

## Statistics 600 Problem Set 2

Due in class on Tuesday, October 27th

1. “Total least squares” (TLS) aims to identify a hyperplane  $\mathcal{P}$  that minimizes

$$\sum_i d(V_i, \mathcal{P})^2,$$

where the  $V_i$  are a collection of points in  $\mathcal{R}^p$ , and  $d(Q, \mathcal{P})$  is the minimum distance in  $\mathcal{R}^p$  between the point  $Q \in \mathcal{R}^p$  and any point on the hyperplane  $\mathcal{P} \subset \mathcal{R}^p$ . By writing  $V = (y, x_1, \dots, x_p)$ , TLS gives an alternative approach to fitting linear functions to point sets that does not treat the response variable differently from any of the predictors.

- (a) Parametrize  $\mathcal{P}$  in the form  $\{Z \in \mathcal{R}^p | B'(Z - W) = 0\}$ , for vectors  $B \in \mathcal{R}^p$  and  $W \in \mathcal{R}^p$  with  $\|B\| = 1$ . Write down an explicit expression for  $d(Q, \mathcal{P})$  in terms of  $B$  and  $W$ . Note that  $W$  can be any point in  $\mathcal{P}$  and is therefore not uniquely identified; also  $B$  is not identified up to a factor of  $\pm 1$ .
- (b) Based on your expression in part (a), show that the TLS solution passes through the center of the data  $\bar{V}$ , and use this to define a minimizing value for  $W$ .
- (c) Building on (a) and (b), construct a quadratic form whose minimizing value subject to  $\|B\| = 1$  solves the TLS problem for  $B$ .

2. If  $x \in \mathcal{R}^p$  is a column vector, construct a column vector  $y \in \mathcal{R}^p$  such that

$$(I + xx')^{-1} = I - yy'.$$

3. This exercise aims to illustrate the effect of outliers in least squares fitting. Suppose we observe data that follows a linear model with  $p = 1$  covariate:  $Y = \alpha + \beta X + \epsilon$ . Specifically, consider a triangular array of data  $Y_{in}, X_{in}$ , where  $i = 1, \dots, n$ . There is also a random indicator  $\delta_{in}$ , that we do not observe, such that  $\text{var}(\epsilon_{in} | X, \delta_{in} = 1) = k_n \sigma^2$ , and  $\text{var}(\epsilon_{in} | X, \delta_{in} = 0) = \sigma^2$  (the errors are centered, so that  $E(\epsilon | X, \delta) \equiv 0$ ). Suppose  $X$  is sampled from a population with variance  $\sigma_X^2$ , and  $P(\delta_{in} = 1) = p_n$ . Note that  $n \cdot \text{var}(\hat{\beta})$  has a finite limit when  $k_n \equiv 1$ . Derive conditions on  $k_n$  and  $p_n$  such that (i)  $n \cdot \text{var}(\hat{\beta})$  has a finite limit, and (ii)  $n \cdot \text{var}(\hat{\beta})$  has the same limit that would occur if  $k_n \equiv 1$ .
4. Suppose that  $V_1, \dots, V_n$  are independent random variables with mean zero and variance one. Calculate the variance of  $\sum_i V_i^2$  when the  $V_i$  follow standardized versions of the following distributions: (i) uniform, (ii) Laplace, (iii) Student's  $t$ , (iv) logistic. Feel free to get the moment formulas for these distributions from Wikipedia rather than deriving them all from scratch.

5. Suppose we have a bivariate regression in which  $\bar{X}_1 = \bar{X}_2 = 0$ ,  $\widehat{\text{var}}(X_1) = \widehat{\text{var}}(X_2) = 1$ , and  $\widehat{\text{cor}}(X_1, X_2) = r$ . We are interested in estimating a contrast  $\theta'\beta$ , where  $\theta = (0, \theta_1, \theta_2)'$  is a unit vector in  $R^3$ . Which values of  $\theta$  give the least and greatest values of  $\text{var}(\hat{\beta}'\theta)$ ? Note whether the answer to this question depends on  $\beta$ .
6. Suppose we have a bivariate regression where  $X_1$  and  $X_2$  are standardized. We fit a linear model  $\hat{Y} = \hat{\alpha}_s + \hat{\beta}_{1s}X_1 + \hat{\beta}_{2s}X_2$  sequentially, by first letting  $\hat{\alpha}_s$  and  $\hat{\beta}_{1s}$  be the intercept and slope from simple linear regression of  $Y$  on  $X_1$ . We then set  $\hat{\beta}_{2s}$  equal to the value of  $\beta_2$  that minimizes the least squares loss function when  $\alpha$  and  $\beta_1$  are fixed at  $\hat{\alpha}_s$  and  $\hat{\beta}_{1s}$ . Under what conditions is  $\hat{\beta}_{2s}$  greater (in signed value) than the multiple least squares estimate  $\hat{\beta}_2$ ?
7. Suppose we are interested in a contrast  $\theta'\beta$ , for a fixed vector of coefficients  $\theta \in \mathcal{R}^{p+1}$ . Let  $Z$  denote the Z-score for  $\theta'\beta$ , i.e. the point estimate of  $\theta'\beta$  divided by its standard error. Now suppose we reparametrize the problem by using a new design matrix  $\tilde{X} = XB$ , where  $B$  is an invertible matrix. (i) What vector  $\tilde{\theta}$  estimates the same contrast as  $\theta$  under the new parameterization? (ii) Are the Z-scores always the same in the two parameterizations?
8. Suppose we have a design matrix  $X$  and are interested in the partial  $R^2$  for  $X_k$  given  $X_1, \dots, X_{k-1}$ . Let  $B$  be an invertible matrix such that  $e'_k B = e_k$  (where  $e_k$  is the indicator vector of the  $k^{\text{th}}$  position in the vector). Let  $\tilde{X} = XB$ . What conditions on  $X$ , if any, are required so that the partial  $R^2$  of  $\tilde{X}_k$  given  $\tilde{X}_1, \dots, \tilde{X}_{k-1}$  is equal to the partial  $R^2$  of  $X_k$  given  $X_1, \dots, X_{k-1}$ ?
9. Suppose we estimate a multiple regression model as follows. First, we regress  $Y$  on each covariate separately using simple linear regression. This yields coefficient estimates  $\hat{\beta}_{1u}, \dots, \hat{\beta}_{pu}$ . For simplicity, suppose that  $Y$  and all the covariates have been standardized, and all models are fit without an intercept. As a second step, let  $\beta^* = (\hat{\beta}_{1u}, \dots, \hat{\beta}_{pu})'$ , and use least squares to estimate the scalar  $\lambda$  minimizing  $\sum_i (Y_i - \lambda\beta^{*'}X_i)^2$ . For the bivariate case, compare the fit achieved by this approach to the fit achieved using multiple linear regression. You can approach this using either numerical calculation, or simulation. Produce a few graphs showing how the relative difference in fits relates to the model parameters.