Exam Review Notes (MGSC310)

-linear regression:

-continuous predictor

Increase one unit of x makes the y variable add/subtract by the coefficient number

-factor or binary predictor

Always set one factor to the baseline, and compare the coefficients to the baseline

Ex) euro cars compared to us cars have 2.42 mpg more than the american cars

Logistic regression

Probability of y/ probability of not y

Must exponentiate the coefficients to get the odds ratio

1 - 50% of something happening, greater than 1 means likelihood is greater than 50%

Less than 1 means less than 50% chance of something happening

Works for continuous variable or for classification

*don't train with the test set; do new data = test_data in predict function to make a prediction for the test data

Logistic regression (con)

Glm function (family = binomial) ; lm is linear regression

Family = binomial -> classification model, then the outcome variable is binary

Group summary statistics: group_by() and summarize() - col_avg, col_max, col_cnt

Df %>% group_by(group_var) %>% summarize(col_avg = mean(col, na.rm = TRUE),

col_max = max(col, na.rm = T), col_cnt = n())

-be familiar with ggplot, any graphing commands, etc.

Loocv vs kfold

-loocv is more computationally expensive

-loocv provides test mse estimates with lower variance

-kfold cross validation might have more mse bc it has higher variance

In fold cv, estimate test data mse by building models using k samples of training data

Goal of ridge regression is to reduce model variance

In ridge regression you put a penalty on the slope coefficients to reduce magnitude

In ridge regression if lambda is 0 (penalty rate), its the same as OLS (least squares)

As you add additional predictor variables to a lm model, the mse will go down bc the complexity of the model increases; sum of squared residuals will also decrease bc that and the mse are very similar

Lambda min is the one that minimizes error
Lambda use is the one that minimizes mse plus 1 standard error

Test to have higher variance than train, cross validation is in the middle