**Math 229 Final Exam Due: Friday, May 10 at 10 pm.**

This is a take-home exam. ***Please insert your answers into this document.*** For this exam you may use only R, a calculator, your notes, and my notes. You may not discuss any aspect of this exam with anybody but the instructor. Neatness and clarity will be appreciated. Do show the major steps in your work and only as much R output as is necessary to support your answers. Correct answers may not receive full credit without supporting work. Please write and sign (electronically) the Honor Pledge [I have neither given nor received help on this exam] on your exam. Please rename the document Exam3_Your name, and submit this assignment electronically (as an attachment please) in the Dropbox for Final Exam.

**Please do not write on this page.**

Honor Pledge

**I have neither given nor received help on this exam.**

Signed

**Ellen Stanton**

1. In the 1980s researchers in The Hague, Netherlands, suspected an association between keeping birds as pets and an increased risk of lung cancer. To investigate bird-keeping as a risk factor, researchers conducted a case-control study of patients in 1995 at four hospitals in The Hague (population 450,000). They identified 49 cases of lung cancer among patients who were registered with a general practice, who were aged 65 or younger, and who had resided in the city since 1975. They also selected 98 controls from a population of residents having the same general age structure.

The following data are stored in the data set *Birdkeeping.csv*.

```
Count  Name
147    LC   1 = lung cancer, 0 = Control
147    SEX  1 = female, 0 = male
147    SS   Socio-economic status; 1 = high, 0 = low
147    BK   Bird-keeping; 1 = yes, 0 = no
147    AG   Age
147    YR   years smoked
147    CD   cigarettes per day
```

The response variable is LC.

(a) Obtain a contingency table with LC as the rows and BK as the columns. Paste in the table here.

```
tally(LC~BK,data=BirdKeeping)
   BK
LC   0  1
  0 64 34
  1 16 33
```

Use the counts in the table to compute:

(i) the odds of lung cancer for those keeping a bird

```
33/34
[1] 0.9705882
```

**The odds of lung cancer for those keeping a bird are .9706.**

(ii) the odds of lung cancer for those not keeping a bird

```
16/64
[1] 0.25
```

**The odds of lung cancer for those not keeping a bird are .2500.**

(iii) an odds ratio (greater than one, please).

```
.9705882/.25
[1] 3.882353
```

**The odds ratio of lung cancer for those keeping a bird over those not keeping a bird is 3.88.**

Interpret your answer in part (iii).

**For those keeping a bird, the odds of lung cancer are 3.88 times the odds of lung cancer for those not keeping a bird.**

(b) Obtain output for a logistic regression predicting the probability of lung cancer as a logistic function of BK; paste in the output below. Insert BK = 1 and then BK = 0 into your expression and show that you obtain your answers in (a) (i) and (ii).

```
lcbk <- glm(LC~BK,data=BirdKeeping,family=binomial)
lcbk

Call:  glm(formula = LC ~ BK, family = binomial, data = BirdKeeping)

Coefficients:
(Intercept)           BK
     -1.386        1.356

Degrees of Freedom: 146 Total (i.e. Null);  145 Residual
Null Deviance:        187.1
Residual Deviance: 172.9      AIC: 176.9
```

**Probability of lung cancer(Y=1)^ = (e^(-1.386+1.356(BirdKeeping))) / (1+(e^(-1.386+1.356(BirdKeeping)))**
**Odds of lung cancer^ = e^(-1.386+1.356(BirdKeeping))**

**To get the answers from (a)(i) and (a)(ii), which find odds of lung cancer, we must plug the respective values for BirdKeeping into the above formula for odds of lung cancer.**

```
exp(-1.386+(1.356*0))
[1] 0.2500736
```

```
exp(-1.386+(1.356*1))
[1] 0.9704455
```

**The odds of lung cancer for those who keep birds (BirdKeeping=1) is .9704, approximately equal to the value found in part (a)(i).**
**The odds of lung cancer for those who do not keep birds (BirdKeeping=0) is .2501, approximately equal to the value found in part (a)(ii).**

(c) Compute a 90% confidence interval for the population odds ratio associated with keeping a bird.

```
confint(lcbk,level=.9)
Waiting for profiling to be done...
                   5 %       95 %
(Intercept) -1.8695851 -0.9455434
BK           0.7559191  1.9807513
exp(.7559191)
[1] 2.129568
exp(1.9807513)
[1] 7.248187
```

**A 90% confidence interval for the population odds ratio associated with keeping a bird is 2.130 to 7.248.**

(d) How would you use your interval to test to test whether the probability of lung cancer is independent of whether or not there is a bird in the house? Be sure to state the hypotheses and your conclusion.

**To use the confidence interval to test whether the probability of lung cancer is independent of whether or not there is a bird in the house, I would perform the following two-sided hypothesis test at the 10% level of significance (since the interval is a 90% confidence interval).**

**H0: Lung Cancer and BirdKeeping are independent, ß=0, OR=1**
**Ha: Lung Cancer and BirdKeeping are dependent, ß≠0, OR≠1**

**Because the interval for OR, 2.13 to 7.25, does not contain 1, we can reject the null hypothesis at the 10% level of significance. The data suggest that the population odds ratio associated with BirdKeeping is not equal to 1, and that Lung Cancer and BirdKeeping are not independent.**

(e) Obtain output for a logistic regression predicting the probability of lung cancer as a logistic function of CD; paste in the output below.

```
lccd <- glm(LC~CD,data=BirdKeeping,family=binomial)
lccd

Call:  glm(formula = LC ~ CD, family = binomial, data = BirdKeeping)

Coefficients:
(Intercept)           CD
   -1.53541      0.05113

Degrees of Freedom: 146 Total (i.e. Null);  145 Residual
Null Deviance:	    187.1
Residual Deviance: 179.6 	AIC: 183.6
```

(f) Write down an expression for the predicted probability of lung cancer as a function of CD.

**P(Predicted Lung Cancer^=1) = (e^(-1.53541 + .05113(CigarettesDaily))) / (1 + e^(-1.53541 + .05113(CigarettesDaily)))**

(g) What is the predicted probability of lung cancer for someone who did not smoke?

```
CD <- 0
num <- exp(-1.53541+(.05113*CD))
denom <- 1+num
num/denom
[1] 0.1772035
```

**The predicted probability of lung cancer for someone who did not smoke is .177.**

(h) What is the predicted probability of lung cancer for someone who smokes 40 cigarettes a day?

```
CD <- 40
num <- exp(-1.53541+(.05113*CD))
denom <- 1+num
num/denom
[1] 0.6247572
```

**The predicted probability of lung cancer for someone who smokes 40 cigarettes a day is .625.**

(i) Compute and interpret the odds ratio in this case.

```
exp(-1.53541+(.05113*0))
```

```
[1] 0.2153674
exp(-1.53541+(.05113*40))
[1] 1.664942
1.664942/.2153674
[1] 7.730706
exp(.05113)^40
[1] 7.730705
```

**The odds ratio in this case is 7.731. The odds of getting lung cancer for someone who smoked 40 cigarettes per day is 7.731 times the odds of lung cancer for someone who didn't smoke. For each additional 40 cigarettes smoked per day, the predicted odds of getting lung cancer change by a factor of 7.731.**

(j) What is the odds ratio associated with smoking 15 rather than 5 cigarettes per day? Interpret your answer.

```
exp(.05113*15)
[1] 0.4637267
exp(.05113*5)
[1] 0.278104
.4637267/.278104
[1] 1.667458
exp(.05113)^10
[1] 1.667457
```

**The odds ratio associated with smoking 15 rather than 5 cigarettes per day is 1.667. The odds of lung cancer for someone who smoked 15 cigarettes per day is 1.667 times the odds of lung cancer for someone who smoked 5 cigarettes per day. For each additional 10 cigarettes smoked per day, the predicted odds of getting lung cancer change by a factor of 1.667.**

(k) Regard the variables Sex, SS, Age, YR, and CD as potential confounding variables. Obtain output for a logistic regression predicting the probability of lung cancer as a logistic function of BK and these five variables. Interpret the odds ratio associated with BK in this case. Do these data suggest that after adjusting for these five variables, there is a significant relationship between the probability of lung cancer and whether or not a bird is kept in the house?

```
lcall <- glm(LC~BK+SEX+SS+AG+YR+CD,data=BirdKeeping,family=binomial)
lcall
```

```
Call:  glm(formula = LC ~ BK + SEX + SS + AG + YR + CD, family = binomial,
    data = BirdKeeping)
```

```
Coefficients:
(Intercept)           BK          SEX           SS           AG           YR           CD
   -1.93736      1.36259      0.56127      0.10545     -0.03976      0.07287      0.02602

Degrees of Freedom: 146 Total (i.e. Null);  140 Residual
Null Deviance:	    187.1
Residual Deviance: 154.2 	AIC: 168.2
```

```
exp(1.36259)
[1] 3.906298
```

**After adjusting for Sex, Socioeconomic Status, Age, Years Smoked, and Cigarettes per Day, for those keeping a bird, the odds of lung cancer are 3.91 times the odds of lung cancer for those not keeping a bird.**

```
summary(lcall)

Call:
glm(formula = LC ~ BK + SEX + SS + AG + YR + CD, family = binomial,
    data = BirdKeeping)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5642  -0.8333  -0.4605   0.9808   2.2460

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.93736    1.80425  -1.074 0.282924
BK           1.36259    0.41128   3.313 0.000923 ***
SEX          0.56127    0.53116   1.057 0.290653
SS           0.10545    0.46885   0.225 0.822050
AG          -0.03976    0.03548  -1.120 0.262503
YR           0.07287    0.02649   2.751 0.005940 **
CD           0.02602    0.02552   1.019 0.308055
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 187.14  on 146  degrees of freedom
Residual deviance: 154.20  on 140  degrees of freedom
AIC: 168.2
```

```
Number of Fisher Scoring iterations: 5
```

**These data suggest that even after adjusting for these five variables, there is a significant relationship between the probability of lung cancer and whether or not a bird is kept in the house. If looking at the Wald Z Test in the summary output above, the p-value for BK is .000923, so at the 1% level of significance we can conclude that after adjusting for the other variables, LC and BK are not independent. So, despite potential confounding variables, it is still worth keeping BK in the model containing the other five variables.**

(l) Did these five potential confounding variables prove to seriously affect the impact of BK on LC? Explain.

**These five potential confounding variables did not seriously affect the impact of BK on LC. Without the confounding variables, the odds ratio for BK was 3.88, and with them, it is 3.91. These two values are similar, showing that both with and without the confounding variables, the odds of lung cancer for someone who has a bird is approximately 3.9 times the odds of lung cancer for someone who doesn't have a bird.**

2. The data set *credit.csv* contains anonymous credit card data for 400 credit card holders. The variables in the data set are listed below.

| Variable | Description |
|---|---|
| Income | Annual income in 1000's dollars |
| Limit | Credit limit |
| Rating | Credit rating |
| Cards | Number of credit cards |
| Age | Age in years |
| Education | Years of education |
| Gender | 1 = female, 0 = male |
| Student | 1 = yes, 0 = no |
| Married | 1 = yes, 0 = no |
| Ethnicity | Caucasian, African-American, or Asian |

| | |
|---|---|
| Balance | Average credit card debt |

The object in this question is to construct and evaluate a model predicting the response variable Balance.

(a) What are the 'individuals' in this case?

**In this case, the individuals are the 400 credit card holders.**

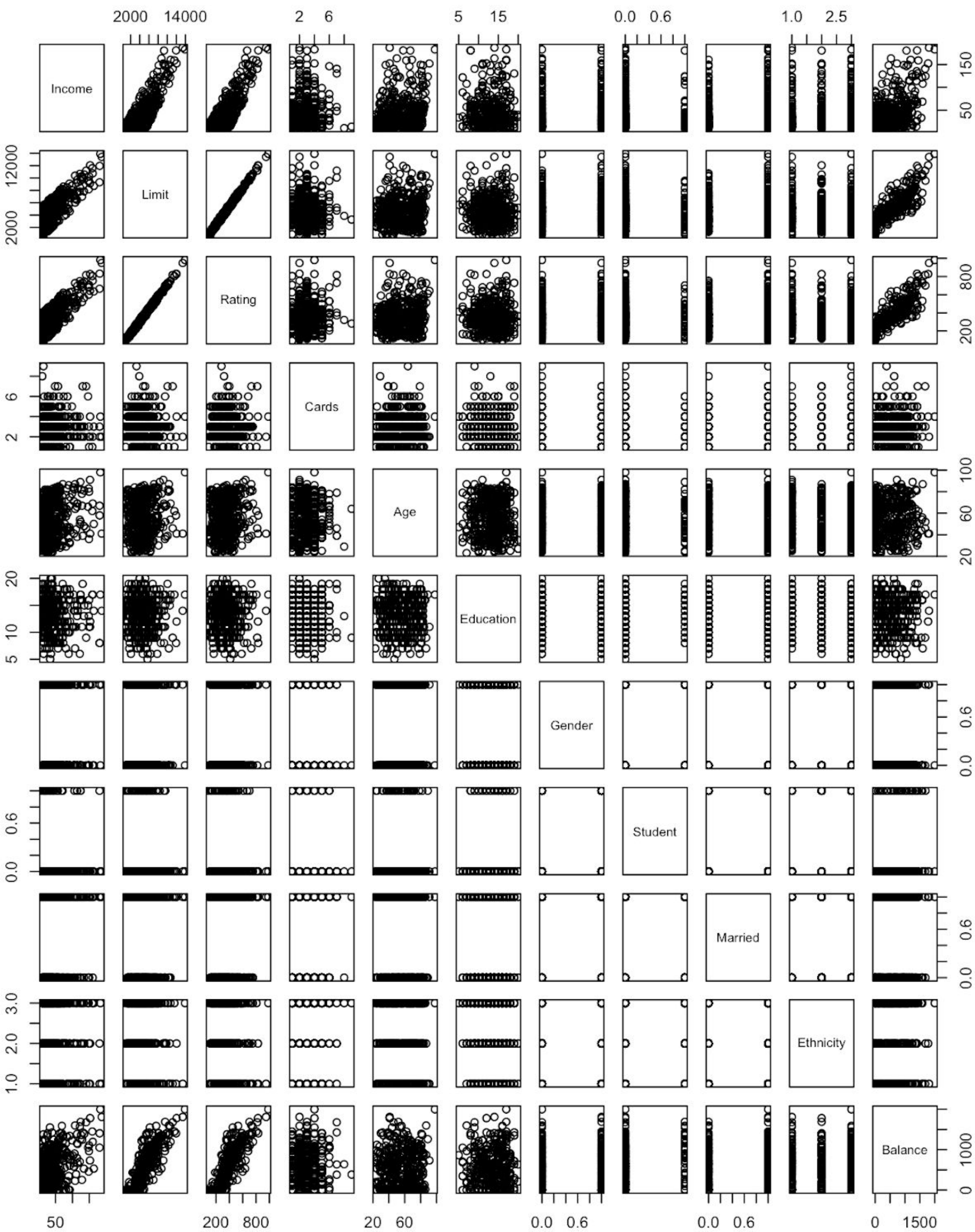
(b) (i) Obtain a correlation matrix for nine of the ten potential predictor. Please make the font small enough to be easily read. Which variable is most highly correlated with Balance?

```
d <- credit[,-c(10)]
round(cor(d),3)
          Income  Limit Rating  Cards    Age Education Gender Student Married Balance
Income     1.000  0.792  0.791 -0.018  0.175    -0.028 -0.011   0.020   0.036   0.464
Limit      0.792  1.000  0.997  0.010  0.101    -0.024  0.009  -0.006   0.031   0.862
Rating     0.791  0.997  1.000  0.053  0.103    -0.030  0.009  -0.002   0.037   0.864
Cards     -0.018  0.010  0.053  1.000  0.043    -0.051 -0.023  -0.026  -0.010   0.086
Age        0.175  0.101  0.103  0.043  1.000     0.004  0.004  -0.030  -0.073   0.002
Education -0.028 -0.024 -0.030 -0.051  0.004     1.000 -0.005   0.072   0.049  -0.008
Gender    -0.011  0.009  0.009 -0.023  0.004    -0.005  1.000   0.055   0.012   0.021
Student    0.020 -0.006 -0.002 -0.026 -0.030     0.072  0.055   1.000  -0.077   0.259
Married    0.036  0.031  0.037 -0.010 -0.073     0.049  0.012  -0.077   1.000  -0.006
Balance    0.464  0.862  0.864  0.086  0.002    -0.008  0.021   0.259  -0.006   1.000
```

**The variable most highly correlated with balance is Rating, with a correlation of approximately .864.**

(ii) Obtain a single graphical display that shows the relationship between Balance and the 10 potential predictors.
[You may want to use the code below to make Ethnicity a factor rather than a character variable.
credit$Ethnicity <- as.factor(credit$Ethnicity)]

```
credit$Ethnicity <- as.factor(credit$Ethnicity)
pairs(credit)
```




(c) I found the three quantitative variables most highly correlated with Balance were Income, Limit, and Rating. However, I discovered that I did not need both Limit and Rating and included just Income and Rating. Explain why did I not need both Rating and Limit.

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

```
[1] 0.9968797
.9968797^2
[1] 0.9937691
1-.9937691
[1] 0.0062309
1/.0062309
[1] 160.490
```

**Income, Limit, and Rating are indeed most highly correlated with Balance, as both the correlation matrix and the pairs plot show. But, Rating and Limit are not both needed, because they are so highly correlated with each other that they create the problem of multicollinearity. The correlation between Rating and Limit, .997, have an associated Variance Inflation Favor of approximately 160, which is far greater than 5, showing that multicollinearity is a problem. Thus, including them both would lead to poorly estimated regression coefficients, so we drop one of the variables.**

(d) As well as these two quantitative variables, I wanted to include at least one qualitative predictor. Which of the four qualitative variables should I include? Hint: For each of the four variables compare the mean Balance for each category of the variable.

```
mean(Balance~Ethnicity,data=credit)
African American            Asian        Caucasian
        531.0000         512.3137         518.4975
mean(Balance~Married,data=credit)
       0        1
523.2903 517.9429
mean(Balance~Student,data=credit)
       0        1
480.3694 876.8250
mean(Balance~Gender,data=credit)
       0        1
509.8031 529.5362
```

**After looking at the mean Balance for each category of the four qualitative variables, you should choose Student as your qualitative variable to add. This is because the mean values of Balance per category of Student are the most different, at approximately 480 for non-students and 877 for students. The values among different ethnicities, between married and nonmarried individuals, and between males and females are nowhere near this different from each other.**

(e) Build the linear model with Income and Rating and the qualitative variable you selected in part (d) above. Write down your model below.

```
bairs <- lm(Balance~Income+Rating+Student,data=credit)
bairs

Call:
lm(formula = Balance ~ Income + Rating + Student, data = credit)

Coefficients:
(Intercept)       Income       Rating      Student
   -581.079       -7.875        3.987      418.760
```

**Predicted Balance^ = -581.079 - 7.875(Income in $1000s) + 3.987(Rating) + 418.760(Student)**

(f) For your model in (e), interpret the slope associated with Income and the slope associated with the qualitative variable you selected.

**After adjusting for Rating and Student, for each additional $1000 of income, the predicted balance of an individual decreases by 7.875.**

**After adjusting for Income and Rating, the predicted balance for a student is 418.760 higher than the predicted balance for a non-student.**

(h) How much of the variability in Balance can be associated with your model in part (e)?

```
summary(bairs)

Call:
lm(formula = Balance ~ Income + Rating + Student, data = credit)

Residuals:
     Min       1Q   Median       3Q      Max
-226.126  -80.445   -5.018   65.192  293.234

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -581.07889   13.83463  -42.00   <2e-16 ***
Income        -7.87493    0.24021  -32.78   <2e-16 ***
Rating         3.98747    0.05471   72.89   <2e-16 ***
Student      418.76028   17.23025   24.30   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 103.3 on 396 degrees of freedom
Multiple R-squared:  0.9499,	Adjusted R-squared:  0.9495 
F-statistic:  2502 on 3 and 396 DF,  p-value: < 2.2e-16
```

**The percentage of variability in balance that can be associated with the model in part (e) is 94.99%.**

(i) As you may recall, the base version of R does not compute VIF values. This maybe because it is fairly straightforward to compute them from their definition. Compute the three VIF values for the three variables in your model in part (f). Show your code and do not use the *car*package. Do you see any problems with these three VIF values? Explain.

```
summary(lm(Income~Rating+Student,data=credit))

Call:
lm(formula = Income ~ Rating + Student, data = credit)

Residuals:
    Min      1Q  Median      3Q     Max 
-40.235 -17.581  -0.313  14.880  77.021 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -19.017453   2.728413  -6.970 1.33e-11 ***
Rating        0.180276   0.006985  25.810  < 2e-16 ***
Student       2.491805   3.597835   0.693    0.489    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.59 on 397 degrees of freedom
Multiple R-squared:  0.6267,	Adjusted R-squared:  0.6248 
F-statistic: 333.3 on 2 and 397 DF,  p-value: < 2.2e-16

1/(1-.6267)
[1] 2.678811
```

**The VIF value for Income is 2.68.**

```
summary(lm(Rating~Income+Student,data=credit))

Call:
lm(formula = Rating ~ Income + Student, data = credit)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-174.832  -75.239    0.478   78.910  171.023 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 198.6777     7.8524  25.302   <2e-16 ***
Income        3.4757     0.1347  25.810   <2e-16 ***
Student      -9.0508    15.8007  -0.573    0.567    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 94.79 on 397 degrees of freedom
Multiple R-squared:  0.6266,  Adjusted R-squared:  0.6247 
F-statistic: 333.1 on 2 and 397 DF,  p-value: < 2.2e-16

1/(1-.6266)
[1] 2.678093
```

**The VIF value for Rating is 2.68.**

```
summary(lm(Student~Income+Rating,data=credit))

Call:
lm(formula = Student ~ Income + Rating, data = credit)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.13156 -0.10651 -0.09809 -0.08853  0.91693 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  1.105e-01  3.991e-02   2.768   0.0059 **
Income       4.843e-04  6.993e-04   0.693   0.4890   
Rating      -9.124e-05  1.593e-04  -0.573   0.5671   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3009 on 397 degrees of freedom
Multiple R-squared:  0.001211,Adjusted R-squared:  -0.003821 
F-statistic: 0.2407 on 2 and 397 DF,  p-value: 0.7862

1/(1-.001211)
```

```
[1] 1.001212
```

**The VIF value for Student is 1.001.**

**I do not see any problem with these three VIF values. I would flag a VIF value if it was greater than 5, but none of these three are.**

(j) Please test whether there is any benefit to adding the variable Ethnicity to your model in part (e). Carefully state your result. Let R do the work! Be sure to state your conclusion.

```
bairse <- lm(Balance~Income+Rating+Student+Ethnicity,data=credit)

bairse

Call:
lm(formula = Balance ~ Income + Rating + Student + Ethnicity,
    data = credit)

Coefficients:
       (Intercept)              Income              Rating             Student
EthnicityAsian  EthnicityCaucasian
          -592.231              -7.875               3.990             417.944
21.100              10.205
```

**To do the hypothesis test below, we should understand that R has broken Ethnicity into indicator variables and has chosen EthnicityAsian and EthnicityCaucasian, leaving AfricanAmerican as the reference group.**

```
anova(bairs,bairse)
Analysis of Variance Table

Model 1: Balance ~ Income + Rating + Student
Model 2: Balance ~ Income + Rating + Student + Ethnicity
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    396 4227219
2    394 4204897  2     22322 1.0458 0.3524
```

**H0: For a model using Income, Rating, and Student to predict Balance, there is no predictive value to adding the variable Ethnicity to the model. The coefficients in the model for the indicator variables created by R are equal to zero.**

**Ha: For a model using Income, Rating, and Student to predict Balance, there is significant predictive value to adding the variable Ethnicity to the model. Some of the coefficients in the model for EthnicityCaucasian and EthnicityAsian are not equal to zero.**

**With a p-value of .3524 for adding Ethnicity (which is done by adding EthnicityAsian and EthnicityCaucasian as a block) to a model already predicting Balance with Income, Rating, and Student, we fail to reject the null hypothesis at the 5% level of significance. It does not seem as if there is any benefit to adding Ethnicity to the model.**