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.

Models with Interaction Effects (1)

- Linear regression models may also contain interaction effects:

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

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

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

  which has the form of linear regression model as (2).
- Note that, **any regression model that is linear in the parameters (the β'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.

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.

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 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ 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):

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

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

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

- 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):

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

Confidence Interval Estimates (2)

- Hence,

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

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

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

- Note that,

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

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

- We then have

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

- 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 $e_i = y_i - \hat{y}_i$ be the residual, then the n residuals can be conveniently written in matrix notation as

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

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

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

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right] = E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\right] \\ &= E\Big[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})\Big] = \boldsymbol{\beta} \end{aligned}$$

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

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

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

$$\begin{aligned} \text{Cov}(\hat{\boldsymbol{\beta}}) = \text{Var}(\hat{\boldsymbol{\beta}}) &= \text{Var}\Big[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\Big] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\text{Var}(\mathbf{y})\Big[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Big]' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

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

- It can be shown that

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

- This implies that

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

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

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

- The least squares normal equations are:

$$n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^n x_{i1} + \hat{\beta}_2\sum_{i=1}^n x_{i2} + ... + \hat{\beta}_k\sum_{i=1}^n x_{ik} = \sum_{i=1}^n y_i$$

$$\hat{\beta}_0\sum_{i=1}^n x_{i1} + \hat{\beta}_1\sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2\sum_{i=1}^n x_{i1}x_{i2} + ... + \hat{\beta}_k\sum_{i=1}^n x_{i1}x_{ik} = \sum_{i=1}^n x_{i1}y_i$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot$$

$$\hat{\beta}_0\sum_{i=1}^n x_{ik} + \hat{\beta}_1\sum_{i=1}^n x_{ik}x_{i1} + \hat{\beta}_2\sum_{i=1}^n x_{ik}x_{i2} + ... + \hat{\beta}_k\sum_{i=1}^n x_{ik}^2 = \sum_{i=1}^n x_{ik}y_i$$

- There are $p = k + 1$ normal equations.
- We use $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$, $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$, $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$.

- We wish to find the vector of LSE that minimizes

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

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

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

- The hypotheses for testing the significance of any individual regression coefficients, β_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 $\hat{\beta}_j$

$SS_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)$ is independent of MS_{Res}, and the null $H_0: \boldsymbol{\beta}_2 = 0$ can be tested using the test statistic:

$$F_0 = \frac{SS_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)/r}{MS_{Res}}$$

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

At significance level α, $H_0: \boldsymbol{\beta}_2 = 0$ is rejected if $F_0 > F_{r,n-p}(\alpha)$, which implies that at least one of the regressor $x_{k-r+1}, ..., x_k$ in $\mathbf{X}_2$ contribute significantly to the model.

- 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 $\boldsymbol{\beta}$ is

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

  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.

- 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.

2 5 Steps of Hypothesis Testing in General

- Step 1: Assumptions
- Step 2: Hypotheses
- Step 3: Test Statistic
- Step 4: p-Value
- Step 5: Conclusion
- Summary

- This is an unbiased estimator of $E(y|\mathbf{x}_0)$, since $E(\hat{y}_0) = \mathbf{x}_0'\,\boldsymbol{\beta} = E(y|\mathbf{x}_0)$, 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 $x_{01}, x_{02}, ...x_{0k}$ is

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

Consider the model

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \end{aligned}$$

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

- 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.

- 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.

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

The residual mean square is

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

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

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

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

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

- At a particular point, such as $x_{01}, x_{02}, ...x_{0k}$, a $100(1-\alpha)\%$ prediction interval for a future observation is

$$\hat{y}_0 - t_{n-p}(\alpha/2)\sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0'\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \le y_0 \le \hat{y}_0 + t_{n-p}(\alpha/2)\sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0'\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{x}_0)}$$