Project 01: Campus Recruitment Data Analysis

Table of Contents

Ch. 1 Introduction

1.1 Overview

- Campus placement or campus recruiting is a program conducted within universities or other educational institutions to provide jobs to students nearing completion of their studies. In this type of program, the educational institutions partner with corporations who wish to recruit from the student population.

1) About Dataset

- This data set consists of Placement data of students in a XYZ campus. It includes secondary and higher secondary school percentage and specialization. It also includes degree specialization, type and work experience and Salary offers to the placed students.
- Link for source: https://www.kaggle.com/benroshan/factors-affecting-campus-placement

2) Variable name description

- Variables in this dataset have arbitrary acronyms created by the provider. According to his description, each variable refers to as follows:
    - SSC_P: Secondary Education (%) 10th Grade
    - SSC_B: 10th Board of Education
    - HSC_P: Higher Secondary Education (%) 12th Grade
    - HSC_B: 12th Board of Education
    - HSC_S: Specialization in Higher Secondary Education
    - Degree_P: Undergraduate (%)
    - Degree_T: Undergraduate degree type
    - WorkExp: Work experience
    - Ptest_P: Placement test (%)
    - Specialisation: MBA Specialisation
    - MBA_P: MBA (%)
    - Status: Hiring Status
    - Salary in US dollars

3) Explore Dataset

- Check codes and results of **01_Dataset Exploration.sas**
- Some missing values are spotted in the Salary due to non-placement.

1.2 Objective

- **The objective of this project is to find out what factors influenced a candidate in getting placed by conducting statistical tests such as ANOVA, linear regression, and logistic regression.** I will first set up an explanatory model to figure out what relationships exit between factors.

Ch. 2 ANOVA and Regression

2.1 Overview

- When we look at the campus recruitment data, several variables can affect the student's placement and Salary. There are 7 categorical variables, including binary variables, and 5 continuous (number) variables except for Status and Salary. Because so many potential predictors could be important in modeling placement and Salary, I need tools to explore and help me choose which predictors might be important.
- I will start my analysis using some graphical tools that can help me determine which predictors are likely or unlikely to be useful. These graphical explorations can be augmented with correlation analysis that describe linear relationships between potential predictors and our response variable. After I determine potential predictors, tools like ANOVA and regression will help me assess the quality of the relationship between the response and predictors.

2.2 Graphical Analysis of Association

1) Scenario

- I want to start by exploring the relationships between the Salary of students who get placed and other predictors, such as Specialization in Higher Secondary Education, Undergraduate degree type, or Work experience. To graphically explore these relationships, or associations between variables in the data, I will use boxplots and scatter plots.

2) Identifying associations in ANOVA with boxplots

- Check the codes and results of **Part A in 02_Graphical Analysis.sas**
    - We can see the differences in Salary in each categorical variable. The Salary is slightly higher for students who are male, has specialization in science in higher secondary education, has science and technology major as an undergraduate, has work experience, and has specialization in marketing and finance in their MBA.

3) Identifying associations in linear regression with scatter plots

- Check the codes and results of **Part B in 02_Graphical Analysis.sas**
    - We can see almost all regression lines fit to the scatter plot are horizontal, which means there seems to be no association between each percentage and Salary except for the MBA (%).

2.3 One-way ANOVA

1) Scenario

- I am interested in finding whether the major of undergraduate students affects the Salary of students who get employed.
- I use one-way analysis of variance to determine whether the mean Salary of students who get employed is equivalent for all three levels of the categorical variable, Degree_T, the undergraduate degree type.
- In this ANOVA, the goal is to determine whether there are significant differences among the group means.

2) Performing a one-way ANOVA

- Check the codes and results of **03_One-Way ANOVA.sas**
- Before I can trust the p-value for our model, I need to assess the assumptions, so let's look at the diagnostics plots.
    - The first assumption, which is independent observations, is presumed satisfied.
    - Normality of error terms can be checked by looking at the residual histogram and Q-Q plot. The histogram is slightly right skewed, but, considering a few outliers, the histogram is approximately symmetric.

- Equal variances can be confirmed by looking at the Levene's Test for Homogeneity of Salary Variance Table. The p-value of 0.9129 is not smaller than the alpha level of 0.05, and therefore, I do not reject the null hypothesis that the variances are equal for all degree types. Evidence suggests that the variances within each group of degree types are not significantly different.

- Because I have met the model assumptions of independence, normal residuals, and constant variance, now I can trust the results of my analysis. Since the p-value of F statistics is 0.1118, which is larger than the alpha level of 0.05, I conclude that there are not statistically significant differences in Salary of students who get employed among degree types with different majors. At this point, I can conclude that among degrees with communication & management, others, and science & technology, none of these group is different.

2.4 Pearson Correlation

1) Scenario

- In the previous sections, I used ANOVA to assess the importance of a categorical predictor to a continuous response variable, Salary. I would like similar tools for assessing the importance of continuous predictor variables as well.
- I want to build linear regression models that relate a continuous response variable to several continuous predictors. Before building such models, it can be helpful to use correlation analysis to test for linear associations among the continuous variables.
- I want to determine which continuous variables in the dataset are correlated with Salary before I create a regression model. Because correlation is a measure of the linear association between two variables, identifying correlations provides information about how well a continuous predictor will explain the response within a regression analysis.
- I will use the Pearson correlation coefficient as my correlation statistic.

2) Producing correlation statistics and scatter plots

- Check the codes and results of **Part A in 04_Correlation Analysis.sas**
  - When looking at the Pearson Correlation Coefficients Table, I see the correlation coefficient of each predictor variables rank-ordered from highest to lowest left to right. Considering that the value becomes close to -1 or 1, it shows stronger correlation, all correlation coefficients given here range from -0.019 to 0.178 indicating no or very little correlations with the response variable Salary. Relatively, the placement test percentage and MBA percentage show higher correlation compared to the rest predictor variables. Also, we can confirm these findings looking at each scatter plot presenting almost horizontal line. Overall, the correlation and scatter plot analyses indicate that none of the variables might be good predictors for Salary.

3) Producing correlation statistics among potential predictors

- When preparing to conduct a regression analysis, it is always a good practice to examine the correlations among the potential predictor variables. Because strong correlations among predictors included in the same model can cause a problem such as multicollinearity.
- I want to produce a correlation matrix to help me compare the relationships between predictor variables. The correlation matric shows correlations and p-values for all combinations of the predictor variables. I will limit my attention to the strongest 3 correlations with each predictor.
- Check the codes and results of **Part B in 04_Correlation Analysis.sas**
  - The Pearson Correlation Coefficients Table indicates that there are moderately strong correlations between SSC_P and MBA_P, 0.43056, between Degree_P and MBA_P, 0.49409, between HSC_P and MBA_P, 0.32998.

2.5 Simple Linear Regression

1) Scenario

- Simple linear regression can be used to more fully describe the relationship between two continuous variables, as opposed to scatter plots and Pearson correlations. Regression model parameter estimates not only define the line of best fit corresponding to the linear association between variables, but also describe how a change in a predictor corresponds to a change in the response.
- In simple linear regression, the goal is to identify the equation that characterizes the linear association between the predictor variable and the response variable and use the model to then estimate the response for a given value of the predictor.
- To practice performing simple linear regression, I will build a model using MBA_P as the predictor and Salary as the response in order to determine how exactly MBA_P and Salary are linearly related.

2) Performing Simple Linear Regression

- Check the codes and results of **05_Simple Linear Regression.sas**
- The p-value and R-square of the predictor variables

Variable	P-value (F statistic = t statistic)	R-square
SSC_P	0.6699	0.0012
HSC_P	0.3534	0.0059
Degree_P	0.8162	0.0004
Ptest_P	**0.0301**	**0.0318**
MBA_P	**0.0334**	**0.0306**

- The ANOVA table reports the p-value to evaluate the null hypothesis. Since the p-value of the t statistics and the p-value of the F statistics are identical in the simple linear regression, the p-value indicates how each simple linear regression model fits the data than the baseline model and whether there is a significant linear relationship between the response and predictor variable.
- Ptest_P and MBA_P have the p-value smaller than the alpha level of 0.05 meaning that there is a significant linear relationship between Ptest_P and Salary and between MBA_P and Salary.
- The R-square is a measure of the proportion of variability in the response variables explained by the predictor variables in the analysis. For example, the R-square of MBA_P, 0.0306 indicates that the MBA_P explains 3.06% of the variability in the response variable Salary.
- The Parameter Estimates table specifies the individual pieces of the model equation based on the data.
    - For example, the parameter estimate for the intercept is 199762 and the parameter estimate for the slope of Ptest_P is 1213.76231. So, the regression equation is Salary = 199762 + 1213.76231*Ptest_P. The model indicates that each additional percentage of placement test is associated with an approximately $1231.76 higher Salary.
- For a simple linear regression analysis to be valid, 4 assumptions need to be met.
    - The first assumption is that the mean of the response variable is linearly related to the value of the predictor variable. When looking at the residual plot of Ptest_P and MBA_P, even though there are a few outliers above 200000, the data hovers around the regression line, so a regression line does in fact adequately describe the data.
    - The rest of assumptions are about the error term. When looking at the residual vs. predicted value plot, except a few outliers above 200000, there is a random scatter of the residual values above and below the reference line at 0. This indicates the equal variances of the error terms.
    - When looking at the residual histogram and the Q-Q plot, I can verify that the errors are normally distributed.

Ch. 3 Categorial Analysis

3.1 Overview

- In the previous chapters, I used the ANOVA and the simple regression for the coding practice rather than for answering the question I raised, which is the ultimate purpose of this project. In this chapter, however, I can answer the question that I set up myself, "what factors influenced a candidate in getting placed?" by using a logistic regression.
- Logistic regression is used to model the relationship between a binary response variable and a set of predictor variables. It is used to estimate the probability of the response according to the various continuous and categorical predictors.
- In this final chapter, I will first look for associations between predictors and a binary response ('Placed', 'Not placed'). I will then build a logistic regression model and discuss how to characterize the relationship between the response and predictors.

3.2 Describing Categorical Data

1) Scenario

- In the placement dataset, I chose predictor variables that I assume would heavily affect the job placement of students. For categorical variables, I selected Gender, degree type, work experience, and specialization at MBA. For continuous variables, degree percentage, placement test percentage, and MBA percentage were chosen.
- I want to first look for associations between those predictors to see which variables should be considered for model inclusion. Then I want to use the logistic regression to determine which students have a high probability of getting themselves placed.

2) Association between categorical variables

- By examining distributions of categorical variables, I can determine the frequencies of data values and possible associations among variables.
- An association exists between two categorical variables if the distribution of one variable changes when the value of the other variable changes. However, if there is no association, the distribution of the first variable is the same, regardless of the level of the other variable.
- To look for associations, I examine the frequencies of values across the combinations of variables.
- Which variables are associated with the response variable Status?
    - Is a student's Gender associated with their employment? Is a student's work experience associated with their employment? Is a student's major associated with their employment?

3) Examining the distribution of categorical & continuous variables

- Categorical variables
    - Check the codes and results of **Part A in 06_Categorical Analysis.sas**
    - Gender by Status
        - It seems there might be no association between variables Gender and Status.
        - With the unequal group sizes, the row percentages might not easily display if Gender is associated with placement Status.
    - Work experience by Status
        - It seems there might be association between variables WorkExp and Status.
        - Students without work experiences are much more likely not to get placed, at about 85%. Even though the percentage of students without work experiences are slightly likely to get placed, but the percentage dramatically decreased.
    - Degree type by Status

        - It seems there is no association between degree type and Status.
        - Regardless of the placement Status, the percentage of students with the same degree type are the same.

- Continuous variables
    - Check the codes and results of **Part B in 06_Categorical Analysis.sas**
    - Degree_P by Status
        - There certainly appears to be an association between the degree percentage and placement Status. The students with more than 70% are more likely to get placed. The histograms are different and centered in different locations. The median degree percentage of students who get placed is 7% higher than that of students who do not get placed.

3.3 Tests of Association

1) Scenario

- I explored the distribution of the variables, Status, Gender, WorkExp, and Degree_P and saw some possible associations of the two predictors with the response using crosstabulation tables and histograms.
- I need to assess whether the differences between the percentages of Salary across levels of the predictors is greater than would be expected by chance.
    - To be certain that the variables are associated, I need to run a formal test of association, the chi-square test.
    - To measure the magnitude of an association, I will use measures of association, such as Cramer's V statistic and an odds ratio.

2) The Pearson Chi-Square Test (Cramer's V Statistic, Odds ratio)

- I will start with my null hypothesis that there is no association between the variables WorkExp and Status, meaning that the probability of getting placed is identical regardless of work experience.
- The alternative hypothesis is that there is an association between WorkExp and Status, meaning that the probability of getting placed is not the same for students with work experience and without work experience.
- Check the codes and results of **07_Tests of Associations.sas**
    - It seems that the cell for WorkExp, Yes and Status, Not Placed relatively contributes the most to the chi-square statistic, with a Cell Chi-Square value of 7.3969.
    - The next table shows the chi-square test and Cramer's V. Because the p-value for the Chi-Square statistic is less than .0001, I reject the null hypothesis at the 0.05 level and conclude that there is evidence of an association between WorkExp and Status.
    - The Cramer's V value of 0.2761 indicates that the association detected with the chi-square test is relatively weak.
    - Before interpreting what the odds ratio means and there a few ways to interpret the odds ratio, it is helpful to understand what I am trying to answer the question by calculating it. That is, **"Are candidates with NO work experience is more likely NOT to get placed than candidates with work experience?"**
        - The odds ratio value of 4.3429 indicates that the candidates with NO work experience is about 4 times more likely NOT to get placed than the candidates with work experience.
        - In addition, the 95% odds ratio interval goes from 2.0586 to 9.1619, which does not include 1.

- This confirms the statistically significant result of the Pearson chi-square test of association. A confidence interval that includes the value of 1 would indicate equality of odds and would not be a significant result.

3.4 Logistic Regression

1) Scenario

- Logistic regression is used to model the relationship between a binary response and a set of predictor variables. Because the response is categorial, we estimate the probability of the response given the various categorical and continuous predictors.
- For example, I have discussed tests to find significant associations between the Status variable and a few candidate predictors.
  - But what is the increase in the probability of getting placed if an identical candidate is female or male, or has work experience or not, or has different degree percentage?
- Once I build my logistic regression model, I can then use the estimated probabilities to predict whether a candidate will get placed.

2) Fitting a Binary Logistic Regression Model

- I try to fit a binary logistic regression model to characterize the relationship between the continuous variable Degree_P and the categorical response, Status.
- Check the codes and results of **08_Fitting Logistic Regression.sas**
  - First, the Response Profile table shows the response variable values listed according to their ordered values. Because I used the EVENT= option, the model is based on the probability of getting placed (Placed = 1). This table also shows frequencies of response value. In this sample of 215 candidates, 148 candidates get placed and 67 candidates do not.
    - Also, I should always check that the modeled response level is the one I intended.
  - Next, the Model Convergence Status simple indicates that the convergence criterion was met. it is always significant to see that the convergence criterion is satisfied.
  - The Model Fit Statistics table reports the results of three tests: AIC, SC, which is also known as Shcwarzs Bayesian Criterion, or SBC, and -2 Long L, which is -2 times the natural log of the likelihood.
    - The AIC, SC, and -2 Long L are goodness-of-fit measures. These statistics measure relative fit and are used only to compare models.
  - The Global Tests table, Testing Global Null Hypothesis: BETA=0, provides three statistics to test the null hypothesis that all regression coefficients of the model are 0.
    - A significant p-value for these tests provides evidence that at least one of the regression coefficients for a predictor variable is significantly different from 0.
  - The Parameter Estimates table, Analysis of Maximum Likelihood Estimates, lists the estimated model parameters, their standard errors, Wald Chi-Square values, and p-values.
    - The parameter estimates are the estimated coefficients of the fitted logistic regression model. For this example, the logistic regression equation is:
      - logit(p-hat) = -11.9376 + 0.1996*Degree_P.
    - The Wald Chi-Square and its associated p-value tests whether the parameter estimate is significantly different from 0. The p-value of less than .0001 for the variable Degree_P is significant at the 0.05 alpha level.
  - The estimated model is displayed on the probability scale in the effect plot. The sigmoidal shape of the estimated probability curve is observed and the probability of getting placed increases as the degree percentage increases.

- For a continuous predictor variable, such as Degree_P, the odds ratio measures the increase or decrease in odds associated with a one-unit difference of the predictor variable.
  - The odds ratio for Degree_P indicates that the odds of getting placed increase by 21.7% for each increase in 1% of a candidate's degree percentage.
  - Because the 95% confidence interval, 1.148 to 1.300, does not include 1.000, the odds ratio is significant at the 0.05 alpha level and therefore, the predictor Degree_P is significantly different from 0.
- The Odds Ratio plot displays the results of the Odds Ratio table graphically. This plot is obtained by applying the parameter estimates from the logistic model to values of the predictors, and then converting the predictions to the probability scale.
  - A reference line shows the null hypothesis, an odds ration equal to 1.
    - When the confidence interval crosses the reference line, the effect of the variable is not significant.
- The Association of Predicted Probabilities and Observed Responses table displays the four rank correlation indices that are computed from the numbers of concordant, discordant, and tied pairs of observations.
  - In general, a model with higher values for those indices has better predictive ability than a model with lower values.
  - The c, concordance statistic, which is most commonly used, estimates the probability of an observation with the event having a higher predicted probability than an observation without the event.
    - The range of possible value is 0.5 to 1.0, where 1.0 is perfect prediction. The value of 0.808 shows a strong ability of Degree_P to discriminate between candidates getting placed and candidates not getting placed.

3.5 Multiple Logistic Regression

1) Scenario

- As a linear regression I can also include several predictors, both continuous and categorical, into my logistic regression model. My goal is to build the best model I can explain the relationship between getting placed and candidate's factors.
- To do this, I want to consider a more complex model with many possible predictors to model the relationship jointly, and account for possible interactions between the effects.
- I will add in the categorical variables Gender and WorkExp. **How do these three variables contribute simultaneously estimate the probability of getting placed for a candidate?**

2) Fitting a Multiple Logistic Regression Model with Categorical Predictors

- I will fit a multiple logistic regression model that characterizes the relationship between the response Status, and the variables Degree-P, Gender, and WorkExp.
- Check the codes and results of **Part A in 09_Fitting Multiple Logistic Regression.sas**
  - The Model information and Response Profile are the same as for the binary logistic regression model that I ran in the previous chapter.
  - The Class Level Information table includes the predictor variables in the Class statement.
    - Because I used the PARAM=REF and REF='No' options, this table reflects my choice of WorkExp = 'No' as the reference level. The design variable is 1 when WorkExp='Yes' and 0 when WorkExp='No'. The reference level for Gender is Female, which is 0.
  - The SC value in the Degree_P only model was 219.023. Here, it is 210.474. Considering that smaller values imply better fit, I can conclude that this model fits better.

- In the Global Tests table, Testing Global Null Hypothesis: BETA=0, I see that this mode is statistically significant, indicating at least one of the predictors in the model is significantly different from 0.
- The Type 3 Analysis of Effects table is generated when I use the Class statement.
  - This table displays the significance of each of the effects individually, adjusting for other predictors included in the model.
  - According to each p-value of all three predictor variables, which is less than 0.05 alpha level, I can conclude that all predictor variables are statistically significant.
- In the Parameter Estimates table, Analysis of Maximum Likelihood Estimates, for CLASS variables, effects are displayed for each of the design variables. Because reference cell coding was used, each effect is measured against the reference level.
  - For example, the estimate for Gender | M shows the difference in logits between male and female candidates, which is 0.9339.
  - All effects are statistically significant.
- In the Association of Predicted Probabilities and Observed Response table, the c statistic value is 0.849 and the percent concordant is 84.7% for this model, indicating that 84.7% of the positive and negative response pairs are correctly sorted using Degree_P, Gender, and WorkExp.
- The Odds Ratio Estimates table shows that, adjusting for other predictors variables,
  - candidates with work experience had about 4 times the odds of getting placed than candidates without work experience.
  - male candidates had about 2.5 times the odds of getting placed than female candidates.
  - the table shows that for each 10 percentage increase in degree percentage, the odds of getting placed increases by 712.8% [(8.128 – 100)*100 = 7.128*100 = 712.8]
- Notice that the confidence interval for all variables' odds ratios do not cover a value of 1, indicating the odds ratios are statistically significant for all variables.
- The final plot is an effect plot of Predicted Probabilities for Status=1.
  - It shows the estimated probability of getting placed across different degree percentage, given the different combinations of the categorical variables Gender and WorkExp.
  - For example, the left most probability curve corresponds to male candidates with work experience.
    - This means that as the values of Degree_P increases, the probability of getting placed increases faster than any other combinations of the categorical variables.
    - However, as the degree percentage increases past 70%, each curve shows a high probability of getting placed.

3.6 Stepwise Selection with Interaction and Predictions

1) Scenario

- I consider more complex logistic regression models to find the best fit, most predictive, and most generalizable model. I might start by adding more predictors into the model and considering possible interaction effects between them.
- For example, does the effect of Degree_P depend on the variable Gender, WorkExp, or both? As I build more variables into the model, I need to be careful of overfitting the data. To systematically build the model and remove ineffective predictors, I can use stepwise selection methods.

2) Fitting a Multiple Logistic Regression Model with Interactions

- I fit a multiple logistic regression model using backward elimination method.
- Check the codes and results of **Part B in 09_Fitting Multiple Logistic Regression.sas**
  - The Class Level Information table is the same before, and this model converged, so I can trust the results here.

- Let me jump to Summary of Backward Elimination table to view the stepwise process.
  - At Step 1, the interaction between Gender and WorkExp was removed because of its least significant p-value of 0.9024, which is larger than the alpha level of 0.10
  - At Step 2, the interaction between Degree_P and WorkExp was removed for the same reason.
  - The procedure stops after the two interactions involving WorkExp are removed.
- The next table, the Joint Tests test that all the parameters associated with that effect are 0 for the final model.
  - I notice that the main effects are included, but only the interaction for Degree_P by Gender is included. Also, individually, each effect is significant at the alpha = .10 level.
- The Parameter Estimate table, Analysis of Maximum Likelihood Estimates, displays the estimates and significance for each parameter in the final model.
  - I notice that all left parameters have the same p-values in the Joint Tests table.
- In the Association Statistics table, Association of Predicted Probabilities and Observed Responses, I see that the c value is a slight improvement over the previous model that only included the main effects;
  - C statistic: 0.849 > 0.862
  - Percent Concordant: 84.7% > 85.9%
- Odds rations are not calculated for effects involved in interactions.
  - Any single odds ratio for Degree_P or for Gender would be misleading, because the effects vary for each at different levels of the other variable.
  - The odds of getting placed are more than 3 times the odds for candidates with work experience compared to the odds for candidates without work experience.
- In the effect plot, Predicted Probabilities for Status=1, I notice the sigmoidal probability curves overlap and are not all equally shaped. This visually displays the Degree_P by Gender interaction discovered with backward elimination. The effect of Degree_P depends on the gender of a candidate, causing an interaction effect.

3) Fitting a Multiple Logistic Regression Model with All Odds Ratios

- In this sub-chapter, I want to refine the multiple logistic regression model that I fit in the last sub-chapter. Now I want to produce the odds ratios for each value of the variables that are involved in an interaction from the final model.
- Check the codes and results of **Part C in 09_Fitting Multiple Logistic Regression.sas**
  - Let me jump to the Odds Ratios. There are four odds ratios displayed for the interaction effects.
    - The first two show the odds ratios comparing candidates with a difference of 10 percentage of their degree holding the gender constant.
      - For example, the odds of getting placed for a male candidate are more than 20 times greater than a male candidate with 10% less of the degree percentage.
    - The last two odds ratio compare the odds ratios for female candidates compared to male candidates when holding the degree percentage constant.
      - For instance, the odds of getting placed are more than 7 times greater for male candidates vs female candidates when holding the degree percentage at 70%.
  - From the effect plot, it is clear that the gender effect is different at different values of degree percentage.
    - The gender effect is highly significant when the Degree_P is set to 70%, but not when Degree_P is set to 50%, as I can see the confidence interval for the odds ratio convers a value of 1.

4) Generating Predictions

- I use the multiple logistic regression model with the same effects as before, but this time, I add in the STORE statement to save the model information and score new data.
- Check the codes and results of **10_Generating Predictions.sas**
    - As expected, the Predictions table shows that a candidate with the highest predicted probability of getting placed (0.99669) is male, has work experience and a degree percentage of 76.
    - The candidate with the lowest predicted probability (0.07912) is also male, has work experience but has a degree percentage of 49.