7 Lecture 7: Feb 3

Last time

- Statistical model of SLR

Today

- Properties of the LS estimators
- 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:

$$\mathbf{E}\left(\hat{\beta}_0\right) = \beta_0$$
$$\mathbf{E}\left(\hat{\beta}_1\right) = \beta_1$$

  Proof:

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

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_0) =& \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2} \\ \mathrm{Var}(\hat{\beta}_1) =& \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2}\end{aligned}$$

  Proof:

- Rewrite the formula for $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{(\mathrm{n}-1)\mathrm{S}_\mathrm{X}^2}$, 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,

$$\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2})$$
$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2})$$

Statistical inference of the SLR model

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

$$\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2})$$
$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2}).$$

However, σ_ϵ is never known in practice. Instead, an 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 $\hat{\beta}_1$

$$\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2})$$
$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2})$$

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):

$$\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}_0 | \mathrm{X} = \mathrm{x}_0) = \sigma^2 (\frac{1}{\mathrm{n}} + \frac{(\mathrm{x}_0 - \bar{\mathrm{x}})^2}{\mathrm{S}_{\mathrm{xx}}})$$

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 <u>predicted value</u> of Y based on the linear regression is given by $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

The <u>95% prediction interval</u> has the form

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