Regression Analysis

Simple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Regression Concepts:
Statistical Inference

Georgia Tech

1

About This Lesson

Georgia Tech

2

Regression Estimators: Properties

For the slope parameter β_1, we can show

$$\hat{\beta}_1 = \frac{\Sigma(x_i-\bar{x})Y_i}{S_{xx}} \text{ but } x_i \text{ fixed} \rightarrow \frac{x_i-\bar{x}}{S_{xx}} = c_i \text{ fixed}$$

$$E[\hat{\beta}_1] = E\left[\sum_{i=1}^{n} c_i y_i\right] = \sum_{i=1}^{n} c_i E[y_i]$$

$$= \sum_{i=1}^{n} c_i(\beta_0+\beta_1 x_i) = \beta_0 \underbrace{\sum_{i=1}^{n} c_i}_{=0} + \beta_1 \underbrace{\sum_{i=1}^{n} c_i x_i}_{=1}$$

$$= \beta_1 \rightarrow E[\hat{\beta}_1] = \beta_1$$

$$\boxed{\begin{gathered} E(\hat{\beta}_1) = \beta_1 \\ Var(\hat{\beta}_1) = \sigma^2 / S_{xx} \end{gathered}}$$

3

Regression Estimators: Properties

Furthermore, $\hat{\beta}_1$ is a linear combination of $\{Y_1,\ldots,Y_n\}$. If we assume that $e_i \sim$ Normal $(0, \sigma^2)$, then $\hat{\beta}_1$ is also distributed as

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

$\hat{\beta}_1 = \sum_{i=1}^{m} c_i Y_i$ a linear combination of normally distributed random variables

$\hat{\beta}_1$ ~ Normally distributed

4

Regression Estimators: Properties

Sampling Distribution of $\hat{\beta}_1$:

We do not know σ^2. We can replace it by MSE, but then the sampling distribution becomes the t-distribution with **n-2** df.

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

$$\hat{\sigma}^2 = \mathrm{MSE} = \frac{\sum \hat{\epsilon}_i^2}{\mathrm{n}-2} \sim \chi^2_{n-2}$$

$$\rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\mathrm{MSE}}{\mathrm{S_{xx}}}}} \sim \mathrm{t}_{n-2}$$

Georgia Tech

5

Inference for Slope Parameter

Given the sampling distribution of $\hat{\beta}_1$, we can derive confidence intervals and perform hypothesis testing for β_1:

$$\left(\hat{\beta}_1 - \left(t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\mathrm{MSE}}{\mathrm{S}_{xx}}}\right), \hat{\beta}_1 + \left(t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\mathrm{MSE}}{\mathrm{S}_{xx}}}\right)\right)$$

Georgia Tech

6

Confidence Interval Derivation

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\text{MSE}}{S_{xx}}}} \sim t_{n-2} \rightarrow \text{t–interval for } \beta_1$$

1 - α
Confidence interval

$$\rightarrow \underbrace{\hat{\beta}_1}_{\text{Estimate of } \beta_1} \pm \underbrace{t_{\frac{\alpha}{2},n-2}}_{\text{t-critical point}} \underbrace{\sqrt{\frac{\text{MSE}}{S_{xx}}}}_{\text{Standard Deviation/Error of } \hat{\beta}_1}$$

Sampling distribution of $\hat{\beta}_1$ is t_{n-2}

$V[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}$
$\sigma^2 \leftarrow \text{MSE}$

Georgia Tech

Testing the Overall Regression

One way we can test statistical significance is to use the t-test for

H_o: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$

$$\boxed{\text{t-value} = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{\hat{\beta}_1\sqrt{S_{xx}}}{\hat{\sigma}}}$$

We reject H_o if |t-value| is large. If the null hypothesis is rejected, we interpret this as β_1 being **statistically significant**.

Georgia Tech

Testing Regression at Different Levels

How will the procedure change if we test:
$H_0: \beta_1 = c$ vs. $H_A: \beta_1 \neq c$
for some known c?

t-value = $\frac{\hat{\beta}_1 - c}{se(\hat{\beta}_1)}$ how large to reject $H_0: \beta_1 = c$?

For significance level α, Reject if |t-value| $> t_{\frac{\alpha}{2}, n-2}$

Alternatively, compute P-value = $2P(T_{n-2} > |\text{t-value}|)$

If P-value small **(p-value < 0.01)** → **Reject**

Georgia Tech

9

Testing Regression at Different Levels (cont'd)

How will the procedure change if we test:
$H_o: \beta_1 = 0$ **versus** $H_A: \beta_1 > 0$
OR
$H_o: \beta_1 = 0$ **versus** $H_A: \beta_1 < 0$?

What if we want to test for positive relationship
$H_o: \beta_1 \leq 0$ **versus** $H_A: \beta_1 > 0$?
P-value = $P(T_{n-2} > \text{t-value})$

What if we want to test for negative relationship
$H_o: \beta_1 \geq 0$ **versus** $H_A: \beta_1 < 0$?
P-value = $P(T_{n-2} < \text{t-value})$

Georgia Tech

10

Inference for Intercept Parameter

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

$$E(\hat{\beta}_0) = E(\bar{Y}) - E(\hat{\beta}_1)\bar{x} = \beta_0$$

$$Var(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)$$

Confidence interval:

$$\left(\hat{\beta}_0 - \left(t_{\frac{\alpha}{2},n-2}\sqrt{\mathrm{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)}\right), \hat{\beta}_0 + \left(t_{\frac{\alpha}{2},n-2}\sqrt{\mathrm{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)}\right)\right)$$

11

Summary

12