UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
Qualifying Exam: Advanced Statistics I
Friday, January 21, 2022

1. Suppose that $(X_1, X_2, X_3, X_4)$ follow a multivariate normal distribution with mean $(1, -1, 0, 1)$ and covariance

$$\begin{pmatrix} 1 & 0.2 & 0.5 & 0 \\ 0.2 & 1 & 0.1 & 0 \\ 0.5 & 0.1 & 1 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix}.$$

   (a) What is the joint distribution of $(X_2 - 0.1X_3 + 1, X_3)$?

   (b) Find the distribution of

$$\frac{\left[(X_2 - 0.1X_3 + 1)^2 / 0.99 + X_3^2\right] / 2}{(X_4 - 1)^2/4}.$$

   (Hint: What are the distributions of the numerator and denominator? Use part (a) for the numerator. You may use known facts and do not necessarily need to derive a pdf for this problem.)

   (c) Find the distribution of

$$\frac{X_4 - 1}{2|X_1 - 1|}.$$

   (Hint: What are the distributions of the numerator and denominator? Again, you may use known facts and do not necessarily need to derive a pdf for this problem.) What is $E\left[\frac{X_4-1}{2|X_1-1|}\right]$?

   (d) What is the conditional distribution of $X_1$ given $X_2 = x_2$ and $X_4 = x_4$ (in terms of $x_2$ and $x_4$)?

2. Suppose that for each $i \in \{1, \ldots, n\}$, $Y_i = \beta_0 + \beta_1 U_i + \varepsilon_i$, where $\varepsilon_1, \ldots, \varepsilon_n$ are IID mean-zero with variance $\sigma_\varepsilon^2$, and $U_1, \ldots, U_n$ are IID with variance $\sigma_U^2$. We wish to estimate the regression coefficient $\beta_1$. However, the $U_i$ are unobserved; instead, we observe the noisy observation $X_i = U_i + \eta_i$, where $\eta_1, \ldots, \eta_n$ are IID mean-zero with variance $\sigma_\eta^2$. Also assume that for each $i$, $U_i$ and $\eta_i$ are uncorrelated $U_i$ and $\varepsilon_i$ are uncorrelated, and $\eta_i$ and $\varepsilon_i$ are uncorrelated. This is a simple *errors-in-variables* model.

   (a) What is $\mathrm{Var}(X_1)$ (in terms of $\beta_0$, $\beta_1$, $\sigma_\varepsilon, \sigma_U$, and/or $\sigma_\eta$).

   (b) What is $\mathrm{Cov}(X_1, Y_1)$ (in terms of $\beta_0$, $\beta_1$, $\sigma_\varepsilon, \sigma_U$, and/or $\sigma_\eta$)?

   (c) Suppose we fit a simple linear regression of $Y_1, \ldots, Y_n$ on $X_1, \ldots, X_n$. Denote the estimated regression coefficient from this fit $\hat{\beta}_1$. What does $\hat{\beta}_1$ converge to in probability? Is $\hat{\beta}_1$ consistent for $\beta_1$? If not, what can be said about the limiting value of $\hat{\beta}_1$ in relation to the truth $\beta_1$? (Hint: use consistency of empirical covariances/variances to population covariances/variances and parts (a) and (b).)

3. Observed data are $\{x_{i1}, x_{i2}, x_{i3}, x_{i4}, y_i\}_{i=1}^n$ and suppose that two potential models are

$$\text{model 1: } y_i = \beta_0 + \sum_{j=1}^{4} \beta_j x_{ij} + \varepsilon_i \text{ with } \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2) \text{ and}$$

$$\text{model 2: } y_i = \gamma_0 + \sum_{j=1}^{2} \gamma_j x_{ij} + \tilde{\varepsilon}_i \text{ with } \tilde{\varepsilon}_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \tilde{\sigma}^2).$$

(a) Let $\boldsymbol{\beta}^\top = (\beta_0, \ldots, \beta_4)$. Define $\mathbf{X}$ and state conditions on it so that the MLE of $\boldsymbol{\beta}$ is $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ and show that it is unbiased for $\boldsymbol{\beta}$. (For the rest of the problem, you may assume your conditions on $\mathbf{X}$.)

(b) Find a matrix $\mathbf{H}$ so that $\mathbf{H}\mathbf{y} = \begin{pmatrix} \hat\beta_0 + x_{11}\hat\beta_1 + \ldots + x_{14}\hat\beta_4 \\ \vdots \\ \hat\beta_0 + x_{n1}\hat\beta_1 + \ldots + x_{n4}\hat\beta_4 \end{pmatrix}$ and show that $\mathbf{H}$ is idempotent.

(c) Consider $\hat\sigma^2(d) = \mathbf{y}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{y}/d = \text{SSE}(\mathbf{X})/d$. Find a $d$ so that $\hat\sigma^2(d)$ is an unbiased estimator for $\sigma^2$ (and show that it is unbiased). (We call that estimator MSE.)

(d) Let $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{H}}$ be appropriate matrices for model 2. Show that $\mathbf{H}\tilde{\mathbf{H}} = \tilde{\mathbf{H}}\mathbf{H} = \tilde{\mathbf{H}}$.

(e) Use the results from the previous parts to show that $\text{SSE}(\tilde{X}) - \text{SSE(X)} \geq 0$. (Hint: you may use that idempotent matrices are semipositive definite.)

(f) Use the ANOVA equation to explain (in words) why the result from part (e) makes intuitive sense.

(g) Suppose you fit models 1 and 2 and get MSEs (error variance estimates) for both models. Can you say that one of those MSEs is necessarily larger than the other? Why or why not. Does that makes sense? Why or why not.