STA 302f17 Assignment Ten[1]

These questions are preparation for the quiz in tutorial on Thursday November 30th, and are not to be handed in.

1. This question explores the practice of centering quantitative independent variables in a regression by subtracting off the mean. Geometrically, this should not alter the configuration of data points in the multi-dimensional scatterplot. All it does is shift the axes. Thus, the intercept of the least squares plane should change, but the slopes should not.

   Consider a simple experimental study with an experimental group, a control group and a single quantitative covariate. Independently for $i = 1, \ldots, n$ let

   $$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \epsilon_i,$$

   where x_i is the covariate and d_i is an indicator dummy variable for the experimental group. If the covariate is "centered," the model can be written

   $$y_i = \beta_0^* + \beta_1^*(x_i - \overline{x}) + \beta_2^* d_i + \epsilon_i,$$

   where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

   (a) Express the β^* quantities in terms of the original β quantities.

   (b) Let's generalize this. For the general linear model in matrix form suppose $\boldsymbol{\beta}^* = A\boldsymbol{\beta}$, where A is a square matrix with an inverse. This makes $\boldsymbol{\beta}^*$ a one-to-one function of $\boldsymbol{\beta}$. Of course X is affected as well. Show that $\mathbf{b}^* = A\mathbf{b}$.

   (c) Give the matrix A for this $k = 2$ model.

   (d) If the independent variable x is centered, what is $E(y|x)$ for the experimental group, and what is $E(y|x)$ for the control group? Give your answer in terms of the β^* values of the centered model.

   (e) In terms of the β^* values of the centered model, give $E(y|x)$ for the experimental group and the control group when x equals the average (sample mean) value.

[1] Copyright information is at the end of the last page.

2. In the following model, there are k quantitative independent variables. The un-centered version is

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \epsilon_i,$$

and the centered version is

$$y_i = \beta_0^* + \beta_1^*(x_{i,1} - \overline{x}_1) + \ldots + \beta_k^*(x_{i,k} - \overline{x}_k) + \epsilon_i,$$

where $\overline{x}_j = \frac{1}{n}\sum_{i=1}^n x_{i,j}$ for $j = 1, \ldots, k$.

   (a) What is β_0^* in terms of the β quantities? Show your work.

   (b) In terms of the β quantities, what is β_j^* for $j = 1, \ldots, j$?

   (c) What is b_0^* in terms of the b quantities? Note that Problem 1b lets you just write this down.

   (d) In terms of the b quantities, what is b_j^* for $j = 1, \ldots, j$?

   (e) Referring again to Problem 1b, give the A matrix for this k-variable model.

   (f) Using $\sum_{i=1}^n \widehat{y}_i = \sum_{i=1}^n y_i$, show that $b_0^* = \overline{y}$.

3. In the usual multiple regression model, the X matrix is an $n \times (k+1)$ matrix of known constants. But in practice, the independent variables are often random and not fixed. Clearly, if the model holds *conditionally* upon the values of the independent variables, then all the usual results hold, again conditionally upon the particular values of the independent variables. The probabilities (for example, p-values) are conditional probabilities, and the F statistic does not have an F distribution, but a conditional F distribution, given $\mathcal{X} = X$. Here, the $n \times (k+1)$ matrix $\mathcal{X}$ is used to denote the matrix containing the random independent variables. It does not have to be *all* random. For example the first column might contain only ones if the model has an intercept.

   (a) Show that the least-squares estimator $(X'X)^{-1}X'\mathbf{y}$ is conditionally unbiased. You've done this before.

   (b) Show that $\mathbf{b} = (\mathcal{X}'\mathcal{X})\mathcal{X}'\mathbf{y}$ is also unbiased unconditionally.

   (c) A similar calculation applies to the significance level of a hypothesis test. Let F be the test statistic (say for an F-test comparing full and reduced models), and f_c be the critical value. If the null hypothesis is true, then the test is size α, conditionally upon the independent variable values. That is, $P(F > f_c | \mathcal{X} = X) = \alpha$. Using the Law of Total Probability (see lecture slides), find the *unconditional* probability of a Type I error. Assume that the independent variables are discrete, so you can write a multiple sum.

4. Consider the following model with random independent variables. Independently for $i = 1, \ldots, n$,

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i \\ &= \alpha + \boldsymbol{\beta}' \mathbf{x}_i + \epsilon_i, \end{aligned}$$

where

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{pmatrix}$$

and $\mathbf{x}_i$ is independent of ϵ_i.

Note that in this notation, α is the intercept, and $\boldsymbol{\beta}$ does not include the intercept. The "independent" variables $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})'$ are not statistically independent. They have the symmetric and positive definite $k \times k$ covariance matrix $\Sigma_x = [\sigma_{ij}]$, which need not be diagonal. They also have the $k \times 1$ vector of expected values $\boldsymbol{\mu}_x = (\mu_1, \ldots, \mu_k)'$.

(a) Let Σ_{xy} denote the $k \times 1$ matrix of covariances between y_i and x_{ij} for $j = 1, \ldots, k$. Calculate $\Sigma_{xy} = cov(\mathbf{x}_i, y_i)$. Stay with matrix notation and don't expand.

(b) From the equation you just obtained, solve for $\boldsymbol{\beta}$ in terms of Σ_x and Σ_{xy}.

(c) Based on your answer to the last part and and letting $\widehat{\Sigma}_x$ and $\widehat{\Sigma}_{xy}$ denote matrices of *sample* variances and covariances, what would be a reasonable estimator of $\boldsymbol{\beta}$ that you could calculate from sample data? If you are not sure, check the lecture notes in which we centered y_i and well as the independent variables, and fit a regression through the origin.

5. In the following regression model, the independent variables x_1 and x_2 are random variables. The true model is

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i,$$

independently for $i = 1, \ldots, n$, where $\epsilon_i \sim N(0, \sigma^2)$.

The mean and covariance matrix of the independent variables are given by

$$E\begin{pmatrix} x_{i,1} \\ x_{i,2} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad cov\begin{pmatrix} x_{i,1} \\ x_{i,2} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix}$$

Unfortunately $x_{i,2}$, which has an impact on y_i and is correlated with $x_{i,1}$, is not part of the data set. Since $x_{i,2}$ is not observed, it is absorbed by the intercept and error term, as follows.

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i \\ &= (\beta_0 + \beta_2\mu_2) + \beta_1 x_{i,1} + (\beta_2 x_{i,2} - \beta_2\mu_2 + \epsilon_i) \\ &= \beta_0^* + \beta_1 x_{i,1} + \epsilon_i^*. \end{aligned}$$

It was necessary to add and subtract $\beta_2\mu_2$ in order to obtain $E(\epsilon_i^*) = 0$. And of course there could be more than one omitted variable. They would all get swallowed by the intercept and error term, the garbage bins of regression analysis.

(a) What is $Cov(x_{i,1}, \epsilon_i^*)$? This is a scalar calculation.

(b) Calculate $Cov(x_{i,1}, y_i)$; it's easier if you use the starred version of the model. This is another scalar calculation. Is it possible to have non-zero covariance between $x_{i,1}$ and y_i when $\beta_1 = 0$?

(c) Suppose we want to estimate β_1 using the usual least squares estimator b_1 (see formula sheet). As $n \to \infty$, does $b_1 \to \beta_1$? You may use the fact that like sample means, sample variances and covariances converge to the corresponding Greek-letter versions as $n \to \infty$ (except possibly on a set of probability zero) like ordinary limits, and all the usual rules of limits apply. So for example, defining $\widehat{\sigma}_{xy}$ as $\frac{1}{n-1}\sum_{i=1}^{n}(x_{i,1} - \overline{x}_1)(y_i - \overline{y})$, we have $\widehat{\sigma}_{xy} \to Cov(x_i, y_i)$.
