Simple Linear Regression

One of the first known uses of regression was to study inheritance of traits from generation to generation in a study in the UK from 1893-1898. E.S. Pearson organized the collection of heights of mothers and one of their adult daughters over 18. The mother's height was the predictor variable (x) and the daughter's height was the response variable (y).

The goal of a linear regression is to quantify the relationship between one independent variable and a single dependent variable. A simple linear regression can be represented by the following:

- y_hat = $\beta_0 + \beta_1 x$ + Error; E(Error) = 0; V(Error) = σ^2
- E(y) = $\beta_0 + \beta_1 x$; V(y) = σ^2

As with correlation, a strong association in a regression analysis does **NOT** imply causality

Additionally, prediction values outside the range of values observed for x is not reliable, and is called **extrapolation.**

Least Squares Estimation

OLS/LS - Ordinary Least Squares is a method that looks for the line that minimizes the "residual sum of squares"

Residual = Observed - Predicted = y - ŷ

So we can set up an equation for sum of squared residuals: $\sum(y - \hat{y})^2$

Then substitute the linear regression equation, take the derivative, set to zero and solve.The solution comes out to:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}} = r_{XY}\sqrt{\frac{S_{YY}}{S_{XX}}}$$

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

The fitted equation will pass through x bar, y bar (the center of data)

$$E(\hat{\beta}_1) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} E(y_i) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} (\beta_0 + \beta_1 x_i) = \beta_1;$$

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

$$E(\hat{\sigma}^2) = \sigma^2$$

Estimating Variances of LSE

The square root of an estimated variance is called **standard error** represented by s or se().

$$\widehat{\sigma^2} = \mathrm{RSS}/(\mathrm{n}\text{-}2) = \sum(y_i - \hat{y}_i)^2 \ /(\mathrm{n} - 2)$$

E(Y |X = x) = a function that depends on the value of x

$$Var(\widehat{\beta_1}|\mathrm{X}) = \frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}$$

$$Var(\widehat{\beta_0}|\mathrm{X}) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x - x_i)^2}\right)$$

LSE Assuming Normally Distributed Data

The distributions of estimates are used to make predictions and hypothesis testing. Since variance of a fixed variable is 0, we can estimate the error of a normal distribution as follows:

$$y = \beta_0 + \beta_1 x + \varepsilon;$$

$$\varepsilon \sim N(0, \sigma^2) \Rightarrow \ y \sim N(\beta_0 + \beta_1 x; \sigma^2)$$

$$\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} \sim N\left(\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, \sigma^2\begin{pmatrix}\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}} & -\frac{\bar{x}}{S_{XX}}\\ -\frac{\bar{x}}{S_{XX}} & \frac{1}{S_{XX}}\end{pmatrix}\right)$$

The estimates are correlated with covariance -x_bar/S_{xx}

$$\frac{(n-2)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-2}$$

It can be shown that the Least Squares Estimate is also the Maximum Likelihood Estimate.

Confidence Intervals

The distribution of β_1 is normal if the variance is known.

When we use estimated variance, we use a Student's t distribution on n-2 degrees of freedom to estimate parameters:

$$\hat{\beta}_1 \pm t_{(\text{crit, two-tailed}\,\alpha\,,\text{n-2 df})} \times se(\hat{\beta}_1)$$

By finding the standard error of y_hat we can also calculate a confidence interval for a fitted value, or a given of x

$$V(\hat{y}) = V(\hat{\beta}_0) + x^2 V(\hat{\beta}_1) + 2x\,\text{cov}(\hat{\beta}_0, \hat{\beta}_1) \Rightarrow$$

$$V(\hat{y}) = \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}+\frac{x^2}{S_{XX}}-\frac{2x\times\bar{x}}{S_{XX}}\right) = \sigma^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{XX}}\right)$$

The interval width is:

$$2\sqrt{V(\hat{y})} = 2\hat{\sigma}\sqrt{\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{XX}}\right)}$$

And width will increase as the distance between observed and expected increases.

ANOVA

Though ANOVA is usually used in determining if there is a difference in variance with data containing 3 or more categories, linear regression under standard conditions is a special case of ANOVA. The name **analysis of variance** is derived from a partitioning of total variability into its component parts.

ANOVA For Simple Linear Regression

TABLE 2.3 The Analysis of Variance Table for Simple Regression

Source	df	SS	MS	F	p-value
Regression	1	$SSreg$	$SSreg/1$	$MSreg/\hat{\sigma}^2$	
Residual	$n-2$	RSS	$\hat{\sigma}^2 = RSS/(n-2)$		
Total	$n-1$	SYY			

Global Null Hypothesis

$$\boldsymbol{H_0}: \textbf{Model fit not significant } (\boldsymbol{SS_{reg} = 0}); \textbf{ or } \boldsymbol{E(Y) = \beta_0}$$
$$\boldsymbol{H_1}: \textbf{Model fit significant } (\boldsymbol{SS_{reg} > 0});$$

which is equivalent to $\boldsymbol{H_0: \beta_1 = 0}$, $\boldsymbol{H_1: \beta_1 \neq 0}$

$$\frac{\boldsymbol{SS_{reg}}}{\widehat{\boldsymbol{\sigma^2}}} = \left(\frac{\widehat{\boldsymbol{\beta_1}}}{\boldsymbol{se}(\widehat{\boldsymbol{\beta_1}})}\right)^2$$

This formula uses an F-distribution with 1 and n − 2 df in place of the t-distribution to correct for the simultaneous inference from our estimate of Beta_1.

Prediction of New Observations

When using points not in the dataset (extrapolation) use the following adjusted formulas for variance of V(y):

$$\sigma^2\left(1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{XX}}\right)$$

Thus, the confidence interval for a new observation would be significantly wider

Other notations commonly seen in our textbook

TABLE 2.1 Definitions of Symbols[a]

Quantity	Definition	Description
$\overline{x}$	$\sum x_i/n$	Sample average of x
$\overline{y}$	$\sum y_i/n$	Sample average of y
SXX	$\sum(x_i-\overline{x})^2=\sum(x_i-\overline{x})x_i$	Sum of squares for the x's
SD_x^2	$SXX/(n-1)$	Sample variance of the x's
SD_x	$\sqrt{SXX/(n-1)}$	Sample standard deviation of the x's
SYY	$\sum(y_i-\overline{y})^2=\sum(y_i-\overline{y})y_i$	Sum of squares for the y's
SD_y^2	$SYY/(n-1)$	Sample variance of the y's
SD_y	$\sqrt{SYY/(n-1)}$	Sample standard deviation of the y's
SXY	$\sum(x_i-\overline{x})(y_i-\overline{y})=\sum(x_i-\overline{x})y_i$	Sum of cross-products
s_{xy}	$SXY/(n-1)$	Sample covariance
r_{xy}	$s_{xy}/(\mathrm{SD}_x\mathrm{SD}_y)$	Sample correlation

[a]In each equation, the symbol $\sum$ means to add over all the n values or pairs of values in the data.

Revision #6
Created 8 September 2022 22:12:13 by Elkip
Updated 15 September 2022 21:15:13 by Elkip