

Statistics 600 Midterm 1

October 15, 2009

1. Suppose we have a sequence of design matrices M_1, M_2, \dots for a simple linear regression analysis. The data are generated from a linear model whose parameters α , β and σ are the same for all of the design matrices. The i^{th} matrix M_i is obtained from the $i - 1^{\text{st}}$ matrix M_{i-1} by appending two rows:

$$\begin{pmatrix} 1 & i^{-\gamma} \\ 1 & -i^{-\gamma} \end{pmatrix}$$

where γ is a real constant. What conditions on γ are required for $\text{var}(\hat{\beta}) \rightarrow 0$ as $i \rightarrow \infty$?

Solution: The variance of $\hat{\beta}$ fit using design matrix M_k is

$$\frac{\sigma^2}{\sum_{i=1}^k (X_i - \bar{X})^2} = \frac{\sigma^2}{2 \sum_{i=1}^k i^{-2\gamma}}.$$

Thus the variance converges to zero if and only if $\sum_{i=1}^k i^{-2\gamma}$ converges to infinity. Since

$$\sum_{i=1}^k i^{-2\gamma} \sim \int_1^k x^{-2\gamma} dx \sim k^{-2\gamma+1}$$

we need $-2\gamma + 1 > 0$, so $\gamma < 1/2$.

2. Suppose u_1, \dots, u_n are independent and identically distributed random variables. Derive an expression for $\text{cor}(u_i - \bar{u}, u_j - \bar{u})$.

Solution: Let $\sigma^2 = \text{var}(u_i)$. Then

$$\begin{aligned} \text{cov}(u_i - \bar{u}, u_j - \bar{u}) &= \sigma^2 \mathcal{I}_{i=j} - 2\text{cov}(u_i, \bar{u}) + \text{cov}(\bar{u}, \bar{u}) \\ &= \sigma^2 \mathcal{I}_{i=j} - 2\sigma^2/n + \sigma^2/n \\ &= \sigma^2 \mathcal{I}_{i=j} - \sigma^2/n. \end{aligned}$$

So if $i \neq j$,

$$\text{cor}(u_i - \bar{u}, u_j - \bar{u}) = -\frac{1}{n-1}.$$

and if $i = j$, the correlation is one.

3. (a) The matrix P given below is a projection matrix. What is the dimension of the space that it projects onto?

$$P = \begin{pmatrix} 0.48474926 & -0.14873891 & -0.45908052 & -0.12995857 \\ -0.14873891 & 0.73418057 & -0.17743776 & 0.37623402 \\ -0.45908052 & -0.17743776 & 0.58191562 & -0.03241529 \\ -0.12995857 & 0.37623402 & -0.03241529 & 0.19915456 \end{pmatrix}$$

Solution: The dimension of the space that a projection matrix projects to is the trace of the matrix, which is 2.

- (b) Consider a matrix of the form

$$F = I - 2\frac{uu'}{u'u}$$

where $u \in \mathcal{R}^p$. (i) Is F sometimes, always, or never an orthogonal matrix?; (ii) Is F sometimes, always, or never a projection matrix?

Solution: By direct calculation, $F'F = I$, so F is always orthogonal (unless $u = 0$, in which case F is not defined. Since F is square and orthogonal, F is full rank, but a full rank projection matrix must be the identity matrix. Since there is no value of u for which F is the identity matrix, it can never be a projection. Just for your information, F is called a “Householder reflection.” Geometrically, it is a reflection in the plane with normal vector u .

4. Suppose we have the following design matrix, for a data generating model that satisfies $E(Y|X) = X\beta$ and $\text{var}(Y|X) = \sigma^2I$:

$$X = \begin{pmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \\ 1 & 2 & 1 \\ 1 & -1 & -2 \end{pmatrix}.$$

- (a) Approximately what will be the width of a 95% confidence interval for $\beta_1 - \beta_2$ (where $\beta = (\beta_0, \beta_1, \beta_2)'$)?

Solution:

$$X'X = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 6 & 4 \\ 0 & 4 & 6 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 3/10 & -1/5 \\ 0 & -1/5 & 3/10 \end{pmatrix}$$

Let $d = (0, 1, -1)'$. Then

$$\text{var}(\hat{\beta}_1 - \hat{\beta}_2) = \sigma^2 d'(X'X)^{-1}d = \sigma^2.$$

The 95% confidence interval for $\beta_1 - \beta_2$ is $\hat{\beta}_1 - \hat{\beta}_2 \pm 2\text{SE}(\hat{\beta}_1 - \hat{\beta}_2)$. Thus the width is around 4σ in this case.

- (b) We then collect additional data by extending X with k copies of the following two cases:

$$\begin{pmatrix} 1 & -1 & -1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Derive an expression for $\text{var}(\hat{\beta}_1 + \lambda\hat{\beta}_2)$ as a function of k and λ . Calculate the limit of this expression as $k \rightarrow \infty$ for fixed λ . For what value (or values) of λ is the limit equal to zero?

Solution: Let X_k denote the extended design matrix. Then

$$\begin{aligned} X_k'X_k &= \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 3/10 & -1/5 \\ 0 & -1/5 & 3/10 \end{pmatrix} + k \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 2 \\ 0 & 2 & 2 \end{pmatrix} \\ &= \begin{pmatrix} 1/4 + 2k & 0 & 0 \\ 0 & 3/10 + 2k & -1/5 + 2k \\ 0 & -1/5 + 2k & 3/10 + 2k \end{pmatrix} \end{aligned}$$

$$(X_k'X_k)^{-1} = \begin{pmatrix} 4/(1+8k) & 0 & 0 \\ 0 & (3/10+2k)/f & (1/5-2k)/f \\ 0 & (1/5-2k)/f & (3/10+2k)/f \end{pmatrix}$$

where

$$f = (3/10 + 2k)^2 - (1/5 - 2k)^2 = 1/20 + 2k.$$

Let $d = (0, 1, \lambda)'$ as in part (a), then

$$d'(X_k'X_k)^{-1}d = \frac{3/10 + 2k(1 + \lambda^2 - 2\lambda) + 3\lambda^2/10 + 2\lambda/5}{1/20 + 2k}.$$

This will converge to zero as $k \rightarrow \infty$ if and only if $\lambda^2 - 2\lambda + 1 = (\lambda - 1)^2 = 0$, which is equivalent to $\lambda = 1$.

The intuition here is that if you consider the two rows being added to the design matrix, they both have $X_1 = X_2$. Thus $\beta_1X_1 + \beta_2X_2 = (\beta_1 + \beta_2)X_1$ - we can learn about the sum of the β 's but not about their difference.

5. Suppose we collect data $Y_i, X_i, i = 1, \dots, n$, where $Y_i \in \mathcal{R}$ and $X_i \in \mathcal{R}$. The X_i follow the alternating sequence $1, -1, 1, -1, \dots$. The regression function is linear: $E(Y_i|X_i) = \alpha + \beta X_i$. The variance function is $\text{var}(Y_i|X_i) = \sigma_1^2$ if i is even, and $\text{var}(Y_i|X_i) = \sigma_2^2$ if i is odd, where σ_1^2 and σ_2^2 are two possibly different non-negative values. The error terms are uncorrelated across i . Suppose we carry out ordinary least squares regression and calculate the usual estimate $\hat{\sigma}^2$ of σ^2 . What is $E\hat{\sigma}^2$? You can assume that n is even.

Solution: We know that

$$E\hat{\sigma}^2 = \text{tr}((I - P)\Sigma_\epsilon)/(n - 2),$$

where P is the projection matrix onto the design space, which has the following form:

$$P = n^{-1} \begin{pmatrix} 2 & 0 & 2 & 0 & \cdots \\ 0 & 2 & 0 & 2 & \cdots \\ 2 & 0 & 2 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The diagonal of $(I - P)\Sigma_\epsilon$ is $(1 - 2/n)\sigma_1^2, (1 - 2/n)\sigma_2^2, (1 - 2/n)\sigma_1^2, \dots$. Thus the trace of $P\Sigma_\epsilon$ is $(n/2 - 1)\sigma_1^2 + (n/2 - 1)\sigma_2^2$, so

$$E\hat{\sigma}^2 = (\sigma_1^2 + \sigma_2^2)/2.$$

6. Suppose we fit a regression model involving 20 covariates. The covariates are mostly orthogonal, except for the pairs (1,2), (3,4), etc. The upper left 5×5 block of $X'X/n$ is

$$X'X/n = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & r & 0 & 0 & \cdots \\ 0 & r & 1 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 1 & r & \cdots \\ 0 & 0 & 0 & r & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

and the same pattern continues throughout the rest of the matrix. The data generating model is $E(Y|X) = X\beta$ and $\text{var}(Y|X) = \sigma^2 I$.

We are interested in identifying any $j > 0$ for which $\beta_j \neq 0$, and we plan to do this by carrying out a hypotheses test for each covariate, using the Bonferroni method to adjust for multiple comparisons. What is the power (the probability of rejecting the null hypothesis) for detecting $\beta_1 = 1$? Your answer can depend on the standard normal CDF F and its inverse F^{-1} .

Solution: Since the matrix has a block diagonal form, the diagonal elements of $(X'X)^{-1}$ are all equal to $\sigma^2/(n(1-r^2))$. Using the Bonferroni method, we set the rejection region for the test of $\beta_1 \neq 0$ based on

$$P_0(|\hat{\beta}_1| > T) = \alpha/20,$$

where α is the level of the test, and P_0 denotes the probability under the null hypothesis. Thus we need

$$P_0(\hat{\beta}_1 > T) = \alpha/40.$$

Since

$$P_0(\sqrt{n(1-r^2)}\hat{\beta}/\sigma > T\sqrt{n(1-r^2)}/\sigma) = P(Z > T\sqrt{n(1-r^2)}/\sigma) = \alpha/40,$$

we get $T = F^{-1}(1 - \alpha/40)\sigma/\sqrt{n(1-r^2)}$. The power is therefore

$$P_1(|\hat{\beta}_1| > T) = P_1(\hat{\beta}_1 > T) + P_1(\hat{\beta}_1 < -T)$$

where P_1 denotes the probability under the alternative hypothesis.

$$\begin{aligned} P_1(\hat{\beta}_1 > T) &= P_1((\hat{\beta}_1 - 1)\sqrt{n(1-r^2)}/\sigma > (T - 1)\sqrt{n(1-r^2)}/\sigma) \\ &= P(Z > F^{-1}(1 - \alpha/40) - \sqrt{n(1-r^2)}/\sigma) \end{aligned}$$

$$\begin{aligned}
P_1(\hat{\beta}_1 < -T) &= P_1((\hat{\beta}_1 - 1)\sqrt{n(1-r^2)}/\sigma < (-T - 1)\sqrt{n(1-r^2)}/\sigma) \\
&= P(Z < -F^{-1}(1 - \alpha/40) + \sqrt{n(1-r^2)}/\sigma)
\end{aligned}$$

The power is the sum of these two pieces, but only one will contribute a non-negligible amount to the power, depending on whether $\beta_1 > 0$ or $\beta_1 < 0$.