**Part A. ChemReaction Dataset (35 pts). CSC 423 students only.**

**1. Write down the matrix form of the regression equation, after plugging the actual data values for the problem.**





**2. Obtain the matrix normal equation (use the xpx option in the SAS model statement).**





**3. Obtain the matrix equation β^ = (XT X)-1 XT y (use the SAS option i).**





**4. Obtain the vector of parameter estimates.  Show your work on this question.  Multiply the matrix (XT X)-1 by the matrix XT y to obtain the result.**

**5. Obtain the vector of predicted values (use the SAS option p).**

Dependent Predicted Std Error Std Error Student Cook's

Obs Variable Value Mean Predict Residual Residual Residual -2-1 0 1 2 D

1 5.7500 5.3202 0.2745 0.4298 0.356 1.208 | |\*\* | 0.290

2 4.7900 5.3202 0.2745 -0.5302 0.356 -1.490 | \*\*| | 0.441

3 5.4400 5.3198 0.2341 0.1202 0.384 0.314 | | | 0.012

4 9.0900 8.8406 0.3177 0.2494 0.318 0.785 | |\* | 0.205

5 8.5900 8.8406 0.3177 -0.2506 0.318 -0.789 | \*| | 0.207

6 5.0900 5.1088 0.4453 -0.0188 0.0600 -0.314 | | | 1.806

Sum of Residuals 0

Sum of Squared Residuals 0.60564

Predicted Residual SS (PRESS) 2.82641

**6. Obtain the residual vector (use the SAS option r).**

Dependent Predicted Std Error Std Error Student Cook's

Obs Variable Value Mean Predict Residual Residual Residual -2-1 0 1 2 D

1 5.7500 5.3202 0.2745 0.4298 0.356 1.208 | |\*\* | 0.290

2 4.7900 5.3202 0.2745 -0.5302 0.356 -1.490 | \*\*| | 0.441

3 5.4400 5.3198 0.2341 0.1202 0.384 0.314 | | | 0.012

4 9.0900 8.8406 0.3177 0.2494 0.318 0.785 | |\* | 0.205

5 8.5900 8.8406 0.3177 -0.2506 0.318 -0.789 | \*| | 0.207

6 5.0900 5.1088 0.4453 -0.0188 0.0600 -0.314 | | | 1.806

Sum of Residuals 0

Sum of Squared Residuals 0.60564

Predicted Residual SS (PRESS) 2.82641

**7. If you try a new experiment with the amount 1.00 of monomer and 0.8 of dimer, what is the predicted reaction rate. Show your work.**

Parameter Estimates

Parameter Standard

Variable DF Estimate Error t Value Pr > |t|

Intercept 1 -1.72310 0.90919 -1.90 0.1544

monomer 1 1.49069 0.33460 4.46 0.0210

dimer 1 8.95400 1.05228 8.51 0.0034

If monomer= 1.00 and dimer= 0.8, then

Rate=-1.72310+1.49069\*1.00 + 8.95400\*0.8=6.93079

### Part B. Banking Dataset (45 pts.)

**1. Create and print a SAS dataset or R dataframe named Banking.**

**PARTLY Show:**

predict the average bank balance as a function of the other variables 5

00:28 Thursday, February 23, 2012

Home

Obs Age Education Income Val Wealth Balance

1 35.9 14.8 91033 183104 220741 38517

2 37.7 13.8 86748 163843 223152 40618

3 36.8 13.8 72245 142732 176926 35206

4 35.3 13.2 70639 145024 166260 33434

5 35.3 13.2 64879 135951 148868 28162

6 34.8 13.7 75591 155334 188310 36708

7 39.3 14.4 80615 181265 201743 38766

8 36.6 13.9 76507 149880 189727 34811

9 35.7 16.1 107935 276139 211085 41032

10 40.5 15.1 82557 182088 220782 41742

11 37.9 14.2 58294 123500 132432 29950

12 43.1 15.8 88041 194369 267556 51107

13 37.7 12.9 64597 119305 186156 34936

14 36.0 13.1 64894 141011 160017 32387

15 40.4 16.1 61091 194928 113559 32150

16 33.8 13.6 76771 159531 197264 37996

17 36.4 13.5 55609 123085 105582 24672

18 37.7 12.8 74091 143750 217869 37603

19 36.2 12.9 53713 112649 117441 26785

20 39.1 12.7 60262 126928 161322 32576

21 39.4 16.1 111548 230893 331009 56569

22 36.1 12.8 48600 105737 106671 26144

23 35.3 12.7 51419 104149 111168 24558

24 37.5 12.8 51182 106898 88370 23584

25 34.4 12.8 60753 95869 143115 26773

26 33.7 13.8 64601 103737 134223 27877

27 40.4 13.2 62164 114257 144038 28507

28 38.9 12.7 46607 94576 114799 27096

29 34.3 12.7 61446 122619 161538 28018

30 38.7 12.8 62024 134430 149351 31283

31 33.4 12.6 54986 105647 126929 24671

32 35.0 12.0 48182 114436 102732 25280

33 38.1 12.7 47388 92820 118016 24890

34 34.9 12.5 55273 102468 126959 26114

35 36.1 12.9 53892 92968 129176 27570

36 32.7 12.6 47923 104539 88384 20826

37 37.1 12.5 46176 92654 101964 23858

38 23.5 13.6 33088 105430 44223 20834

39 38.0 13.6 53890 108446 95013 26542

40 33.6 12.7 57390 111836 134434 27396

41 41.7 13.0 48439 100788 124474 31054

42 36.6 14.1 56803 149138 101695 29198

43 34.9 12.4 52392 93875 133101 24650

44 36.7 12.8 48631 95490 105202 23610

45 38.4 12.5 52500 105377 139199 29706

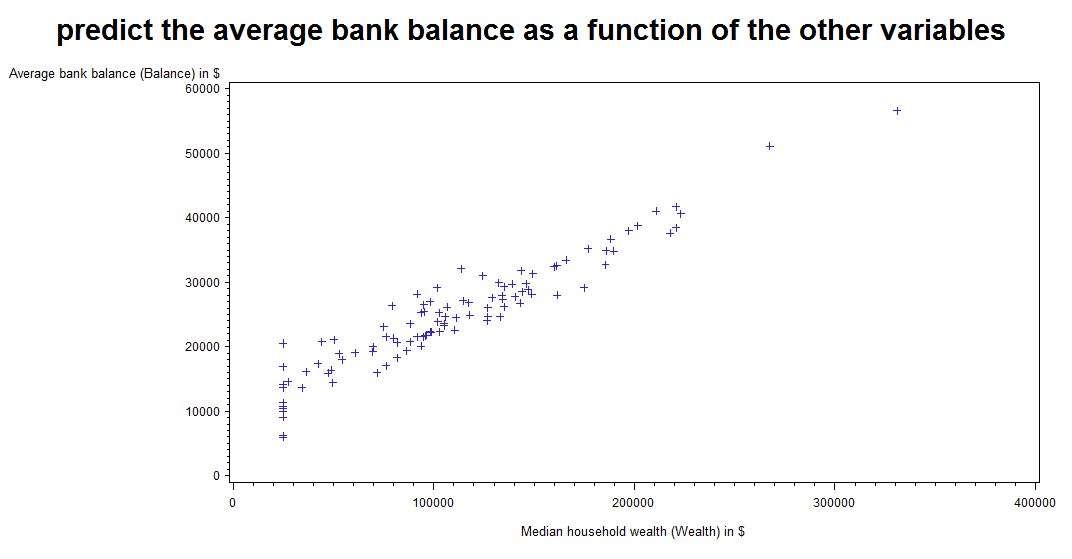
46 34.8 12.5 42401 106478 94867 21572

47 33.6 12.7 64792 116071 185714 32677

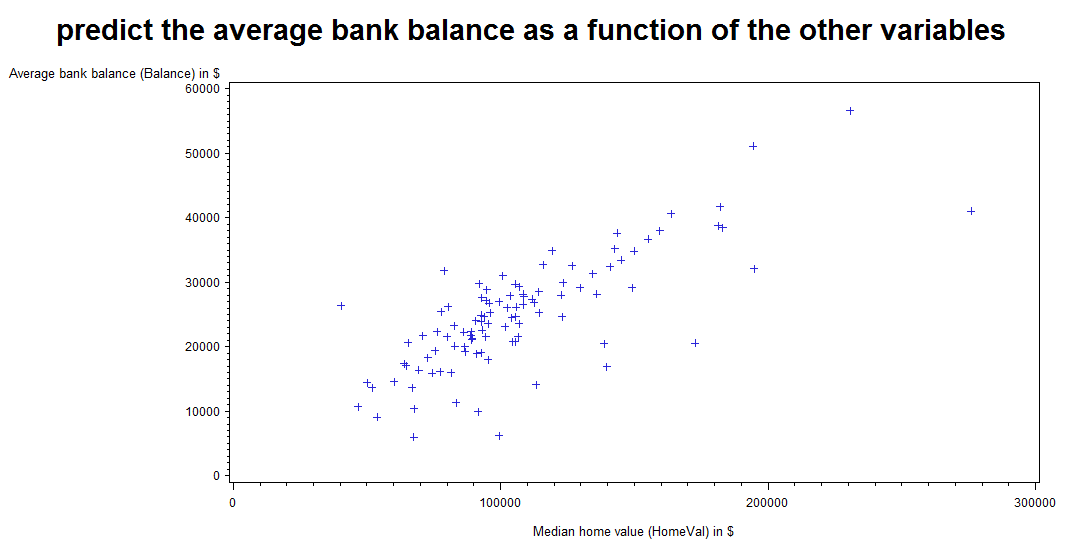
48 37.0 14.1 59842 106949 135329 29347

49 34.4 12.7 65625 129688 175000 29127

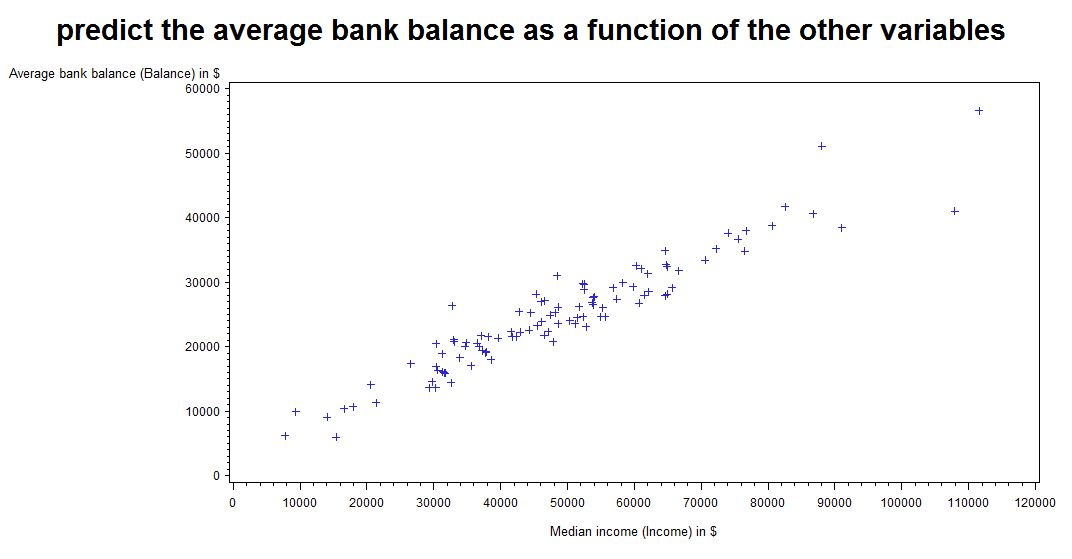
50 37.2 12.5 54044 108654 140726 27753

**2. Create scatterplots to visualize the associations between bank balance and the other five variables. Do the associations appear to be linear? **

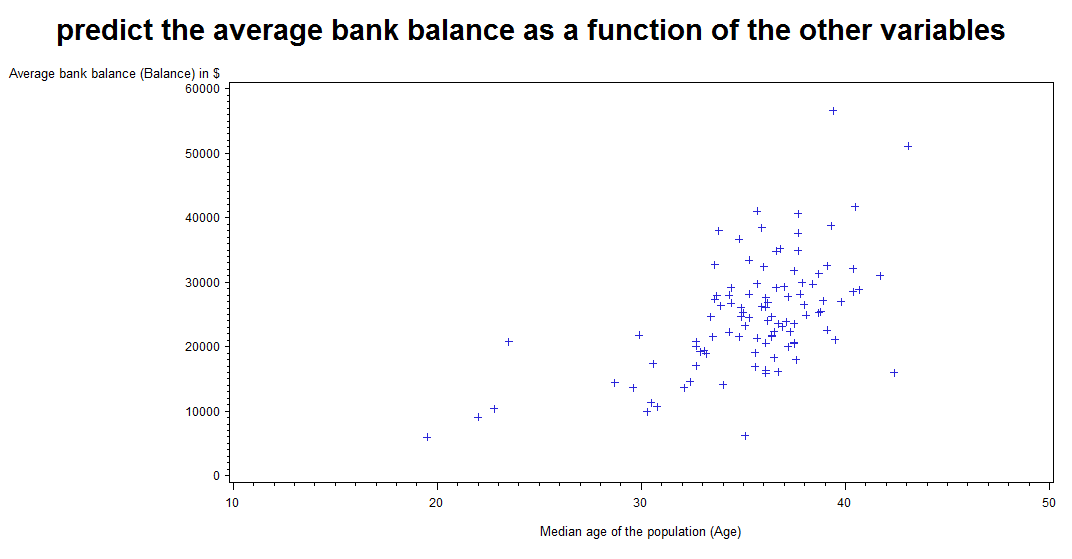
They are linear!

****

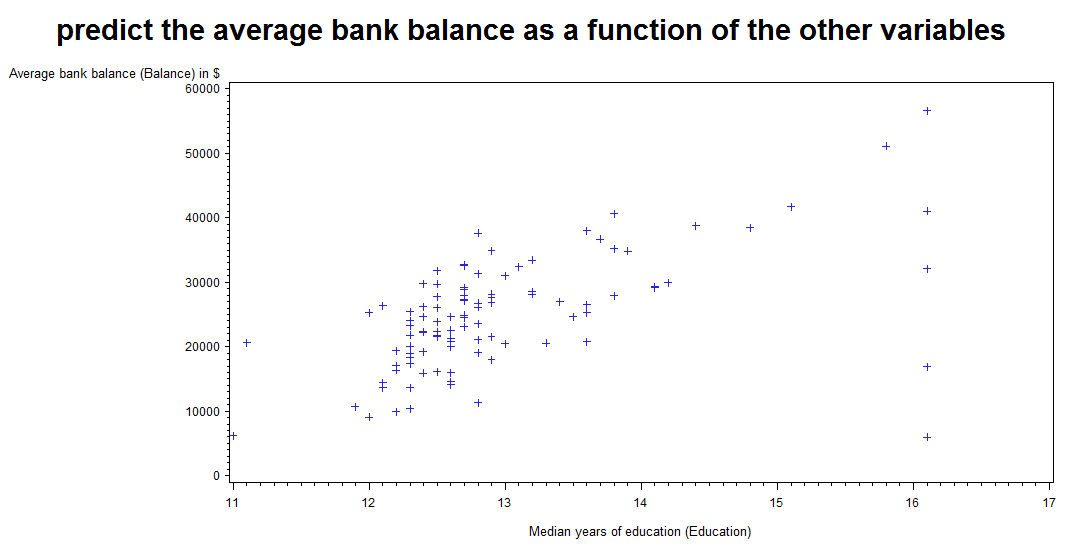
They are almost linear

****

They are almost linear

****

They are not linear

****

They are not linear

**3. Compute correlation values of bank balance vs the other variables. Interpret the correlation values. Which variables appear to be strongly associated.**

Correlations:

balance\*age 0.56547

balance\*education 0.55219

balance\*income 0.95168

balance\*Homeval 0.76639

balance\*Wealth 0.94871

The correlation measures the relative strength of a linear relationship between two numerical variables.

The correlations for balance\*income, balance\*Homeval and balance\*Wealth are much stronger than correlations for balance\*age and balance\*education.

**4. Fit a regression model of balance vs the other five variables. Write the expression of the estimated regression model.**

The equation is

Balance= -10331 +317.46 Age +590.28 Education +0.1468 Income +0.0099 HomeVal +0.0741 Wealth

**5. Are there any influence points for this model?**

Absolute values of rstudent for observations 21,59, 77 are less than 2, so they could be counted as wild or marginal influence points. Observations 9, 85, 102 are influence points.

Also we could further check DFFITS DFBETAS, which could be found in output.

Obs Residual RStudent H Ratio DFFITS

9 -3697 -2.2176 0.3174 1.1524 -1.5121

21 1690 0.9196 0.2048 1.2697 0.4666

59 -1318 -0.7277 0.2294 1.3365 -0.3971

77 2676 1.4655 0.2038 1.1695 0.7414

85 -4187 -2.9282 0.4796 1.2184 -2.8110

102 -5068 -2.9570 0.2510 0.8382 -1.7116

**6. Which of the five predictors have a significant effect on balance? (α=.05)**

Age, income and wealth have significant effects on balance, according to the following stat.

Parameter Standard

Variable Label DF Estimate Error t Value Pr > |t|

Intercept Intercept 1 -10331 4219.45906 -2.45 0.0162

Age Median age of the population 1 317.45845 61.03733 5.20 <.0001

(Age)

Education Median years of education 1 590.28154 315.12121 1.87 0.0641

(Education)

Income Median income (Income) in $ 1 0.14684 0.04083 3.60 0.0005

HomeVal Median home value (HomeVal) in 1 0.00986 0.01099 0.90 0.3716

$

Wealth Median household wealth 1 0.07414 0.01120 6.62 <.0001

(Wealth) in $

**7. A good model should only contain significant independent variables, so remove the variable with the largest p-value (>0.05) and refit the regression model of balance vs the remaining four predictors. Write down the expression of the new regression model.    Do NOT consider dropping more than one insignificant variables at one time, but rather remove one variable at a time. In fact, when one variable is removed from a regression model, it often happens that non-significant variables in the original model become significant in the reduced model.**

Variance

Variable Label DF Tolerance Inflation

Intercept Intercept 1 . 0

Age Median age of the population 1 0.74567 1.34108

(Age)

Education Median years of education 1 0.41467 2.41157

(Education)

Income Median income (Income) in $ 1 0.06714 14.89517

HomeVal Median home value (HomeVal) in 1 0.23089 4.33107

The new equation is

Balance = -12140 +324.2 Age +749.8 Education +0.1615 Income +0.0726 Wealth

**8. Analyze if all four predictors have a significant association with balance? (α=.05)   If not continue to remove one insignificant variable at a time until all of the remaining predictors are significant.**The left four predictors do have a significant association with balance.

Variance

Variable Label DF Tolerance Inflation

Intercept Intercept 1 . 0

Age Median age of the population 1 0.74567 1.34108

(Age)

Education Median years of education 1 0.41467 2.41157

(Education)

Income Median income (Income) in $ 1 0.06714 14.89517

HomeVal Median home value (HomeVal) in 1 0.23089 4.33107

**9. Interpret each of the regression coefficients for the final model.**

Balance = -12140 +324.2 Age +749.8 Education +0.1615 Income +0.0726 Wealth

Intercept is -12140.

Coefficient of age is 324.2. Age. Holding all other constant, the value of age increases 1, the value of Balance tends to increase 324.2.

Coefficient of Education is 749.8. Holding all other constant, the value of Education increases 1, the value of Balance tends to increase 749.8.

Coefficient of Income is 0.1615. Holding all other constant, the value of Income increases 1, the value of Balance tends to increase 0.1615.

Coefficient of Wealth is 0.0726. Holding all other constant, the value of Wealth increases 1, the value of Balance tends to increase 0.0726.

**10. Compute the standardized coefficients (SAS option stb in the model statement).  Discuss which variable has the strongest influence on balance?**

Variance

Variable Label DF Tolerance Inflation

Intercept Intercept 1 . 0

Age Median age of the population 1 0.74567 1.34108

(Age)

Education Median years of education 1 0.41467 2.41157

(Education)

Income Median income (Income) in $ 1 0.06714 14.89517

HomeVal Median home value (HomeVal) in 1 0.23089 4.33107

the standardized coefficients of wealth, which standardized coefficient is 0.49978, has the strongest influence on balance.

**11. Discuss the coefficient of determination, R-squared for the final model.**

Root MSE 2056.59928 R-Square 0.9463

Dependent Mean 24888 Adj R-Sq 0.9441

Coeff Var 8.26346

The r-square for final model is close to 1. Showing 94.63% of variation in the dependent variable can be explained by the variation in the independent variable.

**12. Discuss the five steps of the overall F-test for regression for the final model.**Here are the five steps of the test of hypothesis:

* State the null and alternative hypotheses:
  + - H0:β1 =β2 =β3 =β4 =0   
      H1: βj ≠ 0, for at least one value of j，j=1,2,3,4
* Compute the test statistic assuming that the null hypothesis is true:
  + - F = MSM / MSE = 1807643222 / 4229601=427.379
* Find a (1 – 0.05)100% confidence interval I for (DFM, DFE) degrees of freedom using an F-table or statistical software.   
  the confidence interval is about [0,2.46]
* F ∉[0,2.46] reject null hypothesis
* Use statistical software to determine the p-value.
* From sas, P value< 0.0001

n=102 p=5 α=0.05

DFE = n - p = 102 - 5 = 97

DFM = p - 1 = 5 - 1 = 4

SSM= 410271257

msm= 4229601

**13. Are there any influence points for your final regression model?**

Absolute values of Rstudent for observations 9,21,59 are less than 2, so they could be counted as wild or marginal influence points. Observation 85 is influence point.

Hat Diag Cov -----DFBETAS------

Obs Residual RStudent H Ratio DFFITS Intercept

Age Education Income Wealth

**9** -3243 **-1.8517 0.2568** 1.1891 -1.0886 0.0895 0.2789 -0.0975 -0.8230 0.7107

**21** 1799  **0.9786 0.2013** 1.2547 0.4912 -0.1745 -0.0487 0.2271 -0.1786 0.2757

**59** -1349 **-0.7452 0.2291** 1.3274 -0.4063 0.3306 -0.1008 -0.3620 0.0878 0.0297

**85**  -4600 **-3.0891 0.4295** 1.1493 -2.6801 0.7813 1.2677 -2.0185 1.2209 -0.9230