Midterm Review - Part 2

D. Pluta

2024-01-11

Midterm Information

- Held in class on March 5th. Exam will last entire class session.
- Bring pens or pencils. No calculator needed.
- One *handwritten* page of notes (8in x 10in, front and back) is allowed. Note sheets may include formulas, worked examples and derivations, and summaries of lecture notes.
- Questions will be similar in style and content to the homework and problems given below.
- No R programming or writing of pseudocode will be required.
- No phones.

STAT 7630, SPRING 2024

1. Normal Distribution

    a. For a sample $y_1, \ldots, y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, write out the likelihood for (μ, σ^2).
    b. What is the corresponding log-likelihood?
    c. Suppose now that the sample is drawn such that $y_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma^2)$, that is, each observation has its own mean μ_i. Write the log-likelihood for this sample.
    d. In the scenario of (c), find the MLEs of the μ_i, or explain why they cannot be calculated.
    e. How do the above settings differ from those assumed in ANOVA and linear regression?

2. One-way ANOVA

    The balanced one-way ANOVA model can be stated as

    $$Y_{ij} = \mu_j + \varepsilon_{ij}, \varepsilon + ij \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n; j = 1, \ldots, m.$$

    a. In terms of the model parameters, write out the formal null hypothesis for the ANOVA F-test.

    b. What is the distribution of Y_{ij}?

    c. What is the MLE for μ_{ij} in terms of the sample data? Write the mathematical expression.

    d. Find the MLE of σ^2.

    e. What is the unbiased estimator of σ^2? (Consider the degrees of freedom of the model.)

    f. If we run a two sample t-test that indicates $\mu_1 \neq \mu_2$, do we then know that the ANOVA F-test will also reject the null? Does this change in the *unbalanced* situation, i.e., when there are potentially unequal sample sizes across groups?

    g. Suppose we are analyzing data from an experiment on a new drug, with 3 dosage groups labeled A, B, and C. Group A receives a dose of 10mg, group B receives a dose of 30mg, and group C receives a placebo. The researcher is primarily interested in determining if $\mu_A > \mu_B$. From a statistical standpoint, is the data from group C of any use for the test of interest? From a scientific standpoint, why is it important to collect data from group C?

    h. For the setting of part (g) again, what are the potential costs and benefits of collecting data from an additional group (labeled D) with a dosage of 50 mg. Will this data be useful in testing the hypothesis $\mu_A > \mu_B$?

i. You conduct three two-sample t-tests to compare the means of each of groups A, B, C. The P-values you calculated are $PA > B = 0.043, P_{A>C} = 0.026, P_{B>C} = 0.037$. The lead researcher is happy to see significant results at the 0.05 level. What is your interpretation of these P-values and the significance of the tests?

3. Linear Regression Concepts

a. Four assumptions of linear regression.
b. What are the potential consequences if the equality of variance assumption is violated?
ii. What plot(s) can be used to assess the equality of variance assumption?
iii. What plot or plots can be used to assess the assumption of normally distributed errors?
iv. A study of environmental effects on blood sugar collects data from a cohort of 50 married couples (male/female), for a total of $n = 100$ subjects. The response variable is the blood sugar level at the time of visit. Covariates include Age, Sex, Zip Code. Which assumption(s) of linear regression is this data set likely to violate?
b. In the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, what are the distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$?
c. Write the expression for the 95% confidence interval of $\hat{\beta}_1$.

4. Multiple Linear Regression Concepts

$$Y = X\beta + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

a. Write out the likelihood in matrix form.
b. Write out the log-likelihood in matrix form.
c. The MLE for the coefficient vector β is $\hat{\beta} = (X^T X)^{-1} X^T Y$. Calculate the variance of $\hat{\beta}$.
d. Show $\hat{\beta}$ is unbiased for β.
e. Explain, using properties of the normal distribution, why $\hat{\beta}$ is normally distributed.
f. Linear regression, as we've formulated so far, requires that our sample size n be larger than the number of covariates p. What problem do we encounter if $p > n$? (Hint: Refer to the formula for $\hat{\beta}$.)
g. Suppose we have a set of observations Y_1 with covariance $\sigma^2 I_n$, and an independent set of observations Y_2 with variance $2\sigma^2 * I_n$. Write the covariance matrix for the concatenated response vector $Y = (Y_1, Y_2)$.
h. Assuming it is known that the variance of Y_2 is twice that of Y_1, suggest a data transformation that will remedy the violation of the constant variance assumption.