Problem Set 2

Due Sep 26, before class

1. BLUE estimator for the mean

Assume that y_i has mean μ and variance σ^2, and y_i $(i = 1, \ldots, n)$ are uncorrelated. A linear estimator of the mean μ has the form $\hat{\mu} = \sum_{i=1}^n a_i y_i$, which is unbiased as long as $\sum_{i=1}^n a_i = 1$. Find the BLUE for μ and prove why it is BLUE.

2. Univariate OLS and optimal design

Assume the Gauss-Markov model $y_i = \alpha + \beta x_i + \epsilon_i$ $(i = 1, \ldots, n)$ with a scalar x_i. Show that the variance of the OLS coefficient $\hat{\beta}$ equals

$$\text{var}(\hat{\beta}) \;=\; \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2.$$

Assume x_i must be in the interval $[0, 1]$ and n is an even number. We would like to choose their values to minimize $\text{var}(\hat{\beta})$. Find the minimizer x_i's.

Hint: You may find the following result useful. For $x_i \in [0, 1]$

$$\sum_{i=1}^n (x_i - \bar{x})^2 \;=\; \sum_{i=1} x_i^2 - n\bar{x}^2 \;\le\; \sum_{i=1} x_i - n\bar{x}^2 \;=\; n\bar{x}(1 - \bar{x}) \le n/4.$$

Think about when the equality holds.

3. Consequence of useless regressors

Partition the covariate matrix and parameter into

$$X \;=\; (X_1, X_2), \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},$$

where $X_1 \in \mathbb{R}^{n\times k}$, $X_2 \in \mathbb{R}^{n\times l}$, $\beta_1 \in \mathbb{R}^k$, and $\beta_1 \in \mathbb{R}^l$ with $k + l = p$. Assume the Gauss-Markov model with $\beta_2 = 0$. Let $\hat{\beta}$ be the first k coordinates of $\hat{\beta} = (X^T X)^{-1} X^T Y$ and $\tilde{\beta} = (X_1^T X_1)^{-1} X_1^T Y$ be the coefficient based on the OLS fit of Y on X_1 only. Show that

$$\mathrm{cov}(\hat{\beta}_1) \succeq \mathrm{cov}(\tilde{\beta}_1).$$

Hint: Try to use Gauss-Markov Theorem.

4. MLE under the Normal linear model

Under the Normal linear model, show that the maximum likelihood estimator (MLE) for β is the OLS estimator, but the MLE for σ^2 is $\tilde{\sigma}^2 = \mathrm{RSS}/n$, where $\mathrm{RSS} = \sum_{i=1} \hat{\epsilon}_i^2$. Compare the mean squared errors (MSEs) of $\hat{\sigma}^2 = \mathrm{RSS}/(n-p)$ and $\tilde{\sigma}^2$ for estimating σ^2. Note that the definition of the two MSEs are: $\mathbb{E}(\hat{\sigma}^2 - \sigma^2)^2$ and $\mathbb{E}(\tilde{\sigma}^2 - \sigma^2)^2$.

Hint: You can use the fact that a Chi-squared distribution with k degrees of freedom has mean k and variance $2k$.

5. Two-sample problem under the Normal linear model

Assume that $z_1, \ldots, z_m \overset{\text{IID}}{\sim} \mathrm{N}(\mu_1, \sigma^2)$ and $w_1, \ldots, w_n \overset{\text{IID}}{\sim} \mathrm{N}(\mu_2, \sigma^2)$. We would like to test $H_0 : \mu_1 = \mu_2$.

(a) Show that under H_0, the following t statistic with pooled variance estimator follows a t distri-

bution:

$$t_{\text{equal}} \;=\; \frac{\bar{z}-\bar{w}}{\sqrt{\hat{\sigma}^2(m^{-1}+n^{-1})}} \sim t_{m+n-2},$$

where

$$\hat{\sigma}^2 \;=\; \frac{(m-1)S_z^2+(n-1)S_w^2}{m+n-2}$$

with the sample means

$$\bar{z} \;=\; m^{-1}\sum_{i=1}^{m} z_i, \quad \bar{w} \;=\; m^{-1}\sum_{i=1}^{n} w_i,$$

and the sample variances

$$S_z^2 \;=\; (m-1)^{-1}\sum_{i=1}^{m}(z_i-\bar{z})^2, \quad S_w^2 \;=\; (n-1)^{-1}\sum_{i=1}^{n}(w_i-\bar{w})^2.$$

Remark: The name "equal" is motivated by the "var.equal" parameter of the R function *t.test*.

(b) We can write the above problem as testing hypothesis $H_0 : \beta_1 = 0$ in the linear regression $Y = X\beta + \epsilon$ with

$$Y \;=\; \begin{pmatrix} z_1 \\ \vdots \\ z_m \\ w_1 \\ \vdots \\ w_n \end{pmatrix}, \quad X \;=\; \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix}, \quad \beta \;=\; \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \epsilon \;=\; \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \\ \epsilon_{m+1} \\ \vdots \\ \epsilon_{m+n} \end{pmatrix}.$$

Based on the Normal linear model, we can compute the t statistic. Show that it is identical to t_{equal}.

6. Two-sample problem continued

Assume that $z_1, \ldots, z_m$ are IID with mean μ_1 and variance σ_1^2 and $w_1, \ldots, w_n$ are IID with mean μ_2 and variance σ_2^2 . We would like to test $H_0 : \mu_1 = \mu_2$. The following t statistic has a Normal distribution asymptotically:

$$t_{\text{unequal}} = \frac{\bar{z} - \bar{w}}{\sqrt{S_z^2/m + S_w^2/n}} \to \text{N}(0, 1),$$

We can write the above problem as testing hypothesis $H_0 : \beta_1 = 0$ in the heteroskedastic linear regression. Based on the EHW standard error, we can compute a test statistic. Show that it is identical to t_{unequal} under HC2 correction. (For the HC2 correction, h_{ii} is the i-th diagonal element of the hat matrix H.)

Remark: The name "unequal" is motivated by the "var.equal" parameter of the R function *t.test*.

7. Real data analysis

The R package *sampleSelection* describes the dataset *RandHIE* as follows: "The RAND Health Insurance Experiment was a comprehensive study of health care cost, utilization and outcome in the United States. It is the only randomized study of health insurance, and the only study which can give definitive evidence as to the causal effects of different health insurance plans." You can find more detailed information about other variables in this package. The main outcome of interest *lnmeddol* means the log of medical expenses. Use linear regression to investigate the relationship between the outcome and various important covariates.

Note that the solution of this problem is not unique, but you need to justify your choice of covariates and model, and interpret the results.