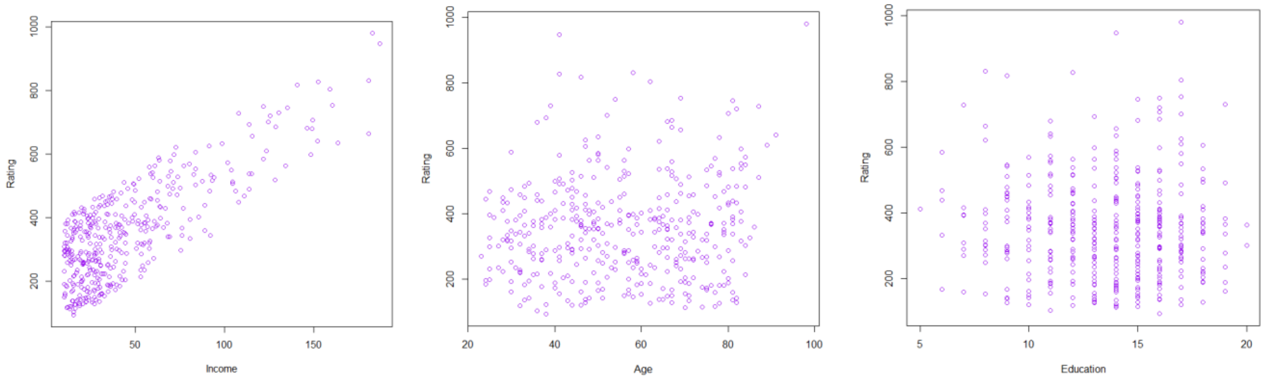
ADMN 872: Predictive Analytics

Homework 1

Team Hotel

Alejandro Robles, Alden Finver

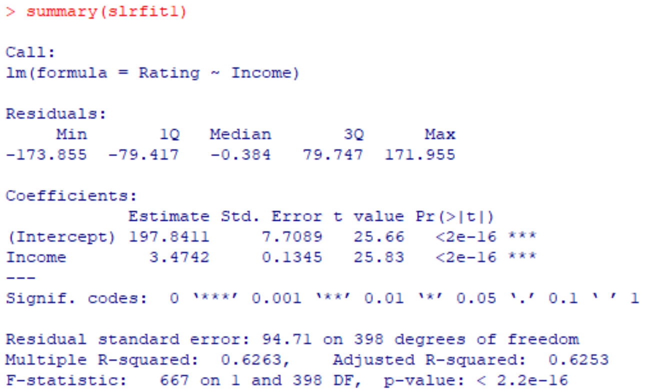
1) Explore if a linear relationship is viable between Rating (dependent variable) and all the other variables by obtaining the scatter plots and the respective correlations.



Rating Vs income displays a positive linear relationship while Rating vs Age and Rating vs Education does not.

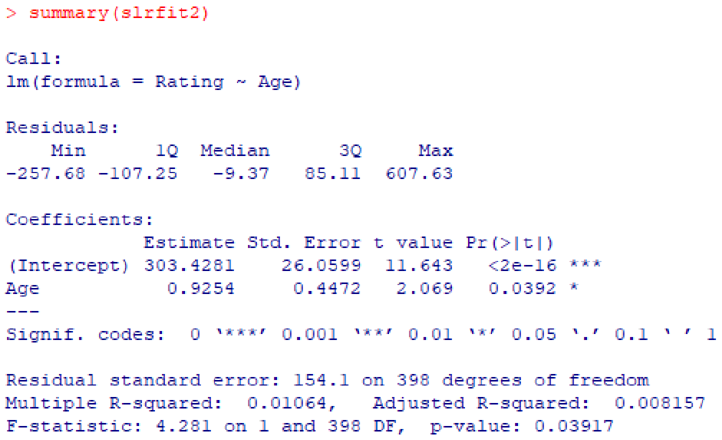
2) Estimate 3 simple linear regression models between Rating (as the dependent variable) and all the other variables. Report your regression models.

Regression 1



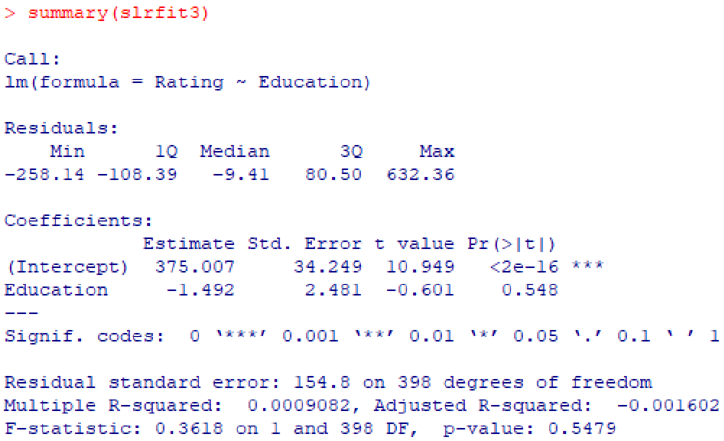
```
> summary(slrfit1)

Call:
lm(formula = Rating ~ Income)

Residuals:
     Min       1Q   Median       3Q      Max 
-173.855  -79.417   -0.384   79.747  171.955 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 197.8411     7.7089   25.66   <2e-16 ***
Income        3.4742     0.1345   25.83   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 94.71 on 398 degrees of freedom
Multiple R-squared:  0.6263,	Adjusted R-squared:  0.6253 
F-statistic:   667 on 1 and 398 DF,  p-value: < 2.2e-16
```

Regression 2



```
> summary(slrfit2)

Call:
lm(formula = Rating ~ Age)

Residuals:
    Min      1Q  Median      3Q     Max 
-257.68 -107.25   -9.37   85.11  607.63 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 303.4281    26.0599  11.643   <2e-16 ***
Age           0.9254     0.4472   2.069   0.0392 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 154.1 on 398 degrees of freedom
Multiple R-squared:  0.01064,	Adjusted R-squared:  0.008157 
F-statistic: 4.281 on 1 and 398 DF,  p-value: 0.03917
```

Regression 3



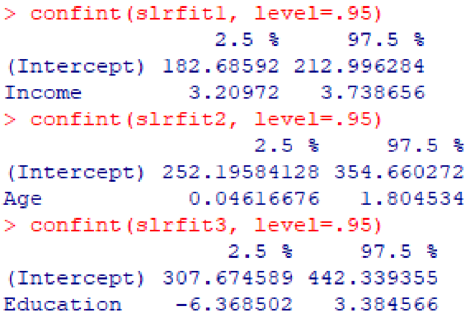
```
> summary(slrfit3)

Call:
lm(formula = Rating ~ Education)

Residuals:
    Min      1Q  Median      3Q     Max 
-258.14 -108.39   -9.41   80.50  632.36 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  375.007     34.249  10.949   <2e-16 ***
Education     -1.492      2.481  -0.601    0.548    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 154.8 on 398 degrees of freedom
Multiple R-squared:  0.0009082,	Adjusted R-squared:  -0.001602 
F-statistic: 0.3618 on 1 and 398 DF,  p-value: 0.5479
```

3) Comment on the intercept and slope coefficients in the context of the problem. What do they represent? Also obtain the 95% confidence intervals for the intercept and slope coefficients for all three models.

For instance in **Regression 1**, 212.99 (intercept estimate) represents the **expected** ratings if the company spends zero dollars in annual income of a customer. 3.73 (slope estimate) represents the **expected** increase in ratings if the company were to increase their annual income of a customer by 1 unit.

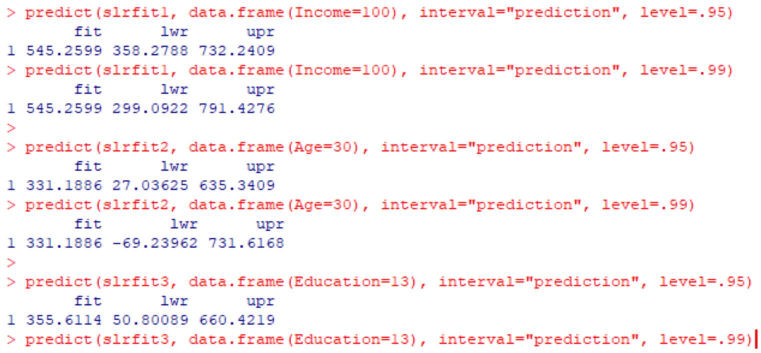
For instance in **Regression 2**, 354.66 (intercept estimate) represents the **expected** ratings if the company have an age of 0. 1.80 (slope estimate) represents the **expected** increase in ratings if the company were to increase the age by 1.

For instance in **Regression 3**, 442.33 (intercept estimate) represents the **expected** ratings if the company doesn't have years of education. 3.38 (slope estimate) represents the **expected** increase in ratings if the company were to earn one year of education.



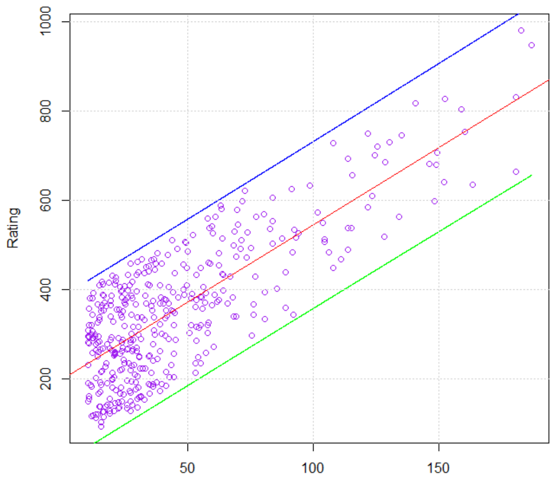
```
> confint(slrfit1, level=.95)
                2.5 %     97.5 %
(Intercept) 182.68592 212.996284
Income        3.20972   3.738656
> confint(slrfit2, level=.95)
                   2.5 %     97.5 %
(Intercept) 252.19584128 354.660272
Age           0.04616676   1.804534
> confint(slrfit3, level=.95)
                 2.5 %     97.5 %
(Intercept) 307.674589 442.339355
Education    -6.368502   3.384566
```

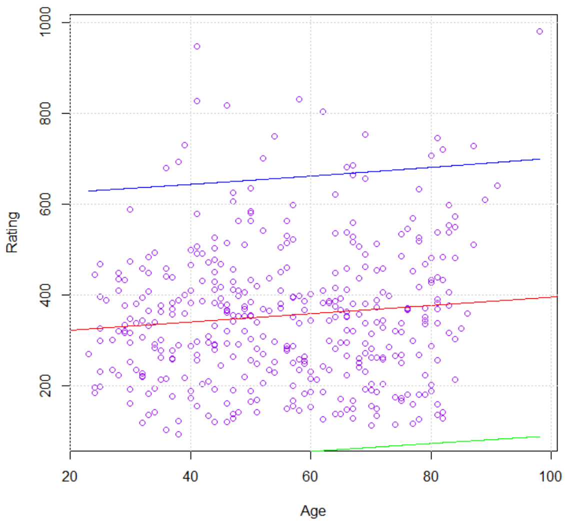
4) Obtain the predictions for a customer rating whose income is 100,000, whose age is 30, whose education level is 13 (for each model separately). In doing so, also obtain 95% and 99% prediction intervals and comment on your findings.

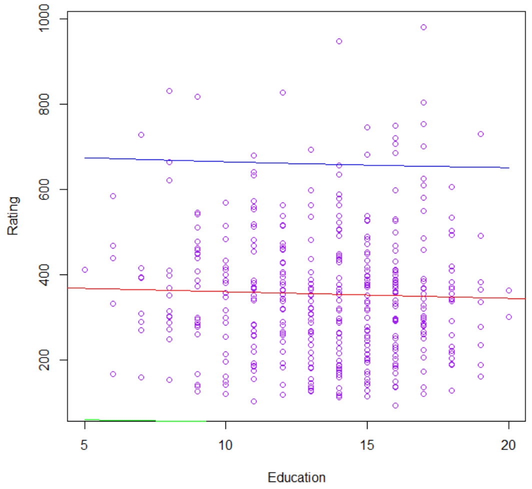


```
> predict(slrfit1, data.frame(Income=100), interval="prediction", level=.95)
       fit      lwr      upr
1 545.2599 358.2788 732.2409
> predict(slrfit1, data.frame(Income=100), interval="prediction", level=.99)
       fit      lwr      upr
1 545.2599 299.0922 791.4276
>
> predict(slrfit2, data.frame(Age=30), interval="prediction", level=.95)
       fit      lwr      upr
1 331.1886 27.03625 635.3409
> predict(slrfit2, data.frame(Age=30), interval="prediction", level=.99)
       fit       lwr      upr
1 331.1886 -69.23962 731.6168
>
> predict(slrfit3, data.frame(Education=13), interval="prediction", level=.95)
       fit      lwr      upr
1 355.6114 50.80089 660.4219
> predict(slrfit3, data.frame(Education=13), interval="prediction", level=.99)|
```

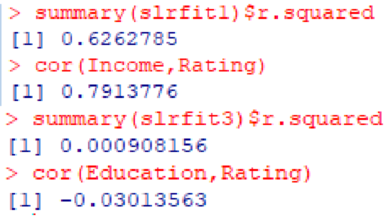
5) Obtain plots of the actual "Rating" variable versus your three fitted simple linear regression models and the corresponding 95% prediction intervals.



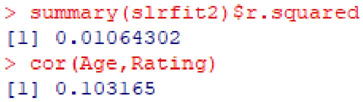




6) Obtain the R-squared estimates for all three models and using the R-squared estimates obtain the correlations and comment on your findings.



```
> summary(slrfit1)$r.squared
[1] 0.6262785
> cor(Income,Rating)
[1] 0.7913776
> summary(slrfit3)$r.squared
[1] 0.000908156
> cor(Education,Rating)
[1] -0.03013563
```



```
> summary(slrfit2)$r.squared
[1] 0.01064302
> cor(Age,Rating)
[1] 0.103165
```

For 1 we were able to capture 62% of the deviation of the credit rating deviation by fitting this regression line

For 3 our correlation is negative be because there is negative linear relationship between education and credit rating.

7) Which model (among the three simple linear regressions) provides the best fit to data? What are the measures you are using to compare the three models?

The model that provides the best fit is model 1 based on the R squared of .62 and the lower standard error of 94