

Statistics 600 Problem Set 3

Due in class on Tuesday, November 17th

1. Prove the following identity involving the fitted slope vector $\hat{\beta}_{(i)}$ calculated when the i^{th} case has been removed from the data set.

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{R_i}{1 - P_{ii}}(X'X)^{-1}X'_i,$$

where R_i is the i^{th} residual, X_i is the i^{th} row of the design matrix, and P is the projection matrix onto the design space. Both R and P are calculated with respect to all observations (including the i^{th} observation).

2. Suppose we have a bivariate regression model with

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

and we use the Bonferonni approach to produce simultaneous confidence intervals for β_1 and β_2 with $\geq 95\%$ coverage.

- (a) Derive an analytic expression giving the actual coverage of this interval. State which of the following affect the result: n , r , σ^2 .
 - (b) Numerically evaluate your expression from part (a) for a range of values of the relevant parameters. Based on your findings, describe some circumstances under which the actual coverage is either particularly close to 0.95, or particularly far from 0.95.
3. Suppose we observe a design matrix $X \in \mathcal{R}^{n \times p+1}$ (with the first column being identically 1). The rows of X (excluding the intercept) are sampled independently from a distribution G_x with mean 0 and covariance Σ . Then we observe a response vector Y such that $EY = X\beta$ for some coefficient vector β , and such that $\text{var}(Y|X) \propto I$. We regress Y on X using OLS regression to estimate β . For this problem, expected values and variances should only be conditioned on things that are explicitly given.
 - (a) Determine an expression for $\text{var}[y^* - x^{*'}\hat{\beta}|X, x^*]$, where x^* is a fixed vector in \mathcal{R}^{p+1} with a 1 in the first position.
 - (b) Determine an expression for $\text{var}[y^* - x^{*'}\hat{\beta}|X]$.
 - (c) Consider the decomposition

$$\text{var}(y^* - x^{*'}\hat{\beta}) = \text{var}_X E(y^* - x^{*'}\hat{\beta}|X) + E_X \text{var}(y^* - x^{*'}\hat{\beta}|X),$$

which can be converted to proportions as

$$1 = \text{var}_X E(y^* - x^{*\prime} \hat{\beta} | X) / \text{var}(y^* - x^{*\prime} \hat{\beta}) + E_X \text{var}(y^* - x^{*\prime} \hat{\beta} | X) / \text{var}(y^* - x^{*\prime} \hat{\beta}).$$

Use simulation or numerical calculation to approximate these proportions. You will need to choose values for the distribution G_x , the sample size, β , etc. Can you identify any attributes of Σ that seem to exert a strong influence on these proportions?

Note that these two proportions have an interesting interpretation. The first is the proportion of prediction error due to variation in the training design matrix X . The second is the proportion of prediction error due to variation in the training responses Y .

4. Suppose we have a design matrix X with $p = 2$ covariates, and a covariate vector $\beta = (0, \lambda \sin(\theta), \lambda \cos(\theta))'$, for some scalar values of λ and θ . The response vector is generated in the usual way as $Y = X\beta + \epsilon$, where $\text{var}(\epsilon | X) = I$.

We don't observe X , rather we observe $Z = X + E$ where the first column of E is identically zero, and the second and third columns of E consist of iid centered normal measurement errors with variance τ^2 .

Determine (a)-(c) below, and note whether the results appear to depend on θ .

- (a) Determine analytically a value of λ so that the population R^2 if X is observed (i.e. the squared correlation coefficient between Y and $X\beta$) is 0.7.
 - (b) For the value of λ found in part (a), use simulation to approximate an error variance τ^2 so that the population R^2 based on Z is 0.65. You can do this by averaging $\hat{\beta}$ obtained over many replicates of Y, E to estimate $E\hat{\beta}$, then calculating the squared correlation coefficient between $Z \cdot E\hat{\beta}$ and Y .
 - (c) For the values of λ and τ^2 found in parts (a) and (b), use simulation to consider the squared correlation coefficient between observed values and fitted values (regressing Y on Z).
5. Suppose we are using the C_p statistic for model selection, and we are considering its limiting behavior as $n \rightarrow \infty$ (everything else if fixed). You can assume that $Y = X\beta + \epsilon$ with $E(\epsilon | X) = \sigma^2 I$, $\hat{\beta} \rightarrow \beta$ as $n \rightarrow \infty$, and the rows of X are sampled independently from a non-degenerate probability distribution (except for the 1 in the first position).
- (a) If we have a procedure such that $\sigma^* \rightarrow \sigma$, explain why the C_p statistic identifies the correct model consistently.
 - (b) Now suppose that the "large design matrix" X^* that we use to construct σ^{*2} does not actually contain EY . Also, the set of models considered as candidates

also does not include the correct model. However the “large design matrix” does include all the candidate models. Which of the candidate models is selected by the C_p statistic in the limit?