Statistics 600 Final Exam

December 20, 2021

There are 5 questions, each with multiple parts. Each of the 5 questions is worth 20 points. Try to complete all five questions. Partial credit will be given, show your work where appropriate.

1. Suppose we have data for a sample of subjects representing a population, where y_i is a person's resting heart rate, a_i is their age in years, $f_i = 1$ if a person is female, and $f_i = 0$ if a person is male.

   (a) Suppose we fit a model of the form $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 a + \hat{\beta}_2 f$. State two distinct aspects of this model that may limit its ability to describe a real human population.

   Solution: In this model, the sex-specific conditional mean functions $E[y|a, f = 0]$ and $E[y|a, f = 1]$ are both linear functions of age, and furthermore are parallel. The true conditional mean functions may be non-linear, and/or may not be parallel. There are a number of other possible answers such as the model having a homoscedastic variance structure and the model not accounting for any possibility of correlations between people.

   We did not deduct for this but it is not very informative to simply state that the model does not include an interaction. It is more informative to directly express what is deficient about the model as given.

   We also accepted without deduction that the model may exclude some covariates. However this is true of any model, and in many settings the goal of a regression analysis is to estimate the conditional mean for a given set of explanatory variables, not for all possible explanatory variables. Arguably, if all possible variables are included then the system becomes deterministic and statistical approaches are no longer needed. In other words, the conditional mean $E[y|a, f]$ is a well-defined estimand in spite of the fact that other covariates may exist.

   (b) Suppose we obtain the fitted model $\hat{y} = 60 + 0.1 \cdot a + 5 \cdot f - 5 \cdot a \cdot f/70$. Describe in 2-3 sentences the relationship according to this model

between female and male heart rates over a typical adult human age range of say 18 to 80.

Solution: Under this model, the expected heart rates for both females and for males are increasing linear functions of age. The sex-specific conditional mean lines intersect at age 70, so that below age 70 women have higher heart rate and above age 70 men have higher heart rate.

(c) Suppose we construct a plot of residuals on fitted values for a model fit to this data (not specifically either of the models discussed in parts a or b). Describe in 1-2 sentences what we might hope to learn by doing this.

Solution: The main use of a plot of residuals on fitted values is to assess whether the conditional variance is approximately constant (i.e. the residual variation is homoscedastic).

(d) Suppose we construct a partial residual plot for age using a model fit to this data (again, not specifically one of the models presented above). Describe in 1-2 sentences how this plot is constructed and what might be a purpose of doing this.

Solution: A partial residual plot is a scatterplot of $\hat{\beta}_j x_{ij} + r_i$ for a specific covariate j, where i indexes the observations. The main use for a partial residual plot is to visualize a dataset that would arise if all covariates other than covariate j were held at fixed values rather than varying.

2. Suppose we have a regression model with a mean structure $E[y|x] = x+x^3$ and $\text{var}[y|x] = \sigma^2$, where furthermore x follows a standard normal distribution. Note that a standard normal random variable z has the property that Ez^p is equal to 0 if p is odd, and if p is even then Ez^p is the product of all odd integers between 1 and $p-1$.

(a) Suppose we fit the simple linear regression model $\hat{y} = \hat{\alpha}+\hat{\beta}x$ using ordinary least squares (OLS). What are the limiting values of $\hat{\alpha}$ and $\hat{\beta}$?

Solution:

$$\begin{aligned}\hat{\beta} &= \sum y_i(x_i - \bar{x}) / \sum (x_i - \bar{x})^2 \\ &= \sum (x_i + x_i^3 + \epsilon_i)(x_i - \bar{x}) / \sum (x_i - \bar{x})^2 \\ &\to E[x^2] + E[x^4] \\ &= 4.\end{aligned}$$

Since $E[y] = E[x] = 0$, it follows that $\hat{\alpha} \to 0$.

(b) What is the limiting R^2 if we use the simple linear regression model from part (a)?

Solution:

$$\text{var}(\hat{y}) = 16\text{var}(x) = 16.$$

$$\begin{aligned}\text{var}(y) &= \text{var}(x) + \text{var}(x^3) + 2\text{cov}(x, x^3) + \text{var}(\epsilon) \\ &= 1 + 15 + 2 \cdot 3 + \sigma^2 \\ &= 22 + \sigma^2.\end{aligned}$$

Thus the limiting R^2 is $16/(22 + \sigma^2)$.

3. Suppose we have data on n matched pairs of people. Let $y_{ij} \in \mathcal{R}$ for $i = 1, \ldots, n$ and $j = 1, 2$ denote the responses, and suppose that $E[y_{ij}|x_{ij}] = \beta' x_{ij}$, for covariates $(1, x_{ij}) \in \mathcal{R}^2$. Further, suppose that the y_{ij} are independent within and between pairs, and that the variance structure is $\text{var}(y_{i1}|x_{i1}) = \sigma^2$, $\text{var}(y_{i2}|x_{i2}) = f\sigma^2$ for some $f > 0$. For simplicity, take $\sum_i x_{i1} = \sum_i x_{i2} = 0$, and $\sum_i x_{i1}^2 = \sum_i x_{i2}^2 = n$.

(a) Suppose that σ^2 amd f are known and we use GLS to estimate β. Derive a concise expression for $\text{cov}(\hat{\beta})$.

Solution:

$$\text{cov}(\hat{\beta}_{\text{gls}}) = \frac{\sigma^2 f}{n(1+f)} I_{2\times 2}.$$

(b) What is the ratio of $\text{cov}(\hat{\beta}_1)$ (where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$) for a given value f to the value obtained when $f = 1$?

Solution: $2f/(1+f)$

(c) Suppose in one study we have a sample size n and $f = 1$, and in a second study with the same σ^2 we have a sample size n' and a given value f. If the two studies have the same standard error for $\text{var}(\hat{\beta}_1)$, what is the ratio n'/n?

Solution: $2f/(1+f)$

(d) Suppose we estimate the model parameters using OLS, ignoring the possible heteroscedasticity. What is the ratio $\text{var}(\hat{\beta}_1^{\text{ols}})/\text{var}(\hat{\beta}_1^{\text{gls}})$?

Solution: Since

$$\hat{\beta}_{\text{ols}} = (\sum_i X_i' X_i)^{-1} \sum X_i' y_i,$$

therefore

$$\text{cov}(\hat{\beta}_{\text{ols}}) = (\sum_i X_i' X_i)^{-1} \cdot \sum X_i' \text{cov}(y_i) X_i \cdot (\sum_i X_i' X_i)^{-1}.$$

Based on what is given above,

$$\sum_i X_i' X_i = 2nI_2.$$

Therefore

$$\text{cov}(\hat{\beta}_{\text{ols}}) = (4n^2)^{-1} \sum X_i' \text{cov}(y_i) X_i.$$

Since

$$\sum X_i' \text{cov}(y_i) X_i = n\sigma^2(1+f) I_2.$$

Therefore $\text{cov}(\hat{\beta}_{\text{ols}}) = (4n^2)^{-1} n\sigma^2(1+f) = \sigma^2(1+f)/(4n)$.

Thus

$$\text{var}(\hat{\beta}_{\text{ols},1})/\text{var}(\hat{\beta}_{\text{gls},1}) = (1+f)^2/(4f).$$

4. Suppose we are interested in the relationship between the means EX and EY, for a population defined by two random variables X and Y. We have IID samples of data $x_1, \ldots, x_n \sim F_X$ and $y_1, \ldots, y_n \sim F_Y$.

   (a) Taking X and Y to be independent, and $\sigma_x \equiv \text{SD}(X)$ and $\sigma_y \equiv \text{SD}(Y)$ to be known, provide a 95% confidence interval for $EX - EY$.

   Solution: By the central limit theorem, $\bar{x}$ and $\bar{y}$ are approximately Gaussian. The standard error of $\bar{x} - \bar{y}$ is $\sqrt{\sigma_x^2/n + \sigma_y^2/n}$, so an approximate 95% confidence interval is

   $$\bar{x} - \bar{y} \pm 2\sqrt{\sigma_x^2/n + \sigma_y^2/n}.$$

   (b) Now suppose that the data may be correlated, with $\text{cor}(X, Y) = r$ (the x_i remain mutually independent of each other, as do the y_i). Taking σ_x, σ_y, and r to be known, provide a 95% confidence interval for $EX - EY$ in this setting.

   Solution: In this case we can estimate $EX - EY$ as $\bar{d}$, where $d_i = x_i - y_i$. The standard error of $\bar{d}$ is $\sqrt{\sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y}/\sqrt{n}$, so the interval has the form

   $$\bar{x} - \bar{y} \pm 2\sqrt{\sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y}/\sqrt{n}.$$

   (c) Under what conditions can we be certain that the interval from part (b) will be narrower than the interval from part (a)?

   Solution: The second interval will be narrower than the first interval if and only if $r > 0$.

   (d) Now suppose that our data are non-negative and follow a quasi-Poisson like distribution, meaning that $\text{var}(X) = \phi \cdot EX$ and $\text{var}(Y) = \phi \cdot EY$, for the same value of ϕ, and we wish to obtain

a 95% confidence interval for the ratio EY/EX in the presence of possible correlation between X and Y. Describe how this can be achieved using methods covered in this course.

Solution: We can use quasi-Poisson GEE. Place the data into a vector of length $2n$: $(x_1, y_1, x_2, y_2, \ldots)$, and let the design matrix be an $n \times 2$ array with first column identically 1 and second column $(0, 1, 0, 1, \ldots)$. Under this model, $E[X] = \exp(\beta_0)$ and $E[Y] = \exp(\beta_0 + \beta_1)$. Thus $E[Y]/E[X] = \exp(\beta_1)$. We can group the data using group labels $(1, 1, 2, 2, \ldots)$ and use GEE to estimate the model parameters. Let s denote the GEE standard error for β_1, which accommodates correlation between X and Y. Then a 95% confidence interval for $E[Y]/E[X]$ is $\exp(\hat{\beta}_1 - 2s), \exp(\hat{\beta}_1 + 2s)$.

5.

(a) Suppose we observe data $y_i \in \mathcal{R}^n$, $x_i \in \mathcal{R}^{n \times p}$ for $i = 1, \ldots, m$, where $E[y_i|x_i] = x_i\beta$ and $\mathrm{cov}[y_i|x_i] = \Sigma_i \in \mathcal{R}^{n \times n}$. The Σ_i are known, and $y_i, y_{i'}$ are uncorrelated when $i \neq i'$. Let $\Sigma_i = U_i D U_i'$ be the spectral decomposition of Σ_i, so that $U_i \in \mathcal{R}^{n \times n}$ is orthogonal and $D \in \mathcal{R}^{n \times n}$ is diagonal. Note that D is common to all groups. What is the relationship between $\hat{\beta}_{\mathrm{ols}}$ and $\hat{\beta}_{\mathrm{gls}}$ when $x_i = c_i U_i$ for all i, with $c_i \in \mathcal{R}$?

Solution:

$$\sum x_i' \Sigma_i^{-1} x_i = \sum c_i^2 \cdot D,$$

and

$$\sum x_i' \Sigma^{-1} = D \cdot \sum c_i U_i' y_i$$

Thus

$$\hat{\beta}_{\mathrm{gls}} = (\sum c_i^2)^{-1} \sum c_i U_i' y_i.$$

Also,

$$\sum x_i' x_i = \sum c_i^2 I_p$$

$$\sum x_i' y_i = \sum c_i U_i' y_i.$$

So we see that $\hat{\beta}_{\text{ols}} = \hat{\beta}_{\text{gls}}$.

(b) Suppose that $M = I + b \cdot 1_d 1_d'$, where I is the $d \times d$ identity matrix, $1_d \in \mathcal{R}^d$, and $b > 0$ is a scalar. Show that every eigenvector v of M satisfies either $v'1 = 0$ or $v \propto 1_d$.

Solution: Suppose that v, λ are an eigenvector and eigenvalue of M. Since $Mv = \lambda v$

$$(I + b \cdot 1_d 1_d')v = v + c \cdot 1_d,$$

where $c = b \cdot (1_d' v) \in \mathcal{R}$.

Thus we have a linear equation relating v and 1_d. This implies that either $v \propto 1$ or the linear equation is degenerate, and since $\lambda \neq 0$ by definition, the latter implies that $v' 1_d = 0$.

(c) Suppose that $n = 2$ and

$$\Sigma_i = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}.$$

What are U_i and D in this case, where $\Sigma_i = U_i D U_i'$ as above?

Solution: The columns of U are the eigenvectors of Σ, which has the compound symmetry structure from part (c). Therefore the eigenvectors are $(1, 1)/\sqrt{2}$ and $(1, -1)/\sqrt{2}$. We have

$$U = 2^{-1/2} \cdot \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

The corresponding eigenvalues are $1 + r$, $1 - r$ so

$$D = \begin{pmatrix} 1+r & 0 \\ 0 & 1-r \end{pmatrix}.$$