Biost/Stat 533
Emerson, Spr 99
Homework #4 Key
Jun 5, 1999

1. Suppose W_i is a categorical variable taking on one of the values $a_1, a_2, \ldots, a_p$. Consider a linear regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ in which $\vec{\beta}^T = (\beta_0, \ldots, \beta_{p-1})$ and $\vec{W}$ is modeled with dummy variables. That is, we consider a model where the first column of the design matrix is filled with 1's (so we are fitting an intercept), and the jth column of the design matrix is an indicator that $W_i = a_j$ for $j = 2, \ldots, p$ (so $X_{i1} = 1$, and for $j = 2, \ldots, p$, $X_{ij} = 1$ if $W_i = a_j$ and $X_{ij} = 0$ otherwise). Assume that $var(\vec{\epsilon}) = \sigma^2 \mathbf{I}_n$.

   a. Find expressions for $\widehat{\vec{\beta}}$ in terms of the group sample means $\overline{Y}_j$ where $\overline{Y}_j = \sum_{i=1}^n Y_i 1_{[W_i = a_j]} / \sum_{i=1}^n 1_{[W_i = a_j]}$ for $j = 1, \ldots, p$.

   Ans: Notationally, let $\vec{X}_{\cdot j} = (X_{1j} \ \ldots \ X_{nj})^T$ and $\vec{n} = (n_2 \ \ldots \ n_p)^T$, where $n_j = \sum_{i=1}^n 1_{[W_i = a_j]}$ counts the number of observations having $W_i = a_j$. Then $\mathbf{X} = (\vec{1}_n \ \vec{X}_{\cdot 2} \ \ldots \ \vec{X}_{\cdot p})$ and we note that $\vec{1}_n^T \vec{X}_{\cdot j} = n_j$, $\vec{X}_{\cdot j}^T \vec{X}_{\cdot j} = n_j$, and for $j \neq k$ $\vec{X}_{\cdot j}^T \vec{X}_{\cdot k} = 0$. Hence

$$\mathbf{X}^T\vec{Y} = \begin{pmatrix} n\overline{Y} \\ n_2\overline{Y}_2 \\ \vdots \\ n_p\overline{Y}_p \end{pmatrix}$$
$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \vec{n}^T \\ \vec{n} & diag(\vec{n}) \end{pmatrix}$$

   where $diag(\vec{n})$ is a diagonal matrix having $\vec{n}$ on the diagonal and zeroes elsewhere. Using the formula for the inverse of a partitioned matrix as given on page 6 of the key to homework #3 (where $\mathbf{A} = n$ $\mathbf{B} = \vec{n}^T$, and $\mathbf{D} = diag(\vec{n})$ so $\mathbf{D}^{-1} = diag((1/n_2 \ \ldots \ 1/n_p))$, $\mathbf{G} = n - \sum_{j=2}^p n_j = n_1$, and $\mathbf{J} = \vec{1}_{p-1}^T$), we therefore find that

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n_1} & -\frac{1}{n_1}\vec{1}_{p-1}^T \\ -\frac{1}{n_1}\vec{1}_{p-1} & diag((1/n_2 \ \ldots \ 1/n_p)) + \frac{1}{n_1}\vec{1}_{p-1}\vec{1}_{p-1}^T \end{pmatrix}$$

   We then obtain OLSE

$$\widehat{\vec{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{Y} = \begin{pmatrix} \overline{Y}_1 \\ \overline{Y}_2 - \overline{Y}_1 \\ \vdots \\ \overline{Y}_p - \overline{Y}_1 \end{pmatrix}$$

   b. Show that an asymptotic test of $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ is equivalent to a one-way analysis of variance to compare $H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$, where it is assumed that independent observations $Y_i \sim (\mu_j, \sigma^2)$ when $W_i = a_j$.

   Ans: We can write the null hypothesis as $H_0 : \mathbf{A}\vec{\beta} = \vec{0}$, where $\mathbf{A} = (\vec{0}_{p-1} \quad \mathbf{I}_{p-1})$. We then have

$$\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T = diag((1/n_2 \ \ldots \ 1/n_p)) + \frac{1}{n_1}\vec{1}_{p-1}\vec{1}_{p-1}^T$$

   and we can find the inverse to be (I did this by considering the case $p = 3$, inverting that simple 2 by 2 matrix, guessing the general form by induction, and then checking that the formula did indeed work)

$$(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1} = diag(\vec{n}) - \frac{1}{n}\vec{n}\vec{n}^T$$

The quadratic form is then

$$
\begin{aligned}
Q &= \frac{1}{\sigma^2}(\mathbf{A}\widehat{\vec{\beta}})^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\widehat{\vec{\beta}}) \\
&= \frac{1}{\sigma^2}\left[(\mathbf{A}\widehat{\vec{\beta}})^T diag(\vec{n})(\mathbf{A}\widehat{\vec{\beta}}) - (\mathbf{A}\widehat{\vec{\beta}})^T\frac{1}{n}\vec{n}\vec{n}^T(\mathbf{A}\widehat{\vec{\beta}})\right] \\
&= \frac{1}{\sigma^2}\left[\sum_{j=2}^{p} n_j(\overline{Y}_j - \overline{Y}_1)^2 - \frac{1}{n}\left(\sum_{j=2}^{p} n_j(\overline{Y}_j - \overline{Y}_1)\right)^2\right] \\
&= \frac{1}{\sigma^2}\left[\sum_{j=1}^{p} n_j(\overline{Y}_j - \overline{Y}_1)^2 - \frac{1}{n}\left(\sum_{j=1}^{p} n_j(\overline{Y}_j - \overline{Y}_1)\right)^2\right] \\
&= \frac{1}{\sigma^2}\left[\sum_{j=1}^{p} n_j\overline{Y}_j^2 - 2n\overline{Y}\overline{Y}_1 + n\overline{Y}_1^2 - \frac{1}{n}(n\overline{Y} - n\overline{Y}_1)^2\right] \\
&= \frac{1}{\sigma^2}\left[\sum_{j=1}^{p} n_j\overline{Y}_j^2 - 2n\overline{Y}\overline{Y}_1 + n\overline{Y}_1^2 - n\overline{Y}^2 + 2n\overline{Y}\overline{Y}_1 - n\overline{Y}_1^2\right] \\
&= \frac{1}{\sigma^2}\left[\sum_{j=1}^{p} n_j\overline{Y}_j^2 - n\overline{Y}^2\right]
\end{aligned}
$$

which, after the estimate for σ^2 is substituted, is the form of the traditional statistic for one-way ANOVA.

2. Let independent random vectors (X_i, Y_i) for $i = 1, \ldots, n$ be distributed according to a bivariate normal distribution with $X_i \sim (\mu, \sigma^2)$, $Y_i \sim (\nu, \tau^2)$, and $corr(X_i, Y_i) = \rho$. Let $\vec{X} = (X_1, \ldots, X_n)^T$ and $\vec{Y} = (Y_1, \ldots, Y_n)^T$.

   a. Derive the conditional distribution of $Y_i|X_i = x$ and $X_i|Y_i = y$.

   Ans: If

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} \mu \\ \nu \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix}\right)$$

   then the conditional density for Y given X can be found from

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

   which in this case leads to

$$Y|X = x \sim \mathcal{N}\left(\nu + \frac{\rho\tau}{\sigma}(x - \mu), \tau^2(1 - \rho^2)\right)$$

   b. Suppose we fit linear regression model $\vec{Y} = \beta_0 + \beta_1\vec{X} + \vec{\epsilon}$. Is asymptotic inference for OLSE of the regression parameters valid for this model? Justify your answer. For what function of parameters μ, ν, σ^2, τ^2, and ρ is OLSE $\widehat{\vec{\beta}}$ an unbiased estimator?

   Ans: We use the results of part (a) to find the conditional distribution of the Y_i's given the X_i's. Since the Y_i's conditional on the X_i's are independent with equal variance for all i, and since $E[Y_i|X_i] = \nu + \frac{\rho\tau}{\sigma}(X_i - \mu)$, the necessary assumptions for asymptotic inference based on OLSE are met (and in fact due to normality, even the assumptions necessary for small sample inference are met). In the regression model, $\beta_0 = \nu - \frac{\rho\tau}{\sigma}\mu$, and $\beta_1 = \frac{\rho\tau}{\sigma}$. The OLSE $\hat{\beta}_0$ and $\hat{\beta}_1$ are therefore consistent for β_0 and β_1 as given above, respectively.

c. Suppose we fit linear regression model $\vec{X} = \gamma_0 + \gamma_1 \vec{Y} + \vec{\delta}$. Is asymptotic inference for OLSE of the regression parameters valid for this model? Justify your answer. For what function of parameters μ, ν, σ^2, τ^2, and ρ is OLSE $\widehat{\vec{\beta}}$ an unbiased estimator?

Ans: We use the results of part (a) to find the conditional distribution of the X_i's given the Y_i's. Since the X_i's conditional on the Y_i's are independent with equal variance for all i, and since $E[X_i|Y_i] = \mu + \frac{\rho\sigma}{\tau}(Y_i - \nu)$, the necessary assumptions for asymptotic inference based on OLSE are met (and in fact due to normality, even the assumptions necessary for small sample inference are met). In the regression model, $\gamma_0 = \mu - \frac{\rho\sigma}{\tau}\nu$, and $\gamma_1 = \frac{\rho\sigma}{\tau}$. The OLSE $\hat{\beta}_0$ and $\hat{\beta}_1$ are therefore consistent for γ_0 and γ_1 as given above, respectively.

d. Under what conditions will $y = \hat{\beta}_0 + \hat{\beta}_1 x$ and $x = \hat{\gamma}_0 + \hat{\gamma}_1 y$ be the same line?

Ans: Rewriting the second linear equation to solve for y, we have $y = -\hat{\gamma}_0/\hat{\gamma}_1 + x/\hat{\gamma}_1$. Thus for the two lines to be coincident, we must have that $\hat{\beta}_0 = -\hat{\gamma}_0/\hat{\gamma}_1$ and $\hat{\beta}_1 = 1/\hat{\gamma}_1$. Now in simple linear regression, $\hat{\beta}_1 = S_{XY}/S_{XX}$ and $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$. We would also therefore have $\hat{\gamma}_1 = S_{XY}/S_{YY}$ and $\hat{\gamma}_0 = \overline{X} - \hat{\gamma}_1 \overline{Y}$. I note that if $\hat{\beta}_1 = 1/\hat{\gamma}_1$, we necessarily have $\hat{\beta}_0 = -\hat{\gamma}_0/\hat{\gamma}_1$. In order for $\hat{\beta}_1 = 1/\hat{\gamma}_1$, we must have $S_{XY}^2/(S_{XX}S_{YY}) = 1$ which in turn implies that the sample correlation r_{XY} is either 1 or -1. It should be noted that this result carries over from the sample space to the parameter space. That is, the lines being estimated by the OLSE in a consistent manner are coincident only if $\rho = 1$ or $\rho = -1$ (in which case we would also have that $r_{XY} = 1$ or $r_{XY} = -1$, respectively, in every sample).

3. Consider an "error in the variables" model in which there is a true relationship between response Y and predictor W given by $Y = \beta_0 + \beta_1 W + \epsilon$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ totally independent. Suppose that W is unobserved, and we instead have Z, an imprecise measurement of W which follows the relation $Z = \alpha_0 + \alpha_1 W + \delta$, with $\delta_i \sim \mathcal{N}(0, \tau^2)$ totally independent of each other and the ϵ's. We further assume that W_i, δ_i, and ϵ_i are jointly normally distributed and totally independent. We then fit a regression model $E[Y] = \gamma_0 + \gamma_1 Z$, and use this model to make inference about an association between Y and W.

a. Under what conditions is OLSE $\hat{\gamma}_1$ unbiased for β_1?

Ans: We have that

$$\vec{U}_i = \begin{pmatrix} W_i \\ \epsilon_i \\ \delta_i \end{pmatrix} \sim \mathcal{N}_3 \left(\begin{pmatrix} \mu \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} v^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & \tau^2 \end{pmatrix} \right)$$

and

$$\begin{pmatrix} Y_i \\ Z_i \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \alpha_0 \end{pmatrix} + \begin{pmatrix} \beta_1 & 1 & 0 \\ \alpha_1 & 0 & 1 \end{pmatrix} \vec{U}_i$$

yielding

$$\begin{pmatrix} Y \\ Z \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \beta_0 + \beta_1 \mu \\ \alpha_0 + \alpha_1 \mu \end{pmatrix}, \begin{pmatrix} \beta_1^2 v^2 + \sigma^2 & \alpha_1 \beta_1 v^2 \\ \alpha_1 \beta_1 v^2 & \alpha_1^2 v^2 + \tau^2 \end{pmatrix} \right)$$

We thus obtain conditional distribution

$$E[Y_i|Z_i] = \beta_0 + \beta_1 \mu + \frac{\alpha_1 \beta_1 v^2}{\alpha_1^2 v^2 + \tau^2}(Z_i - \alpha_0 - \alpha_1 \mu)$$

$$var(Y_i|Z_i) = \frac{\alpha_1^2 v^2 \sigma^2 + \beta_1^2 v^2 \tau^2 + \sigma^2 \tau^2}{\alpha_1^2 v^2 + \tau^2}$$

From this, we see that $\hat{\gamma}_1$ is an unbiased estimate of

$$E[\hat{\gamma}_1] = \frac{\alpha_1 \beta_1 v^2}{\alpha_1^2 v^2 + \tau^2}$$

which is equal to β_1 if $\alpha_1 v^2 = \alpha_1^2 v^2 + \tau^2$. This latter condition is satisfied when $\alpha_1 = 1$ and $\tau^2 = 0$, among other less interesting possibilities. It should be noted that $E[\hat{\gamma}_1]$ will be of the same sign as β_1 so long as $\alpha_1 > 0$. Furthermore, in the most interesting case in which $\alpha_1 = 1$ approximately (so our surrogate predictor variable is approximately the same scale as the true predictor), any measurement error will tend to attenuate the slope estimate by bringing it closer to 0.

More generally, we can consider $\mathbf{W} = (\vec{1}_n \quad \vec{W})$ and $\mathbf{Z} = (\vec{1}_n \quad \vec{Z})$. Then

$$\begin{aligned} E[\widehat{\vec{\gamma}}] &= (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T E[\vec{Y}] \\ &= (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{W}\vec{\beta} \end{aligned}$$

Thus $\hat{\gamma}_1$ is unbiased for $\hat{\beta}_1$ precisely when

$$(0 \quad 1)(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{W} = (0 \quad 1).$$

By straightforward manipulation of these 2 by 2 matrices, we find this condition reduces to $S_{ZZ} = S_{WZ}$. Now

$$\begin{aligned} \frac{1}{n}S_{WZ} &= \frac{1}{n}\sum_{i=1}^{n} Z_iW_i - \overline{ZW} \\ &= \frac{1}{n}\sum_{i=1}^{n}(\alpha_0 + \alpha_1 W_i + \delta_i)W_i - (\alpha_0 + \alpha_1\overline{W} + \overline{\delta})\overline{W} \\ S_{WZ} &= \alpha_1 S_{WW} + S_{W\delta} \\ \frac{1}{n}S_{ZZ} &= \frac{1}{n}\sum_{i=1}^{n} Z_i^2 - (\overline{Z})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\alpha_0 + \alpha_1 W_i + \delta_i)^2 - (\alpha_0 + \alpha_1\overline{W} + \overline{\delta})^2 \\ S_{ZZ} &= \alpha_1^2 S_{WW} + S_{\delta\delta} + 2\alpha_1 S_{W\delta} \end{aligned}$$

Hence, the conditions to guarantee that $\hat{\gamma}_2$ to be unbiased for β_2 are again that $\alpha_2 = 1$ and $\tau^2 = 0$.

b. How does the standard error of $\hat{\gamma}_1$ compare to the standard error of $\hat{\beta}_1$ (if we had W)? What does this suggest about our ability to test for associations in such a model? How much do we lose by having errors in the predictors?

Ans: Using the results given above under the assumption of normally distributed W's, we would find that

$$\begin{aligned} var(\widehat{\vec{\gamma}}) &= (\mathbf{Z}^T\mathbf{Z})^{-1}\frac{\alpha_1^2 v^2\sigma^2 + \beta_1^2 v^2\tau^2 + \sigma^2\tau^2}{\alpha_1^2 v^2 + \tau^2} \\ var(\widehat{\vec{\beta}}) &= (\mathbf{W}^T\mathbf{W})^{-1}\sigma^2 \end{aligned}$$

These then give

$$\begin{aligned} var(\hat{\gamma}_1) &= \frac{1}{S_{ZZ}}\frac{\alpha_1^2 v^2\sigma^2 + \beta_1^2 v^2\tau^2 + \sigma^2\tau^2}{\alpha_1^2 v^2 + \tau^2} \\ var(\hat{\beta}_1) &= \frac{1}{S_{WW}}\sigma^2 \end{aligned}$$

From the results of (a), it can be seen that S_{WW} can be larger or smaller than S_{ZZ}. In the most interesting case where α_1 is approximately 1 (so our surrogate predictor variable is approximately the same scale as the true predictor) and the sample correlation between W_is and δ_is are approximately zero, then $S_{ZZ} \doteq S_{WW} + n\tau^2$. To consider the loss of power associated with errors in the predictors in this setting, we can consider the ratios

$$\frac{\beta_1^2}{Var(\hat{\beta}_1)} \doteq \frac{nv^2\beta_1^2}{\sigma^2}$$
$$\frac{(E[\hat{\gamma}_1])^2}{Var(\hat{\gamma}_1)} \doteq \frac{nv^2\beta_1^2}{\sigma^2} \frac{v^2\sigma^2}{v^2\sigma^2 + \beta_1^2 v^2\tau^2 + \sigma^2\tau^2}$$

Because the second such ratio is smaller than the first, there will be a decrease in the statistical power to detect nonzero β_1 when using Z_i instead of W_i.

When the bias and variability are considered jointly in this manner, it should be clear that there can be marked attenuation of the association when using predictors measured with error. To the extent that estimates of α_2, τ^2, and v^2 can be obtained, better estimates of β_2 can be derived from $\hat{\gamma}_2$.