STAT 22400 Takeaway

Assumptions of MLR

1. The model
   - Have a linear relationship: $\mathbb{E}[X \mid Y] = \beta x$
2. The predictors
   - Independent from each other
   - They are nonrandom fixed values
3. The errors
   - Independent of time
   - With a constant variance, mean = 0

Violations of MLR assumptions & Detection

1. Non-linear relationship: non-linearity
   - Plot Y against X directly
   - Plot residual again fitted value -> there's a pattern
2. Predictors are linearly dependent: multicollinearity
   - Variance inflated factor, $VIF_j = \frac{1}{1-R_j^2} \geq 10$
     - Explanation: R_j^2 is the **coefficient of determination** from regressing the predictor X_j on all the **other predictors** in the model. When R_j is around 0.95, which means 95% of the variance in R_j could be explained by other predictors. Therefore VIF is large, therefore there's multicollinearity.
3. Errors dependent on time: autocorrelation
   - Residual time plot (surprisingly smooth or rough)

- Durbin-Watson Test for AR(1)
- ACF plot (shows the autocorrelation at different lags)
- Lag plot

4. Errors have a heteroscedasticity issue
    - Residual plot, residual plot, residual plots…

What are the remedies

1. Non-linearity:
    - Variable transformation: either predictor or response
2. Multicollinearity:
    - Remove predictor(s)
    - Ridge, Lasso
3. Autocorrelation:
    - Oscutt Method (iteration) -> remove AR(1)
4. Heteroscedasticity:
    - Response variable transformation
        - Box-Cox
    - WLS: if the errors demonstrate a significant pattern

Interview questions so far...

1. What's the basic assumption for OLS?
2. Have you heard about regularization/Lasso?
3. The biggest challenge for developing a model is solving non-linearity and model selection, so I guess lots of things will expand on that…

Concepts should know crystal clear

1. What is an influential observation? What's the commonality/difference between influential points, high leverage points, and outliers?
    - An influential observation has large effect on some part of the model (measure Cook's distance, DEFITs) (measure the regression coefficient no matter it's slope or intercept)
    - An outlier is a point that this model fails to explain (measure residuals)

    - Little influence on the slope unless it's high leverage
    - May or may not be influential
  - An high leverage point is far from the mean of predictors (only consider X) (measure leverage, AKA hat value)
    - It pulls the model towards itself; it might no be an outlier
    - May or may not be influential

2. What is interaction?
   - When two or more **categorical** variables are combined together
3. How do you penalize the model for extra variables?
   - Score: AIC, BIC, smaller is preferred
   - Adjusted R square: larger is preferred