Discussion section 6

Math 181B

1 Review

1.1 Estimation inference

Let $(x_1, Y_1), (x_2, Y_2), \ldots$, and (x_n, Y_n) be a set of points satisfying the simple linear model, $E(Y \mid x) = \beta_0 + \beta_1 x$. And $Y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right), Y_i$ independent. The maximum likelihood estimators for β_0, β_1, and σ^2 are given by

- $$\hat{\beta}_1 = \frac{n \sum_{i=1}^n x_i Y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n Y_i\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2}$$
- $$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$
- $$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \hat{Y}_i\right)^2$$

  where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \ldots, n$.

Moreover, we have

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are both normally distributed.
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are both unbiased: $E\left(\hat{\beta}_0\right) = \beta_0$ and $E\left(\hat{\beta}_1\right) = \beta_1$.
- $\operatorname{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$.
- $\operatorname{Var}\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]$.
- $\hat{\beta}_1$,$\bar{Y}$ and $\hat{\sigma}^2$ are mutually independent.
- $\frac{n\hat{\sigma}^2}{\sigma^2}$ has a chi square distribution with $n-2$ degrees of freedom.
- $\frac{n}{n-2} \cdot \hat{\sigma}^2$ is an unbiased estimator for σ^2.

1.2 Inference for mean response

Given a new observation x, we are interested in the mean response of Y corresponding with this specific x, i.e., we want to figure out the distribution of $E(Y|x)$. We can estimate this quantity by $E(\hat{Y}|x)$. Moreover, we can build a confidence interval to cover $E(Y|x)$, i.e.,

Theorem 1.1 *Let $(x_1, Y_1), (x_2, Y_2), \ldots,$ and (x_n, Y_n) be a set of points that satisfy the assumptions to perform a simple linear regression. A $100(1-\alpha)\%$ confidence interval for $E(Y \mid x) = \beta_0 + \beta_1 x$ is given by $(\hat{y} - w, \hat{y} + w)$, where*

$$w = t_{\alpha/2,n-2} \cdot s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

and $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$; $s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$.

The proof idea is based on the normal distribution of $\hat{Y}$.

1.3 Inference for new response

Rather than building a confidence interval for the mean response, we can also make a confidence interval to cover the single future response, i.e.,

Theorem 1.2 *Let $(x_1, Y_1), (x_2, Y_2), \ldots,$ and (x_n, Y_n) be a set of n points that satisfy the assumptions of the simple linear model. A $100(1-\alpha)\%$ prediction interval for Y at the fixed value x is given by $(\hat{y} - w, \hat{y} + w)$, where*

$$w = t_{\alpha/2,n-2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

and $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$; $s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$.

The proof idea is based on the normal distribution of $\hat{Y} - Y$. Note that $\hat{Y}$ and Y are independent.

1.4 Anova test

Our goal is to test $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ and H_1 : At least one μ_i is different. Also, we assume that Y_{ij} are independent and normally distributed with mean $\mu_j, j = 1, 2, \ldots, k$, and variance σ^2(constant for all j). Suppose, for each level, we have independent samples with size $n_1, \ldots, n_k$. Some useful statistics are listed below:

- Treatment sum of squares:

$$\mathrm{SSTR} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\bar{Y}_{.j} - \bar{Y}_{..}\right)^2 = \sum_{j=1}^{k} n_j\left(\bar{Y}_{.j} - \bar{Y}_{..}\right)^2 = \sum_{j=1}^{k} n_j\left(\bar{Y}_{.j} - \mu\right)^2 - n\left(\bar{Y}_{..} - \mu\right)^2.$$

- Error sum of squares:

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1) S_j^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(Y_{ij} - \bar{Y}_{.j}\right)^2 .$$

- Total sum of squares:

$$\text{SSTOT} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(Y_{ij} - \bar{Y}_{..}\right)^2 .$$

Remark:

- $\text{SSTOT} = \text{SSTR} + \text{SSE}$.
- $E(\text{SSTR}) = (k-1)\sigma^2 + \sum_{j=1}^{k} n_j \left(\mu_j - \mu\right)^2$.
- Under $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ is true, SSTR $/\sigma^2 \sim \chi^2_{k-1}$.
- Whether or not $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ is true, $\text{SSE}/\sigma^2 \sim \chi^2_{n-k}$.
- Whether or not $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ is true, SSE and SSTR are independent.
- Under $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ is true, $\frac{\text{SSTOT}}{\sigma^2} \sim \chi^2_{n-1}$.