Final Exam: 2025

Part 1 - Multinomial Regression

The first part of this exam focuses on a dataset with car sales. The source dataset is the same as what we used in the midterm, but we will address a different research question.

```
subaru <- read_csv('https://raw.githubusercontent.com/STAT439/Exam/refs/heads/main/subaru.csv
  mutate(model = factor(model))
```

Our research question will be to model the probability that a sold vehicle is one of four different Subaru models: `Legacy`, `Forester`, `Impreza`, `Outback`. For this framework, we will use the following variables:

- odometer: mileage
- sellingprice: price of the sale
- transmission: type of transmission

1. Data Visualization (4 points)

Create a set of figures to visualize the research question stated above. Include a summary paragraph describing your findings.

2. (4 points) Statistical Model

Write out a statistical model for estimating probabilities of the four vehicle models as a function of `odometer`, `sellingprice`, and `transmission`. Be specific with your notation.

3 (4 points) Model Fit

Fit your model and print out the results.

4 (4 points) Model Summary

Talk about each of the three predictors and how they impact the probability that the car being sold in each of the four vehicle models.

5 (4 points) Probability Estimations

For the two scenarios estimate the probability that the car is each of the four models.

- odometer = 100,000, transmission = automatic, sellingprice = 10,000
- odometer = 10,000, transmission = automatic, sellingprice = 20,000

Part 2 - Random Forest Classification

Now we will use the same dataset in a classification setting to build a model for predicting the type of vehicle being sold. For this framework use the following variables

- year: year the vehicle was made
- odometer: mileage
- sellingprice: price of the sale
- transmission: type of transmission
- color: color of the vehicle
- interior: color of the vehicle interior

6 (4 points) Test & Training Set

Construct a test & training set `subaru` dataset. Include a written description of *why* and *how* you do this.

7 (4 points) Fit a Random Forest Model

Fit a random forest model on the training set and briefly describe how the random forest works.

8 (4 points) Classification Error

Use the random forest model to make prediction for the vehicle make on the test set. Use a zero-one loss function and report the classification error (% of incorrect predictions).

Part 3

Consider a synthetic dataset that has one predictor variable (`x`), a group identifier (`id`), and a count response.

```
exam_data <- read_csv('https://raw.githubusercontent.com/STAT439/Exam/refs/heads/main/exam_da
```

```
Rows: 200 Columns: 3
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (3): y, id, x

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(exam_data)
```

```
# A tibble: 6 x 3
      y    id      x
  <dbl> <dbl>  <dbl>
1     0     1 -0.779
2     6     1  0.263
3     0     1  0.532
4     1     1 -0.770
5     1     1 -0.537
6     2     1  0.443
```

9 (4 points) Data Visualization

Assume the goal is to model `y`. Create a data visualization that informs the relationship between `x` and `y` and includes `id`.

10 (4 points) Write out a model

Assume you decide to fit a generalized linear mixed model. Write out the notation for this model.

11 (4 points) Fit the model

Fit the generalized linear mixed model you've specified in Q10. Include a written interpretation of the parameters in the model.

12 (4 points) Visualize Model Fit

Create an updated visualization that includes the model fit from Q11.