BST 140.753
Problem Set 2
Due Feb. 24th 11:59PM 2011

Should $X$ be included in a regression model?

Building on the example in the lecture stream, for $i = 1, \ldots, n; j = 1, 2$ let

$$[Y_{ij} \mid \alpha, \beta, \sigma] = \alpha + \beta X_j + \epsilon_{ij} \tag{1}$$

$\epsilon_{ij}$ iid $N(0, 1)$

$X_1 = -1$

$X_2 = +1$

**Part i:** Let $(\hat{\alpha}, \hat{\beta})$ be the least squares estimates.
Show that irrespective of whether $X$ is included in the model:

$$\hat{\alpha} = \bar{Y} = Y_+ = \frac{Y_++}{2n}, \text{ the overall mean}$$

and if $X$ is included,

$$\hat{\beta} = (\bar{Y}_2 - \bar{Y}_2)/2 = (Y_2 - Y_1)/2.$$

**Part ii:** Separately, for the cases when $X$ is/(is not) included in the model, compute $\hat{Y}_{ij}$.

**Part iii:** Consider using the prediction model that uses $a\hat{\beta}$ rather than $\hat{\beta}$. That is, use $\hat{Y}_{ij} = \hat{\alpha} + a\hat{\beta}X_j$.
The MSE of $\hat{\alpha} + a\hat{\beta}$ for estimating $(\alpha + \beta)$ (and since $|X| = 1$ for predicting the $Y_{ij}$) is:

$$MSE(\hat{\alpha} + a\hat{\beta}) = (1 + a^2)\frac{\sigma^2}{2n} + (1 - a)^2\beta^2.$$

If you are choosing between not including $X$ ($a = 0$) or including $X$ ($a = 1$) show that you should use $a = 1$ iff $\beta^2 > \frac{\sigma^2}{2n}$.

**Part iv:** More generally, show that the optimal $a$ is:

$$a(\sigma^2, \beta, n) = \frac{\beta^2}{\frac{\sigma^2}{2n} + \beta^2} = \frac{2n\beta^2}{\sigma^2 + 2n\beta^2} = \frac{t^2}{1 + t^2} \tag{2}$$

$$t^2 = \frac{(2n)\beta^2}{\sigma^2}.$$

Note that $t^2$ has the form of the square of the t-statistic testing $\beta = 0$.

**Part v:** Simulation
Conduct a simulation to study performance of various candidate rules for including/excluding/(partially including) $X$. Write up your results into a nice report.

**Scenarios:** Without loss of generality (wlog), in generating data set $\alpha = 0, \sigma^2 = 1$. Use $n = 10$ for a total sample size of 20 and simulate for 4 $\beta$s chosen so that the optimal $a = (0, 0.25, 0.50, 0.90)$ ($a = 0$ produces $\beta = 0$). Use $nreps = 500$. 
Dataset generation: For each of the \textit{nreps} datasets, generate the 20 values for the \( \epsilon_{ij} \) as independent \( \mathcal{N}(0,1) \) and then \( \tilde{Y}_{ij} = \beta X_j + \epsilon_{ij} \). Reuse the the \( \epsilon_{ij} \) for each of the 4 \( \beta \) values. This is efficient, but more importantly, makes the performance for the different \( \beta \)s more comparable.

Decision rules: Study the performance of the following rules. Each starts by producing the LSE estimates of \( (\alpha, \beta) \) and the estimate of \( \sigma^2 \). Each decision rule always uses \( \hat{\alpha} \), the LSE. For each dataset each decision rule produces an estimated \( \beta \). For rules that either include or exclude \( X \), the estimated \( \beta \) is either the LSE or 0. For the attenuation rules, it is the attenuated LSE.

\textbf{Full knowledge}: Use the true \( (\alpha, \beta) \). This rule is the best case and included to calibrate the simulation.

\textbf{P-value based (6 rules)}: Include \( X \) if the two-sided P-value for testing \( H_0 : \beta = 0 \) versus \( H_1 : \beta \neq 0 \) is less than \( (0, .05, .10, .32, .50, 1.00) \). The first and last rules are “never add \( X \)” and “always add \( X \).”

\textbf{MSE based}: Add \( X \) if the MSE from the regression with \( X \) included is less than the MSE when \( X \) is excluded. (This is equivalent to adding \( X \), if the \( R^2_{adj} \) is smaller when \( X \) is included that when excluded).

\textbf{Partial inclusion (semi-full knowledge)}: Let \( a(\beta) = a(\sigma^2 = 1, \beta, n = 20) \). For the slope on \( X \) use \( a(\beta)\hat{\beta} \). Note that this rule uses the true \( \beta \) in \( a(\beta) \), multiplying \( \hat{\beta} \), the LSE estimate.

\textbf{Partial inclusion (fully empirical)}: Use \( a(\hat{\sigma}^2, \hat{\beta}, n = 20)\hat{\beta} \) for the slope on \( X \). This is equivalent to computing the t-statistic (you need this for the P-value rules anyway) and using \( a = \frac{t^2}{1 + t^2} \).

Assessments: For each of the scenarios and each of the decision rules, summarize performance of the \( nreps = 500 \) estimates (\( \beta_{est} \) is a generic estimate):

- bias = \( \text{avg}(\beta_{est} - \beta) \)
- variance = sample variance of the \( \beta_{est} \)
- \( \text{MSE} = \text{avg}(\beta_{est} - \beta)^2 \) \( (= \text{variance} + (\text{bias})^2) \)

Also, report the histogram of the \( \beta_{est} \) for \( \hat{\beta} \) (the LSE), for the P-value based with \( P = .05 \) and for the partial inclusion (fully empirical) attenuated slope estimate.

Discuss your results and include your recommendation on how to decide on including/excluding/(partially including \( X \)).

\textbf{End note}: The Prediction Sum of Squares(PRESS) criterion is a very attractive alternative to other include/exclude rules. For the model without \( X \) and then for the model with \( X \), PRESS successively excludes a case, re-estimates the regression based on the reduced dataset, predicts the \( Y \) for the excluded case and averages the squared errors. \( X \) is included if its PRESS is smaller that the intercept-only model. Though the PRESS computation is closed form in this simple case, I haven’t asked you to evaluate it because it will be similar to the MSE based rule and you have enough to do!