# Part 1 – ANOVA (10.5 points)

The General Social Survey collects data on demographics, education, and work, among many other characteristics of US residents. Using ANOVA, we can consider educational attainment levels for all 1,172 respondents at once. Below are the distributions of hours worked by educational attainment and relevant summary statistics that will be helpful in carrying out this analysis.

Chart, box and whisker chart

Description automatically generated

Table

Description automatically generated

1. (2 points) Write the null and alternative hypotheses for evaluating whether the average number of hours worked varies across the five groups.

**Null Hypothesis:**

**Alternate Hypothesis:** At least 1 group’s average number of hours varies from the other groups.

1. Using the information provided, assess whether the following conditions necessary to accurately perform an ANOVA F test are met:
   1. (0.5 point) Are the observations in the study independent?

The observations in the study are independent between and within groups as the data for each group is randomly sampled.

* 1. (0.5 point) Are the sample sizes sufficiently large? (Hint: the *n* row of the table above provides the sample sizes of each group.)

Yes, the sample sizes for each individual group are sufficiently large.

* 1. (0.5 point) Is the variation in the groups about equal from one group to the next? (Hint: use the spread of the boxplots and standard deviation values from the table to assess this condition.)

The variation among the groups is roughly equal to one another as the boxplot is quite spread out and the median of the boxplots are roughly equal. The standard deviations between the groups are close to each other.

1. To assess whether there is a significant difference in the average number of hours worked between one or more of the groups, we need to determine the mean squares between groups (MSG) and the mean squares within groups (MSE). Each of these values has an associated degrees of freedom.
   1. (1 point) Determine the degrees of freedom associated with the MSG.

* 1. (1 point) Determine the degrees of freedom associated with the MSE.

1. (1 point) An ANOVA was performed in R. The estimate for the mean squares between groups is MSG = 501.54 and the resulting *F* statistic is equal to 2.189. Determine the average variation within each group. That is, calculate the MSE.
2. (2 points) Using the *F* statistic from question 4 and the two values for the degrees of freedom in question 3, calculate the p-value for this test.

p-value = 1 – pf(2.189, 4, 1167) = 0.068.

1. (2 points) Using the p-value calculated in question 5, write a conclusion for this ANOVA F test using a significance level of .. (Hint: your conclusion should include a statement of evidence in favor of the alternative and a statement as to whether the null hypothesis is rejected or not.)

There is slightly suggestive to suggest that the average number of hours worked varies across the five groups. So, we fail to reject the null hypothesis at the 0.05 significance level.

# Part 2 – Simple Linear Regression (12.5 points)

In this lab we’ll be looking at data from all 30 Major League Baseball teams in 2011 and examining the linear relationship between runs scored in a season and the number of at-bats for each time. Our aim will be to summarize this relationship both graphically and numerically in order to determine if the number of at-bats helps us predict a team’s runs scored in a season.

For this part of the assignment, please refer to the DA8\_ SLR.R script and the mlb.csv file available on the Canvas page for this assignment. Be sure to read through all comments in the R script to understand how we visualize and fit the least squares regression line in R.

1. (1 point) We are interested in determining if the number of at-bats is a good predictor of the number of runs a team scores in the 2011 season. Identify the explanatory and response variables for this question of interest.

The explanatory variable is the “number of runs a team scores”, and the response variable is the “number of at-bats”.

1. (2.5 points) Using the R script, construct a scatterplot of the at-bat and runs variables. Paste your plot below. Using the scatterplot, describe the relationship between the number of at-bats and runs. Include the strength, direction, and form of the relationship. Comment on whether or not there are any outliers present in the data.

Scatter chart

Description automatically generated

There is a moderately strong, positive, linear relationship between the number of at-bats and runs. There are 3 outliers present in the data.

1. (0.5 points) Calculate and report the correlation coefficient between the two quantitative variables.



1. (2 points) Using the lm() and summary() functions provided to you in the R script, fit the least squares regression line that models the linear relationship between the number of at-bats and the number of runs scored. Write out the estimated least squares regression model. (Hint: your answer should be in the form where and are replaced with the estimates for the intercept and slope.

Text

Description automatically generated

1. (1 point) Interpret the slope estimate in the context of the problem.

For a 1 unit increase in x, we expect y to change by a value of 0.6305 (i.e., since = 0.6305 > 0, we expect y to increase with x).

1. Using the R script, construct a plot of the residuals for this model.
   1. (1 point) Paste your residual plot below.

Chart, scatter chart

Description automatically generated

* 1. (1.5 points) Use the residual plot to assess whether the conditions are met: linearity, normality in the response values, and constant variance about the regression line. Comment on whether or not you think each of these conditions are met and use the residual plot to support your comments.

**Linearity:** the relationship is linear (no “u” shape in residual plot)

**Normality:** random scatter above and below the 0 line in residual line

**Constant Variance:** no “funnel” shape in residuals

1. (1 point) Suppose a team expects 5,600 at-bats in one season. Use the least squares regression line from question 4 to predict the number of runs the team will score.

1. (1 point) Suppose the team in question 7 actually scores 728 runs in the season. Compute the residual for this team.

1. (1 point) Now suppose that a new team is joining the MLB and only expects 4,900 at-bats. Why is inappropriate to use this model to predict this team’s runs scored?

The minimum of the dataset is 5450 and 4900 is less than the minimum. Hence, it would be inappropriate to use this model to predict this team’s runs scored.

**Gradescope Page Matching (2 points)**

When you upload your PDF file to Gradescope, you will need to match each question on this assignment to the correct pages. Video instructions for doing this are available in the Start Here module on Canvas on the page “Submitting Assignments in Gradescope”. Failure to follow these instructions will result in a 2-point deduction on your assignment grade. Match this page to outline item “Gradescope Page Matching”.