Stat 104 - Fall 2004

Assignment 2 **Solution keys**

Written Assignment (20pts total)

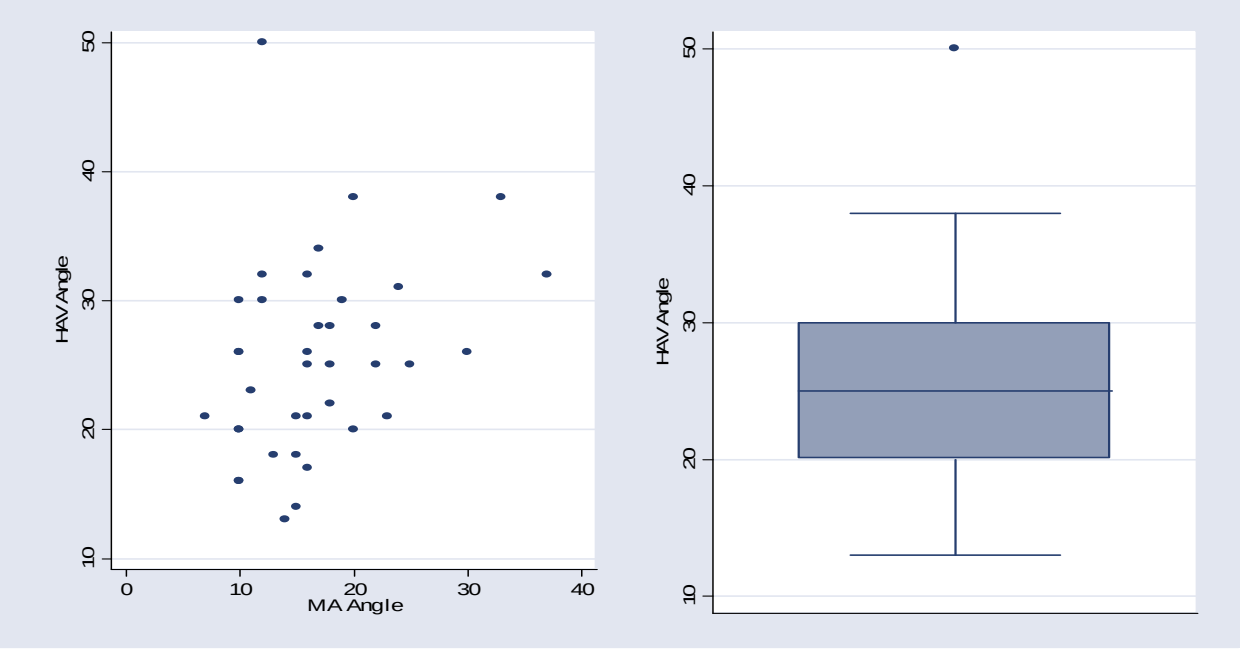
2.4 (2pts)

(a) Stock returns range from -30% to 50%. Treasury bills range from 1% to 15%. **(1pt for both)**

(b) There seems to be no apparent pattern. Based on the data, we cannot say high interest rates are bad for stocks. The relationship is weak. **(1pt for description)**

2.8 (3pts)

(a) MA is the explanatory variable as it is used to predict HAV. Thus, the x-axis of the scatterplot needs to be MA, as shown below. **(1pt for the right scatterplot)**



(b) The relationship is linear, moderate, and positive. There is one suspected outlier, as shown in the boxplot of HAV (the boxplot is not required). **(1pt for description)**

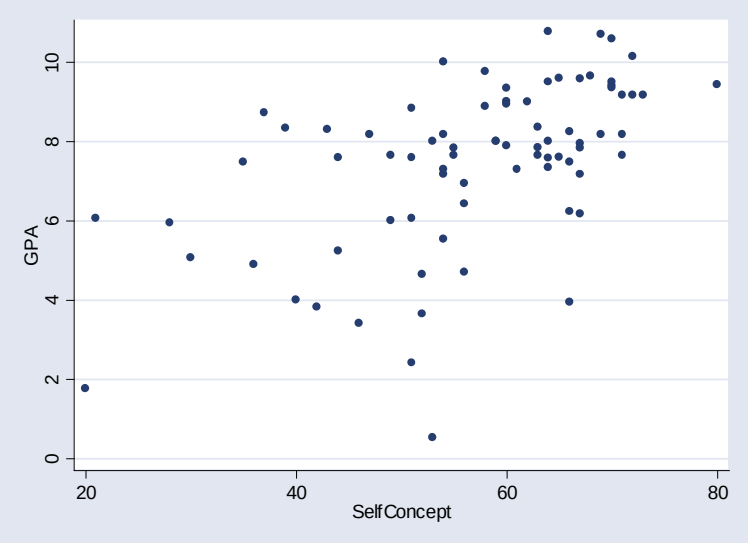
(c) The data seem to suggest predicting HAV with MA, but the regression won't be very precise. **(1pt for any reasonable thought)**

2.30 (3pts)

The scatterplot is shown below. **(1pt for the scatterplot)**

Overall pattern is a weak positive association. **(1pt for description)**

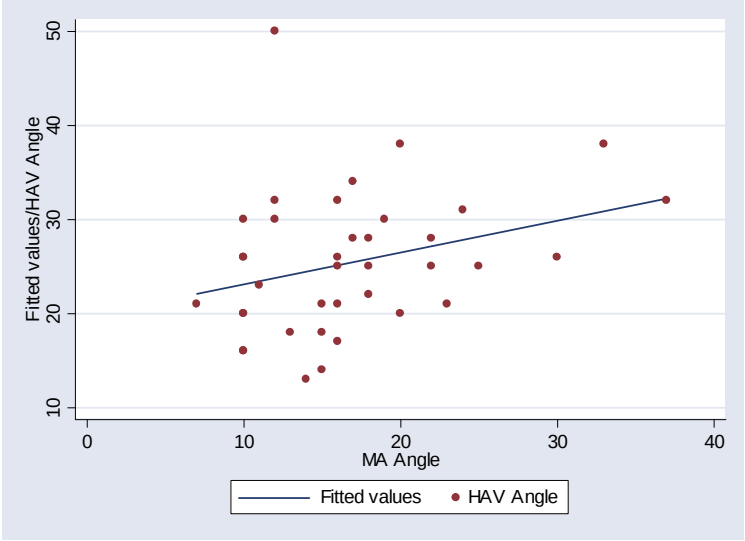
Correlation is 0.5418, so there is a slight tendency for large GPA to be associated with high self esteem and vice versa. **(1pt for a right numerical summary and its interpretation)**



2.42 (3pts)

(a) $\hat{HAV} = 19.72327 + 0.3388354 \times MA$ **(1pt for the regression equation)**

```
------------------------------------------------------------------------------
       HAV |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
        MA |   .3388354   .1781753     1.90   0.065    -.0225208    .7001916
     _cons |   19.72327   3.217168     6.13   0.000     13.19855    26.24799
------------------------------------------------------------------------------
```

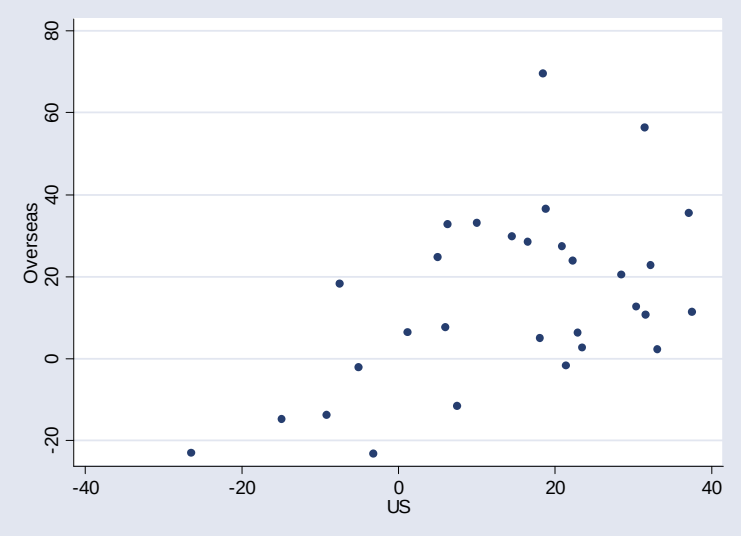


(b) Predicted value with MA angle 25 degrees = 19.72327 + 0.3388354 x 25 = 28.194155. **(1pt for the predicted value)**

(c) The prediction is not accurate because the correlation is low, i.e., 0.3021. **(1pt for description)**

2.46 (3pts)

(a) The scatterplot is shown below and US returns should be x-axis. **(1pt for the scatterplot)**



(b) Correlation, r = 0.5034 and r^2 = 0.2534. **(1pt for the right values)**

There is a weak positive association between US and overseas returns. US returns can explain about 25% of variation in overseas returns by the regression. **(1pt for the interpretation)**

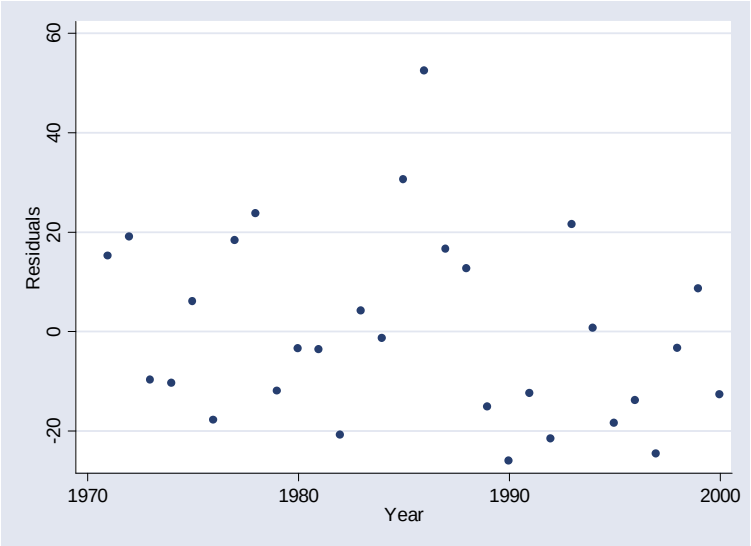
2.66 (3pts)

(a) The largest residual is in 1986. **(1pt for detecting 1986)**

Including this year, the regression equation is Overseas = 4.758 + 0.6628 x US.

Excluding this year, the regression equation is Overseas = 3.369 + 0.6337 x US.

Because there a big change in the equation, 1986 must have been influential. **(1pt for the appropriate reasoning to determine the influential point)**



(b) The plot of residuals against year is above. We cannot find any suspicious pattern. **(1pt for the residual plot and description)**

2.78 (2pts)

• How higher income can cause better health: higher income can give better nutrition and better working conditions, which causes better health. **(1pt for this)**

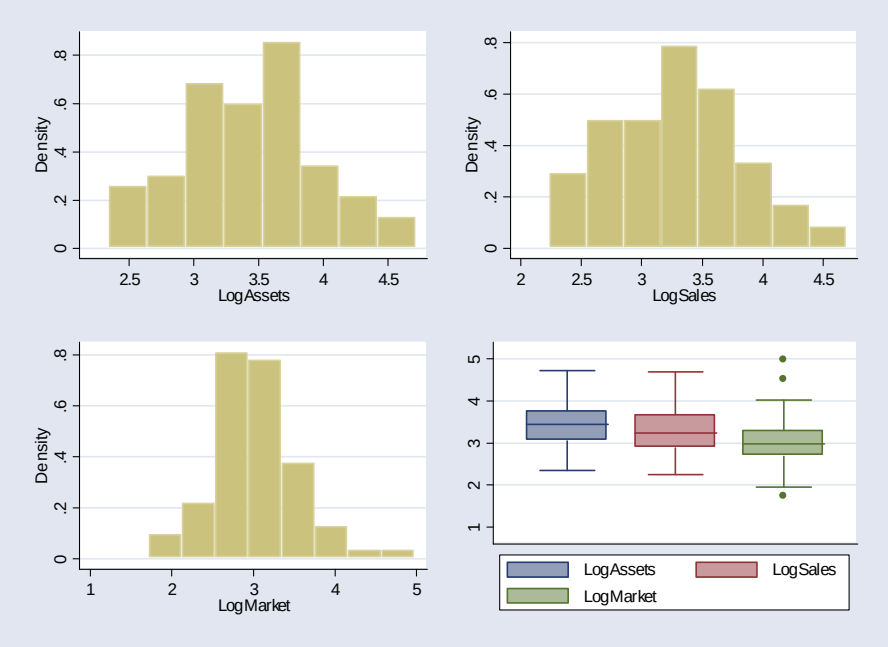
• How better health can cause higher income: better health reduces the chance of losing income due to sickness and can make people more productive, which causes higher income. **(1pt for this)**

2.86 (1pt)

A better explanation is that people who are heavier use artificial sweeteners in order to lose weight. **(1pt for the appropriate reasoning)**

Computer problems (20pts total)

1. (5pts) Histograms and boxplots are shown below. **(1pt for histograms; 1pt for boxplots)** These distributions look fairly symmetric and unimodal. **(1pt for description)**



LogAssets

	Percentiles	Smallest		
1%	2.348305	2.348305		
5%	2.514548	2.401401		
10%	2.794488	2.444045	Obs	79
25%	3.048053	2.514548	Sum of Wgt.	79
50%	3.445293		Mean	3.454578
		Largest	Std. Dev.	.5270205
75%	3.766041	4.42213		
90%	4.134209	4.523825	Variance	.2777506
95%	4.42213	4.650657	Skewness	.1373513
99%	4.721266	4.721266	Kurtosis	2.725436

LogSales

	Percentiles	Smallest		
1%	2.245513	2.245513		
5%	2.432969	2.311754		
10%	2.564666	2.421604	Obs	79
25%	2.874482	2.432969	Sum of Wgt.	79
50%	3.24403		Mean	3.298177
		Largest	Std. Dev.	.5227886
75%	3.679519	4.209435		
90%	3.958277	4.233605	Variance	.273308

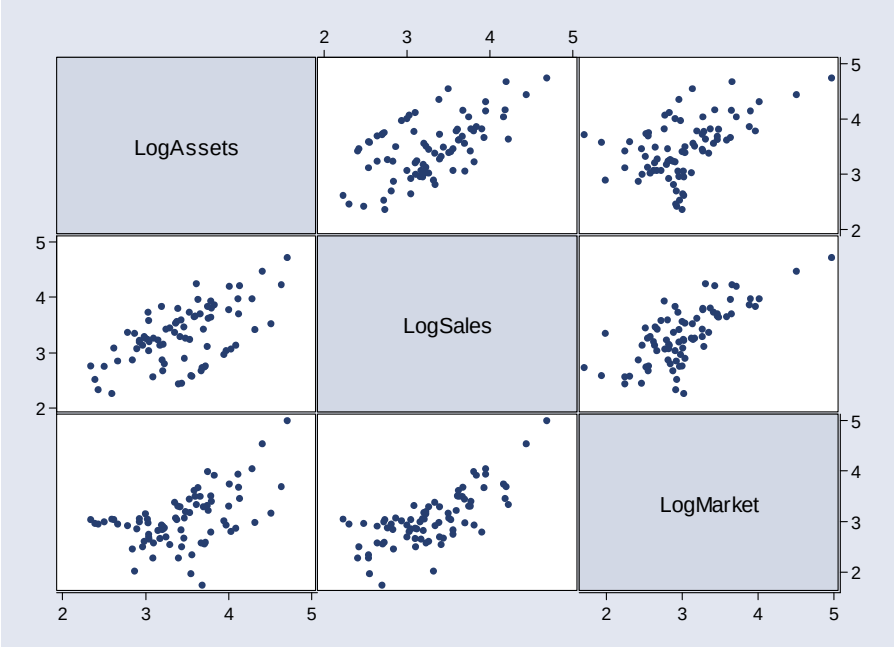
```
95%    4.209435      4.451556      Skewness       .204422
99%    4.699456      4.699456      Kurtosis      2.672689

                      LogMarket
-------------------------------------------------------------
     Percentiles     Smallest
 1%    1.724276      1.724276
 5%    2.257679      1.954242
10%    2.482874      2.004321      Obs                79
25%    2.683947      2.257679      Sum of Wgt.         79

50%    2.974972                    Mean          3.032384
                    Largest       Std. Dev.      .5343874
75%    3.301464      3.975983
90%    3.667733      4.026778      Variance       .2855699
95%    3.975983      4.520772      Skewness       .7016034
99%    4.980898      4.980898      Kurtosis       4.839255
```

From the output, the means and medians are very similar, which implies that these distributions are fairly symmetric. The log transformation seems to solve the skewness problem. **(1pt for outputs of the means and medians; 1pt for description)**

2. (3pts) The correlation matrix shows that LogMarket and LogSales have the strongest positive relationship and then LogAssets and LogSales have the second strongest positive relationship. **(1pt for the correlation matrix; 1pt for description)**

```
(obs=79)
             | LogAssets LogSales LogMarket
-------------+------------------------------
   LogAssets |   1.0000
    LogSales |   0.5823   1.0000
   LogMarket |   0.4999   0.7270   1.0000
```

The scatterplot matrix below confirms this relationship. That is, these three variables are positively associated with each other. **(1pt for the scatterplot)**



3. (1pt) The scatterplots are the same as the scatterplot matrix above. The scatterplots say that the relationship between each pair of three variables is linear and positive. **(1pt for description)**

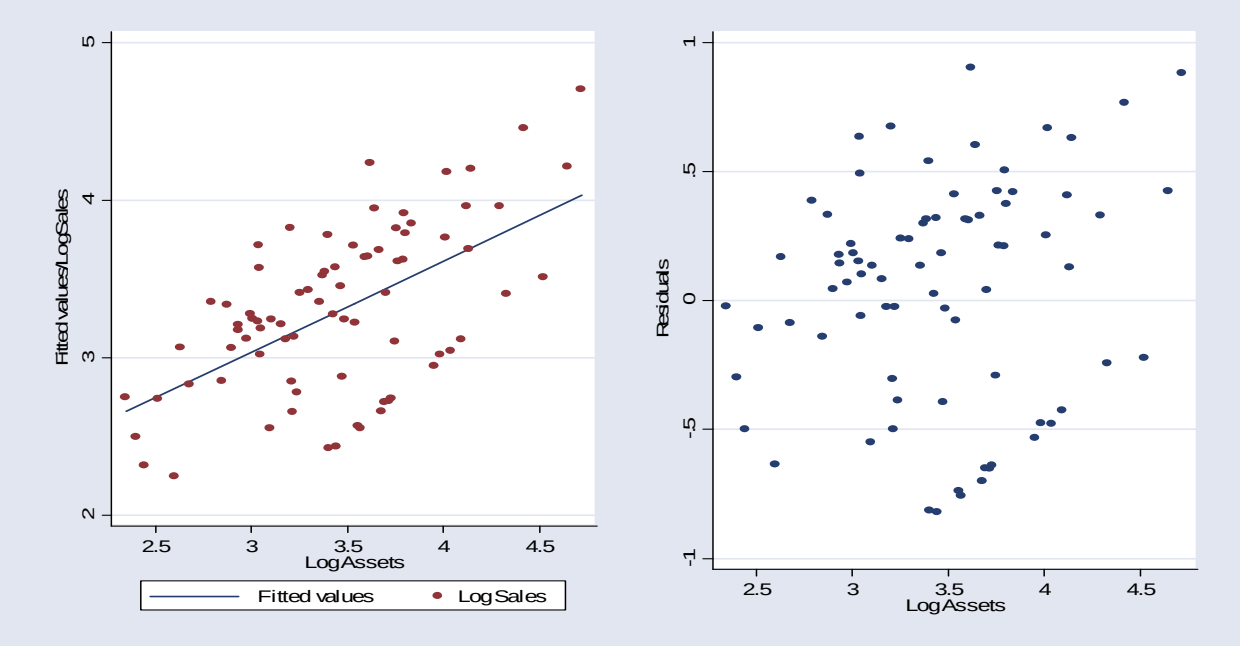
4. (1pt) Yes, LogSales and LogAssets appear to be linearly and positively related. **(1pt for description)**

5. (1pt) Regression equation: LogSales = 1.302895 + 0.577576 * LogAssets **(1pt for the equation)**

```
      Source |       SS       df       MS              Number of obs =      79
-------------+------------------------------           F(  1,    77) =   39.49
       Model |  7.22716389     1  7.22716389           Prob > F      =  0.0000
    Residual |  14.0908578    77  .182998154           R-squared     =  0.3390
-------------+------------------------------           Adj R-squared =  0.3304
       Total |  21.3180217    78  .273307971           Root MSE      =  .42778

------------------------------------------------------------------------------
    LogSales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   LogAssets |    .577576    .091907     6.28   0.000     .3945658    .7605862
       _cons |   1.302895   .3211271     4.06   0.000     .6634498    1.942341
------------------------------------------------------------------------------
```

6. (2pts) On average, LogSales is increased by 0.577576 with a unit increase of LogAssets. This is expected because a company with more assets tends to have more sales. **(1pt for interpretation of the slope; 1pt for the explanation)**

7. (1pt) The scatterplot with the regression line is presented below. **(1pt for the scatterplot)**



8. (2pts)

(1) LogAssets is 2.5 : LogSales = 1.302895 + 0.577576 * 2.5 = 2.746835

(2) LogAssets is 3.5 : LogSales = 1.302895 + 0.577576 * 3.5 = 3.324411

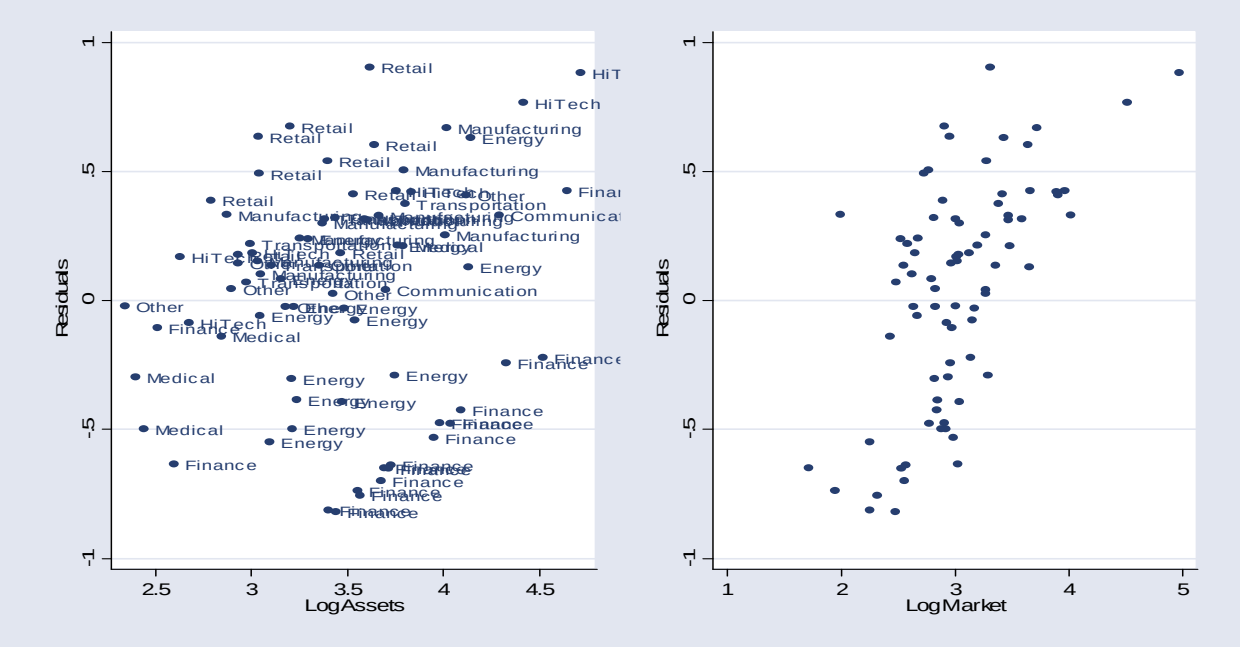
(3) LogAssets is 4.8 : LogSales = 1.302895 + 0.577576 * 4.8 = 4.075260

(1pt for the predictions)

When LogAssets is 4.8, the value of LogSales is the most unreliable because we have few observations corresponding to LogAssets of 4.8 in the data. **(1pt for detecting LogAssets of 4.8 is the most unreliable)**

9. (1pt) The residual plot is presented above. **(1pt for the residual plot)**

10. (1pt) The scatterplot with the sector information is presented below. **(1pt for the scatterplot)**



11. (1pt) The scatterplot for residuals against LogMarket is presented above. This shows that the residuals have a systematic trend according to LogMarket, which implies it is not a good idea to use only LogAssets to describe the LogSales data. To remove this trend, we may need to include LogMarket in the regression. **(1pt for the residual plot and interpretation)**

12. (1pt) Both Profit and Cash in the data set take some negative values. Since the log function is not defined for negative numbers, LogProfit and LogCash will contain missing values for these cases. When running analyses using these variables, you will get misleading answers since they only look at a subset of the data. **(1pt)**

Challenge Problem (2pts)

1. The residuals can be written as the followings. **(1pt for writing out the residuals)**

$$Residuals^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_1 x_i)^2$$

2. The derivative of the sum with respect to b_1 is as follows.

$$\frac{\partial \text{Re}\, siduals^2}{\partial b_1} = -2\sum_{i=1}^{n}\left(y_i - b_1 x_i\right)x_i = -2\sum_{i=1}^{n} y_i x_i + 2b_1\sum_{i=1}^{n} x_i^2 = 0$$

3. When solving the equation, you would get the following least squares coefficient. **(1pt for taking a derivative and solving it for b_1)**

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$