Linear Regression

Stephen B. Holt, Ph.D.

Rockefeller College of Public Affairs & Policy
University at Albany State University of New York

May 4, 2022

Multiple Regression Setup

- Up to now, we have considered, in detail, the linear regression model of outcome Y using one explanatory variable, X:

$$\widehat{Y} = \beta_0 + \beta_1 X_1 \tag{1}$$

Multiple Regression Setup

- Up to now, we have considered, in detail, the linear regression model of outcome Y using one explanatory variable, X:

$$\widehat{Y} = \beta_0 + \beta_1 X_1 \tag{1}$$

- We know, of course, that for predicting most outcomes or studying most effects of a particular X, the population model will likely need to account for more factors than a single X, particularly in the absence of random assignment.

Multiple Regression Setup

- Up to now, we have considered, in detail, the linear regression model of outcome Y using one explanatory variable, X:

$$\widehat{Y} = \beta_0 + \beta_1 X_1 \tag{1}$$

- We know, of course, that for predicting most outcomes or studying most effects of a particular X, the population model will likely need to account for more factors than a single X, particularly in the absence of random assignment.
- In multiple regression, the outcome Y depends on many explanatory variables in the population, denoted as $X_1, X_2, X_3, ...X_k$:

$$\widehat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \tag{2}$$

Data structure for Multiple Regression

- The data for a simple linear regression problem consists of n observations with data points at (x_i, y_i) of the two variables in the model.

Data structure for Multiple Regression

- The data for a simple linear regression problem consists of n observations with data points at (x_i, y_i) of the two variables in the model.
- Data for multiple linear regression consists of the value of outcome variable Y and k explanatory (or independent) variables $(X_1, X_2, X_3, ...X_k)$ for n observations.

Data structure for Multiple Regression

- The data for a simple linear regression problem consists of n observations with data points at (x_i, y_i) of the two variables in the model.
- Data for multiple linear regression consists of the value of outcome variable Y and k explanatory (or independent) variables $(X_1, X_2, X_3, ...X_k)$ for n observations.
- The data should be structured in the software as:

	Independent Variables				Dependent Variables
Case	X_1	X_2	X_3	X_4	Y
1	$x1_1$	$x1_2$	$x1_3$	$x1_4$	y1
2	$x2_1$	$x2_2$	$x2_3$	$x2_4$	y2
3	$x3_1$	$x3_2$	$x3_3$	$x3_4$	y3
n	xn_1	xn_2	xn_3	xn_4	yn

Multiple Linear Regression Model

- For k number of explanatory variables, we can express the population mean response (the outcome or μ_y) as a linear equation:

$$\mu_y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \tag{3}$$

Multiple Linear Regression Model

- For k number of explanatory variables, we can express the population mean response (the outcome or μ_y) as a linear equation:

$$\mu_y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \tag{3}$$

- The statistical model for n sample data $(i = 1, 2, ...n)$ is then:

$$\begin{aligned} Data &= \quad fit \quad + residual \\ y_i &= (\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}) + (\varepsilon_i) \end{aligned}$$

where the ε_i are independent and normally distributed $N(0, \sigma)$.

Multiple Linear Regression Model

- For k number of explanatory variables, we can express the population mean response (the outcome or μ_y) as a linear equation:

$$\mu_y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \tag{3}$$

- The statistical model for n sample data $(i = 1, 2, ...n)$ is then:

$$\begin{aligned} Data &= \quad fit \quad + residual \\ y_i &= (\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}) + (\varepsilon_i) \end{aligned}$$

  where the ε_i are independent and normally distributed $N(0, \sigma)$.

- Multiple linear regression assumes equal variance σ^2 of y.

Multiple Linear Regression Model

- For k number of explanatory variables, we can express the population mean response (the outcome or μ_y) as a linear equation:

$$\mu_y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \tag{3}$$

- The statistical model for n sample data $(i = 1, 2, ...n)$ is then:

$$\begin{aligned} Data &= \quad fit \quad + residual \\ y_i &= (\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}) + (\varepsilon_i) \end{aligned}$$

  where the ε_i are independent and normally distributed $N(0, \sigma)$.

- Multiple linear regression assumes equal variance σ^2 of y.
- $\beta_{0,1,...k}$ are parameters of the population model we try to estimate with our sample of n observations.

How It Works

The multivariate regression line is the line that minimizes the average squared residuals $(y_i - \widehat{y}_i)$ for the relationship between all x variables in the model and outcome y. In the case of a model with two x variables, the line can be found with:

$$\beta_1 = \frac{(\sum(X_{i2} - \overline{X_2})^2(\sum(X_{i1} - \overline{X}_1)(Y_i - \overline{Y})) - (\sum(X_{i1} - \overline{X}_1)(X_{i2} - \overline{X}_2))(\sum(X_{i2} - \overline{X}_2)(Y_i - \overline{Y}))}{(\sum(X_{i1} - \overline{X_1})^2(\sum(X_{i2} - \overline{X_2})^2 - (\sum(X_{i1} - \overline{X}_1)(X_{i2} - \overline{X}_2))^2} \quad (4)$$

$$\beta_2 = \frac{(\sum(X_{i1} - \overline{X_1})^2(\sum(X_{i2} - \overline{X}_2)(Y_i - \overline{Y})) - (\sum(X_{i1} - \overline{X}_1)(X_{i2} - \overline{X}_2))(\sum(X_{i1} - \overline{X}_1)(Y_i - \overline{Y}))}{(\sum(X_{i1} - \overline{X_1})^2(\sum(X_{i2} - \overline{X_2})^2 - (\sum(X_{i1} - \overline{X}_1)(X_{i2} - \overline{X}_2))^2} \quad (5)$$

$$\beta_0 = \overline{Y} - \beta_1\overline{X}_1 - \beta_2\overline{X}_2 \quad (6)$$

Estimation of the parameters

- From a simple random sample of n individuals for which we collect data on $k+1$ variables $(x_1, ...x_k, y)$, the least-squares regression method estimates the line that minimizes the sum of squared deviations $(e_i(= y_i - \widehat{y_i}))$ to express y as the linear function of k explanatory variables:

$$\widehat{y_i} = b_0 + b_1 x_{1i} + ... + b_k x_{ki} \tag{7}$$

Estimation of the parameters

- From a simple random sample of n individuals for which we collect data on $k+1$ variables $(x_1, ...x_k, y)$, the least-squares regression method estimates the line that minimizes the sum of squared deviations $(e_i(= y_i - \widehat{y}_i))$ to express y as the linear function of k explanatory variables:

$$\widehat{y}_i = b_0 + b_1 x_{1i} + ... + b_k x_{ki} \tag{7}$$

- As is the case with simple linear regression, the constant b_0 is the intercept of the least-squares line of y.

Estimation of the parameters

- From a simple random sample of n individuals for which we collect data on $k+1$ variables $(x_1, ...x_k, y)$, the least-squares regression method estimates the line that minimizes the sum of squared deviations $(e_i(= y_i - \widehat{y}_i))$ to express y as the linear function of k explanatory variables:

$$\widehat{y}_i = b_0 + b_1 x_{1i} + ... + b_k x_{ki} \tag{7}$$

- As is the case with simple linear regression, the constant b_0 is the intercept of the least-squares line of y.
- The coefficients $(b_1 \textit{ through } b_k)$ reflect the unique association of each independent variable in the model with outcome y, analogous to the slope of the simple linear model. They provide unbiased estimates of the population parameters.

Confidence Intervals for Regression Parameters

- Estimating the regression parameters $\beta_0, ..., \beta_k$ is a case of one-sample inference with unknown population variance.

Confidence Intervals for Regression Parameters

- Estimating the regression parameters $\beta_0, ..., \beta_k$ is a case of one-sample inference with unknown population variance.
- We rely on the t distribution with $n - k - 1$ degrees of freedom.

Confidence Intervals for Regression Parameters

- Estimating the regression parameters $\beta_0, ..., \beta_k$ is a case of one-sample inference with unknown population variance.
- We rely on the t distribution with $n - k - 1$ degrees of freedom.
- A level C confidence interval for the slope (β_j) is proportional to the standard error of the least-squares estimate of β_j:

$$b_j \pm t * SE_{bj} \tag{8}$$

Confidence Intervals for Regression Parameters

- Estimating the regression parameters $\beta_0, ..., \beta_k$ is a case of one-sample inference with unknown population variance.
- We rely on the t distribution with $n - k - 1$ degrees of freedom.
- A level C confidence interval for the slope (β_j) is proportional to the standard error of the least-squares estimate of β_j:

$$b_j \pm t * SE_{bj} \tag{8}$$

- We rely on statistics software to estimate SE_{bj}

Confidence Intervals for Regression Parameters

- Estimating the regression parameters $\beta_0, ..., \beta_k$ is a case of one-sample inference with unknown population variance.
- We rely on the t distribution with $n - k - 1$ degrees of freedom.
- A level C confidence interval for the slope (β_j) is proportional to the standard error of the least-squares estimate of β_j:

$$b_j \pm t * SE_{bj} \tag{8}$$

- We rely on statistics software to estimate SE_{bj}
- Note that t* is the t-critical value for the $t(n - k - 1)$ distribution with area C between -t* and +t*.

Confidence Intervals for Regression Parameters

- Estimating the regression parameters $\beta_0, ..., \beta_k$ is a case of one-sample inference with unknown population variance.
- We rely on the t distribution with $n - k - 1$ degrees of freedom.
- A level C confidence interval for the slope (β_j) is proportional to the standard error of the least-squares estimate of β_j:

$$b_j \pm t * SE_{bj} \tag{8}$$

- We rely on statistics software to estimate SE_{bj}
- Note that t* is the t-critical value for the $t(n - k - 1)$ distribution with area C between -t* and +t*.
- As before, our t-score for b_j is again calculated as a ratio of the coefficient to the standard error:

$$t = \frac{b_j}{SE_{bj}} \tag{9}$$

with a t distribution of $n - k - 1$ degrees of freedom.

Predictions Using Regressions

Once we estimate a line of best fit, we can use the line of best fit to make predictions based on our model and our sample. Note that predictions for out-of-sample characteristics are generally not meaningful.

Example, estimating a model of mental health using a score where higher scores is worse mental health, we get a estimated linear regression:

$MentalHealth = 0.495 + -0.006HrsWrk + -0.336College + -0.019Age + -0.024Rent + 0.123Povert + 0.417Married$

A single 30 year old with no rental assistance, no college education, who works 20 hours per week, living under the poverty line would be predicted to have a mental health score of -0.072 or just a little better than the average American $(0.495 + (-0.006 * 20) + (-0.019 * 30) + (0.123 * 1))$. By comparison, a 45 year old with a college degree working 40 hours per week who is married and not living in poverty or receiving rental assistance would be predicted to have a mental health score of -0.519, which is even better than the average American $(0.495 + (-0.006 * 40) + (-0.019 * 45) + (-0.336 * 1) + (0.417 * 1))$.

Coefficient of Determination (R^2)

- The coefficient of determination, generally referred to as R^2 or the square of the correlation coefficient, measures the percentage of the variance in y (vertical scatter from the regression line) that can be explained by changes in x. In multiple regression, the calculation and interpretation is the same *except* the predicted y ($\widehat{y}$) of the model includes all explanatory variables taken together.

Coefficient of Determination (R^2)

- The coefficient of determination, generally referred to as R^2 or the square of the correlation coefficient, measures the percentage of the variance in y (vertical scatter from the regression line) that can be explained by changes in x. In multiple regression, the calculation and interpretation is the same *except* the predicted y ($\widehat{y}$) of the model includes all explanatory variables taken together.
- More formally:

$$R^2 = \frac{\sum(\widehat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y}_i)^2} = \frac{SSM}{SST} \tag{10}$$

Attendance