PSTAT220A Homework 2

Due 10/26/2025

Problem 1 (35 pt)

The dataset `Salaries` in library `carData` concerns the salaries for Professors in 2008-2009. Please read the `Salaries` documentation for a full description of the data.

a) Make a numerical and graphical summary of the data, commenting on any features that you find interesting

b) Fit a linear regression model with the salary as the response. Which variables, if any, are significantly associated with salary?

c) Compute LS estimates in R using the matrix solution to the least squares problem and confirm you get the same estimates as those in (b). To generate the covariates matrix it may help to use the `model.matrix` function.

d) What percentage of variation in the response is explained by the covariates? Explain whether you use the unadjusted or adjusted measure in your answer and why.

e) Which observation has the largest absolute residual (give the row number)?

f) Report separate 99% confidence intervals for the coefficients associated with `yrs.since.phd` and `yrs.service`.

g) Plot a 95% confidence region for the coefficients associated with `yrs.since.phd` and `yrs.service`. Comment on the resulting shape and why this makes sense.

h) Construct pointwise and simultaneous 95% confidence band for the prediction of future mean response and the prediction of future observations

i) Compute the partial coefficient of determination for `yrs.since.phd`. Interpret the meaning of this quantity.

j) Construct the EHW heteroskedasticity-consistent standard errors for the regression coefficients. Comment on the comparison between these standard errors to those returned by `lm`. In you response, reference any evidence for (or against) heteroskedasticity.

k) What are the highest leverage and highest influence points?

l) Are the residuals approximately normally distributed? If not suggest a transformation of the outcome that might improve the model ift.

```
## Your code here
```

Problem 2 (25 pt)

A series of $n+1$ observations y_i $(i = 1, \dots, n+1)$ are taken from a normal distribution with unknown variance σ^2. After the first n observations it is suspected that there is a sudden change in the mean of the distribution. That is, assume the first n observations are iid $y_1, ..., y_n \sim N(\mu, \sigma^2)$ the $y_{(n+1)} \sim N(\mu + \delta, \sigma^2)$.

a) Write this model in the matrix form $y = X\beta + \epsilon$

b) Derive the LS estimates of μ and δ

c) Derive a test statistic for testing the hypothesis that the $(n+1)$st observation has the same population mean as the previous observations, that is, the two mean parameters are equal.

d) Assume that $\sigma^2 = 1$ and $\delta = 2$. Simulate the distribution of the test statistic under this alternative hypothesis and compute the power of the test to detect $\delta \neq 0$ by counting the fraction of times the test statistic rejects. Assume you design your test with Type I error of 5% and are conducting a 2-sided test.

```
## Your code here
```

Problem 3 (25 pt)

In this problem we'll conduct a simulation to confirm and explore some important theoretical results.

a) Use simulation to confirm that $\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{(X^T X)^{-1}_{ii}}} \sim t_{n-p}$.

b) Compute the coverage of the associated confidence intervals for parameters. Do they have the desired coverage?

c) Evaluate the performance of hypothesis tests of parameters. Discuss type I errors and powers.

d) Repeat a)-c) assuming that the Gaussianity assumption is violated by generating non-Gaussian random error. Run one set of simulations with symmetric but heavy-tailed residual distribution and another with a skewed residual distribution. How did these violations influence your results? Which violation appeared worse (heavy-tailed or skewed)?

```
## Your code here
```

Problem 4 (15 pts)

This problem concerns the `divusa` data in the `faraway` library.

a) Make a well-constructed visualization showing how divorce rate is changing over time. Does it appear to be steady, going up, or going down?

b) Fit a regression model with divorce as the response and remaining variables as covariates. Interpret the coefficient on `year` (include units). How can you reconcile this result with the answer to the previous part?

c) Why might observations be correlated? Make two graphical checks for correlated errors. What do you conclude?

d) Conduct a statistical test the presence of autocorrelation.

```
## Your code here
```