Homework 6

Erik Lee

Sun Aug 5 19:12:51 2018

Problem 1

A real estate agent wishes to determine the selling price of residences using the size (square feet) and whether the residence is a condominium or a single- family home.

Load the data into R using the command

`mydata=read.csv("http://people.fas.harvard.edu/ mparzen/stat104/condo1.csv")`

a) Produce a regression equation using R to predict the selling price for residences (the y variable) using a model of the following form: $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, where x_1 is square footage and x_2 is 1 if a condo and 0 otherwise.

```
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/condo1.csv")
fit=lm(mydata$price~mydata$sqfeet+mydata$condo)
summary(fit)
```

```
##
## Call:
## lm(formula = mydata$price ~ mydata$sqfeet + mydata$condo)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -42627 -15636  -7005  10441  98485
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    66001.3    48476.3    1.36   0.1911
## mydata$sqfeet     90.4       29.5    3.06   0.0071 **
## mydata$condo    3629.5    15891.2    0.23   0.8221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32200 on 17 degrees of freedom
## Multiple R-squared:  0.406,  Adjusted R-squared:  0.337
## F-statistic: 5.82 on 2 and 17 DF,  p-value: 0.0119
```

Regression Equation: y_i = 66001.3 + 90.4x_1 + *3629.5*x_2

x_1 = mydata$sqrfeet x_2 = *mydata*condo

b_0 = 66001.3 b_1 = 90.4 b_2 = 3629.5 e = 0

b) Interpret the parameters b1 and b2 in the model given in part a.

b_1: For every increase of one square foot for the residence, the price of the residence increases by $90.4.

b_2: If the residence is a condo (condo==1), the price of the base price of residence increases by $3629.50. Otherwise (condo==0), the base price does NOT increase by the $3629.50.

c) Fit a new regression model now including the interaction term x1* x2

```
x_3 = mydata$sqfeet*mydata$condo
fit=lm(mydata$price~mydata$sqfeet+mydata$condo+x_3)
summary(fit)
```

```
##
## Call:
## lm(formula = mydata$price ~ mydata$sqfeet + mydata$condo + x_3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24320 -14900  -7791   9482 101302
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    148863.9    64449.6    2.31    0.035 *
## mydata$sqfeet      38.4       39.8    0.97    0.349
## mydata$condo  -167044.2    95189.9   -1.75    0.098 .
## x_3               100.7       55.5    1.82    0.088 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30200 on 16 degrees of freedom
## Multiple R-squared:  0.508,  Adjusted R-squared:  0.416
## F-statistic:  5.5 on 3 and 16 DF,  p-value: 0.00861
```

Regression Equation: y_i = 148863.9 + 38.4x_1 + *(-167044.2)*x_2 + 100.7*x_3

x_1 = mydata$sqrfeet$x_2 = mydata$condo x_3 = mydata$sqfeet * mydata$condo

b_0 = 148863.9 b_1 = 38.4 b_2 = -167044.2 b_3 = 100.7 e = 0

d) Describe what including this interaction term accomplishes.

The interaction term has changed the b values for the intercept (b_0) and slope of the weights (b_1,b_2,b_3). Adding the interaction variable (x_3), b_0 increased from 66001.3 to 148863.9 (increase of 2863). b_1 decreased from 90.4 to 38.4 (decrease of 52). b_2 decreased from 3629.5 to -167044.2 (decrease of 170674). And b_3 was added as 100.7.

With the interaction term, the Residual Standard Error went down from 32200 to 30200 (decrease of 2000), which is a good sign for the model. And the R-sq increased from 0.406 to 0.508, which is also helpful for the model.

However, the biggest issue with the interaction term, and what makes the model poor is that the t-value and p-value for sqfeet changes, indicating a poor relationship with price (y). The t-value=0.97 less than 1.96 and p-value=0.349 greater than 0.05. Both indicate a true value of 0 for beta_1, and accepting the null hypothesis that there is no association between b_1 and y. Adding this term makes the regression model worse.

e) Conduct a test of hypothesis to determine if the relationship between the selling price and the square footage is different between condominiums and single- family homes (that is do you need the interaction term in the model?).

H_o: beta_3 = 0 H_a: beta_3 != 0

```
x_3 = mydata$sqfeet*mydata$condo
fit=lm(mydata$price~mydata$sqfeet+mydata$condo+x_3)
summary(fit)
```

```
##
```

```
## Call:
## lm(formula = mydata$price ~ mydata$sqfeet + mydata$condo + x_3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24320 -14900  -7791   9482 101302
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    148863.9    64449.6    2.31    0.035 *
## mydata$sqfeet      38.4       39.8    0.97    0.349
## mydata$condo  -167044.2    95189.9   -1.75    0.098 .
## x_3               100.7       55.5    1.82    0.088 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30200 on 16 degrees of freedom
## Multiple R-squared:  0.508,  Adjusted R-squared:  0.416
## F-statistic:  5.5 on 3 and 16 DF,  p-value: 0.00861
```

Based on the regression that includes x_3 (sqfeet*condo), the interaction term is not needed in the model. This can be determined with a two tailed hypothesis test of 5% level of significance, the null hypothesis that the true value of beta_3 (slope for the interaction term) is 0 is supported. The t-value for b_3 is 1.82, less than 1.96. The p-value is 0.088, greater than 0.05. Based on these two statistics, it is plausible to support the null hypothesis that the true slope of the interaction term x_3 (sqfeet*condo) is 0. This concludes that the interaction term is not needed in the model.

Problem 2

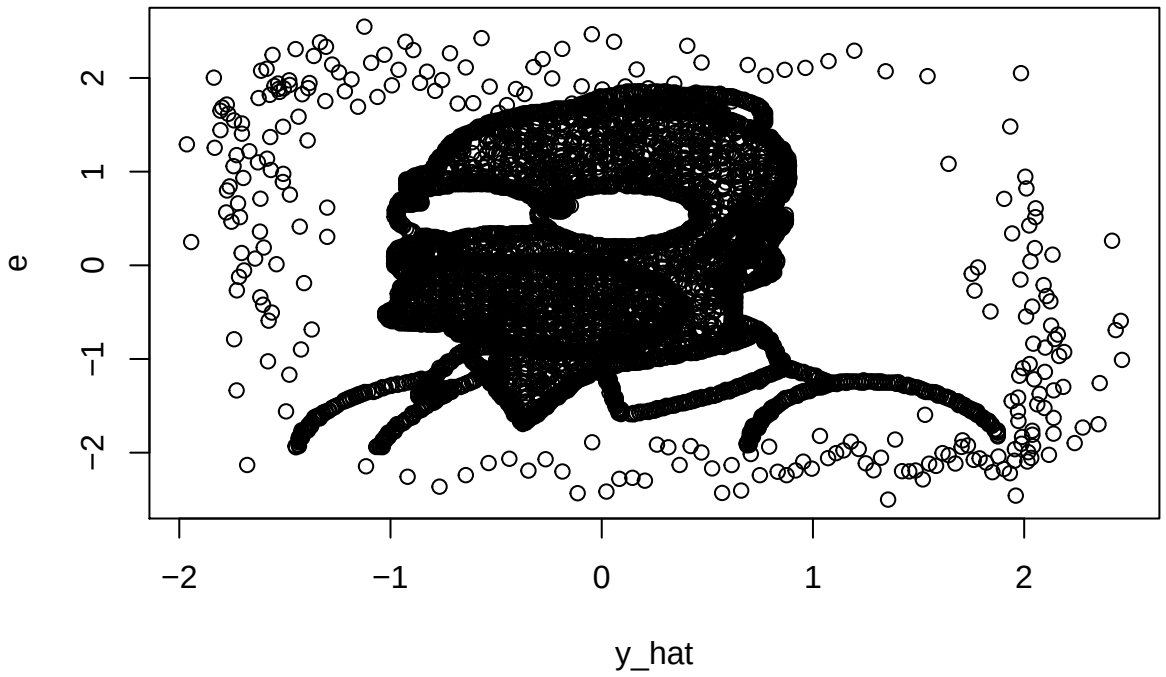
Consider the homer data set shown in class.

Load the data into R using the command

```
mydata=read.csv("http://www.datadescant.com/stat104/homer.csv")
```

a) Run a multiple regression model of Y against all the X variables and then produce the residuals and fitted values. Create the residual diagnostic plot of residuals on the Y axis against the fitted values on the X axis.

```
mydata=read.csv("http://www.datadescant.com/stat104/homer.csv")
fit=lm(mydata$y~mydata$x1+mydata$x2+mydata$x3+mydata$x4+mydata$x5+mydata$x6)
#summary(fit)
y_hat=fitted(fit)
e=residuals(fit)
plot(x=y_hat,y=e)
```



b) What is the result of the test for seeing if the residuals are normally distributed?

```
fit=lm(mydata$y~mydata$x1+mydata$x2+mydata$x3+mydata$x4+mydata$x5+mydata$x6)
library(nortest)
ad.test(residuals(fit))
```

```
##
##  Anderson-Darling normality test
##
## data:  residuals(fit)
## A = 26, p-value <2e-16
```

Normality Test: H_o: residuals are normally distributed H_a: residuals are not normally distributed

Conducting the Anderson-Darling normality test, the p-value is less than 2e-16. With a p-value less than 0.05, the null hypothesis is not supported and the alternative hypothesis can be supported. According to the alternative hypothesis, the residuals are not normally distributed.

c) What is the result of the test for heteroscedasticity?

```
fit=lm(mydata$y~mydata$x1+mydata$x2+mydata$x3+mydata$x4+mydata$x5+mydata$x6)
library(car)
```

```
## Loading required package: carData
```

```
ncvTest(fit)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 29.9    Df = 1     p = 4.555e-08
```

H_o: homoskedasticity H_a: heteroskedasticity

Based on the test for heteroscedasticity, p-value = 4.555e-08. According to this p-value, which is less than 0.05, the null hypothesis can not be supported and the alternative hypothesis can be supported. The alternative hypothesis claims that there is heteroskedasticity in the regression model.

Problem 3

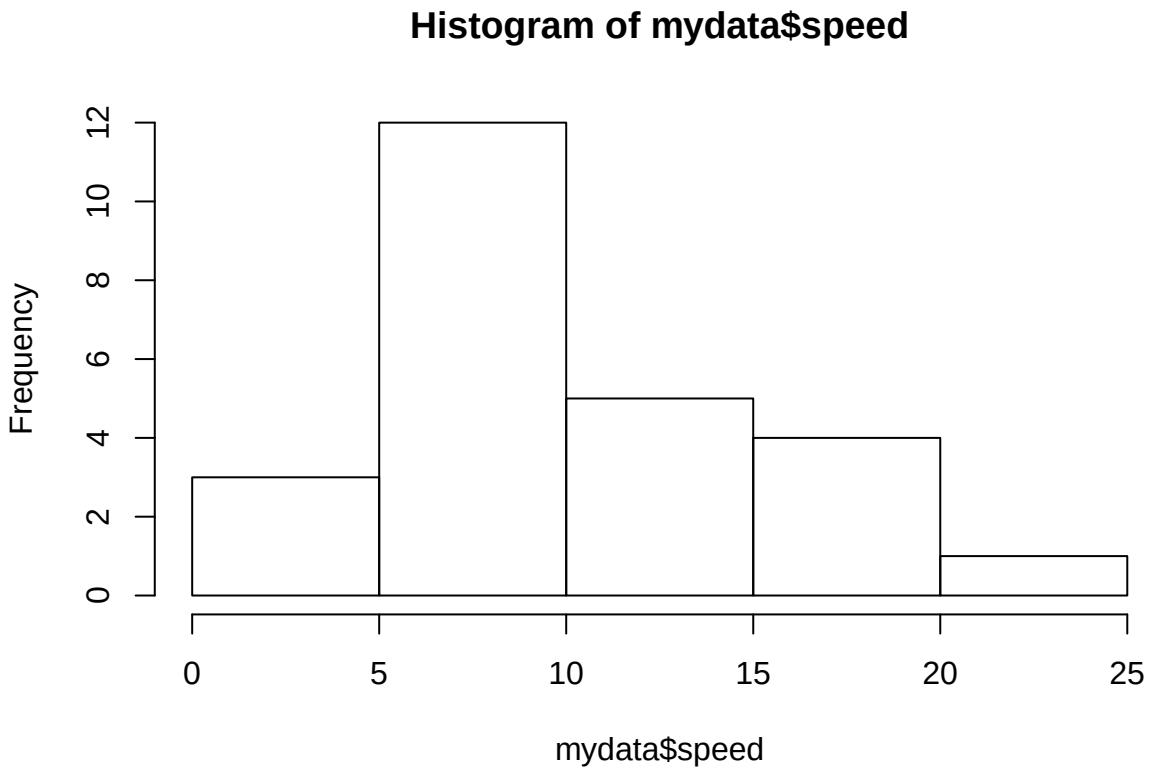
The data in running.csv contains data from a 1963 study to assess energy expenditure while running. Researchers asked athletes to run on a treadmill at various speeds and inclines and assessed energy expenditure (computed indirectly via oxygen consumption and individual body measurements). Speed is measured in kilometers per hour and treadmill incline as either downhill, flat, or uphill. Energy is measured in units of Cal / kg hour.

The goal for this analysis is to create a model that can predict energy expenditure from running speed and treadmill incline; this kind of model would be useful for programming the software on a computerized treadmill that displays how many calories have been burned during an exercise session. Read the data into R using

```
mydata=read.csv("http://www.datadescant.com/stat104/running.csv")
```

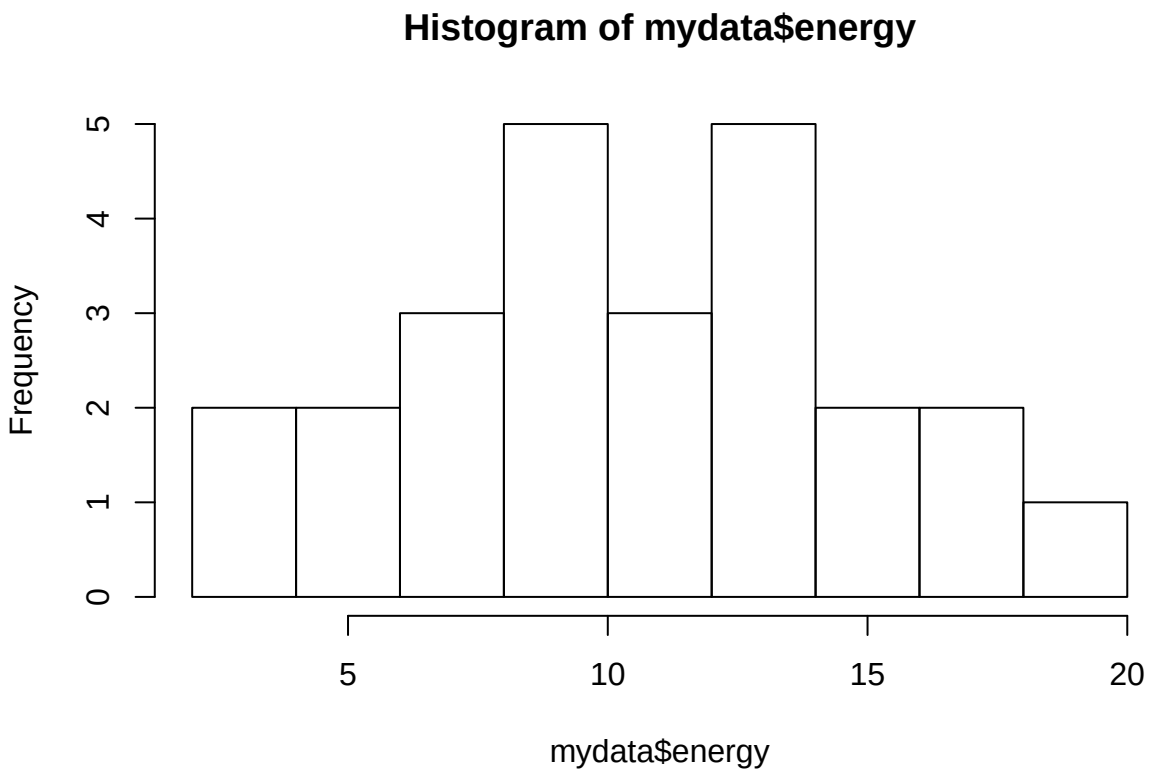
a) Explore the data using numerical and graphical summaries. Describe the distributions of Speed, Energy, and Incline.

```
mydata=read.csv("http://www.datadescant.com/stat104/running.csv")
hist(x=mydata$speed)
```



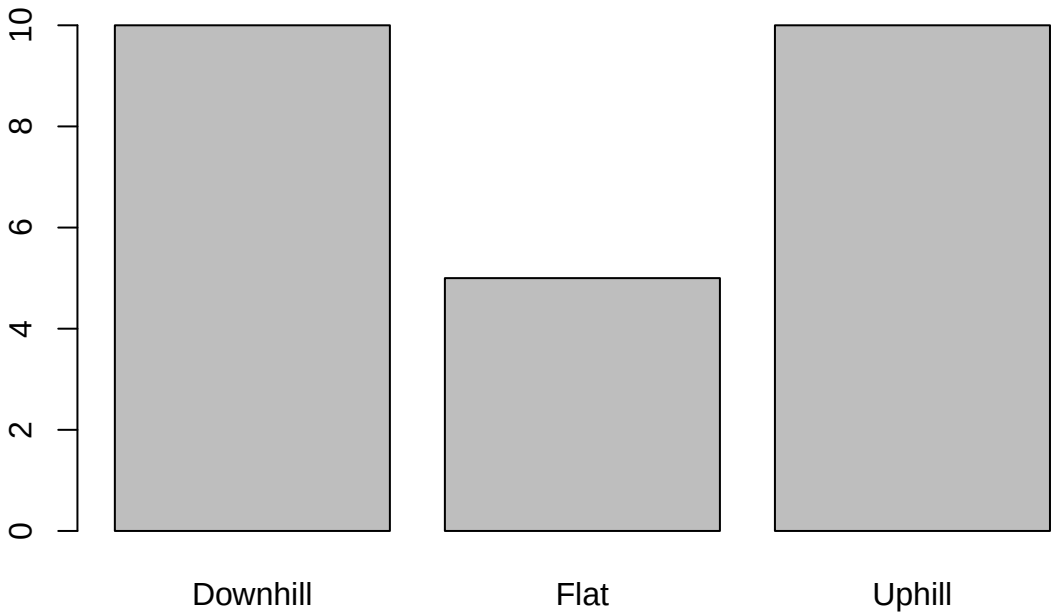
The historgram of speed looks normally distributed with a skew to the right. The values at 20-25 speed look like potential outliers.

```
hist(x=mydata$energy)
```



The historgram of energy expenditure looks pretty normally distributed. The distribution is pretty symmetrical and may have potential outliers at the far right or far left sides of the distribution.

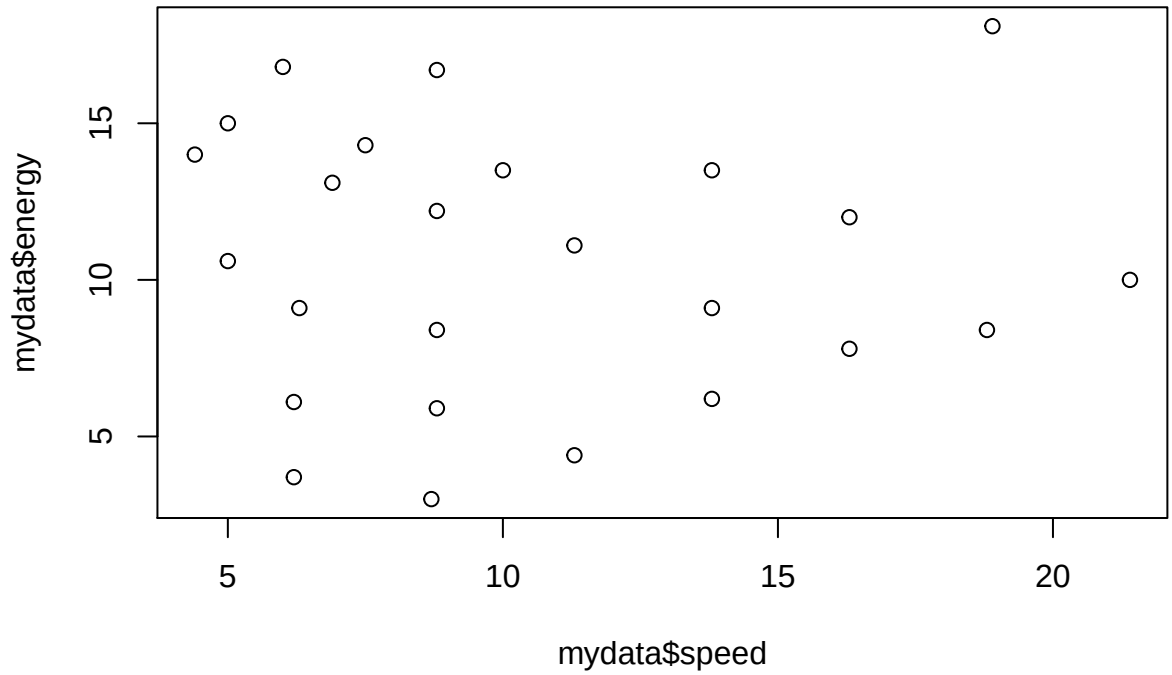
```
plot(mydata$incline)
```



Incline is has equal number of Downhill and Uphill subjects and about half the amount of Flat incline participants as there are Downhill or Uphill.

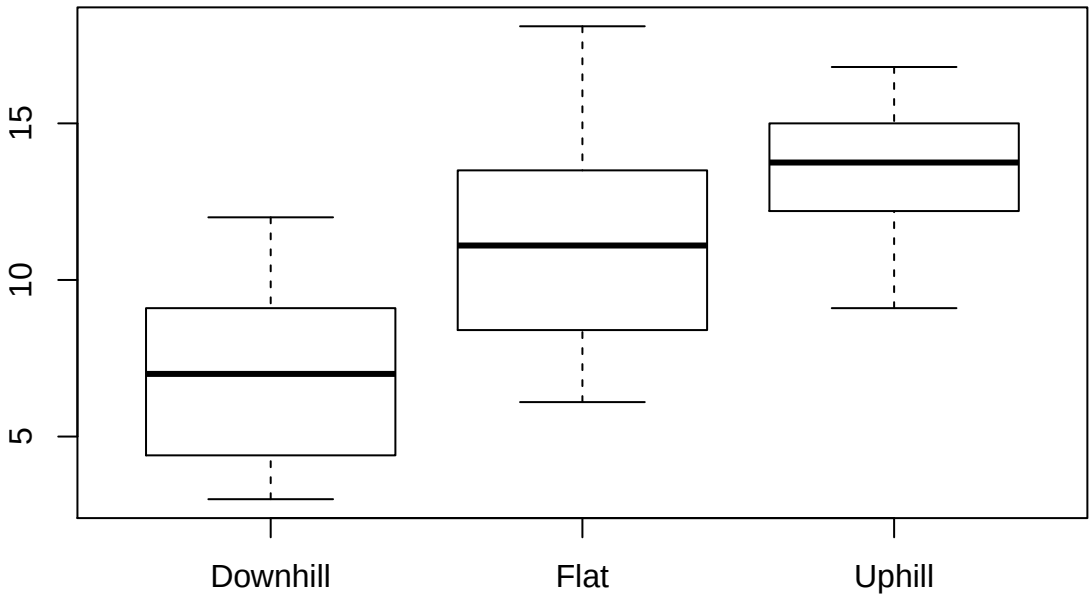
b) Use graphical summaries to explore the relationships between the variables. Describe what you see.

```
plot(x=mydata$speed,y=mydata$energy)
```



The data points for speed vs. energy are evenly spread out. It is hard to discern a linear relationship among the points. It is important to note the scale of the graph. Both speed and energy have ranges of approximately 5-20.

```
plot(x=mydata$incline,y=mydata$energy)
```



There are three boxplots representing each category of incline and their relationships to energy. If we look at the midean values for each level of incline, Downhill has the lowest median, Flat has the next largest median, and Uphill has the largest median energy expenditure.

c) R can neatly create dummy variables for you automatically from a categorical variable. Fit and interpret the following model in R. Which category is defined as the baseline category?

```
      fit=lm(energy~speed+incline,data=mydata)
```

```
fit=lm(energy~speed+incline,data=mydata)
summary(fit)
```

```
##
## Call:
## lm(formula = energy ~ speed + incline, data = mydata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.095 -1.668 -0.448  1.897  3.781
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.906      1.758   -0.52  0.61181
## speed            0.588      0.119    4.92  7.2e-05 ***
## inclineFlat      5.412      1.217    4.45  0.00022 ***
## inclineUphill   10.399      1.262    8.24  5.1e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.19 on 21 degrees of freedom
## Multiple R-squared:  0.767,  Adjusted R-squared:  0.734
## F-statistic:   23 on 3 and 21 DF,  p-value: 7.65e-07
```

The baseline category is Downhill incline.

d) One can include interaction variables by adding the term speed*incline to the model as follows

```
fit=lm(energy~speed*incline,data=mydata)
```

This model is actually fitting three models at the same time; write out what the three models are. Is there evidence that the interaction terms are needed?

```
fit=lm(energy~speed*incline,data=mydata)
summary(fit)
```

```
##
## Call:
## lm(formula = energy ~ speed * incline, data = mydata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.337 -0.976 -0.110  0.872  3.608
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.467      2.027    0.23   0.8203
## speed                  0.486      0.142    3.43   0.0028 **
## inclineFlat           -0.278      3.335   -0.08   0.9345
## inclineUphill         11.945      3.283    3.64   0.0017 **
## speed:inclineFlat      0.467      0.254    1.84   0.0813 .
## speed:inclineUphill   -0.323      0.391   -0.83   0.4178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.05 on 19 degrees of freedom
```

```
## Multiple R-squared:  0.815,  Adjusted R-squared:  0.766
## F-statistic: 16.7 on 5 and 19 DF,  p-value: 2.23e-06
```

Three Models:

Incline=Downhill y_hat = 0.467 + 0.486*speed

Incline=Flat: y_hat = 0.467 + 0.486*speed + (-0.278)*(1) + 0.467*(speed) = 0.189 + 0.953*speed

Incline=Uphill: y_hat = 0.467 + 0.486*speed + (11.945)*(1) + (-0.323)*(speed) = 12.41 + 0.163*speed

There does not seem to be evidence that the interaction term is needed. For speed:inclineFlat t-value=1.84 and p-value=0.0813. Since t is less than 1.96 and p is greater than 0.05, the true value for speed:inclineFlat is hypothesized to be 0. Similarly, speed:inclineUphill has a t-value=-0.83 and p-value=0.4178. The t is less than -1.96 and p is greater than 0.05. This supports the null hypothesis that the true value for speed:inclineUphill can be 0. Since both interaction terms support the null hypothesis that their values can equal 0, the evidence shows that the interaction term is not needed.

Problem 4

Researchers at General Motors collected data on 60 U.S. Standard Metropolitan Statistical Areas (SMSA's) in a study of whether or not air pollution contributes to mortality. The dependent variable for analysis is age adjusted Mortality rate. Several (potential) explanatory variables include measures of demographic characteristics of the cities, of climate characteristics and the air pollution potential of three different chemicals. Specifically:

- JanT Mean January temperature (degrees Fahrenheit)
- JulyT Mean July temperature (degrees Fahrenheit)
- RelHum Relative Humidity
- Rain Annual rainfall (inches)
- Edu Median education
- PopD Population density
- NonWht Percentage of non whites
- WC Percentage of white collar workers
- Pop Population
- PHouse Population per household
- Income Median income
- HCPot HC pollution potential
- NOxPot Nitrous Oxide pollution potential
- SO2Pot Sulfur Dioxide pollution potential

This question is open-ended and will be graded loosely. Using stepwise regression fit a regression model to this data set. You are free to transform any variables you want but do not use any interaction terms. For your final model comment on the residual diagnostic plot.

Here is some R code to get you started [it fits the full model-there is a variable NOx in the dataset we don't need so we first remove it along with the city variable.]

```
> mydata=read.csv("http://www.datadescant.com/stat104/SMSA.csv")
> mynewdata=subset(mydata,select=-c(city,NOx))
> fit=lm(Mortality~.,data=mynewdata)
```

```
mydata=read.csv("http://www.datadescant.com/stat104/SMSA.csv")
mynewdata=subset(mydata,select=-c(city,NOx))
fit=lm(Mortality~.,data=mynewdata)
summary(fit)
```

```
##
## Call:
## lm(formula = Mortality ~ ., data = mynewdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -64.33 -17.32  -1.61  14.82  93.47
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.40e+03   2.82e+02    4.97  1.1e-05 ***
## JanTemp     -1.44e+00   7.64e-01   -1.89    0.066 .
## JulyTemp    -2.95e+00   1.94e+00   -1.52    0.135
## RelHum       1.36e-01   1.15e+00    0.12    0.906
## Rain         9.70e-01   5.85e-01    1.66    0.105
## Education   -1.10e+01   9.00e+00   -1.23    0.226
## PopDensity   4.72e-03   4.34e-03    1.09    0.283
## X.NonWhite   5.30e+00   9.06e-01    5.85  5.6e-07 ***
## X.WC        -1.49e+00   1.23e+00   -1.21    0.232
## pop          3.40e-06   4.12e-06    0.83    0.413
## pop.house   -3.80e+01   4.03e+01   -0.94    0.350
## income      -4.25e-04   1.29e-03   -0.33    0.743
## HCPot       -6.71e-01   4.56e-01   -1.47    0.148
## NOxPot       1.18e+00   9.14e-01    1.29    0.204
## SO2Pot       8.46e-02   1.36e-01    0.62    0.536
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35 on 44 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.762,  Adjusted R-squared:  0.686
## F-statistic: 10.1 on 14 and 44 DF,  p-value: 1.56e-09
```

Test for multicoliniearity within the variables with vif() function. Remove any with a vif > 10.

```
vif(fit)
```

```
##    JanTemp   JulyTemp     RelHum       Rain  Education PopDensity
##      2.851      3.761      1.819      2.175      2.779      1.854
## X.NonWhite       X.WC        pop  pop.house     income      HCPot
##      3.155      1.852      1.910      2.574      1.582     84.479
##     NOxPot     SO2Pot
##     86.360      3.526
```

NOxPot has a high vif of 86.360. This will be removed first.

```
mynewdata=subset(mydata,select=-c(city,NOx,NOxPot))
fit=lm(Mortality~.,data=mynewdata)
```

```
vif(fit)
```

```
##    JanTemp   JulyTemp     RelHum       Rain  Education PopDensity
##      2.851      3.757      1.809      2.048      2.768      1.853
## X.NonWhite       X.WC        pop  pop.house     income      HCPot
##      3.137      1.815      1.903      2.568      1.577      3.511
##     S02Pot
##      1.766
```

After removing NOxPot and checking the vif() function, all other values are below 10. It is interesting to note that HCPot had a high vif of 84.479 with NOxPot. But after NOxPot was removed, HCPot has a new vif of 3.511. It is possible that HCPot and NOxPot were related in some way.

Now as a pre-check, we can test for the heteroskedacity of the current model.

```
ncvTest(fit)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.8932    Df = 1     p = 0.3446
```

Since the p-value = 0.3446 and greater than 0.05, the null hypothesis is supported that the regression is homoskedastic. This result means that no transformation of y is needed and we can proceed to check each x variable in realtion to y.

Now, after testing for multicollinearity and homoskedacity, the variables can be tested and refitted with a backward stepwise regression.

```
mynewdata=subset(mydata,select=-c(city,NOx,NOxPot,RelHum,income,pop,pop.house,HCPot,Education,JulyTemp,
fit=lm(Mortality~.,data=mynewdata)
summary(fit)
```

```
##
## Call:
## lm(formula = Mortality ~ ., data = mynewdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67.04 -22.87   2.01  16.13 114.37
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 870.1622    24.8197   35.06  < 2e-16 ***
## JanTemp      -1.8181     0.5335   -3.41  0.00123 **
## Rain          1.6154     0.4381    3.69  0.00052 ***
## X.NonWhite    4.4349     0.6479    6.85  6.8e-09 ***
## S02Pot        0.3223     0.0781    4.13  0.00013 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.9 on 55 degrees of freedom
## Multiple R-squared:  0.689,  Adjusted R-squared:  0.666
## F-statistic: 30.5 on 4 and 55 DF,  p-value: 2.24e-13
```

Backward-step removal order:

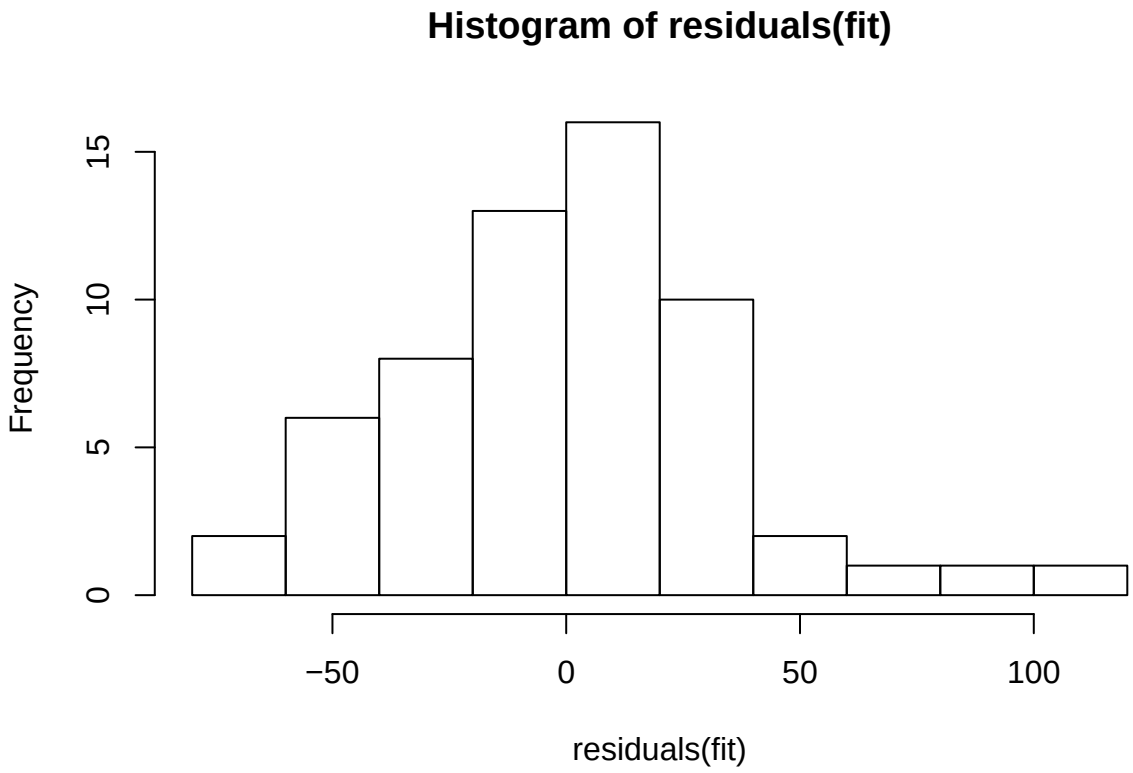
1. RelHum s_e = 35.2, R-sq = 0.753, t-val = 0.02, p-val = 0.982
2. income s_e = 34.8, R-sq = 0.753, t-val = -0.26, p-val = 0.800

3. pop s_e = 34.5, R-sq = 0.753, t-val = 0.74, p-val = 0.466
4. pop.house s_e = 34.2, R-sq = 0.748, t-val = -0.94, p-val = 0.352
5. HCPot s_e = 34.2, R-sq = 0.744, t-val = -0.71, p-val = 0.479
6. Eduction s_e = 34, R-sq = 0.741, t-val = -1.02, p-val = 0.3119
7. JulyTemp s_e = 34, R-sq = 0.736, t-val = -1.36, p-val = 0.1811
8. PopDensity s_e = 34.3, R-sq = 0.727, t-val = 1.86, p-val = 0.0687
9. X.WC s_e = 35.1, R-sq = 0.709, t-val = -1.92, p-val = 0.05986

Listed in order (1-8) are the variables with the highest t-values and p-values removed from the model. The remaining variables are JanTemp, Rain, X.NonWhite, and S02Pot. These values have absolute t-values greater than 1.96 and p-values less than 0.05.

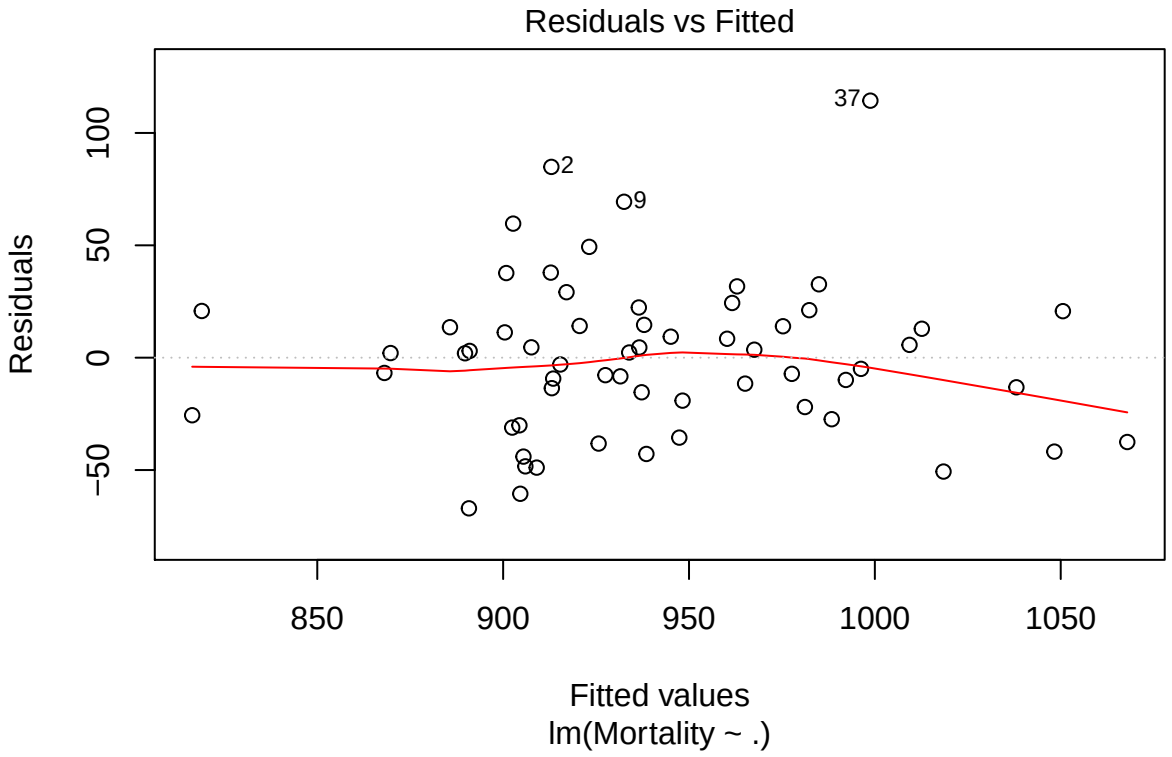
Now that the variables have been narrowed down, we can check the residuals to see if they are normally distrubed.

```
hist(residuals(fit))
```



The histogram of the residuals looks normally distributed with a skew to the right. It may be helpful to figure out which residuals are outliers, with a scatter plot, and remove them.

```r
plot(fit,which=1)
```



Based on the plot of Residuals vs Fitted values there are three outliers as indicated by the scatterplot (2,9,37). These values can dreastically affect R-sq and s_e. The next step would be to remove these outliers.

```r
new_data=subset(mydata,abs(rstudent(fit))<2)
fit3=update(fit,.~.,data=new_data)
summary(fit3)
```

```
##
## Call:
## lm(formula = Mortality ~ JanTemp + Rain + X.NonWhite + SO2Pot,
##     data = new_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -62.14 -20.02   2.16  17.27  62.12
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 869.2478    20.7310   41.93  < 2e-16 ***
## JanTemp      -1.7860     0.4406   -4.05  0.00017 ***
## Rain          1.5308     0.3553    4.31  7.3e-05 ***
## X.NonWhite    4.0763     0.5321    7.66  4.4e-10 ***
## SO2Pot        0.3685     0.0635    5.80  4.0e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29 on 52 degrees of freedom
## Multiple R-squared:  0.77,   Adjusted R-squared:  0.752
```

```
## F-statistic: 43.5 on 4 and 52 DF,  p-value: 5.37e-16
```

After removing the residuals s_e went down to 29 and R-sq went up to 0.77.

Finally, we can check the heterskedasticity and normality of residuals before accepting the final fit.

```
ncvTest(fit3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.9367    Df = 1     p = 0.3331
```

```
ad.test(residuals(fit3))
```

```
##
##  Anderson-Darling normality test
##
## data:  residuals(fit3)
## A = 0.17, p-value = 0.9
```

The final model is homoskedastic with a p-value=0.3331 and residuals are normally distributed with p-value=0.9.

```
summary(fit3)
```

```
##
## Call:
## lm(formula = Mortality ~ JanTemp + Rain + X.NonWhite + S02Pot,
##     data = new_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -62.14 -20.02   2.16  17.27  62.12
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 869.2478    20.7310   41.93  < 2e-16 ***
## JanTemp      -1.7860     0.4406   -4.05  0.00017 ***
## Rain          1.5308     0.3553    4.31  7.3e-05 ***
## X.NonWhite    4.0763     0.5321    7.66  4.4e-10 ***
## S02Pot        0.3685     0.0635    5.80  4.0e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29 on 52 degrees of freedom
## Multiple R-squared:  0.77,   Adjusted R-squared:  0.752
## F-statistic: 43.5 on 4 and 52 DF,  p-value: 5.37e-16
```

This is the summary data for the final regression model. The dependent variable of Mortality depends on JanTemp, Rain, X.NonWhite, and S02Pot.

The following questions are multiple choice questions from former final exams. The answers are out on the internet-however, we want you to show your work.

Problem 5

Show using the fact that SST=SSR+SSE and the definition of R^2 that if $R^2 = 0$, SSE must be what value?

R^2 = SSR/SST = 1 - (SSE/SST) 0 = 1 - (SSE/SST) SSR/SST = 1 SSE = SST

R^2 = 1 - (SSE/SST) if SSE = SST, R^2 = 1 - (SST/SST) = 1 - 1 = 0

If R^0 = 0, then SSE = SST

The SSE (error sum of squares) equals SST (total sum of squares). Based on this calculation, when the SSE is exatly equal to the SST, then R^2 = 0.

Problem 6

Suppose we have the following regression model where X is a continuous variable and Male=1 if male and 0 otherwise: `Yhat = 10 + 2X-3Male + 5X*Male`. If we run the regression on the same data but now with a dummy variable for Female, what would the estimated model be?

Y_hat = 10 + 2X - 3*Male* + *5X*Male

Male=1: Y_hat = 10 + 2X - 3*(1)* + *5X*(1) = 13 + 7X

Female=0: Y_hat = 10 + 2X - 3*(0)* + *5X*(0) = 10 + 2X

Dummy Variable for Female=0:

Male=0: Y_hat = 13 + 7X - 3*Female* - *5X*Female Y_hat = 13 + 7X - 3*(0)* - *5X*(0) = 13 + 7X

Female=1: Y_hat = 13 + 7X - 3*Female* - *5X*Female Y_hat = 13 + 7X - 3*(1)* - *5X*(1) = 10 + 2X

Estimated Model: Y_hat = 13 + 7X - 3*Female* - *5X*Female, where Female=1 and Male=0

Problem 7

A study attempted to establish a linear relationship between IQ score and musical aptitude. The following table is a partial printout of the regression analysis and is based on a sample of 30 individuals.

```
  MusApt |    Coef.   Std. Err.   t     P>|t|
---------+-------------------------------------------
      IQ |   .4925   .1215
   _cons |   -22.26   12.94      -1.72   0.102
```

What is the value of the test statistic for $H_o : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ is?

H_o: beta_1 = 0 H_a: beta_1 != 0

n = 30, size of sample b_1 = 0.4925, guess of beta_1 se_b1 = 0.1215, std. err. of b_1

beta_1* = 0, hypothesis test value of beta_1

t-value = (b_1 - beta_1*)/se_b1 = (b_1 - 0)/se_b1 = b_1/se_b1 = 0.4925/0.1215 = 4.05

Problem 8

For the following, tell if each statement is true or false. Justify your answers.

Consider the error term ϵ in the regression model:

- The expected value of the error term is one.

- The variance of the error term is the same for all values of x.
- The values of the error term are independent.
- The error term is normally distributed.

1. false, the expected value of the error term is zero. In the assumption of normal distribution of error, e~N(0,sigmaˆ2), the mean or expectation is 0 since the model aims to minimize the errors/residuals when summed.
2. true, a multiple regression model has a single epsilon value. The epsilon does not depend on any of the values of x, so the variance of the error is the same for all the values of x. Epsilon depends strictly on the error calculated by the difference in observed and expected y.
3. true, the error term does not depend on the value(s) of the x. They depend only on Y values for expected and observed, and are independent of any x values.
4. true, the errors are expected to be a normally distributed random variable with a mean of 0 and variance of sigmaˆ2. A normal distribution checks if the random errors follow the stated confidence levels from the inferences made about errors.
