New York University Shanghai

Problem Set 1

BUSF-SHU 210: Business Analytics (Spring 2018)

Due at 10:00PM on Sunday, Feb 11

Instructions

Whenever you are asked to build a linear regression model, you should write out the model in the form of

$$Y_i \approx \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij},$$

and clearly explain what are the variables Y_i, X_{i1}, X_{i2}, …, X_{ip}. After running the model to get the estimated values of $\hat{\beta}_0$, $\hat{\beta}_1$, …, $\hat{\beta}_p$, please report the results in the form of

$$Y_i \approx \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ij},$$

and interpret what the coefficients $\hat{\beta}_0$, $\hat{\beta}_1$, …, $\hat{\beta}_p$ mean in the problem of interest.

1. Greenhouse Effect

You will attempt to study the relationship between average global temperature and several other factors. The file `greenhouse.csv` contains the climate data from 1983 to 2008. The variables are listed as follows:

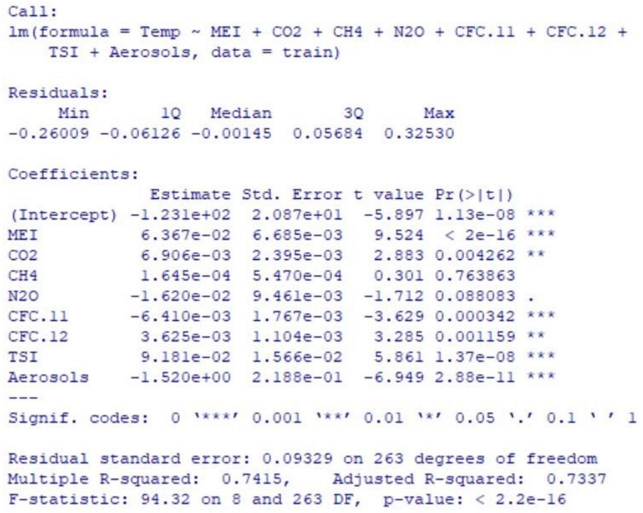
- $Year$: the observation year.
- $Month$: the observation month.
- $Temp$: the difference between the average global temperature and a reference value in that month.
- $CO2$, $N2O$, $CH4$, $CFC.11$, $CFC.12$: concentrations of carbon dioxide (CO2), nitrous oxide (N2O), methane (CH4), trichlorofluoromethane (CCl3F; commonly referred to as CFC-11) and dichlorodifluoromethane (CCl2F2; commonly referred to as CFC-12), respectively, in the atmosphere. $CO2$, $N2O$ and $CH4$ are expressed in ppmv (parts per million by volume, i.e., 397 ppmv of CO2 means that CO2 constitutes 397 millionths of the total volume of the atmosphere)
- $CFC.11$ and $CFC.12$ are expressed in ppbv (parts per billion by volume).
- $Aerosols$: the mean stratospheric aerosol optical depth at 550 nm. This variable is linked to volcanoes.

- TSI: the total solar irradiance (TSI) in W/m^2 (the rate at which the sun's energy is deposited per unit area).
- MEI: multivariate El Nino Southern Oscillation index (MEI), a measure of the strength of the El Nino/La Nina-Southern Oscillation.

(a) Read the dataset `greenhouse.csv` into R. Then, split the data into a training set, consisting of all the observations up to and including 2005, and a testing set consisting of the remaining years. A training set is the data that will be used to build the model, and a testing set is the data to test the predictive ability.

Build a linear regression model to establish the association between the dependent variable $Temp$, and the independent variables MEI, $CO2$, $CH4$, $N2O$, $CFC.11$, $CFC.12$, TSI, and $Aerosols$. Use the training set to build the model.

Please show a screen shot of your linear regression model using the "summary" function and interpret the results. Clearly state the significance, the sign, and the magnitude of the association between the dependent variable and each independent variable. Figure 1



```
Call:
lm(formula = Temp ~ MEI + CO2 + CH4 + N2O + CFC.11 + CFC.12 +
    TSI + Aerosols, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26009 -0.06126 -0.00145  0.05684  0.32530

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.231e+02  2.087e+01  -5.897 1.13e-08 ***
MEI          6.367e-02  6.685e-03   9.524  < 2e-16 ***
CO2          6.906e-03  2.395e-03   2.883 0.004262 **
CH4          1.645e-04  5.470e-04   0.301 0.763863
N2O         -1.620e-02  9.461e-03  -1.712 0.088083 .
CFC.11      -6.410e-03  1.767e-03  -3.629 0.000342 ***
CFC.12       3.625e-03  1.104e-03   3.285 0.001159 **
TSI          9.181e-02  1.566e-02   5.861 1.37e-08 ***
Aerosols    -1.520e+00  2.188e-01  -6.949 2.88e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09329 on 263 degrees of freedom
Multiple R-squared:  0.7415,    Adjusted R-squared:  0.7337
F-statistic: 94.32 on 8 and 263 DF,  p-value: < 2.2e-16
```

Figure 1: Question 1(a)

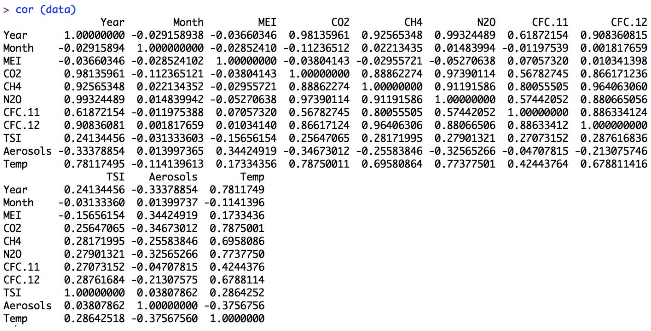
Temp = -123.1 + 0.06367*MEI + 0.006906 * CO2 + 0.0001645 * CH4 - 0.01620 * N2O C 0.00641 * CFC.11 + 0.003625 * CFC.12 + 0.09181 * TSI -1.52 * Aerosols

- MEI : significant, positive, one unit change in MEI is associate with 0.064 increase in Temp
- $CO2$: significant, positive, one unit change in CO2 is associate with 0.007 increase in Temp
- $CH4$: insignificant, positive, one unit change in CH4 is associate with 0.0002 increase in Temp

- *N2O* : insignificant, negative, one unit change in NO2 is associate with 0.016 decrease in Temp
- *CFC*.11 : significant, negative, one unit change in CFC.11 is associate with 0.006 decrease in Temp
- *CFC*.12 : significant, positive, one unit change in CFC.12 is associate with 0.004 increase in Temp
- *TSI* : significant, positive, one unit change in TSI is associate with 0.092 increase in Temp
- *Aerosols* : significant, negative, one unit change in Aerosols is associate with 1.520 decrease in Temp

(b) It is widely believed that nitrous oxide and CFC-11 are greenhouse gases. From the regression analysis in part (a), is there anything counter-intuitive about the regression results? Please provide some explanation supported by quantitative evidence for such counter-intuitive phenomenon. (*Hint*: You may want to check the correlations between different independent variables.)

In the regression model, the coefficients for N2O and CFC.11 are negative, meaning that the higher N2O and CFC.11 are, the lower the temperature is, which is counter-intuitive. In addition, if we use p value to determine significance, N2O is not that significant while it actually is a greenhouse gas. Both problems are due to the multicollinearity between the selected variables. Figure 2



```
> cor (data)
                Year        Month         MEI         CO2         CH4         N2O      CFC.11       CFC.12
Year      1.00000000 -0.029158938 -0.03660346  0.98135961  0.92565348  0.99324489  0.61872154  0.908360815
Month    -0.02915894  1.000000000 -0.02852410 -0.11236512  0.02213435  0.01483994 -0.01197539  0.001817659
MEI      -0.03660346 -0.028524102  1.00000000 -0.03804143 -0.02955721 -0.05270638  0.07057320  0.010341398
CO2       0.98135961 -0.112365121 -0.03804143  1.00000000  0.88862274  0.97390114  0.56782745  0.866171236
CH4       0.92565348  0.022134352 -0.02955721  0.88862274  1.00000000  0.91191586  0.80055505  0.964063060
N2O       0.99324489  0.014839942 -0.05270638  0.97390114  0.91191586  1.00000000  0.57442052  0.880665056
CFC.11    0.61872154 -0.011975388  0.07057320  0.56782745  0.80055505  0.57442052  1.00000000  0.886334124
CFC.12    0.90836081  0.001817659  0.01034140  0.86617124  0.96406306  0.88066506  0.88633412  1.000000000
TSI       0.24134456 -0.031333603 -0.15656154  0.25647065  0.28171995  0.27901321  0.27073152  0.287616836
Aerosols -0.33378854  0.013997365  0.34424919 -0.34673012 -0.25583846 -0.32565266 -0.04707815 -0.213075746
Temp      0.78117495 -0.114139613  0.17334356  0.78750011  0.69580864  0.77377501  0.42443764  0.678811416
                 TSI    Aerosols       Temp
Year      0.24134456 -0.33378854  0.7811749
Month    -0.03133360  0.01399737 -0.1141396
MEI      -0.15656154  0.34424919  0.1733436
CO2       0.25647065 -0.34673012  0.7875001
CH4       0.28171995 -0.25583846  0.6958086
N2O       0.27901321 -0.32565266  0.7737750
CFC.11    0.27073152 -0.04707815  0.4244376
CFC.12    0.28761684 -0.21307575  0.6788114
TSI       1.00000000  0.03807862  0.2864252
Aerosols  0.03807862  1.00000000 -0.3756756
Temp      0.28642518 -0.37567560  1.0000000
```

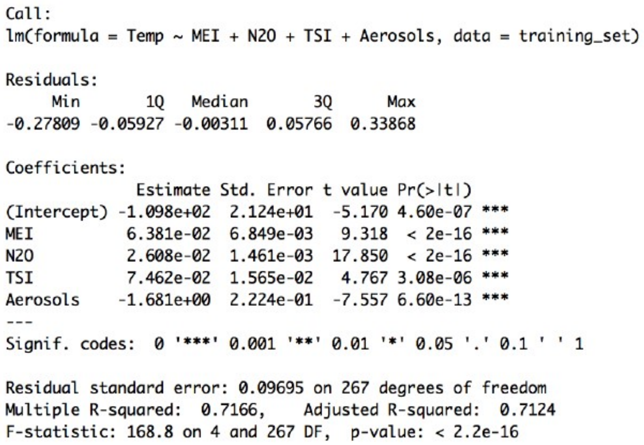
Figure 2: Question 1(b)

Here we regard correlation as high if the absolute value of correlation > 0.7, so N2O is highly correlated with CO2, CH4 and CFC.12; CFC.11 is highly correlated with CH4 and CFC.12.

(c) Based on your analysis of parts (a) and (b), build another (possibly simpler with fewer covariates) linear regression model. Test both the full model (i.e., the one developed in part (a)) and the new model using the testing data. Report the respective out-of-sample R^2's for both models. Which one performs better?

Since there are strong correlation among the variables N2O, CO2, CH4, CFC.11 and CFC.12, we would only select one of the above variables. The new regression model (see Figure 3) is as follows:

Temp = - 109.8 + 0.0638 * MEI + 0.02608 * N2O + 0.07462 * TSI -1.681 * Aerosols



```
Call:
lm(formula = Temp ~ MEI + N2O + TSI + Aerosols, data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.27809 -0.05927 -0.00311  0.05766  0.33868 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.098e+02  2.124e+01  -5.170 4.60e-07 ***
MEI          6.381e-02  6.849e-03   9.318  < 2e-16 ***
N2O          2.608e-02  1.461e-03  17.850  < 2e-16 ***
TSI          7.462e-02  1.565e-02   4.767 3.08e-06 ***
Aerosols    -1.681e+00  2.224e-01  -7.557 6.60e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09695 on 267 degrees of freedom
Multiple R-squared:  0.7166,	Adjusted R-squared:  0.7124 
F-statistic: 168.8 on 4 and 267 DF,  p-value: < 2.2e-16
```

Figure 3: Question 1(c)

Out-of-sample R-squared for model 1 = 0.7631
Out-of-sample R-squared for model 2 = 0.6582
Model 1 performs better, partly because model 2 loses quite a lot information by eliminating 4 variables.

2. Student Assessment Scores

In this problem, you will predict the reading scores of students on the 2009 Assessment.

The datasets `Assessment_train.csv` and `Assessment_test.csv` contain information about the students taking the exam.

Each row in the datasets `Assessment_train.csv` and `Assessment_test.csv` represents one student taking the Assessment. The datasets have the following variables:

- *grade*: The grade in school of the student (most 15-year-olds in America are in 10th grade)
- *male*: Whether the student is male (1/0)
- *raceeth*: The race/ethnicity composite of the student
- *preschool*: Whether the student attended preschool (1/0)
- $expectBachelors$: Whether the student expects to obtain a bachelor's degree (1/0)
- $motherHS$: Whether the student's mother completed high school (1/0)
- $motherBachelors$: Whether the student's mother obtained a bachelor's degree (1/0)
- $motherWork$: Whether the student's mother has part-time or full-time work (1/0)
- $fatherHS$: Whether the student's father completed high school (1/0)
- $fatherBachelors$: Whether the student's father obtained a bachelor's degree (1/0)
- $fatherWork$: Whether the student's father has part-time or full-time work (1/0)
- $selfBornUS$: Whether the student was born in the United States of America (1/0)
- $motherBornUS$: Whether the student's mother was born in the United States of America (1/0)
- $fatherBornUS$: Whether the student's father was born in the United States of America (1/0)
- $englishAtHome$: Whether the student speaks English at home (1/0)
- $computerForSchoolwork$: Whether the student has access to a computer for schoolwork (1/0)
- $read30MinsADay$: Whether the student reads for pleasure for 30 minutes/day (1/0)
- $minutesPerWeekEnglish$: The number of minutes per week the student spend in English class
- $studentsInEnglish$: The number of students in this student's English class at school
- $schoolHasLibrary$: Whether this student's school has a library (1/0)

- *publicSchool*: Whether this student attends a public school (1/0)
- *urban*: Whether this student's school is in an urban area (1/0)
- *schoolSize*: The number of students in this student's school
- *readingScore*: The student's reading score, on a 1000-point scale

There are some missing values in this data set, which we need to remove before building the linear regression model. Use the function `na.omit()` to remove the missing values in both the training data and the testing data. To apply this function, you may type: `AssesmentTrain = na.omit(AssessmentTrain)` and `Assessment = na.omit(AssessmentTest)` in the R console.

(a) Factor variables are variables that take on a discrete set of values, like the "Region" variable in the WHO data set from the exercises of Session 1.An ordered factor has a natural ordering between the levels (an example would be the classifications "large," "medium," and "small"). Which of the variables in the Assessment data set are unordered factors with at least three levels? Which of the variables are ordered factors with at least three levels?

From summary(AssessmentTrain), we can see that raceeth is the unordered factor with at least three levels, and grade is the ordered factor with at least three levels. (minutesPerWeekEnglish, studentsInEnglish, schoolSize, and readingscore could be counted as ordered factors as well)

(b) To include unordered factors in a linear regression model, we define one level as the "reference level" and add a binary variable for each of the remaining levels. In this way, a factor with n levels is replaced by $n-1$ binary variables. The reference level is typically selected to be the most frequently occurring level in the dataset.

As an example, consider the unordered factor variable "place", with levels "Shanghai", "Beijing", and "Guangzhou". If "Beijing" were the reference level, then we would add binary variables "placeShanghai" and "placeGuangzhou" to a linear regression problem. All Shanghai examples would have placeShanghai=1 and placeGuangzhou=0. All Guangzhou examples would have placeShanghai=0 and placeGuangzhou=1. All Beijing examples would have placeShanghai=0 and placeGuangzhou=0.

Now, consider the variable "raceeth" in our problem, which has levels "American Indian/Alaska Native", "Asian", "Black", "Hispanic", "More than one race", "Native Hawaiian/Other Pacific Islander", and "White". Because it is the most common in the data set, we will select White as the reference level. Which binary variables should be included in the regression model? For a student who is Asian, which of the binary variables you created above should be set to 0, which of them should be set to 1?

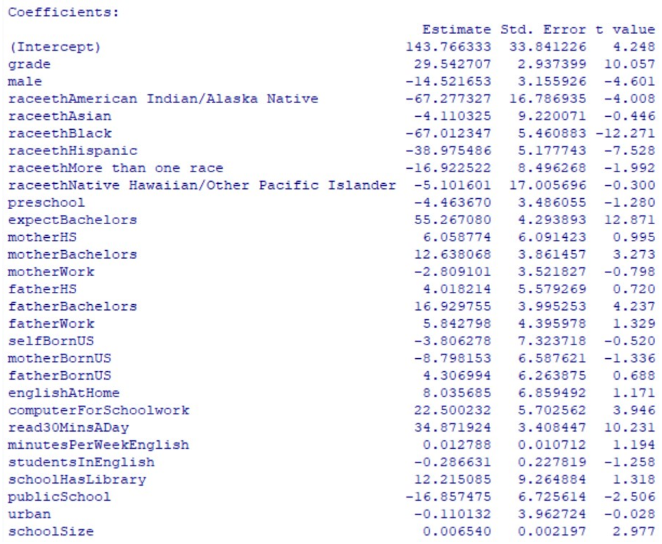
raceethAmerican Indian/Alaska Native", raceethAsian", raceethBlack", raceethHispanic", raceethMore than one race", and raceethNative Hawaiian/Other Pacific Islander should be included.
For a student who is Asian, raceethAsian" should be set to 1; All other binary variables, raceethAmerican Indian/Alaska Native", raceethBlack", raceethHispanic", raceethMore than one race", and raceethNative Hawaiian/Other Pacific Islander should be set to 0.

(c) Because the race variable takes on text values, by default R selects the first level alphabetically ("American Indian/Alaska Native") as the reference level, instead of the most common level ("White"). To reset the reference level of the factor, we can use the following two commands in the R console: `AssessmentTrain$raceeth = relevel(AssessmentTrain$raceeth, "White")` and `AssessmentTest$raceeth = relevel(AssessmentTest$raceeth, "White")`. Build a linear regression model to predict the reading score of students using all remaining variables as covariates. Use the training set to build the model.

Please show a screen shot of your linear regression model using the "summary" function. Find the independent variables that are significant (with $p-$value, $Pr(> |t|)$, smaller than 0.05). Consider two students A and B. They have all variable values the same, except that student A is in grade 11 and student B is in grade 9. What is the predicted reading score of student A minus the predicted reading score of student B? Please interpret the estimated coefficient of the variable "Asian". Figure 4

readingScore = 143.77 + 29.54 * grade -14.52*male -67.28 * raceethAmerican Indian/Alaska Native -4.11*raceethAsian -67.01*raceethBlack -38.98*raceethHispanic -16.92 * raceethMore than one race - 5.10 * raceethNative Hawaiian/Other Pacific Islander - 4.46 * preschool + 55.27 * expectBachelors + 6.05 * mothers + 12.64 * motherBachelors -2.81 * motherWork + 4.02 * fathers + 16.93 * fatherBachelors + 5.84 * fatherWork -3.81 * selfBornUS - 8.80 * motherBornUS + 4.31 * fatherBornUS +8.04 * englishAtHome + 22.50 * computerForSchoolwork + 34.87 * read30MinsADay +0.01 * minutesPerWeekEnglish + 12.22 * schoolHasLibrary -16.86 * publicSchool -0.11* urban +0.01 * schoolSize

- $raceethAmericanIndian/AlaskaNative$: one unit change in raceethAmerican Indian/Alaska Native is associate with 67.28 decrease in readingScore
- $raceethAsian$: one unit change in raceethAsian is associate with 4.11 decrease in readingScore
- $raceethBlack$: one unit change in raceethBlack is associate with 67.01 decrease in readingScore
- $raceethHispanic$: one unit change in raceethHispanic is associate with 38.98 decrease in readingScore
- $raceethMorethanonerace$: one unit change in raceethMore than one race is associate with 16.92 decrease in readingScore
- $raceethNativeHawaiian/OtherPacificIslander$: one unit change in raceethNative Hawaiian/Other Pacific Islander is associate with 5.10 decrease in readingScore



```
Coefficients:
                                                Estimate Std. Error t value
(Intercept)                                   143.766333  33.841226   4.248
grade                                          29.542707   2.937399  10.057
male                                          -14.521653   3.155926  -4.601
raceethAmerican Indian/Alaska Native          -67.277327  16.786935  -4.008
raceethAsian                                   -4.110325   9.220071  -0.446
raceethBlack                                  -67.012347   5.460883 -12.271
raceethHispanic                               -38.975486   5.177743  -7.528
raceethMore than one race                     -16.922522   8.496268  -1.992
raceethNative Hawaiian/Other Pacific Islander  -5.101601  17.005696  -0.300
preschool                                      -4.463670   3.486055  -1.280
expectBachelors                                55.267080   4.293893  12.871
motherHS                                        6.058774   6.091423   0.995
motherBachelors                                12.638068   3.861457   3.273
motherWork                                     -2.809101   3.521827  -0.798
fatherHS                                        4.018214   5.579269   0.720
fatherBachelors                                16.929755   3.995253   4.237
fatherWork                                      5.842798   4.395978   1.329
selfBornUS                                     -3.806278   7.323718  -0.520
motherBornUS                                   -8.798153   6.587621  -1.336
fatherBornUS                                    4.306994   6.263875   0.688
englishAtHome                                   8.035685   6.859492   1.171
computerForSchoolwork                          22.500232   5.702562   3.946
read30MinsADay                                 34.871924   3.408447  10.231
minutesPerWeekEnglish                           0.012788   0.010712   1.194
studentsInEnglish                              -0.286631   0.227819  -1.258
schoolHasLibrary                               12.215085   9.264884   1.318
publicSchool                                  -16.857475   6.725614  -2.506
urban                                          -0.110132   3.962724  -0.028
schoolSize                                      0.006540   0.002197   2.977
```

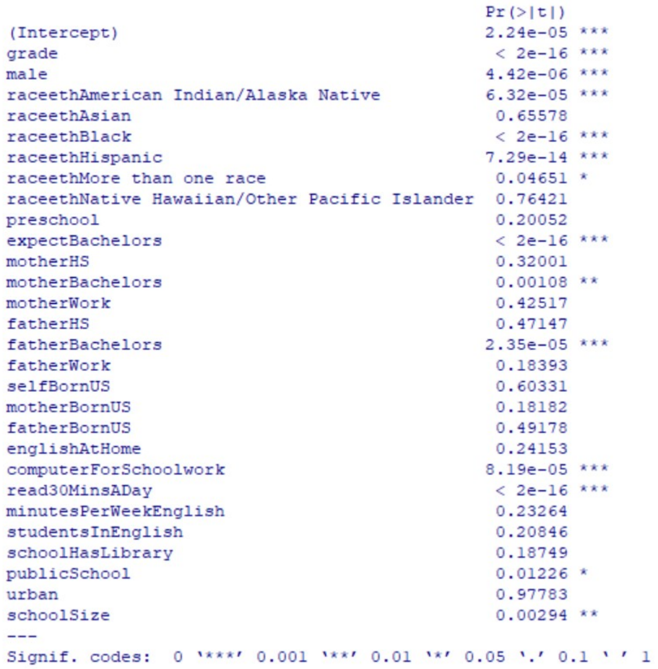
Significant variables: grade, male, raceeth, expectBachelors, motherBachelors, fatherBachelors, computerForeSchoolwork, read30MinsADay, publicSchool, and schoolSize.

Predicted score difference between student A and B is 29.54*2 = 59.08.

The estimated coefficient of the variable raceethAsian equals -4.11, meaning that, if all else equal, Asian students are estimated to have reading scores 4.11 points lower than the white students.

(d) If we remove all the independent variables that are not significant (with $p-$value, $Pr(>|t|)$, no smaller than 0.05), we can build a new linear regression model. Test both the full model (i.e., with all the variables in the data set as independent variables) and the new model (i.e., with the insignificant variables removed) using the testing data. Report the respective out-of-sample R^2's for both models.

The out-of-sample R-squared for the full model is 0.2615.
The out-of-sample R-squared for the simplified model is 0.2657.



```
                                              Pr(>|t|)
(Intercept)                                   2.24e-05 ***
grade                                          < 2e-16 ***
male                                          4.42e-06 ***
raceethAmerican Indian/Alaska Native          6.32e-05 ***
raceethAsian                                   0.65578
raceethBlack                                   < 2e-16 ***
raceethHispanic                               7.29e-14 ***
raceethMore than one race                      0.04651 *
raceethNative Hawaiian/Other Pacific Islander  0.76421
preschool                                      0.20052
expectBachelors                                < 2e-16 ***
motherHS                                       0.32001
motherBachelors                                0.00108 **
motherWork                                     0.42517
fatherHS                                       0.47147
fatherBachelors                               2.35e-05 ***
fatherWork                                     0.18393
selfBornUS                                     0.60331
motherBornUS                                   0.18182
fatherBornUS                                   0.49178
englishAtHome                                  0.24153
computerForSchoolwork                         8.19e-05 ***
read30MinsADay                                 < 2e-16 ***
minutesPerWeekEnglish                          0.23264
studentsInEnglish                              0.20846
schoolHasLibrary                               0.18749
publicSchool                                   0.01226 *
urban                                          0.97783
schoolSize                                     0.00294 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4: Question 2(c)

3. Forecasting Auto Sales

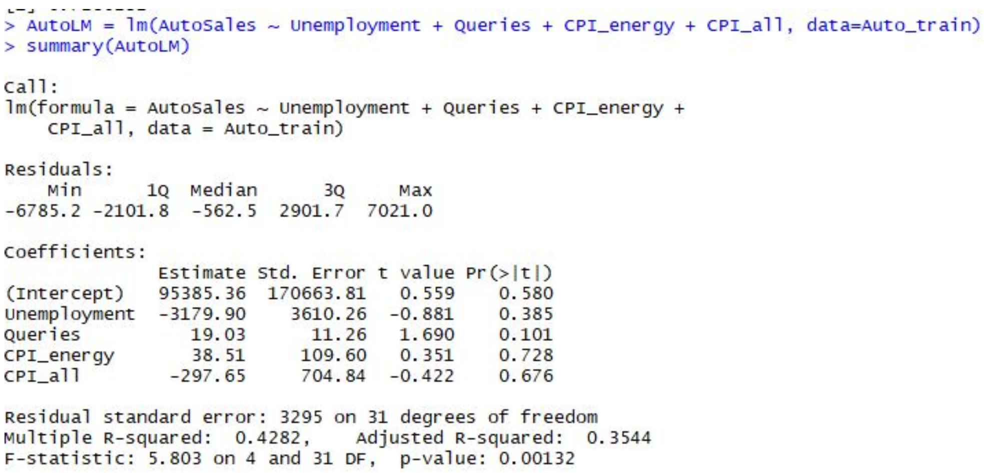
In this problem, we will try to predict monthly sales of an Auto Brand.

The file `Auto.csv` contains data for the problem. Each observation is a month, from January 2010 to February 2014. For each month, we have the following variables:

- $Month$ = the month of the year for the observation (1 = January, 2 = February, 3 = March, ...).
- $Year$ = the year of the observation.
- *AutoSales* = the number of units of the Auto sold in the United States in the given month.
- $Unemployment$ = the estimated unemployment percentage in the United States in the given month.
- *Queries* = a (normalized) approximation of the number of Google searches for "Auto" in the given month.
- CPI_energy = the monthly consumer price index (CPI) for energy for the given month.
- CPI_all = the consumer price index (CPI) for all products for the given month; this is a measure of the magnitude of the prices paid by consumer households for goods and services (e.g., food, clothing, electricity, etc.).

Load the data set into R and split the data set into training and testing sets as follows: Place all observations for 2012 and earlier in the training set, and all observations for 2013 and 2014 into the testing set.

(a) Build a linear regression model to predict monthly Auto sales using Unemployment, CPI_all, CPI_energy and Queries as the independent variables. Use all of the training set data to do this. Please show a screen shot of your linear regression model using the "summary" function. Clearly state the significance, the sign, and the magnitude of the association between the dependent variable and each independent variable. Figure 5



```
> AutoLM = lm(AutoSales ~ Unemployment + Queries + CPI_energy + CPI_all, data=Auto_train)
> summary(AutoLM)

Call:
lm(formula = AutoSales ~ Unemployment + Queries + CPI_energy +
    CPI_all, data = Auto_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-6785.2 -2101.8  -562.5  2901.7  7021.0 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   95385.36  170663.81   0.559    0.580
Unemployment  -3179.90    3610.26  -0.881    0.385
Queries          19.03      11.26   1.690    0.101
CPI_energy       38.51     109.60   0.351    0.728
CPI_all        -297.65     704.84  -0.422    0.676

Residual standard error: 3295 on 31 degrees of freedom
Multiple R-squared:  0.4282,	Adjusted R-squared:  0.3544 
F-statistic: 5.803 on 4 and 31 DF,  p-value: 0.00132
```

Figure 5: Question 3(a)

AutoSales = 95385.36 -3179.90Unemployment + 19.03Queries + 38.51CPI_energy -297.65CPI_all

- *Unemployment* : insignificant, negative, one unit change in Unemployment is associate with 3179.90 decrease in Auto Sales
- *Queries* : insignificant, positive, one unit change in Queries is associate with 19.03 increase in Auto Sales
- *CPI_energy* : insignificant, positive, one unit change in CPI_energy is associate with 38.51 increase in Auto Sales
- *CPI_all* : insignificant, negative, one unit change in Unemployment is associate with 297.65 decrease in Auto Sales

(b) We would now like to improve the model by incorporating seasonality. Seasonality refers to the fact that demand is often cyclical/periodic in time. For example, demand for warm outerwear (like jackets and coats) is higher in fall/autumn and winter than in spring and summer.

In our problem, since our data includes the month of the year in which the units were sold, it is feasible for us to incorporate monthly seasonality. From a modeling point of view, it may be reasonable that the month plays an effect in how many Auto units are sold.

To incorporate the seasonal effect due to the month, build a new linear regression model that predicts monthly Auto sales using Month as well as Unemployment, CPI_all, CPI_energy and Queries. Do not modify the training and testing data frames before building the model. Based on the model estimation results, how do you evaluate the new model compared with the original one?

After creating the new model AutoLM_new by adding the Month Variable, the model is not performing better because the adjusted R-squared has gone down and none of the variables are very significant. The adjusted R-Squared is adjusted to take into account the number of variables. If the adjusted R-Squared is lower, then this indicates that our model is not better and in fact may be worse.

(c) In the new model, given two monthly periods that are otherwise identical in Unemployment, CPI_all, CPI_energy and Queries, what is the absolute difference in predicted Auto sales given that one period is in January and one is in March? Consider again the new model, given two monthly periods that are otherwise identical in Unemployment, CPI_all, CPI_energy and Queries, what is the absolute difference in predicted Auto sales given that one period is in January and one is in May? Is there anything you feel uncomfortable about this finding?

110.69 * (3 - 1) = 110.69 * 2 = 221.38

110.69 * (5 - 1) = 110.69 * 4 = 442.76

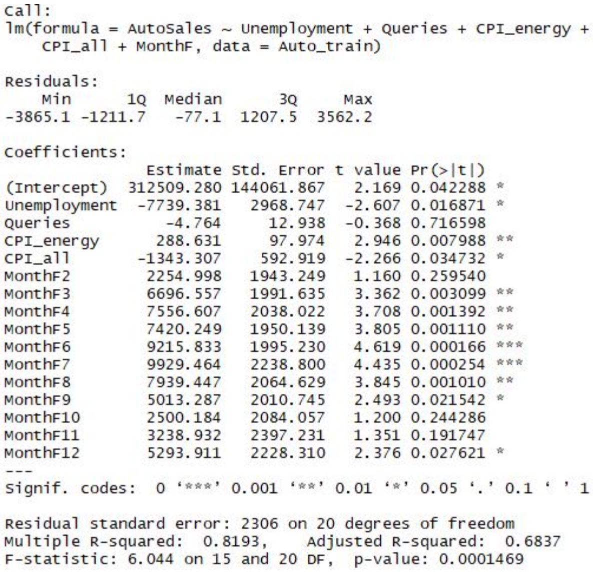
There is something not quite right in how we have modeled the effect of the calendar month on the monthly sales of autos. In particular, we added Month as a variable, but Month is an ordinary numeric variable. We must convert Month to a factor variable before adding it to the model.

(d) Alternatively, we consider Month as a factor variable, instead of a numeric variable. Then, we can use the binary variable technique introduced in Problem 2 to build a linear regression model. Why do you think we should use the factor variable instead of the numeric variable to represent month? To convert a numeric variable into a factor variable, you may use the function `as.factor()`. To apply this function, you may type:
`Auto_train$MonthF=as.factor(Auto_train$Month)` and
`Auto_test$MonthF=as.factor(Auto_test$Month)` in the R console. In this way, you will not overwrite the original numeric variable Month.

Because by modeling Month as a factor variable, the effect of each calendar month is not restricted to be linear in the numerical coding of the month.

(e) Re-run the regression with the Month variable modeled as a factor variable. (Create a new variable that models the Month as a factor. From the new regression results, what seasonality pattern have you observed?

Let us call the new model AutoLM_newF, with switching the Month Variable in the previous model to MonthF.
After the looking at the summary of this model (see Figure 6), we can see that summer, especially June and July, is the peak season for auto sales.



```
Call:
lm(formula = AutoSales ~ Unemployment + Queries + CPI_energy + 
    CPI_all + MonthF, data = Auto_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-3865.1 -1211.7   -77.1  1207.5  3562.2 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  312509.280 144061.867   2.169 0.042288 *  
Unemployment  -7739.381   2968.747  -2.607 0.016871 *  
Queries          -4.764     12.938  -0.368 0.716598    
CPI_energy      288.631     97.974   2.946 0.007988 ** 
CPI_all       -1343.307    592.919  -2.266 0.034732 *  
MonthF2        2254.998   1943.249   1.160 0.259540    
MonthF3        6696.557   1991.635   3.362 0.003099 ** 
MonthF4        7556.607   2038.022   3.708 0.001392 ** 
MonthF5        7420.249   1950.139   3.805 0.001110 ** 
MonthF6        9215.833   1995.230   4.619 0.000166 ***
MonthF7        9929.464   2238.800   4.435 0.000254 ***
MonthF8        7939.447   2064.629   3.845 0.001010 ** 
MonthF9        5013.287   2010.745   2.493 0.021542 *  
MonthF10       2500.184   2084.057   1.200 0.244286    
MonthF11       3238.932   2397.231   1.351 0.191747    
MonthF12       5293.911   2228.310   2.376 0.027621 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2306 on 20 degrees of freedom
Multiple R-squared:  0.8193,	Adjusted R-squared:  0.6837 
F-statistic: 6.044 on 15 and 20 DF,  p-value: 0.0001469
```

Figure 6: Question 3(e)

(f) Another peculiar observation about the regression results (with month as a factor variable) is that the signs of the Queries variable and the CPI_energy variable. Why their signs are counter-intuitive? Please try to give an explanation for such phenomenon and find a way to address this issue. You may need to remove some independent variables and re-build the linear regression model.

It seems counter-intuitive that Queries should expect a positive sign and CPI should expect a negative sign but both of them are assigned an opposite sign in the model as expected.
It is a matter of multicolinearity. There are high correlations among Queries, CPI_all and CPI_energy. Therefore, we need a reduced model by removing Queries, which has the highest p-values. Here we create a new reduced model called AutoLM_newF_reduced with removing Queries from the AutoLM_newF model in (e). Also, after checking the correlation of this model, we find CPI_energy and CPI_all are still highly correlated. Therefore, it is better to remove these two variables from the model AutoLM_newF_reduced and create another model AutoLM_newF_reduced2, which only contains Unemployment and MonthF.
Specifically,
AutoLM_newF_reduced = lm(AutoSales ∼ Unemployment + CPI_energy + CPI_all + MonthF, data=Auto_train)
AutoLM_newF_reduced2 = lm(AutoSales ∼ Unemployment + MonthF, data=Auto_train)
We are going to do out-of-sample tests to all the models in (g).

(g) Use out-of-sample test to evaluate all your models built to estimate the sales of Auto. Report the out-of-sample R^2 of each model and discuss which model you would like recommend to this Auto Brand for their sales forecasting.

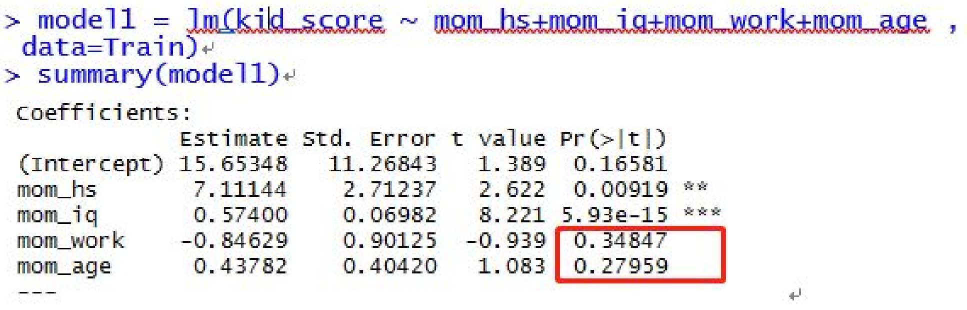
In the model AutoLM, out-of-sample $R^2 = 0.4975116$
In the model AutoLM_new, out-of-sample $R^2 = 0.4662344$
In the model AutoLM_newF, out-of-sample $R^2 = 0.7426902$
In the model AutoLM_newF_reduced, out-of-sample $R^2 = 0.7280232$
In the model AutoLM_newF_reduced2, out-of-sample $R^2 = 0.8793167$
Here we recommend the last model because it 1) has a higher out-of-sample R^2 than the other models; 2) it includes Month as factor instead of numeric, which makes the seasonality more interpretable; 3) it has dropped highly correlated variable Queries, CPI_all and CPI_energy, which can deal with the problem of multicolinearity.

4. Child IQ

Build a model to predict the IQ of a child based on covariates about his/her mom using the data set `kidiq.csv`. Below are the variables contained therein

- *kid_score*: IQ score of the kid
- *mom_hs*: whether the mom has attained high school (1/0)
- *mon_iq*: mom's IQ score
- *mon_work*: a numerical variable ranges from 1 to 4,
  - 1 =did not work in the first three years of the child's life
  - 2 =worked in the 2nd or 3rd year of child's life
  - 3 =worked part-time in the first year of child's life
  - 4 =worked full-time in the first year of child's life
- *mom_age*: age of the mom when delivering the child.

Which variables do you recommend to be included into the independent variables? Please report all the steps and results of your analysis. Please also interpret your results.
After testing several models, we have decided to choose the following model:
kid_score =-11.48+51.27 mom_hs+0.97mom_iq-0.48mom_hs*mom_iq,
where if mom_hs = 1, it is associated with 51.27 increase in kid_score and one unit change in mom_iq is associate with 0.49 increase in kid_score;
if mom_hs = 0, one unit change in mom_iq is associate with 0.97 increase in kid_score
Following are the steps and analysis:
1. Plot Scatter between kid_score and each variable.
2. Split data into training set and testing set
3. Construct model: regress kid_score on the rest of variables and observe the result.
model1 = lm(kid_score ~ mom_hs+mom_iq+mom_work+mom_age , data=kidIQ) Figure 7

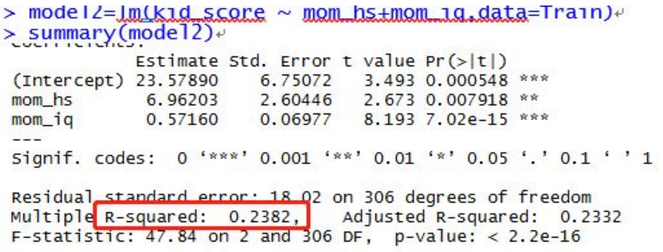


```
> model1 = lm(kid_score ~ mom_hs+mom_iq+mom_work+mom_age ,
  data=Train)
> summary(model1)
 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
 (Intercept) 15.65348   11.26843   1.389  0.16581
 mom_hs       7.11144    2.71237   2.622  0.00919 **
 mom_iq       0.57400    0.06982   8.221 5.93e-15 ***
 mom_work    -0.84629    0.90125  -0.939  0.34847
 mom_age      0.43782    0.40420   1.083  0.27959
 ---
```

Figure 7: Question 4(step3)

From above picture, we find that the P-value of the mom work and mom age are greater than 0.05. It implies that mom work and mom age are insignificant variables. We build a new model by drop these two variables.

4. Regress kid_score on mom_hs and mom_iq.
See Figure 8



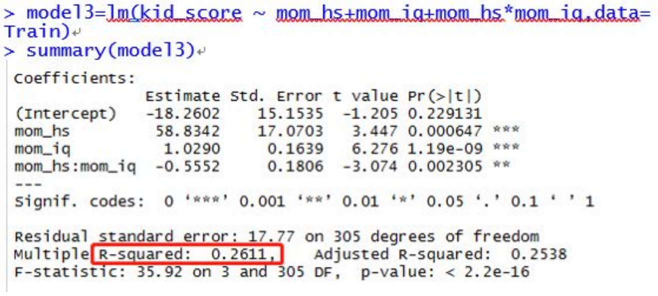
```
> model2=lm(kid_score ~ mom_hs+mom_iq,data=Train)
> summary(model2)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.57890    6.75072   3.493 0.000548 ***
mom_hs       6.96203    2.60446   2.673 0.007918 **
mom_iq       0.57160    0.06977   8.193 7.02e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.02 on 306 degrees of freedom
Multiple R-squared:  0.2382,    Adjusted R-squared:  0.2332
F-statistic: 47.84 on 2 and 306 DF,  p-value: < 2.2e-16
```

Figure 8: Question 4(step4)

5. We find variables are significant. Then, we want to improve the R squared. The interaction term is taken into consideration. Figure 9

R squared is improved.

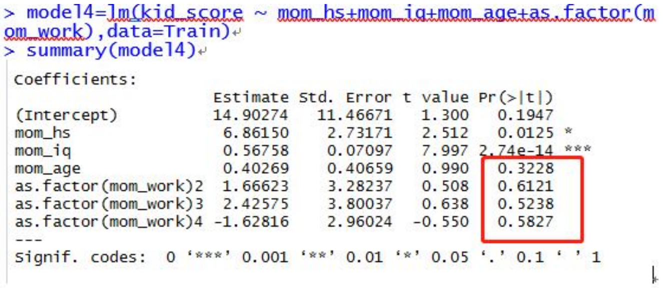


```
> model3=lm(kid_score ~ mom_hs+mom_iq+mom_hs*mom_iq,data=
Train)
> summary(model3)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -18.2602    15.1535  -1.205 0.229131
mom_hs         58.8342    17.0703   3.447 0.000647 ***
mom_iq          1.0290     0.1639   6.276 1.19e-09 ***
mom_hs:mom_iq  -0.5552     0.1806  -3.074 0.002305 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.77 on 305 degrees of freedom
Multiple R-squared:  0.2611,    Adjusted R-squared:  0.2538
F-statistic: 35.92 on 3 and 305 DF,  p-value: < 2.2e-16
```

Figure 9: Question 4(step5)

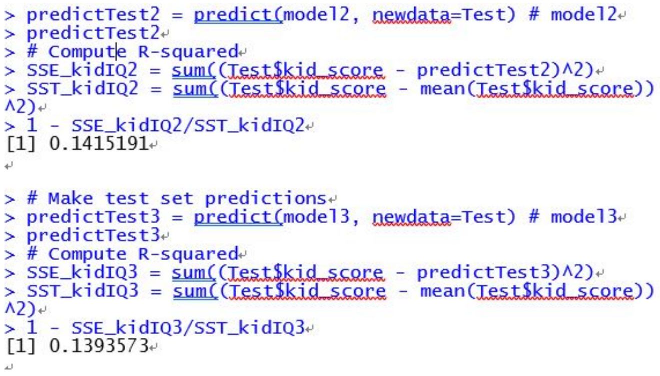
6. We observed that mom_work can be regraded as factor variables, then we could construct a new model. See Figure 10, mom_work are also not significant variables.



```
> model4=lm(kid_score ~ mom_hs+mom_iq+mom_age+as.factor(m
om_work),data=Train)
> summary(model4)
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           14.90274   11.46671   1.300   0.1947
mom_hs                 6.86150    2.73171   2.512   0.0125 *
mom_iq                 0.56758    0.07097   7.997 2.74e-14 ***
mom_age                0.40269    0.40659   0.990   0.3228
as.factor(mom_work)2   1.66623    3.28237   0.508   0.6121
as.factor(mom_work)3   2.42575    3.80037   0.638   0.5238
as.factor(mom_work)4  -1.62816    2.96024  -0.550   0.5827
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10: Question 4(step6)

7. We use testing set to test model2 and model3: See 11



```
> predictTest2 = predict(model2, newdata=Test) # model2
> predictTest2
> # Compute R-squared
> SSE_kidIQ2 = sum((Test$kid_score - predictTest2)^2)
> SST_kidIQ2 = sum((Test$kid_score - mean(Test$kid_score))
^2)
> 1 - SSE_kidIQ2/SST_kidIQ2
[1] 0.1415191


> # Make test set predictions
> predictTest3 = predict(model3, newdata=Test) # model3
> predictTest3
> # Compute R-squared
> SSE_kidIQ3 = sum((Test$kid_score - predictTest3)^2)
> SST_kidIQ3 = sum((Test$kid_score - mean(Test$kid_score))
^2)
> 1 - SSE_kidIQ3/SST_kidIQ3
[1] 0.1393573
```

Figure 11: Question 4(step7)

From the test data, we should choose the model:
kid_score= 23.57890+ 6.96203*mom_hs+ 0.57160* mom_iq;
In fact, the difference between model2 and model3 is very small. You also could choose model3.
kid_score= -18.2602+ 58.8342*mom_hs+ 1.0290* mom_iq -0.5552* mom_hs*mom_iq.