

9 Lecture 9: Feb 14

Last time

- Simple correlation
- The statistical model of the SLR (JF chapter 6)

Today

- Properties of the Least-Squares estimator
- Inference of SLR model

Properties of the Least-Squares estimator

Under the strong assumptions of the simple regression model, the sample least squares coefficients $\hat{\beta}_{ls}$ have several desirable properties as estimators of the population regression coefficients β_0 and β_1 :

- The least-squares intercept and slope are *linear estimators*, in the sense that they are linear functions of the observations y_i .

Proof:

- The sample least-squares coefficients are *unbiased estimators* of the population regression coefficients:

$$\mathbf{E}(\hat{\beta}_0) = \beta_0$$

$$\mathbf{E}(\hat{\beta}_1) = \beta_1$$

Proof:

- Both $\hat{\beta}_0$ and $\hat{\beta}_1$ have simple sampling variances:

$$\text{Var}(\hat{\beta}_0) = \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2}$$

Proof:

- Rewrite the formula for $\text{Var}(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{(n-1)S_X^2}$, we see that the sampling variance of the slope estimate will be small when

- The error variance σ_ϵ^2 is small

- The sample size n is large
- The explanatory-variable values are spread out (i.e. have a large variance, S_X^2)
- (Gauss-Markov theorem) Under the assumptions of linearity, constant variance, and independence, the least-squares estimators are BLUE (Best Linear Unbiased Estimator), that is they have the smallest sampling variance and are unbiased. (show this)
Proof:

- Under the full suite of assumptions, the least-squares coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are the maximum-likelihood estimators of β_0 and β_1 . (show this)
Proof:

- Under the assumption of normality, the least-squares coefficients are themselves normally distributed. Summing up,

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2}\right)$$

Statistical inference of the SLR model

Now we have the distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2}\right).$$

However, σ_ϵ is never known in practice. Instead, an *unbiased* estimator of σ_ϵ^2 is given by

$$\hat{\sigma}_\epsilon^2 = MS[E] = \frac{SS[E]}{n - 2}.$$

Proof:

Confidence intervals

Now we substitute $\hat{\sigma}_\epsilon^2$ into the distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\hat{\sigma}_\epsilon^2}{\sum (x_i - \bar{x})^2}\right)$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\hat{\sigma}_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}\right)$$

to get the estimated standard errors:

$$\widehat{SE}(\hat{\beta}_1) = \sqrt{\frac{MS[E]}{\sum(x_i - \bar{x})^2}}$$

$$\widehat{SE}(\hat{\beta}_0) = \sqrt{MS[E] \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right)}$$

And the $100(1 - \alpha)\%$ confidence intervals for β_1 and β_0 are given by

$$\hat{\beta}_1 \pm t(n - 2, \alpha/2) \sqrt{\frac{MS[E]}{S_{xx}}}$$

$$\hat{\beta}_0 \pm t(n - 2, \alpha/2) \sqrt{MS[E] \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}$$

where $S_{xx} = \sum(x_i - \bar{x})^2$

Confidence interval for $\mathbf{E}(Y|X = x_0)$

The conditional mean $\mathbf{E}(Y|X = x_0)$ can be estimated by evaluating the regression function $\mu(x_0)$ at the estimates $\hat{\beta}_0, \hat{\beta}_1$. The conditional variance of the expression isn't too difficult (already shown):

$$\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 | X = x_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

This leads to a confidence interval of the form

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t(n - 2, \alpha/2) \sqrt{MS[E] \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$$

Prediction interval

Often, prediction of the response variable Y for a given value, say x_0 , of the independent variable of interest. In order to make statements about future values of Y , we need to take into account

- the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$
- the randomness of a future value Y .

We have seen the predicted value of Y based on the linear regression is given by $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

The 95% prediction interval has the form

$$\hat{Y}_0 \pm t(n - 2, \alpha/2) \sqrt{MS[E] \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}.$$

Hypothesis test

To test the hypothesis $H_0 : \beta_1 = \beta_{slope_0}$ that the population slope is equal to a specific value β_{slope_0} (most commonly, the null hypothesis has $\beta_{slope_0} = 0$), we calculate the test statistic (T -statistics) with $df = n - 2$

$$t_0 = \frac{\hat{\beta}_1 - \beta_{slope_0}}{\widehat{SE}(\hat{\beta}_1)} \sim t_{n-2}$$