Stat 513
Emerson, Win 16
Homework #2
January 13, 2016

Written solutions to the homework problems are due on Wednesday, January 20, 2015 at the beginning of class.

The homework problems are divided into "regular" and "more involved" problems. In order to facilitate multiple graders, you should hand in these categories of problems separately. That is, hand in one paper that contains only the "regular" problems, and another paper that contains only the "more involved" problems.
As noted on the syllabus, copying of homework solutions is not allowed and, when detected, will be investigated as an infraction of the academy integrity policy of the University of Washington. While it is permissible to discuss problems with other students, TAs, or the instructor in order to learn how to solve a problem, your written solutions must be prepared without directly referencing any notes or solutions derived from other students or sources found on the internet.

REGULAR PROBLEMS

1. The χ^2, t, and F distributions are important "sampling distributions" commonly used in statistical inference. These distributions are derived as the exact distribution of certain statistics computed on normally distributed data. We are often ultimately interested the distribution-free interpretation of these statistics.

   (a) Rigorously show that as n becomes large, a normal distribution provides a good approximation to the χ^2_n distribution. Make clear the parameters of the normal distribution, as well as the sense in which the approximation is valid.

   (b) Rigorously show that as n becomes large, a normal distribution provides a good approximation to the t_n distribution. Make clear the parameters of the normal distribution, as well as the sense in which the approximation is valid.

   (c) Rigorously show that as n becomes large, a χ^2 distribution provides a good approximation to the $F_{m,n}$ distribution. Make clear the parameters of the χ^2 distribution, as well as the sense in which the approximation is valid.

2. Suppose n-vector $\vec{\epsilon}$ has $E[\vec{\epsilon}] = \vec{0}$ and $Cov[\vec{\epsilon}] = \mathbf{V}$ with $rank(\mathbf{V}) = n$. Let $\widehat{\vec{\beta}} = (\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\vec{Y}$ be the ordinary least squares estimator of $\vec{\beta}$ and $\widehat{\vec{\beta}}_G = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}^T\mathbf{V}^{-1}\vec{Y}$ be the generalized least squares estimator of $\vec{\beta}$ in regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$.

   (a) Find the mean and variance of estimators $\vec{a}^T\widehat{\vec{\beta}}$ and $\vec{a}^T\widehat{\vec{\beta}}_G$ of estimable function $\vec{a}^T\beta$.

(b) Show that a best linear unbiased estimator of estimable function $\vec{a}^T\vec{\beta}$ is $\vec{a}^T\widehat{\vec{\beta}}_G$.

3. Consider a "two sample" setting in which $Y_i \sim (\mu_0, \sigma^2)$ for $i = 1, \ldots, n_0$ and $Y_i \sim (\mu_1, \sigma^2)$ for $i = n_0 + 1, \ldots, n = n_0 + n_1 = 2n_0$, except observations within each group are correlated. That is, we have $Cov(Y_i, Y_j) = \rho\sigma^2$ for $i, j = 1, \ldots, n_0; i \neq j$, $Cov(Y_i, Y_j) = \rho\sigma^2$ for $i, j = n_0+1, \ldots, n; i \neq j$, and $Cov(Y_i, Y_j) = 0$ for $i = 1, \ldots, n_0; j = n_0+1, \ldots, n$. For notational convenience, let $\vec{w}$ be an n-vector such that $w_i = 1$ for $1 \leq i \leq n_0$ and $w_i = 0$ otherwise, and let $\vec{z} = \vec{1}_n - \vec{w}$. Consider linear regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ with $\mathbf{X} = (\vec{w} \quad \vec{z})$ and $\vec{\epsilon} \sim (\vec{0}, \mathbf{V})$. We are interested in estimating $\vec{a}^T\vec{\beta} = \mu_1 - \mu_0$.

(a) Show that in this "balanced design" setting in which $n_0 = n_1$, the ordinary least squares estimator $\widehat{\vec{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{Y}$ is equal to the generalized least squares estimator $\widehat{\vec{\beta}}_G = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\vec{Y}$. What are the mean and variance of these estimators?

(b) Provide an estimate of the variance of $\widehat{\vec{\beta}}_G$ and $\vec{a}^T\widehat{\vec{\beta}}_G$ assuming that ρ is known.

(c) Provide an estimate of the variance of $\widehat{\vec{\beta}}$ and $\vec{a}^T\widehat{\vec{\beta}}$ under the assumption that the observations are independent. How do they compare to the answers in b?

4. Now consider the setting in which $Y_i \sim (\mu_0, \sigma^2)$ for $i = 1, \ldots, n_0$ and $Y_i \sim (\mu_1, \sigma^2)$ for $i = n_0 + 1, \ldots, n = n_0 + n_1 = 2n_0$, except observations are paired across groups. That is, we have $Cov(Y_i, Y_i) = \sigma^2$ for $i = 1, \ldots, n$, $Cov(Y_i, Y_{n_0+i}) = \rho\sigma^2$ for $i = 1, \ldots, n_0$, and $Cov(Y_i, Y_j) = 0$ otherwise. For notational convenience, let $\vec{w}$ be an n-vector such that $w_i = 1$ for $1 \leq i \leq n_0$ and $w_i = 0$ otherwise, and let $\vec{z} = \vec{1}_n - \vec{w}$. Consider linear regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ with $\mathbf{X} = (\vec{w} \quad \vec{z})$ and $\vec{\epsilon} \sim (\vec{0}, \mathbf{V})$. We are interested in estimating $\vec{a}^T\vec{\beta} = \mu_1 - \mu_0$.

(a) Show that in this "balanced design" setting in which $n_0 = n_1$, the ordinary least squares estimator $\widehat{\vec{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{Y}$ is equal to the generalized least squares estimator $\widehat{\vec{\beta}}_G = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\vec{Y}$. What are the mean and variance of these estimators?

(b) Provide an estimate of the variance of $\widehat{\vec{\beta}}_G$ and $\vec{a}^T\widehat{\vec{\beta}}_G$ assuming that ρ is known.

(c) Provide an estimate of the variance of $\widehat{\vec{\beta}}$ and $\vec{a}^T\widehat{\vec{\beta}}$ under the assumption that the observations are independent. How do they compare to the answers in b?

(d) How does the effect of correlated observations affect an ordinary least squares analysis differ when the correlated observations are within groups sharing the same predictor values versus when the correlated observations have different predictor values?

MORE INVOLVED PROBLEMS

5. Consider linear regression models relating response $\vec{Y}$ to an intercept and up to two predictor vectors $\vec{W}$ and $\vec{Z}$ (so a full design matrix $\mathbf{X} = (\vec{1}_n \quad \vec{W} \quad \vec{Z})$ has $X_{i1} \equiv 1$ for

$i = 1, \ldots, n$ and $X_{i2} = W_i$ and $X_{i3} = Z_i$ and $\vec{\beta} = (\beta_0, \beta_1, \beta_2)^T$). Assume $E[\vec{\epsilon}] = \vec{0}$ and $var(\vec{\epsilon}) = \sigma^2 \mathbf{I}_n$. Our primary target of inference is the association between Y and W. We consider "adjusted" linear regression model in which

$$\vec{Y} = \beta_0 + \vec{W}\beta_1 + \vec{Z}\beta_2 + \vec{\epsilon}$$

and "unadjusted" model

$$\vec{Y} = \gamma_0 + \vec{W}\gamma_1 + \vec{\epsilon}^*$$

.

(a) Under what conditions is the OLS estimate $\hat{\beta}_1$ equal to the OLS estimate $\hat{\gamma}_1$?

(b) Under what conditions is the standard error of $\hat{\beta}_1$ equal to the standard error of $\hat{\gamma}_1$.

(c) Under what conditions is the estimated standard error of $\hat{\beta}_1$ equal to the estimated standard error of $\hat{\gamma}_1$.

(d) Under what conditions is $\hat{\gamma}_1$ unbiased for β_1?

(e) Under what conditions is $\hat{\gamma}_1$ BLUE for β_1?

(f) Suppose in particular that $\beta_1 = 0$ and $\beta_2 \neq 0$. What is the impact of this situation on the distribution of $\hat{\gamma}_1$, and how would $\hat{\gamma}_1$ compare to $\hat{\beta}_1$ from the full model? Compare this situation to the setting in which $\beta_2 = 0$ and $\beta_1 \neq 0$.