1 Notes on Regression Analysis

Updated February 11, 2026.

Linear regression is a technique for summarizing data as a linear equation, predicting values of an outcome variable Y as a linear combination of independent variables X_1, X_2, ..., X_K. When there are only two data points (or observations), there is only one line that goes through both points – so the process of choosing the line is straightforward. With more than two data points, we cannot usually choose a linear combination of X variables that perfectly predicts all of the observations. In linear regression, we choose the line that minimizes the **residual sum of squares**.

Consider a data set that contains N observations. For each observation, we have information about outcome variable Y and K other variables, X_1, X_2, ..., X_K, that we will use to predict Y. A linear equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K$$

assigns each independent variable X_k a scalar weight β_k. β_0 is the intercept in the equation for the line, though we can also think of this term as the weight on an (implicit) variable X_0 which is a vector of ones. For any vector $\beta = \{\beta_0, \beta_1, \beta_2, \ldots, \beta_K\}$, the predicted value of Y_i is:

$$\hat{Y}_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_K X_{i,K}.$$

In most data sets with $N > 2$, there is no set of linear coefficients β_0, β_1, ..., β_K that will fit the data perfectly. The **residual** is the gap between the predicted value of Y_i and the true, observed value of Y_i, $Y_i - \hat{Y}$. In linear regression, we choose the vector of linear coefficients β that minimizes the residual sum of squares across all N observations.

Taking the partial derivatives of the residual sum of squares

$$RSS(\beta) = \sum_{i=1}^{N} \left(Y_i - \sum_{k=0}^{K} \beta_k X_{i,k} \right)^2$$

with respect to β_0, β_1, ..., β_K yields a system of $K + 1$ first-order conditions.[1] Solving this system of equations gives us unique, explicit analytical expressions characterizing the linear regression coefficients as long as columns of X are not linearly dependent.

[1] Henceforth, we will assume that our linear model includes a constant term, β_0, and thus that X is an $N \times (K + 1)$ matrix that includes X_0, a vector of ones, plus the K independent variables of interest, X_1, X_2, ..., X_K. We will continue to write the regression equation as $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K$, omitting X_0 – but it is implied.

1.1 Regression on a Constant

The simplest possible regression is a regression of outcome Y on (only) a **constant**, X_0 (i.e. a vector of ones). When we regress Y on a constant, we do not use any independent variables that vary across observations, so the predicted value of Y is the same for all observations:

$$\hat{Y}_i = \beta_0.$$

We refer to the value of β_0 that minimizes the sum of squared residuals (in the given data set) as the **estimated regression coefficient** or $\hat{\beta}_0$. In this special case, the value of $\hat{\beta}_0$ that minimizes the sum of squared errors is the mean of Y, $\bar{Y}$.

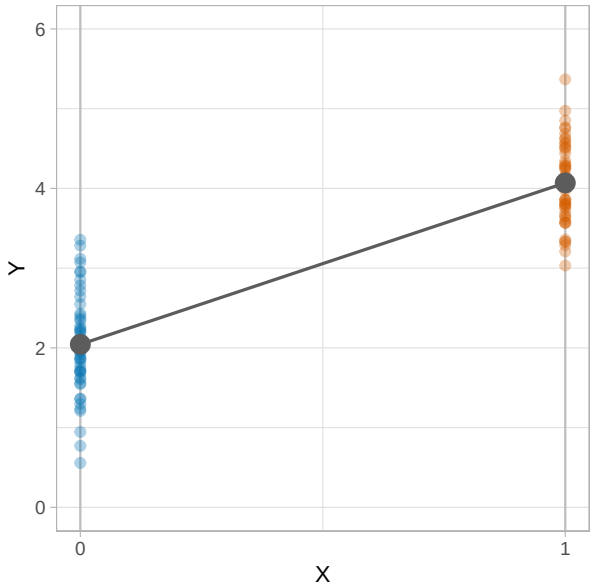
Practice Problem 1 *Show that when $\hat{Y}_i = \beta_0$ and the regression includes only a constant, $\hat{\beta}_0 = \bar{Y}$ minimizes the residual sum of squares.*

1.2 Dummy Variables

1.2.1 Regression on a Single Dummy Variable

Next, consider the case when we regress Y on a single dummy variable X_1 plus a constant X_0. As Figure 1.2.1 illustrates, all the data points fall on one of the two possible values of the dummy variable X_1, 1 or 0. The regression line runs from the cluster of points on the vertical line $X_1 = 0$ to the cluster of points on the vertical line $X_1 = 1$.

Figure 1: Bivariate Regression on a Dummy Variable (axis labels should be X and Y)



For any possible regression line

$$Y = \beta_0 + \beta_1 X_1,$$

there are only two predicted values of Y that are of particular interest: the predicted value for $X_1 = 0$ and the predicted value for $X_1 = 1$. When $X_1 = 0$, the predicted $\hat{Y} = \beta_0$; and when $X_1 = 1$, the predicted $\hat{Y} = \beta_0 + \beta_1$. The chosen regression line which minimizes the residual sum of squares connects these two predicted values.

Key Result 1 *The value of $\hat{Y}$ that minimizes the residual sum of squares across all observations with $X_1 = 0$ is the conditional mean of Y among observations with $X_1 = 0$, which we denote $\bar{Y}_{X_1=0}$, and the value of $\hat{Y}$ that minimizes the residual sum of squares across all observations with $X_1 = 1$ is the conditional mean of Y among observations with $X_1 = 1$, which we denote $\bar{Y}_{X_1=1}$.*

We can establish this formally by solving for the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the residual sum of squares

$$\sum_{i=1}^{N} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right)^2,$$

which yields $\hat{\beta}_0 = \bar{Y}_{X=0}$ and $\hat{\beta}_1 = \bar{Y}_{X=1} - \bar{Y}_{X=0}$.

Practice Problem 2 *Demonstrate Key Result 1 by minimizing the residual sum of squares.*

Practice Problem 3 *Find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the residual sum of squares if X_1 is either 1 or 2 (instead of being either 0 or 1).*

1.3 Mutually Exclusive Dummy Variables

The same logic extends to the case when we regress Y on multiple mutually-exclusive dummy variables, for example if we want to compare a control group to multiple distinct treatments.

Consider the simple case where $K = 2$, so X contains two mutually exclusive dummy variables X_1 and X_2. In this example, we are comparing three groups: a group of observations with $X_1 = X_2 = 0$, which we will refer to as the control group; a group with $X_1 = 1$ and $X_2 = 0$; and a group with $X_1 = 0$ and $X_2 = 1$. The so-called control group is particularly important: if there were not some observations with $X_1 = X_2 = 0$, the sum of X_1 and X_2 would be colinear with X_0, the implicit vector of ones associated with the constant, and there would not be a unique solution for the regression coefficients $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$.

We want to fit the regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

As in the bivariate regression case discussed above, there are only a few predicted values of Y that are of interest, since X_1 and X_2 are dummies and they are never both equal to 1. The predicted $\hat{Y}$ when X_1 and X_2 are both zero is $\hat{\beta}_0$, and when we minimize the residual sum of squares $\hat{\beta}_0$ will be equal to the average value of Y among observations with $X_1 = X_2 = 0$, which we denote $\bar{Y}_{X_1=X_2=0}$. The predicted $\hat{Y}$ when $X_1 = 1$ and $X_2 = 0$ is $\hat{\beta}_0 + \hat{\beta}_1$, and $\hat{\beta}_1 = \bar{Y}_{X_1=1,X_2=0} - \bar{Y}_{X_1=X_2=0}$, the difference between the average value of Y among observations with $X_1 = 1$ (and $X_2 = 0$) and the average value of Y among observations with $X_1 = X_2 = 0$. Similarly, $\hat{\beta}_2 = \bar{Y}_{X_1=0,X_2=1} - \bar{Y}_{X_1=X_2=0}$.

Key Result 2 *Let X_1, X_2, ..., X_K be a set of mutually exclusive dummy variables, and assume that their exists some subset of observations with $X_k = 0$ for all k. When we regress outcome Y on X_1, X_2, ..., X_K plus a constant, estimating the equation*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K,$$

$\hat{\beta}_0 = \bar{Y}_{X_k=0}$, the mean of Y among observations with all of the dummy variables equal to 0, and for $k = 1, \ldots K$, $\hat{\beta}_k = \bar{Y}_{X_k=1} - \bar{Y}_{X_k=0}$.

Practice Problem 4 *Consider two regressions: the regression of Y on mutually exclusive dummies X_1 and X_2 that we discussed above and a bivariate regression of Y on $Z = X_1 + X_2$. Assume both regressions also include a constant. Let $\hat{\beta}_Z$ denote the coefficient on Z from the second regression. Using Key Results 1 and 2, show that*

$$\hat{\beta}_Z = \left(\frac{N_{X_1=1}}{N_{X_1=1} + N_{X_2=1}} \right) \hat{\beta}_1 + \left(\frac{N_{X_2=1}}{N_{X_1=1} + N_{X_2=1}} \right) \hat{\beta}_2$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the coefficients on X_1 and X_2 in the first regression and $N_{X_1=1}$ and $N_{X_2=1}$ are, respectively, the numbers of observations with $X_1 = 1$ and $X_2 = 1$. In other words, show that $\hat{\beta}_Z$ is a weighted average of the regression coefficients from the multivariate regression, where the weights are proportional to the sample size of the treatment arms.

1.3.1 Fixed Effects

We often include a set of mutually exclusive dummy variables as a way of removing variation that is attributable to some categorical variable – for example, location (e.g. state of

residence), time period, or demographic group. We refer to such dummies as **fixed effects**.[2] We do this by including mutually exclusive dummies for all but one of the observed values of the categorical variable.[3] The constant is equal to the mean in the omitted category, and the other coefficients on the fixed effects dummies capture the difference in the mean of Y between the group represented by a particular dummy and the omitted category. As we discuss below, we typically include fixed effects as controls, and in such cases the estimated regression coefficients may not be of interest in their own right, so it does not matter which value of the categorical variable is chosen as the omitted category.

1.4 Interactions Between Dummy Variables

Practice Problem 5 *Suppose $K = 2$ but X_1 and X_2 are not mutually exclusive: assume $X_2 = 1 \Rightarrow X_2 = 1$, but not vice versa. Thus, there are three distinct types of observations in the sample: those with $X_1 = X_2 = 0$, those with $X_1 = 1$ and $X_2 = 0$, and those with $X_1 = X_2 = 1$. Consider two regressions. The first is a regression of Y on X_1 and X_2:*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

The second is a regression of Y on Y on Z_1 and Z_2:

$$Y = \gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2$$

where $Z_1 = X_1(1 - X_2)$ and $Z_2 = X_1 \times X_2$. Characterize the relationship between $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\gamma}_1$, and $\hat{\gamma}_2$.

2 One Continuous Independent Variable

When we fit a bivariate linear regression

$$Y = \beta_0 + \beta_1 X_1,$$

[2] It is not entirely clear whether the term **fixed effects** refers to the set of mutually exclusive dummies representing all but one of the values of a categorical variable or the regression coefficients associated with such a set of dummy variables.

[3] As discussed above, we cannot include a dummy for *all* of the observed values of the categorical variable of interest – because then every observation would have one of the relevant dummies equal to one, so the sum of all the dummies would always be one and hence equal to X_0. When independent variables are colinear in this way, unique values for the regression coefficients cannot be calculated. To address this, we typically include fixed effects for all but one of the observed values of the categorical variable that we wish to control for, though one could also omit the constant term and include dummies for all of the observed values.

we choose the values of $\hat{\beta}_0$ and $\hat{\beta}_0$ that minimize the residual sum of squares,

$$\sum_{i=1}^{N} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i,1} \right)^2.$$

Taking the partial derivatives of this expression with respect to β_0 and β_0 and solving the resulting first-order conditions yields an explicit expression for $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} X_{i,1} \left(Y_i - \bar{Y} \right)}{\sum_{i=1}^{N} X_{i,1} \left(X_{i,1} - \bar{X}_1 \right)},$$

which is equivalent to

$$\frac{COV\left(Y_i, X_{i,1}\right)}{VAR\left(X_{i,1}\right)},$$

and

$$\frac{\sum_{i=1}^{N} Y_i \left(X_{i,1} - \bar{X}_1 \right)}{\sum_{i=1}^{N} X_{i,1} \left(X_{i,1} - \bar{X}_1 \right)},$$

and

$$\sum_{i=1}^{N} w_i Y_i$$

where $w_i = \left(X_{i,1} - \bar{X}_1 \right) / V_X$ and $V_X = \sum_{i=1}^{N} X_{i,1} \left(X_{i,1} - \bar{X}_1 \right)$. These last two formulations emphasize a particularly important fact: the bivariate regression coefficient is a linear combination of the observed values of Y, and the weights in the linear combination are proportional to $X_{i,1} - \bar{X}_1$, the deviations from the mean of X_1. Observations with above-mean values of X_1 receive positive weight while those with below-mean X_1 values receive negative weight. Any observation with X_1 equal to the sample mean receives zero weight in the calculation of the regression coefficient, $\hat{\beta}_1$.

3 Multivariate Regression

Now, we consider the more general case where we want to estimate a regression of Y on X_1, X_2, ..., X_K, fitting the line:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K.$$

There is probably more to say here.

3.1 Adding Fixed Effects

As discussed above, fixed effects control for differences in Y across the observed values of a categorical variable – for example, geographic region or year of birth. Consider a regression of Y on X that also includes fixed effects Z_1, Z_2, ..., Z_M representing all but one of the $M+1$ observed values of some categorical variable. Let $m = 0, 1, \ldots, M$ index the categories defined by the categorical variable: $m = 0$ for observations with $Z_1 = Z_2 = \ldots = Z_M = 0$, and for $m = 1, \ldots, M$ the category m refers to the observations with $Z_m = 1$ and $Z_{\neq m} = 0$.

When we regress Y on **only** the fixed effects, we have seen that the regression coefficient $\hat{\gamma}_m$ captures the difference in the mean of Y between the group of observations with $Z_m = 1$ (category m) and the group chosen as the omitted category (category 0). Fixed effects function in a similar way in a multivariate regression, but they capture the difference in the mean of Y after controlling for X (in a way that we make precise below).

We are often more interested in understanding how the coefficient on X_1 changes with the inclusion of fixed effects that in the fixed effects coefficients themselves. When we include fixed effects for the observed values of a categorical variable, that is equivalent to subtracting off category-specific means of both Y and X. For $m = 0, 1, \ldots, M$, let $\bar{Y}_m$ and $\bar{X}_m$ denote the means of Y and X, respectively, in category m. The fixed effects regression of Y on X and dummy variables Z_1, Z_2, ..., Z_M,

$$Y = \beta_0 + \beta_1 X_1 + \gamma_1 Z_1 + \ldots + \gamma_M Z_M,$$

yields the same estimate of $\hat{\beta}_1$ as a bivariate regression of $\tilde{Y}$ on $\tilde{X}$ where

$$\tilde{Y}_i = Y_i - \bar{Y}_m$$

and

$$\tilde{X}_i = Y_i - \bar{X}_m.$$

The standard error of $\hat{\beta}_1$ is also mechanically equal across the two regressions.

We will sometimes refer to $\tilde{Y}$ and $\tilde{X}$ as normalized values of Y and X. We can also refer to them as **residualized** values of Y and X. The reason for this is that $\tilde{Y}$ and $\tilde{X}$ are the residuals from regressions of Y and X on the set of fixed effects, Z_1, Z_2, ..., Z_M (plus a constant). There are, therefore, three equivalent approaches to calculating $\hat{\beta}_1$:

1. Regress Y on X while including fixed effects Z_1, Z_2, ..., Z_M;
2. Regress Y and X on the fixed effects Z_1, ..., Z_M and define $\tilde{Y}$ and $\tilde{X}$ as the residuals from those regressions, and then regress $\tilde{Y}$ and $\tilde{X}$; or

3. Calculate category-specific means, construct normalized variables $\tilde{Y}$ and $\tilde{X}$ by subtracting of the category-specific means, and then regress $\tilde{Y}$ and $\tilde{X}$.

3.1.1 When X Is a Dummy Variable

3.2 The Frisch-Waugh-Lovell Theorem

The Frisch-Waugh-Lovell Theorem states that the coefficient $\hat{\beta}_1$ from the regression

$$Y = \beta_0 + \beta_1 X_1 + \gamma_1 Z_1 + \ldots + \gamma_M Z_M,$$

is equal to the coefficient on $\tilde{X}$ from a regression of $\tilde{Y}$ and $\tilde{X}$ where $\tilde{Y}$ and $\tilde{X}$ are the residuals from regressions of Y and X on Z_1, ..., Z_M. Thus, the relationship between the first and second approaches to estimating the coefficient on X in a regression including fixed effects (described above) is not specific to the case where Z_1, ..., Z_M are dummy variables. This is, instead, a direct result of the application of the Frisch-Waugh-Lovell theorem to the case where all but one of the variables are dummies.
