CAP 5302, Questions for Q & A ("Oral Exam")

1. Power Analysis, Sample Size Calculations

1. What is meant by **statistical power of a study**? E.g. "this study has a power of 0.80 when true value is p_0"?
2. Give examples where **high statistical power** is more important than **low Type I error**, and make sure to explain why there's always a trade-off between the two.
3. What ways do we have to increase statistical power of a study (name at least three)?

2. Association Between Categorical Variables.

1. Describe what's meant by independence (or, vice versa, dependence) of two categorical variables (potentially using a contingency table example as reference).
2. Name the statistical test that allows to test for independence of two categorical variables. Explain its steps, intuition.
3. Comment on practical vs statistical significance of independence between two categorical variables. How can we measure practical effect size?

3. Simple Linear Regression.

1. Explain the goal of simple linear regression (visually). What are we trying to minimize?
2. State the general formula of simple linear regression. Be able to interpret the intercept and the slope in linear regression.
3. What is the goal of statistical inference? **Hint**: What is the difference between $\hat{\beta}_1$ and β_1?
4. Under what modeling assumptions are we allowed to conduct inference? Be able to formulate these assumptions for both the **error terms** ϵ_i, and the **response values** y_i.
5. Interpret the following statement: "$\hat{\beta}_1$ is an unbiased estimate of β_1".
6. What is meant by standard error of the estimate (e.g. $SE(\hat{\beta}_1)$)?
7. "Sampling distribution of $\hat{\beta}_1$ is normal" - explain what that statement means.
8. When conducting inference on β_1, what distribution is used? Why not normal?
9. Interpret the following statement: "95% confidence interval for β_1 (in $Y = \beta_0 + \beta_1 X$ regression) is $(0.2, 0.4)$". E.g. what does it mean
   - to have "95% confidence"?
   - for the relationship between X and Y?
10. Interpret: "For TV budget value of $10K\$$, the predicted value of sales is $\hat{Y} = 300$."
11. When making predictions via linear regression, explain the difference between 95% confidence bands and 95% prediction bands. Which ones are wider? Why?
12. Walk through the steps of hypothesis test for slope parameter in simple linear regression (from formulating hypotheses down to the decision about H_0). Define the concept of p-value.
13. Name two quality-of-fit metrics for linear models. Explain what each of them is trying to capture, what values are indicative of a good fit.

4. Multiple Linear Regression.

1. State the general formula of multiple linear regression. Be able to interpret the intercept and slope coefficients. What is the most critical difference in slope interpretations between simple and multiple linear regression?
2. If the result of fitting simple linear regression is a line, then what's the result of fitting a two-predictor multiple regression? Three-predictor or more?
3. In case of two-predictor multiple regression, when would we fail to have a uniquely defined least squares solution $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$? Be able to interpret intuition behind this fact.
4. Explain why is that possible for a predictor to
   - be significant in simple linear regression (e.g. *newspaper* in $Sales \sim newspaper$), but
   - become insignificant in multiple linear regression (e.g. *newspaper* in $Sales \sim newspaper + radio$).
5. When conducting inference on β_1 (for **multiple** linear regression), what distribution is used?
6. Be able to interpret confidence intervals for slopes in multiple linear regression.
7. Given a summary table of fitted multiple linear regression model, be able to explain all the entries.
8. Formulate the hypotheses and explain the main idea/steps behind the procedure to test the overall model significance in multiple linear regression. What values of the test statistic indicate good model?
9. Why refer to procedure in previous q. to test the model significance, when one could simply see if at least one individual predictor t-test was significant?
10. Explain what is meant by collinearity in the case of two predictors, what issues it causes, and how it is remedied.
11. For the case of general number p of predictors, explain what is meant by multi-collinearity, how it can be diagnosed and remedied.
12. Describe the process of backward variable selection via AIC. Explain the main idea behind AIC.
13. Whenever we drop variables from the model, what happens to model's SSE (AKA "RSS")? R^2?
14. How does one incorporate categorical predictors with K categories?
15. Be able to interpret effect of categorical predictors on the response.
16. How do we test for significance of a categorical predictor with $K > 2$ categories? Name the test, and outline the main idea/steps behind testing procedure.

5. Multiple Linear Regression: Extensions and Diagnostics.

1. What is meant by "additivity" assumption of a linear regression model? Give an example. How can we relax that assumption?
2. Explain the main idea behind interaction effect for
   - two quantitative predictors (e.g. factory lines and workers, their effect on production)
   - one quantitative and one categorical predictor (Explain visually, use an example).
3. How does one test for interaction in cases when it is represented by
   - single term
   - multiple terms
4. In case there's a significant interaction between two predictors, should one test/interpret their **main effects** (also, explain what's meant by "main effects")? Why?
5. Is interaction and association between two predictors the same thing? Why?
6. How to model a non-linear relationship between response and a predictor? How to test significance of that potential non-linear relationship?
7. Why shouldn't we use polynomials of very high degrees (e.g. 5 and beyond)?
8. What does the **principle of marginality** state? Give examples.
9. Why are the residual plots important, and what assumptions do they allow us to diagnose?
10. What issues does the residuals-vs-fitted plot allow us to diagnose? How could we remedy those issues?
11. When stabilizing the variance, why is the log-transformation preferred to $\sqrt{}$?
12. How could one apply log-transform in case response takes on negative values?
13. What issue does the QQ-norm plot allow us to diagnose? When does that issue actually present a big deal? How can we address that issue?
14. Explain what is meant by "regression outlier". Why do we need to be wary of regression outliers? What measure allows us to detect them?
15. Explain what is meant by "leverage" of an observation. Why do we need to be wary of leverage? What metric is used to measure it?
16. What constitutes an influential observation? What metric is used to measure influence of an observation? What should we do with such observations?

6. Logistic Regression.

1. For binary response, what distribution is appropriate?
2. Let $Y = 0/1$ - our binary response, X - our predictor, $p(X) = P(Y = 1 \mid X)$. We need to represent probability $p(X)$ as a function of X. What is the issue with the following:

$$p(X) = \beta_0 + \beta_1 X \quad ? \tag{1}$$

   How do we fix this issue?
3. How do we define the *odds* of $Y = 1$ given X? How do we use the odds to obtain a proper logistic regression *equation* (instead of (1))?
4. Name three components to a generalized linear model. Show them on the example of logistic regression.
5. Be able to provide/explain full GLM formulation of a logistic regression model.
6. Name the method used to estimate parameters of logistic regression. What's the main idea behind it (AKA why is it called that way)?
7. Be able to provide interpretations of predictor effects on:
   - probability of $P(Y = 1 \mid X)$,
   - log-odds of $P(Y = 1 \mid X)$
   - odds of $P(Y = 1 \mid X)$.
8. Having estimated the logistic regression coefficients $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$, how do we predict

$$P(Y = 1 \mid \mathbf{X}_{new} = (X_{1,new}, \ldots, X_{p,new}))$$

   for a new observation $\mathbf{X}_{new}$?
9. Walk through the steps of hypothesis test for slope parameter in logistic regression (from formulating hypotheses down to the decision about H_0). What distribution is used for test statistic?
10. Be able to interpret confidence intervals for slopes in multiple linear regression.
11. For **logistic regression**, explain the main idea/steps behind the procedure to test the overall model significance. What's the sampling distribution of test statistic? What values of the test statistic indicate good model?

7. Multinomial Logit.

1. Explain the difference between multinomial logit model and classic binary logistic regression. Provide details on how it (multinomial logit) operates.
2. Be able to interpret effects of predictors in fitted multinomial logit model.
3. What distribution/approximation is used to obtain p-values and confidence intervals for effects of individual predictors in multinomial logit model?
4. In multinomial logit model, what does it mean to test a predictor "as a whole"? Be able to roughly outline steps of such a test.
5. How do predictions look like for multinomial logit models? E.g. what kind of values are expected from *predict*() function applied to a multinomial logit fit.