Regression Analysis

Linear regression is one of the most widely used tools in statistics. Suppose we were jobless college students interested in finding out how big (or small) our salaries would be 20 years from now. There's no way to pin down this number for sure, but we know that there are many factors that contribute to how much money a college graduate will make. For example, a naive observation (but a good starting point) is that students with higher GPAs earn more money 20 years from now. In this case, we assume that there is some true distribution that governs the behavior of the random variables

$$X \doteq \text{GPA}$$
$$Y \doteq \text{Salary 20 years from now}$$

where X and Y are not independent. In this case, we call X *a predictor* of Y. Another way that people refer to X and Y are as independent and dependent variables (nothing to do with *probabilistic* independence), since Y *depends* on X. In the following sections, we set up a linear model to describe the relationship between Y and X, which we can then use to predict our own future salary, based on some sample data.

Ordinary Least Squares

The Linear Model

Since X and Y seem to have some relationship, it would be reasonable to assume that given some value of X, we have a better idea about what Y is. Intuitively, we would expect students with higher GPAs to have a larger future salary, so we could model the relationship between X and Y using a line. That is, for some real numbers w_0 and w_1,

$$Y = w_0 + w_1 X$$

This is our familiar $y = mx + b$ relationship from high school algebra, but with different names for m and b.

Note that this is an extremely simple model that is likely to miss most of the nuances in predicting someone's salary 20 years from now. There are in fact many more predictors than someone's GPA that affect their future salary. Also notice that we can express the above relationship using the following vector form.

$$Y = \mathbf{X} \cdot \mathbf{w} \doteq (1, X) \cdot (w_0, w_1)$$

where "$\cdot$" represents the dot product. This form is why the method is called *linear* regression.

Exercise 0.0.5. *Verify the function $f : ^2 \to$ defined by*

$$f(\boldsymbol{w}) = \boldsymbol{X} \cdot \boldsymbol{w}$$

is linear in $\boldsymbol{w}$.

Solution. Remember that the term *linear* was used to describe the "Expectation" operator. The two conditions we need to check are

(a) For any vectors $\mathbf{w}, \mathbf{v} \in ^2$, we have

$$f(\mathbf{u} + \mathbf{v}) = f(\mathbf{w}) + f(\mathbf{v})$$

(b) For any vector $\mathbf{w} \in ^2$ and constant $c \in$,

$$f(c\mathbf{w}) = cf(\mathbf{w})$$

To show (a), we know that $\mathbf{w}$ and $\mathbf{v}$ are vectors of the form

$$\mathbf{w} \doteq (w_0, w_1)$$
$$\mathbf{v} \doteq (v_0, v_1)$$

so that

$$\begin{aligned}
f(\mathbf{w} + \mathbf{v}) &= f((w_0, w_1) + (v_0, v_1)) \\
&= f((w_0 + v_0, w_1 + v_1)) \\
&= \mathbf{X} \cdot (w_0 + v_0, w_1 + v_1) \\
&= (1, X) \cdot (w_0 + v_0, w_1 + v_1) && \text{(Definition of } \mathbf{X}\text{)} \\
&= (w_0 + v_0) + X(w_1 + v_1) && \text{(Definition of dot product)} \\
&= (w_0 + Xw_1) + (v_0 + Xv_1) && \text{(Rearranging)} \\
&= \mathbf{X} \cdot \mathbf{w} + \mathbf{X} \cdot \mathbf{v} \\
&= f(\mathbf{w}) + f(\mathbf{v})
\end{aligned}$$

For (b), observe that if $\mathbf{w} \in^2$ and $c \in$,

$$\begin{aligned} f(c\mathbf{w}) &= \mathbf{X} \cdot (cw_0, cw_1) \\ &= (1, X) \cdot (cw_0, cw_1) \\ &= cw_0 + cw_1 X \\ &= c(w_0 + w_1 X) \\ &= c\mathbf{X} \cdot \mathbf{w} \\ &= cf(\mathbf{w}) \end{aligned}$$

This completes the proof. □

The observation that f is linear in $\mathbf{w}$ as opposed to linear in $\mathbf{X}$ is an extremely important distinction. Take a moment and let it sink in. This means that we can transform $\mathbf{X}$ in crazy nonlinear ways while maintaining the linearity of this problem. For example, the proof above implies that we could replace $\mathbf{X}$ with $\log(\mathbf{X})$ or $\sin(\mathbf{X})$ and we still have a linear relationship between Y and $\mathbf{w}$.

The above example not realistic in the sense that its extremely unlikely that if we sampled n college graduates and their actual salaries 20 years after college, all their GPAs fall on a perfect line when plotted against their salaries. That is, if we took n sample points, written

$$\text{Sample} = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$$

and plotted these points in the plane with "GPA" on the x-axis and "Salary" on the y-axis, the points would almost surely not fall on a perfect line. As a result, we introduce an error term ϵ, so that

$$Y = \mathbf{X} \cdot \mathbf{w} + \epsilon \tag{5}$$

All of this hasn't yet told us how to predict our salaries 20 years from now using only our GPA. The subject of the following section gives a method for determining the best choice for w_0 and w_1 given some sample data. Using these values, we could plug in the vector $(1, \text{our GPA})$ for $\mathbf{X}$ in equation (2) and find a corresponding predicted salary Y (within some error ϵ).

Method of Least Squares

Our current model for X and Y is the relationship

$$Y = \mathbf{X} \cdot \mathbf{w} + \epsilon$$

where ϵ is some error term. Suppose we go out and ask a bunch of 50 year olds for their college GPAs and their salaries 20 years out of college. We can pair these quantities and record this sample data as

$$\text{Data} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

Remember that we assume these samples come from the relationship

$$y_i = (1, x_i) \cdot (w_0, w_1) + \epsilon_i$$

and we are trying to find w_0 and w_1 to best fit the data. What do we mean by "best fit"? The notion we use is to find w_0 and w_1 that minimize the sum of squared errors $\sum_{i=1}^{n} \epsilon_i^2$. Rearranging the above equation for ϵ_i, we can rewrite this sum of squared errors as

$$E(\mathbf{w}) \doteq \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i \cdot \mathbf{w})^2$$

where the vector $\mathbf{x}_i$ is shorthand for $(1, x_i)$. As we can see above, the error E is a function of $\mathbf{w}$. In order to minimize the squared error, we minimize the function E with respect to $\mathbf{w}$. E is a function of both w_0 and w_1. In order to minimize E with respect to these values, we need to take partial derivatives with respect to w_0 and w_1. This derivation can be tricky in keeping track of all the indices so the details are omitted. If we differentiate E with respect to w_0 and w_1, we eventually find that minimizing $\mathbf{w}$ can be expressed in matrix form as

$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \left(\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

This can be written in the following concise form,

$$[box = \boxed{\ }align* \mathbf{w}^T = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{y}$$

where $\mathbf{D}$ is the matrix made by stacking the sample vectors $\mathbf{x}_i$,

$$\mathbf{D} \doteq \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

and $\mathbf{y}$ is the column vector made by stacking the observations y_i,

$$\mathbf{y} \doteq \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

A sketch of the derivation using matrices is given in the following section for those who cringed at the sentence "This derivation can be tricky in keeping track of all the indices so the details are omitted." Some familiarity with linear algebra will also be helpful going through the following derivation.

Linear Algebra Derivation

We can write the error function E as the squared norm of the matrix difference $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{D}\mathbf{w}^T$.

$$E(\mathbf{w}) = \|\mathbf{y} - \mathbf{D}\mathbf{w}^T\|^2 = \|\mathbf{D}\mathbf{w}^T - \mathbf{y}\|^2$$

Differentiating with respect to $\mathbf{w}$, the two comes down from the exponent by the power rule, and we multiply by $\mathbf{D}^T$ to account for the chain rule. We get

$$\nabla E = 2\mathbf{D}^T(\mathbf{D}\mathbf{w}^T - \mathbf{y})$$

We set $\nabla E = \mathbf{0}$ (we use a bold "o" since it is actually a vector of zeros) so that

$$2\mathbf{D}^T(\mathbf{D}\mathbf{w}^T - \mathbf{y}) = \mathbf{0}$$

Dividing by 2 on both sides and distributing the $\mathbf{D}^T$ across the difference gives

$$\mathbf{D}^T\mathbf{D}\mathbf{w}^T - \mathbf{D}^T\mathbf{y} = \mathbf{0}$$

Adding $\mathbf{D}^T\mathbf{y}$ to both sides gives

$$\mathbf{D}^T\mathbf{D}\mathbf{w}^T = \mathbf{D}^T\mathbf{y}$$

Multiplying on the left by the inverse of the matrix $\mathbf{D}^T\mathbf{D}$ on both sides of the above equation finally yields the famous linear regression formula,

$$\mathbf{w}^T = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{y}$$

Now, assuming salaries are related to college GPAs according to the relation

$$Y = w_0 + w_1 X + \epsilon,$$

we can plug in our GPA for X, and our optimal w_0 and w_1 to find the corresponding predicted salary Y, give or take some error ϵ. Note that since we chose w_0 and w_1 to minimize the errors, it is likely that the corresponding error for our GPA and predicted salary is small (we assume that our (GPA, Salary) pair come from the same "true" distribution as our samples).

Generalization

Our above example is a simplistic one, relying on the very naive assumption that salary is determined solely by college GPA. In fact

there are many factors which influence someones salary. For example, earnings could also be related to the salaries of the person's parents, as students with more wealthy parents are likely to have more opportunities than those who come from a less wealthy background. In this case, there are more predictors than just GPA. We could extend the relationship to

$$Y = w_0 + w_1 X_1 + w_2 X_2 + w_3 X_3 + \epsilon$$

where X_1, X_2, and X_3 are the GPA, Parent 1 salary, and Parent 2 salary respectively.

By now it is clear that we can extend this approach to accomodate an arbitrary number of predictors $X_1, \ldots, X_d$ by modifying the relationship so that

$$Y = w_0 + w_1 X_1 + w_2 X_2 + \cdots + w_d X_d + \epsilon$$

or more concisely,

$$Y = \mathbf{X} \cdot \mathbf{w} + \epsilon$$

where the vectors $\mathbf{X}, \mathbf{w} \in^{d+1}$ are the extensions

$$\mathbf{X} \doteq (1, X_1, X_2, \ldots, X_d)$$
$$\mathbf{w} \doteq (w_0, w_1, w_2, \ldots, w_d)$$

the parameters w_i can be thought of as "weights" since the larger any particular weight is, the more influence its attached predictor has in the above equation. Recall that in Exercise 6.1, we verified the function $f(\mathbf{w}) = \mathbf{X} \cdot \mathbf{w}$ was linear in the vector $\mathbf{w} \in^2$. In fact, when we extend $\mathbf{w}$ to be a vector in $^{d+1}$, the function $f(\mathbf{w}) = \mathbf{X} \cdot \mathbf{w}$ is still linear in $\mathbf{w}$.

The linear regression formula still holds, i.e. that the optimal weights are given by

$$\mathbf{w}^T = (\mathbf{D}^T \mathbf{D})^{-1} \mathbf{D}^T \mathbf{y}$$

where the matrix $\mathbf{D}$ is still constructed by stacking the observed samples,

$$\mathbf{D} \doteq \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(d)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(d)} \\ \vdots & & & & \\ 1 & x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(d)} \end{bmatrix}$$

where the i^{th} sample is written

$$\mathbf{x}_i \doteq (1, x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(d)})$$

Correlation

Throughout the past chapters, we often made the assumption that two random variables are independent in various exercises and methods. In reality, most random variables are not actually independent. In this section we give some measures to quantify how "related" a collection of random variables are.

The example in the Linear Regression chapter began with the observation that GPAs are positively correlated with future salaries. That is, we assumed that as college GPA increased, future salary also increased. Qualitatively, this was enough to motivate the problem of regression. However, there were other predictors that contributed to the future salary, some of which were also positively correlated to the projected salary. The fact that some variables contributed "more positively" than others was manifested in the size of the weights that were attached to the variables in the equation $Y = \mathbf{X} \cdot \mathbf{w} + \epsilon$. If one X_i were more predictive of Y than another, then its corresponding weight was larger. In the following section we examine the covariance of two random variables, which is another attempt to quantify the relationship between random variables.

Covariance

Suppose we have two random variables X and Y, not necessarily independent, and we want to quantify their relationship with a number. This number should satisfy two basic requirements.

(a) The number should be positive when X and Y increase/decrease together.

(b) It should be negative when one of X or Y decreases while the other increases.

Consider the following random variable.

$$(X - EX)(Y - EY)$$

Consider the possible realizations of the random variables $X = x$ and $Y = y$. The collection of these pairs is the sample space Ω. We can think of the outcomes of sampling an X and a Y as pairs $(x, y) \in \Omega$. Suppose the probability distribution governing X and Y on Ω assigns most of the probability mass on the pairs (x, y) such that $x > EX$ and $y > EY$. In this case, the random variable $(X - EX)(Y - EY)$ is likely to be positive most of the time. Similarly, if more mass were placed on pairs (x, y) such that $x < EX$ and $y < EY$, the product $(X - EX)(Y - EY)$ would be a negative number times a negative

number, which means it would still be positive most of the time. Hence the product $(X - EX)(Y - EY)$ being positive is indicative of X and Y being mutually more positive or mutually more negative.

By similar reasoning, the product $(X - EX)(Y - EY)$ is more often negative if the distribution assigns more mass to pairs (x, y) that have $x < EX$ and $y > EY$, or that satisfy $x > EX$ and $y < EY$. In either case, the product $(X - EX)(Y - EY)$ will be a product of a positive and negative number, which is negative.

We are almost done. Remember at the beginning of this discussion we were searching for a number to summarize a relationship between X and Y that satisfied the requirements (a) and (b). But $(X - EX)(Y - EY)$ is a random variable, (that is, a function mapping Ω to) not a number. To get a number, we take the expectation. Finally we arrive at the definition of covariance.

Definition 0.0.19. *The **covariance** of two random variables X and Y, written $Cov(X, Y)$, is defined*

$$Cov(X, Y) = E[(X - EX)(Y - EY)]$$

This definition may look similar to the definition for variance of a random variable X, except we replace one of the terms in the product with the difference $Y - EY$. Similar to Proposition 2.11 (c), there is another useful form of the covariance.

Proposition 0.0.3. *Let X and Y be two random variables with means EX and EY respectively. Then*

$$Cov(X, Y) = E[XY] - E[X]E[Y]$$

Proof. By the definition of covariance, we can foil the product inside the expectation to get

$$\begin{aligned} \text{Cov}(X, Y) &= E[XY - XEY - YEX + EXEY] \\ &= E[XY] - E[XEY] - E[YEX] + E[EXEY] \quad \text{(linearity of } E\text{)} \\ &= E[XY] - EYEX - EXEY + EXEY \quad \text{(linearity of } E\text{)} \\ &= E[XY] - EXEY \end{aligned}$$

□

The Correlation Coefficient

The covariance quantity we just defined satisfies conditions (a) and (b), but can become arbitrarily large depending on the distribution of X and Y. Thus comparing covariances between different pairs of random variables can be tricky. To combat this, we normalize the quantity to be between -1 and 1. The normalized quantity is called the correlation, defined below.

Definition 0.0.20. *The **correlation coefficient** between two random variables X and Y with standard deviations σ_x and σ_y, is denoted ρ and is defined*

$$\rho_{xy} = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

Exercise 0.0.6. *Verify that for given random variables X and Y, the correlation ρ_{xy} lies between -1 and 1.*

Heuristic. The rigorous proof for this fact requires us to view X and Y as elements in an infinite-dimensional normed vector space and apply the Cauchy Schwartz inequality to the quantity $E[(X - EX)(Y - EY)]$. Since we haven't mentioned any of these terms, we instead try to understand the result using a less fancy heuristic argument.

Given a random variable X, the first question we ask is,

What is the random variable *most positively* correlated with X?

The random variable that correlates most positively with X should increase *exactly* with X and decrease *exactly* with X. The only random variable that accomplishes this feat is X itself. This implies that the correlation coefficient between X and any random variable Y is less than that between X and itself. That is,

$$\rho_{xy} \leq \rho_{xx} = \frac{\text{Cov}(X, X)}{\sigma_x \sigma_x} = \frac{\text{Var}(X)}{\text{Var}(X)} = 1$$

By now you've probably guessed the second question we need to ask.

What is the random variable *least positively* correlated with X?

In other words, we are looking for a random variable with which the correlation between X and this random variable is the most negative it can be. This random variable should increase *exactly* as X decreases, and it should also decrease *exactly* as X increases. The candidate that comes to mind is $-X$. This would imply that the correlation coefficient between X and any random variable Y is greater than that between X and $-X$.

This implies that

$$\rho_{xy} \geq \rho_{x,-x} = \frac{\text{Cov}(X, -X)}{\sigma_x \sigma_{-x}}$$

By Proposition 6.3, the expression on the right becomes

$$= \frac{E[X(-X)] - E[X]E[-X]}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(-X)}} = \frac{-(E[X^2] - (EX)^2)}{\text{Var}(X)} = \frac{-\text{Var}(X)}{\text{Var}(X)} = -1$$

Hence, we conclude that $-1 \leq \rho_{xy} \leq 1$. □

Interpretation of Correlation

The correlation coefficient between two random variables X and Y can be understood by plotting samples of X and Y in the plane. Suppose we sample from the distribution on X and Y and get

$$\text{Sample} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

There are three possibilities.

Case 1: $\rho_{xy} > 0$. We said that this corresponds to X and Y increasing mutually or decreasing mutually. If this is the case, then if we took n to be huge (taking many samples) and plotted the observations, the best fit line would have a positive slope. In the extreme case if $\rho_{xy} = 1$, the samples (X_i, Y_i) would all fall perfectly on a line with slope 1.

Case 2: $\rho_{xy} = 0$. This corresponds to X and Y having no *observable* relationship. However, this does not necessarily mean that X and Y have no relationship whatsoever. It just means that the measure we are using the quantify their relative spread (the correlation) doesn't capture the underlying relationship. We'll see an example of this later. In terms of the plot, the samples (X_i, Y_i) would look scattered on the ^{2} plane with no apparent pattern.

Case 3: $\rho_{xy} < 0$. We said that this case corresponds to one of X or Y decreasing while the other increases. If this were the case, then the best fit line is likely to have a negative slope. In the extreme case when $\rho_{xy} = -1$, all samples fall perfectly on a line with slope -1.

Independence vs Zero Correlation

There is a commonly misunderstood distinction between the following two statements.

1. "X and Y are independent random variables."

2. "The correlation coefficient between X and Y is 0."

The following statement is always true.

Proposition 0.0.4. *If X and Y are independent random variables, then $\rho_{xy} = 0$.*

The converse is not. That is, $\rho_{xy} = 0$ does not *necessarily* imply that X and Y are independent.

In "Case 2" of the previous section, we hinted that even though $\rho_{xy} = 0$ corresponded to X and Y having no observable relationship, there could still be some underlying relationship between the random variables, i.e. X and Y are still not independent. First let's prove Proposition 6.6

Proof. Suppose X and Y are independent. Then functions of X and Y are independent. In particular, the functions

$$f(X) \doteq X - EX$$
$$g(Y) \doteq Y - EY$$

are independent. By the definition of correlation,

$$\begin{aligned}
\rho_{xy} &= \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y} \\
&= \frac{E[(X-EX)(Y-EY)]}{\sigma_x \sigma_y} \\
&= \frac{E[f(X)g(Y)]}{\sigma_x \sigma_y} \\
&= \frac{E[f(X)]E[g(Y)]}{\sigma_x \sigma_y} && \text{(independence of } f(X) \text{ and } g(Y)) \\
&= \frac{0 \cdot 0}{\sigma_x \sigma_y} && (E[f(X)] = E(X - EX) = 0) \\
&= 0
\end{aligned}$$

Hence if X and Y are independent, $\rho_{xy} = 0$. □

Now let's see an example where the converse does not hold. That is, an example of two random variables X and Y such that $\rho_{xy} = 0$, but X and Y are *not* independent.

Example 0.0.11. *Suppose X is a discrete random variable taking on values in the set $\{-1, 0, 1\}$, each with probability $\frac{1}{3}$. Now consider the random variable $|X|$. These two random variables are clearly not independent, since once we know the value of X, we know the value of $|X|$. However, we can show that X and $|X|$ are uncorrelated. By the definition of correlation and Proposition 6.3,*

$$\rho_{x,|x|} = \frac{E(X \cdot |X|) - EX \cdot E|X|}{\sigma_x \sigma_{|x|}} \tag{6}$$

Let's compute the numerator. By looking at the distribution of X, we can see that the product $X \cdot |X|$ can only take on three possible values. If $X = 0$, then $|X| = 0$ so $X \cdot |X| = 0$. If $X = -1$, then $|X| = 1$ and $X \cdot |X| = -1$. Finally if $X = 1$, then $|X| = 1$ and $X \cdot |X| = 1$. Each of these cases occur with probability $\frac{1}{3}$. Hence,

$$X \cdot |X| \sim \mathit{Uniform}\{-1, 0, 1\}$$

It follows that the expectation of $X \cdot |X|$ is

$$E(X \cdot |X|) = \frac{1}{3} \cdot (-1) + \frac{1}{3} \cdot (0) + \frac{1}{3} \cdot (1) = 0.$$

Also by the definition of expectation,

$$E[X] = \frac{1}{3} \cdot (-1) + \frac{1}{3} \cdot (0) + \frac{1}{3} \cdot (1) = 0.$$

Plugging these values into the numerator in expression (3), we get $\rho_{x,|x|} = 0$. *Thus, the two random variables* X *and* $|X|$ *are certainly not always equal, they are not independent, and yet they have correlation 0. It is important to keep in mind that zero correlation does not necessarily imply independence.*