1) Andrew James Tillmann

2) My statement, "I have completed this work independently. The solutions given are entirely my own work."

3) Short Essay Questions

3a) Explain what cross-validation is and describe some of specific implementations. How is it used to validate a regression model?

Cross-validation is used by dividing the data in two groups. One is the training set and the other is the testing set. This process then can be repeated again and again to the nth time. Each time the training set and the testing set should be different. The average result of the testing set is the data looked at for K-fold cross-validation. It would be better if the process is done more than once to help take reduce the variability. However, doing once or more using this method of a subtest of data training and test set is the cross-validation method. The error you are trying to avoid is over fitting the model. For example the model looks great on using it on past data but it does not work as well for the future data.

3b)When building a model, you make four assumptions about the residuals. Explain what they are and how you can verify that your assumptions are correct.

First of the residuals of the regression model are the random error at each plot.

1) They are normally distributed.

   To verify normal distribution you could use a normal probability plot. This plot would show the distribution of the data by their standard deviation. The plot should show a straight line if not then the data is not normally distributed.

2) They have a mean of 0 (unbiased).

   To verify if the mean is 0 you would need to add all the residual values together and divide by the number of residuals if these number is 0 then there is no bias.

3) Have a constant variance of errors (Homoscedasticity).

   To verify if the constant of variance of errors you would need to make and look at a residual plots. If the plots data has an underlining trend and not just noise then the residual is not Homoscedasticity.

4) They are uncorrelated.

   To verify that the inputs(x) are uncorrelated with output(y) you could do a lag plot(a plot of each residual value[vertical axis=Y_i] and

residual value offset by lag[horizontal axis=Y_{i-1}]. If the plot shows an underlining trend then the data is bias thus correlated.

3c) When judging the quality of a model...

Why do we prefer adj-R^2 to R^2?

Adj-R^2 is better when looking at a model in its entirety if it has more than one beta that is not constant. The reason is that R^2 does not take into account for more than one non-constant beta where as adj-R^2 does.

Why do we prefer an F-test to a collection of t-test?

F-test is better when looking at a model in its entirety if it has more than one beta that is not constant. The reason is that t-test does not take into account for more than one non-constant beta where as the F-test does.

What are the units of MSE and RMSE? Explain how you can use them to evaluate a model.

Mean squared error (MSE) equals the squared of (value of the data points minus value of the fitted line). Root mean square error (RMSE) is the square root of the MSE. The higher the values of MSE and RMSE means that the values of the data points are of greater distance from the fitted line. The lower the values of MSE and RMSE means that the values of the data points are of lesser distance from the fitted line.

4a) Use the Banking dataset for this question, found under content on the D2L. This dataset consists of data acquired from banking and census record 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:

i. Median age of the population(Age) = 36.1

ii. Median years of education(Education) = 12.7

iii. Median income (Income) in $ = 47655.50

iv. Median home value (HomeVal) in $ = 97743.50

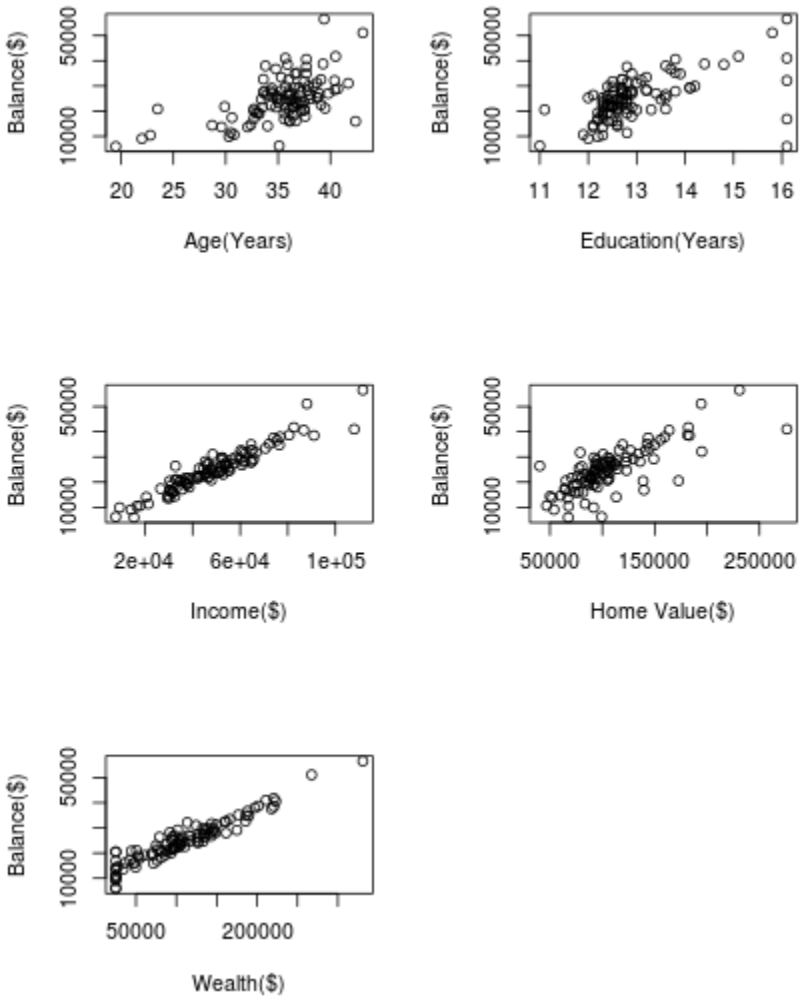
v. Median household wealth (Wealth) in $ = 102348

vi. Average bank balance (Balance) in $ = 24887.88

4b) Load the data into R. (Took some effort but done)

4c) In R, you can create a scatterplot by using the plot command, i.e. plot(x,y). Create scatterplots to visualize the associations between bank balance and the other five variables. Paste them into your submission. Describe the relationships.

Scatterplots:



Description:

Looking at the relationships you can see that there is a high positive relationship with wealth and income (this is show by the bottom left to top right movement with the data). It is interesting that Education shows some of the data at the upper bounds of years of education actually have a negative relationship with the balance. However, prior to that years of education seems to have a positive relationship with the balance. Home value seems to have some positive relationship. Age seems to

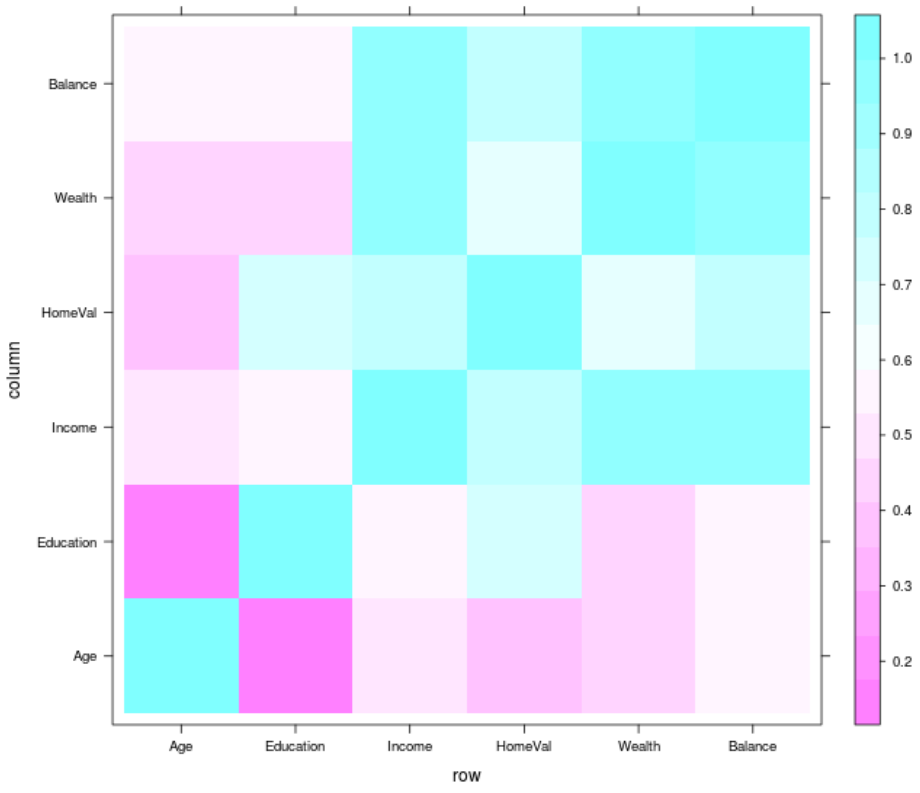
have a positive relationship after the age of 34 or so. Prior to that no correlation is clear.

4d) In R, you can compute correlations between two variables by using the cor command, i.e. cor(x,y) where x and y are the names of your variables, or you can compute pair-wise correlations by using cor(T), where T is the name of your table. Compute correlations found in the bank data. Interpret the correlation values. Paste them into your submission. Describe which variables appear to be strongly associated?

```
Table:
               Age Education    Income   HomeVal    Wealth   Balance
Age       1.0000000 0.1734611 0.4771464 0.3864931 0.4680918 0.5654668
Education 0.1734611 1.0000000 0.5731480 0.7489426 0.4681199 0.5521889
Income    0.4771464 0.5731480 1.0000000 0.7953565 0.9466652 0.9516842
HomeVal   0.3864931 0.7489426 0.7953565 1.0000000 0.6984778 0.7663871
Wealth    0.4680918 0.4681199 0.9466652 0.6984778 1.0000000 0.9487117
Balance   0.5654668 0.5521889 0.9516842 0.7663871 0.9487117 1.0000000
```

HeatMap visual for above table:

Description: Wealth and Income are strong positive associated with each other as well as the balance. Age and education seem to have a strong negative association with each other.



4e) Fit a regression model of balance vs the other five variables. Write the expression of the estimated regression model and evaluate it. Recall that you can build a linear regression model by using the lm command and display the model by using the summary command.

```
> regression=lm(formula=Balance ~ Age + Education + Income +HomeVal + Wealth, data=data)
> summary(regression)

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

Residuals:
    Min      1Q  Median      3Q     Max
-5365.5 -1102.6   -85.9   868.9  7746.4

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.033e+04  4.219e+03  -2.449 0.016160 *
Age          3.175e+02  6.104e+01   5.201 1.12e-06 ***
Education    5.903e+02  3.151e+02   1.873 0.064085 .
Income       1.468e-01  4.083e-02   3.596 0.000512 ***
HomeVal      9.864e-03  1.099e-02   0.898 0.371599
Wealth       7.414e-02  1.120e-02   6.620 2.06e-09 ***
---
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: < 2.2e-16
```

4f) Which of the five predictors have a significant effect on balance?(a=.05)Explain.

The independent variables where significant effect on balance are Age,Income,Wealth. They have a significant effect where a=0 this is indicated by the '***' next to the p-value. When a=.05 the it is indicated by the '*' and if a=.01 it is '**'.

4g) 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 versus the remaining four predictors. Write down the expression of the new regression model.

```
> regression=lm(formula=Balance ~ Age + Education + Income + Wealth, data=data)
> summary(regression)

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

Residuals:
    Min      1Q  Median      3Q     Max
-5403.9 -1234.1   -75.0   998.6  7430.6

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.214e+04  3.704e+03  -3.278  0.00145 **
Age          3.242e+02  6.051e+01   5.358 5.68e-07 ***
```

```
Education    7.498e+02  2.600e+02   2.884  0.00484 **
Income       1.615e-01  3.738e-02   4.321 3.75e-05 ***
Wealth       7.265e-02  1.106e-02   6.566 2.56e-09 ***
---
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: < 2.2e-16
```

4h) Analyze if all four predictors have a significant association with balance? (a=.05) If not continue to remove one insignificant variable at a time until all of the remaining predictors are significant.

All remaining predictors have a significant association with balance at lest a=.05 .

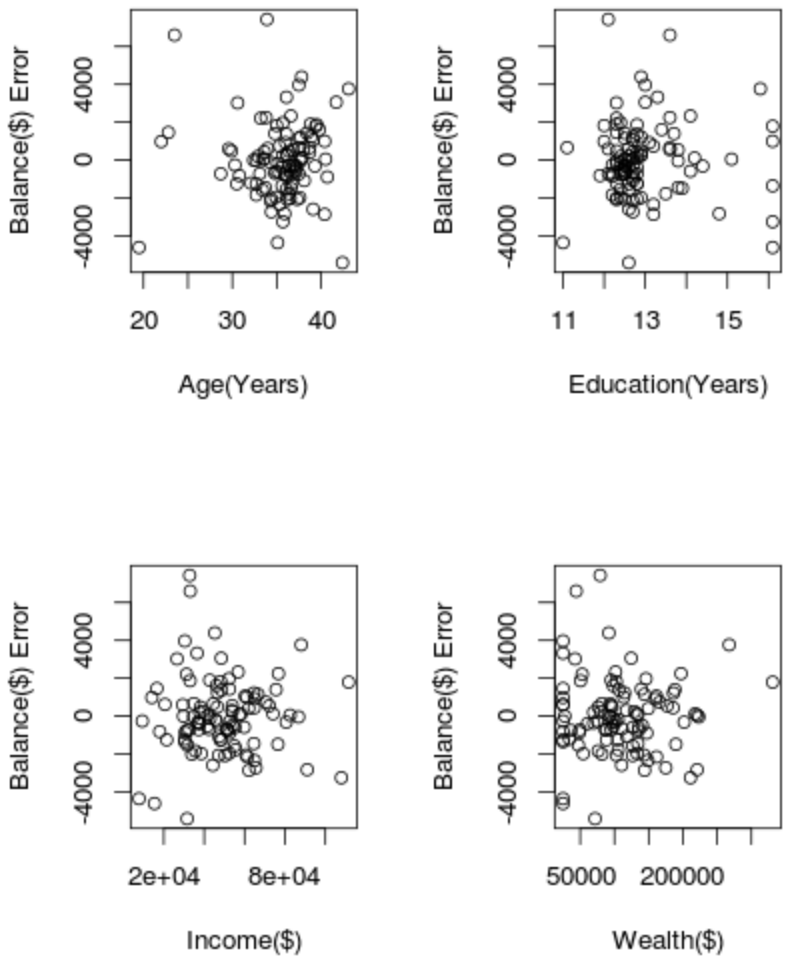
4i) Interpret each of the regression coefficients for the final model.

Balance = -12140+ 324.2*(age)+749.8*(education)+.1615*(Income) + .07265*(Wealth)

4j) Discuss the adj-R^2 for the final model.

The adj-R^2 is .9441 thus 94.41% of the balance is explained by the model.

4k)Create residual plots. Paste them into your submission. You can calculate the residuals by using the resid(m) command, where m is your model name, and then create a plot.



4l)Analyze the residual plots.

From looking at the residual plot they appear to have a mean of 0 and of constant variance. Their also appears to have no bias/direction in the plot.

4m) Are there any influence points for your final regression model? Explain what impact an influence point might have.

There seems to be a influence point at age 30 since the ages before that do not seem to have a mean of 0 (more data will be needed). Influence points can have a major impact on the model since the trend of the data is changing after that point. It would be better to use two different models or two different betas. One beta for data before the point and one for the data after the point. This is due to two different trends one before and one after the point.