**CSC 423 -- Project 3**

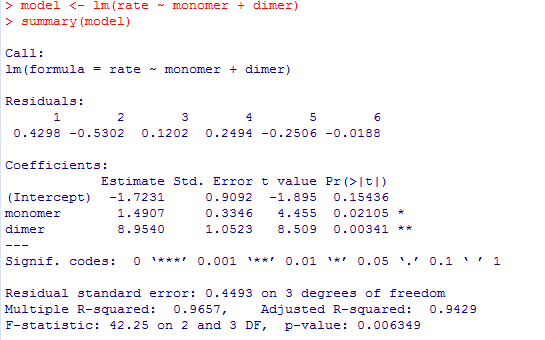
The total points for this project is listed as 145, this raw score will be multiplied by 100/145 to obtain the scaled score.

**Part A. ChemReaction Dataset (35 pts).**

* **Create a regression model for predicting the rate of a chemical reaction based on the amount of the monomer (molecule containing a single cluster of atoms) and the amount of the dimer (molecule containing two identical clusters of atoms. Here is the data:**

|  |  |  |
| --- | --- | --- |
| **monomer** | **dimer** | **rate** |
| 0.34 | 0.73 | 5.75 |
| 0.34 | 0.73 | 4.79 |
| 0.58 | 0.69 | 5.44 |
| 1.26 | 0.97 | 9.09 |
| 1.26 | 0.97 | 8.59 |
| 1.82 | 0.46 | 5.09 |

* Use SAS or R to help you do the following:  
  1. **Find the regression equation for predicting rate from monomer and dimer.**



Regression Equation: y = 1.4907X1 + 8.9540X2 – 1.7231

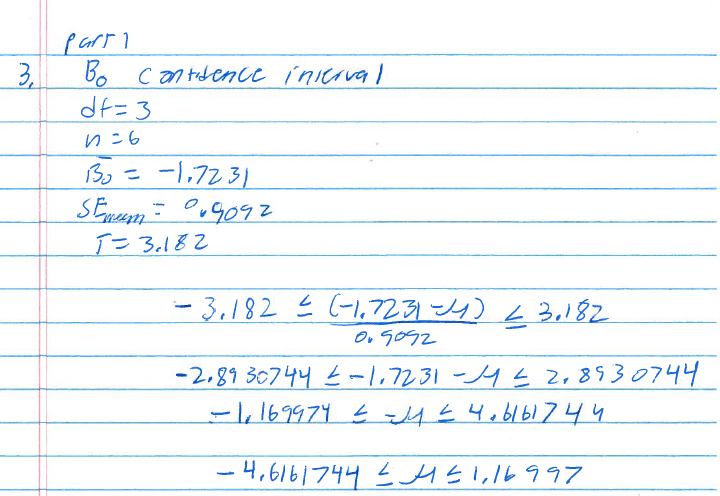
* 1. Obtain the estimated parameters from the SAS or R output.

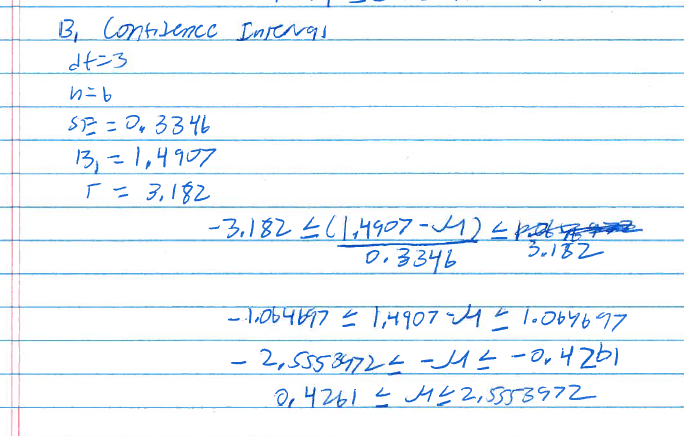
B0 = -1.7231

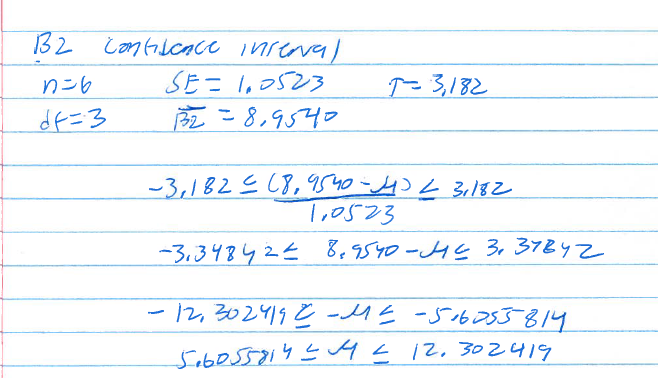
B1 = 1.4907

B2 = 8.9540

* 1. Find a 95% confidence interval for each estimated parameter compute them by hand from the estimated parameter value and standard error. Check your answer with SAS or R.







> confint(model, level = 0.95)

2.5 % 97.5 %

(Intercept) -4.6165626 1.170360

monomer 0.4258423 2.555546

dimer 5.6051844 12.302825

* 1. **Obtain the predicted values from the SAS or R output.**

 > pred <- predict(model)

> print(pred)

1 2 3 4 5 6

5.320158 5.320158 5.319764 8.840558 8.840558 5.108804

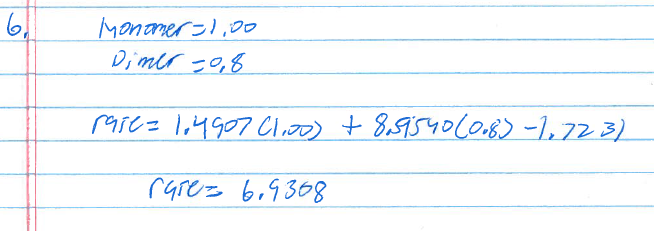
* 1. **Obtain the residuals from the SAS or R output.**

> resid <- residuals(model)

> print(resid)

1 2 3 4 5 6

0.42984207 -0.53015793 0.12023562 0.24944225 -0.25055775 -0.01880425

* 1. **If you try a new experiment with the amount 1.00 of monomer and 0.8 of dimer, what is the predicted reaction rate? Obtain the predicted reaction rate by hand and also using SAS or R.**  
      

> print(pred2)

fit lwr upr

1 5.320158 3.644532 6.995784

2 5.320158 3.644532 6.995784

3 5.319764 3.707430 6.932098

4 8.840558 7.089292 10.591823

5 8.840558 7.089292 10.591823

6 5.108804 3.095636 7.121973

7 6.930796 5.377652 8.483941

You can see that the new predicted value for the new row is 6.93

* 1. Obtain the 95% confidence and prediction intervals using SAS or R.

> predict\_interval <- predict(model2, newdata, interval="predict")

> print(predict\_interval)

fit lwr upr

1 5.320158 3.644532 6.995784

2 5.320158 3.644532 6.995784

3 5.319764 3.707430 6.932098

4 8.840558 7.089292 10.591823

5 8.840558 7.089292 10.591823

6 5.108804 3.095636 7.121973

7 6.930796 5.377652 8.483941

> confidence\_interval <- predict(model2, newdata, interval="confidence")

> print(confidence\_interval)

fit lwr upr

1 5.320158 4.446613 6.193703

2 5.320158 4.446613 6.193703

3 5.319764 4.574795 6.064734

4 8.840558 7.829475 9.851641

5 8.840558 7.829475 9.851641

6 5.108804 3.691693 6.525916

7 6.930796 6.324483 7.537110

The 95% prediction interval for the new value is (5.377652, 8.483941)

The 95% confidence interval for the new value is (6.324483, 7.537110)

**Part B. Banking Dataset (65 pts.)**

* Use the Banking Dataset [banking.txt](http://facweb.cdm.depaul.edu/sjost/csc423/projects/banking.txt) for this part. This dataset consists of data acquired from banking and census records for different zip codes in the bank's current market. Such information can be useful in targeting advertising for new customers or for choosing locations for branch offices.
* The fields in the dataset:
  1. Median age of the population (Age)
  2. Median years of education (Education)
  3. Median income (Income) in $
  4. Median home value (HomeVal) in $
  5. Median household wealth (Wealth) in $
  6. Average bank balance (Balance) in $
* **Goal:** to define a regression model to predict the average bank balance as a function of the other variables.
* **Problems**   
  1. **Create and print a SAS dataset or R dataframe named Banking.**

> banking <- read.table("c:/DataSets/banking.txt", header=T)

> print(banking)

Age Education Income HomeVal 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

51 35.7 12.6 39707 89552 80124 21345

52 37.8 12.9 45286 108431 91928 28174

53 35.6 12.8 37784 92712 60721 19125

54 35.7 12.4 52284 92143 146028 29763

55 34.3 12.4 42944 86192 98778 22275

56 39.8 13.4 46036 99508 98343 27005

57 36.2 12.3 50357 90750 126613 24076

58 35.1 12.3 45521 82720 105346 23293

59 35.6 16.1 30418 139739 24999 16854

60 40.7 12.7 52500 94792 147222 28867

61 33.5 12.5 41795 94456 91806 21556

62 37.5 12.5 66667 78906 143750 31758

63 37.6 12.9 38596 95364 54453 17939

64 39.1 12.6 44286 93103 110465 22579

65 33.1 12.2 37287 75561 86591 19343

66 36.4 12.9 38184 80099 76438 21534

67 37.3 12.5 47119 88958 102993 22357

68 38.7 13.6 44520 96112 93915 25276

69 36.9 12.7 52838 101705 75040 23077

70 32.7 12.3 34688 82870 93750 20082

71 36.1 12.4 31770 74525 47446 15912

72 39.5 12.8 32994 89223 50592 21145

73 36.5 12.3 33891 72739 81880 18340

74 32.9 12.4 37813 86667 69643 19196

75 29.9 12.3 46528 88889 96591 21798

76 32.1 12.3 30319 67083 34367 13677

77 36.1 13.3 36492 172768 24999 20572

78 35.9 12.4 51818 80357 135185 26242

79 32.7 12.2 35625 64737 76321 17077

80 37.2 12.6 36789 86563 69764 20020

81 38.8 12.3 42750 77717 95192 25385

82 37.5 13.0 30412 138911 24999 20463

83 36.4 12.5 37083 70909 95833 21670

84 42.4 12.6 31563 81597 71759 15961

85 19.5 16.1 15395 67500 24999 5956

86 30.5 12.8 21433 83456 24999 11380

87 33.2 12.3 31250 91049 52976 18959

88 36.7 12.5 31344 77541 36510 16100

89 32.4 12.6 29733 60252 27531 14620

90 36.5 12.4 41607 76270 98455 22340

91 33.9 12.1 32813 40313 79167 26405

92 29.6 12.1 29375 52096 24999 13693

93 37.5 11.1 34896 65357 81818 20586

94 34.0 12.6 20578 113239 24999 14095

95 28.7 12.1 32574 50244 49662 14393

96 36.1 12.2 30589 69375 48890 16352

97 30.6 12.3 26565 64038 42543 17410

98 22.8 12.3 16590 67850 24999 10436

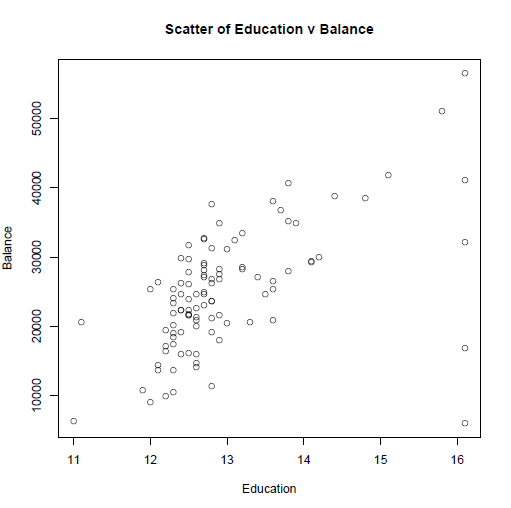
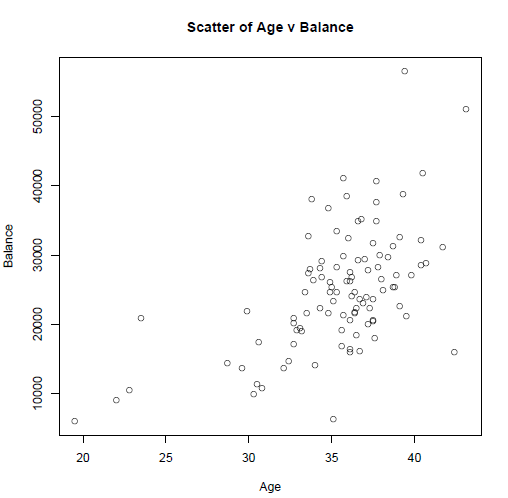
99 30.3 12.2 9354 91708 24999 9904

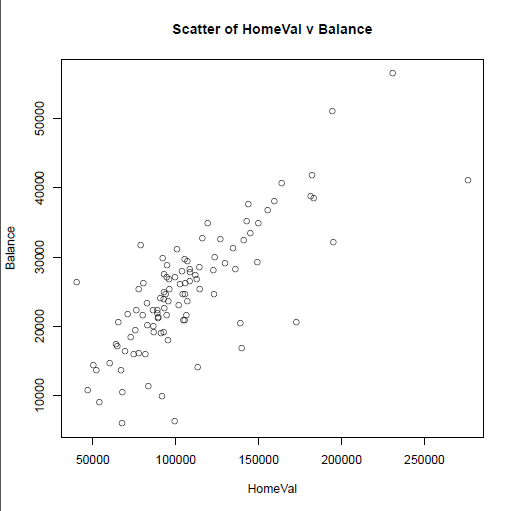
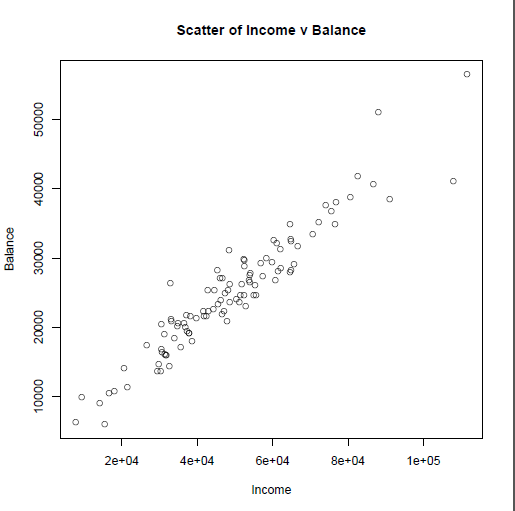
100 22.0 12.0 14115 53923 24999 9071

101 30.8 11.9 17992 46885 24999 10679

102 35.1 11.0 7741 99375 24999 6207

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







Yes, the scatterplots of the associations appear to be linear.

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

> cat("Correlation between Age and Balance:\n")

Correlation between Age and Balance:

> cor(Age,Balance)

[1] 0.5654668

> cat("Correlation between Education and Balance:\n")

Correlation between Education and Balance:

> cor(Education,Balance)

[1] 0.5521889

> cat("Correlation between Income and Balance:\n")

Correlation between Income and Balance:

> cor(Income,Balance)

[1] 0.9516845

> cat("Correlation between HomeVal and Balance:\n")

Correlation between HomeVal and Balance:

> cor(HomeVal,Balance)

[1] 0.7663871

> cat("Correlation between Wealth and Balance:\n")

Correlation between Wealth and Balance:

> cor(Wealth,Balance)

[1] 0.9487117

Wealth and Income seem to be highly correlated with bank Balance.

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

> model <- lm(Balance ~ Age + Education + Income + HomeVal + Wealth)

> summary(model)

Call:

lm(formula = Balance ~ Age + Education + Income + HomeVal + Wealth)

Residuals:

Min 1Q Median 3Q Max

-5365.5 -1102.6 -85.9 868.9 7746.5

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -10331.372191 4219.459063 -2.449 0.016160 \*

Age 317.458452 61.037332 5.201 0.00000111952 \*\*\*

Education 590.281539 315.121208 1.873 0.064085 .

Income 0.146844 0.040832 3.596 0.000512 \*\*\*

HomeVal 0.009864 0.010989 0.898 0.371591

Wealth 0.074142 0.011200 6.620 0.00000000206 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2059 on 96 degrees of freedom

Multiple R-squared: 0.9468, Adjusted R-squared: 0.944

F-statistic: 341.4 on 5 and 96 DF, p-value: < 0.00000000000000022

Balance = 317.458452X1 + 590.281539X2 + 0.146844X3 + 0.009864X4 + 0.074142X5 -10.3314372191

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

Yes, it appears that there are some influence points for this model: See this R output below. Those points marked with an asterisk \* are influence points

> cat("Influence points for Model :\n")

Influence points for Model :

> print(influence.measures(model))

Influence measures of

lm(formula = Balance ~ Age + Education + Income + HomeVal + Wealth) :

dfb.1\_ dfb.Age dfb.Edct dfb.Incm dfb.HmVl dfb.Wlth dffit

1 0.0151843 0.1414529 -0.0464916 -0.1400364 -0.004424 0.0607850 -0.400437

2 -0.0012355 0.0015630 0.0013685 -0.0017314 -0.000520 -0.0002982 -0.007442

3 -0.0041511 -0.0063264 0.0051633 0.0107664 -0.004470 -0.0002992 0.045394

4 0.0312426 -0.0251806 -0.0295793 0.0220844 0.018517 -0.0128489 0.060799

5 -0.0828113 0.0695219 0.0770940 -0.0707409 -0.049555 0.0480137 -0.187713

6 0.0457676 -0.0671179 -0.0298628 0.0196935 0.033176 0.0076882 0.133268

7 0.0076478 -0.0048982 -0.0011436 0.0043303 -0.019400 -0.0062105 -0.051307

8 0.0015126 0.0301840 -0.0072944 -0.0317639 0.006537 -0.0008200 -0.131069

9 -0.2170438 0.4271858 0.2717421 -0.6779568 -0.660407 0.7801130 -1.512134

10 -0.0011162 0.0005108 0.0009600 -0.0007000 0.000170 0.0008425 0.002097

11 -0.0203977 0.0095712 0.0195008 0.0010032 -0.011121 -0.0004855 0.025545

12 -0.5831358 0.3080778 0.5322409 -0.5929446 0.027857 0.6657018 1.015098

13 0.0117535 0.0056795 -0.0110391 -0.0578313 0.001155 0.0897629 0.127110

14 0.0312578 -0.0174203 -0.0316519 -0.0023136 0.031946 0.0082624 0.062781

15 -0.1268362 0.0711060 0.1021595 -0.0207002 0.042936 -0.0293245 0.194944

16 0.1200620 -0.1518757 -0.0801994 0.0060403 0.089262 0.0449937 0.255389

17 0.0264288 -0.0159972 -0.0139609 -0.0893710 0.016138 0.0998924 -0.139306

18 -0.0021920 0.0007363 0.0022419 0.0023183 -0.001924 -0.0035211 -0.005661

19 0.0033776 0.0007202 -0.0046456 0.0064207 0.000404 -0.0058008 0.016807

20 0.0180189 0.0253541 -0.0339681 -0.0354022 0.032379 0.0409556 0.086222

21 -0.1152193 -0.0530732 0.1421738 -0.1785651 0.061118 0.2658033 0.466636

22 0.0097462 0.0108324 -0.0145137 0.0074950 0.004516 -0.0097494 0.069757

23 -0.0128562 0.0048901 0.0124136 -0.0164494 0.002782 0.0132268 -0.036677

24 0.0001650 -0.0186250 0.0129669 -0.0485737 0.009226 0.0538728 -0.066330

25 -0.0291197 0.0552899 0.0030911 -0.0940983 0.099353 0.0349741 -0.174006

26 0.0368807 0.0368680 -0.0604378 -0.1064879 0.106459 0.0716206 -0.151995

27 0.1077451 -0.1560386 -0.0419953 -0.0816190 0.103358 0.0560599 -0.247000

28 -0.0343527 0.0682088 0.0144624 -0.0323944 -0.020591 0.0326954 0.102837

29 -0.1276694 0.0992008 0.0984515 0.0334361 -0.061419 -0.0785182 -0.206815

30 0.0073240 0.0051446 -0.0126356 0.0001094 0.010350 -0.0000552 0.022039

31 -0.0736063 0.0668623 0.0516171 -0.0399948 0.000764 0.0147063 -0.118690

32 0.1325638 -0.0338799 -0.1438874 0.0181087 0.102182 -0.0306242 0.167878

33 0.0213053 -0.0400315 -0.0118912 0.0218595 0.020715 -0.0265885 -0.073350

34 -0.0202822 0.0116587 0.0177669 -0.0167266 0.004651 0.0082464 -0.039118

35 -0.0056129 0.0024875 0.0074565 0.0066120 -0.019383 0.0012730 0.029534

36 -0.0638238 0.0539011 0.0503960 -0.0661561 -0.013198 0.0663355 -0.112364

37 -0.0001518 -0.0056672 0.0022255 -0.0018300 0.002761 0.0016513 -0.013514

38 0.4294239 -1.1150183 0.1054819 0.1129486 0.208893 -0.1833125 1.339747

39 -0.0423598 0.0321928 0.0307088 0.0558914 -0.040771 -0.0591911 0.085210

40 0.0031520 -0.0028877 -0.0022685 0.0015410 0.000318 -0.0004550 0.005079

41 -0.2041723 0.3041461 0.1087862 -0.1437715 -0.067701 0.1288094 0.375123

42 -0.0529091 0.0248197 0.0280183 0.0891069 0.062789 -0.1419824 0.240381

43 -0.0575096 0.0333471 0.0401589 0.0050713 0.024990 -0.0427809 -0.138621

44 0.0123022 -0.0230623 -0.0062621 -0.0220145 0.029094 0.0179663 -0.068295

45 0.0102773 0.0432678 -0.0273369 -0.0486273 0.012271 0.0584353 0.113623

46 -0.0297202 0.0066277 0.0286991 0.0189814 -0.033267 -0.0109362 -0.056422

47 0.0234302 -0.0247251 -0.0129748 -0.0138070 0.003186 0.0293005 0.050614

48 0.0219445 -0.0065796 -0.0233592 -0.0086949 0.022869 0.0040142 -0.030754

49 -0.1939628 0.1462753 0.1572163 0.0401055 -0.100981 -0.1078690 -0.303050

50 -0.0124209 -0.0056644 0.0156466 0.0134926 -0.007671 -0.0185394 -0.040035

51 -0.0002102 0.0040207 -0.0004393 0.0004383 -0.001446 -0.0022209 0.014298

52 -0.0536832 0.1692538 -0.0189967 0.0095349 0.045414 -0.0737361 0.296853

53 0.0033997 -0.0073709 -0.0012377 -0.0089206 0.002205 0.0149129 -0.026555

54 0.0411000 -0.0130877 -0.0232560 -0.0702942 -0.027791 0.1173311 0.183014

55 -0.0017306 0.0008904 0.0010062 0.0003075 0.000721 -0.0008257 -0.004833

56 -0.1329742 0.1358477 0.0961830 -0.0077011 -0.069533 -0.0092149 0.179307

57 -0.0417881 -0.0080494 0.0416359 0.0094331 0.028385 -0.0400250 -0.142650

58 -0.0016170 0.0002356 0.0012048 -0.0008188 0.002031 -0.0002123 -0.005815

59 0.2901771 -0.0994128 -0.2996408 0.0731563 0.013595 0.0307189 -0.397050

60 0.0320026 -0.0559720 -0.0164050 0.0385060 0.023047 -0.0480293 -0.090334

61 -0.0000562 0.0000395 0.0000362 0.0000121 -0.000021 -0.0000111 -0.000106

62 -0.0145603 0.0202426 0.0028135 0.2119530 -0.195686 -0.1386616 0.273219

63 0.0568366 -0.0980204 -0.0171202 -0.0801985 0.024822 0.1255753 -0.192684

64 0.0639509 -0.1484267 -0.0194698 0.0907627 0.021863 -0.0830893 -0.220324

65 -0.0192058 0.0141266 0.0083255 0.0101355 0.005998 -0.0143710 -0.047652

66 -0.0250655 0.0223329 0.0235696 0.0011002 -0.025485 -0.0040186 0.048521

67 0.0100059 -0.0567802 0.0096450 -0.0367767 0.051041 0.0286357 -0.131950

68 -0.0715797 0.0562443 0.0614122 -0.0045240 -0.043199 -0.0020008 0.088891

69 -0.0009937 -0.0025676 0.0037930 -0.0174366 0.004631 0.0181680 -0.019853

70 -0.0029286 0.0023774 0.0009740 0.0060129 -0.001819 -0.0061660 -0.008702

71 0.0193244 -0.0519095 -0.0057528 -0.0289144 0.028672 0.0521198 -0.121110

72 -0.0985982 0.1657993 0.0407072 -0.0005003 -0.014804 -0.0582875 0.217095

73 0.0259451 -0.0685878 -0.0207445 0.0784753 0.030001 -0.0716054 -0.163648

74 0.0062231 -0.0045322 -0.0041022 0.0033531 0.000652 -0.0042688 0.012866

75 0.0374146 -0.0419961 -0.0210363 0.0213734 -0.002336 -0.0128187 0.054842

76 -0.0208789 0.0217709 0.0067172 -0.0580149 0.030593 0.0671670 -0.107476

77 0.1846131 0.0431225 -0.2946407 0.0247076 0.577304 -0.2959065 0.741433

78 -0.0022005 -0.0011230 -0.0018236 0.0002730 0.027173 -0.0155503 -0.049533

79 -0.0305477 0.0352798 -0.0015300 -0.0092296 0.058726 -0.0072245 -0.131682

80 0.0061236 -0.0139485 -0.0020108 0.0003222 0.002946 0.0042854 -0.023639

81 -0.0442431 0.1176811 0.0062244 0.0128765 -0.073871 -0.0096189 0.179932

82 0.0654498 0.1847986 -0.1978059 -0.0617793 0.471005 -0.2108069 0.671001

83 0.0049473 -0.0059482 -0.0054580 0.0077837 0.006552 -0.0087920 -0.016840

84 0.4431935 -0.7538814 -0.2049606 0.3561091 0.061494 -0.1984422 -0.881622

85 1.1153227 1.1367669 -2.1670168 0.7464531 0.908874 -0.7708813 -2.811042

86 -0.0192772 0.0440290 -0.0151689 0.0452669 -0.039614 -0.0077001 -0.140425

87 0.0807519 -0.0289725 -0.0687897 -0.0228120 0.079120 -0.0190708 0.164938

88 0.0142079 -0.0315590 -0.0033425 -0.0252076 0.013458 0.0406272 -0.067988

89 -0.0029342 -0.0017616 0.0060205 0.0154672 -0.013793 -0.0169746 0.026431

90 0.0044133 -0.0102577 -0.0037877 0.0035210 0.013456 -0.0063153 -0.028853

91 -0.1642002 0.0528969 0.3254836 -0.0349255 -0.736039 0.2195706 1.001637

92 0.0293740 -0.0440005 -0.0064460 0.0731935 -0.045575 -0.0706594 0.112136

93 0.0320677 0.0227135 -0.0438727 -0.0077584 0.006881 0.0071798 0.072912

94 0.0073957 0.0017334 -0.0083677 -0.0164979 0.024551 0.0044939 0.031940

95 -0.0207860 0.0359343 0.0008776 -0.0316187 0.030603 0.0232046 -0.064349

96 0.0056975 -0.0277856 0.0011847 -0.0094426 0.015201 0.0199751 -0.064198

97 0.0766734 -0.1178682 0.0080724 0.0019126 -0.045739 -0.0165701 0.270060

98 0.1333299 -0.2080189 -0.0253229 -0.0699563 0.061357 0.0599106 0.259052

99 -0.0474406 0.0276644 0.0264600 0.1270102 -0.102260 -0.0894285 -0.158612

100 0.0984851 -0.1504969 -0.0181719 -0.0531496 0.029000 0.0542869 0.187744

101 -0.0174890 0.0183852 -0.0018160 0.0149493 0.012573 -0.0087962 -0.075459

102 -0.7230464 -0.1390913 0.8291359 1.1966901 -1.333473 -0.7591720 -1.711631

cov.r cook.d hat inf

1 1.010 0.02643531055 0.0725

2 1.123 0.00000932676 0.0516

3 1.087 0.00034676204 0.0254

4 1.098 0.00062192531 0.0362

5 0.999 0.00584768188 0.0244

6 1.081 0.00297803937 0.0406

7 1.115 0.00044311427 0.0478

8 1.065 0.00287763497 0.0321

9 1.152 0.36614843503 0.3174 \*

10 1.152 0.00000074092 0.0753

11 1.103 0.00010988621 0.0355

12 1.017 0.16610700201 0.1950 \*

13 1.087 0.00271078529 0.0425

14 1.084 0.00066281844 0.0267

15 1.255 0.00638728082 0.1619 \*

16 1.053 0.01085927194 0.0560

17 1.045 0.00324374311 0.0262

18 1.148 0.00000539783 0.0721

19 1.077 0.00004756143 0.0128

20 1.087 0.00124927862 0.0338

21 1.270 0.03635003666 0.2048 \*

22 1.047 0.00081572436 0.0108

23 1.073 0.00022633546 0.0139

24 1.101 0.00074015711 0.0391

25 1.047 0.00505501000 0.0350

26 1.124 0.00387853738 0.0708

27 0.983 0.01008299932 0.0326

28 1.065 0.00177361817 0.0256

29 1.020 0.00711139637 0.0335

30 1.094 0.00008179274 0.0278

31 1.051 0.00235796364 0.0233

32 1.070 0.00471553825 0.0431

33 1.073 0.00090389741 0.0222

34 1.082 0.00025752030 0.0206

35 1.084 0.00014683844 0.0205

36 1.071 0.00211755068 0.0308

37 1.081 0.00003075669 0.0155

38 0.570 0.26649838333 0.1233 \*

39 1.110 0.00122107321 0.0493

40 1.091 0.00000434388 0.0239

41 0.963 0.02309390717 0.0534

42 1.031 0.00960676288 0.0446

43 1.030 0.00320684560 0.0215

44 1.060 0.00078296289 0.0147

45 1.062 0.00216361742 0.0266

46 1.074 0.00053523418 0.0189

47 1.125 0.00043125881 0.0563

48 1.121 0.00015926805 0.0513

49 0.981 0.01514496262 0.0434

50 1.084 0.00026974213 0.0221

51 1.078 0.00003442641 0.0135

52 0.810 0.01413669911 0.0183 \*

53 1.086 0.00011872508 0.0216

54 1.033 0.00558216197 0.0323

55 1.080 0.00000393317 0.0145

56 1.058 0.00537196848 0.0406

57 1.022 0.00339206526 0.0202

58 1.083 0.00000569414 0.0173

59 1.337 0.02640421795 0.2294 \*

60 1.112 0.00137220249 0.0518

61 1.081 0.00000000189 0.0146

62 1.139 0.01248639869 0.1024

63 1.047 0.00619366296 0.0392

64 0.993 0.00804077661 0.0296

65 1.080 0.00038199606 0.0204

66 1.085 0.00039611746 0.0241

67 1.023 0.00290403362 0.0184

68 1.101 0.00132836005 0.0433

69 1.164 0.00006638064 0.0858

70 1.106 0.00001275234 0.0371

71 1.061 0.00245711870 0.0279

72 1.059 0.00786324740 0.0497

73 1.035 0.00446741105 0.0285

74 1.084 0.00002787893 0.0184

75 1.111 0.00050621761 0.0457

76 1.086 0.00193939365 0.0375

77 1.170 0.09053817059 0.2038 \*

78 1.093 0.00041288198 0.0306

79 1.044 0.00289910653 0.0242

80 1.087 0.00009408723 0.0218

81 1.031 0.00539534537 0.0311

82 0.976 0.07320612186 0.1168

83 1.102 0.00004775958 0.0339

84 0.719 0.12074323410 0.0886 \*

85 1.218 1.22068363981 0.4796 \*

86 1.081 0.00330556441 0.0423

87 1.023 0.00453201712 0.0254

88 1.102 0.00077759223 0.0403

89 1.123 0.00011764735 0.0524

90 1.086 0.00014015183 0.0216

91 0.407 0.14262619208 0.0541 \*

92 1.135 0.00211418074 0.0716

93 1.117 0.00089443959 0.0522

94 1.167 0.00017179377 0.0882

95 1.134 0.00069694602 0.0646

96 1.085 0.00069307208 0.0274

97 0.949 0.01198804928 0.0302

98 1.185 0.01124775319 0.1270

99 1.243 0.00423079212 0.1510 \*

100 1.227 0.00592337608 0.1437 \*

101 1.100 0.00095761222 0.0399

102 0.838 0.45183361819 0.2510 \*

>  
 > summary(influence.measures(model))

Potentially influential observations of

lm(formula = Balance ~ Age + Education + Income + HomeVal + Wealth) :

dfb.1\_ dfb.Age dfb.Edct dfb.Incm dfb.HmVl dfb.Wlth dffit cov.r cook.d hat

9 -0.22 0.43 0.27 -0.68 -0.66 0.78 -1.51\_\* 1.15 0.37 0.32\_\*

12 -0.58 0.31 0.53 -0.59 0.03 0.67 1.02\_\* 1.02 0.17 0.19\_\*

15 -0.13 0.07 0.10 -0.02 0.04 -0.03 0.19 1.25\_\* 0.01 0.16

21 -0.12 -0.05 0.14 -0.18 0.06 0.27 0.47 1.27\_\* 0.04 0.20\_\*

38 0.43 -1.12\_\* 0.11 0.11 0.21 -0.18 1.34\_\* 0.57\_\* 0.27 0.12

52 -0.05 0.17 -0.02 0.01 0.05 -0.07 0.30 0.81\_\* 0.01 0.02

59 0.29 -0.10 -0.30 0.07 0.01 0.03 -0.40 1.34\_\* 0.03 0.23\_\*

77 0.18 0.04 -0.29 0.02 0.58 -0.30 0.74 1.17 0.09 0.20\_\*

84 0.44 -0.75 -0.20 0.36 0.06 -0.20 -0.88\_\* 0.72\_\* 0.12 0.09

85 1.12\_\* 1.14\_\* -2.17\_\* 0.75 0.91 -0.77 -2.81\_\* 1.22\_\* 1.22\_\* 0.48\_\*

91 -0.16 0.05 0.33 -0.03 -0.74 0.22 1.00\_\* 0.41\_\* 0.14 0.05

99 -0.05 0.03 0.03 0.13 -0.10 -0.09 -0.16 1.24\_\* 0.00 0.15

100 0.10 -0.15 -0.02 -0.05 0.03 0.05 0.19 1.23\_\* 0.01 0.14

102 -0.72 -0.14 0.83 1.20\_\* -1.33\_\* -0.76 -1.71\_\* 0.84 0.45 0.25\_\*

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

> model <- lm(Balance ~ Age + Education + Income + HomeVal + Wealth)

> summary(model)

Call:

lm(formula = Balance ~ Age + Education + Income + HomeVal + Wealth)

Residuals:

Min 1Q Median 3Q Max

-5365.5 -1102.6 -85.9 868.9 7746.5

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -10331.372191 4219.459063 -2.449 0.016160 \*

Age 317.458452 61.037332 5.201 0.00000111952 \*\*\*

Education 590.281539 315.121208 1.873 0.064085 .

Income 0.146844 0.040832 3.596 0.000512 \*\*\*

HomeVal 0.009864 0.010989 0.898 0.371591

Wealth 0.074142 0.011200 6.620 0.00000000206 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2059 on 96 degrees of freedom

Multiple R-squared: 0.9468, Adjusted R-squared: 0.944

F-statistic: 341.4 on 5 and 96 DF, p-value: < 0.00000000000000022

Looking at the significant codes of the model, it looks like Age, Education, Income and Wealth have a significant effect on balance.

* 1. **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.**

> model2 <- lm(Balance ~ Age + Education + Income + Wealth)

> summary(model2)

Call:

lm(formula = Balance ~ Age + Education + Income + Wealth)

Residuals:

Min 1Q Median 3Q Max

-5403.9 -1234.1 -75.0 998.6 7430.7

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -12140.04991 3703.64135 -3.278 0.00145 \*\*

Age 324.20337 60.51247 5.358 0.00000056800 \*\*\*

Education 749.79573 259.98358 2.884 0.00484 \*\*

Income 0.16152 0.03738 4.321 0.00003753763 \*\*\*

Wealth 0.07265 0.01106 6.566 0.00000000257 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2057 on 97 degrees of freedom

Multiple R-squared: 0.9463, Adjusted R-squared: 0.9441

F-statistic: 427.4 on 4 and 97 DF, p-value: < 0.00000000000000022

New Regression Model:

Balance = 324.20337X1 + 749.79573X2 + 0.16152X3 + 0.07265X3 – 12140.04491

* 1. **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.**

> model2 <- lm(Balance ~ Age + Education + Income + Wealth)

> summary(model2)

Call:

lm(formula = Balance ~ Age + Education + Income + Wealth)

Residuals:

Min 1Q Median 3Q Max

-5403.9 -1234.1 -75.0 998.6 7430.7

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -12140.04991 3703.64135 -3.278 0.00145 \*\*

Age 324.20337 60.51247 5.358 0.00000056800 \*\*\*

Education 749.79573 259.98358 2.884 0.00484 \*\*

Income 0.16152 0.03738 4.321 0.00003753763 \*\*\*

Wealth 0.07265 0.01106 6.566 0.00000000257 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2057 on 97 degrees of freedom

Multiple R-squared: 0.9463, Adjusted R-squared: 0.9441

F-statistic: 427.4 on 4 and 97 DF, p-value: < 0.00000000000000022

Currently, it appears that all of the predictors have a significant effect on the balance, as the p-values are all fairly low, and the t-statistics are significant at the .05 level.

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

For an increase of 1 year of age, bank account balance will increase by 324.20337

For an increase of 1 year of Education, bank account balance will increase by 749.79573

For an increase of 1 dollar of education, bank account balance will increase by 0.16152

For an increase of 1 dollar of Wealth, bank account balance will increase by 0.07265

* 1. **Discuss the coefficient of determination, R-squared for the final model.**  
      The R-squared for the final model is 0.9463, and the adjusted r-squared for the final model is 0.9441. These are fairly high r-squared coefficients, and very close to 1, which means that the predictors are very good at explaining the variation in the dependent variable.
  2. **Discuss the five steps of the overall F-test for regression for the final model.**

1. State the null and alternative hypotheses:

H0: β1 = β2 = ... = βp-1 = 0

H1: βj ≠ 0, for at least one value of j

1. Compute the test statistic assuming that the null hypothesis is true:

F = MSM / MSE = (explained variance) / (unexplained variance)

(SSM / DFM) / (SSE / DFE) = MSM / MSE

F-statistic: 427.4 on 4 and 97 DF

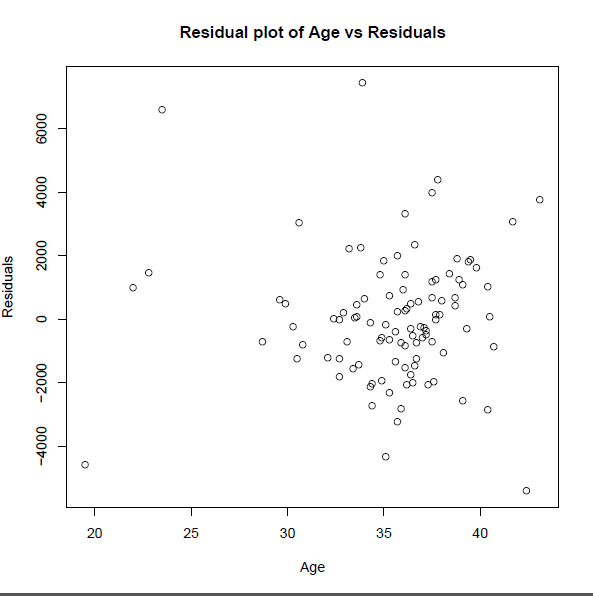
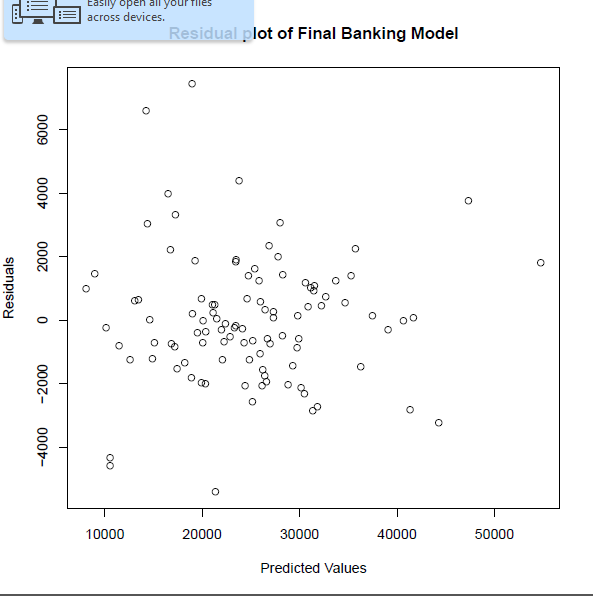
1. Find a (1 - α)100% confidence interval I for (DFM, DFE) degrees of freedom using an F-table or statistical software.

qf(0.95, 4, 97)

[1] 2.46548

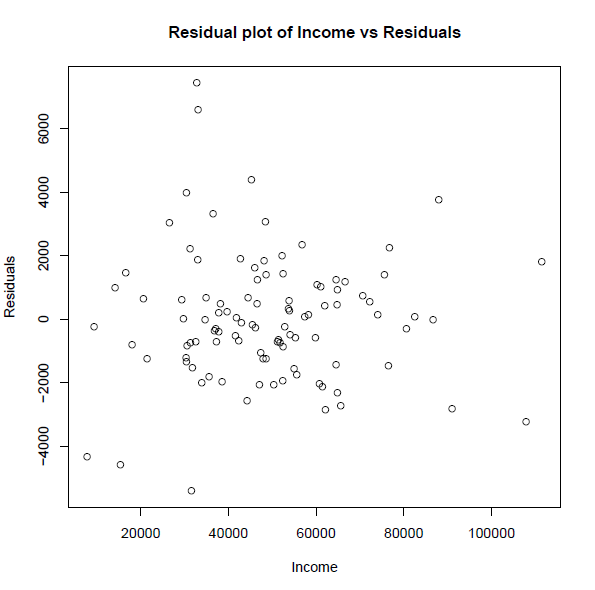
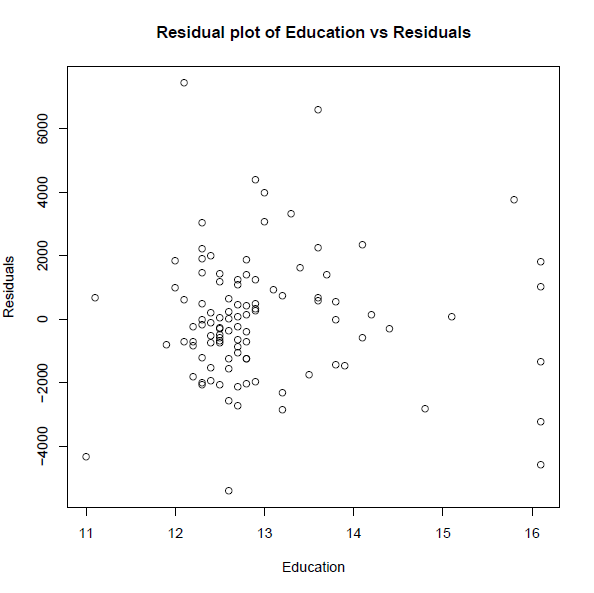
1. Accept the null hypothesis if F ∈ I; reject it if F ∉ I.

F-Statistic is not in, reject H0.

1. Use statistical software to determine the p-value.  
   p-value: < 0.00000000000000022
   1. **For the final model, create residual plots (r.\*p. and r.\*x1, ... r.\*xn, where x1, ... , xn are the independent variables) and the normal plot of the residuals. Interpret these plots.**   
       

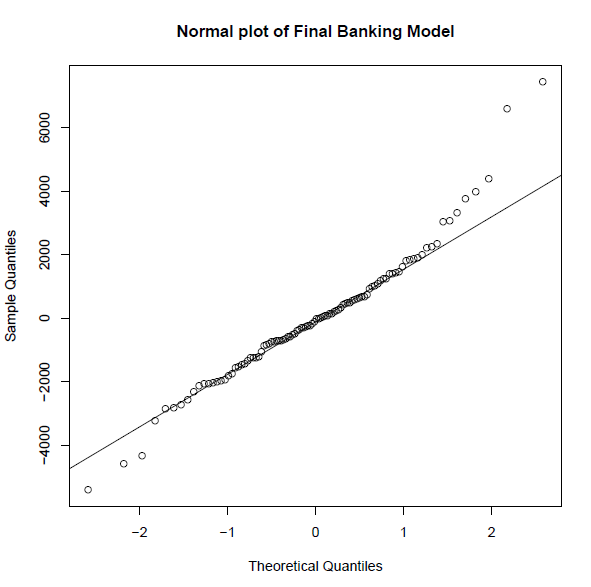
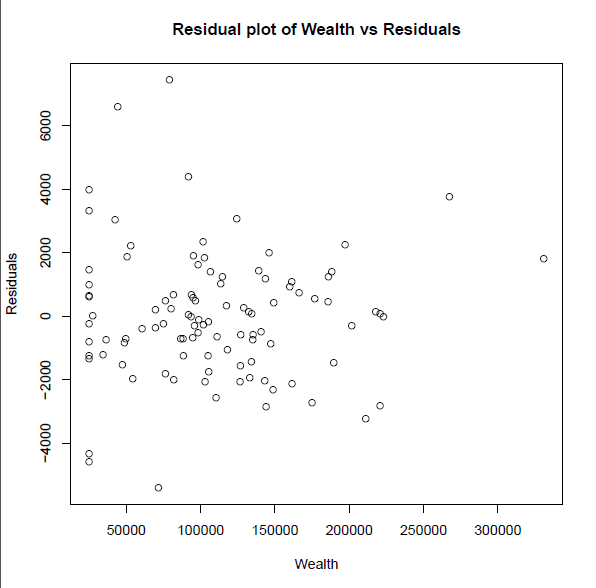
The residual plot vs the predicted values is unbiased in my opinion, though appears to be a bit heteroscedastic, with the values being located mostly on the left side of the plot.

As well, the residual plot vs the independent variable Age is unbiased and heteroscedastic.



The residual plot against the independent variable Education reveals another unbiased yet somewhat heteroscedastic plot.

However, the plot of residuals vs Income looks to be a bit more homoscedastic, and similarly unbiased.



Last, the residual plot of Wealth vs Residuals is again unbiased and heteroscedastic.

The normal plot appears to have thick tails.

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

Yes, there are influence points for my final regression model, as evidenced by the R script here:

> summary(influence.measures(model2))

Potentially influential observations of

lm(formula = Balance ~ Age + Education + Income + Wealth) :

dfb.1\_ dfb.Age dfb.Edct dfb.Incm dfb.Wlth dffit cov.r cook.d hat

9 0.09 0.28 -0.10 -0.82 0.71 -1.09\_\* 1.19\_\* 0.23 0.26\_\*

12 -0.68 0.32 0.67 -0.64 0.67 1.02\_\* 1.05 0.20 0.19\_\*

15 -0.20 0.09 0.18 0.00 -0.04 0.23 1.23\_\* 0.01 0.15\_\*

21 -0.17 -0.05 0.23 -0.18 0.28 0.49 1.25\_\* 0.05 0.20\_\*

38 0.38 -1.11\_\* 0.27 0.22 -0.22 1.34\_\* 0.63\_\* 0.32 0.12

52 -0.09 0.18 0.01 0.03 -0.08 0.30 0.84\_\* 0.02 0.02

59 0.33 -0.10 -0.36 0.09 0.03 -0.41 1.33\_\* 0.03 0.23\_\*

84 0.47 -0.76 -0.21 0.42 -0.21 -0.89\_\* 0.77\_\* 0.15 0.09

85 0.78 1.27\_\* -2.02\_\* 1.22\_\* -0.92 -2.68\_\* 1.15 1.32\_\* 0.43\_\*

91 0.20 -0.04 -0.10 -0.33 0.31 0.63 0.51\_\* 0.07 0.02

98 0.13 -0.22 0.01 -0.06 0.06 0.28 1.16\_\* 0.02 0.12

100 0.11 -0.17 0.00 -0.05 0.06 0.21 1.21\_\* 0.01 0.14

102 -0.07 -0.21 0.07 0.51 -0.40 -0.75\_\* 0.90 0.11 0.10

We can see that the dfBetas for observation 38 on age, as well as the dffits for that observation are too high.

For observation 85, the cook’s D is far too high at 1.32, and the hat-value is in the cutoff range of .2 to .5 at .43.

Those are the clearest examples of influence points listed, though there are other potentially influential points in the data set as well.

**Part C. SalarySurvey Dataset (45 pts.)**

* Use the dataset [salary-survey.txt](http://facweb.cdm.depaul.edu/sjost/csc423/projects/salary-survey.txt) to answer the following questions.   
  1. **Input the data from the salary-survey file. Create dummy variables to represent educ and mgt. You should have two dummy variables for educ and one for mgt.**

> salary <- read.table("c:/DataSets/salary-survey.txt",header=T)

> print(salary)

exper educ mgt salary dummy\_col dummy\_ad

1 1 1 1 58418 0 0

2 1 3 0 48870 0 1

3 1 3 1 78731 0 1

4 1 2 0 47501 1 0

5 1 3 0 49539 0 1

6 2 2 1 87871 1 0

7 2 2 0 49560 1 0

8 2 1 0 44352 0 0

9 2 3 0 51341 0 1

10 3 2 0 51838 1 0

11 3 1 1 63045 0 0

12 3 2 1 89972 1 0

13 3 3 1 83358 0 1

14 4 1 0 48066 0 0

15 4 3 1 85307 0 1

16 4 3 0 55703 0 1

17 4 2 0 54242 1 0

18 5 2 0 55761 1 0

19 5 3 0 57580 0 1

20 5 1 1 67213 0 0

21 6 1 0 51935 0 0

22 6 3 1 89892 0 1

23 6 2 0 58262 1 0

24 6 2 1 96342 1 0

25 7 1 1 71477 0 0

26 8 2 0 62321 1 0

27 8 1 1 73271 1 0

28 8 3 1 93395 0 1

29 8 1 0 57037 0 0

30 10 1 0 60906 0 0

31 10 2 0 67116 1 0

32 10 3 1 97563 0 1

33 10 2 1 100114 1 0

34 11 2 1 106976 1 0

35 11 1 0 62565 0 0

36 12 2 0 71073 1 0

37 12 3 1 101764 0 1

38 13 1 0 67318 0 0

39 13 2 1 110849 1 0

40 14 2 0 75565 1 0

41 15 3 1 108134 0 1

42 16 2 1 117194 1 0

43 16 2 0 79308 1 0

44 16 1 0 73603 0 0

45 17 2 0 80861 1 0

46 20 1 0 81447 0 0

> exper <- salary\_data$exper

> educ <- salary\_data$educ

> mgt <- salary\_data$mgt

> salary <- salary\_data$salary

> dummy\_col <- as.numeric(educ == 2)

> print(dummy\_col)

[1] 0 0 0 1 0 1 1 0 0 1 0 1 0 0 0 0 1 1 0 0 0 0 1 1 0 1 0 0 0 0 1 0 1 1 0 1 0 0

[39] 1 1 0 1 1 0 1 0

> dummy\_ad <- as.numeric(educ == 3)

> print(dummy\_ad)

[1] 0 1 1 0 1 0 0 0 1 0 0 0 1 0 1 1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 0 0 1 0

[39] 0 0 1 0 0 0 0 0

* 1. **Create a regression model that predicts salary from exper, educ, and mgt.**

> model <- lm(salary ~ exper + dummy\_col + dummy\_ad + mgt)

> summary(model)

Call:

lm(formula = salary ~ exper + dummy\_col + dummy\_ad + mgt)

Residuals:

Min 1Q Median 3Q Max

-7934.3 -2751.7 93.6 3556.8 7226.3

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 33829.4 1627.9 20.780 < 2e-16 \*\*\*

exper 2299.5 128.5 17.897 < 2e-16 \*\*\*

dummy\_col 13236.3 1523.9 8.686 7.73e-11 \*\*\*

dummy\_ad 12614.7 1733.5 7.277 6.71e-09 \*\*\*

mgt 28980.0 1321.6 21.928 < 2e-16 \*\*\*

---

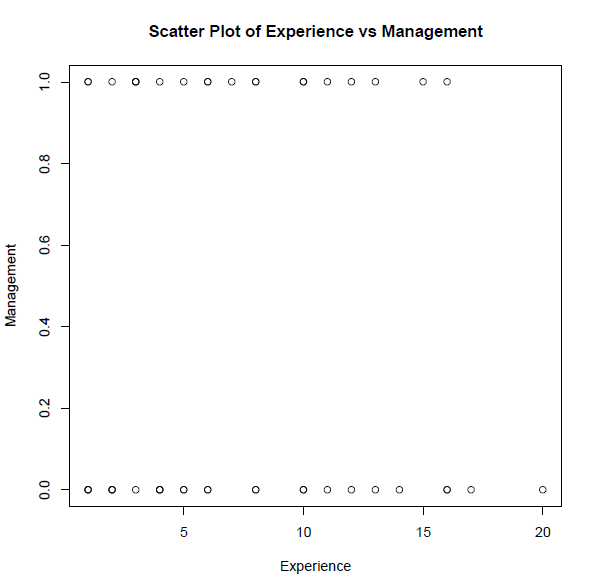
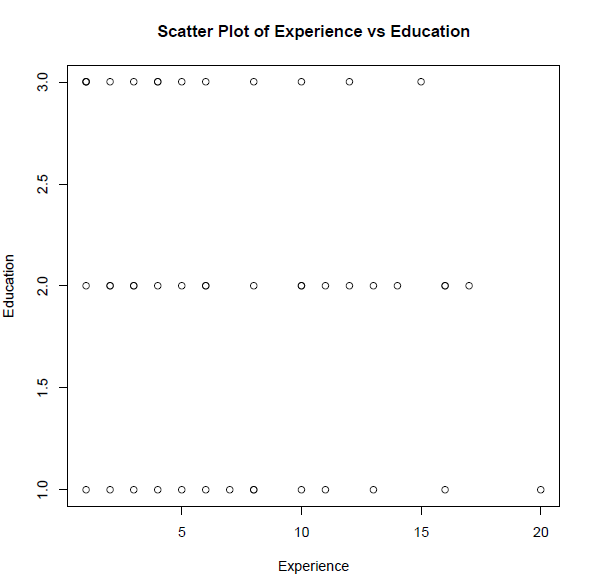
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

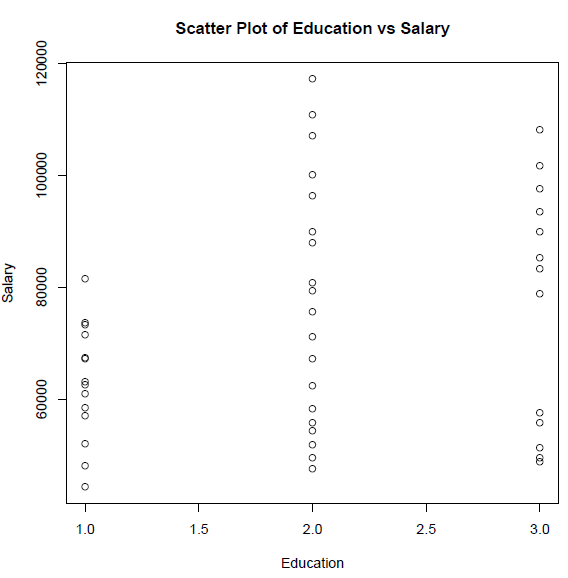
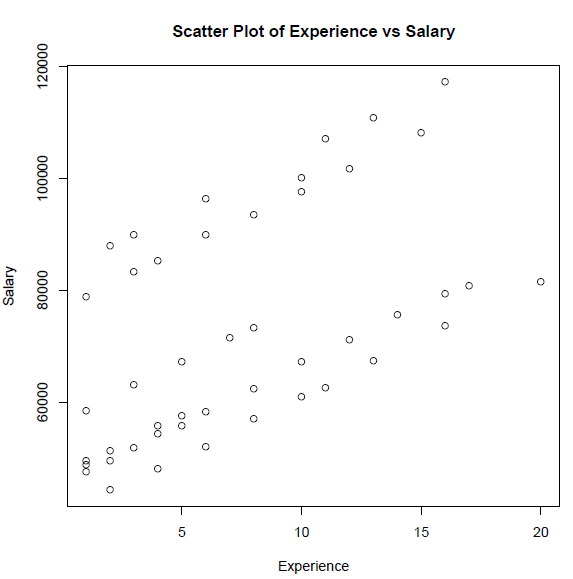
Residual standard error: 4325 on 41 degrees of freedom

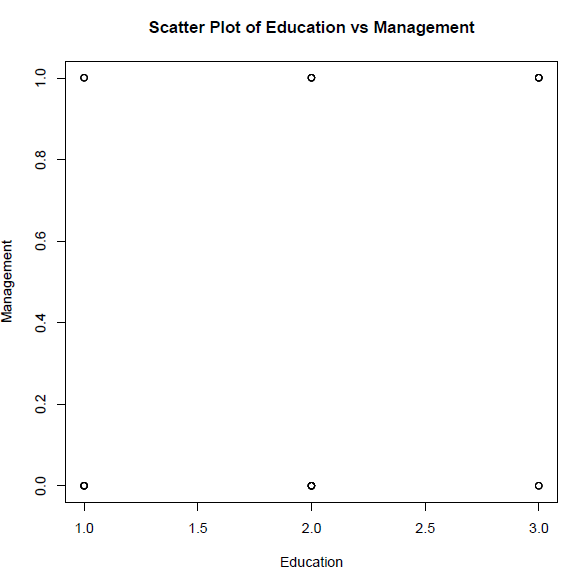
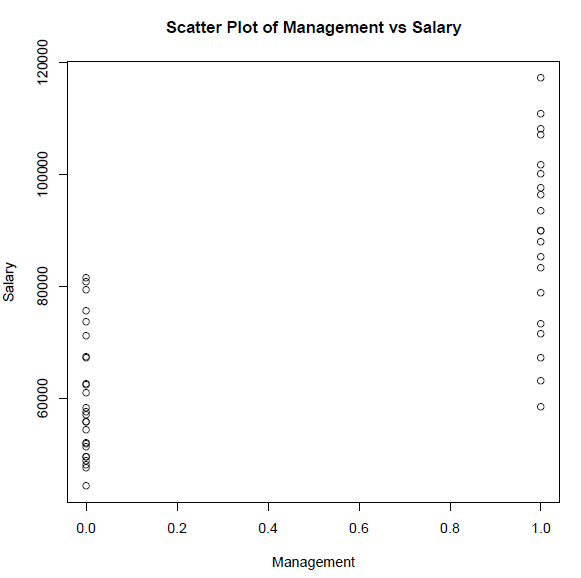
Multiple R-squared: 0.9568, Adjusted R-squared: 0.9526

F-statistic: 226.8 on 4 and 41 DF, p-value: < 2.2e-16

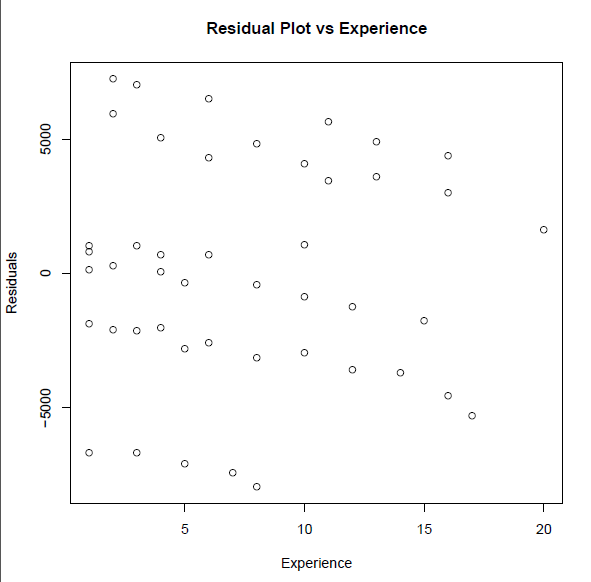
* 1. **Create the six pairwise scatterplots using the original variables of the dataset?  (Dummy variables don't usually make sense for scatterplots if they are for a variable with more than one level.)**

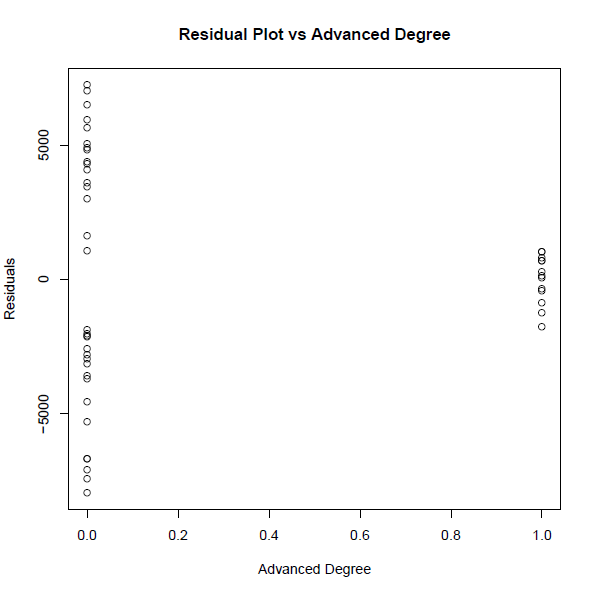
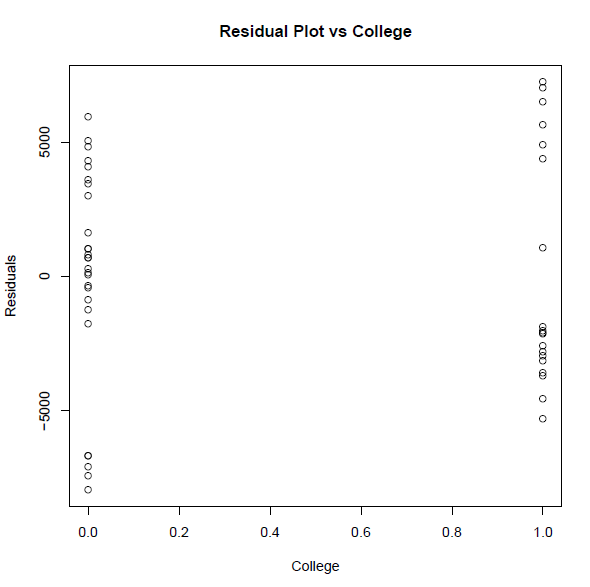


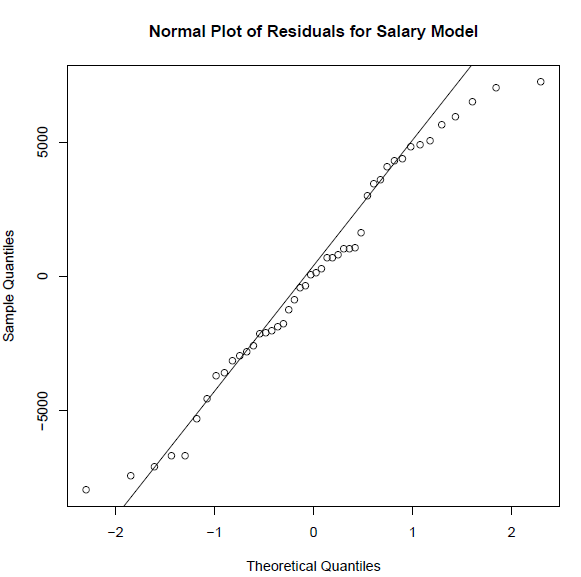
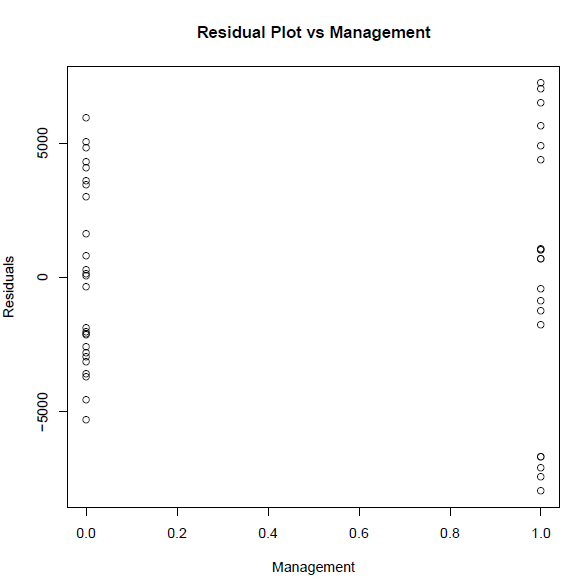




* 1. **Create the residual and normal plots of the residuals. For residual plots plot the residuals vs. predicted values and the residuals vs. each independent variable.**







* 1. **How much of an increase in salary is one additional year of experience likely to produce?**

An additional year of experience is likely to increase salary by $2,299.50.

* 1. **How much higher is the predicted salary of a college graduate than the salary of a person with only a highschool degree?**

The salary of a college graduate is predicted to be $13236.30 higher than the salary of a person with only a high school degree.

* 1. **For a high school graduate with 3 years of experience with no management responsibilities, what is the predicted salary?**

For a High School graduate with 3 years of experience with no management responsibilities, the predicted salary is $40,727.90.

* 1. **Does a person with an advanced degree have a higher predicted salary than a college graduate for this dataset?**

A person with an advanced degree does not have a higher salary than a college graduate for this dataset.

* 1. **Find 95% prediction interval for the person in Problem 7.**

> pred <- predict(model2, newdata=new\_data, interval="prediction")

> print(pred)

fit lwr upr

1 65108.93 55719.52 74498.34

2 48743.66 39456.64 58030.67

3 77723.67 68524.34 86922.99

4 49365.19 40149.26 58581.12

5 48743.66 39456.64 58030.67

6 80644.68 71375.54 89913.82

7 51664.67 42498.95 60830.39

8 38428.40 29198.17 47658.63

9 51043.14 41785.28 60301.00

10 53964.15 44841.54 63086.76

11 69707.89 60407.34 79008.44

12 82944.16 73717.92 92170.40

13 82322.63 73175.46 91469.80

14 43027.36 33872.27 52182.45

15 84622.11 75490.06 93754.16

16 55642.10 46420.90 64863.30

17 56263.63 47176.93 65350.34

18 58563.11 49505.01 67621.21

19 57941.58 48727.80 67155.36

20 74306.85 65066.83 83546.87

21 47626.32 38517.38 56735.27

22 89221.07 80097.18 98344.96

23 60862.59 51825.74 69899.45

24 89842.60 80701.97 98983.24

25 78905.81 69697.42 88114.20

26 65461.56 56444.91 74478.20

27 81205.29 72001.78 90408.81

28 93820.03 84674.81 102965.26

29 52225.29 43133.06 61317.51

30 56824.25 47719.14 65929.36

31 70060.52 61034.24 79086.80

32 98419.00 89223.15 107614.84

33 99040.53 89911.45 108169.61

34 101340.01 92195.39 110484.62

35 59123.73 50001.10 68246.35

36 74659.48 65593.82 83725.14

37 103017.96 93742.69 112293.23

38 63722.69 54543.11 72902.27

39 105938.97 96741.38 115136.56

40 79258.44 70124.05 88392.84

41 109916.40 100469.65 119363.15

42 112837.41 103506.65 122168.18

43 83857.40 74625.56 93089.25

44 70621.13 61302.43 79939.83

45 86156.89 76865.83 95447.94

46 79819.06 70219.33 89418.79

47 40727.88 31538.80 49916.96

The 95% prediction interval for this example is ($31,538.80, $49,916.96)