Homework - Chapter 11

Tracy Holsclaw

July 2013

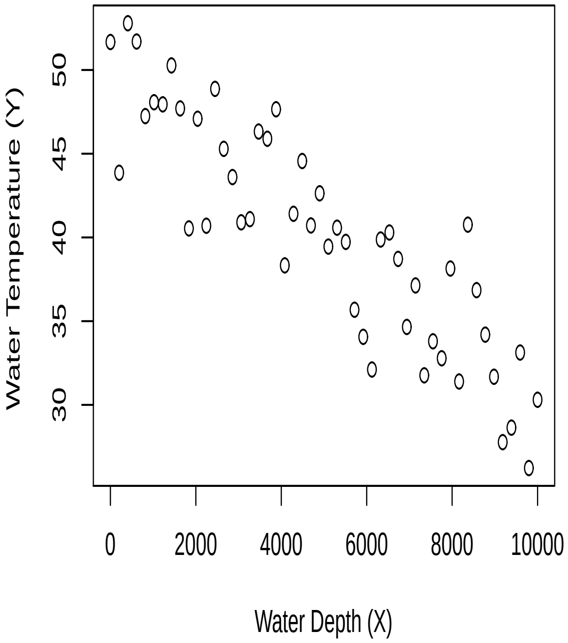
1 Problem 1

I am including the R code for generating the Example in class with X (water depth) and Y (water temperature). Usually the data (X,Y) would be collected from a study. But for the point of a clean example, I generated this data so I know it is Normally distributed. The > is the command prompt in R. The pound sign is used for comments.

```
> set.seed(154)                # set the random number generator seed
> n=50                         # set the number of observations
> b0=50                        # set b0 to generate data
> b1=-.002                     # set b1 to generate data
> x=seq(1,10000,length=n)      #generate an evenly spaced sequence of X (water depth)
> y=b0+b1*x+rnorm(n,0,sd=3)    #generate the mean fit b0+b1*x and add random Normal noise
>                              #  rnorm() to get randomly generated Y (water temperature)

> x

 [1]     1.0000   205.0612   409.1224   613.1837   817.2449  1021.3061  1225.3673  1429.4286
[10]  1837.5510  2041.6122  2245.6735  2449.7347  2653.7959  2857.8571  3061.9184  3265.9796
[19]  3674.1020  3878.1633  4082.2245  4286.2857  4490.3469  4694.4082  4898.4694  5102.5306
[28]  5510.6531  5714.7143  5918.7755  6122.8367  6326.8980  6530.9592  6735.0204  6939.0816
[37]  7347.2041  7551.2653  7755.3265  7959.3878  8163.4490  8367.5102  8571.5714  8775.6327
[46]  9183.7551  9387.8163  9591.8776  9795.9388 10000.0000

> y

 [1] 51.66951 43.86143 52.79129 51.69816 47.24592 48.07322 47.94730 50.27228 47.70287 40.537
[12] 40.69745 48.87059 45.29188 43.60320 40.90172 41.09186 46.32451 45.89228 47.65380 38.330
[23] 44.56249 40.71576 42.63473 39.45306 40.58191 39.73388 35.68130 34.06277 32.11162 39.875
[34] 38.71599 34.65571 37.13193 31.76396 33.79624 32.78296 38.14314 31.39671 40.76280 36.851
[45] 31.68024 27.77360 28.64064 33.12556 26.22646 30.30604
```

We can go ahead and plot X and Y. There are a few extra parameters in the plot function that I like to use. xlab and ylab are used to label the axis. If you ever want to look up a built in R function, you can just append a question mark to the front of the function (?plot)

```
> par( mar=c(4,4,1,1))
> plot(x,y, xlab="Water Depth (X)", ylab="Water Temperature (Y)")     #plot x vs y
```



A) The regression assumptions include that the Y values are independent, the variance is constant for all values of X, the regression is linear, and the noise of Y after fitting the mean (regression curve) is Normally distributed. If we did not know how the data was generated (pretend this is real data you collected) which of these can we assess at this point in the analysis?

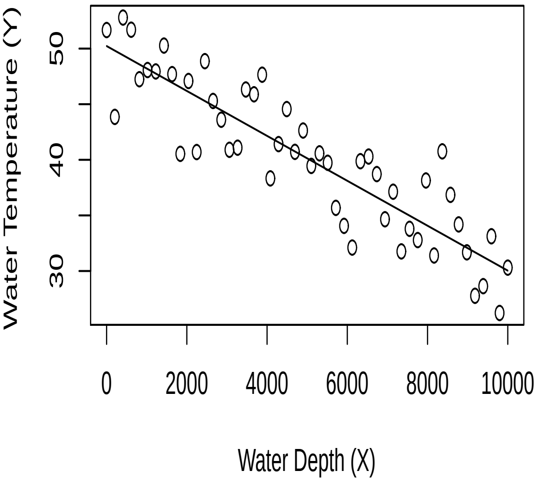
Next we run the regression using the built in R function for linear modeling (lm). I am calling the object created by the lm function "my.reg". The summary function is used to give a summary table for the object created by lm.

```
> my.reg=lm(y~x)
> summary(my.reg)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9824 -2.4945  0.2309  2.3904  7.4308

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.231557   0.882153   56.94   <2e-16 ***
x           -0.002020   0.000152  -13.29   <2e-16 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

Residual standard error: 3.165 on 48 degrees of freedom
Multiple R-squared: 0.7862,        Adjusted R-squared: 0.7818
F-statistic: 176.5 on 1 and 48 DF,  p-value: < 2.2e-16
```

B) What is the point estimate for β_0 and β_1?
C) What is the p-value for β_0? What is it testing?
D) What is the p-value for β_1? What is it testing?
E) What is the distribution used to obtain these p-values?
F) What is the R^2? What is the interpretation? Does this say anything about how linear the relationship is?
G) What is the mean fit of the regression in terms of X?
H) What would be the best point estimate for the temperature at water depth 4000?
I) What would be the best point estimate for the temperature at water depth 15000?

We plot the X and Y. Then we want to use the mean fit (sample mean) for the regression and plot it as a line through the data. The object my.reg has some properties, you can see the properties by using the function "names", try names(my.reg). I already know the property we want is fitted.values(), which gives the mean fit for each x value. We connect these points with the lines() plotting function. plot() is used first and then lines() are added to it.

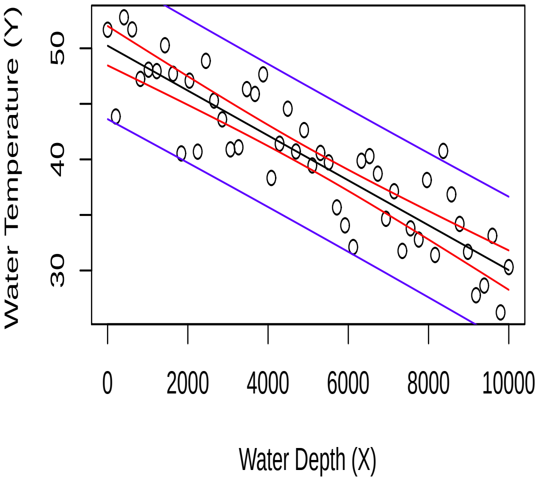
```
> plot(x,y, xlab="Water Depth (X)", ylab="Water Temperature (Y)")        #plot x vs y
> lines(x,fitted.values(my.reg))
```



J) What was minimized to get this "best" fitting line through the data points?

Next we want to add the 95% confidence bands to the mean fit in red (col=2 in the plot function) and 95% bands for the data in blue (col=4). The default color used is black (col=1). R tends to throw out warning messages with the predict.lm function, in this case they are not important.

```
> plot(x,y, xlab="Water Depth (X)", ylab="Water Temperature (Y)")        #plot x vs y
> lines(x,fitted.values(my.reg))
> lines(x,predict.lm(my.reg,interval="confidence")[,2], col=2)
> lines(x,predict.lm(my.reg,interval="confidence")[,3], col=2)
> lines(x,predict.lm(my.reg,interval="prediction")[,2], col=4)
> lines(x,predict.lm(my.reg,interval="prediction")[,3], col=4)
```



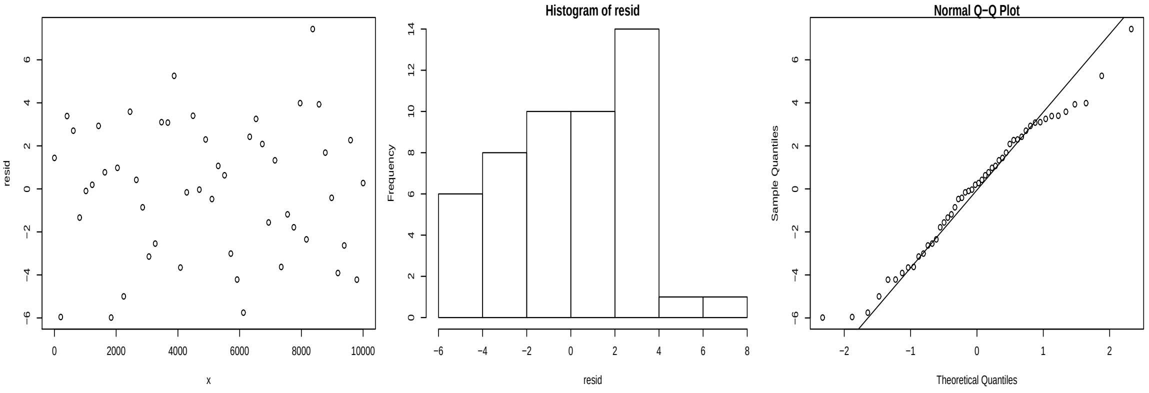
K) How many points should be outside the blue lines on a 95% confidence band if there are n=50 data points?
L) How many points do you observe outside the blue lines for this one randomly generated data set?

At this point we want to plot and check the residuals. The residuals tell us the most about the assumptions of the regression and we can see if any of the assumptions are violated. Residuals are calculated by taking the difference between the observed Y and the mean fitted line at the same X location

```
> resid=y-fitted.values(my.reg)    #or resid(my.reg) will give them automatically
```

We include three possible residual plots: a X vs residual plot, a histogram of the residuals, and a QQ-plot of the residuals. (The par() function sets the plotting parameters and we want three plots to show up with 1 row and 3 columns, so the figures are in a line.)

```
> par(mfrow=c(1,3), mar=c(4,4,1,1))
> plot(x,resid)
> hist(resid)
> qqnorm(resid)
> qqline(resid)
```

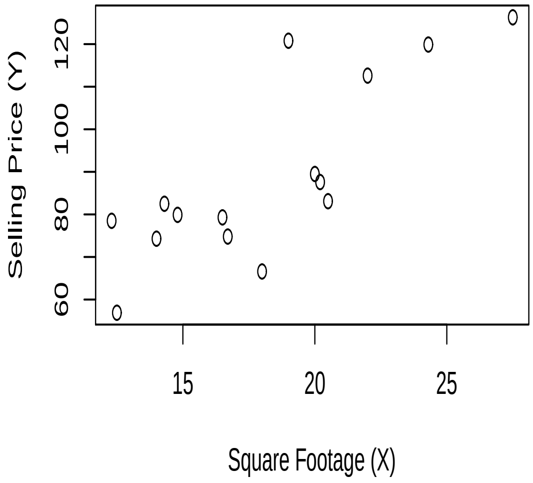


M) What assumption are we assessing in the first residual plot?
N) What assumption are we assessing in the second residual plot?
O) What assumption are we assessing in the third residual plot?
P) How do we feel about the regression assumptions at this point? Was this regression analysis we performed valid?

2 Problem 2

We want to look at the square footage of a house (X) and its selling price (Y). We have some data in a text file; the first column is X and the second is Y. The square footage is divided by 100 and the selling price is divided by 1000 just so the numbers are not as large.

```
> infile="C://Documents and Settings//Owner//Desktop//Statistics67//Notes//9-Chapter11//Hmwk
> data=read.table(infile)
> x=data[,1]
> o1=order(x)    #find the order of the x
> x=sort(x)      #re-order the x from highest to lowest (makes plots nice)
> x

 [1] 12.3 12.5 14.0 14.3 14.8 16.5 16.7 18.0 19.0 20.0 20.2 20.5 22.0 24.3 27.5

> y=data[o1,2]   #order the y data, so it lines up with the x data
> y

 [1]  78.5  56.9  74.3  82.5  79.9  79.3  74.8  66.6 120.8  89.5  87.6  83.1 112.6 119.9 126

> plot(x,y, xlab="Square Footage (X)", ylab="Selling Price (Y)")     #plot x vs y
```



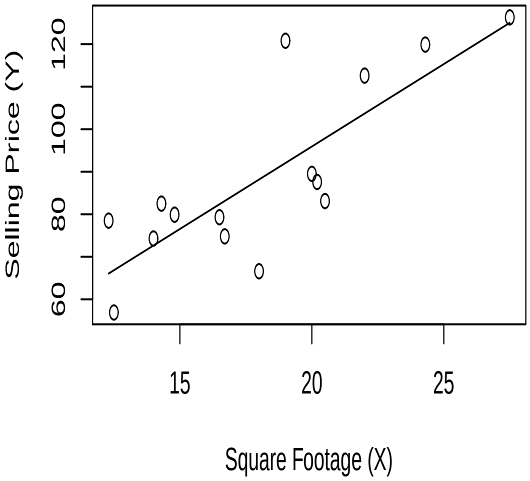
A) What assumptions can we assess at this point in the regression?

```
> my.reg=lm(y~x)
> summary(my.reg)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-21.568  -8.713   1.286   7.990  28.754

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.3538    14.8077   1.239 0.237080
x             3.8786     0.7936   4.887 0.000297 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

Residual standard error: 13 on 13 degrees of freedom
Multiple R-squared: 0.6476,	Adjusted R-squared: 0.6204
F-statistic: 23.89 on 1 and 13 DF,  p-value: 0.0002968
```

B) What is the point estimate for β_0 and β_1?
C) What is the p-value for β_0? What is it testing?
D) What is the p-value for β_1? What is it testing?
E) What is the distribution used to obtain these p-values?
F) What is the R^2? What is the interpretation? Does this say anything about how linear the relationship is?
G) What is the mean fit of the regression in terms of X?

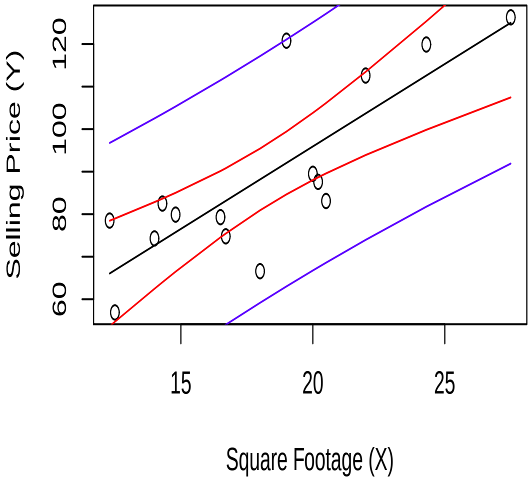
H) What would be the best point estimate for the price of a house with square footage 19?
I) What would be the best point estimate for the price of a house with square footage 50?

```
> plot(x,y, xlab="Square Footage (X)", ylab="Selling Price (Y)")       #plot x vs y
> lines(x,fitted.values(my.reg))
```



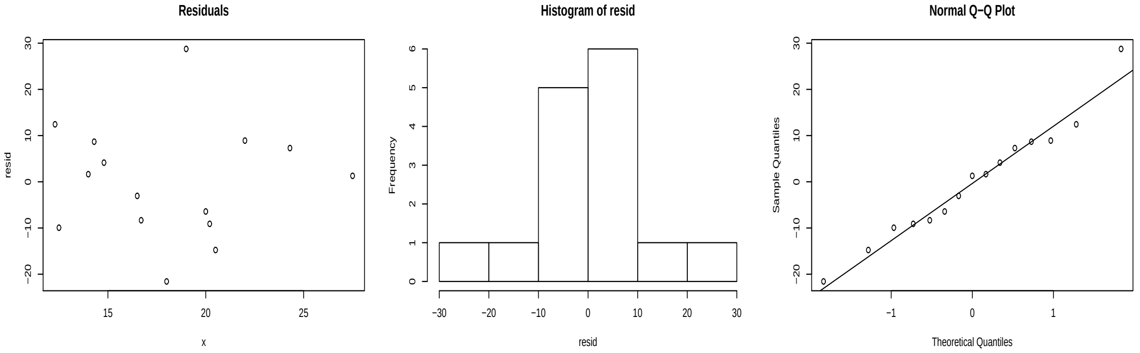
J) What was minimized to get this "best" fitting line through the data points?

```
> plot(x,y, xlab="Square Footage (X)", ylab="Selling Price (Y)")       #plot x vs y
> lines(x,fitted.values(my.reg))
> lines(x,predict.lm(my.reg,interval="confidence")[,2], col=2)
> lines(x,predict.lm(my.reg,interval="confidence")[,3], col=2)
> lines(x,predict.lm(my.reg,interval="prediction")[,2], col=4)
> lines(x,predict.lm(my.reg,interval="prediction")[,3], col=4)
```



K) How many points should be outside the blue lines on a 95% confidence band if there are n=50 data points?
L) How many points do you observe outside the blue lines for this one randomly generated data set?

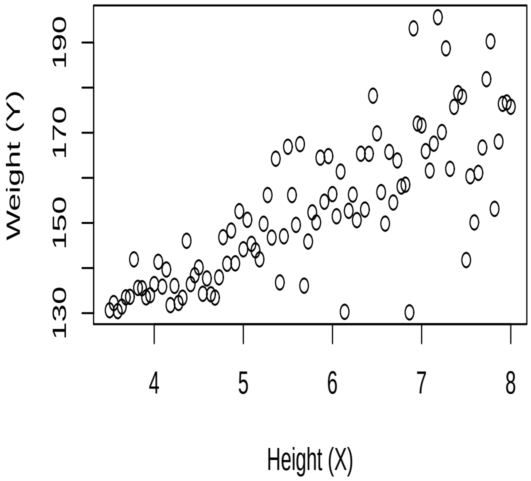
```
> resid=y-fitted.values(my.reg)    #or resid(my.reg) will give them automatically
> resid

        1          2          3          4          5          6          7          8
12.440023  -9.935687   1.646486   8.682921   4.143645  -3.049891  -8.325602 -21.567718  28.
       10         11         12         13         14         15
-6.424820  -9.100530 -14.764095   8.918078   7.297411   1.286048

> par(mfrow=c(1,3))
> plot(x,resid, main="Residuals")
> hist(resid)
> qqnorm(resid)
> qqline(resid)
```



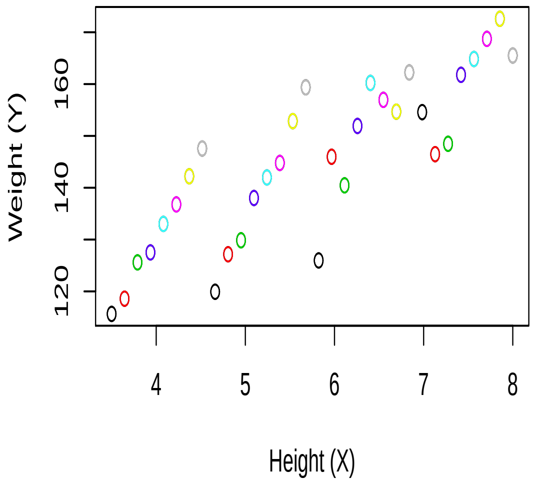
M) What assumption are we assessing in the first residual plot?
N) What assumption are we assessing in the second residual plot?
O) What assumption are we assessing in the third residual plot?
P) How do we feel about the regression assumptions at this point? Was this regression analysis we performed valid?

3 Problem 3

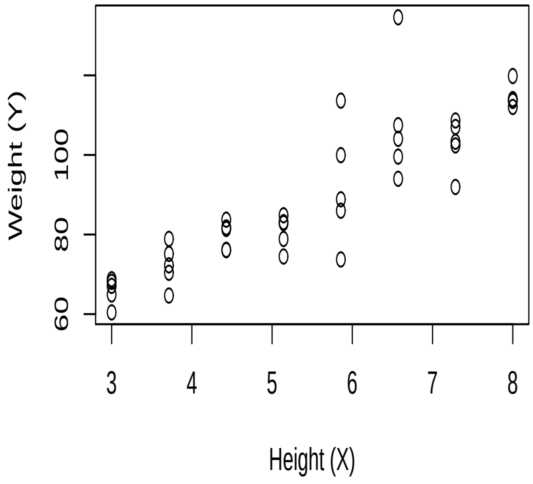
We show some different scenarios of data (X,Y), determine whether the assumptions for linear regression are met.
A) We collect the height (X) and weight (Y) of 100 randomly selected people.



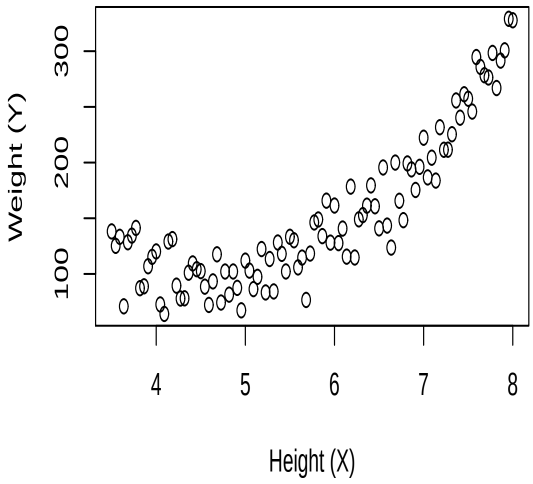
B) We collect the height (X) and weight (Y) of 8 families. Each family has four members. We plot each family in a different color.



C) We collect the height (X) and weight (Y) of 40 people. We choose specific heights of interest and randomly collect 5 people with that height and take their weight measurement.



D) We collect the height (X) and weight (Y) of 100 randomly selected people.



4 Extra Credit

Part 1:
You can get extra credit for doing this next problem. It is extra as you need to use statistical software. R is the statistics package that I have been using in this homework. It is free and can be downloaded at: http://cran.r-project.org/. The data set is from the UCI Machine Learning repository. http://archive.ics.uci.edu/ml/datasets/Auto+MPG We will use column 5 which is the weight of vehicles as our X and column 1 which is miles per gallon (mpg) for our Y. We will study how the weight of a car affect its miles per gallon?

The infile syntax I am using is for Windows. For Unix it is infile=”C:\Documents and Settings\Owner\Desktop\Statistics67\Notes\9-Chapter11\Hmwk\data2.txt” You will have to set the correct pathname and file syntax to read in the file. I am providing the first few lines of code to help open the text file and order x and y. The ordering and sorting is so the plotting functions will look good.

```
> infile="C://Documents and Settings//Owner//Desktop//Statistics67//Notes//9-Chapter11//Hmwk
> data=read.table(infile)
> x=data[,5]      #use the 5th column of data (weight of the vehicle)
> o1=order(x)     #find the order of the x
> x=sort(x)       #re-order the x from highest to lowest (makes plots nice)
> y=data[o1,1]    #order the y data, so it lines up with the x data, use the first column of
```

Provide output for:
1. Plot the data and 95% bands.
2. Perform a regression analysis - provide a summary Table.
3. Do some residual plots.
Write answers for:
4. Is there a linear relationship between the weight of a vehicle (X) and mpg (Y)?
5. Are the assumptions of the regression met (they are close enough, but which ones might be somewhat questionable)?

Part 2:
Generate some data. I chose the number 400 to set the seed of the random number generator.

Change this number to anything (DO NOT USE 400). (This should be different from anyone elses number in the class, so your data will be different from everyone elses.)

```
> set.seed(400)                # set the random number generator seed
> n=20                         # set the number of observations
> b0=50                        # set b0 to generate data
> b1=-.002                     # set b1 to generate data
> x=seq(1,10000,length=n)      #generate an evenly spaced sequence of X (water depth)
> y=b0+b1*x+rnorm(n,0,sd=3)    #generate the mean fit b0+b1*x and add random Normal noise
```

Provide output for:
What number did you use to set the seed?
1. Plot the data and 95% bands.
2. Perform a regression analysis - provide a summary Table.
3. Do some residual plots.
Write answers for:
4. Is there a linear relationship between X and Y?
5. Are the assumptions of the regression met (they are close enough, but which ones might be somewhat questionable)?

5 Solutions

5.1 Problem 1

A) The sample is random, so the assumption about independence is met. The variance for small values of X seems to be the same as for large values of X. For example, X near zero has a range of 42 to about 55 which is a range of about 13. At X near 8000 the range is about 32 to about 43 which is a range of about 11. It seems that the variance is pretty constant across all ranges of X. We cannot say much at this time about the Normal distribution property and if that assumption is valid.
B) $\hat{\beta_0} = 50.23$ and $\hat{\beta_1} = -0.00202$
C) The p-value is less than $2 * 10^{-16}$. It is testing the hypothesis that the intercept of the regression (where the mean regression line crosses the y-axis) is zero
D) The p-value is less than $2 * 10^{-16}$. It is testing the hypothesis that the slope of the regression line is zero. In practical terms, it is testing if there is no linear relationship between water depth and water temperature. This is a small p-value, so we fail to retain H_0 and conclude that water temperature is related to water depth.
E) t-distribution for β_0 and β_1 because we have to estimate s^2. The t-value given is the test statistic and the p-value is found from this test statistic (the computer looked it up on the Table for you.)
F) $R^2 = 0.7818$ (you can use either R^2). R^2 represents the percent variation in Y that can be explained by the regression model (changes in X.) In this case the R^2 is pretty high, so the changes in X explains about 80% of the change in Y, the part X cannot explain is random noise. R^2 can only be used on linear relationships and does not tell us if the relationship is linear. G) $\hat{y} = \hat{\beta_0} + \hat{\beta_1}X = 50.23 - 0.00202 * X$
H) $\hat{y} = 50.23 - 0.00202 * (4000) = 42.15$
I) $\hat{y} = 50.23 - 0.00202 * 15000 = 19.93$, while we can find this value, it is not valid. The X range is only from 0 to 10000. 15000 is well outside this range and we cannot accurately comment on what will happen at a depth X we have no data around.
J) Sum of the squares of the errors.
K) .05*50=2.5, we would expect to see around 2 or 3 points outside the blue bands.
L) one outside and few very near or one the blue bands, this is pretty close to our estimate of 2-3.

M) constant variance
N) Normality
O) Normality
P) The second and third residual plot are a bit weak on the Normality but this is a small sample of $n = 50$. Since this is simulated data we know the actual underlying assumptions are valid. There is random noise in the model and so no residual plots are perfect but they can be close enough. The bins on the histogram can occasionally be misleading with small data sets. The constant variance and Normality assumption (especially in the QQplot) seem reasonable. The Normality assummption might be questioned because of residual plot two.

5.2 Problem 2

A) The sample is random, we are not taking multiple measures of the same houses and we are assuming it is random, so we do not have clusters of houses from good or bad neighborhoods that might be correlation. The assumption about independence seems to be met. We cannot comment on Normality at this time. The relationship seems linear, eventhough there is little data. And there is no reason to question the constant variance assumption.
B) $\hat{\beta_0} = 18.36$ and $\hat{\beta_1} = 3.879$
C) p-value=0.237. It is testing the hypothesis that the intercept of the regression (where the mean regression line crosses the y-axis) is zero. We would conclude that the regression line probably does go through the point (0,0). In practical terms, a zero square foot house would cost zero dollars.
D) -value=0.00297. It is testing the hypothesis that the slope of the regression line is zero. In practical terms, it is testing if there is linear relationship between square footage and housing price. The pvalue is small, so we fail to retain H_0 (reject H_0) and conclude that the square footage does related to the price of a house.
E) t-distribution for β_0 and β_1 because we have to estimate s^2. The t-value given is the test statistic and the p-value is found from this test statistic (the computer looked it up on the Table for you.)
F) $R^2 = 0.6204$ (you can use either R^2). R^2 represents the percent variation in Y that can be explained by the regression model (changes in X.) In this case the R^2 is pretty high, so the changes in X explains about 62% of the change in Y, the part X cannot explain is random noise. R^2 can only be used on linear relationships and does not tell us if the relationship is linear. G) $\hat{y} = \hat{\beta_0} + \hat{\beta_1}X = 18.35 + 3.8786 * X$
H) $\hat{y} = 18.35 + 3.8786 * (19) = 92.04$
I) $\hat{y} = 18.35 + 3.8786 * (50) = 212.28$, while we can find this value, it is not valid. The X range is only from 5 to 30. 50 is well outside this range and we cannot accurately comment on what will happen if a house has 50 sqft, we have no data around that value of X.
J) Sum of the squares of the errors.
K) .05*15=0.75, we would expect to see around 0 or 1 points outside the blue bands.
L) one on the band, this is pretty close to our estimate of one.
M) constant variance, there are so few points to assess this well. But nothing obvious enough to reject the assumption of constant variance
N) Normality
O) Normality
P) The regression assumptions seem pretty good. It can be hard to tell with such a small sample size. But the residual plots look close to Normal and there is not enough evidence to seriously question the non-constant variance. These things tend to be judgement calls on the part of the person doing the analysis. After generating random data and staring at hundreds of these plots it gets easier to distinguish between true patterns and just noise.

5.3 Problem 3

A) violates constant variance assumption, shorter people have less weight range than tall people. B) violates the independence assumption, people from the same family will tend to be the same height. For example, all four members of the gray family are quite tall. And the green family is quite a bit shorter. C) This is fine. X is a fixed quantity and not a variable in regression. We can set specific values of X before the study if we wish. This tends to happen often in drug studies, where several set doses of medication are used as the X and then random patients are given specified doses. For example, a dose of 50mg is administered to 8 people and a dose of 70mg is administered to 5 people, and a dose of 100mg is administered to 10 people. As long as the people are different for each does, this is random. D) X and Y are not linearly related. There is a curve. We cannot fit a regression line to this. We would need to do something more complex and fit a curve.