data$V5 + abalone_data$V6 + abalone_data$V7 + abalon
e_data$V8, data = abalone_data) #Log of Age
anova(size_to_age_model_log) #Let's just check out the summary
```

```
## Analysis of Variance Table
##
## Response: log(abalone_data$V9)
##                   Df  Sum Sq Mean Sq  F value    Pr(>F)
## abalone_data$V2    1 182.221 182.221 4298.572 < 2.2e-16 ***
## abalone_data$V3    1   8.739   8.739  206.143 < 2.2e-16 ***
## abalone_data$V4    1   6.715   6.715  158.417 < 2.2e-16 ***
## abalone_data$V5    1   2.977   2.977   70.231 < 2.2e-16 ***
## abalone_data$V6    1  44.547  44.547 1050.855 < 2.2e-16 ***
## abalone_data$V7    1   3.091   3.091   72.915 < 2.2e-16 ***
## abalone_data$V8    1   1.316   1.316   31.039 2.689e-08 ***
## Residuals       4169 176.729   0.042
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
abalone.res = resid(size_to_age_model_log) #Let us get the residuals
abalone.predict = predict(size_to_age_model_log, data.frame(abalone_data[c(1:7)]))
plot(abalone.predict, abalone.res,pch = 14, cex = .3, col = "green",xlab="Fitted",
ylab="Residuals", main="Residual vs Fitted Values") #Plot
```

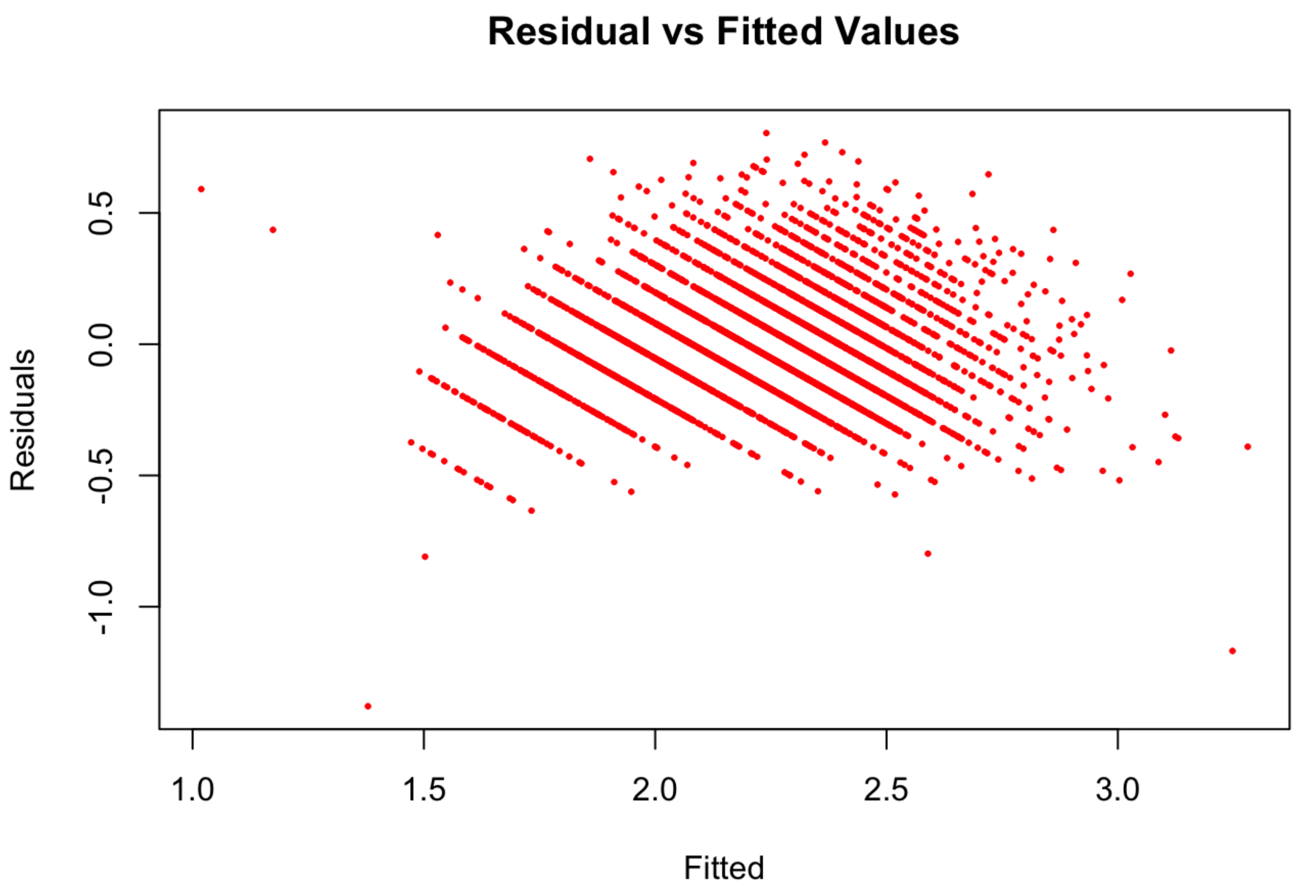


d. Now build a linear regression predicting the log age from the measurements, including gender, represented as above. Plot the residual against the fitted values.

```
#We can use part b because we are looking at gender
abalone_data <- read.csv("~/Desktop/abalone.data.csv", header = FALSE)
abalone_data$V1 <- factor(abalone_data$V1) ##Using factor to create the categorical gender as numeric
size_to_age_model_genl = lm(log(abalone_data$V9) ~ abalone_data$V1 + abalone_data$V2 + abalone_data$V3 + abalone_data$V4 + abalone_data$V5 + abalone_data$V6 + abalone_data$V7 + abalone_data$V8, data = abalone_data)
anova(size_to_age_model_genl) #Let's just check out the summary
```

```
## Analysis of Variance Table
##
## Response: log(abalone_data$V9)
##                   Df  Sum Sq Mean Sq  F value    Pr(>F)
## abalone_data$V1    2 103.228  51.614 1258.222 < 2.2e-16 ***
## abalone_data$V2    1  89.301  89.301 2176.937 < 2.2e-16 ***
## abalone_data$V3    1   6.238   6.238  152.071 < 2.2e-16 ***
## abalone_data$V4    1   5.277   5.277  128.643 < 2.2e-16 ***
## abalone_data$V5    1   3.942   3.942   96.105 < 2.2e-16 ***
## abalone_data$V6    1  42.344  42.344 1032.246 < 2.2e-16 ***
## abalone_data$V7    1   3.666   3.666   89.370 < 2.2e-16 ***
## abalone_data$V8    1   1.401   1.401   34.161 5.463e-09 ***
## Residuals       4167 170.936   0.041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
abalone.res = resid(size_to_age_model_genl) #Resids
abalone.predict = predict(size_to_age_model_genl, data.frame(abalone_data[c(0:7)])
) #Include the first (index[0])
plot(abalone.predict, abalone.res,pch = 10, cex = .3, col = "red",xlab="Fitted", y
lab="Residuals", main="Residual vs Fitted Values")
```



d. Now build a linear regression predicting the log age from the measure- ments, including gender, represented as above. Plot the residual against the fitted values.

e. It turns out that determining theage of anabalone is possible,but difficult (you section the shell, and count rings). Use your plots to explain which regression you would use to replace this procedure, and why.

f. Can you improve these regressions by using a regularizer? Use glmnet to obtain plots of the cross-

Loading [Contrib]/a11y/accessibility-menu.js