Statistics 600 Problem Set 3

Due Monday November 1st at midnight

1. Suppose that A is a square $p \times p$ matrix, and v is a p-dimensional unit vector. (i) Derive an expression for $(A + \lambda vv')^{-1}$ in terms of A^{-1}. (ii) Provide a derivation for the deleted slope vector (the least squares estimate of β in which a single observation has been deleted), as discussed in the course notes.

2. Suppose we have a covariate x and an outcome y that are generated according to the model $y = x + \epsilon$, with $\epsilon \sim [0, \sigma^2]$ (i.e. ϵ has mean zero and variance σ^2). For simplicity take $x \sim [0, 1]$. In addition we have another covariate z that is generated according to $z = \theta x + \gamma y + \eta$, where $\eta \sim [0, \tau^2]$. The values x, ϵ, η are mutually independent. We fit a working model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 z$ using iid samples of (x, y, z). Derive an expression for the limiting value of $\hat{\beta}_1$ as n tends to infinity. Taking the causal effect of x with respect to y to be 1, state conditions under which $\hat{\beta}_1$ is consistent for its intended target.

3. Suppose we have a data-generating model of the form $E[y|x, z] = \beta x + \gamma z$, and $E[z|x] = \theta x$. Consider the following three "effects":

   - Total effect: $E[y|x = 1, z = E[z|x = 1]] - E[y|x = 0, z = 0]$.
   - Direct effect: $E[y|x = 1, z = E[z|x = 0]] - E[y|x = 0, z = 0]$
   - Indirect effect: $E[y|x = 0, z = E[z|x = 1]] - E[y|x = 0, z = 0]$

   Update (11/1): alternative definitions:

   - Total effect: $E[y|x = 1, z = E[z|x = 1]] - E[y|x = 0, z = E[z|x = 0]]$.
   - Direct effect: $E[y|x = 1, z = E[z|x = 0]] - E[y|x = 0, z = E[z|x = 0]]$
   - Indirect effect: $E[y|x = 0, z = E[z|x = 1]] - E[y|x = 0, z = E[z|x = 0]]$

   Derive expressions for each of these effects in terms of the model parameters β, γ, and θ. If all the zeros in these definitions were replaced with a constant c, and all the ones were replaced with $c + 1$, would these effects change?

4. Continuing with the previous exercise, suppose now that there is an unobserved variable u such that $E[y|x,z,u] = \beta x + \gamma z + \delta u$, $E[z|x,u] = \theta x + \zeta u$, and $E[u|x] = \lambda x$. Derive expressions for the direct and indirect effects in this setting, and discuss how these results are biased relative to what we would observe if we could control u (i.e select a sample in which $u = 0$ for all units).

5. Suppose we are considering models with interactions of the form $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 z + \hat{\beta}_3 xz$. (i) Suppose we transform x and/or z by translation, i.e. we replace x with $x - c_x$ or z with $z - c_z$. How do the parameter estimates in the new model relate to the parameter estimates in the original model? Describe how to translate x and z so that in the new model, there are no main effects.

6. Suppose we are working with a quadratic "response surface" model of the form $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 z + \hat{\beta}_3 x^2 + \hat{\beta}_4 z^2 + \hat{\beta}_5 xz$. Our goal is to identify (x, z) that minimizes $E[y|x,z]$. Derive an expression for the estimate of this value $(\hat{x}, \hat{z})$. Supposing that the $\hat{\beta}_j$ are jointly Gaussian, conduct a small simulation study to assess the sampling distribution of $(\hat{x}, \hat{z})$.

7. Suppose we have a single-index model of the form $y = f(\beta' x) + \epsilon$ where $x \in \mathcal{R}^p$ and $\epsilon \in \mathcal{R}$ are random and independent of each other, and f is an unknown function. Without loss of generality, we can take $E[x] = 0$ and $\mathrm{cov}[x] = I$. One method for estimating x is to construct the $p \times p$ matrix $M = \sum_i (y_i - \bar{y}) x_i' x_i$, where $x_i \in \mathcal{R}^p$ is a row vector. It is claimed that the dominant eigenvector of M (the eigenvector corresponding to the eigenvalue that is greatest in magnitude) estimates β. Conduct a small simulation study to demonstrate a setting in which this approach is effective.

8. Suppose we are analyzing a dependent variable y and two categorical independent variables x_1 and x_2, with k_1 and k_2 levels respectively. Let y_{ij} denote the observation made when x_1 is at level i and x_2 is at level j. Consider first the "additive model" $E[y_{ij}] = \mu + \alpha_i + \beta_j$. (i) State a natural constraint on the model parameters so that they become identified, (ii) Derive explicit forms for the least squares estimates of these parameters, subject to the constraint that you defined in (i). Next, suppose that we are interested in exploring interactions, but are not sure which interactions might be relevant. Moreover, the most general model including all possible interactions is quite high-dimensional. To

address this situation, first fit the additive model $\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$ based on (i)-(ii) above, and form residuals $r_{ij} = y_{ij} - \hat{y}_{ij}$. Then consider the model $\tilde{r}_{ij} = \hat{\mu} + \hat{\lambda}\hat{\alpha}_i\hat{\beta}_j$, where the $\hat{\alpha}_i$ and $\hat{\beta}_j$ are held fixed at their estimates from the additive fit, and $\hat{\lambda}$ is estimated using least squares. Answer the following questions using simulation: (iii) What is the type 1 error rate for λ when there is no interaction for the population?, (iv) what is the power of the test when the interaction truly has the hypothesized form?, (v) construct an interaction structure for which this test should have very low power.

9. Suppose we have two covariates x_1 and x_2 with $\bar{x_1} = \bar{x_2} = 0$, $\bar{y} = 0$, $x_1'x_1 = 1$, and $x_2'x_2 = 1$ Let $r_j = x_j'y$, and for simplicity take $r_1 > r_2 > 0$. Consider an adjusted dependent variable $y_\lambda = y - \lambda x_1$. Find the value $\hat{\lambda}$ such that $y_{\hat{\lambda}}'x_1 = y_{\hat{\lambda}}'x_2$. Describe situations in which $\hat{\lambda}$ will be less than, equal to, or greater than the simple least squares slope of y on x_1. Do the same thing in comparison to the OLS slope of y on x_1 and x_2.

10. Suppose we take the following approach to fitting a multiple regression model. First, we center and standardize all covariates, so that $X'1 = 0$ and $\mathrm{diag}(X'X)/n = (1, 1, \ldots, 1)$. Next, let $b = X'y \in \mathcal{R}^p$. We wish to construct an estimator $\tilde{b} = \lambda b$, where $\lambda \in \mathcal{R}$. We can do this through least squares, by minimizing $\|y - \lambda Xb\|^2$. Derive an explicit form for the least squares estimate of λ, for $\tilde{b}$, and for the corresponding fitted values $X\tilde{b}$. Then obtain the large sample limit for $\tilde{b}$, call this b^*. Finally, consider the expected prediction error $E\|y - Xb^*\|^2$ and compare this to the corresponding "oracle" prediction error $E\|y - X\beta\|$. Use a computer to calculate these values for a few examples (note that this does not involve simulation, you are only using the computer to do deterministic calculations).