Statistics 600 Problem Set 2

Due Wednesday October 20th at midnight

1. Suppose we have a least squares analysis in which the design matrix satisfies the structure shown on slide 94 of the least squares notes. The model coefficients are β_0, β_1, and β_2, and the main interest lies in the difference $\theta = \beta_2 - \beta_1$. We plan to conduct a Wald test of the null hypothesis $\theta = 0$ (equivalently, $\beta_2 = \beta_1$). The false positive probability of this test will be set to 0.05. The power of this test is the probability that the null hypothesis is rejected for given values of θ, n, r, and σ^2. For simplicity treat $\hat{\beta}$ as Gaussian and σ^2 as known. Derive an expression for the power of the test as a function of n, r, σ^2 and θ, and discuss how the power varies with each of these values.

2. Suppose we have a least squares analysis yielding fitted values $\hat{y} = x'\hat{\beta}$, based on a data set of n observations. We wish to construct a 95% prediction interval for a new observation y^* taken at a given x^* (which need not be one of the x_i in the data used to estimate β). Construct a prediction interval in the following way: (i) append x^* to the bottom of the design matrix, which now has $n+1$ rows, (ii) append a candidate value $\tilde{y}$ to the bottom of the vector y of responses, so that it now has length $n+1$, (iii) use least squares to fit a linear model to the $n+1$ observations constructed in steps (i)-(ii), (iv) calculate the $n+1$ residuals r_i, (v) $\tilde{y}$ is in the prediction interval if and only if $\sum_{i=1}^{n+1} \mathcal{I}(|r_{n+1}| \geq |r_i|)/(n+1) \leq 0.95$. You can assume that the set of y accepted in step (v) is an interval, and approximate the end points of this interval by selecting an appropriate grid. Conduct a simulation study to assess the coverage probability of this procedure.

3. Suppose we have p covariates and an intercept in our model, and we calculate the partial R^2 for adding covariate x_1 to the model that contains $x_2, \ldots, x_p$. and an intercept. Show how this partial R^2 value can be monotonically related to an F statistic.

4. Suppose we are applying the Bonferroni procedure in a setting where the endpoints of the confidence intervals are mutually independent, i.e. if (L_i, U_i) are the lower and upper limits of the interval for the i^{th} target value, then $\{(L_i, U_i)\}$ is a collection of independent 2-vectors (note that L_i and U_i are not independent of each other). Moreover, each individual

confidence interval has coverage probability α. Derive an expression for the simultaneous coverage probability of the collection of intervals. Then use a limiting argument to demonstrate that we do not need to worry about the Bonferroni procedure being too conservative in this setting.

5. Suppose we are conducting a multiple regression analysis with p covariates, where $X'X/n = (1-r)I_n + r1_n1_n'$, i.e. the sample correlation coefficient for each pair of covariates is $r \geq 0$ and the covariates are standardized (for this problem it is irrelevant to consider whether there is an intercept in the design). The sample size is large enough that we can treat σ^2 as known and for simplicity take $\sigma^2 = 1$, also we can treat $\hat{\beta}$ as being Gaussian. We use the Bonferroni procedure to construct a collection of simultaneous coverage confidence intervals for the p coefficients in β. (a) Use simulation to estimate the simultaneous coverage probability of this collection of intervals. Display your results as a graph of simultaneous coverage probability versus r, for at least three values of p. (b) Repeat part a for the collection of $\binom{p}{2}$ confidence intervals for $\beta_j - \beta_k$, where $j < k$.

6. Suppose A and B are symmetric matrices, and A, B, and $A + B$ are all idempotent. Show that $AB \equiv 0$.

7. Show that if u and v are jointly Gaussian random values, then there exists a constant c such that $u - cv$ is independent of v. Then, suppose that we have two covariates x_1 and x_2, and for this exercise treat them as jointly Gaussian random values each with mean zero and unit variance, and suppose that $r = \mathrm{cor}(x_1, x_2)$. We then decide to include the product interaction x_1x_2 as a third covariate in the model. What are the values of $E[x_1x_2]$, $\mathrm{var}[x_1x_2]$, and $\mathrm{cor}(x_1, x_1x_2)$?

8.

   (a) Suppose that U and V are bivariate Gaussian, and g is a function. Show that

$$\mathrm{cov}(g(U), V) = \mathrm{cov}(U, V)\mathrm{cov}(g(U), U)/\mathrm{var}(U).$$

   (b) Suppose that we have a "single index model" in which

$$y = g(\alpha + \beta' x) + \epsilon,$$

Based on part (a), show that if x is multivariate Gaussian with a non-singular covariance matrix, then OLS regression of y on x yields $\hat{y} = \hat{\alpha} + \hat{\beta}' x$, where $\hat{\beta}$ consistently (as the sample size n goes to infinity) estimates $k\beta$ for a constant $k \in \mathcal{R}$.

9. Suppose we partition our nonsingular design matrix $X = [X_0 \ X_1]$, where X_0 is $n \times p$, X_1 is $n \times q$, and X is $n \times p + q$. We wish to test the null hypothesis $E[y] \in \text{col}(X_0)$ versus the alternative hypothesis $E[y] \in \text{col}(X)$. Let $r = (I - P_0)y$ denote the residuals when regressing y on X_0, and consider the vector $X_1' r$. Briefly describe why $X_1' r$ having large elements (in magnitude) is evidence against the null hypothesis. Then derive the covariance matrix V of $X_1' r$. How can V and $X_1' r$ be combined into a test statistic appropriate for this setting? What is the reference distribution of this test statistic? For simplicity, treat σ^2 as known and take $y - E[y]$ to follow a Gaussian distribution with covariance $\sigma^2 I_{n \times n}$ under the null hypothesis.

10. Suppose we have a design matrix as in slide 94 of the least squares notes. We plan to construct a confidence interval for $\cos(\theta)\beta_1 + \sin(\theta)\beta_2$, for $0 \leq \theta < 2\pi$. Taking σ^2 to be known, which values of θ give the narrowest and widest confidence intervals? What is the ratio of the lengths of the widest and narrowest intervals?