# STAT 210
# Applied Statistics and Data Analysis:
# Homework 8

Due on November 23/2025

You cannot use artificial intelligence tools to solve this homework.

**Show complete solutions to get full credit. Writing code is not enough to answer a question. Your comments are more important than the code. Do not write comments in chunks. Label your graphs appropriately**

**For all tests in this homework use a significance level of $\alpha = 0.02$.**

## Question 1

A labor economist studies weekly wages (`Y`) and how they depend on several worker attributes. The researcher collected a sample of workers and recorded 7 candidate predictors. Your job is to build a regression model. The data is available in the file `HW825FQ1.csv`.

(a) Do a exploratory analysis of this data, including a scatterplot matrix and a graphical representation of the correlation matrix. Comment on your results.

(b) Fit a complete model for `Y` including all the other variables. Produce a summary table and interpret the $t$ tests in the table. What is the $p$-value for the overall significance test for the regression?

(c) Check for multicollinearity and drop variables as needed until this problem is resolved. Use a threshold value of 2.

(d) Starting with the model obtained in section (c), get a minimal model using a backward selection procedure with a critical $\alpha$ of 0.1. Use the function `drop1` for this.

(e) Starting with the full model obtained in section (b), fit a model using Akaike's Information Criterion (AIC). You can use the function `stepAIC` in the `MASS` library or the function `step` in the base package. Check the resulting model for multicollinearity. Compare with the model obtained in (d).

(f) Starting with the full model obtained in section (b), fit a model by maximizing the adjusted $R^2$. Check the resulting model for multicollinearity. Compare with the models obtained in (d) and (e). Are all the models the same? If not, which one would you choose and why?

(g) Plot the standard diagnostic graphs for the model that you fitted in (d) and comment on what you observe. Use also the Shapiro-Wilk and ncv tests and comment on the results.

(h) Predict the `Y` value for a subject with covariates

$$(\texttt{X1, X2, X3, X4, X5, X6, X7}) = (15.2, 9.6, 10.7, 9.2, 5.33, 1, -0.8)$$

using the model you fitted in (d). Add a confidence interval at level 98%.

---

## Question 2

A city transportation department is studying bike rental duration. They believe that the effect of age on rental duration differs between:

- Casual riders (`member` = 0)
- Registered members (`member` = 1)

To study this, they collect data on 30 riders. The data is in the file `HW825FQ2.csv`. Remember to transform `member` into a factor.

(a) Fit a simple regression model for `duration` in terms of `age`. Print the summary table and comment on the results. Draw a scatterplot and add the regression line. Comment. Using diagnostic plots and test, check whether the assumptions for the model are satisfied. Predict the value of `duration` for a value of `age = 45` with this model and include confidence intervals at the 98% level.

(b) Draw a new scatterplot and color the points according to the value of `member`. Comment on what you observe.

(c) Fit a new model for `duration` as a function of `age` and `member`, including an interaction term. Print an anova table for this model and interpret the $p$-values in the table. If necessary, fit a new model dropping the terms that have a non-significant $p$-value. Print a summary table for the final model and interpret the coefficients. What is the value for the estimated variance of the errors? What is the $R^2$, how do they compare with the previous model?

(d) Draw a scatterplot and color the points according to the value of type. Add the regression lines corresponding to the model you fitted in (c). Write down an equation for this model.

(e) Plot the standard diagnostic graphs for the model that you fitted in (c) and comment on what you observe. Use also the Shapiro-Wilk and ncv tests and comment on the results.

(f) Predict the value of `duration` for a value of `age = 45` and for the two levels of `member`. Include confidence intervals at the 98% level. Compare with the prediction in (a).