6 Regression Analysis

One of the fundamental themes in statistics is using observed data to make an inference about some underlying "truth". For example, in the Estimation sections, we observed some samples and used them to estimate parameters (points) and intervals (confidence intervals). Regression fits nicely into this framework. Instead of estimating points or intervals, regression attempts to estimate a relationship between variables using only observed samples (assumed to be generated according to the "true" relationship).

6.1 Linear Regression

Linear regression is one of the most widely used tools in statistics. Suppose we were 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

$$\begin{aligned} X &\doteq \text{GPA} \\ Y &\doteq \text{Salary 20 years from now.} \end{aligned}$$

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.

6.1.1 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 6.1. Verify the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

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

is linear in $\mathbf{w}$.

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 $\mathbf{w}, \mathbf{v} \in \mathbb{R}^2$, we have

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

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

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

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 $\mathbf{w} \in \mathbb{R}^2$ and $c \in \mathbb{R}$,

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

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. This means that we can transform $\mathbf{X}$ in nonlinear ways while maintaining the linearity of this problem (linear problems are much easier to solve than nonlinear ones!). For example, the proof above implies that we could replace $\mathbf{X}$ with $\log(\mathbf{X})$ or $\sin(\mathbf{X})$ and we would still have a linear relationship between Y and $\mathbf{w}$.

The above example is 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

$$Y = \mathbf{X} \cdot \mathbf{w} + \varepsilon. \tag{2}$$

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).

6.1.2 Method of Least Squares

Our current model for X and Y is the relationship

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

where ε is some error term. Suppose we go out and ask n people 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

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

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 $\sum_{i=1}^{n} \varepsilon_i^2$. Rearranging the above equation for ε_i, we can rewrite this sum of squared errors as

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

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. The details of this derivation involve a lot of keeping track of indices, so they are omitted, but a clean matrix algebra justification is given in the following section (Section 6.1.3).

If we differentiate E with respect to w_0 and w_1, we eventually find that the minimizing $\mathbf{w}$ can be expressed in matrix form as

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

This can be written in the following concise form,

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

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. Some familiarity with linear algebra will also be helpful going through the following derivation.

6.1.3 Linear Algebra Derivation

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

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

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 "0" 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

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

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).

6.1.4 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

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

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

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

or more concisely,

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

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

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

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 $\mathbf{w} \in \mathbb{R}^2$. In fact, when we extend $\mathbf{w}$ to be a vector in $\mathbb{R}^{d+1}$, 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

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