Statistics 600 Midterm Exam

November 3rd, 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.

(a) Is the product of square orthogonal matrices always orthogonal? Prove the statement or provide a counterexample.

Solution:

Suppose A and B are square and orthogonal, so that $A'A = I$ and $B'B = I$. Therefore $(AB)'AB = B'A'AB = I$, so the statement is true.

(b) Is the product of projection matrices always a projection matrix? Prove the statement or provide a counterexample.

Solution:

The statement is false. Suppose we have rank 1 projections $P = vv'$ and $Q = uu'$, where $\|v\| = \|u\| = 1$. Thus

$$PQ = (v'u) \cdot vu'$$

which is not symmetric or idempotent unless $v = u$. Since projection matrices are always symmetric and idempotent, this is a counterexample.

(c) Suppose we have a $n \times p$ matrix X, each column of which sums to 0. Let $\tilde{X}$ be an $n \times p$ matrix obtained by adding a fixed vector $v \in \mathcal{R}^p$ to each row of X. Derive a concise expression for the difference of Gram matrices $\tilde{X}'\tilde{X} - X'X$.

Solution:

Let $x_i \in \mathcal{R}^p$, $i = 1, \ldots, n$, denote the rows of X. Then

$$\begin{aligned}\tilde{X}'\tilde{X} &= \sum_i (x_i + v)'(x_i + v) \\ &= \sum_i x_i'x_i + \sum_i x_i'v + \sum_i v'x_i + nv'v \\ &= X'X + nv'v.\end{aligned}$$

Thus $\tilde{X}'\tilde{X} - X'X = nv'v$.

(d) Suppose that P is a projection matrix on $\mathcal{R}^p$. Prove that for any $x \in \mathcal{R}^p$, $\|Px\| \leq \|x\|$.

Solution:

We can write $x = Px + (I - P)x$, and since $P(I - P) \equiv 0$, it follows that Px and $(I - P)x$ are orthogonal, hence $\|x\|^2 = \|Px\|^2 + \|(I - P)x\|^2$, and it follows that $\|Px\|^2 \leq \|x\|^2$.

(e) Suppose $V \in \mathcal{R}^{n\times p}$ is orthogonal. For what values $\lambda \in \mathcal{R}$ is the matrix $I_{n\times n} + \lambda VV'$ positive definite?

Solution:

For a unit vector $x \in \mathcal{R}^p$

$$\begin{aligned}x'(I + \lambda VV')x &= 1 + \lambda\|V'x\|^2 \\ &= 1 + \lambda\|VV'x\|^2\end{aligned}$$

Since VV' is a projection matrix, using part (e) we know that $\|VV'x\| \leq \|x\| = 1$. Thus $\lambda > -1$ is sufficient. If $x \in \text{col}(V)$, $\|x\| = \|Vx\|$, so the condition is also necessary.

2. Suppose we are conducting an experiment to understand the relationship between a response y, and two predictor variables x_1 and x_2 which satisfy $\bar{x}_1 = \bar{x}_2 = 0$ and $x_1'x_1/n = x_2'x_2/n = 1$. Our goal is to precisely estimate β_1 in the mean structure $E[y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Since this is an experiment, we determine the values of x_1 and x_2, but are only given two options for doing so. One option is to have $\widehat{\text{cor}}(x_1, x_2) = 0.4$

at sample size $n = 30$. Another option is to have $\widehat{\text{cor}}(x_1, x_2) = 0.2$ at a sample size $\tilde{n} < 30$. How small can $\tilde{n}$ be so that the second option is preferable to the first option? Simplify where practical, but your final answer can be an expression, not a number.

Solution:

The variance of $\hat{\beta}_1$ is

$$\frac{\sigma^2}{n(1-r^2)}$$

Thus we need

$$\frac{\sigma^2}{30(1-0.4^2)} \geq \frac{\sigma^2}{\tilde{n}(1-0.2^2)}$$

so

$$\tilde{n} \geq \frac{30(1-0.4^2)}{1-0.2^2} \approx 26.$$

3. Suppose we have a multiple regression analysis involving p covariates $x_1, \ldots, x_p$ and a response y. We regress y on $x_1, \ldots, x_p$. Then, we rescale the variables to obtain a new set of p variables $\tilde{x}_j = f_j x_j$, where the $f_j > 0$ are known constants. Briefly describe how each of the following quantities changes in doing this:

   (a) The fitted values $\hat{y}$ using the original covariates compared to the fitted values $\tilde{y}$ using the rescaled covariates.

   Solution:

   Since $\text{span}(x_1, \ldots, x_p) = \text{span}(\tilde{x}_1, \ldots, \tilde{x}_p)$, the fitted values do not change.

   (b) The regression coefficient estimates $\hat{\beta}_j$ for the original covariates compared to the regression coefficient estimates $\tilde{\beta}_j$ for the rescaled covariates.

   Solution:

The fitted values are the same, which can be achieved by ensuring that $\hat{\beta}x_j = \tilde{\beta}\tilde{x}_j$. Thus, $\tilde{\beta}_j = \hat{\beta}_j/f_j$.

(c) The residuals based on the original covariates compared to the residuals based on the rescaled covariates.

Solution:

Since the fitted values do not change, the residuals do not change.

(d) The mean squared error $\hat{\sigma}^2$ based on the original covariates compared to the mean squared error based on the rescaled covariates.

Solution:

Since the residuals do not change, the mean squared error does not change.

(e) The standard errors of the coefficient estimates based on the original covariates compared to the standard errors of the coefficient estimates based on the rescaled covariates.

Solution:

Since $\tilde{\beta}_j = \hat{\beta}_j/f_j$, it follows that $\mathrm{SD}(\tilde{\beta}_j) = \mathrm{SD}(\hat{\beta}_j)/f_j$.

(f) The Z-scores for the regression coefficient estimates (the estimates divided by their standard error) using the original covariates compared to the Z-scores using the rescaled covariates.

Solution:

$\tilde{\beta}_j/\mathrm{SD}(\tilde{\beta}_j) = \hat{\beta}_j/\mathrm{SD}(\hat{\beta}_j)$, so the Z-scores do not change.

4. Suppose we have covariates x_1 and x_2 that are random variables with zero mean, unit variance, and $\mathrm{cor}(x_1, x_2) = r$. We observe data from the linear model $y = x_1 + \beta x_2 + \epsilon$, where $\mathrm{var}(\epsilon|x) = \sigma^2$. Let Q_1 denote the limiting partial R^2 for the model including both covariates relative to the model including only x_2. Let Q_2 denote the limiting partial R^2 for the model including both covariates relative to the model including only x_1. In all cases, the intercept is also included in the model.

(a) What is the (unconditional) variance of y?

Solution: $1 + \beta^2 + 2r\beta + \sigma^2$.

(b) What are the values of Q_1 and Q_2 in the special case where $r = 0$?

Solution: The full model R^2 is

$$\frac{1+\beta^2}{1+\beta^2+\sigma^2}.$$

The R^2 for the model including x_1 only is

$$\frac{1}{1+\beta^2+\sigma^2}.$$

The R^2 for the model including x_2 only is

$$\frac{\beta^2}{1+\beta^2+\sigma^2}.$$

Thus,

$$Q_1 = \frac{1}{1+\sigma^2}$$

$$Q_2 = \frac{\beta^2}{\beta^2+\sigma^2}.$$

(c) What are the limiting values of Q_1 and Q_2 as $|r| \to 1$?

Solution: Both are equal to zero, since all the information in one of x_1, x_2 is already contained in the other. Thus there is no gain in explained variance when including them both when one is already in the model.

5. Suppose we are planning a study in which an $n \times 11$ design matrix X_n satisfying $X_n'X_n = nI$ will be generated, and responses y will be then be obtained that follow a linear model meeting the usual conditions. We aim to produce a set of 95% simultaneous confidence intervals for $\beta_1, \ldots, \beta_{10}$ using the Bonferroni procedure.

(a) Given that $\sigma^2 = 1$ (assumed known), how large should n be so that the intervals are all 0.5 units wide? You can use the following table of normal quantiles:

p	$q : P(Z \leq q) = p$
0.95000	1.64
0.97500	1.96
0.99000	2.32
0.99500	2.58
0.99750	2.81
0.99950	3.29
0.99975	3.48

Solution:

The standard errors are all equal to $\sigma/\sqrt{n} = 1/\sqrt{n}$. Suppose we want to have at least 95% simultaneous coverage for the 10 intervals. Then the intervals will have the form $\hat{\beta}_j \pm 2.81/\sqrt{n}$, since 2.81 is the $1 - .025/10 = 0.9975$ quantile of the standard normal distribution. Thus the intervals are $5.62/\sqrt{n}$ units wide. We set $5.62/\sqrt{n} = 0.5$ and solve to get $n \approx 126$.

(b) Now consider the more general setting where we have p mutually orthonormal covariates, a sample size n, and a known value of σ^2. Let w denote the width of each interval within a collection of p simultaneous 95% coverage intervals obtained using the Bonferroni method. (i) Calculate the widths of these intervals using the crude approximation to the standard normal quantile function $Q(1 - u) \approx \sqrt{-2 \log u}$. (ii) Now suppose we have q mutually orthogonal covariates, instead of $p = 10$ mutually orthogonal covariates. How large must q be in order for the width of the simultaneous CI based on q covariates to be twice the width based on $p = 10$ covariates, when σ^2 and n are held fixed?

Solution:

The intervals have the form $f \cdot \mathrm{SE} = f\sigma/\sqrt{n}$, where f is the $1 - 0.025/p$ quantile of the standard normal distribution. Using the given approximation, $Q(1 - 0.025/p) \approx \sqrt{-2\log(0.025/p)} = \sqrt{-2\log(0.025) + 2\log(p)}$. Thus the widths for part (i) are

$$w = 2\sqrt{-2\log(0.025) + 2\log(p)} \cdot \sigma/\sqrt{n}.$$

For part (ii), we need to solve:

$$\begin{aligned} 2 &= \frac{2\sqrt{-2\log(0.025) + 2\log(q)}}{2\sqrt{-2\log(0.025) + 2\log(p)}} \\ &= \sqrt{\frac{k + 2\log(q)}{k + 2\log(p)}} \end{aligned}$$

where $k = -2\log(0.025)$.

Thus

$$4 = \frac{k + 2\log(q)}{k + 2\log(p)}$$

and

$$q = \exp(3k/2)p^4.$$