homework2

Sunny Lee

2/5/2021

1)

a) To fit a linear model onto balance, we can simply call the lm() command with Balance as the response and a dot to get a linear model for each of the predictors. With this, we can also call the summary function in order to get some statistics on each of the linear models.

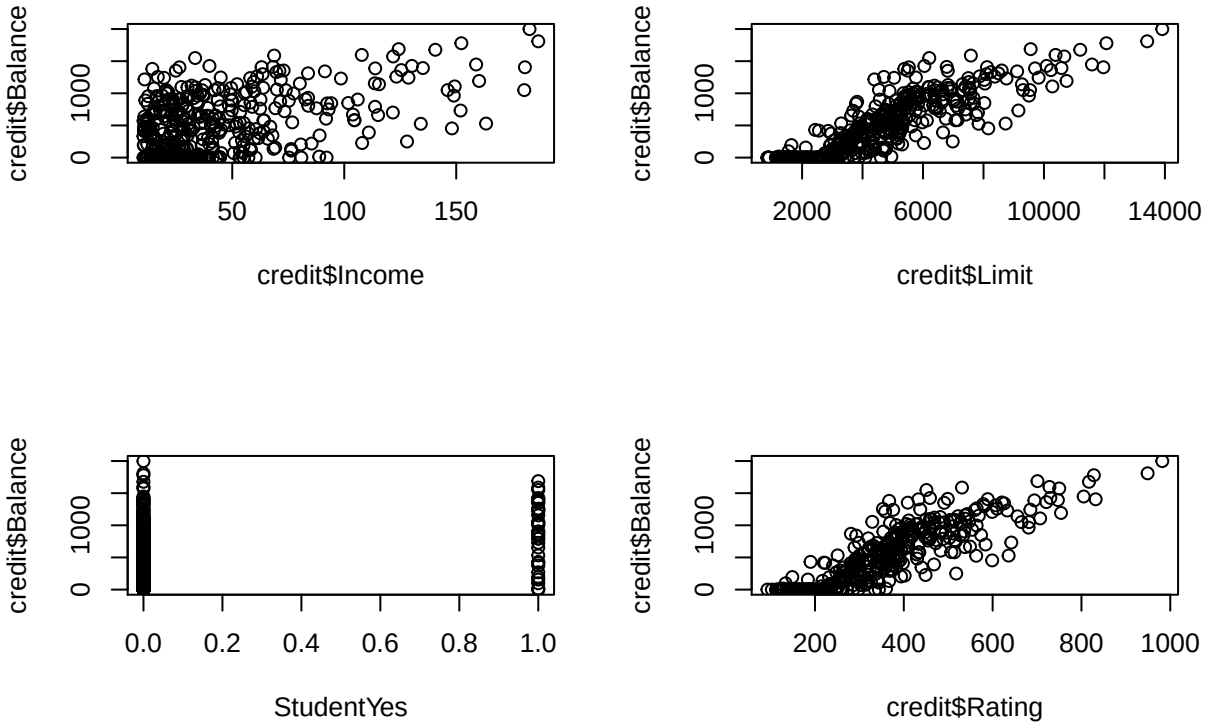
```
credit <- read.csv("Credit.csv")

coeff <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 0, 1)

result.income <- lm(Balance~Income, data = credit)
result.limit <- lm(Balance~Limit, data = credit)
result.rating <- lm(Balance~Rating, data = credit)
result.cards <- lm(Balance~Cards, data = credit)
result.age <- lm(Balance~Age, data = credit)
result.education <- lm(Balance~Education, data = credit)
result.gender <- lm(Balance~Gender, data = credit)
result.student <- lm(Balance~Student, data = credit)
result.married <- lm(Balance~Married, data = credit)
result.ethnicity <- lm(Balance~Ethnicity, data = credit)

coeff[1] <- result.income$coefficients[2]
coeff[2] <- result.limit$coefficients[2]
coeff[3] <- result.rating$coefficients[2]
coeff[4] <- result.cards$coefficients[2]
coeff[5] <- result.age$coefficients[2]
coeff[6] <- result.education$coefficients[2]
coeff[7] <- result.gender$coefficients[2]
coeff[8] <- result.student$coefficients[2]
coeff[9] <- result.married$coefficients[2]
coeff[10] <- result.ethnicity$coefficients[2]
coeff[11] <- result.ethnicity$coefficients[3]
coeff
```

```
##  [1]   6.0483634   0.1716373   2.5662403  28.9869482   0.0489114  -1.1859636
##  [7] -19.7331231 396.4555556  -5.3474654 -18.6862745 -12.5025126
```

Here we can see the estimates, std error, t value and p-value for each of the predictors. Using the summary, we find that there are some predictors which have a high p-value and some which do not. The intercept, Income, Limit, StudentYes and Rating all have very low p values which would seem to suggest that they contribute to predicting the balance.

```
par(mfrow = c(2, 2))
plot(credit$Income, credit$Balance)
plot(credit$Limit, credit$Balance)
```

```
StudentYes <- (credit$Student == "Yes")*1
plot(StudentYes, credit$Balance)
plot(credit$Rating, credit$Balance)
```



With these plots, though we do not see too much of a correlation between Income and StudentYes, we see a strong correlation in Limit and Rating.

b) By using our multiple linear regression, we find that our p-value is quite small, which confirms the fact that can can definitely reject the null hypothesis. The predictors which have very low p values are Income, Limit, Rating, Cards, Age and StudentYes and these are what we are going to pick to reject the null hypothesis.

```
multifit <- lm(Balance~., data = credit)
summary(multifit)
```

```
##
## Call:
## lm(formula = Balance ~ ., data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -166.48  -77.62  -14.37   56.21  316.52
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -496.62039   36.51325 -13.601  < 2e-16 ***
## X                     0.04105    0.04343   0.945   0.3452
```

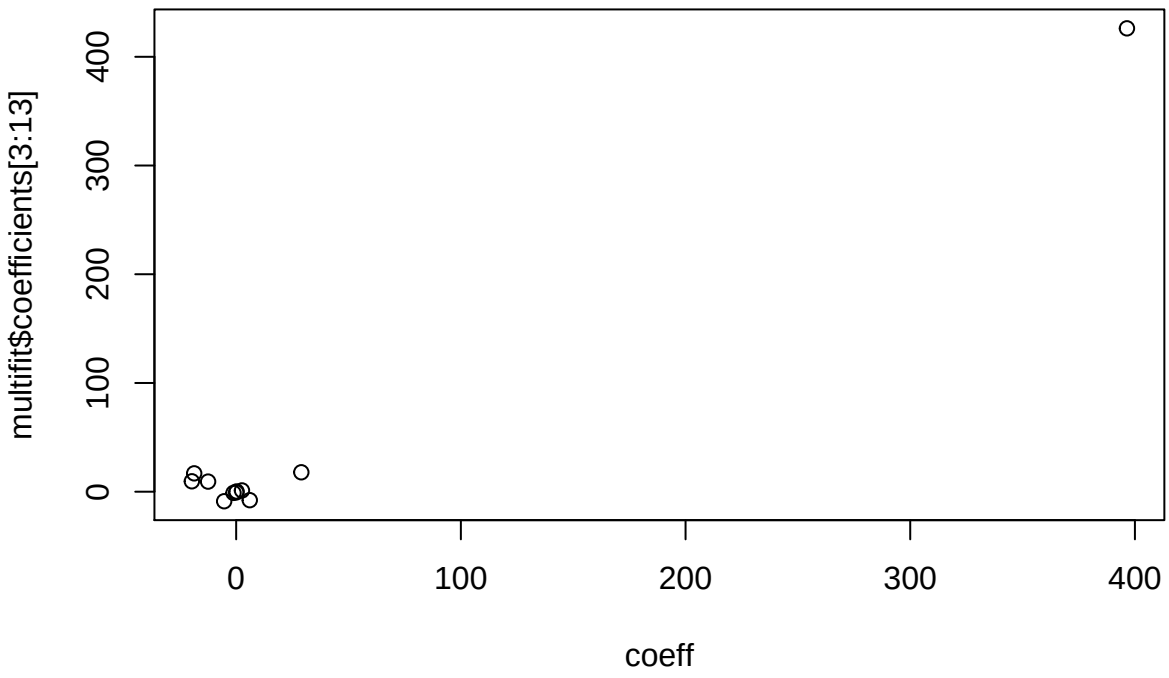
```
## Income               -7.80740    0.23431 -33.321  < 2e-16 ***
## Limit                 0.19052    0.03279   5.811  1.3e-08 ***
## Rating                1.14249    0.49100   2.327   0.0205 *
## Cards                17.83639    4.34324   4.107  4.9e-05 ***
## Age                  -0.62955    0.29449  -2.138   0.0332 *
## Education            -1.09831    1.59817  -0.687   0.4924
## GenderMale            9.54615    9.98431   0.956   0.3396
## StudentYes          426.16715   16.73077  25.472  < 2e-16 ***
## MarriedYes           -8.78055   10.36758  -0.847   0.3976
## EthnicityAsian       16.85752   14.12112   1.194   0.2333
## EthnicityCaucasian    9.29289   12.24194   0.759   0.4483
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.8 on 387 degrees of freedom
## Multiple R-squared:  0.9552, Adjusted R-squared:  0.9538
## F-statistic: 687.7 on 12 and 387 DF,  p-value: < 2.2e-16
```

c) The results from (a) and (b) are somewhat different. Income, Limit, Rating and StudentYes still have low p values, however Cards and Age appear to have low p values in the multiple linear regression but rather large p values in the simple linear regression.

```
multifit$coefficients
```

```
##        (Intercept)                  X             Income              Limit
##      -496.62039189          0.04104764         -7.80739871          0.19052127
##             Rating              Cards                Age          Education
##          1.14248766         17.83638753         -0.62954679         -1.09830902
##          GenderMale         StudentYes          MarriedYes     EthnicityAsian
##          9.54615446        426.16715394         -8.78055030         16.85751762
## EthnicityCaucasian
##          9.29289272
```

```
plot(coeff, multifit$coefficients[3:13])
```



d)

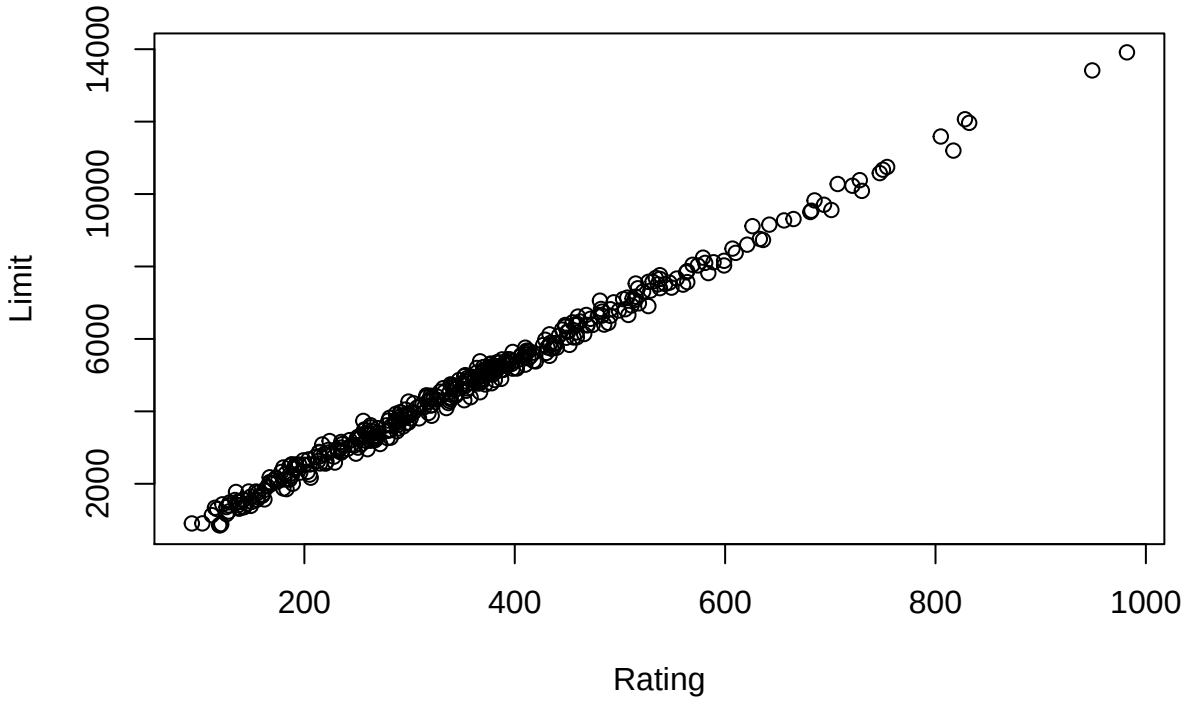
```
library(car)
```

```
## Loading required package: carData
```

```
vif(multifit)
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## X           1.030358  1        1.015066
## Income      2.787231  1        1.669500
## Limit     234.064316  1       15.299161
## Rating    235.887178  1       15.358619
## Cards       1.449767  1        1.204063
## Age         1.054739  1        1.027005
## Education   1.019588  1        1.009747
## Gender      1.019885  1        1.009894
## Student     1.032245  1        1.015994
## Married     1.045300  1        1.022399
## Ethnicity   1.040571  2        1.009992
```

```
plot(Limit~Rating, data = credit)
```



By using the vif function from the car library, we can find the variance inflation factors of each of the predictors under multiple linear regression. Having a high amount of variance inflation is undesirable, as a small change in our data set could drastically change our β_j for those predictors. Thus, from our vif(multifit) function above, we find most of the variables do not have very high variance inflation except for Limit and Rating. To visualize the colinearity between Rating and Limit, we plot Limit and Rating against each other. Finally, we can conclude Limit and Rating are collinear, thus we can either take either Limit or Rating out of our linear regression model, or we can combine them into one predictor.

e)

```
polyfit.income <- lm(Balance~poly(Income, 2), data = credit)
polyfit.limit <- lm(Balance~poly(Limit, 2), data = credit)
polyfit.rating <- lm(Balance~poly(Rating, 2), data = credit)
polyfit.cards <- lm(Balance~poly(Cards, 2), data = credit)
polyfit.age <- lm(Balance~poly(Age, 2), data = credit)
polyfit.education <- lm(Balance~poly(Education, 2), data = credit)

summary(polyfit.income)
```

```
##
## Call:
## lm(formula = Balance ~ poly(Income, 2), data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -782.88 -361.40  -54.98  316.26 1104.39
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          520.0       20.4  25.494   <2e-16 ***
## poly(Income, 2)1    4258.1      407.9  10.438   <2e-16 ***
## poly(Income, 2)2     370.6      407.9   0.908    0.364
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 408 on 397 degrees of freedom
## Multiple R-squared:  0.2166, Adjusted R-squared:  0.2127
## F-statistic: 54.88 on 2 and 397 DF,  p-value: < 2.2e-16
```

```
summary(polyfit.limit)
```

```
##
## Call:
## lm(formula = Balance ~ poly(Limit, 2), data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -675.80 -137.03   -5.41  136.29  750.09
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       520.01      11.56   44.99  < 2e-16 ***
## poly(Limit, 2)1  7913.55     231.17   34.23  < 2e-16 ***
## poly(Limit, 2)2  -707.35     231.17   -3.06  0.00236 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 231.2 on 397 degrees of freedom
## Multiple R-squared:  0.7485, Adjusted R-squared:  0.7472
## F-statistic: 590.6 on 2 and 397 DF,  p-value: < 2.2e-16
```

```
summary(polyfit.rating)
```

```
##
## Call:
## lm(formula = Balance ~ poly(Rating, 2), data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -708.79 -133.94   -1.86  141.48  801.95
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         520.01      11.46  45.395  < 2e-16 ***
## poly(Rating, 2)1   7931.25     229.11  34.618  < 2e-16 ***
## poly(Rating, 2)2   -772.48     229.11  -3.372  0.00082 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.1 on 397 degrees of freedom
## Multiple R-squared:  0.7529, Adjusted R-squared:  0.7517
## F-statistic: 604.9 on 2 and 397 DF,  p-value: < 2.2e-16
```

```
summary(polyfit.cards)
```

```
##
## Call:
## lm(formula = Balance ~ poly(Cards, 2), data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -641.34 -449.38  -53.22  353.41 1464.27
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       520.01      22.93  22.678   <2e-16 ***
## poly(Cards, 2)1   793.99     458.61   1.731   0.0842 .
## poly(Cards, 2)2   459.88     458.61   1.003   0.3166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 458.6 on 397 degrees of freedom
## Multiple R-squared:  0.009982,   Adjusted R-squared:  0.004995
## F-statistic: 2.001 on 2 and 397 DF,  p-value: 0.1365
```

```
summary(polyfit.age)
```

```
##
## Call:
## lm(formula = Balance ~ poly(Age, 2), data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -541.20 -459.35  -54.92  345.85 1429.46
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     520.01      23.04  22.569   <2e-16 ***
## poly(Age, 2)1    16.85     460.82   0.037    0.971
## poly(Age, 2)2   182.97     460.82   0.397    0.692
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.8 on 397 degrees of freedom
## Multiple R-squared:  0.0004003,  Adjusted R-squared:  -0.004635
## F-statistic: 0.07949 on 2 and 397 DF,  p-value: 0.9236
```

```
summary(polyfit.education)
```

```
##
## Call:
## lm(formula = Balance ~ poly(Education, 2), data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -599.30 -463.61  -51.64  338.60 1471.20
##
## Coefficients:
```

```
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            520.01      23.02  22.588   <2e-16 ***
## poly(Education, 2)1    -74.03     460.44  -0.161    0.872
## poly(Education, 2)2    411.44     460.44   0.894    0.372
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.4 on 397 degrees of freedom
## Multiple R-squared:  0.002072,   Adjusted R-squared:  -0.002955
## F-statistic: 0.4122 on 2 and 397 DF,  p-value: 0.6625
```

From the summaries given above, we find that Limit and Rating have low p values for the β_j in the polynomial terms. We might conclude that these predictors have some form of polynomial association.