The Analytics Edge Summer 2022

Test your knowledge of Linear Regression and PCA in R

Exercise: Week 2

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) First install the package `ggfortify` which aids plotting linear models with `ggplot2`. Use the following two commands in R to produce diagnostic plots of the linear regression fit:
   ```
   > library(ggfortify)
   > autoplot(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` variable 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 black residents 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-axis. 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

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

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. 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 reject the hypothesis that the coefficient for the variable `hrain` is zero?

(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.

6. This question involves the use of principal component analysis on the well-known `iris` dataset. The dataset is available in R.

   (a) How many observations are there in the dataset? What are the different fields/attributes in the data set?

   (b) Create a new dataset `iris_data` by removing the `Species` column and store its content as `iris_sp`.

   (c) Compare the various pair of features using a pairwise scatterplot and find correlation coefficients between the features. Which features seem to be highly correlated?

   (d) Conduct a principal component analysis on `iris_data` without standardizing the data. You may use `prcomp(..., scale=F)`.

      (i) How many principal components are required to explain at least 90 % of the variability in the data? Plot the cumulative percentage of variance explained by the principal components to answer this question.

      (ii) Plot the data along the first two principal components and color the different species separately. Does the first principal component create enough separation among the different species? To plot, you may use the function `fviz_pca_ind` or `fviz_pca_biplot` in `library(factoextra)`. Alternatively, you may use `biplot` or construct a plot using `ggplot2` as well.

   (e) Do the same exercise as in (d) above, now after standardizing the dataset. Comment on any differences you observe.

7. This problem involves the dataset `wine_italy.csv` which was obtained from the University of Irvine Machine Learning Repository. These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different *cultivars*. The analysis determined the quantities of 13 constituents found in each of the three types of wines. The first column identifies the cultivars and the next thirteen are the attributes given by:

   - `alcohol`: Alcohol
   - `malic`: Malic acid
   - `ash`: Ash
   - `alkalin`: Alkalinity of ash
   - `mag`: Magnesium
   - `phenols`: Total phenols
   - `flavanoids`: Flavanoids
   - `nonflavanoids`: Nonflavanoid phenols
   - `proanth`: Proanthocyanins
   - `color`: Color Intensity
   - `hue`: Hue

- `od280`: OD280/ OD315 of diluted wines
- `proline`: Proline

(a) Check the relationship between the variables by creating a pair-wise scatterplot of the thirteen attributes.

(b) Conduct a principal component analysis on the standardized data. What proportion of the total variability in the data is explained by the first two components?

(c) Plot the data along the first two principal components and color the different cultivars separately. Also plot the loadings of the different components to show the importance of the different attributes on the first two principal components?

  (i) Which two key attributes differentiate Cultivar 2 from the other two cultivars?

  (ii) Which two key attributes differentiate Cultivar 3 from the other two cultivars?

(d) Use an appropriate plot to find the number of attributes required to explain at least 80% of the total variation in the data. How many principal components would you pick to explain the variability in the data reasonably?