

# Inferential Statistic

- The parameter to which the hypothesis refers to has a point estimate.
- The **test statistic** describes how far that point estimate falls from the parameter value given in the null hypothesis.
- Usually, this distance is measured by the number of standard errors between the point estimate and the parameter value in  $H_0$ .
- In order to compute the value of the test statistic in a given situation, we will need to know:
  - the values in the sample, and
  - the parameter value in the null hypothesis.

The second interpretation is from the view on the contribution of  $x_j$  to the response  $y$  after both  $y$  and  $x_j$  have been linearly adjusted for all other regressors.

- To illustrate, consider model  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$  where we would like to interpret the effect of  $x_2$  on  $y$ .

- Step 1: Model 1 = model  $y$  on  $x_1$  (linearly adjust  $y$  on  $x_1$ ):

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 x_1, \text{ hence the residual is } y - \hat{y} = e_{y,x_1}.$$

- Step 2: Model 2 = model  $x_2$  on  $x_1$  (linearly adjust  $x_2$  on  $x_1$ ):

$$\hat{x}_2 = \hat{\gamma}_0 + \hat{\gamma}_1 x_1, \text{ hence the residual is } x_2 - \hat{x}_2 = e_{x_2,x_1}.$$

- Step 3: Model 3 = model  $e_{y,x_1}$  on  $e_{x_2,x_1}$  (the effect of  $x_2$  after  $y$  and  $x_2$  are linearly adjusted for  $x_1$ ):

$$\hat{e}_{y,x_1} = \hat{\lambda}_0 + \hat{\lambda}_1 e_{x_2,x_1}, \text{ hence the residual is } e_{y,x_1} - \hat{e}_{y,x_1}.$$

- Since  $E(\epsilon) = 0$  and  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$  we have

$$E(\hat{\beta}) = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \epsilon)] \\ = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon] = \beta$$

Thus,  $\hat{\beta}$  is an unbiased estimator of  $\beta$  if the model is correct.

- We have

$$\text{Cov}(\hat{\beta}) = E\{[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]'\},$$

a  $p \times p$  symmetric matrix, where whose the  $j$ th diagonal element is  $\text{Var}(\hat{\beta}_j)$  and  $(ij)$ th off-diagonal element is  $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j)$ . Hence

$$\text{Cov}(\hat{\beta}) = \text{Var}(\hat{\beta}) = \text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}(\mathbf{y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \\ = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- Denote  $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ , then  $\text{Var}(\hat{\beta}_j) = \sigma^2 C_{jj}$  and  $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}$ .

- We wish to find the vector of LSE that minimizes

$$S(\beta) = \sum_{i=1}^n \varepsilon_i^2 = \mathbf{\varepsilon}'\mathbf{\varepsilon} = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \\ = \mathbf{y}'\mathbf{y} - \beta'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta \\ = \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta$$

- Note that  $\beta'\mathbf{X}'\mathbf{y}$  is a  $1 \times 1$  matrix, or a scalar, hence the least squares estimators must satisfy

$$\left[ \frac{\partial S}{\partial \beta} \right]_{\hat{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} = 0 \iff \mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \quad (5)$$

- Equations (5) are the least squares equations. To solve the normal equations, multiply both sides of (5) by the inverse of  $\mathbf{X}'\mathbf{X}$ , we get the LSE of  $\beta$  is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

provided that the inverse matrix  $(\mathbf{X}'\mathbf{X})^{-1}$  exists.

- $(\mathbf{X}'\mathbf{X})^{-1}$  always exists if the regressors are linearly independent, that is, no column of the  $\mathbf{X}$  matrix is a linear combination of the other columns.

- This is an unbiased estimator of  $E(y|\mathbf{x}_0)$ , since  $E(\hat{y}_0) = \mathbf{x}'_0 \beta = E(y|\mathbf{x}_0)$ , and the variance of  $\hat{y}_0$  is

$$\text{Var}(\hat{y}_0) = \sigma^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0$$

- Therefore, a  $100(1 - \alpha)\%$  CI on the mean response at the point  $x_{01}, x_{02}, \dots, x_{0k}$  is

$$(\hat{y}_0 - t_{n-p}(\alpha/2)\sqrt{\sigma^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0}; \hat{y}_0 + t_{n-p}(\alpha/2)\sqrt{\sigma^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0}).$$

- Adding variable  $x_2$  has improved the precision of estimation. However, the change in the length of the CI depends on the location of the point in the  $x$  space.

- Consider the point  $x_1 = 16$ ,  $x_2 = 688$  feet. The 95% CI for the simple LR with  $x_1$  only is (35.6, 40.68) with length 5.08 minutes. While the 95% CI for the multiple LR with both variables  $x_1$  and  $x_2$  is (36.11, 40.08) with length 3.97 minutes.

The residual mean square is

$$MS_{Res} = \frac{SS_{Res}}{n-p} \text{ which has } E(MS_{Res}) = \sigma^2.$$

Hence,  $\hat{\sigma}^2 = MS_{Res}$  is an unbiased estimator of  $\sigma^2$ .

## Models with Interaction Effects (1)

- Linear regression models may also contain interaction effects:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon, \quad (3)$$

if we let  $x_3 = x_1 x_2$  then this model becomes

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

which has the form of linear regression model as (2).

- Note that, any regression model that is linear in the parameters (the  $\beta$ 's) is a linear regression model, regardless of the shape of the surface that it generates (meaning regardless of the linearity of the regressors).
- For model (3), the expected change in  $y$  when  $x_1$  changes will depend on the level of  $x_2$ .

## Confidence Interval Estimates (2)

- Hence,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t_{n-p}, \quad j = 0, 1, \dots, k.$$

- A  $100(1 - \alpha)\%$  CI for the regression coefficient  $\beta_j$ ,  $j = 0, 1, \dots, k$  is

$$\hat{\beta}_j - t_{n-p}(\alpha/2)\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{n-p}(\alpha/2)\sqrt{\hat{\sigma}^2 C_{jj}}$$

- Note that,

$$se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$$

is the standard error of the regression coefficient  $\hat{\beta}_j$ .

- It can be shown that

$$\frac{(\hat{\beta} - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta)}{p MS_{Res}} \sim F_{p,n-p}.$$

- This implies that

$$P\left\{ \frac{(\hat{\beta} - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta)}{p MS_{Res}} \leq F_{p,n-p}(\alpha) \right\} = 1 - \alpha$$

- Hence a  $100(1 - \alpha)\%$  joint confidence region for all the parameters in  $\beta$  is

$$\frac{(\hat{\beta} - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta)}{p MS_{Res}} \leq F_{p,n-p}(\alpha)$$

- The hypotheses for testing the significance of any individual regression coefficients,  $\beta_j$  are

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

- Not rejecting  $H_0$  indicates that the regressor  $x_j$  can be deleted from the model.

- The test statistics

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

where  $C_{jj}$  is the diagonal element of  $(\mathbf{X}'\mathbf{X})^{-1}$  corresponding to  $\hat{\beta}_j$

- The refitting must be done since the coefficient estimates for an individual regressor depend on all of the regressors,  $x_j$ .

- However, if the columns are orthogonal to each other, then there is no need to refit.

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \epsilon \\ &= \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon \end{aligned}$$

If the columns  $\mathbf{X}_1$  are orthogonal to the columns in  $\mathbf{X}_2$ , the sum of squares due to  $\beta_2$  that is free of any dependence on the regressors in  $\mathbf{X}_1$ .

- However, for the similar accuracy, a simpler model is preferred, hence we have adjusted  $R^2$ , denoted as  $R^2_{Adj}$  which penalizes you for added terms to the model that are not significant.

$$R^2_{Adj} = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)}.$$

- In Bonferroni method, we set  $\Delta = t_{\alpha/2p, n-p}$  so that the CIs become

$$\hat{\beta}_j \pm t_{\alpha/2p, n-p} se(\hat{\beta}_j), \quad j = 0, 1, \dots, k.$$

- With this method, the probability is at least  $(1 - \alpha)$  that all intervals are correct. For each interval, the confidence level is  $(1 - \alpha/p)$ .

## Step 4: Interpreting a Test Statistic

We shall use a probability to summarise the evidence against the null hypothesis,  $H_0$ .

- First, we presume that  $H_0$  is true.
- Then, we consider the values we would expect to get for the test statistic according to the sampling distribution of the test statistic, presuming  $H_0$  is true.
- If the sample test statistic falls well out in the tail of the sampling distribution, then it is far from what  $H_0$  predicts.
  - If  $H_0$  were indeed true, such a value would be unlikely.
- We summarize how unlikely the observed value is, by computing the tail probability in the sampling distribution.
- The probability that the test statistic assumes a value more extreme than what has been observed is known as the  $p$ -value.
- When  $p$ -value is small, either the assumption " $H_0$  is true" is not correct, or the sample is not representative of the population.
- A small  $p$ -value (close to 0) provides strong evidence against  $H_0$ .

- We then have

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{Hy}$$

- The  $n \times n$  matrix  $\mathbf{H}$  is usually called the **hat matrix**. It maps the vector of observed values into a vector of fitted values.

- The hat matrix plays a central role in regression analysis.

- Let  $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$  be the residual, then the  $n$  residuals can be conveniently written in matrix notation as

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \quad \text{or}$$

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{y} - \mathbf{Hy} = (\mathbf{I} - \mathbf{H})\mathbf{y}.$$

- The least squares normal equations are:

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n x_{ik} &= \sum_{i=1}^n y_i \\ \hat{\beta}_0 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1}x_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n x_{i1}x_{ik} &= \sum_{i=1}^n x_{i1}y_i \end{aligned}$$

- There are  $p = k + 1$  normal equations.

$$\begin{aligned} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \mathbf{e} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}. \end{aligned}$$

$SS_R(\beta_2|\beta_1)$  is independent of  $MS_{Res}$ , and the null  $H_0 : \beta_2 = 0$  can be tested using the test statistic:

$$F_0 = \frac{SS_R(\beta_2|\beta_1)/r}{MS_{Res}}$$

which follows  $F_{r,n-p}$  under  $H_0$ .

</