hw6

Code

1. A developer of vacation homes is considering purchasing a tract of land near a lake.

Hide

```
summary(model_full)
```

```
Call:
lm(formula = df$Price ~ df$Lot.size + df$Trees + df$Distance,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-66.702 -35.272   0.365  28.854  84.966

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.3912    23.5165   2.185   0.0331 *
df$Lot.size   0.6999     0.5589   1.252   0.2156
df$Trees      0.6788     0.2293   2.960   0.0045 **
df$Distance  -0.3784     0.1952  -1.938   0.0577 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.24 on 56 degrees of freedom
Multiple R-squared:  0.2425,    Adjusted R-squared:  0.2019
F-statistic: 5.975 on 3 and 56 DF,  p-value: 0.001315
```

#a. Do the model assumptions appear to be satisfied? If not, which ones are violated? it seems like the model assumptions do not hold. There does not appear to be linear relationship between the independent variables and dependent variables, also the residuals are not independent because we can see that they are correlated– see later input #b. What is R2? What does it tell you? R2 = .1239 and adjusted R2 = .09319, this tells us that about 10% of the variation in price is explained by lot size, distance and trees.

Hide

```
model1 = lm(df$Price ~ df$Lot.size,data=df)



print('model1----------------------------------------------------')
```

```
[1] "model1----------------------------------------------------"
```

Hide

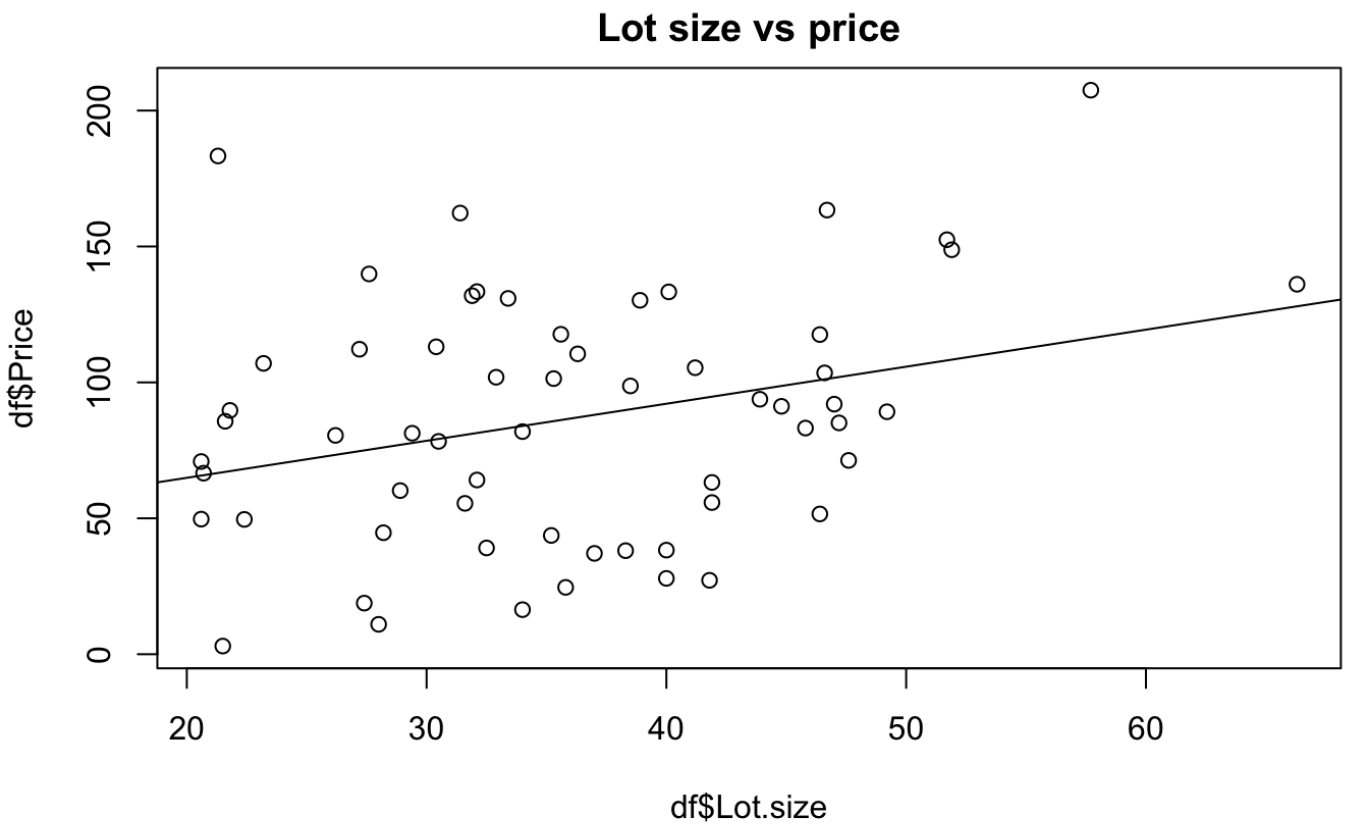
```
summary(model1)
```

```

Call:
lm(formula = df$Price ~ df$Lot.size, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-67.579 -31.417  -0.082  32.162 116.628 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  37.6455    20.9835   1.794   0.0780 .
df$Lot.size   1.3627     0.5617   2.426   0.0184 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.29 on 58 degrees of freedom
Multiple R-squared:  0.09213,   Adjusted R-squared:  0.07647 
F-statistic: 5.886 on 1 and 58 DF,  p-value: 0.0184
```

Hide

```
plot(df$Lot.size,df$Price,main = 'Lot size vs price')+
  abline(model1)
```

```
integer(0)
```



Hide

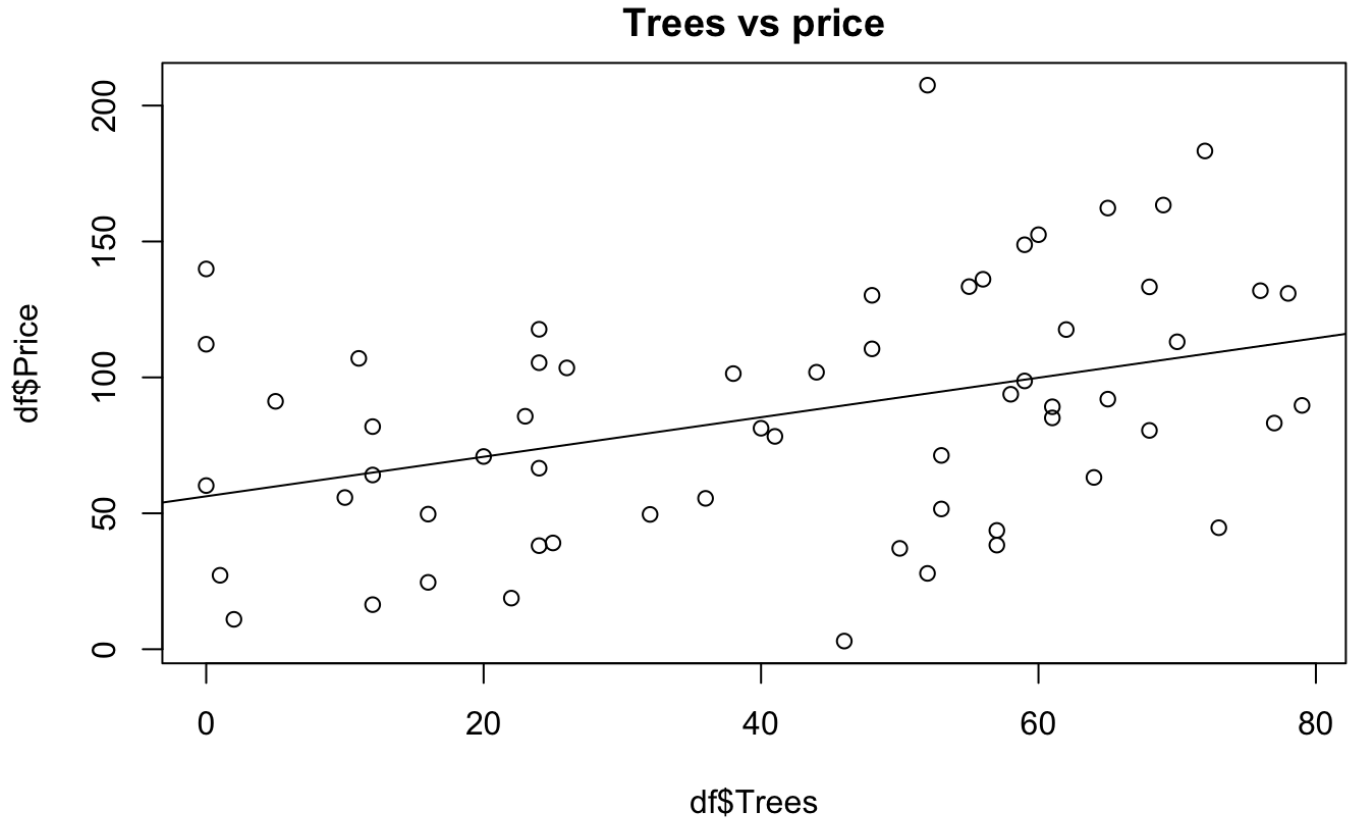
```
# print('model2----------------------------------------------------')
model2 = lm(df$Price ~ df$Trees,data=df)
summary(model2)
```

```

Call:
lm(formula = df$Price ~ df$Trees, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-86.73 -29.83  -2.48  29.07 113.40 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  56.2612    10.9010   5.161 3.13e-06 ***
df$Trees      0.7276     0.2262   3.216  0.00212 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.85 on 58 degrees of freedom
Multiple R-squared:  0.1514,    Adjusted R-squared:  0.1367 
F-statistic: 10.35 on 1 and 58 DF,  p-value: 0.002124
```

Hide

```
plot(df$Trees,df$Price,main = 'Trees vs price')+
  abline(model2)
```

```
integer(0)
```



Hide

```
print('model3----------------------------------------------------')
```

```
[1] "model3----------------------------------------------------"
```

Hide

```
model3 = lm(df$Price ~ df$Distance,data=df)
summary(model3)
```

```
Call:
lm(formula = df$Price ~ df$Distance, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-86.257 -31.384  -5.913  31.431 106.998

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 105.0566    11.5739   9.077  9.9e-13 ***
df$Distance  -0.3795     0.2084  -1.821   0.0737 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44.19 on 58 degrees of freedom
Multiple R-squared:  0.05411,   Adjusted R-squared:  0.0378
F-statistic: 3.318 on 1 and 58 DF,  p-value: 0.07369
```

Hide

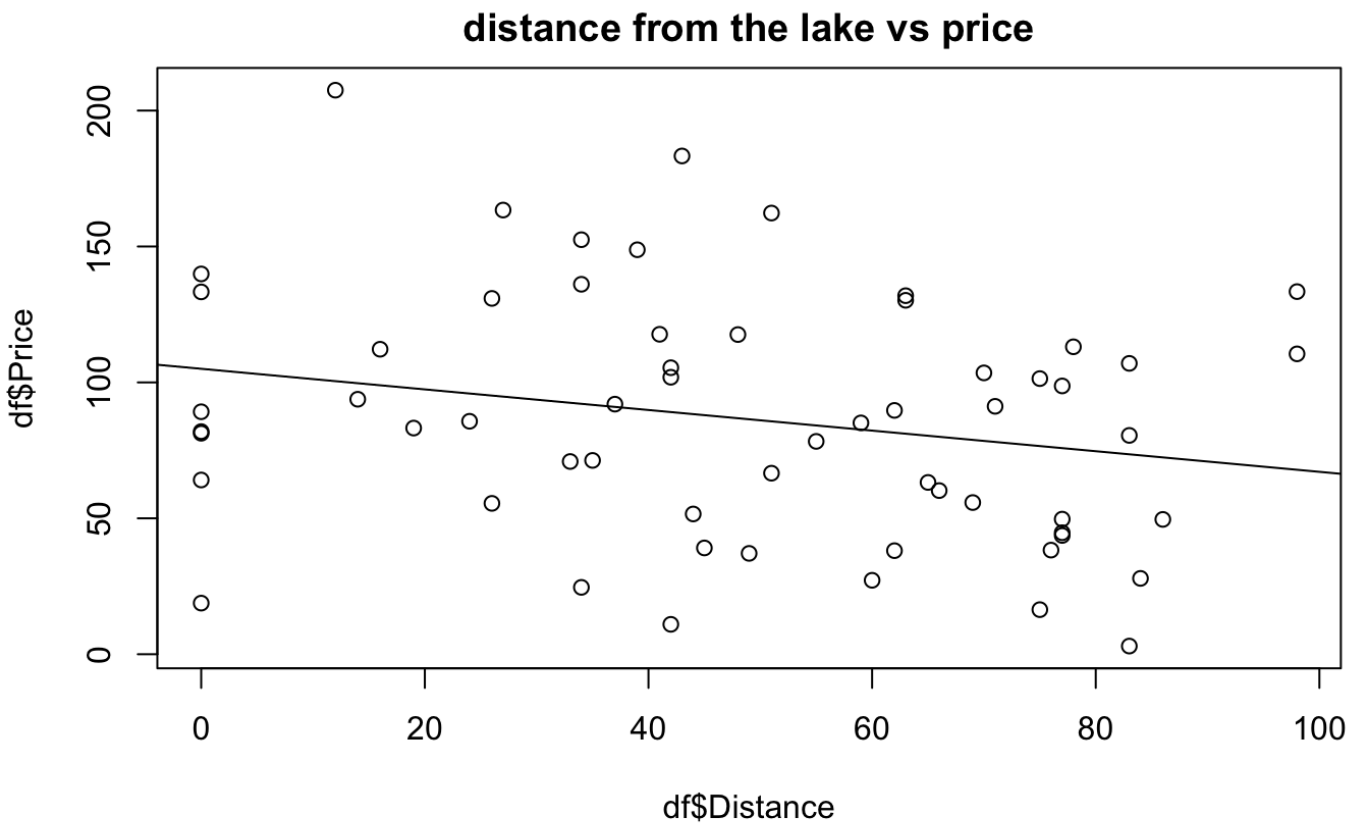
```
model32 = lm(df$Price ~ df$Distance+df$Trees,data=df)
summary(model32)
```

```
Call:
lm(formula = df$Price ~ df$Distance + df$Trees, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-73.600 -33.159  -4.829  33.828  97.281

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  75.5248    13.5464   5.575 7.06e-07 ***
df$Distance  -0.4327     0.1913  -2.262 0.027549 *
df$Trees      0.7671     0.2193   3.498 0.000917 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.44 on 57 degrees of freedom
Multiple R-squared:  0.2213,    Adjusted R-squared:  0.1939
F-statistic: 8.097 on 2 and 57 DF,  p-value: 0.0008031
```

Hide

```
plot(df$Distance,df$Price,main = 'distance from the lake vs price')+abline(model3)
```

```
integer(0)
```



c. Which of the explanatory variables is linearly related to the response variable in this (the original) model?

Trees Adjusted R-squared: 0.1367 F-statistic: 10.35 on 1 and 58 DF,
p-value: 0.002124 # d. If necessary, create a new model by removing insignificant variables. here we go with model2 which is Price = 56.2612 + 0.7276*Trees
e. Interpret the slopes in the new model. houses with no trees are worth $56k relative units…. every tree will add .7276 additional units of price so $7,276 in the bank for a tree :) # f. Predict with 95% confidence the selling price of a 40,000-square foot lot with 50 mature trees that is located 75 feet from the lake. from the output below we can see that the 95% CI for price of houses with 40k sqft lot and 50 mature trees located 75 feet from the lake is (2.79,167) in thousands of dollars so (2790,167000)

Hide

```
a <- df$Price
x1 <- df$Lot.size
x2 <- df$Trees
x3 <- df$Distance

#here you use x as a name
fitPrice <- lm(a ~ x1 + x2 + x3)
# single value
prediction = predict(fitPrice, data.frame(x1=40,x2=50,x3=75), interval = "prediction")
print(prediction)
```

```
       fit      lwr     upr
1 84.95099 2.794942 167.107
```

g. Estimate with 95% confidence the average selling price of all such lots.

from the output below we can see that the 95% CI for average selling price of such lots is (69.12,100.77) in thousands of dollars so (69120,100770)

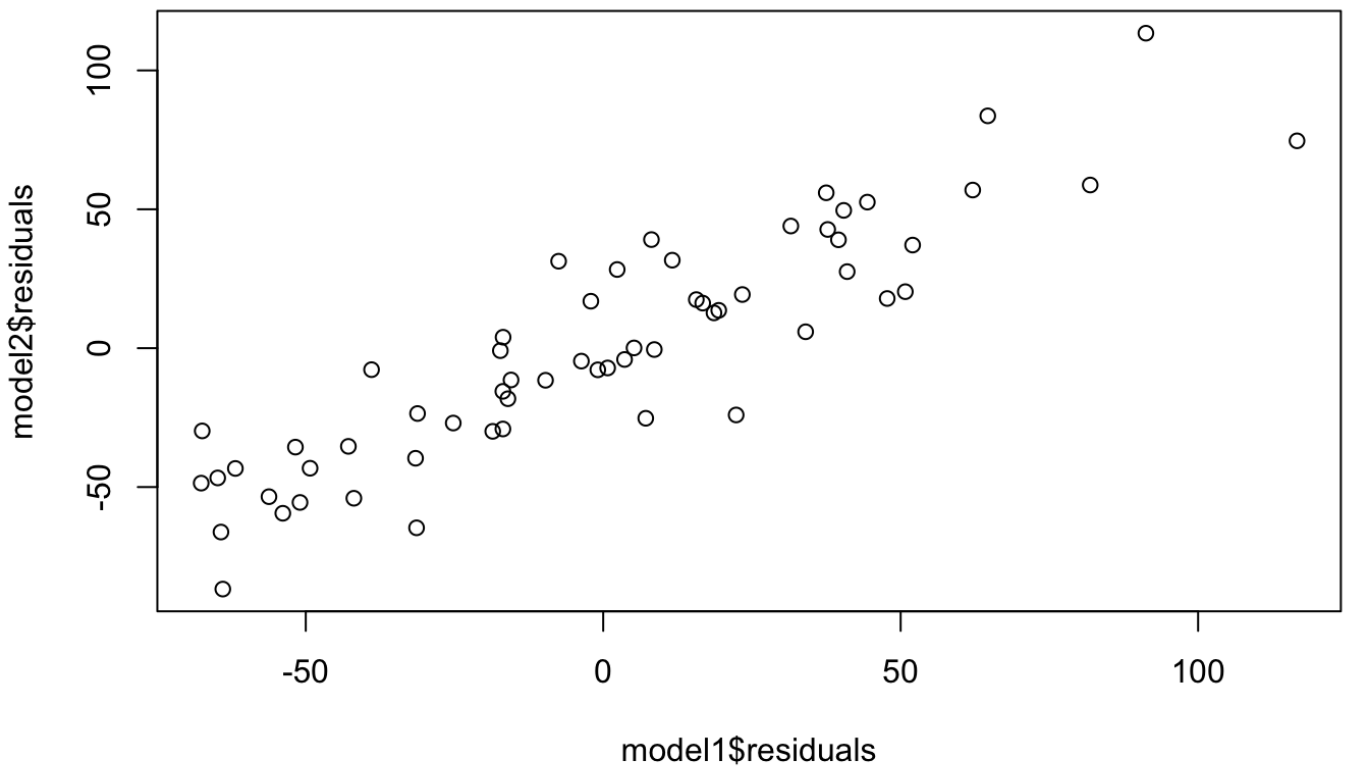
Hide

```
print(prediction_mean)
```

```
       fit      lwr      upr
1 84.95099 69.12573 100.7762
```

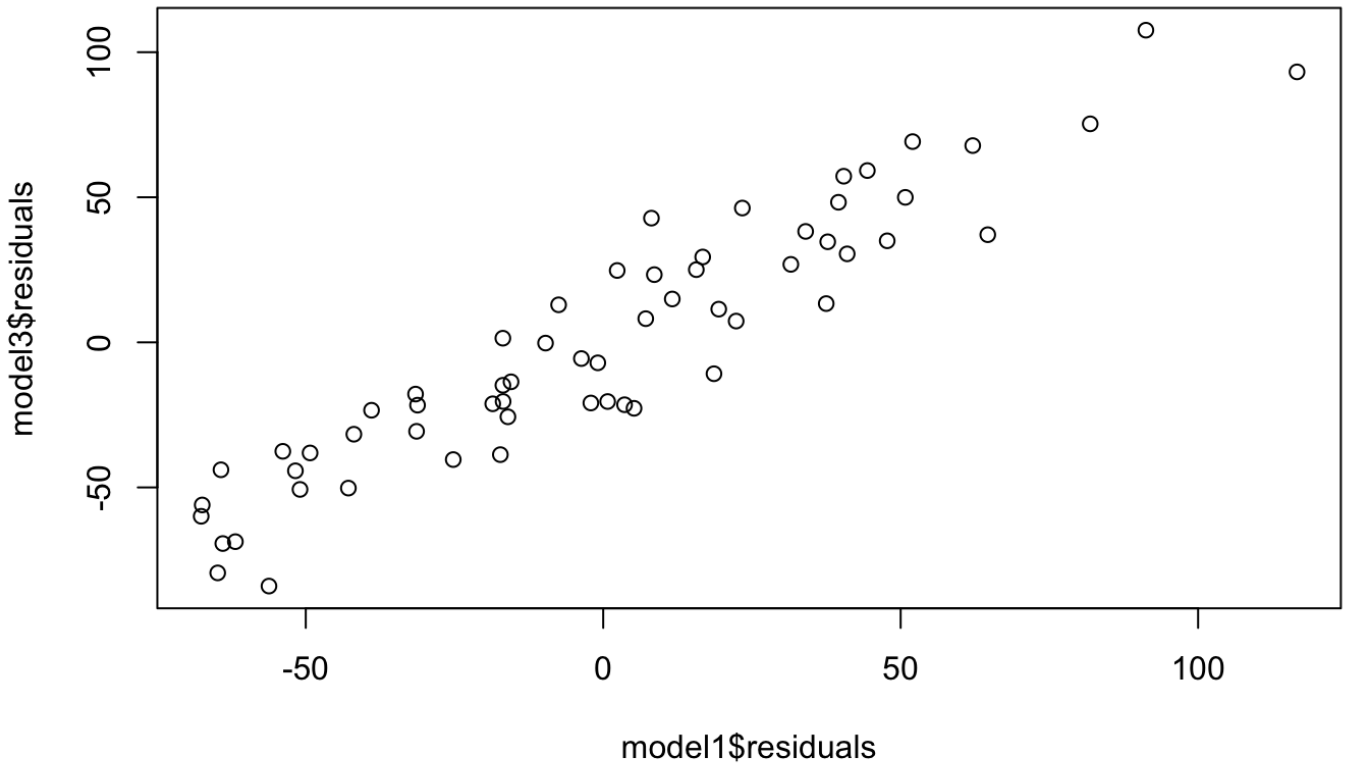
Hide

```
plot(model1$residuals,model2$residuals)
```



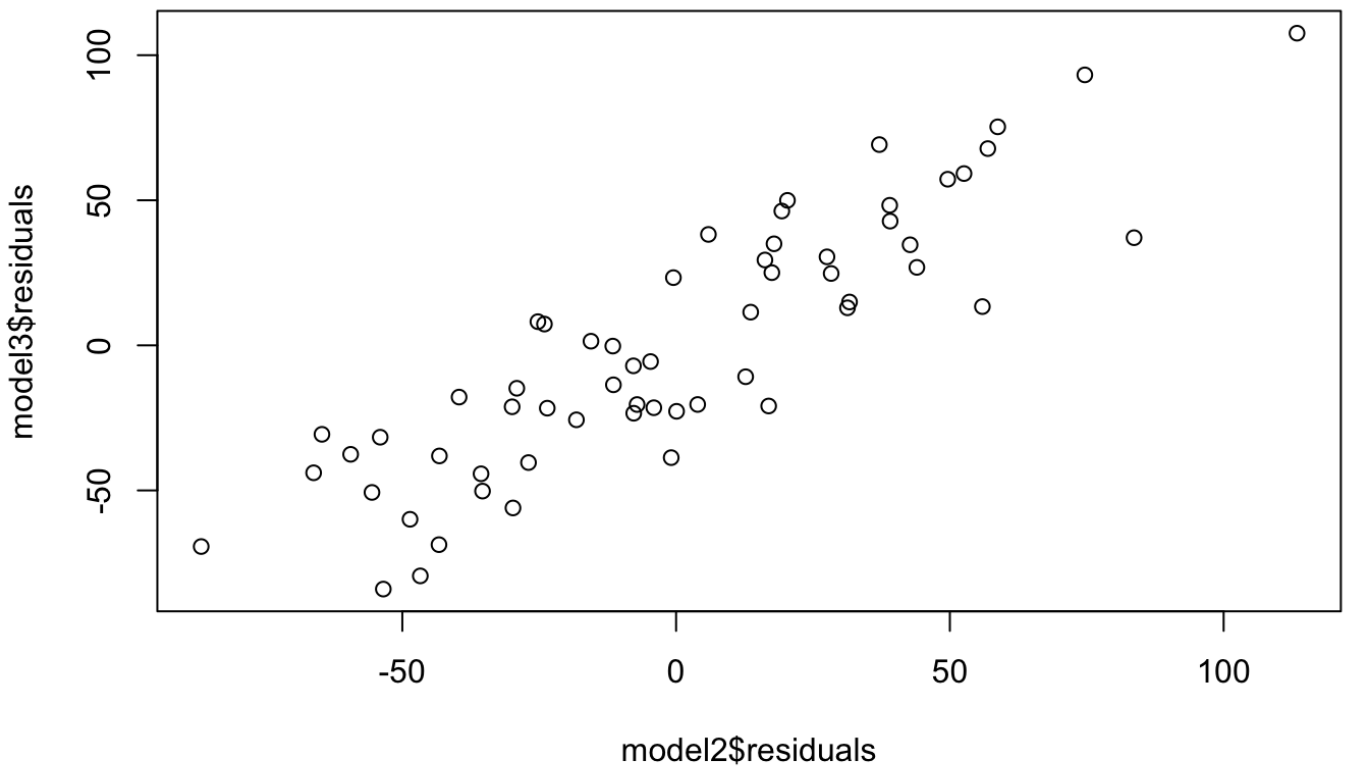
Hide

```
plot(model1$residuals,model3$residuals)
```



Hide

```
plot(model2$residuals,model3$residuals)
```



2. The manager of an amusement park would like to be able to predict daily attendance. After some consideration, he decided that the following three factors are critical, yesterday's attendance, whether it's a weekday or weekend, and the predicted weather. He then took a random sample of 40 days and recorded the data in the file AMUSEMENT. Since two of the variables are qualitative, he created the following sets if dummy variables:

Weekend = 1 (if weekend)= 0 (if not) Sunny = 1 (if mostly sunny is predicted)= 0 (if not) Rain= 1 (if rain is predicted)= 0 (if not)

a. Construct a regression model to predict attendance. Is the model likely to be useful? Include all relevant computer output, organized so that I can follow it.

We can see from the output below that this is a useful model. All of the coefficients are far from zero with low p-values. The only variable that seems unnecessary is Yesterday because the p-value is somewhat high and coefficient is near zero.

Hide

```
AMUSEMENT = read_xls('AMUSEMENT.xls')


model2.full = lm(formula = Attendance ~ Yesterday + Weekend + Sunny + Rain, data=AMUSEME
NT)
model2.2 = lm(formula = Attendance ~ Yesterday + Weekend + Sunny, data=AMUSEMENT)
model2.3 = lm(formula = Attendance ~ Weekend + Sunny, data=AMUSEMENT)
weekend = lm(formula = Attendance ~ Weekend,data=AMUSEMENT)


print('-------------------------------------------------------------------------')
```

```
[1] "-------------------------------------------------------------------------"
```

Hide

```
summary(model2.1)
```

```
Call:
lm(formula = Attendance ~ Yesterday + Weekend + Sunny + Rain,
    data = AMUSEMENT)

Residuals:
     Min       1Q   Median       3Q      Max
-1181.11  -494.98    41.44   487.71  1725.30

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4514.44912  570.67150   7.911 2.66e-09 ***
Yesterday      0.20612    0.08472   2.433  0.02023 *
Weekend      933.90177  311.61848   2.997  0.00499 **
Sunny       1074.83737  322.76468   3.330  0.00205 **
Rain        -727.09560  365.57648  -1.989  0.05457 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 699.6 on 35 degrees of freedom
Multiple R-squared:  0.7768,    Adjusted R-squared:  0.7513
F-statistic: 30.45 on 4 and 35 DF,  p-value: 5.841e-11
```

Hide

```
print('--------------------------------------------------------------------------------')
```

```
[1] "--------------------------------------------------------------------------------"
```

Hide

```
summary(model2.2)
```

```
Call:
lm(formula = Attendance ~ Yesterday + Weekend + Sunny, data = AMUSEMENT)

Residuals:
    Min      1Q  Median      3Q     Max
-1268.0  -389.3   186.8   338.6  2021.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.731e+03  4.296e+02   8.686 2.33e-10 ***
Yesterday   2.953e-01  7.479e-02   3.948 0.000351 ***
Weekend     9.771e+02  3.234e+02   3.022 0.004609 **
Sunny       1.234e+03  3.253e+02   3.794 0.000547 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 727.7 on 36 degrees of freedom
Multiple R-squared:  0.7516,    Adjusted R-squared:  0.7309
F-statistic:  36.3 on 3 and 36 DF,  p-value: 5.551e-11
```

Hide

```
print('--------------------------------------------------------------------------------')
```

```
[1] "--------------------------------------------------------------------------------"
```

Hide

```
summary(model2.3)
```

```
Call:
lm(formula = Attendance ~ Weekend + Sunny, data = AMUSEMENT)

Residuals:
     Min       1Q   Median       3Q      Max
-1671.59  -687.79    58.41   591.66  2061.41

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5305.6      188.3  28.177  < 2e-16 ***
Weekend        587.6      363.6   1.616    0.115
Sunny         1869.8      333.7   5.604 2.16e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 859.2 on 37 degrees of freedom
Multiple R-squared:  0.644, Adjusted R-squared:  0.6248
F-statistic: 33.47 on 2 and 37 DF,  p-value: 5.027e-09
```

Hide

```
print('------------------------------------------------------------------------------')
```

```
[1] "------------------------------------------------------------------------------"
```

Hide

```
summary(weekend)
```

```

Call:
lm(formula = Attendance ~ Weekend, data = AMUSEMENT)

Residuals:
    Min      1Q  Median      3Q     Max 
-2205.8  -717.3  -167.3   674.9  2089.2 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5839.8      217.9  26.806  < 2e-16 ***
Weekend       1767.3      397.8   4.443 7.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1153 on 38 degrees of freedom
Multiple R-squared:  0.3419,    Adjusted R-squared:  0.3246 
F-statistic: 19.74 on 1 and 38 DF,  p-value: 7.425e-05
```

b. Can we conclude that weather is a factor in determining attendance?

yes. In general we can assume that weather affects amusement park attendance. There is a strong correlation between attendence and the variable sunny with a low p-value. We don’t need the rain variable because it is mutually exclusive (usually) with the sunny variable. # c. Determine the best model using the Akaike Information Criterion (AIC) and Mallow’s Cp statistics. How does this affect your choice of final model? Does it change your answer to part (b)?

based on the AIC I would say that the model with all independent variables is the best model. Based on the F-statistic and R2 I would say that the model without Weekend is the best. I will use the model with Yesterday, Weekend and Sunny as the independent variables. No we can still conclude that weather is a factor in attendence. We can see that

AIC = 2p/n + 2ln(RMSE) CP = SSEP/MSEQ - (n - 2*p)

Hide

```
#n=length(summary(model2.1)$residuals)
n=40
p2.1=5
p2.2=4
p2.3=3

RSS2.1 = summary(model2.full)$residuals ** 2
RSS2.2 = summary(model2.2)$residuals ** 2
RSS2.3 = summary(model2.3)$residuals ** 2

# MSE for full model with q=5 parameters
MSEQ = sum(RSS2.1)/n

# RMSE for all models
RMSE2.1 = sqrt(sum(RSS2.1)/n)
RMSE2.2 = sqrt(sum(RSS2.2)/n)
RMSE2.3 = sqrt(sum(RSS2.3)/n)

# SSEQ for the smaller models
SSEP2.2=sum(RSS2.2)
SSEP2.3=sum(RSS2.3)



# Akaike Information Criterion
AIC2.1 = (2*p2.1/n) + 2*log(RMSE2.1)
AIC2.2 = (2*p2.2/n) + 2*log(RMSE2.2)
AIC2.3 = (2*p2.3/n) + 2*log(RMSE2.3)

# Mallows Cp
CP2.2=SSEP2.2/MSEQ - (n-2*p2.2)
CP2.3=SSEP2.3/MSEQ - (n-2*p2.3)


print('AIC:')
```

```
[1] "AIC:"
```

Hide

```
print(AIC2.1)
```

```
[1] 13.21735
```

Hide

```
print(AIC2.2)
```

```
[1] 13.27443
```

Hide

```
print(AIC2.3)
```

```
[1] 13.58412
```

Hide

```
print('Cp')
```

```
[1] "Cp"
```

Hide

```
print(CP2.2)
```

```
[1] 12.52083
```

Hide

```
print(CP2.3)
```

```
[1] 29.79373
```

d. Does this data provide sufficient evidence that weekend attendance is, on average, larger than weekday attendance? Support your answer.

Yes because in the full model the coefficient for weekend is 933 with p-value=.005, so there is evidence to conclude that weekend attendance is positively correlated with attendance. So we can say that attendance is, on average larger than weekdy attendance. I also created a simple linear model with just weekend and the R2 for that was .34 so we can say that weekend explains 34% of the variance in attendance.

3. The general manager of a chain of catalog stores wanted to determine the factors that affect how long it takes to unload a truck delivering orders. A random sample of 50 deliveries to a store was observed. The times

(in minutes) to unload the truck, the total number of boxes, and the total weight (in hundreds of pounds) were recorded in the file CATALOG.

a. Determine the multiple regression equation.

Time = -41 + .6Boxes + .37Weight

Hide

```
model3.full = lm(formula = Time~Boxes+Weight,data=CATALOG)
summary(model3.full)
```

```

Call:
lm(formula = Time ~ Boxes + Weight, data = CATALOG)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.0809  -3.9519   0.5004   3.6927  17.1160 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -28.42712    6.88696  -4.128 0.000149 ***
Boxes         0.60411    0.05568  10.850 2.16e-14 ***
Weight        0.37430    0.08467   4.420 5.78e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.069 on 47 degrees of freedom
Multiple R-squared:  0.8072,    Adjusted R-squared:  0.799 
F-statistic: 98.37 on 2 and 47 DF,  p-value: < 2.2e-16
```

#b. How well does the model fit the data? Yes. We have a high adjusted R2=.799, low p-value=2.2e-16, very hight F-statistic=98.37 # c. Perform diagnostics on the model and report your findings. I don’t even know what this means. # d. Is multicollinearity a problem? If so, propose a solution. It does not appear to be a problem because the adjusted R2 is not much lower than the multiple R2. If it were a problem we could drop one of the variables which correlates with another causing a problem.

e. Construct a regression model that includes the information for the time of day.

the below code will create a new variable

Hide

```
# encoding morning early afternoon and late afternoon and adding that to the data frame
# 1 = morning, 2 = early afternoon, and 3 = late afternoon.
levels = c(1,2,3)
morning_labels = c(1,0,0)
early_afternoon_labels = c(0,1,0)
late_afternoon_labels = c(0,0,1)
labels = c('morning','early_afternoon','late_afternoon')


# this will create a new column called morning which will store 1 as 1 and other values
 as zero

CATALOG$morning = factor(CATALOG$Codes,levels=levels,labels=morning_labels)
CATALOG$early_afternoon = factor(CATALOG$Codes,levels=levels,labels=early_afternoon_labe
ls)
CATALOG$late_afternoon = factor(CATALOG$Codes,levels=levels,labels=late_afternoon_label
s)
# CATALOG$Codes = factor(CATALOG$Codes,levels=levels,labels=morning_labels)

# df = CATALOG[c(1,2,3,5,6,7)]

model3e = lm(formula = Time ~ Boxes + Weight + morning + early_afternoon + late_afternoo
n,data=CATALOG)
summary(model3e)
```

```

Call:
lm(formula = Time ~ Boxes + Weight + morning + early_afternoon +
    late_afternoon, data = CATALOG)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9104 -2.2655  0.2487  2.1936 10.7709

Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -36.47362    3.96515  -9.199 6.66e-12 ***
Boxes              0.61774    0.03091  19.985  < 2e-16 ***
Weight             0.34638    0.04591   7.545 1.60e-09 ***
morning0           6.75565    1.49928   4.506 4.67e-05 ***
early_afternoon1   6.48003    1.44828   4.474 5.17e-05 ***
late_afternoon1         NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.818 on 45 degrees of freedom
Multiple R-squared:  0.9461,    Adjusted R-squared:  0.9414
F-statistic: 197.7 on 4 and 45 DF,  p-value: < 2.2e-16
```

f. Does the time of day affect the unloading time? Explain.

yes it does I am having some issues with the last column. I suspect this has something to do with the fact that there are too many dependent variables. I tried a few different things with encoding these properly, but it did not fix the NA values I am getting for late_afternoon. Specifically I tried to make the codes variable into the morning variable but this did not work and I could not even get a model from that.