247 DAY 20

6.6: Linear Models

In science, social science, and engineering, we are often interested in how several quantities measuring various attributes of some object of interest interact or depend on one another. In the simplest case, one is interested in how two such variables interact.

Example. For example, we may be interested in how well a first-year CSULB student's GPA is predicted by her or his SAT score. This type of data can easily be displayed using a *scatterplot* (show). To further simplify the analysis, let's assume that we are interested in the *linear* relationship between the values, so that a student's "predicted" GPA is a linear function of that student's SAT score. That is, let x_i stand for the i^{th} student's SAT score and y_i for that student's GPA. The model we're exploring is represented by the equation $y = \beta_0 + \beta_1 x$, where β_0 and β_1 are to be determined. As in the last section, unless the data points all happen to lie upon a line, there is no literally correct solution for β_0 and β_1. We will typically refer to the value y_i as an *observed* value and $\beta_0 + \beta_1 x_i$ as the *predicted* value for y_i. We will now turn this problem into a least-squares problem.

How do we do this? Notice that the above linear equations applied to the finite set of ordered pairs (x_i, y_i) give the system of equations

$$\begin{aligned} \beta_0 + \beta_1 x_1 &= y_1 \\ \beta_0 + \beta_1 x_2 &= y_2 \\ &\vdots \\ \beta_0 + \beta_1 x_n &= y_n \end{aligned}$$

which leads to the matrix equation $X\boldsymbol{\beta} = \mathbf{y}$, where $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$, $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$, and $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$. X is called the *design matrix* (really an engineering term), $\boldsymbol{\beta}$ the *parameter vector*, and $\mathbf{y}$ the *vector of observations.* Does everybody understand where the column of 1s in the matrix X comes from? Note that the variables here are β_0 and β_1!

Now, if we solve this least-squares problem, what does the solution give us in terms of the original graph? Well, we're minimizing the distance from the vector $\mathbf{y}$

to the vector $X\boldsymbol{\beta} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_1 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix}$, which is the sum of the squares of the vertical distances from the observed values y_i to the predicted values along the line. The vector $\boldsymbol{\beta}$ minimizing this quantity is called the *least-squares regression line.*

(Question: What would you do if you wanted to minimize the sum of the squares of the *horizontal* distances to the line instead?)

Here are a few actual data points from some undisclosed state university:

SAT	GPA
1232	3.52
1070	2.91
1086	2.4
1287	3.47
1130	3.47
1048	2.37
1121	2.4

Recall that the *normal equations* for this least-squares problem come from the matrix equation $X^T X\boldsymbol{\beta} = X^T\mathbf{y}$. I used Maple to calculate these quantities and got the equation (rounding as necessary)

$$\begin{bmatrix} 7 & 7974 \\ 7974 & 9130334 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 20.54 \\ 23618 \end{bmatrix}.$$

This equation I solved to get $\hat{\mathbf{y}} = \begin{bmatrix} -2.42 \\ 0.0047 \end{bmatrix}$. In other words, the predicted GPA value for a given SAT score x is $-2.4 + .0047x$. As a specific example, for an SAT score of 1200, it predicts a GPA of 3.22.

In fact, we can change the design matrix to fit models other than straight lines and to use more than one predictor variable.

Example. Suppose that we think that the dependency between two variables is actually parabolic rather than linear, so that our model becomes $y = \beta_0 + \beta_1 x + \beta_2 x^2$. Our design matrix X would get more complicated; now $X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}$. Our parameter vector $\boldsymbol{\beta}$ is now the 3-vector $\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$, and our vector $\mathbf{y}$ of observations is unchanged.

Note that this is still called a linear model! This is because we're determining it using linear equations in the variables β_i. We could also do this for higher-degree polynomials, introducing more coefficients, or other equations. Look at the Practice Problem on p 425 for a nice illustration of how linear models can be used.

Example. One can also use linear models for *multivariate regression*, in which there are two or more predictor variables. In the above example, for instance, we might have wanted to consider SAT math and verbal scores as separate variables, say u and v. Then our model would look like $y = \beta_0 + \beta_1 u + \beta_2 v$, our design matrix X becomes $\begin{bmatrix} 1 & u_1 & v_1 \\ 1 & u_2 & v_2 \\ \vdots & \vdots & \vdots \\ 1 & u_n & v_n \end{bmatrix}$, and $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$.

Of course, we could have had more variables and/or more complicated functions of each variable. The idea of a linear model is very flexible!

If time: Problems 5 and 7, p. 425.

7.1: Diagonalization of Symmetric Matrices

Recall: A matrix A is *diagonalizable* if there is an invertible matrix P and a diagonal matrix D such that $D = P^{-1}AP$. In this case, the columns of P are eigenvectors of A, and the entries on the diagonal of D are the corresponding eigenvalues. Once again, the matrix D represents the same linear transformation as A, but with respect to the basis of eigenvectors $\mathcal{P}$ that form the columns of P.

Also recall that if U is an $m \times n$ matrix with orthonormal columns, $U^T U = I_n$. In particular, if U is a square matrix, U is called an *orthogonal* matrix and $U^T = U^{-1}$.

Definition. A matrix A is *symmetric* if $A^T = A$. Notice that a symmetric matrix needs to be square!

Example. Let $A = \begin{bmatrix} 2 & 2 \\ 2 & -1 \end{bmatrix}$. Notice that A is symmetric. It's easy to calculate that 3 is an eigenvalue of A, with corresponding eigenvector $\mathbf{v}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, and that -2 is the other eigenvalue with eigenvector $\mathbf{v}_2 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$. Note that these eigenvectors are orthogonal.

The fact that the eigenvectors were orthogonal in the above example is no coincidence!

Theorem. (Theorem 1, p. 450) Let A be an $n \times n$ symmetric matrix, and let λ_1 and λ_2 be two different eigenvalues of A, with associated eigenvectors $\mathbf{v}_1$ and v_2. Then v_1 and $\mathbf{v}_2$ are orthogonal.

Proof.

$$\begin{aligned} \lambda_1(\mathbf{v}_1 \cdot \mathbf{v}_2) &= (\lambda_1 \mathbf{v}_1) \cdot \mathbf{v}_2 = (\lambda_1 \mathbf{v}_1)^T \mathbf{v}_2 \\ &= (A\mathbf{v}_1)^T \mathbf{v}_2 = (\mathbf{v}_1^T A^T)\mathbf{v}_2 = \mathbf{v}_1^T (A^T \mathbf{v}_2) = \mathbf{v}_1^T (A\mathbf{v}_2) \\ &= v_1^T (\lambda_2 \mathbf{v}_2) = \lambda_2 (\mathbf{v}_1^T \mathbf{v}_2) = \lambda_2 (\mathbf{v}_1 \cdot \mathbf{v}_2). \end{aligned}$$

Since $\lambda_1(\mathbf{v}_1 \cdot \mathbf{v}_2) = \lambda_2(\mathbf{v}_1 \cdot \mathbf{v}_2)$, and $\lambda_1 \neq \lambda_2$, $\mathbf{v}_1 \cdot \mathbf{v}_2$ must be equal to 0. □

Example. To continue the above example, if we use the unit eigenvectors $\mathbf{u}_1 = \begin{bmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{bmatrix}$, $\mathbf{u}_2 = \begin{bmatrix} \frac{-1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix}$ to form our diagonalizing matrix P, then P will be the orthogonal matrix $\begin{bmatrix} \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix}$, and P^{-1} can be calculated as P^T.

Definition. An $n \times n$ matrix A is *orthogonally diagonalizable* if there exists an orthogonal matrix P and a diagonal matrix D such that $D = P^{-1}AP = P^TAP$.

Thus the matrix in the above example is orthogonally diagonalizable. In fact, by the above theorem, as long as a symmetric matrix has a basis of eigenvectors, it will be orthogonally diagonalizable, since for each individual eigenspace it is always possible to find an orthonormal basis.

To go the other direction, what can you say about A if it is orthogonally diagonalizable? Well, then we can write $A = PDP^T$, so

$$A^T = (PDP^T)^T = (P^T)^T D^T P^T = PDP^T = A.$$

In other words, A is automatically symmetric!

In fact, look at Theorem 2. The above discussion proves one half of the "if and only if"; the other half is more difficult and would probably take about a class period. I won't do this proof; if you're interested, there are references after Theorem 3 on p. 452, which is just a restatement of Theorem 2 in detail.

Just a little note on spectral decomposition: If A is symmetric, we can write $A = PDP^T$, where P is orthogonal and D is diagonal. Then, if we write $P = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_n]$ and the corresponding eigenvalues to be $\lambda_1, \lambda_2, ...\lambda_n$,

$$A = PDP^T = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_n] \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \vdots \\ \mathbf{u}_n^T \end{bmatrix}$$

and use the definition of matrix-vector product to write this as

$$[\lambda_1 \mathbf{u}_1, \lambda_2 \mathbf{u}_2, ..., \lambda_n \mathbf{u}_n] \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \vdots \\ \mathbf{u}_n^T \end{bmatrix}.$$

Just by looking at how the product works, it's not hard to see that this is equal to

$$\lambda_1 \mathbf{u}_1 \mathbf{u}_1^T + \lambda_2 \mathbf{u}_2 \mathbf{u}_2^T + ... + \lambda_n \mathbf{u}_n \mathbf{u}_n^T;$$

the result they cite is the "column-row rule" from Section 2.4, which we skipped, but it's a pretty simple fact.

This expresses A as a weighted sum, according to the corresponding eigenvalues of the matrices $\mathbf{u}_i \mathbf{u}_i^T$. What kind of matrices are they? Since each $\mathbf{u}_i$ is a unit vector, it makes a matrix with one orthonormal column, and therefore $\mathbf{u}_i \mathbf{u}_i^T$ is the matrix for the projection onto the span of $\mathbf{u}_i$.

This sum is called the *spectral decomposition* for A.

Groups

- Work out the spectral decomposition of the matrix $A = \begin{bmatrix} 2 & 2 \\ 2 & -1 \end{bmatrix}$ from the class examples. Verify by multiplying out and adding that you get the right answer.
- Look back in Section 6.6; write out the four elements of the matrix X^TX and the two elements of the vector $X^T\mathbf{y}$ in terms of in terms of the given data $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. If time, use your answers to do #15 on p. 426.