Lecture 8b.

Last class:

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}} \sim t_{n-p-1} \quad \text{(e.g. if } p=1 \Rightarrow \text{inference in SLR).}$$

1) $100(1-\alpha)\%$ CI for β_j: $\hat{\beta}_j \pm c\,SE(\hat{\beta}_j)$, $c = t_{1-\frac{\alpha}{2}, n-p-1}$ ($1-\frac{\alpha}{2}$ quantile of t_{n-p-1} dist^n)
2) Test $H_0: \beta_j = 0$ vs. $H_a: \beta_j \neq 0$. If $|t| > c$ then reject H_0

Estimating mean response for $\vec{x}_c = (1, x_{c1}, \dots, x_{cp})^T$: $\hat{\mu}_c = \vec{x}_c^T \hat{\vec{\beta}}$

- As a R.V. $\hat{\mu}_c \sim N(\vec{x}_c^T \vec{\beta}, \sigma^2 x_c^T (X^TX)^{-1} \vec{x}_c)$
- $100(1-\alpha)\%$ CI: $\hat{\mu}_c \pm c\hat{\sigma}\sqrt{\vec{x}_c^T (X^TX)^{-1} \vec{x}_c}$

Predicting response for $\vec{x}_0 = (1, x_{01}, \dots, x_{0p})^T$; $\hat{y}_0 = \vec{x}_0^T \hat{\vec{\beta}}$

- As a R.V. $Y_0 - \hat{Y}_0 \sim N(0, \sigma^2(1 + \vec{x}_0^T (X^TX)^{-1} \vec{x}_0))$
- $100(1-\alpha)\%$ PI: $\hat{y}_0 \pm c\hat{\sigma}\sqrt{1 + \vec{x}_0^T (X^TX)^{-1} \vec{x}_0}$.

Handling Categorical Explanatory Variables

In linear regression, the explanatory variables can be categorical. A categorical variable can take values that fall into several categories. E.g.,

- binary variable: gender (male, female) $\rightarrow$ 0/1 (0 if female, 1 if male).
- ordered variable: mild, medium, severe.
- unordered variable: red, blue, green.

e.g. $E(Y) = \beta_0 + \beta_1 x$, x is binary variable (1 or 0).

- β_0: mean response when $x=0$
- $\beta_0 + \beta_1$: mean response when $x=1$
- β_1: difference in mean response between the two categories.

Approaches for more than two categories:

(1) Convert into indicator/dummy variables

(2) Treat them as numerical (only when it makes sense to do so).

e.g. Hospital data (Applied Linear Models by Kutner et al).

Response: Infection risk (measure of risk of acquiring an infection in a hospital).

Explanatory var.: Stay (average length of hospital stay in days).

Region (Geographic Region: 1 = NE, 2 = NC, 3 = S, 4 = W).

Infct Risk	Stay ($\vec{x}_1$)	Region ($\vec{x}_2$)
4.1	7.13	4
1.6	8.82	2
2.7	8.34	3
5.6	8.95	4
5.7	11.2	1
⋮	⋮	⋮

How should we code "Region"?

Currently, $x_{i2} = \begin{cases} 1, & \text{if NE} \\ 2, & \text{if NC} \\ 3, & \text{if S} \\ 4, & \text{if W.} \end{cases}$ ⇒ not appropriate (unless there is a linear relationship between the response and this particular ordering).

More flexible approach is to use indicator/dummy variables:

$$x_{i2} = \begin{cases} 1 & \text{if NC} \\ 0 & \text{o/w} \end{cases} \quad ; \quad x_{i3} = \begin{cases} 1 & \text{if S} \\ 0 & \text{o/w} \end{cases} \quad ; \quad x_{i4} = \begin{cases} 1 & \text{if W} \\ 0 & \text{o/w.} \end{cases}$$

Why not define x_{i5} for NE?

$$X = \begin{array}{c} \\ \begin{bmatrix} 1 & 4.1 & 0 & 0 & 1 & 0 \\ 1 & 1.6 & 1 & 0 & 0 & 0 \\ 1 & 2.7 & 0 & 1 & 0 & 0 \\ 1 & 5.6 & 0 & 0 & 1 & 0 \\ 1 & 5.7 & 0 & 0 & 0 & 1 \end{bmatrix} \\ \begin{matrix} \vec{1} & \vec{x}_1 & \vec{x}_2 & \vec{x}_3 & \vec{x}_4 & \cancel{\vec{x}_5} \end{matrix} \end{array}$$

$\vec{1} = \vec{x}_2 + \vec{x}_3 + \vec{x}_4 + \vec{x}_5$

→ rank$(X) = 5 < p+1 = 6$

→ linear dependent columns

∴ exclude $\vec{x}_5$.

$$\text{Model}: \quad Y_i = \beta_0 + \beta_1 \underbrace{x_{i1}}_{\text{stay}} + \beta_2 \overset{NC}{x_{i2}} + \beta_3 \overset{S}{x_{i3}} + \beta_4 \overset{W}{x_{i4}} + \varepsilon_i$$

x_{i2}, x_{i3}, x_{i4}: region

Interpretation

- Average risk of acquiring an infection if average hospital stay was x_{i1} and the geographic region was NE: $\beta_0 + \beta_1 x_{i1}$
- " " was NC: $\beta_0 + \beta_1 x_{i1} + \beta_2$
- " " was S: $\beta_0 + \beta_1 x_{i1} + \beta_3$
- " " was W: $\beta_0 + \beta_1 x_{i1} + \beta_4$
- β_2: difference between NC and NE in avg. infection risk, holding avg. hospital stay constant.
- β_3: difference between S and NE " "
- β_4: difference between W and NE " "
- $\beta_2 - \beta_3$: difference between NC and S " "
- $\beta_2 - \beta_4$: difference between NC and W " "
- $\beta_3 - \beta_4$: difference between S and W " "

We know from before $\hat{\vec{\beta}} \sim MVN(\vec{\beta}, \sigma^2 (X^T X)^{-1})$

To test the difference in mean response between NC and NE (i.e. β_2)

we can use $\hat{\beta}_j \sim N(\beta_j, \sigma^2 (X^T X)^{-1}_{jj})$ and $SE(\hat{\beta}_j) = \hat{\sigma}\sqrt{(X^T X)^{-1}_{jj}}$

(similarly for differences given by β_3 and β_4).

How do we test the difference in mean response between NC and S (i.e. $\beta_2 - \beta_3$)?

$$Var(\hat{\beta}_2 - \hat{\beta}_3) = \underbrace{Var(\hat{\beta}_2)}_{\sigma^2 (X^T X)^{-1}_{22}} + \underbrace{Var(\hat{\beta}_3)}_{\sigma^2 (X^T X)^{-1}_{33}} - 2\underbrace{Cov(\hat{\beta}_2, \hat{\beta}_3)}_{\sigma^2 (X^T X)^{-1}_{23}}$$

$$= \sigma^2 (X^T X)^{-1}_{22} + \sigma^2 (X^T X)^{-1}_{33} - 2\sigma^2 (X^T X)^{-1}_{23}$$

$$SE(\hat{\beta}_2 - \hat{\beta}_3) = \hat{\sigma}\sqrt{(X^T X)^{-1}_{22} + (X^T X)^{-1}_{33} - 2(X^T X)^{-1}_{23}}$$

and we can use $SE(\hat{\beta}_2 - \hat{\beta}_3)$ to test for this difference.

* In general, for categorical variables w/ k categories, need $k-1$ indicator variables