HW05

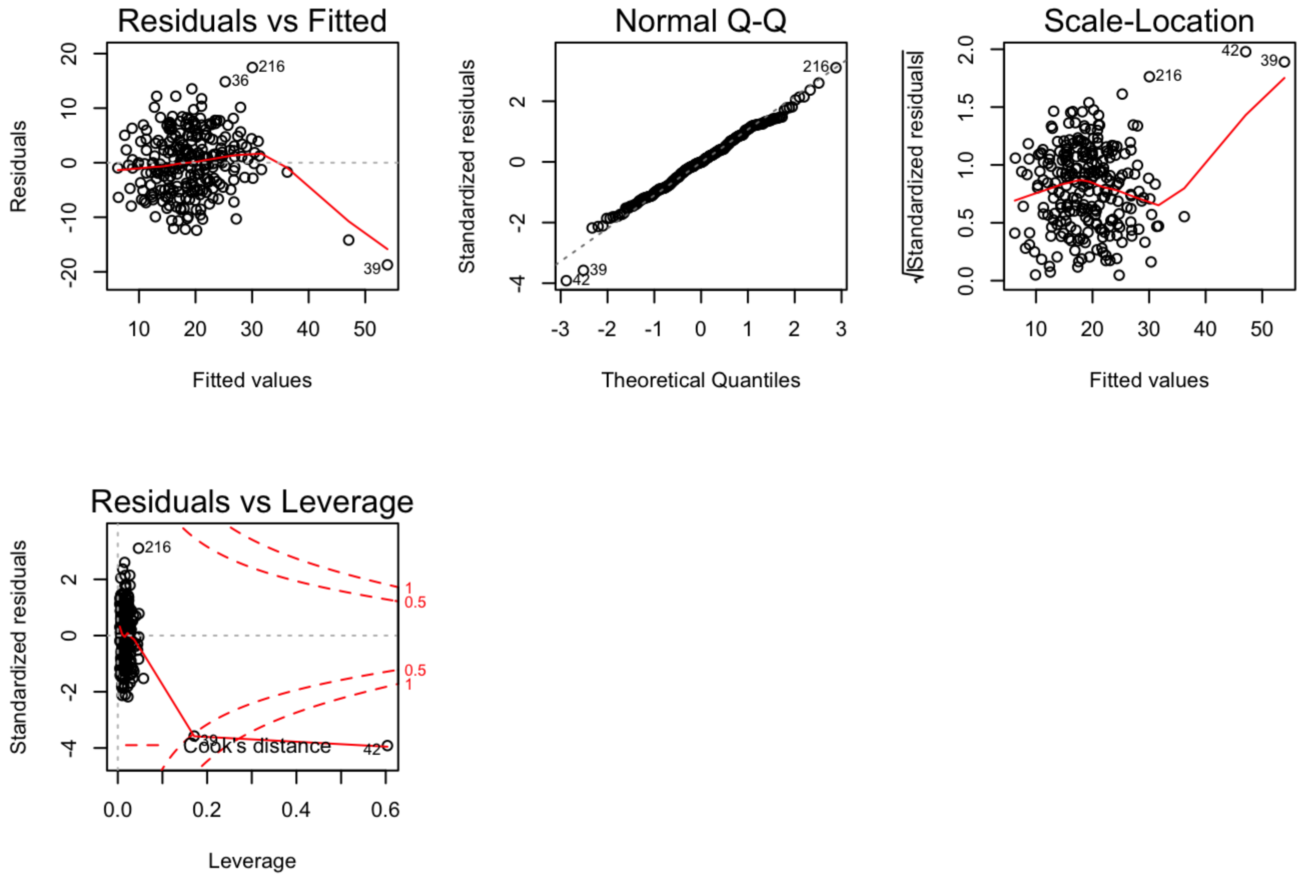
caojilin

10/24/2018

Problem 1

When I run this code, R gives me four plots. Describe each of these plots and explain how to interpret them.(0.75 points)

```
g = lm(bodyfat ~ Age +Weight + Height+Thigh,data=body)
par(mfrow = c(2, 3))
plot(g)
```



First is the residual against fitted values plot. We can use it to check assumption of linearity and the constant variance. The constant variance assupmtion is true if we see there is no pattern for residuals and they are equally spread around x-axis. Linearity assumption fails if we see very large residuals. We expect to see residuals that are not far away from 0. We also expect the red line to be linear. But we see that there are some points pull the red line away from y=0. These are potential outliers.

Second is qqplot. sorted values x_i vs $\Phi^{-1}(\frac{i}{n})$ We can use it to check the normality assupmtion of the error terms. If we observe a stright line, then our assumption is true. We see left tail is heavier than normal. Those points are potential outliers.

Third is scale-location plot. It’s square rooted standardized residual vs. predicted value. This is useful for checking the assumption of homoscedasticity. We are checking to see if there is a pattern in the residuals. Similar to the first graph. Some data points pull the red line up. Those are potential outliers.

Fourth is the residaul vs leverage. The cook’s distance is interpreted by the red contours. It is a measure of the influence of each observation on the regression coefficients. It measures the distance between $\hat{\beta}\ and\ \hat{\beta}_{[i]}$. Any observation for which the cook’s distance is substantially larger than other Cook’s distances requires investigation.

Problem 2

a) Using each of the following methods, perform variable selection to select a subset of the explanatory variables for modeling the response: (0.25 points each = 1.5 points)

i. Backward elimination using the individual p-values.

```
lmod2 = lm(bodyfat ~ Age + Weight + Height + Neck + Chest + Abdomen+ Hip +
             Thigh+ Knee + Ankle + Biceps + Forearm+ Wrist, data=body)
#cirtical value = 0.15
summary(lmod2)
lmod2 = update(lmod2,. ~. -Knee)
summary(lmod2)
lmod2 = update(lmod2,. ~. -Chest)
summary(lmod2)
lmod2 = update(lmod2,. ~. -Height)
summary(lmod2)
lmod2 = update(lmod2,. ~. -Ankle)
summary(lmod2)
lmod2 = update(lmod2,. ~. -Biceps)
summary(lmod2)
lmod2 = update(lmod2,. ~. -Hip)
summary(lmod2)
#Age Weight Neck Abdomen Thigh Forearm  Wrist
lmod2.backward = lmod2
```

ii. Forward Selection using p-values.

```
null=lm(bodyfat ~ 1, body)
full = lm(bodyfat ~ Age + Weight + Height + Neck + Chest + Abdomen+ Hip +
             Thigh+ Knee + Ankle + Biceps + Forearm+ Wrist, data=body)
step(null, scope=list(lower=null, upper=full), direction="forward")
lmod2.forward = lm(bodyfat ~Abdomen+Weight+Wrist+Forearm+Neck+Age+Thigh,data=body)
```

iii. Adjusted R2

```
b<-regsubsets(bodyfat~Age + Weight + Height + Neck + Chest + Abdomen+ Hip +
             Thigh+ Knee + Ankle + Biceps + Forearm+ Wrist,data=body)
rs = summary(b)
#we select the model with highest adjusted R-squares
rs$which[which(rs$adjr2 == max(rs$adjr2)),]
```

```
## (Intercept)         Age      Weight      Height        Neck       Chest 
##        TRUE        TRUE        TRUE       FALSE        TRUE       FALSE 
##     Abdomen         Hip       Thigh        Knee       Ankle      Biceps 
##        TRUE        TRUE        TRUE       FALSE       FALSE       FALSE 
##     Forearm       Wrist 
##        TRUE        TRUE 
```

```
lmod2.adjr2 = lm(bodyfat ~ Age+Weight+Neck+Abdomen+Hip+Thigh+Forearm+Wrist, body)
# plot(b,scale="adjr2")
```

iv. AIC

```
lmod2 = lm(bodyfat ~ Age + Weight + Height + Neck + Chest + Abdomen+ Hip +
             Thigh+ Knee + Ankle + Biceps + Forearm+ Wrist, data=body)
step(lmod2)
lmod2.AIC = lm(formula = bodyfat ~ Age + Weight + Neck + Abdomen + Hip + 
    Thigh + Forearm + Wrist, data = body)
```

v. BIC

```
lmod2 = lm(bodyfat ~ Age + Weight + Height + Neck + Chest + Abdomen+ Hip +
             Thigh+ Knee + Ankle + Biceps + Forearm+ Wrist, data=body)
step(lmod2,k = log(nrow(body)))
lmod2.BIC = lm(formula = bodyfat ~ Weight + Abdomen + Forearm + Wrist, data = body)
```

vi. Mallow’s Cp

```
rs$which
```

```
##   (Intercept)   Age Weight Height  Neck Chest Abdomen   Hip Thigh  Knee
## 1        TRUE FALSE  FALSE  FALSE FALSE FALSE    TRUE FALSE FALSE FALSE
## 2        TRUE FALSE   TRUE  FALSE FALSE FALSE    TRUE FALSE FALSE FALSE
## 3        TRUE FALSE   TRUE  FALSE FALSE FALSE    TRUE FALSE FALSE FALSE
## 4        TRUE FALSE   TRUE  FALSE FALSE FALSE    TRUE FALSE FALSE FALSE
## 5        TRUE FALSE   TRUE  FALSE  TRUE FALSE    TRUE FALSE FALSE FALSE
## 6        TRUE  TRUE   TRUE  FALSE FALSE FALSE    TRUE FALSE  TRUE FALSE
## 7        TRUE  TRUE   TRUE  FALSE  TRUE FALSE    TRUE FALSE  TRUE FALSE
## 8        TRUE  TRUE   TRUE  FALSE  TRUE FALSE    TRUE  TRUE  TRUE FALSE
##   Ankle Biceps Forearm Wrist
## 1 FALSE  FALSE   FALSE FALSE
## 2 FALSE  FALSE   FALSE FALSE
## 3 FALSE  FALSE   FALSE  TRUE
## 4 FALSE  FALSE    TRUE  TRUE
## 5 FALSE  FALSE    TRUE  TRUE
## 6 FALSE  FALSE    TRUE  TRUE
## 7 FALSE  FALSE    TRUE  TRUE
## 8 FALSE  FALSE    TRUE  TRUE
```

```
rs$cp
```

```
## [1] 72.868837 20.690746 14.210205  9.314331  8.559272  7.664855  6.337654
## [8]  6.367146
```

```
#row 7 has smallest cp
lmod2.cp = lm(bodyfat ~ Age + Weight+Neck+Abdomen+Thigh+Forearm+Wrist , data=body)
```

b) Let M1,…,M6 denote the six models selected by each of the six variable selection methods of the previous part. Select one of these models by cross-validation. (0.2 points)

```
lmod2.backward.score = sum((lmod2.backward$residuals/(1 - influence(lmod2.backward)$ha
t))^2)
lmod2.forward.score = sum((lmod2.forward$residuals/(1 - influence(lmod2.forward)$hat))^2
)
lmod2.adjr2.score = sum((lmod2.adjr2$residuals/(1 - influence(lmod2.adjr2)$hat))^2)
lmod2.AIC.score = sum((lmod2.AIC$residuals/(1 - influence(lmod2.AIC)$hat))^2)
lmod2.BIC.score = sum((lmod2.BIC$residuals/(1 - influence(lmod2.BIC)$hat))^2)
lmod2.cp.score = sum((lmod2.cp$residuals/(1-influence(lmod2.cp)$hat))^2)
score = c(lmod2.backward.score,lmod2.forward.score,lmod2.adjr2.score,lmod2.AIC.score,lmo
d2.BIC.score,lmod2.cp.score)
score
```

```
## [1] 4840.639 4840.639 4829.885 4829.885 4908.053 4840.639
```

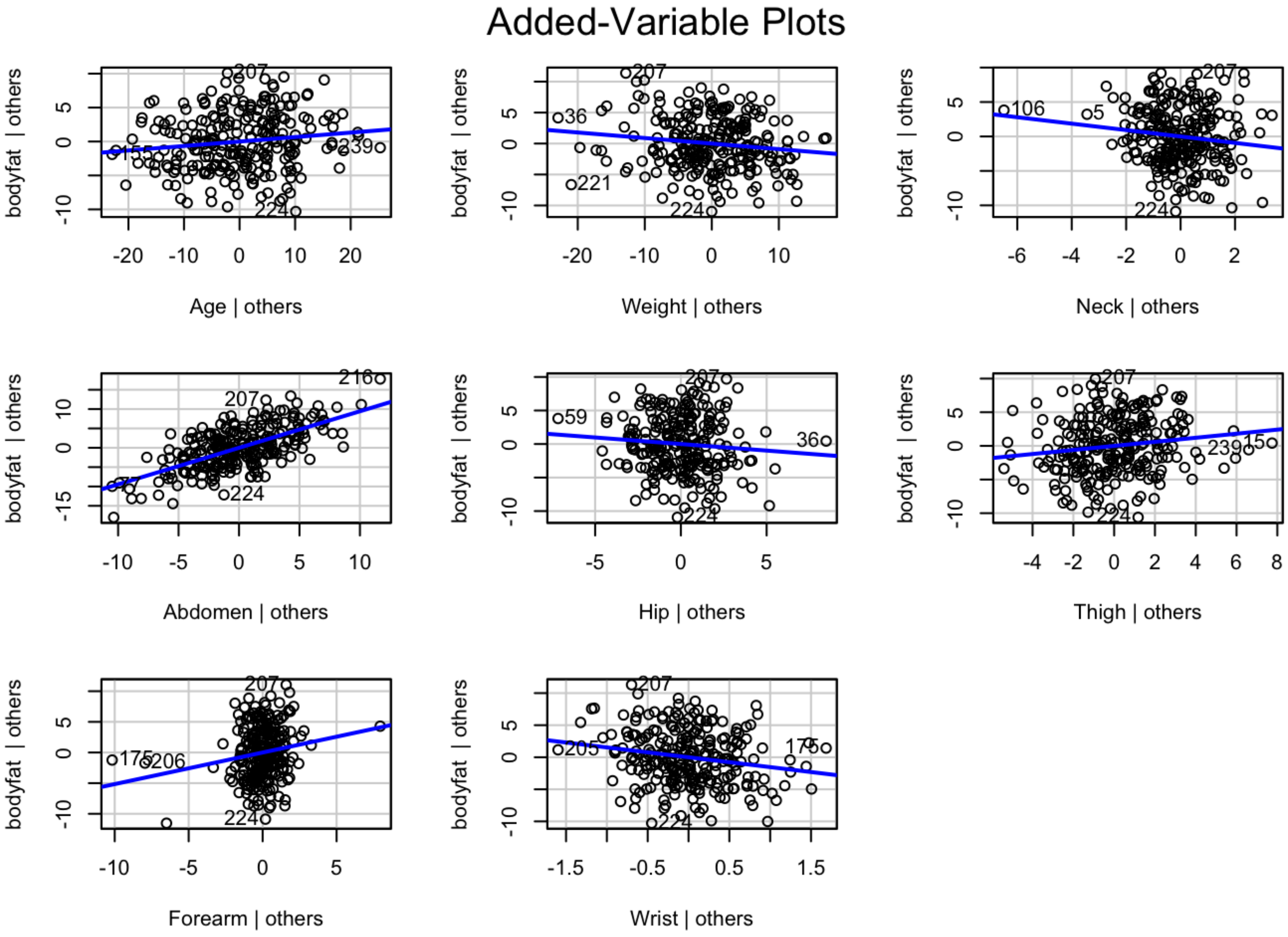
we notice that both AIC and adjusted R-square have the same score and they selected the same variables. So choose either one.

c) Let M be the model selected in the previous part. Fit this model to the data. Perform regression diagnostics. Comment on the validity of the assumptions of the linear model. Identify infuential observations and outliers. Delete them if necessary and re-fit the model.

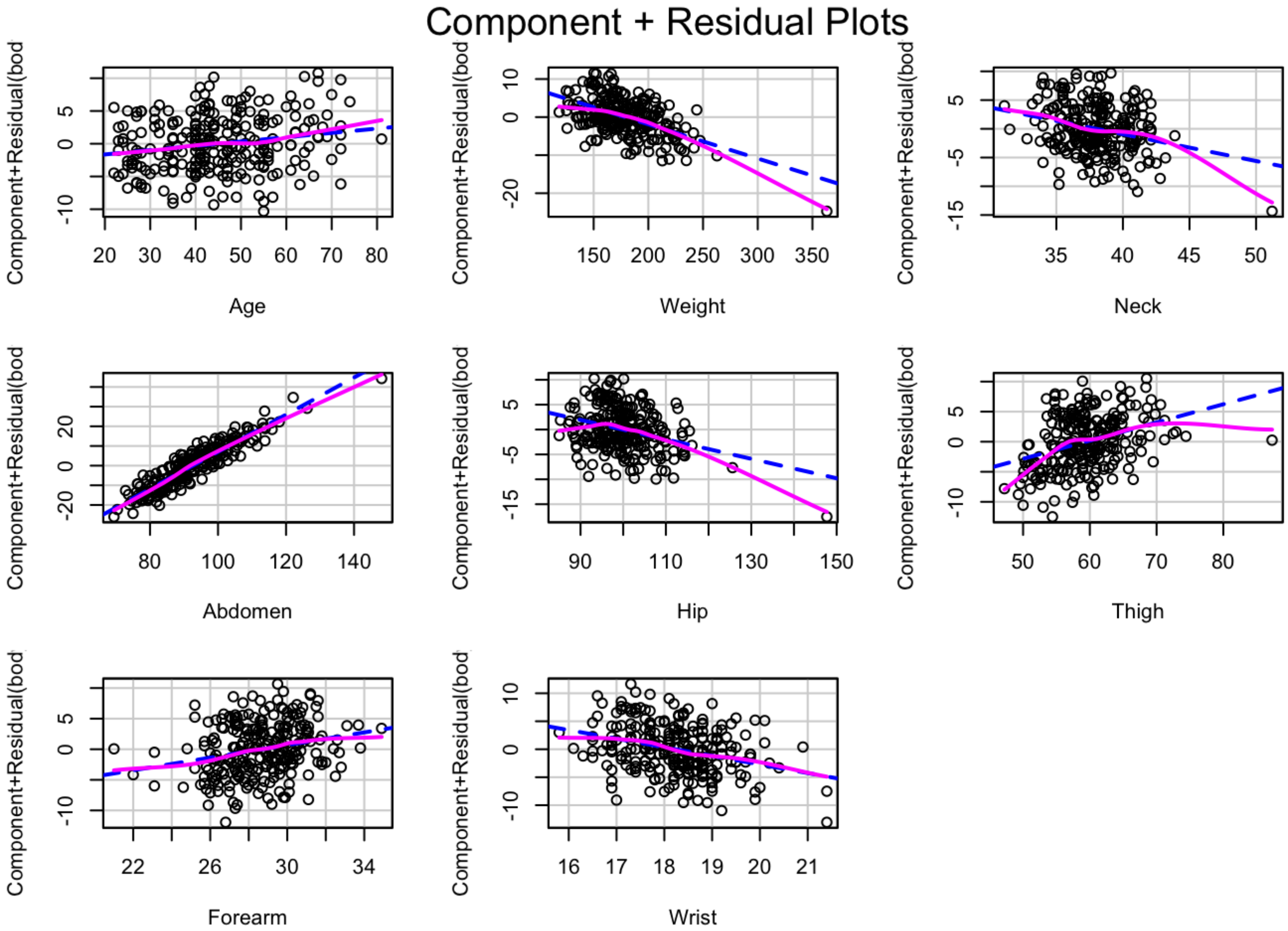
check linear model assumption

We see that there the purple loess line is influnced by some potential outliers.

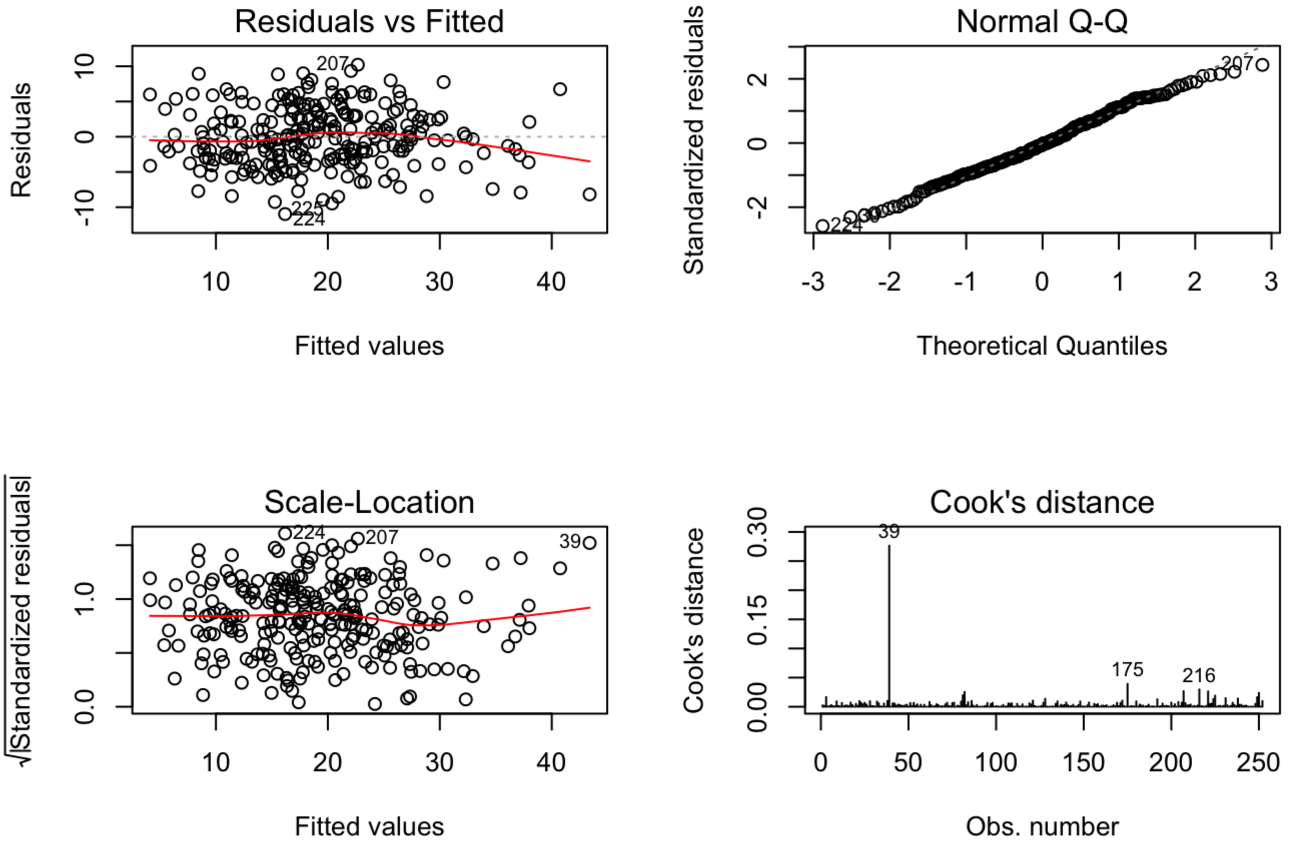
```
# Partial regression plots are useful for identifying points with high leverage and infl
uential data points that might not have high leverage
avPlots(lmod2.AIC)
```



```
# partial residual plot identify the type of relationship between y and each xi (given the effects of the other xj).
crPlots(lmod2.AIC)
```



```
par(mfrow = c(2,2))
plot(lmod2.AIC,which=1:4)
```



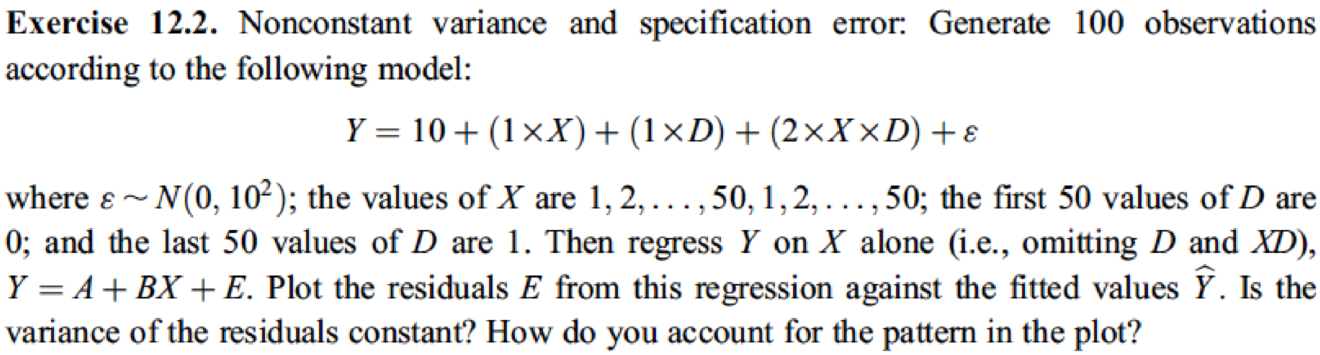
the constant variance assumption is not violated, and we didn’t see a pattern for residuals, so linear assumption is not violated.

```
#check the potential outliers
body[c(39,175,207,216,224,225),]
```

```
##     Density bodyfat Age Weight Height Neck Chest Abdomen   Hip Thigh Knee
## 39   1.0202    35.2  46 363.15  72.25 51.2 136.2   148.1 147.7  87.3 49.1
## 175  1.0414    25.3  36 226.75  71.75 41.5 115.3   108.8 114.4  69.2 42.4
## 207  1.0250    32.9  44 166.00  65.50 39.1 100.6    93.9 100.1  58.9 37.6
## 216  0.9950    47.5  51 219.00  64.00 41.2 119.8   122.1 112.8  62.5 36.9
## 224  1.0874     5.2  55 142.25  67.25 35.2  92.7    82.8  91.9  54.4 35.2
## 225  1.0740    10.9  55 179.75  68.75 41.1 106.9    95.3  98.2  57.4 37.1
##     Ankle Biceps Forearm Wrist
## 39   29.6   45.0    29.0  21.4
## 175  24.0   35.4    21.0  20.1
## 207  21.4   33.1    29.5  17.3
## 216  23.6   34.7    29.1  18.4
## 224  22.5   29.4    26.8  17.0
## 225  21.8   34.1    31.1  19.2
```

These observations seem acceptable. We do not have enough reason to remove these points.

Probelm 3

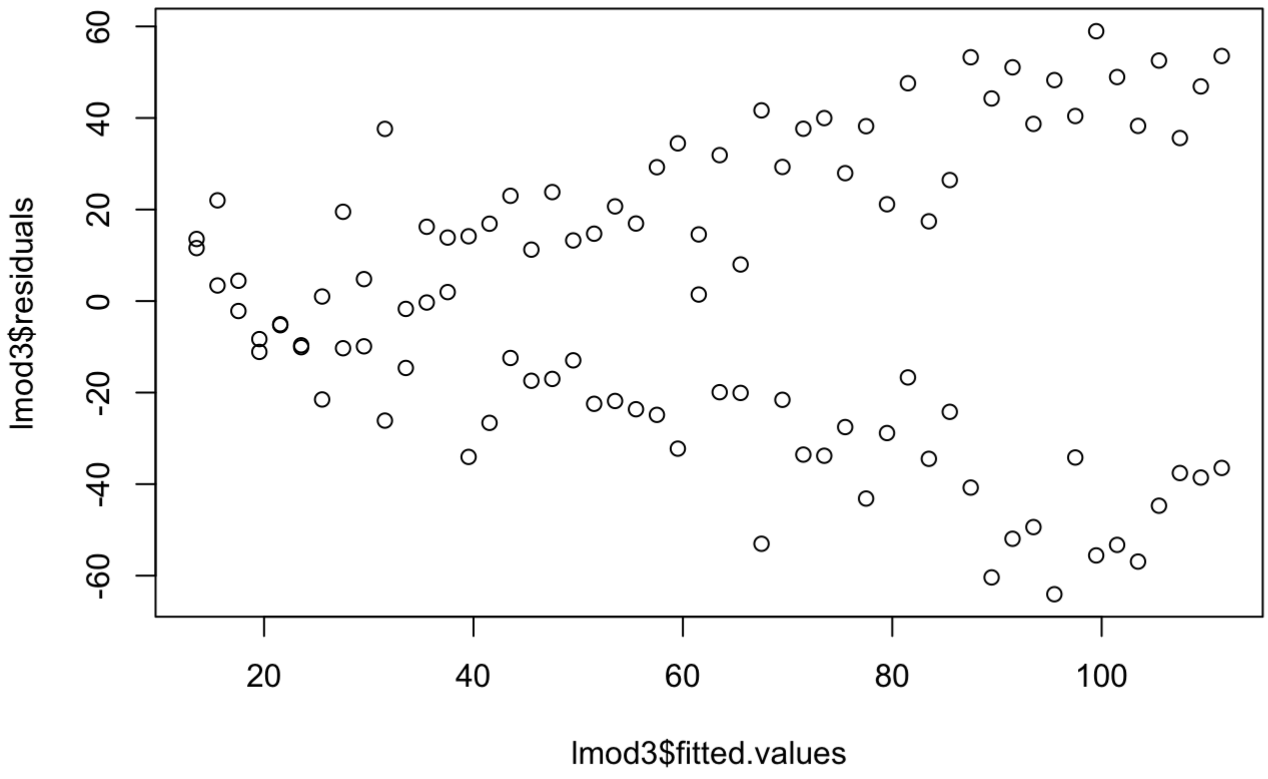


Exercise 12.2. Nonconstant variance and specification error: Generate 100 observations according to the following model:

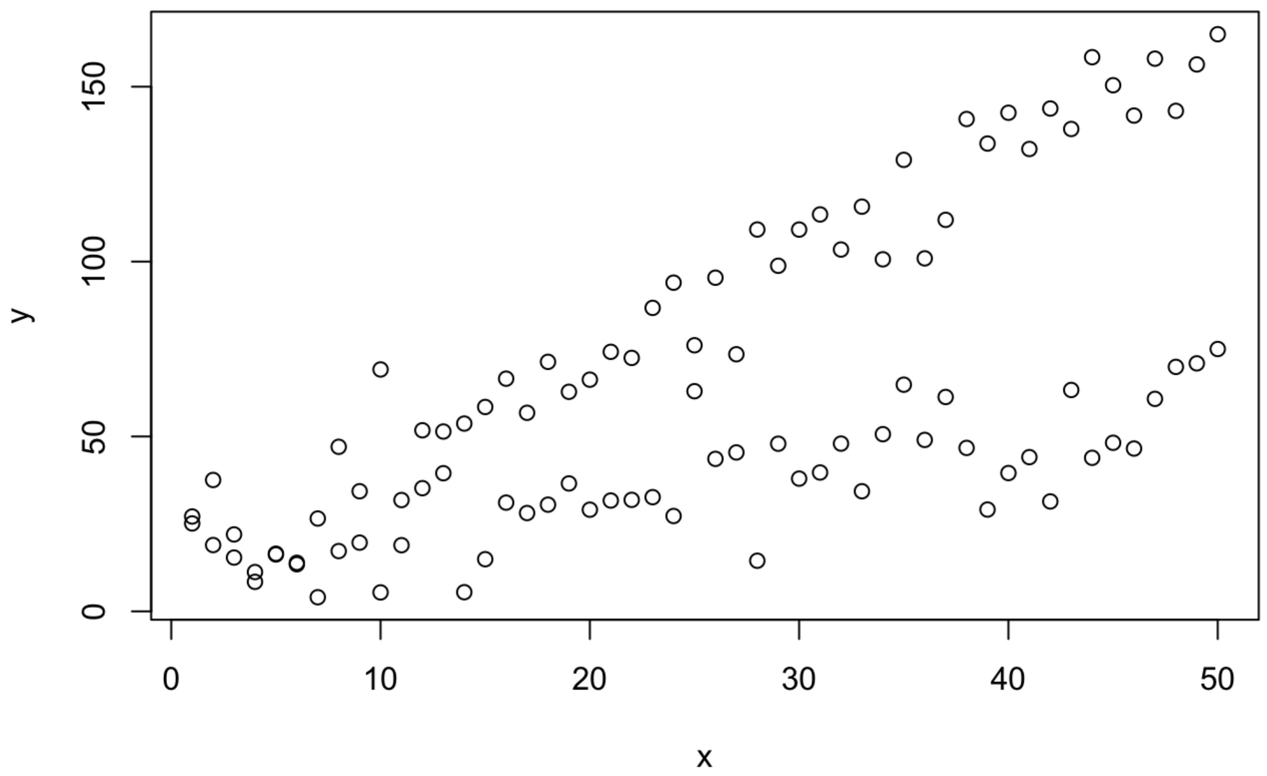
$$Y = 10 + (1\times X) + (1\times D) + (2\times X\times D) + \varepsilon$$

where $\varepsilon \sim N(0, 10^2)$; the values of X are $1, 2, \ldots, 50, 1, 2, \ldots, 50$; the first 50 values of D are 0; and the last 50 values of D are 1. Then regress Y on X alone (i.e., omitting D and XD), $Y = A + BX + E$. Plot the residuals E from this regression against the fitted values $\widehat{Y}$. Is the variance of the residuals constant? How do you account for the pattern in the plot?

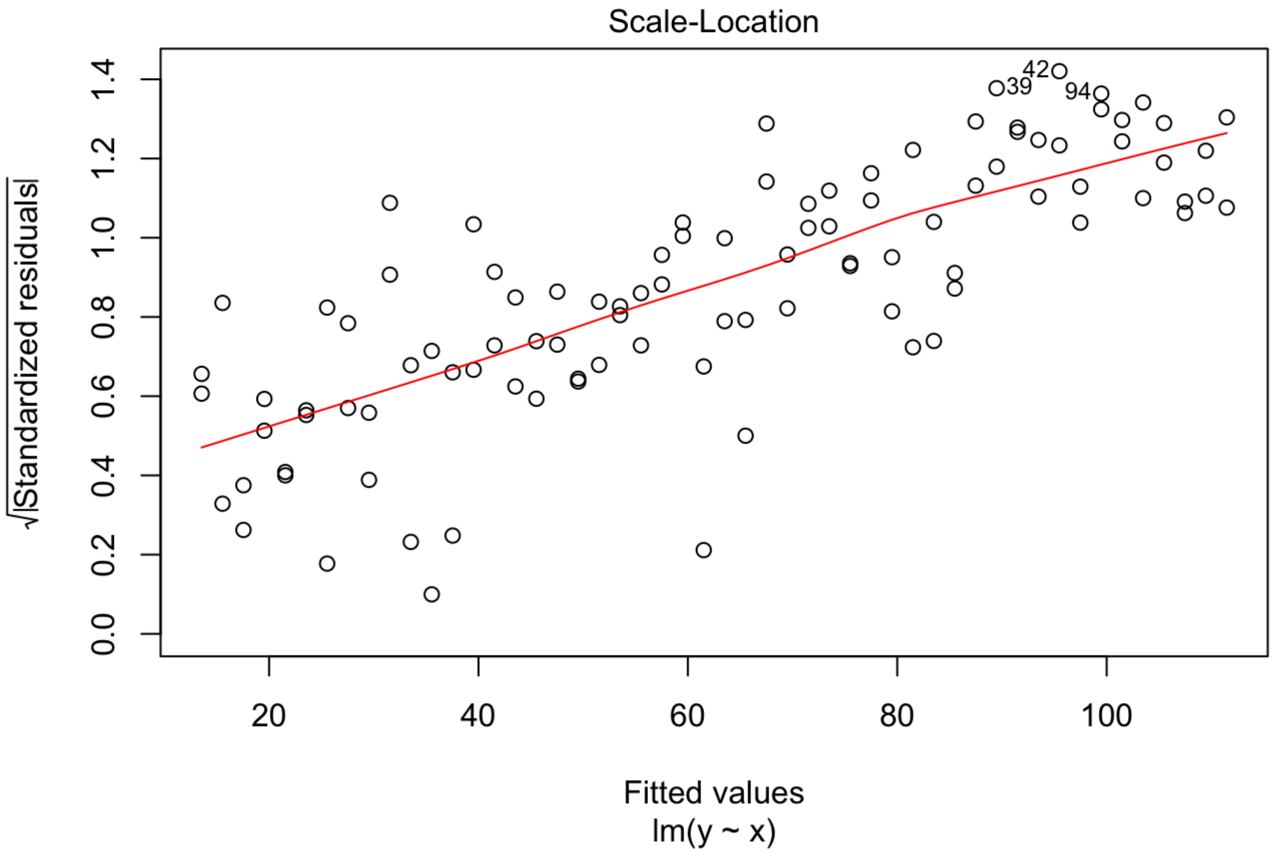
```
x = rep(seq(1,50),2)
d = c(rep(0,50),rep(1,50))
y = 10 + x + d + 2*x*d + rnorm(100,0,10)
lmod3 = lm(y ~ x)
plot(lmod3$residuals ~ lmod3$fitted.values)
```



```
plot(y ~ x)
```

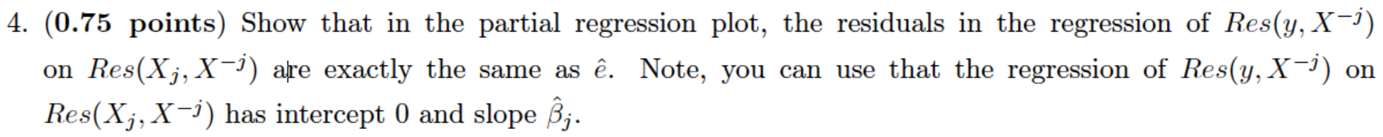


```
plot(lmod3,3)
```

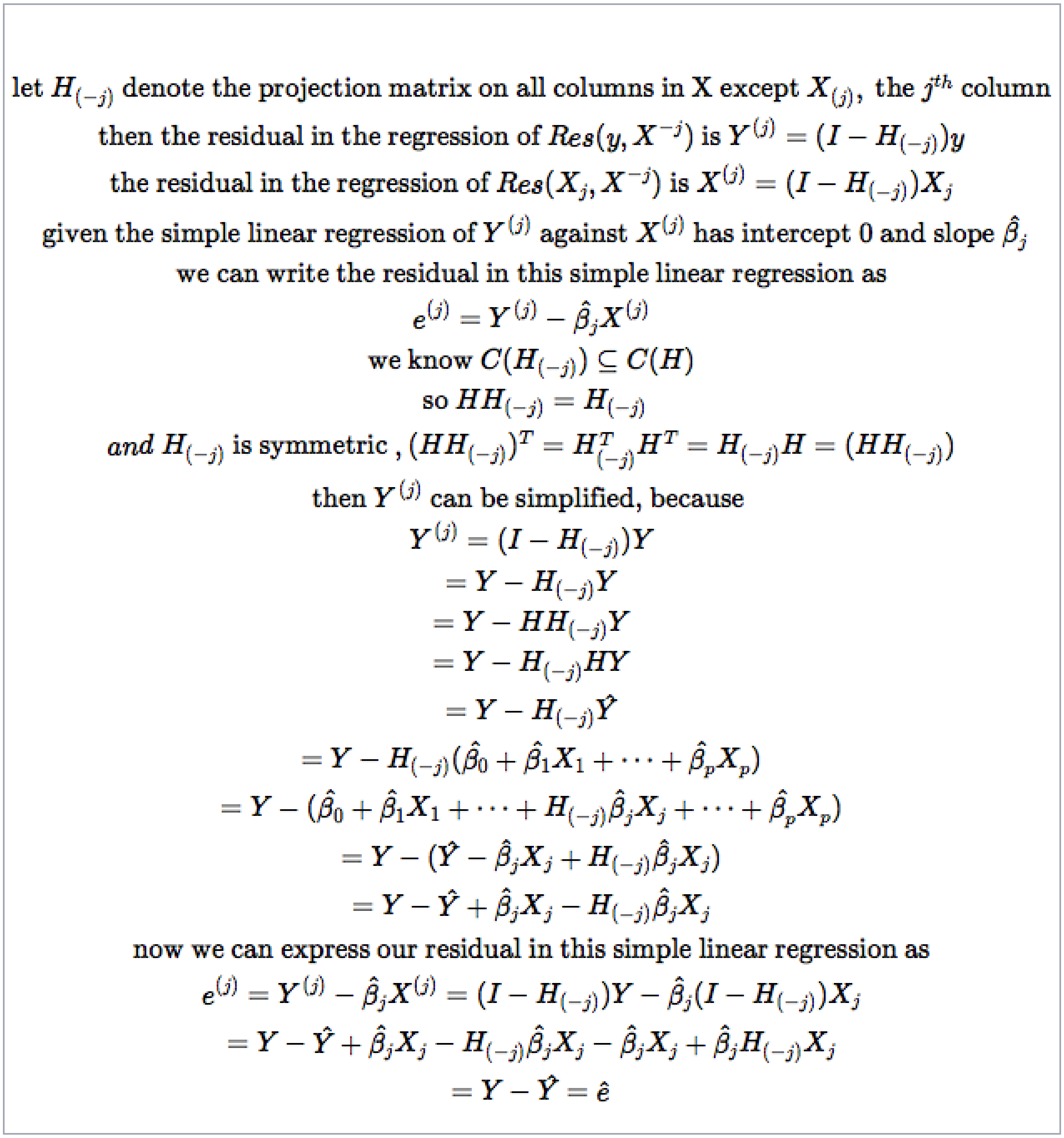


Variance of the residual is not constant, as we see there is a pattern for residuals in this graph. The variance increases as the fitted values increase. We see that X is from [1,…, 50, 1,…, 50]. Y is 10+X for first 50 items and 10 + X + D + 2XD for the rest of 50 items. The regression line tries to minimize the errors, so as X increases, the residuals also increase. We can also see this from Y agains X plot.

Problem 4



4. **(0.75 points)** Show that in the partial regression plot, the residuals in the regression of $Res(y, X^{-j})$ on $Res(X_j, X^{-j})$ are exactly the same as $\hat{e}$. Note, you can use that the regression of $Res(y, X^{-j})$ on $Res(X_j, X^{-j})$ has intercept 0 and slope $\hat{\beta}_j$.



let $H_{(-j)}$ denote the projection matrix on all columns in X except $X_{(j)}$, the j^{th} column

then the residual in the regression of $Res(y, X^{-j})$ is $Y^{(j)} = (I - H_{(-j)})y$

the residual in the regression of $Res(X_j, X^{-j})$ is $X^{(j)} = (I - H_{(-j)})X_j$

given the simple linear regression of $Y^{(j)}$ against $X^{(j)}$ has intercept 0 and slope $\hat{\beta}_j$

we can write the residual in this simple linear regression as

$$e^{(j)} = Y^{(j)} - \hat{\beta}_j X^{(j)}$$

we know $C(H_{(-j)}) \subseteq C(H)$

so $HH_{(-j)} = H_{(-j)}$

and $H_{(-j)}$ is symmetric , $(HH_{(-j)})^T = H^T_{(-j)} H^T = H_{(-j)} H = (HH_{(-j)})$

then $Y^{(j)}$ can be simplified, because

$$\begin{aligned} Y^{(j)} &= (I - H_{(-j)})Y \\ &= Y - H_{(-j)}Y \\ &= Y - HH_{(-j)}Y \\ &= Y - H_{(-j)}HY \\ &= Y - H_{(-j)}\hat{Y} \\ &= Y - H_{(-j)}(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p) \\ &= Y - (\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + H_{(-j)}\hat{\beta}_j X_j + \cdots + \hat{\beta}_p X_p) \\ &= Y - (\hat{Y} - \hat{\beta}_j X_j + H_{(-j)}\hat{\beta}_j X_j) \\ &= Y - \hat{Y} + \hat{\beta}_j X_j - H_{(-j)}\hat{\beta}_j X_j \end{aligned}$$

now we can express our residual in this simple linear regression as

$$\begin{aligned} e^{(j)} = Y^{(j)} - \hat{\beta}_j X^{(j)} &= (I - H_{(-j)})Y - \hat{\beta}_j (I - H_{(-j)})X_j \\ &= Y - \hat{Y} + \hat{\beta}_j X_j - H_{(-j)}\hat{\beta}_j X_j - \hat{\beta}_j X_j + \hat{\beta}_j H_{(-j)} X_j \\ &= Y - \hat{Y} = \hat{e} \end{aligned}$$

Problem 5 Determine if the statement below is true or false. (0.25 points each question)

a. Dropping a variable from the model will cause the other variables to become more significant.
   FALSE, we can see results from backward elimination and observe that dropping a variable can either decrease or increase some other variables's p-value

b. If an observation with high leverage is dropped from the model fitting, then the estimate for the coefficients will become more precise.
   False, if a high leverage observation is not an influential observation, then the estimate for the coefficients doesn't change

c. If AIC and Mallows Cp choose a best model and these both have the same number of parameters, then the models are identical (i.e., they include the same explanatory variables).

$AIC(m) = n\log\left(\frac{RSS(m)}{n}\right) + n\log(2\pi e) + 2(1+p(m)))$ $C_p(m) := \frac{RSS(m)}{\hat{\sigma}^2} - (n - 2(1+p(m)))$

True, comparing two formulas, we see that if p(m) are the same, then the only difference between two methods is a function of RSS. Then given the same number of parameters, both methods would choose the parameters that minimize RSS, so both method will choose sample explanatory variables

d. R2 can be used as a model selection criterion in the linear model.

False, R^2 will always increase as the number of parameters increase. We use adjusted R-square instead

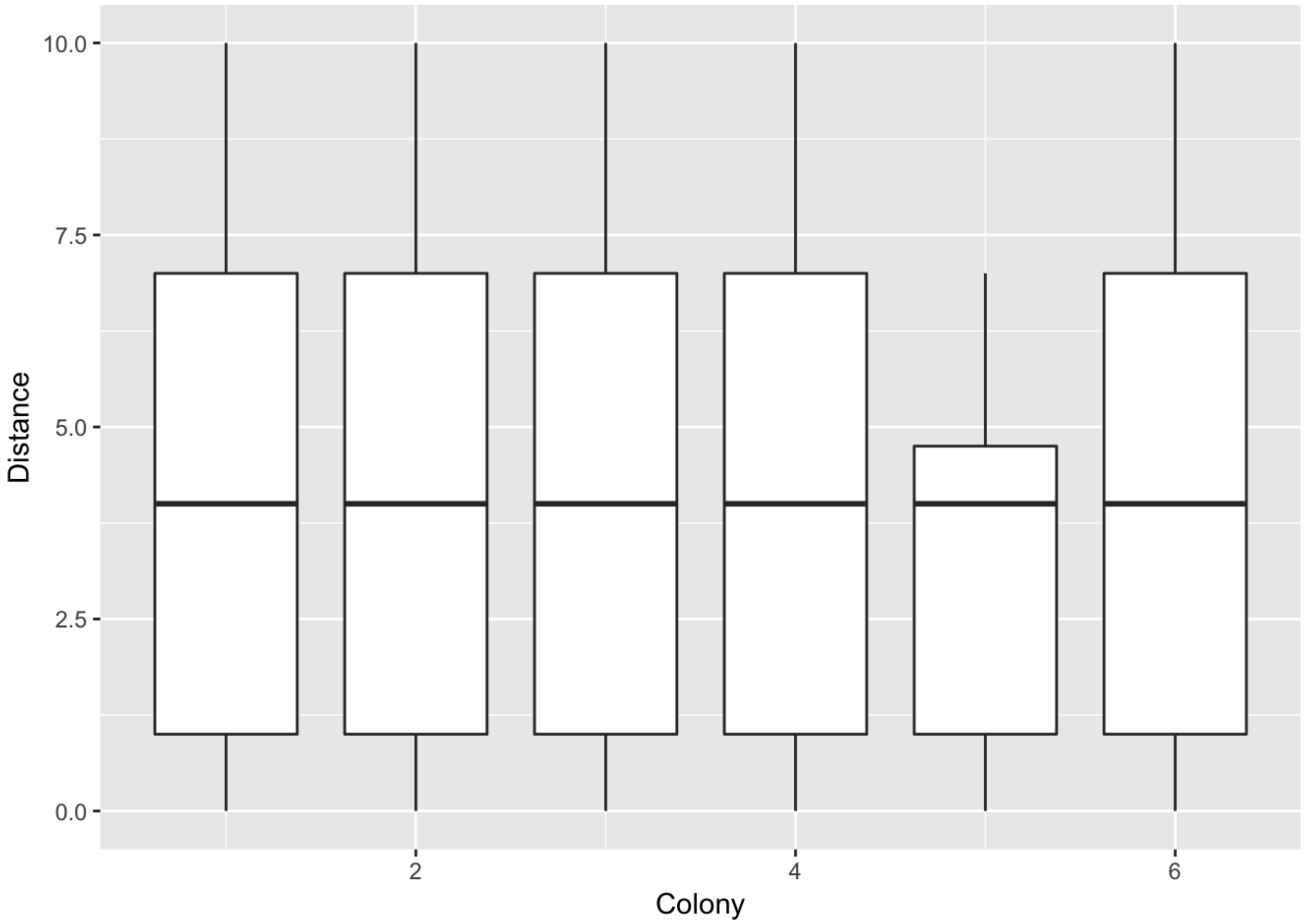
Problem 6

Scientific Questions The principle scientific questions that this data pose are: Do different colonies use different foraging strategies? (e.e. worker-conservative versus energy-conservative) Is there some difference across size classes? Are there differences across colonies in the distribution of sizes or distances of the member ants?

a. (0.5 points) First, examine the data visually, using various plots, including boxplots and coplots.

boxplots on Distance based on different Colony

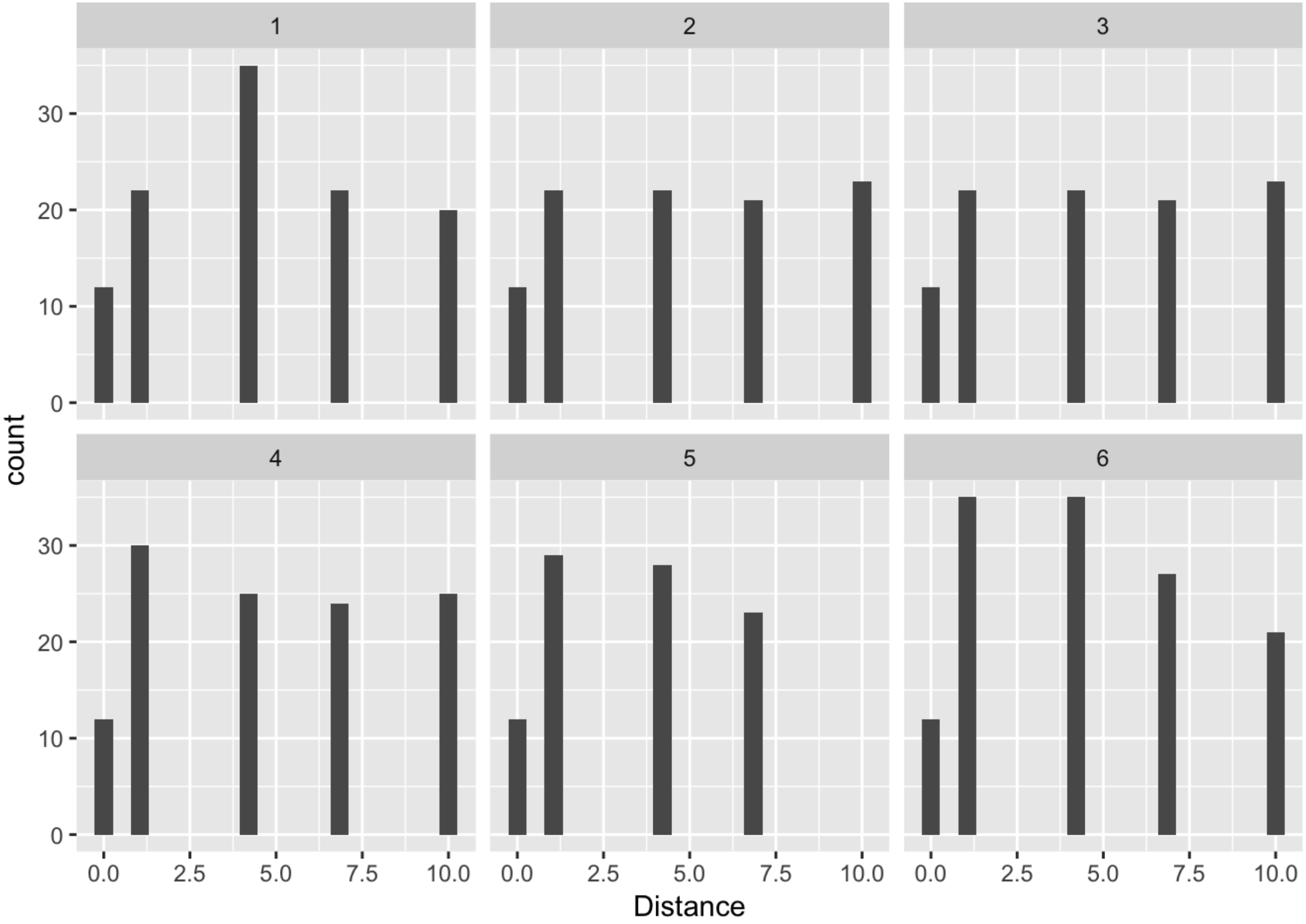
```
ggplot(data = ant6, aes(x= Colony, y=Distance)) +
 geom_boxplot(aes(group=Colony))
```



We see that colony 5 is different from other colonies. As we can see that colony 5 doesn’t have workers whose distance is above 7.

Furthermore, we draw histogram on Distance for different colonies. We can see how many workers on different Distance for each colony.

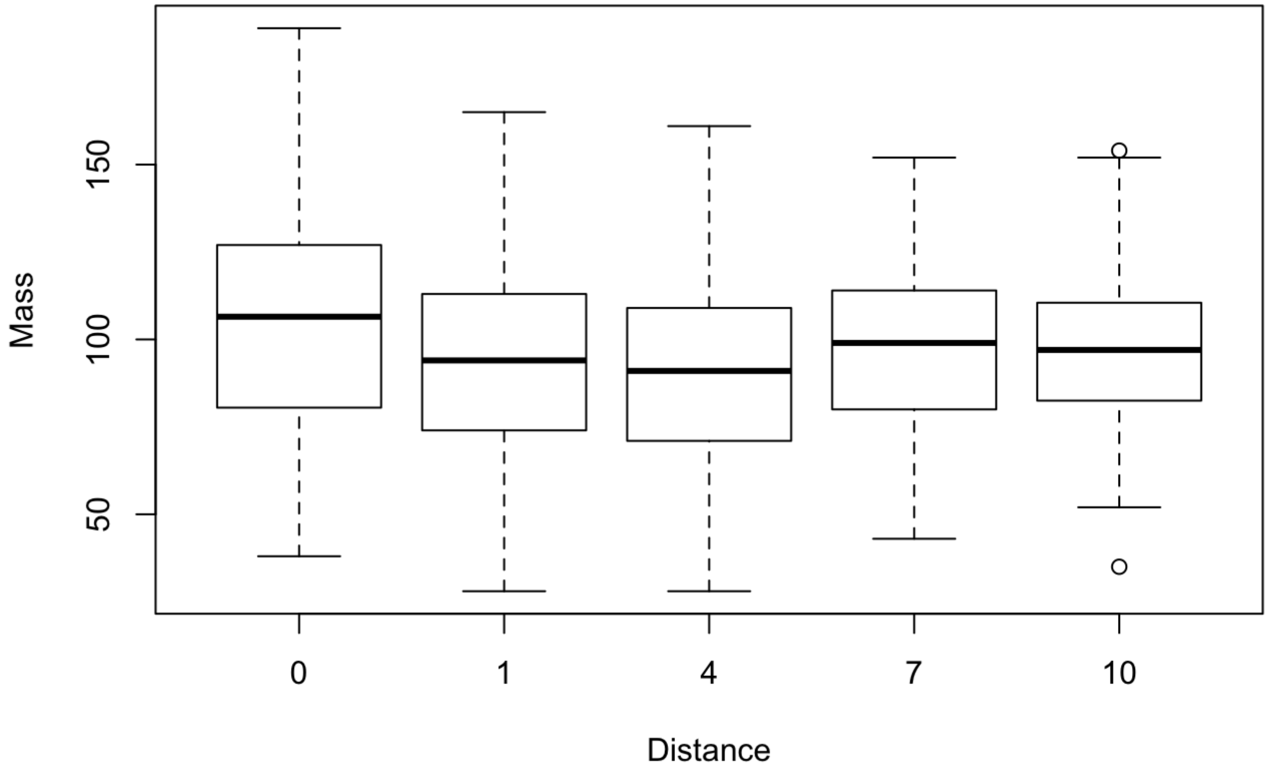
```
ggplot(data=ant6) + geom_histogram(aes(Distance),bins = 20)+facet_wrap(~Colony)
```



we can see relation between Mass and Distance for each Colony

for all colones together

```
boxplot(Mass ~ Distance,xlab="Distance",ylab="Mass",data= ant6)
```



b. (2 points) Perform a regression of the mass on colony, distance, and size, and evaluate the appropriateness of your model using graphical techniques. If you find a transformation needed, justify your choice of transformation.

Because colony is a categorical variable, we turn it into a factor, just as size class is a factor.
assume there’s relation between colony and distance, we add an interaction term

```
skewness(ant6$Mass)
```

```
## [1] 0.0289869
```

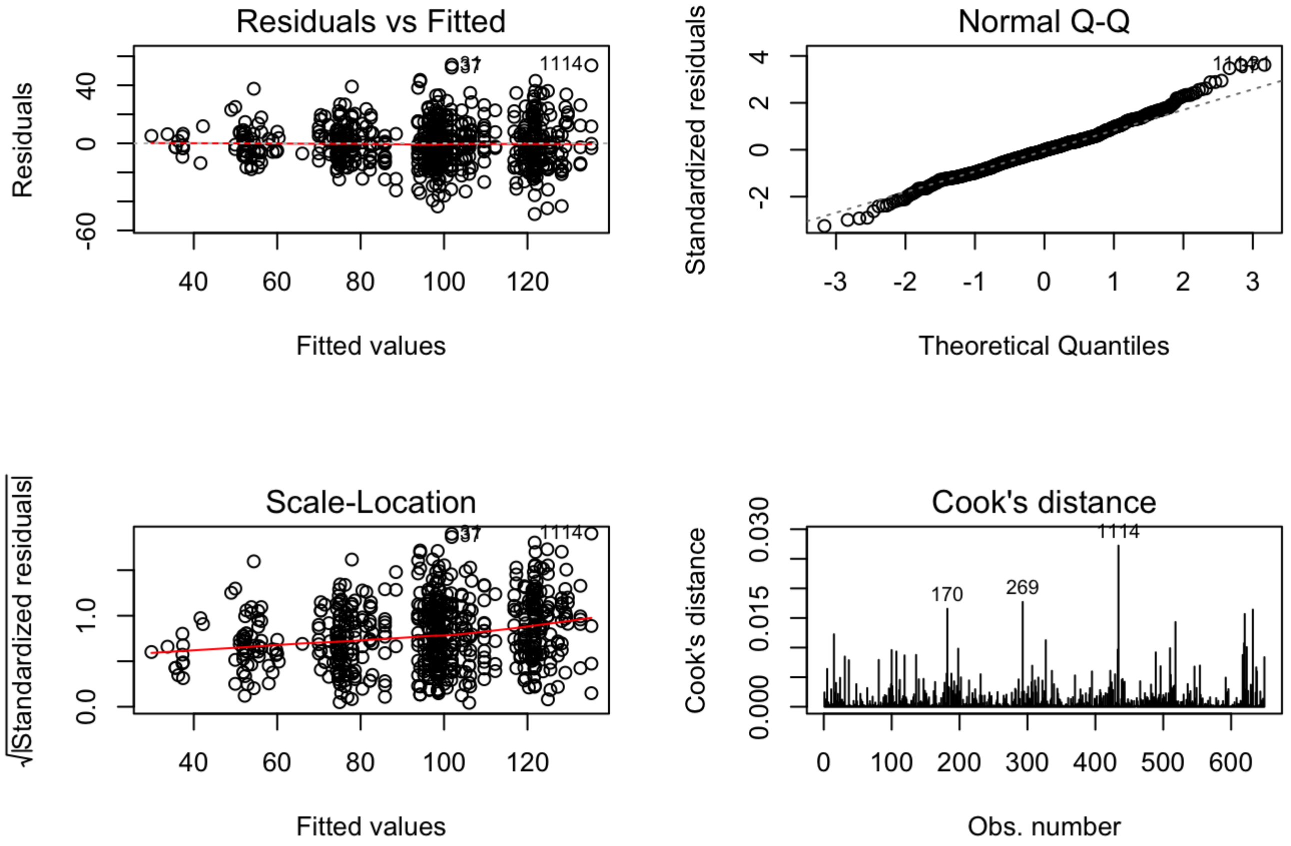
```
#skewness of Mass is not big, so we don't neet to take the log

ant6$Colony = as.factor(ant6$Colony)
#assume there's relation between colony and distance, we add an interaction term
lmod6 = lm(Mass ~ Colony + Distance +Colony*Distance + Size.class, data=ant6)
summary(lmod6)
```

```
##
## Call:
## lm(formula = Mass ~ Colony + Distance + Colony * Distance + Size.class,
##     data = ant6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.723  -9.490  -0.267   7.975  54.151
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        41.8377     4.7932   8.729  < 2e-16 ***
## Colony2            -7.4559     3.5234  -2.116 0.034726 *
## Colony3             0.3637     3.5173   0.103 0.917676
## Colony4            -7.1176     3.3688  -2.113 0.035008 *
## Colony5             6.0151     3.5546   1.692 0.091102 .
## Colony6            -5.5632     3.2935  -1.689 0.091682 .
## Distance           -1.1033     0.4258  -2.591 0.009777 **
## Size.class\x80     38.7903    15.7953   2.456 0.014325 *
## Size.class>43      87.4767     4.3916  19.919  < 2e-16 ***
## Size.class30-34    18.1701     4.6887   3.875 0.000118 ***
## Size.class35-39    40.6138     4.4007   9.229  < 2e-16 ***
## Size.class40-43    64.4250     4.3382  14.851  < 2e-16 ***
## Colony2:Distance    1.0695     0.5937   1.802 0.072096 .
## Colony3:Distance    0.5050     0.5945   0.850 0.395911
## Colony4:Distance    0.8369     0.5770   1.450 0.147470
## Colony5:Distance   -1.5397     0.7459  -2.064 0.039400 *
## Colony6:Distance    0.4601     0.5786   0.795 0.426753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.09 on 632 degrees of freedom
## Multiple R-squared:  0.6962, Adjusted R-squared:  0.6886
## F-statistic: 90.54 on 16 and 632 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lmod6,which=1:4)
```

```
## Warning: not plotting observations with leverage one:
##   460
```



```
ant6[ant6$Size.class == "\x80",]
```

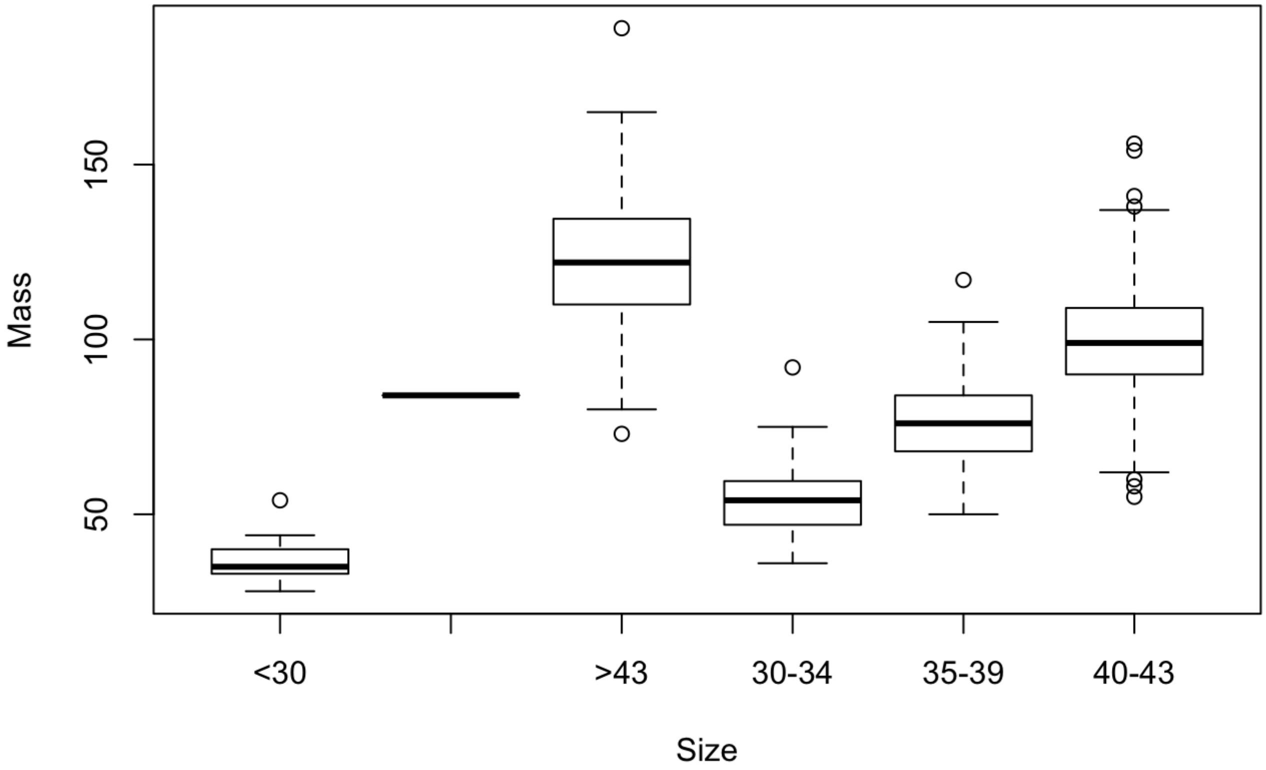
```
##      Colony Distance Mass Headwidth Headwidth..mm. Size.class
## 1140      5        1   84        NA             NA       \x80
```

```
#Only one data with size class = 80
```

We see the variance of residuals gets bigger as the fitted value increases. We didn't see a significant pattern for reisudals , so linear assumption seems ok. In qqplot, the normality assumption looks valid, although both tails are heavy. There are some potential outliers as we can see in cook's distance plot. We also notice there's one point that has very high leverage that needs to be investigated.

c. (1 points) Interpret the coeffcients relative to the scientic contributions and discuss what conclusions you can draw.

```
boxplot(Mass ~ Size.class,xlab="Size",ylab="Mass",data= ant6)
```



Across all colonies, we see a negative relationship between Mass and Distance. This means that the far the workers go, the less food they get. Overrall ants colony use energy conservative strategy.
In summary table, colony and size class are dummy vairables. We can see differences among different colonies. For example, colony 1 is the baseline. The mass of colony 2 on average is less than mass of colony 1 by 2.32 holding all other variables constant. Similar for size classes.
From above boxplot and summary table, we see that on average size<30 has the lowest Mass. This means that in general, ants with the smallest size get the smallest amount of food on average. We also notice that ants whose size=80 don’t get a lot of food. The size between 40-43 has the highest mean Mass, meaning the most food.