

## Statistics 600 Exam 1

October 14, 2008

1. (a) Suppose we fit a simple linear regression between  $Y$  and  $X$ . Our goal is to estimate the coefficient  $\beta$  in the relationship  $E(Y|X) = \alpha + \beta X$  with the least possible variance, where  $\text{cov}(Y|X) = \sigma^2 I$  can be assumed to hold. Consider the following situations:

A The sample size is 20 and  $\widehat{\text{var}}(X) = 2$

B The sample size is 30 and  $\widehat{\text{var}}(X) = 1$

C The sample size is 15 and  $\widehat{\text{var}}(X) = 3$

D The sample size is 10 and  $\widehat{\text{var}}(X) = 4$

Which situation is most favorable? Very briefly justify your answer.

**Solution:** Since  $\text{var}(\hat{\beta}) = \sigma^2 / ((n-1)\widehat{\text{var}}(X))$ , we need to maximize  $(n-1)\widehat{\text{var}}(X)$ , which occurs for option C.

- (b) Construct the projection matrix onto the span of  $(1, 0, 1, 0)'$  and  $(1, 1, 1, 1)'$ .

**Solution:** The span is the same as the orthogonal basis  $(1, 0, 1, 0)'$  and  $(0, 1, 0, 1)'$ .

$$\frac{1}{2} \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

- (c) True or false:  $\widehat{\text{cov}}(Y, \hat{Y}) \geq 0$ ? If true, provide a proof; if false, provide a counterexample.

**Solution:** The statement is true:

$$\begin{aligned} \|Y - \hat{Y}\|^2 &= \|Y - \bar{Y} + \bar{Y} - \hat{Y}\|^2 \\ &= \|Y - \bar{Y}\|^2 + \|\hat{Y} - \bar{Y}\|^2 - 2(Y - \bar{Y})'(\hat{Y} - \bar{Y}) \end{aligned}$$

Note that  $\widehat{\text{cov}}(Y, \hat{Y})$  has the same sign as  $(Y - \bar{Y})'(\hat{Y} - \bar{Y})$ . If  $(Y - \bar{Y})'(\hat{Y} - \bar{Y}) < 0$ , we can replace  $\hat{Y}$  with  $-\hat{Y} + 2\bar{Y}$ , which is also in  $\text{col}(X)$ . This doesn't change the value of  $\|Y - \bar{Y}\|^2$  or  $\|\hat{Y} - \bar{Y}\|^2$ , but the third term becomes positive, giving a lower value for  $\|Y - \hat{Y}\|^2$ . Since the OLS estimator minimizes  $\|Y - \hat{Y}\|^2$ , it follows that  $(Y - \bar{Y})'(\hat{Y} - \bar{Y}) \geq 0$  in the first place.

2. Suppose we observe data from the multivariate linear model  $Y = X\beta + \epsilon$ , where  $Y \in \mathcal{R}^n$ ,  $\beta \in \mathcal{R}^{p+1}$ ,  $X \in \mathcal{R}^{n \times p+1}$ ,  $E(\epsilon|X) = 0$  and  $\text{cov}(\epsilon|X) = \sigma^2 I$ .

Let  $B \equiv (X'X)^{-1}X'$ , so  $\hat{\beta} = BY$ , and let  $\tilde{B} \equiv B + uv'$ , where  $u$  and  $v$  are column vectors, and without loss of generality,  $\|v\| = 1$ .

(a) Under what condition will  $\tilde{\beta} \equiv \tilde{B}Y$  be an unbiased estimate of  $\beta$ ?

**Solution:**

$$\begin{aligned} E\tilde{\beta} &= E\tilde{B}Y \\ &= BEY + uv' EY \\ &= \beta + uv' X\beta \end{aligned}$$

If  $u \equiv 0$  then  $\tilde{\beta}$  is unbiased. Otherwise,  $X\beta$  could be anything in  $\text{col}(X)$ , so we need  $v \in \text{col}(X)^\perp$ , or  $v'X \equiv 0$ .

(b) If condition (a) holds, what is the covariance matrix of  $\tilde{\beta}$  given  $X$ ?

**Solution:**

$$\begin{aligned} \text{cov } \tilde{\beta} &= \text{cov } BY + \text{cov}(BY, uv'Y) + \text{cov}(uv'Y, BY) + \text{cov } uv'Y \\ &= \text{cov } BY + \text{cov}(BY, uv'\epsilon) + \text{cov}(uv'\epsilon, BY) + \text{cov } uv'Y \\ &= \sigma^2(X'X)^{-1} + \sigma^2 Bvu' + \sigma^2 uv'B' + \sigma^2 uu' \\ &= \sigma^2 \left( (X'X)^{-1} + uu' \right). \end{aligned}$$

3. Suppose we observe data from the multivariate linear model  $Y = X\beta + \epsilon$ , where  $Y \in \mathcal{R}^n$ ,  $\beta \in \mathcal{R}^{p+1}$ ,  $X \in \mathcal{R}^{n \times p+1}$  (with first column identically equal to 1),  $E(\epsilon|X) = 0$ , and  $\text{cov}(\epsilon|X) = \sigma^2 I$ . Based on these data, we construct an estimate  $\hat{\beta}$  of  $\beta$  using ordinary least squares. Suppose  $x^* \in \mathcal{R}^{p+1}$  is sampled from a population with mean  $\mu_x^*$  and covariance matrix  $\Sigma_x^*$ , which is a degenerate distribution where  $x_1^* \equiv 1$ . The corresponding value  $y^* = \alpha + x^{*'}\beta + \epsilon^*$ , where  $E(\epsilon^*|x^*) = 0$  and  $\text{var}(\epsilon^*) = \text{var}(\epsilon_1)$ , is not observed. We then predict  $y^*$  using  $x^{*'}\hat{\beta}$ . What is  $E(y^* - x^{*'}\hat{\beta})^2$ , where the expectation is only conditioned on  $X$ ?

**Solution:** Since  $E(y^* - x^{*'}\hat{\beta}|x^*) = 0$ , and given  $x^*$ ,  $y^*$  is independent of  $x^{*'}\hat{\beta}$ , it follows that

$$\begin{aligned} E \left( (y^* - x^{*'}\hat{\beta})^2 | x^* \right) &= \text{var}(y^* - x^{*'}\hat{\beta} | x^*) \\ &= \sigma^2 + \sigma^2 x^{*'}(X'X)^{-1}x^* \\ &= \sigma^2 + \sigma^2 \text{tr} \left( (X'X)^{-1}x^*x^{*'} \right). \end{aligned}$$

Now if we take the expectation over  $x^*$ , we get

$$\sigma^2 \left( 1 + \text{tr} \left( (X'X)^{-1} (\Sigma_x^* + \mu_x^* \mu_x^{*'}) \right) \right).$$

4. Suppose we regress  $Y$  on  $p = 2$  centered covariates with an intercept. If the two covariates are positively correlated, which of the following can be estimated with more precision using the usual regression-based estimate:  $\beta_1 - \beta_2$ , or  $\beta_1 + \beta_2$ ? Briefly explain your reasoning.

**Solution:** If the covariates are positively correlated, then  $\hat{\beta}_1$  and  $\hat{\beta}_2$  are negatively correlated. Therefore  $\hat{\beta}_1 + \hat{\beta}_2$  is less variable than  $\hat{\beta}_1 - \hat{\beta}_2$ , so  $\hat{\beta}_1 + \hat{\beta}_2$  can be estimated with more precision. Another way to think about it is that since the errors are negatively correlated, they will tend to cancel each other out when added, but will tend to amplify each other when subtracted.

5. Suppose we observe data from a simple linear model  $Y = \alpha + \beta X + \epsilon$  where  $X, Y \in \mathcal{R}^n$ ,  $E(\epsilon|X) = 0$  and  $\text{cov}(\epsilon|X) = \sigma^2 I$ . Suppose  $X$  and  $Y$  are partitioned as

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \quad X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},$$

where  $Y_1$  and  $Y_2$  each have half the length of  $Y$ , and  $X_1$  and  $X_2$  each have half the length of  $X$ . Let  $\hat{\beta}_1$  and  $\hat{\beta}_2$  denote the least squares estimates obtained by regressing  $Y_1$  on  $X_1$  and  $Y_2$  on  $X_2$ , respectively, and let  $\tilde{\beta} = (\hat{\beta}_1 + \hat{\beta}_2)/2$ .

As notation for this problem, let  $S_1 = \sum_{i=1}^{n/2} (X_i - \bar{X}_1)^2$ ,  $S_2 = \sum_{i=n/2+1}^n (X_i - \bar{X}_2)^2$ , and  $S = \sum_{i=1}^n (X_i - \bar{X})^2$ .

- (a) If  $\bar{X}_1 = \bar{X}_2 = \bar{X}$ , state a condition such that  $\tilde{\beta}$  has the same variance as the least squares estimate  $\hat{\beta}$  obtained by regressing  $Y$  on  $X$ . Then state whether when this condition holds,  $\tilde{\beta}$  is the least squares estimate, or is a different estimate with the same variance.

**Solution:** The variance of  $\tilde{\beta}$  is

$$\text{var } \tilde{\beta} = \frac{\sigma^2}{4} (1/S_1 + 1/S_2) = \frac{\sigma^2(S_1 + S_2)}{4S_1S_2}.$$

Since  $\bar{X}_1 = \bar{X}_2$ ,  $S_1 + S_2 = S$ , so the variance is  $\sigma^2 S / (4S_1S_2)$ . The variance of  $\hat{\beta}$  is  $\sigma^2/S$ , so equating the variances we get  $S/(4S_1S_2) = 1/S$ , which is equivalent to  $S^2 = 4S_1S_2$ . Substituting  $S = S_1 + S_2$  and expanding the square gives as an equivalent equation  $(S_1 - S_2)^2 = 0$ , so we see that  $S_1 = S_2$  is the condition required for the variances to be equal. When  $S_1 = S_2$ , it follows that  $S_1 = S_2 = S/2$ , and

$$\begin{aligned} \tilde{\beta} &= (\hat{\beta}_1 + \hat{\beta}_2)/2 \\ &= \frac{1}{2} \sum Y_{1i}(X_{1i} - \bar{X}_1)^2/S_1 + \frac{1}{2} \sum Y_{1i}(X_{2i} - \bar{X}_2)^2/S_2 \\ &= \sum Y_{1i}(X_{1i} - \bar{X})^2/S + \sum Y_{1i}(X_{2i} - \bar{X})^2/S \\ &= \sum Y_i(X_i - \bar{X})^2/S \\ &= \hat{\beta}, \end{aligned}$$

so the estimators are identical.

- (b) Now suppose  $\bar{X}_1 = \bar{X}_2 = \bar{X}$ , but the condition you found in part (a) does not hold. Show how an estimate that is linear in  $\hat{\beta}_1$  and  $\hat{\beta}_2$  can be constructed that has the same variance as the least squares estimate.

**Solution:** If we write our estimate as  $c_1\hat{\beta}_1 + c_2\hat{\beta}_2$ , then setting  $c_1 = S_1/S$  and  $c_2 = S_2/S$  yields the least squares estimate.

- (c) Now consider the more general case where  $\bar{X}_1$  and  $\bar{X}_2$  may differ. Show that in this case  $\text{var } \tilde{\beta}$  is always greater than  $\text{var } \hat{\beta}$ , and derive a concise expression for the difference.

**Solution:** Let  $D = (\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2$ , so

$$\begin{aligned} S &= \sum_{i=1}^{n/2} (X_i - \bar{X}_1 + \bar{X}_1 - \bar{X})^2 + \sum_{i=n/2+1}^n (X_i - \bar{X}_2 + \bar{X}_2 - \bar{X})^2 \\ &= \sum_{i=1}^{n/2} (X_i - \bar{X}_1)^2 + (\bar{X}_1 - \bar{X})^2 + \sum_{i=n/2+1}^n (X_i - \bar{X}_2)^2 + (\bar{X}_2 - \bar{X})^2 \\ &= S_1 + S_2 + n(\bar{X}_1 - \bar{X})^2/2 + n(\bar{X}_2 - \bar{X})^2/2 \\ &= S_1 + S_2 + nD/2. \end{aligned}$$

So

$$\text{var}(\tilde{\beta}) = \frac{\sigma^2(S_1 + S_2)}{4S_1S_2} = \frac{\sigma^2(S - nD/2)}{4S_1S_2}.$$

Therefore

$$\begin{aligned}\text{var } \tilde{\beta} - \text{var } \hat{\beta} &= \sigma^2(S - nD/2)/(4S_1S_2) - \sigma^2/S \\ &= \frac{\sigma^2}{4S_1S_2S}(S(S_1 + S_2) - 4S_1S_2) \\ &= \frac{\sigma^2}{4S_1S_2S} \left( (S_1 + S_2)^2 - 4S_1S_2 + nD(S_1 + S_2)/2 \right) \\ &= \frac{\sigma^2}{4S_1S_2S} \left( (S_1 - S_2)^2 + n(S_1 + S_2)D/2 \right).\end{aligned}$$

Note that the two become equal only when  $S_1 = S_2$  and  $D = 0$ , which are the conditions from part (a) above.