HWK 6- Regression Practice

For this assignment you will need to submit the code in an R Markdown file (rmd) and the completed knitted HTML file.

The data: You will need the following data sets to complete this assignment.

1. HWK6 Data, which is in Excel format.

Problem 1: In the spirit of Women’s History Month. A researcher interviews 50 employees of a large manufacturer and collect data on each worker’s hourly wage (wage in $), years of higher education (EDUC), years of experience (EXPER), age (AGE), and Male (1=male, 0=female). The primary goal of collecting this data is to determine how Wage is related to the other variables. Explore the use of linear regression to determine the strongest possible relationships and answer the following questions.

1. Create a R-Markdown file name the File: Lastname\_HWK6
2. Read in the Wage data from the provided Excel File. It will be under the Wage worksheet within the excel file. Assure all the data has was read in correctly into the R environment. Display the first 5 rows of the data in the R-Markdown file.
3. Estimate the model for the data. Print the summary output and write in the estimated values for the coefficients into the model. Provide the R^2 and the Adjusted R^2 statistic for the model. Is it a good model? Explain.
4. Predict the hourly wage of a 40-year-old male employee with 10 years of higher education and 5 years of experience. Predict the hourly wage of a 40-year-old female with the same qualifications. Is there a difference explain this difference? In addition, assure that this explanation is within the document.
5. Interpret the estimated coefficient for Male. IS the variable Male significant at the 5% level. This can be determined through the p-value provided in the summary associated with this variable. Do the data suggest that sex discrimination exists at this firm? Explain?
6. Now, explore one variable at a time. This means you must generate a linear regression model for the following combinations. Using R to estimate each of these models. Create a table in the R markdown, see this page to create the table, [Markdown Tables generator - TablesGenerator.com](https://www.tablesgenerator.com/markdown_tables). The Table will look like this (as an example only):

|  |  |  |
| --- | --- | --- |
| Variable | Coefficients (intercept, Slope) | R Squared Statistic |
| Educ | (12, .05) | .85 |

To get the values in the table. Use the in text code ` r code-to-obtain-number` , thus first compute all the values and name them so that you can recall them in the table. For example, if the R^2 value for the first model is M1R2 then you can recall the number in the table by `r M1R2`. The table can not be in a code block, it must be part of the text in the R-Markdown.

1. Based on task 7, which is the best variable to predict Wage, this can be determined by which model explains the most variation in Wage by itself. Interpret the intercepts of each model. Do the intercepts of each model differ? If so, explain what this means. Lastly, once you’ve determined which model is the best, create a scatter plot and plot the model along with the data as well as the residual analysis plots provided by plot(fit), where fit is the results stored from the lm function.
2. (Moderate Task) Now explore all possible two variable combinations. These are:

Which of these models has the best Adjusted R Squared? Is the highest adjusted R^2 higher than the regular R^2 squared value from task 6. Create a table like the one in task 6 for these three models.

1. (Moderate Task) Now, that you know which additional variable enhanced the model in task 8. Use that model to check all possible three variable combinations. For Example (example only), if the best model is the first model of Education and Expert then the only three you need to check are: (educ, Expert, Age) and (educ, Expert, Male). Report the adjusted R squared values for these three models in a table like in tasks 6 and 8.
2. Overall, which is the best model? Compare all the results including the model that includes all variables in tasks three. Remember, when comparing models with more than one variable we use Adjusted R Squared, when using single variable, we just use the regular R^2.