Discussion section 5

Math 181B

1 Review

1.1 Intuition of the regression curve

Recall that: Given n points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we can use a straight line $y = a + bx$ to depict which minimizes

$$L = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 .$$

However, in practice, it is more appropriate to think of each y value as a random variable Y, which stands for a distribution of possible y-values conditional on every value of x.

Definition 1.1 (Regression curve) *Let $f_{Y|x}(y)$ denote the pdf of the random variable Y for a given value x, and let $E(Y \mid x)$ denote the expected value associated with $f_{Y|x}(y)$. The function*

$$y = E(Y \mid x)$$

is called the regression curve of Y on x.

1.2 Linear model: A special case of regression curve

In this class, we consider a special case of regression curve:

- $f_{Y|x}(y)$ is a normal pdf for all x.
- The standard deviation, σ, associated with $f_{Y|x}(y)$ is the same for all x.
- The means of all the conditional Y distributions are collinear-that is,

$$y = E(Y \mid x) = \beta_0 + \beta_1 x$$

- All of the conditional distributions represent independent random variables.

1.3 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, $E(Y \mid x) = \beta_0 + \beta_1 x$. 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 $E\left(\hat{\beta}_1\right) = \beta_1$.
- $\text{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.
- $\text{Var}\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$.
- $\hat{\beta}_1$,$\bar{Y}$ 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.
- $$T_{n-2} = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$$

  has a Student t distribution with $n-2$ degrees of freedom; where $S^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$. Based on this, we can do a hypothesis test and make a confidence interval for β_1.