The Analytics Edge

Test your knowledge of Linear Regression in R

1. This question involves the use of simple linear regression on the **Auto** dataset. This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset has the following fields:

   - **mpg**: miles per gallon
   - **cylinders** Number of cylinders
   - **displacement**: Engine displacement (cu. inches)
   - **horsepower**: Engine horsepower
   - **acceleration**: Time to accelerate from 0 to 60 mph (sec.)
   - **year**: Model year (modulo 100)
   - **origin**: Origin of car (1. American, 2. European, 3. Japanese)
   - **name**: Vehicle name

   (a) Perform a simple linear regression with **mpg** as the response and **horsepower** as the predictor. Comment on why you need to change the horsepower variable before performing the regression.

   (b) Comment on the output by answering the following questions:

   - Is there a strong relationship between the predictor and the response?
   - Is the relationship between the predictor and the response positive or negative?

   (c) What is the predicted **mpg** associated with a **horsepower** of 98? What is the associated 99% confidence interval? Hint: You can check the predict.lm function on how the confidence interval can be computed for predictions with R.

   (d) Compute the correlation between the response and the predictor variable. How does this compare with the R^2 value?

   (e) Plot the response and the predictor. Also plot the least squares regression line.

   (f) Use the following two commands in R to produce diagnostic plots of the linear regression fit:
   > layout(matrix(1:4,2,2))
   > plot(*your_model_name*)
   Comment on the Residuals versus Fitted plot and the Normal Q-Q plot and on any problems you might see with the fit.

2. This question involves the use of multiple linear regression on the **Auto** dataset building on question 1.

   (a) Produce a scatterplot matrix which includes all the variables in the dataset.

   (b) Compute a matrix of correlations between the variables using the function **cor()**. You need to exclude the **name** variables which is qualitative.

   (c) Perform a multiple linear regression with **mpg** as the response and all other variables except **name** as the predictors. Comment on the output by answering the following questions:

   - Is there a strong relationship between the predictors and the response?
   - Which predictors appear to have a statistically significant relationship to the response?
   - What does the coefficient for the **year** variable suggest?

3. This problem focusses on the multicollinearity problem with simulated data.

   (a) Perform the following commands in R:
   > set.seed(1)
   > x1 <− runif(100)
   > x2 <− 0.5*x1 + rnorm(100)/10
   > y <− 2 + 2*x1 + 0.3*x2 + rnorm(100)
   The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients?

   (b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.

   (c) Using the data, fit a least square regression to predict y using x1 and x2.

   - What are the estimated parameters of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$? How do these relate to the true β_0, β_1 and β_2?
   - Can you reject the null hypothesis $H_0 : \beta_1 = 0$?
   - How about the null hypothesis $H_0 : \beta_2 = 0$?

   (d) Now fit a least squares regression to predict y using only x1.

   - How does the estimated $\hat{\beta}_1$ relate to the true β_1?
   - Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

   (e) Now fit a least squares regression to predict y using only x2.

   - How does the estimated $\hat{\beta}_2$ relate to the true β_2?
   - Can you reject the null hypothesis $H_0 : \beta_2 = 0$?

   (f) Provide an explanation on the results in parts (c)-(e).

4. This problem involves the **Boston** dataset. This data was part of an important paper in 1978 by Harrison and Rubinfeld titled - "Hedonic housing prices and the demand for clean air" published in the Journal of Environmental Economics and Management 5(1): 81-102. The dataset has the following fields:

   - **crim**: per capita crime rate by town
   - **zn**: proportion of residential land zoned for lots over 25,000 sq.ft
   - **indus**: proportion of non-retail business acres per town
   - **chas**: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
   - **nox**: nitrogen oxides concentration (parts per 10 million)
   - **rm**: average number of rooms per dwelling
   - **age**: proportion of owner-occupied units built prior to 1940
   - **dis**: weighted mean of distances to five Boston employment centres
   - **rad**: index of accessibility to radial highways
   - **tax**: full-value property-tax rate per $10,000
   - **ptratio**: pupil-teacher ratio by town
   - **black**: $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
   - **lstat**: lower status of the population (percent)
   - **medv**: median value of owner-occupied homes in $1000s

   We will try to predict the median house value using thirteen predictors

   (a) For each predictor, fit a simple linear regression model using a single variable to predict the response. In which of these models is there a statistically significant relationship between the predictor and the response? Plot the figure of relationship between medv and lstat as an example to validate your finding.

   (b) Fit a multiple linear regression models to predict your response using all the predictors. Compare the adjusted R^2 from this model with the simple regression model. For which predictors, can we reject the null hypothesis $H_0 : \beta_j = 0$?

   (c) Create a plot displaying the univariate regression coefficients from (a) on the X-axis and the multiple regression coefficients from (b) on the Y-axs. That is each predictor is displayed as a single point in the plot. Comment on this plot.

   (d) In this question, we will check if there is evidence of non-linear association between the **lstat** predictor variable and the response? To answer the question, fit a model of the form

   $$\mathbf{medv} = \beta_0 + \beta_1 \mathbf{lstat} + \beta_2 \mathbf{lstat}^2 + \epsilon.$$

   You can make use of the poly() function in R. Does this help improve the fit. Add higher degree polynomial fits. What is the degree of the polynomial fit beyond which the terms no longer remain significant?

5. There have been many studies documenting that the average global temperature has been increasing over the last century. The consequences of a continued rise in global temperature will be dire. Rising sea levels and an increased frequency of extreme weather events will affect billions of people. In this problem, we will attempt to study the relationship between average global temperature and several other factors. The file **climate_change.csv** contains climate data from May 1983 to December 2008. The available variables include:

   - **Year**: Observation year
   - **Month**: Observation month
   - **Temp**: Difference in degrees Celsius between the average global temperature in that period and a reference value. This data comes from the Climatic Research Unit at the University of East Anglia.
   - **CO2, N2O, CH4, CFC.11, CFC.12**: Atmospheric 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. This data comes from the ESRL/NOAA Global Monitoring Division.
     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**: Mean stratospheric aerosol optical depth at 550 nm. This variable is linked to volcanoes, as volcanic eruptions result in new particles being added to the atmosphere, which affect how much of the sun's energy is reflected back into space. This data is from the Godard Institute for Space Studies at NASA.
   - **TSI**: Total solar irradiance (TSI) in W/m2 (the rate at which the sun's energy is deposited per unit area). Due to sunspots and other solar phenomena, the amount of energy that is given off by the sun varies substantially with time. This data is from the SOLARIS-HEPPA project website.
   - **MEI**: Multivariate El Nino Southern Oscillation index (MEI), a measure of the strength of the El Nino/La Nina-Southern Oscillation (a weather effect in the Pacific Ocean that affects global temperatures). This data comes from the ESRL/NOAA Physical Sciences Division.

   (a) We are interested in how changes in these variables affect future temperatures, as well as how well these variables explain temperature changes so far. Read the dataset **climate_change.csv** into R. Then, split the data into a training set, consisting of all the observations up to and including 2006, and a testing set consisting of the remaining years. A training set refers to the data that will be used to build the model, and a testing set refers to the data we will use to test our predictive ability. Build a linear regression model to predict the dependent variable Temp, using MEI, CO2, CH4, N2O, CFC.11, CFC.12,

TSI and Aerosols as independent variables (Year and Month should not be used in the model). Use the training set to build the model. What is the model R^2?

(b) Which variables are significant in the model? We will consider a variable significant in this example only if the p-value is below 0.05.

(c) Current scientific opinion is that nitrous oxide and CFC-11 are greenhouse gases: gases that are able to trap heat from the sun and contribute to the heating of the Earth. However, the regression coefficients of both the N2O and CFC-11 variables are negative, indicating that increasing atmospheric concentrations of either of these two compounds is associated with lower global temperatures. Compute the correlations in the training set. Which of the following is the simplest correct explanation for this contradiction?

- Climate scientists are wrong that N2O and CFC-11 are greenhouse gases - this regression analysis constitutes part of a disproof.
- There is not enough data, so the regression coefficients being estimated are not accurate.
- All of the gas concentration variables reflect human development - N2O and CFC.11 are correlated with other variables in the data set.

(d) Given that the correlations are so high, let us focus on the N2O variable and build a model with only MEI, TSI, Aerosols and N2O as independent variables. Remember to use the training set to build the model. What is the coefficient of N2O in this reduced model? How does this compare to the coefficient in the previous model with all of the variables? What is the model R^2?

(e) We have many variables in this problem, and as we have seen above, dropping some from the model does not decrease model quality. R provides a function **step()**, that will automate the procedure of trying different combinations of variables to find a good compromise of model simplicity and R^2. This trade-off is formalized by the Akaike information criterion (AIC) - it can be informally thought of as the quality of the model with a penalty for the number of variables in the model. The step function has one argument - the name of the initial model. It returns a simplified model. Use the step function in R to derive a new model, with the full model as the initial model. What is the R^2 value of the model produced by the step function? Which of the variable(s) are eliminated from the full model by the step function? It is interesting to note that the step function does not address the collinearity of the variables, except that adding highly correlated variables will not improve the R^2 significantly. The consequence of this is that the step function will not necessarily produce a very interpretable model - just a model that has balanced quality and simplicity for a particular weighting of quality and simplicity (AIC).

(f) We have developed an understanding of how well we can fit a linear regression to the training data, but does the model quality hold when applied to unseen data? Using the model produced from the step function, calculate temperature predictions for the testing data set, using the predict function. What is the test R^2?

6. Orley Ashenfelter in his paper "Predicting the Quality and Price of Bordeaux Wines" published in The Economic Journal showed that the variability in the prices of Bordeaux wines is predicted well by the weather that created the grapes. In this question, you will validate how these results translate to a dataset for wines produced in Australia. The data is provided in the file **winedata.csv**. The dataset contains the following variables:

   - **vintage**: Year the wine was made
   - **price91**: 1991 auction prices for the wine in dollars
   - **price92**: 1992 auction prices for the wine in dollars
   - **temp**: Average temperature during the growing season in degree Celsius
   - **hrain**: Total harvest rain in mm
   - **wrain**: Total winter rain in mm
   - **tempdiff**: Sum of the difference between the maximum and minimum temperatures during the growing season in degree Celsius

   (a) Define two new variables **age91** and **age92** that captures the age of the wine (in years) at the time of the auctions. For example, a 1961 wine would have an age of 30 at the auction in 1991. What is the average price of wines that were 15 years or older at the time of the 1991 auction?

   (b) What is the average price of the wines in the 1991 auction that were produced in years when both the harvest rain was below average and the temperature difference was below average?

   (c) In this question, you will develop a simple linear regression model to fit the **log** of the price at which the wine was auctioned in 1991 with the age of the wine. To fit the model, use a training set with data for the wines up to (and including) the year 1981. What is the R-squared for this model?

   (d) Find the 99% confidence interval for the estimated coefficients from the regression.

   (e) Use the model to predict the **log** of prices for wines made from 1982 onwards and auctioned in 1991. What is the test R-squared?

   (f) Which among the following options describes best the quality of fit of the model for this dataset in comparison with the Bordeaux wine dataset that was analyzed by Orley Ashenfelter?

      - The result indicates that the variation of the prices of the wines in this dataset is explained much less by the age of the wine in comparison to Bordeaux wines.
      - The result indicates that the variation of the prices of the wines in this dataset is explained much more by the age of the wine in comparison to Bordeaux wines.
      - The age of the wine has no predictive power on the wine prices in both the datasets.

(g) Construct a multiple regression model to fit the log of the price at which the wine was auctioned in 1991 with all the possible predictors (**age91, temp, hrain, wrain, tempdiff**) in the training dataset. To fit your model, use the data for wines made up to (and including) the year 1981. What is the R-squared for the model?

(h) Is this model preferred to the model with only the age variable as a predictor (use the adjusted R-squared for the model to decide on this)?

(i) Which among the following best describes the output from the fitted model?

- The result indicates that less the temperature, the better is the price and quality of the wine
- The result indicates that greater the temperature difference, the better is the price and quality of wine.
- The result indicates that lesser the harvest rain, the better is the price and quality of the wine.
- The result indicates that winter rain is a very important variable in the fit of the data.

(j) Of the five variables (**age91, temp, hrain, wrain, tempdiff**), drop the two variables that are the least significant from the results in (g). Rerun the linear regression and write down your fitted model.

(k) Is this model preferred to the model with all variables as predictors (use the adjusted R-squared in the training set to decide on this)?

(l) Using the variables identified in (j), construct a multiple regression model to fit the log of the price at which the wine was auctioned in 1992 (remember to use **age92** instead of **age91**). To fit your model, use the data for wines made up to (and including) the year 1981. What is the R-squared for the model?

(m) Suppose in this application, we assume that a variable is statistically significant at the 0.2 level. Would you would reject the hypothesis that the coefficient for the variable **hrain** is nonzero?

(n) By separately estimating the equations for the wine prices for each auction, we can better establish the credibility of the explanatory variables because:

- We have more data to fit our models with.
- The effect of the weather variables and age of the wine (sign of the estimated coefficients) can be checked for consistency across years.
- 1991 and 1992 are the markets when the Australian wines were traded heavily.

Select the best option.

(o) The current fit of the linear regression using the weather variables drops all observations where any of the entries are missing. Provide a short explanation on when this might not be a reasonable approach to use.

7. In his 2003 bestseller Moneyball, the author Michael Lewis makes an interesting claim that the Oakland Athletics team manager Billy Beane felt that the valuation of skills in the market for baseball players at that time was inefficient and undervalued important batting skills such as on-base percentage while overvalued batting skills such as slugging percentage. In this question, we will verify with data if these claims made by Lewis are indeed true by looking at data on player performances and salaries for the 1996 and 2006 seasons. The data is provided in the file **batters.csv** and contains the following variables:

   - **playerID**: Player identity code
   - **yearID**: Year
   - **teamID**: Team identity code
   - **G**: Number of games in which the player played during the season
   - **AB**: At Bats
   - **R**: Runs
   - **H**: Hits (Times reached base because of a batted, fair ball without error by the defense)
   - **X2B**: Doubles (Hits on which the batter reached second base safely)
   - **X3B**: Triples (Hits on which the batter reached third base safely)
   - **HR**: Homeruns
   - **BB**: Base on balls
   - **HBP**: Hit by pitch
   - **SF**: Sacrifice flies
   - **salary**: Salary for players at the start of the next season

   (a) Read the data into the dataframe **batters**. Which player made the most salary in the 2006 season?

   (b) What is the ratio of the maximum salary to the minimum salary among all players in the 2006 season?

   (c) At the end of the 1996 season, which teams had the set of batters with the minimum and maximum total sum of salaries respectively?

   (d) Write down the R command(s) that you used to answer question (c).

   (e) Plot the histogram of the **salary** variable. What best describes the distribution of player salaries?

   - Most of the salaries are large, with a relatively small number of much smaller salaries (this is referred to as "left-skewed").
   - The salaries are balanced, with equal numbers of unusually large and unusually small salaries.
   - Most of the salaries are small, with a relatively small number of much larger salaries (this is referred to as "right-skewed").

(f) When handling a skewed dependent variable, it is often useful to predict the logarithm of the dependent variable instead of the dependent variable itself - this prevents the small number of unusually large or small observations from having an undue influence on the predictive model. In this problem, you will predict the natural logarithm of the salary variable at the end of a season with the number of runs scored in the season and a constant (intercept). Use the entire dataset to build your model with linear regression. What does your model predict to be the logarithm of the salary of a batter who scores 0 runs in a season?

(g) What is the actual average of the logarithm of the salary of batters who score 0 runs in a season in the dataset? Remember to drop missing entries in computing this number.

(h) Comment on whether the results in questions (f) and (g) are close to each other. If yes provide a brief explanation.

(i) Assume that the number of runs scored by a player increases by 1. Suppose β_1 is the coefficient of the number of runs scored in question (f). What best describes how your model would predict the change in the salary?

- New salary = Old salary + e^{β_1}
- New salary = Old salary $\times$ e^{β_1}
- New salary = Old salary + β_1
- New salary = Old salary $\times$ β_1

(j) We will now compare the effect of two baseball statistics on the salaries of the players. To do this, we need to define two new variables **OBP** (on-base percentage) and **SLG** (slugging percentage) as follows:

$$\text{OBP} = \frac{\text{Hits} + \text{Base on balls} + \text{Hit by pitch}}{\text{At Bats} + \text{Base on balls} + \text{Hit by pitch} + \text{Sacrifice flies}}$$

$$\text{SLG} = \frac{\text{Hits} + \text{Doubles} + 2 \times \text{Triples} + 3 \times \text{Homeruns}}{\text{At Bats}}$$

What is the average on-base percentage in the 2006 season? Drop observations with missing entries.

(k) Perform a two sided t-test to check if the average slugging percentage in the 1996 and 2006 seasons are different. What is the p-value of the test and your conclusion?

(l) We will now use linear regression to predict the logarithm of the salary using the **OBP** and **SLG** variables and the constant (intercept). To build the model, we will consider only batters with at least 130 at-bats, since this is required to qualify as honors for rookie of the year and helps provide an objective cutoff to check the effect of performance on players with relatively large sample of at-bats. Using only data for the year 1996, what is adjusted R-squared for your model?

(m) Is there evidence that you can reject the three null hypothesis H0 : $\beta_j = 0$ for the **OBP**, **SLG** and constant variables? Use a p-value of 0.05 to make your conclusion.

(n) Redo the linear regression from question (l) using only data for 2006. What is the adjusted R-squared for your model?

(o) Billy Beane, the Oakland Athletics coach believed that on-base percentage was much more important than the slugging percentage to help win a game. By looking at the coefficients of the **OBP** and **SLG** variables from the regressions in questions (l) and (n), we can conclude that:

- The market undervalued the **SLG** statistic relative to the **OBP** statistic in 1996 before Moneyball was published. This inefficiency still remained in 2006.
- The market undervalued the **SLG** statistic relative to the **OBP** statistic in 1996 before Moneyball was published. This has been corrected in 2006.
- The market undervalued the **OBP** statistic relative to the **SLG** statistic in 1996 before Moneyball was published. This inefficiency still remained in 2006.
- The market undervalued the **OBP** statistic relative to the **SLG** statistic in 1996 before Moneyball was published. This has been corrected in 2006.

Select the best option.