# Homework 6 Graded

Sheridan Grant

Must be uploaded to Canvas under "Homework 6 Graded" by
**Wednesday, May 13 at 11:59pm**

## Instructions

Format your .RMD file using the template on the course website. **Submit the .RMD file, the knitted .html output, and any other files or folders needed as a single .zip file.**

The grader will be compiling your .RMD file and making sure it knits. Any libraries/packages needed should be near the top of the .RMD file, so the grader can make sure they're installed. Any other files needed to knit the .html should be in the zipped folder you turn in. **If your code does not knit and there is no immediate fix, the grader will grade your HTML for a [-10pts] penalty.**

Any time I ask you to demonstrate something, show something, generate something, etc., you must provide the code and/or text commentary that does so.

Finally, we will be giving [5pts] for code style and cleanliness. For any function you write, include a comment on the line above the function saying what the function expects as input and what it outputs. If you do this and the rest of your code is reasonably neat then this is an easy [5pts].

## 1 Programming Puzzles

(a) Revisit HW 3 Graded, question 2, part (d). You may re-use your code or my solution code from the previous parts of the problem. In this re-visitation, vary both $q \in \{0, 0.01, 0.02, \ldots, 0.99, 1\}$ and $n \in \{5, 20, 100, 500, 2000\}$, again with 1000 repetitions of the experiment. Output a $101 \times 5$ matrix `weks`. Explain how the empirical skewness varies as a function of $n$ as well as $q$. You may **not** use `for` loops or repeat copy-pasted code for this—you **must** use `apply()`-type functions. [3pts]

(b) Revisit HW 5 Graded, question 1, part (c). You may re-use your code or my solution code. Generate $10^4$ data points from the data-generating model for $X$ and $Z$. Then, for $a \in \{1, 5, 20, 100\}$, generate $10^4$ samples from $Y \sim$

$X + 2Z - aXZ + \epsilon$. Generate 99.9% confidence intervals for $\beta_X$ from both model 1 and model 2, and for the different values of $a$ comment on whether or not they contain the true $\beta_X$ and compare the widths of the confidence intervals. You may **not** use `for` loops or repeat copy-pasted code for this—you **must** use `apply()`-type functions. [4pts]

# 2 Extrapolation

(a) Suppose that the SAT data is drawn from two sources: an all-boys' private school, and an all-girls' private school (and that these schools reported all their students' sexes entirely as `male == 0` and `female == 0`, respectively, in the 90s, when the data was collected). Suppose that you are given only the all-boys' school data for training a model, and that you must make predictions about the students from the all-girls' school. Fit two linear models to the all-boys' school data, predicting college GPA from 1) HS GPA or 2) SAT score (use only a first-degree term, so both models have an intercept and one other coefficient). Does one model yield a better fit to the data than the other, and if so which one, or do they both fit the data roughly equally well? Does one model yield better prediction on the all-girls' school data than the other, and if so which one, or do they both predict about equally well? [4pts]

(b) Now suppose I tell you that the rows of this data frame have *already* been randomized, and ask you to pick training and test data sets by taking the first 500 rows and last 500 rows of the data, respectively (no need to use the `sample` function). Fit the two models from the previous part to the training data and predict on the test data. Is the prediction error generally higher or lower than in the previous part of the question? Explain in words why this should be so. [4pts]

# 3 Bae Area COVID-19 and Loss Functions

Just like last week, find all 27 days in March for which the number of SF COVID-19 cumulative confirmed cases was positive (feel free to copy your old code or my solution).

(a) Use the `lapply` function to fit linear models with polynomial terms from degree 1 to 5 to the March data, taking the natural log of the number of confirmed cases as the outcome. Which model has the best $df$-adjusted model fit? Which has the most predictive power for the first week of April in SF? Answer both of these questions on the original scale of the outcome, as well as on the natural log scale. Use the `poly` function in this question! [5pts]

(b) Statisticians employ *loss functions* frequently to determine the predictive quality of a model. Recall that if $X$ is a vector of covariates, then $\hat{Y} = \hat{Y}(X)$ is a function of $X$ and the estimated coefficients $\hat{\beta}$. The loss function $L(Y, \hat{Y})$ is always positive; for example, $L(Y, \hat{Y}) = |Y - \hat{Y}|$ is the absolute loss function. Write a function `squaredLoss(y,yhat)` that returns a loss equal to the squared residual. For the cubic (degree 3) model from the previous part, use `squaredLoss` to compute the Residual Standard Error (you can check your work with `sigma`). [2pts]

(c) Suppose that underestimating the number of confirmed cases by a quantity $q$ yields 4 times as much loss as overestimating by $q$. Which polynomial minimizes this type of loss in prediction on the first week of April SF data? [3pts]

(d) Describe in words how you would modify the `myLM` function from HW 4 Graded to minimize a given loss function `lossFunc` on the training data rather than the mean squared residual. [2pts]

(e) In the first part of this question, you fit a linear regression using the natural log of the number of cases. Does this penalize *under*estimating or *over*estimating the true number of confirmed cases more (i.e., does over- or under-estimating by a quantity $q$ yield larger loss)? Is this appropriate in the context of COVID-19? Answer in words. [3pts]