Final Examination (Take-Home)

Due: Friday, May 4, 2018 at 5:00 pm
Rashid Baishev

1. *Note 1*: Please read the question carefully and answer the all the questions. Show your work as much as possible. Provide all the related outputs (e.g., results, plots, or tables) on 'word' file.
2. *Note 2*: Submit your solution either 'word' or 'pdf' format through Canvas.
3. *Note 3*: Late work will not be accepted!!!. Please keep the due date.

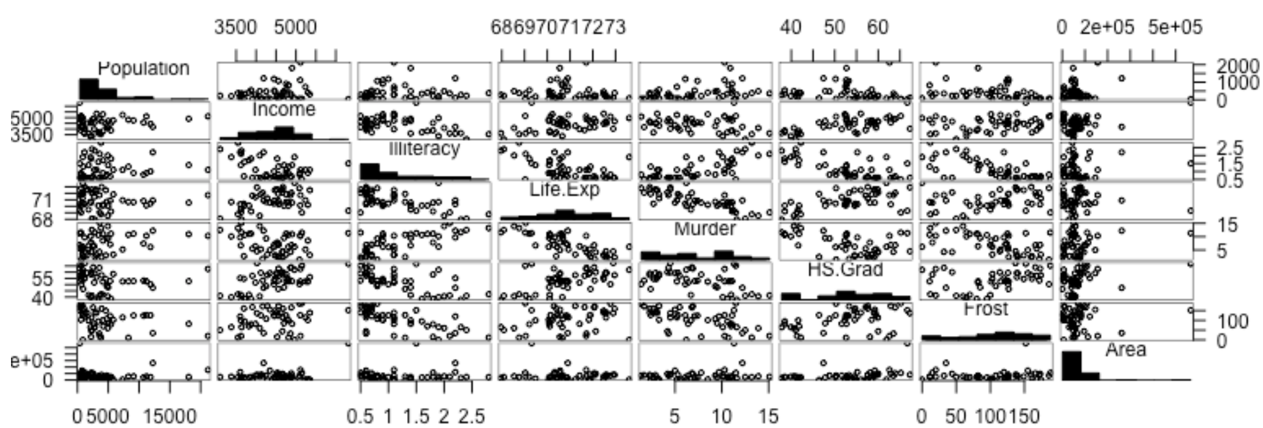
Good Luck!

Question 1

In order to answer this question, we will use the 'state' data set in R. This data was collected from US Bureau of the Census. The variable description is like as following:
We will take life expectancy as the response and the remaining variables as predictors- a fix is necessary to remove spaces in some of the variable names.

Population: Population estimate as of July 1, 1975
Income: Per capita income (1974)
Illiteracy: Illiteracy (1970, percent of the population)
Life Exp: Life expectancy in years (1969–71)
Murder: Murder and non-negligent manslaughter rate per 100,000 population (1976)
HS Grad: Percent high-school graduates (1970)
Frost: Mean number of days with minimum temperature below freezing (1931–1960) in capital or large city
Area: Land area in square miles

1. Create scatterplots of the variables in the 'state' data set and describe **the shape of the distribution of all variables** in the data set (Population, Income, Illiteracy, Murder, HS Grad, Frost, and Area).



Population: unimodal, skewed to the right, has outliers
Income: unimodal, skewed to the left, has outliers
Illiteracy: unimodal, skewed to the right, no outliers
Life Exp: unimodal, slightly symmetric, a little skewed to the left, no outliers
Murder: bimodal with two peaks, not symmetric at center, no outliers
HS Grad: bimodal with two peaks, not symmetric at center, no outliers
Frost: unimodal, skewed to the left, no outliers
Area: unimodal, skewed to the right, multiple outliers

2. Fit a full model and test the hypothesis $H_0: \beta_i = 0$ using t-test statistics and p-values (at $\alpha = .05$).

Hypothesis	Estimated coefficient	P-value	Reject?
$H_0: \beta_{Population} = 0$	**5.180e-05**	**0.0832**	**Fail to Reject**
$H_0: \beta_{Income} = 0$	**-2.180e-05**	**0.9293**	**Fail to Reject**
$H_0: \beta_{Illiteracy} = 0$	**3.382e-02**	**0.9269**	**Fail to Reject**
$H_0: \beta_{Murder} = 0$	**-3.011e-01**	**68e-08**	**Reject**
$H_0: \beta_{HS.Grad} = 0$	**4.893e-02**	**0.0420**	**Reject**
$H_0: \beta_{Frost} = 0$	**-5.735e-03**	**0.0752**	**Fail to Reject**
$H_0: \beta_{Area} = 0$	**-7.383e-08**	**0.9649**	**Fail to Reject**

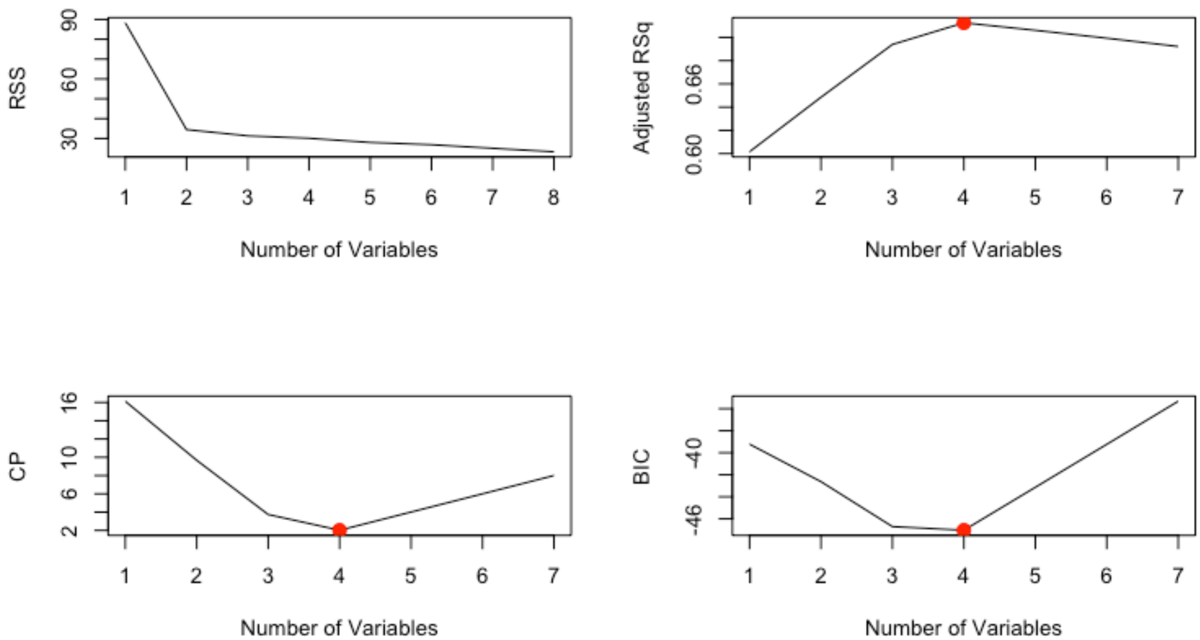
```
Call:
lm(formula = Life.Exp ~ ., data = state1)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.48895 -0.51232 -0.02747  0.57002  1.49447 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
Population   5.180e-05  2.919e-05   1.775   0.0832 .  
Income      -2.180e-05  2.444e-04  -0.089   0.9293    
Illiteracy   3.382e-02  3.663e-01   0.092   0.9269    
Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
HS.Grad      4.893e-02  2.332e-02   2.098   0.0420 *  
Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .  
Area        -7.383e-08  1.668e-06  -0.044   0.9649    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7448 on 42 degrees of freedom
Multiple R-squared:  0.7362,    Adjusted R-squared:  0.6922 
F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

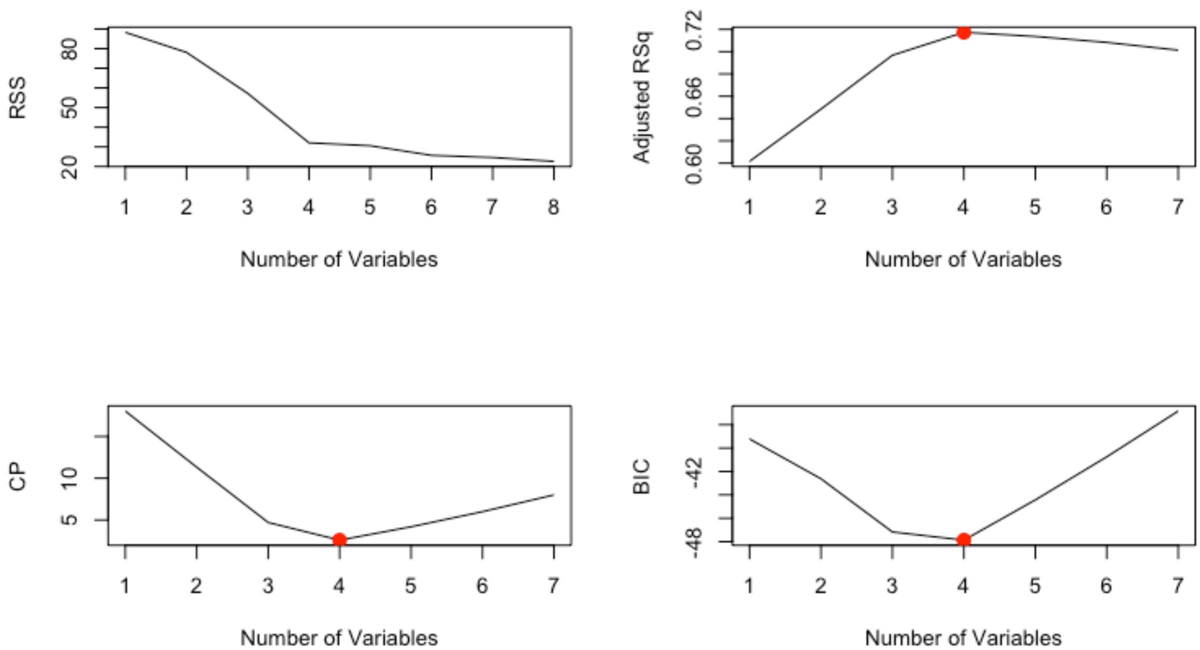
3. Find the best subset model using ‘regsubsets()’ function. . How many variables are in the best model? Which variables are included in the best model? Explain why it is the best model.



	RSS	AdjRSq	Cp	BIC
Best Model # variables	=8	=4	=4	=4

Regression function with 8 number of variables would explain a greater amount of the data, according to RSS plot. Highest AdjRSq equals to 4. Lowest Cp equals to 4 and lowest BIC also equals to 4. Variables included in the best model are the variables which showed significance: Population, Murder, HS. Grad, Frost.

4. Rerun the regsubsets()' function using log-transformed variables. How many variables are in the best model? Which variables are included in the best model?



Based on the graphs above, we can clearly see that the best number of variables is 4 for all models. RSS vs Number of Variables graph shows that after the curve goes below 4 number of variables it does not influence RSS substantially, which is why we choose 4 as best number of variables. For other graphs the best number of variables is still 4. Variables included in the best model are the variables which showed significance: Population, Murder, HS. Grad, Frost.

```
Call:
lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,
    data = state1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47095 -0.53464 -0.03701  0.57621  1.50683

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
Population   5.014e-05  2.512e-05   1.996  0.05201 .
Murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
HS.Grad      4.658e-02  1.483e-02   3.142  0.00297 **
Frost       -5.943e-03  2.421e-03  -2.455  0.01802 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7197 on 45 degrees of freedom
Multiple R-squared:  0.736,     Adjusted R-squared:  0.7126
F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

Best model is showed above.

5. Compare the best model using original variables (in Part 3) and the best model using logged variables (in Part 4) using ANOVA test.

```
Analysis of Variance Table

Model 1: Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
    Frost + Area
Model 2: Life.Exp ~ Population + Murder + HS.Grad + Frost
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     42 23.297
2     45 23.308 -3 -0.010905 0.0066 0.9993
```

According to the result from ANOVA table, there is a very slight improvement when model updated with only significant predictors. However, in the best model, the predictors which had relatively small significance in original model, have more significance in updated model.

6. Interpret the coefficients of the best model in this context.
   Best Updated Model:
   β_1= 5.014e-05. For additional Population unit in July of 1975, we would expect there to be additional 5.014e-05 years in Life Expectancy (1969-71).
   β_1= -3.001e-01. For additional Murder rate percent (1976), we would expect there to be -3.001e-01 years decrease in Life Expectancy (1969-71).
   β_1= 4.658e-02. For additional percent of High School Graduates (1970), we would expect there to be additional 4.658e-02years in Life Expectancy (1969-71).
   β_1= -5.943e-03. For additional Mean number of days with minimum temperature below freezing (1931–1960), we would expect there to be 5.943e-03 years decrease in Life Expectancy (1969-71).

7. Using the better model from the part 5, compute the confidence interval and prediction interval, then compare two intervals. Which interval is wider and why?

```
> predict(best2, new.df1, interval = "confidence")
       fit     lwr      upr
1 73.26022 72.0513 74.46915
> predict(best2, new.df1, interval = "prediction")
       fit      lwr      upr
1 73.26022 71.37272 75.14772
```

   As we can see, prediction interval is wider (75.14772-71.37272=3.775) than confidence interval (74.46915-72.0513=2.41785). Prediction interval is wider because it has more uncertainty than confidence interval.

Question 2

Ridge regression and Lasso regression are a shrinkage model to improve the model prediction accuracy. It improves prediction error by shrinking large regression coefficients and reduce overfitting. Using the MLB dataset which includes player's Name, Team, Position, Height, Weight, and Age.

We may fit a model, e.g., $Weight = \beta_0 + \beta_1 Age + \beta_2 Height$, and compare the results with regularized linear models such as Ridge regression and Lasso regression.

1. Upload the 'data' file. Split the whole data into a training set and a testing set.

2. Fit a regression model in order to predict 'Weight' using two predictors, 'Age' and 'Height'. Then, compute MSE and R^2.

```
Call:
lm(formula = Weight ~ Age + Height, data = data[1:900, ])

Residuals:
    Min      1Q  Median      3Q     Max 
-50.602 -12.399  -0.718  10.913  74.446 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -184.3736    19.4232  -9.492  < 2e-16 ***
Age            0.9799     0.1335   7.341 4.74e-13 ***
Height         4.8561     0.2551  19.037  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.5 on 897 degrees of freedom
Multiple R-squared:  0.3088,	Adjusted R-squared:  0.3072 
F-statistic: 200.3 on 2 and 897 DF,  p-value: < 2.2e-16
```

```
> lm.pred <-  predict(lm.fit, newx = x.test)
> LM.MSE <- mean((y - lm.pred)^2)
> LM.MSE
[1] 305.1995
```

```
> lm.test.r2
[1] 0.2965437
```

3. Fit Ridge regression model with the best lambda (lambda.best1 in R) and write the equation (from the code coef(glmmod)[, 1]). Then, compute the MSE and R^2.

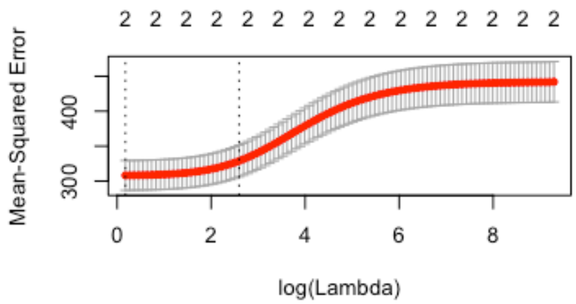
```
> # Part 3
> cv.ridge <-  cv.glmnet(x, y, type.measure="mse", alpha=0, parallel=T)
> ## alpha =1 for lasso only, alpha = 0 for ridge only, and 0<alpha<1 to ble
nd ridge & lasso penalty !!!!
> plot(cv.ridge)
> coef(cv.ridge)
4 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) -30.8276758
(Intercept)   .
Age           0.5609154
Height        2.9359275
> sqrt(cv.ridge$cvm[cv.ridge$lambda == cv.ridge$lambda.1se])
[1] 18.13188
> #plot variable feature coefficients against the shrinkage parameter lambda
.
> glmmod <-glmnet(x, y, alpha = 0)
> plot(glmmod, xvar="lambda")
> grid()
> # report the model coefficient estimates
> coef(glmmod)[, 1]
 (Intercept)  (Intercept)          Age       Height
2.016556e+02 0.000000e+00 8.327372e-37 4.789383e-36
```

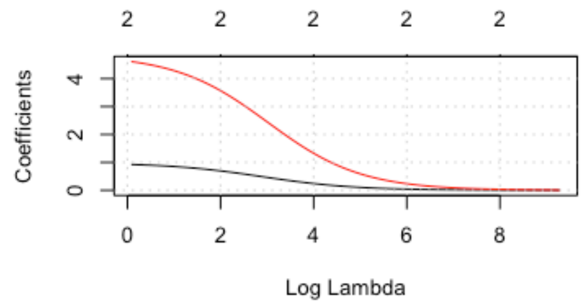
```
> cv.glmmod <- cv.glmnet(x, y, alpha=0)
> plot(cv.glmmod)
> mod.ridge <-  cv.glmnet(x, y, alpha = 0, thresh = 1e-12, parallel = T)
> lambda.best1 <-  mod.ridge$lambda.min
> lambda.best1
[1] 1.192177
> ridge.pred <-  predict(mod.ridge, newx = x.test, s = lambda.best1)
> ridge.MSE <- mean((y.test - ridge.pred)^2)
> ridge.MSE
[1] 264.083
> ridge.test.r2 <-  1 - mean((y.test - ridge.pred)^2)/mean((y.test - mean(y.
test))^2)
> ridge.test.r2
[1] 0.3913134
```

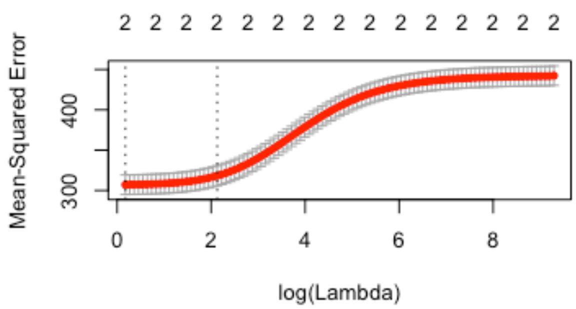
$Weight = (2.0165e+02) + (8.33e-37) * Age + (4.79e-36) * Height$

RIDGE.MSE= 264.083

RIDGE.R2= 0.3913134







4. Fit Lasso regression model with the best lambda (lambda.best2 in R) and write the equation (from the code coef(mod.lasso)[,1]). Then, compute the MSE and R^2.

```
> # Part 4
> mod.lasso <-  cv.glmnet(x, y, alpha = 1, thresh = 1e-12, parallel = T)
> # report the model coefficient estimates
> coef(mod.lasso)[,1]
(Intercept) (Intercept)         Age      Height
-73.3589034   0.0000000   0.3252664   3.6050203
> ## alpha =1 for lasso only, alpha = 0 for ridge only, and 0<alpha<1 for el
astic net, a blend ridge & lasso penalty !!!!
> lambda.best2 <- mod.lasso$lambda.min
> lambda.best2
[1] 0.05933494
> lasso.pred <- predict(mod.lasso, newx = x.test, s = lambda.best2)
> LASSO.MSE <- mean((y.test - lasso.pred)^2)
> LASSO.MSE
[1] 261.8194
> lasso.test.r2 <-  1 - mean((y.test - lasso.pred)^2)/mean((y.test - mean(y.
test))^2)
> lasso.test.r2
[1] 0.3965306
```

$Weight = (-73.3589) + (0.3253) * Age + (3.605) * Height$
LASSO.MSE = 261.8194

LASSO.R2 = 0.3965306

5. What is the best model in terms of MSE and R^2.

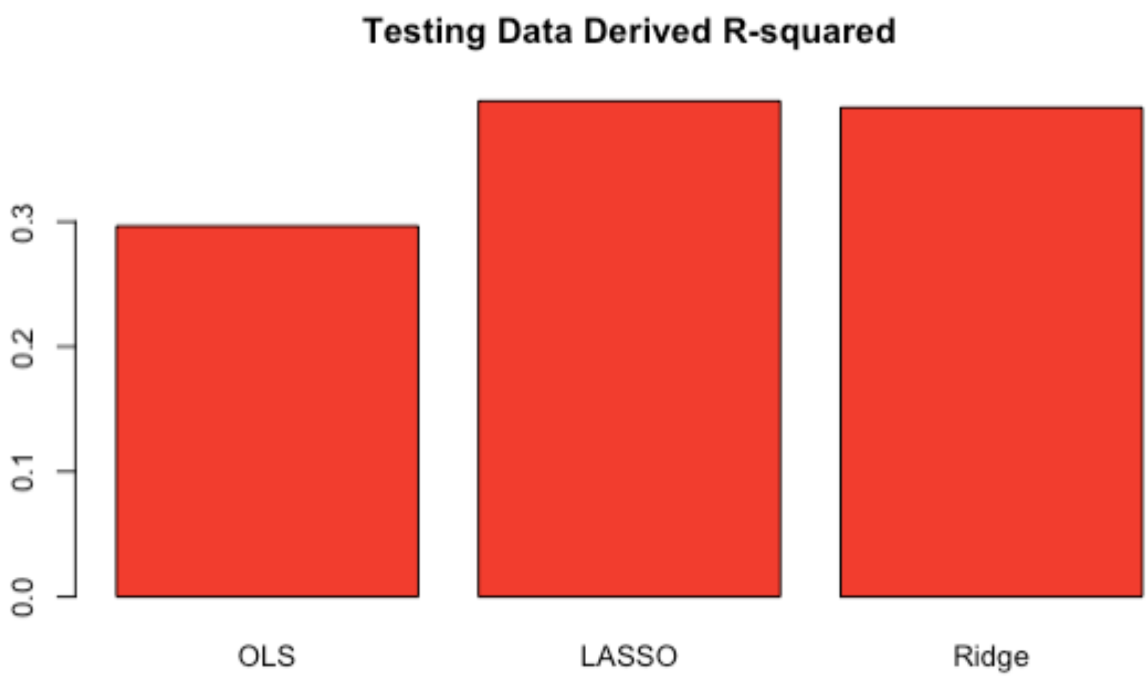


Table: Test Dataset SSE Results

LM	LASSO	Ridge
305.1995	261.8194	264.083

>

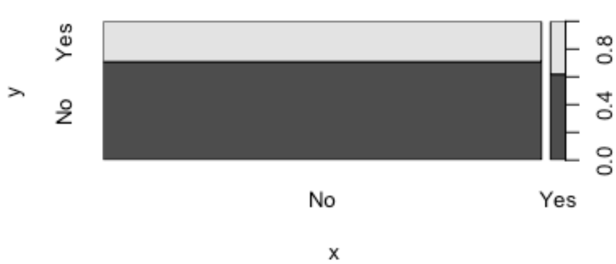
In terms of MSE and R^2, Lasso regression model is better, Lasso has lower MSE (261.8194<264.083) and higher R^2 (39.65>39.13).

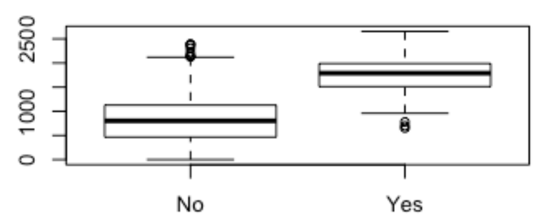
Question 3

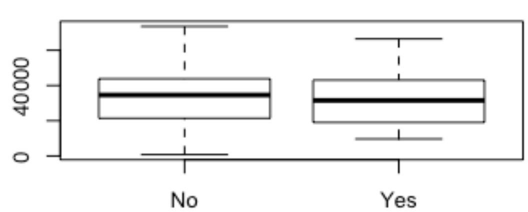
In order to answer this question, we will analyze a Credit Card Default Data in ISLR package. This data is simulated data set containing information on ten thousand customers and includes following 4 variables.

default:	A factor with levels 'No' and 'Yes' indicating whether the customer defaulted on their debt.
student:	A factor with levels 'No' and 'Yes' indicating whether the customer is a student.
balance:	The average balance that the customer has remaining on their credit card after making their monthly payment.
income:	Income of customer

1. By creating plots, investigate the potential associations between outcome variable (=default) and predictors (student, balance, and income). Explain your findings.



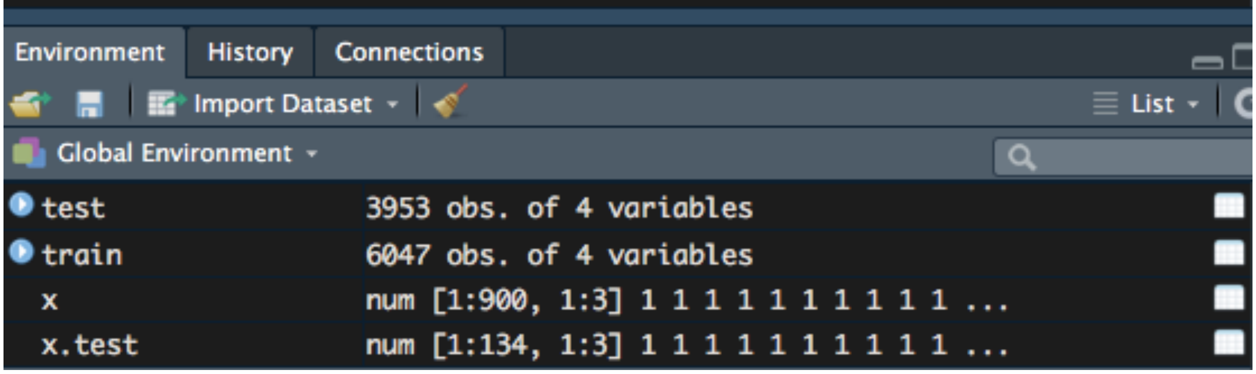




From the plots above, we can conclude several statements. First, the percentage of students who have default status on their Credit Card Account is about 40% and 60% of students does not have default status on the Credit Card Account. Second, people with default status are tend to have higher balance on their accounts, in a range $1500-2000, with some outliers below $1000 mark. Those without default status have lower account balance within range $500-1200 and some outliers above $2000 mark. Third, people with default status and those without it, have about the same income.

2. Split the whole data into a training set and a testing set.

```
> # Part 2
> # Split the whole sample into a training set(60%) and testing set(40%)
> set.seed(123)
> sample <- sample(c(TRUE, FALSE), nrow(default), replace = T, prob = c(0.6,
0.4))
> train <- default[sample, ]
> test <- default[!sample, ]
> View(test)
> View(train)
> View(x)
> View(x.test)
>
```



3. Fit **Simple Logistic Regression models** using student, balance, and income separately. Report e^{β} (from exp(coef(model))) and compare the AIC values.

```
> summary(model1)

Call:
glm(formula = default ~ balance, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2905  -0.1395  -0.0528  -0.0189   3.3346

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.101e+01  4.887e-01  -22.52   <2e-16 ***
balance      5.669e-03  2.949e-04   19.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1723.03  on 6046  degrees of freedom
Residual deviance:  908.69  on 6045  degrees of freedom
AIC: 912.69

Number of Fisher Scoring iterations: 8

> # Assession coefficients
> tidy(model1)
         term      estimate   std.error statistic       p.value
1 (Intercept) -11.006277528 0.488739437 -22.51972 2.660162e-112
2     balance   0.005668817 0.000294946  19.21985  2.525157e-82
> exp(coef(model1))
 (Intercept)      balance
1.659718e-05 1.005685e+00
```

$e^{\beta} = 1.0057$

AIC = 912.69

```
> model2 <- glm(default ~ student, family = "binomial", data = train)
> summary(model2)

Call:
glm(formula = default ~ student, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2951  -0.2951  -0.2376  -0.2376   2.6764

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.55341    0.09337 -38.059  < 2e-16 ***
studentYes   0.44134    0.14927   2.957  0.00311 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1723.0  on 6046  degrees of freedom
Residual deviance: 1714.6  on 6045  degrees of freedom
AIC: 1718.6

Number of Fisher Scoring iterations: 6

> tidy(model2)
         term   estimate  std.error statistic     p.value
1 (Intercept) -3.5534091 0.09336545 -38.05914 0.000000000
2  studentYes  0.4413379 0.14927208   2.95660 0.003110511
> exp(coef(model2))
(Intercept)  studentYes
 0.02862688  1.55478593
```

$e^{\beta} = 1.5548$

AIC = 1718.6

```
> model3 <- glm(default ~ income, family = "binomial", data = train)
> summary(model3)

Call:
glm(formula = default ~ income, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3007  -0.2707  -0.2527  -0.2408   2.7295

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.066e+00  1.883e-01 -16.278   <2e-16 ***
income      -1.033e-05  5.487e-06  -1.883   0.0598 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1723.0  on 6046  degrees of freedom
Residual deviance: 1719.5  on 6045  degrees of freedom
AIC: 1723.5

Number of Fisher Scoring iterations: 6

> tidy(model3)
         term      estimate    std.error statistic      p.value
1 (Intercept) -3.065692e+00 1.883349e-01 -16.27788 1.416988e-59
2      income -1.032929e-05 5.486859e-06  -1.88255 5.976131e-02
> exp(coef(model3))
(Intercept)      income
 0.04662156  0.99998967
```

$e^{\beta} = 1.0000$

AIC = 1723.5

Model 1 has the lowest AIC (=912.69)

4. Using simple logistic regression models in part 3, make a prediction for the new data and fill out a table below.

	Prediction 1	Prediction 2
$default = \beta_0 + \beta_1 * student$	= 0.04261206	= 0.02783019
$default = \beta_0 + \beta_1 * balance$	= 0.004785057	= 0.582089269
$default = \beta_0 + \beta_1 * income$	=0.04452284	= 0.04451405

```
> # Part 4
> # Making prediction
> predict(model1, data.frame(balance = c(1000, 2000)), type = "response")
          1           2 
0.004785057 0.582089269 
> predict(model2, data.frame(student = factor(c("Yes", "No"))), type = "resp
onse")
         1          2 
0.04261206 0.02783019 
> predict(model3, data.frame(income = c(50, 70)), type = "response")
         1          2 
0.04452284 0.04451405 
```

5. Fit **Multiple Logistic Regression models** including student, balance, and income together. Report all e^{β}s(from exp(coef(model))) and AIC.

```
Call:
glm(formula = default ~ balance + income + student, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4556  -0.1344  -0.0499  -0.0174   3.4155

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.091e+01  6.481e-01 -16.830  < 2e-16 ***
balance      5.907e-03  3.102e-04  19.040  < 2e-16 ***
income      -5.013e-06  1.079e-05  -0.465  0.64212
studentYes  -8.095e-01  3.133e-01  -2.584  0.00978 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1723.03  on 6046  degrees of freedom
Residual deviance:  895.02  on 6043  degrees of freedom
AIC: 903.02

Number of Fisher Scoring iterations: 8

> tidy(model4)
         term      estimate    std.error   statistic      p.value
1 (Intercept) -1.090704e+01 6.480739e-01 -16.8299277 1.472817e-63
2     balance  5.907134e-03 3.102425e-04  19.0403764 7.895817e-81
3      income -5.012701e-06 1.078617e-05  -0.4647343 6.421217e-01
4  studentYes -8.094789e-01 3.133150e-01  -2.5835947 9.777661e-03
> exp(coef(model4))
 (Intercept)      balance       income   studentYes
1.832881e-05 1.005925e+00 9.999950e-01 4.450899e-01
```

$Balance: e^{\beta} = 1.0059$

$Income: e^{\beta} = 1.0000$

$StudentYes: e^{\beta} = 0.4509$

AIC = 903.02

6. Using a multiple logistic regression models in part 5, make a prediction for the new data and fill out a table below.

	Prediction 1	Prediction 2
$default = \beta_0 + \beta_1 * student + \beta_2 * balance + \beta_3 * income$	=0.05437124	=0.11440288

```
> # Part 6
> new.df <- tibble(balance = 1500, income = 40, student = c("Yes", "No"))
> predict(model4, new.df, type = "response")
         1          2
0.05437124 0.11440288
```

7. Evaluate models and pick the best model using ANOVA test, R-square, Examining residuals (Cook's distance), and MSE.
ANOVA: 2 stars significance level on model 4 (balance+income+student) indicates that we have enough evidence to consider model 4 the best model and keep it.

```
> anova(model1, model2, model3, test = "Chisq")
Analysis of Deviance Table

Model 1: default ~ balance
Model 2: default ~ student
Model 3: default ~ income
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      6045     908.69
2      6045    1714.59  0  -805.90
3      6045    1719.45  0    -4.86
> anova(model1, model4, test = "Chisq")
Analysis of Deviance Table

Model 1: default ~ balance
Model 2: default ~ balance + income + student
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      6045     908.69
2      6043     895.02  2   13.668 0.001076 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R-square: model 4 is the best, it has the highest R-square.

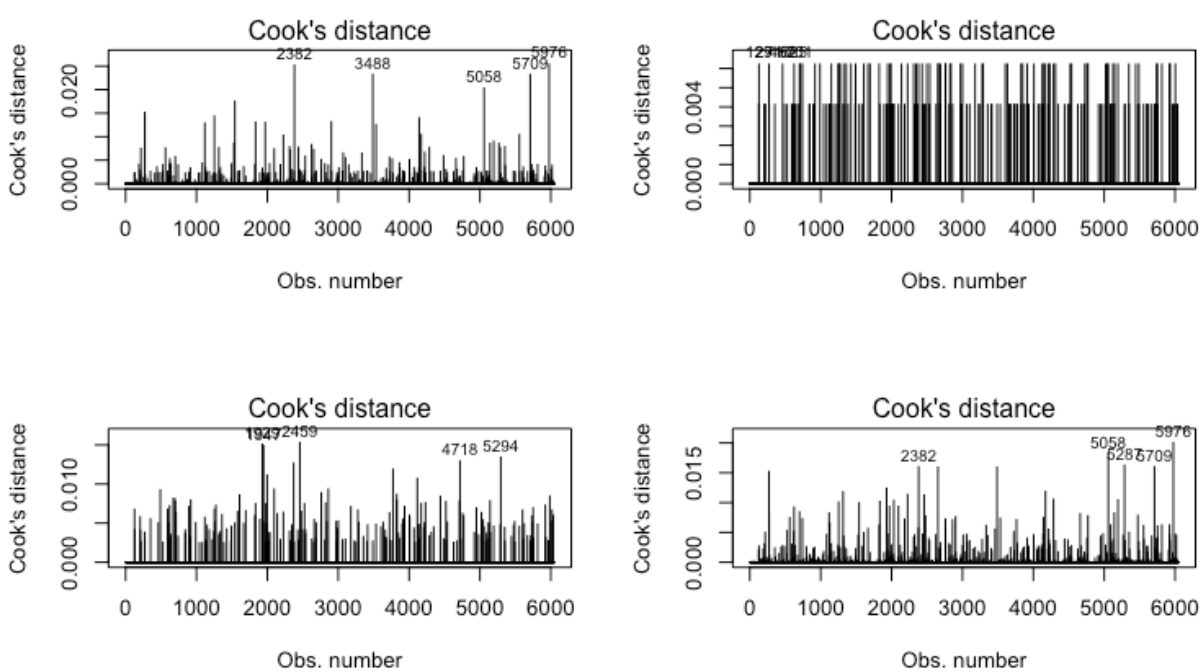
```
$model1
 McFadden
0.4726215

$model2
    McFadden
0.004898314

$model3
    McFadden
0.002075557

$model4
 McFadden
0.4805543
```

Residuals (Cook's Distance): models 1, 3 and 4 are good, when the model 2 shows a lot of outliers. Models 1, 3 and 4 have much less outliers.



```
> # Cook's distance
> plot(model4, which = 4, id.n = 5)
> model4_data %>%
+   top_n(5, .cooksd)
  default   balance    income student   .fitted   .se.fit    .resid
1      No 2388.1740  7832.136     Yes  2.351488 0.2752552 -2.210181
2     Yes 1013.2169 19651.262     Yes -5.829813 0.2704579  3.415479
3     Yes 1323.6281 18820.795      No -3.182531 0.2648255  2.538966
4     Yes  961.7327 27600.416      No -5.364305 0.2541166  3.276881
5      No 2391.0077 50302.910      No  2.964813 0.2965382 -2.455646
          .hat    .sigma    .cooksd .std.resid index
1 0.0060158273 0.3838222 0.01598508  -2.216859  2382
2 0.0002137641 0.3823634 0.01819350   3.415844  5058
3 0.0026823570 0.3834881 0.01625343   2.542378  5287
4 0.0002995307 0.3825639 0.01600773   3.277372  5709
5 0.0041018328 0.3835762 0.02004888  -2.460697  5976
```

MSE: model 1is the best, it has the lowest MSE = 2.78%.

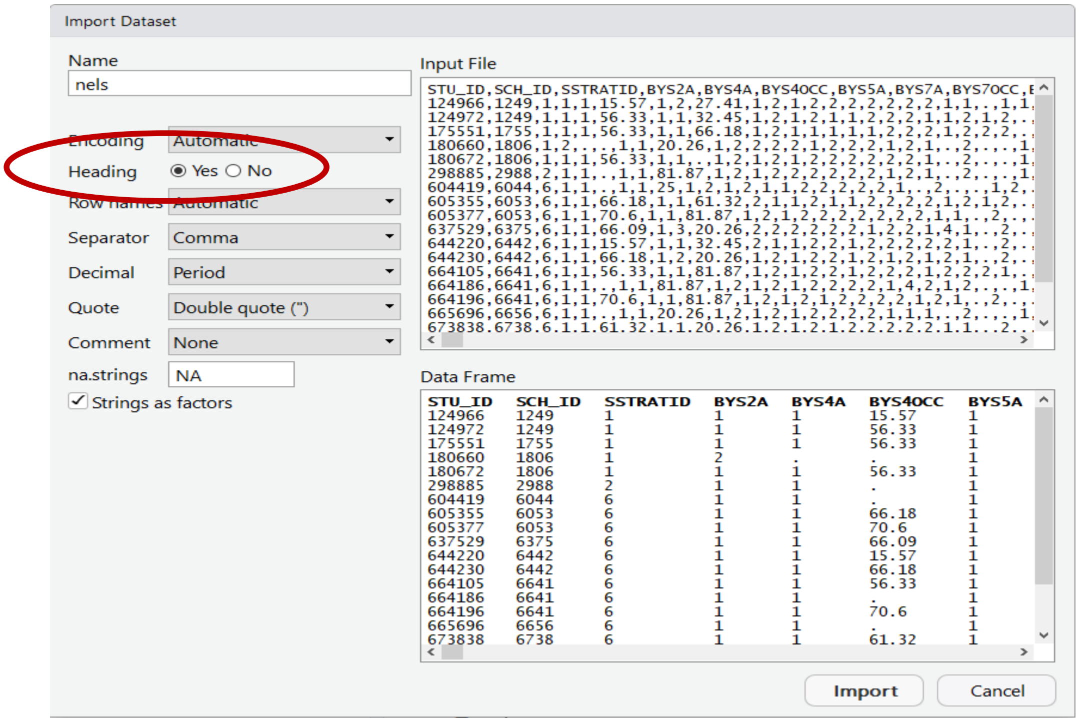
```
# A tibble: 1 x 4
  m1.error m2.error m3.error m4.error
     <dbl>    <dbl>    <dbl>    <dbl>
1   0.0278   0.0349   0.0349   0.0281
```

Question 4

Data for this question were taken from a subset of the National Education Longitudinal Study of 1988 (NELS), provided by Keith (2006). The variables used for this analysis are listed in the table below. We only used the observations with values for each of the variables. The outcome is the number of times a student cut/skipped class (skips), placed into one of five categories.

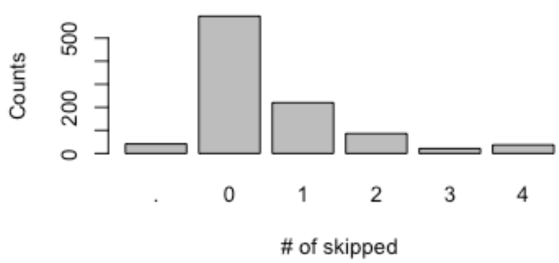
Variable (Name in dataset)	Description	Values
Skips (F1S10B)	Number of times student cut/skipped class (Outcome variable)	0 = 0 times; 1 = 1- 2 times; 2= 3- 6 times; 3 = 7- 9 times; 4= ≥ 10 times
College (F1S51)	Plan on going to college	0 = No; 1 = Yes
Male (BYS12)	Sex	0 = Female; 1 = Male
Race (BYS31A)	Self- described race	0 = White; 1 = Asian; 2 = Hispanic; 3 = Black; 4 = Native American
Achievement (BYTEXCOMP)	Standardized reading and math achievement test composite	Continuous
Self Concept (BYCNCPT1)	The positive self concept, which is a composite of four items	Continuous
SES (BYSES)	Socioeconomic status composite	Continuous

1.
2. Import dataset, nels.dat, and clean the data. Please make sure to check Heading is Yes.



1. Clean the data set; filling in a few missing values and deleting the unnecessary variable.

2. Create plot of the outcome variable, skipped. Based on the plot, propose a model in order to analyze the data and explain why you suggest the model in order to analyze this dataset.



Since the outcome variable is a count variable, I believe, the Poisson model would be the best to predict the number of times a student cut/skipped class in terms of given predictors. Based on the plot above, we also see a lot of “0” values.

3. Discuss your model (model equation, interpret coefficient, evaluate the model using AIC, BIC, and residual plot, and so on).

```
Call:
glm(formula = skipped ~ male + race + college + self.con1.m +
    ses.m + achievement.m, family = poisson, data = count.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2111  -0.3802  -0.2975   0.2621   1.9546

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.409e-01  1.117e-01   3.052  0.00227 **
male             -3.726e-03  2.184e-02  -0.171  0.86456
raceasian         8.895e-03  8.771e-02   0.101  0.91923
racehispanic      9.449e-02  7.009e-02   1.348  0.17761
raceblack         4.159e-03  7.637e-02   0.054  0.95658
racenat.american  7.314e-04  1.132e-01   0.006  0.99484
college1          6.020e-01  1.097e-01   5.486 4.11e-08 ***
self.con1.m      -1.456e-04  7.706e-04  -0.189  0.85012
ses.m            -3.003e-02  3.224e-02  -0.931  0.35162
achievement.m    -1.042e-04  9.311e-05  -1.120  0.26288
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 295.13  on 870  degrees of freedom
Residual deviance: 256.46  on 861  degrees of freedom
AIC: 2686.9

Number of Fisher Scoring iterations: 4
```

```
> # AIC values
> AIC(model1)
[1] 2372
> AIC(model3)
[1] 2686.916
> # BIC values
> BIC(model1)
[1] 2424.466
> BIC(model3)
[1] 2734.612
```

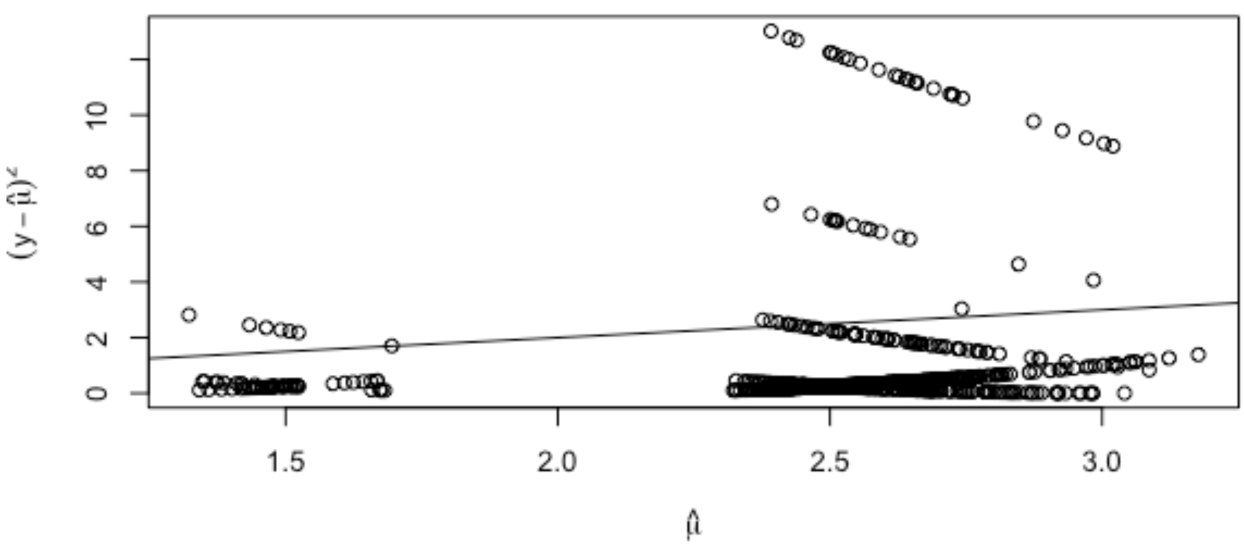
From the summary above, we can see that only one predictor, whether a student is planning on going to college, is highly significant.

$skipped = \beta_0 + \beta_1 * college1$

$\beta_1 = 0.602$. For every additional student who plans to go to college, the expected number of skipped classes increases by 0.602, when other variables are held constant.

Since this model includes count as an outcome variable and not binary, we exclude logistic model as an alternative option. By comparing AIC and BIC for models 1 and 3, we can see that model 1 has lower results, which makes model 1 better. However, considering that our outcome variable is count and not continuous, the Poisson model is the best model in this case.

4. What is a strength of your model? What is a limitation of your model?



On the plot above, we can see overdispersion. Especially on the part from 2.3 to 3.2 (mean).

```
> (dp <- sum(residuals(model3,type="pearson")^2)/model3$df.res)
[1] 0.346414
> summary(model3,dispersion=dp)

Call:
glm(formula = skipped ~ male + race + college + self.con1.m +
    ses.m + achievement.m, family = poisson, data = count.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2111  -0.3802  -0.2975   0.2621   1.9546

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.3409030  0.0657383   5.186 2.15e-07 ***
male             -0.0037262  0.0128572  -0.290   0.7720
raceasian         0.0088946  0.0516262   0.172   0.8632
racehispanic      0.0944871  0.0412508   2.291   0.0220 *
raceblack         0.0041586  0.0449516   0.093   0.9263
racenat.american  0.0007314  0.0666078   0.011   0.9912
college1          0.6020254  0.0645858   9.321  < 2e-16 ***
self.con1.m      -0.0001456  0.0004536  -0.321   0.7482
ses.m            -0.0300319  0.0189768  -1.583   0.1135
achievement.m    -0.0001042  0.0000548  -1.902   0.0571 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 0.346414)

    Null deviance: 295.13  on 870  degrees of freedom
Residual deviance: 256.46  on 861  degrees of freedom
AIC: 2686.9

Number of Fisher Scoring iterations: 4
```

Strengths: Poisson model is best for count variables;
Limitations: There are lots of "0" values in the data set. Possibility of over-dispersion if dispersion value is > 1, in out case it's less than 1 (=0.346414). We can conclude that there is a slight overdispersion.

Question 5

Briefly discuss your project using your words (no more than 1 page); topics, proposed model (weakness and strength of your model), and application of your project (how can apply your finding? Which area? Can we apply your finding into another context?).

Since both of us, me and Hudson are basketball fans, we were curious what factors really determine the eight-digit salary of NBA players. First of all, we decided to use traditional simple linear regression and multiple linear regression models with six, potentially highly significant predictors, as we thought. Our assumption was correct, so we decided to go further. We wanted to avoid potential of multicollinearity from multiple regression and decided to try another model too. In order to maximize the prediction power, using minimum number of predictor variables, we run stepwise regression, because we had multiple independent variables.

After evaluating models using parameters like, AIC, BIC, Cp, R-squared and MSE, we concluded that, number of points per game is not the only factor that determines NBA player's salary. We discovered that salary is determined on factors such as; position, age, assists, and points. In addition, we were surprised that average number of blocks and steals per season are irrelevant in determining NBA player's salary.

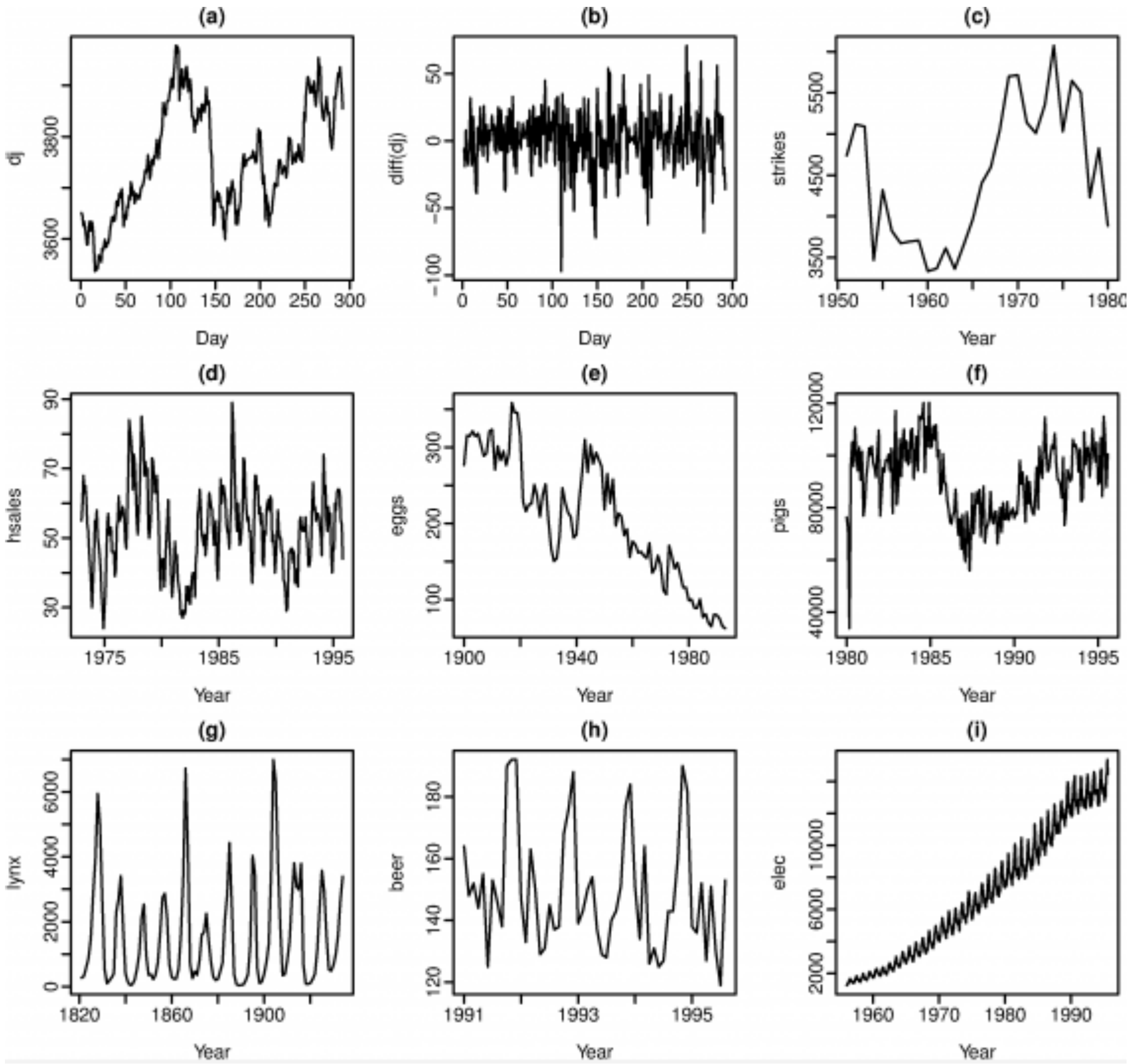
I believe that we can apply our approach to determine salary of players in other competitive sports games similar to basketball in nature. For instance, salary of football players can be determined using our approach. However, every sport is different, there might be some significant variables we don't know about before we run regression model.

Thank you so much for your hard work during this semester.

Dream Big.Sparkle More. Shine Bright

EXTRA CREDITS

1. What are the four components of time series data? Explain each component of the time series data.
   Trend: persistent upward or downward pattern in a time series.
   Seasonal: variation dependent on the time of year; each year shows same pattern.
   Cyclical: up & down movement repeating over long time frame; each year does not show the same pattern.
   Noise or random fluctuations: follow no specific pattern; short duration and non-repeating.
2. Which plot contains trend? Which plot contains random component?
   G, H, I, D and E plots contain trend. A, B, C and F plots contain random component.



3. What is a panel data? What are advantages to use a panel data?
   Panel data refers to the data the combines cross-section and time-series data.
   Advantages to use panel data:
   1. Takes explicit account of individual-specific heterogeneity.
   2. When combining data in two dimensions, panel data gives more data variation, less collinearity and more degrees of freedom.
   3. Better suited than cross-sectional data for studying the dynamics of change.
   4. Better at detecting and measuring effects that cannot be observed in either cross-section or time-series.
   5. Enables the study of more complex behavioral models.
   6. Can minimize the effects of aggregation bias, from aggregating firms into broad groups.