JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

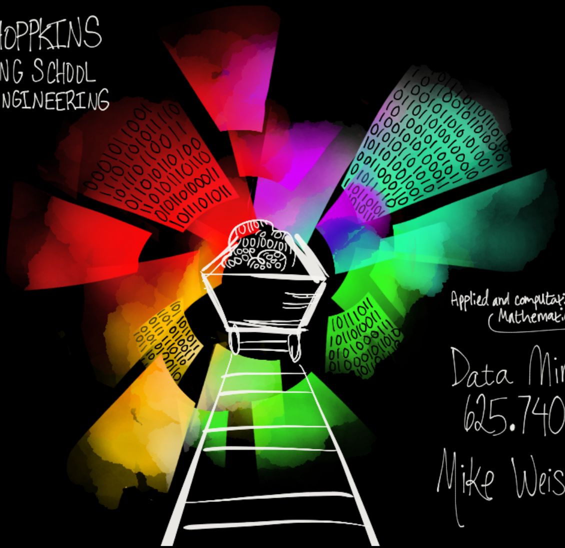
Applied and Computational Mathematics

Data Mining
625.740
Regression

Mike Weisman

email: `data.mining.625.740@gmail.com`

JOHNS HOPPKINS
WHITING SCHOOL
of ENGINEERING



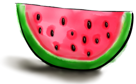
Applied and computational Mathematics

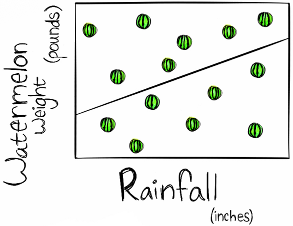
Data Mining 625.740

Mike Weisman

Regression

We wish to predict the response Y to a single predictor variable X and believe the relationship is linear: $Y \approx \beta_0 + \beta_1 X$





For example, X may be rainfall in inches, and Y the average weight of watermelons in our crop. To find estimates of our parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ we fit our sample $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ to a line.

Simple Linear Regression

Letting $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction of Y based on x_i, the i^{th} value of X, then $\varepsilon_i = y_i - \hat{y}_i$ is the i^{th} residual:

$$\varepsilon_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

We define the residual-sum-of-squares to be

$$R = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Estimating the Coefficients: $\hat{\beta}_0$ and $\hat{\beta}_1$

We seek β_0 and β_1 to minimize R. Taking partial derivatives:

$$-\frac{1}{2}\frac{\partial R}{\partial \beta_0} = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)$$

$$-\frac{1}{2}\frac{\partial R}{\partial \beta_1} = \sum_{i=1}^{n}x_i(y_i - \beta_0 - \beta_1 x_i)$$

Setting the partial derivatives equal to zero, yields two equations in the two unknown parameters.

$$\left.\begin{array}{l} \dfrac{\partial R}{\partial \beta_0} = 0 \\ \dfrac{\partial R}{\partial \beta_1} = 0 \end{array}\right\} \quad \Longrightarrow \quad \begin{array}{r} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{array}$$

Estimating the Coefficients: $\hat{\beta}_0$ and $\hat{\beta}_1$

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Applying Cramer's rule:

$$\hat{\beta}_0 = \frac{\begin{vmatrix} \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix}}, \quad \hat{\beta}_1 = \frac{\begin{vmatrix} n & \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i y_i \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix}}$$

Estimating the Coefficients: $\hat{\beta}_0$ and $\hat{\beta}_1$

The sample means are defined as $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Similarly $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ and $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

Then,

$$\hat{\beta}_0 = \frac{\overline{y} \cdot \overline{x^2} - \overline{x} \cdot \overline{xy}}{\overline{x^2} - (\overline{x})^2}, \quad \hat{\beta}_1 = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2}$$

Estimating the Coefficients: $\hat{\beta}_0$ and $\hat{\beta}_1$

Notice that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 &= \frac{1}{n}\sum_{i=1}^{n}x_i^2-2x_i\overline{x}+(\overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2-2\overline{x}\cdot\frac{1}{n}\sum_{i=1}^{n}x_i+(\overline{x})^2 = \overline{x^2}-2\left(\overline{x}\right)^2+(\overline{x})^2 \\ &= \overline{x^2}-(\overline{x})^2\end{aligned}$$

and

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) &= \frac{1}{n}\sum_{i=1}^{n}(x_iy_i-\overline{x}y_i-\overline{y}x_i+\overline{xy}) = \frac{1}{n}\sum_{i=1}^{n}x_iy_i-2\overline{xy}+\overline{xy} \\ &= \overline{xy}-\overline{x}\cdot\overline{y}\end{aligned}$$

We can thus express the parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ as

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \\ \hat{\beta}_0 &= \overline{y}-\hat{\beta}_1\overline{x}\end{aligned}$$

An Example

Fitting this data, the coefficients are $\hat{\beta}_0 = 7.03$ and $\hat{\beta}_1 = 0.0475$. Thus, by spending \$1000 on television advertising, we can expect to sell an additional 47.5 units of the product (assuming the trend continues!).

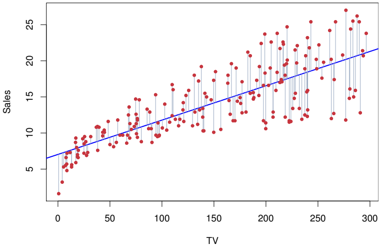


Figure: Sales vs. TV advertising, from **An Introduction to Statistical Learning**,p. 62.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Multiple Regression

With multiple inputs, $X^T = (X_1, X_2, \ldots, X_p)$, the linear regression model is

$$Y \approx \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

The $\hat{\beta}_j$'s are unknown parameters and the input variables X_j can come from different sources.*

- Quantitative inputs
- Transformations of quantitative inputs (e.g. log, square-root, $\cdots$)
- Basis expansions (e.g. $X_2 = X_1^2, \quad X_3 = X_1^3$) leading to a polynomial representation
- Numeric coding of levels of quantitative inputs
- Interactions between variables (e.g. $X_3 = X_1 \cdot X_2$)

*__The Elements of Statistical Learning__, Second Edition, p. 44.

Multiple Regression*

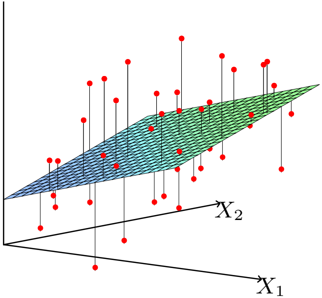


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

The Elements of Statistical Learning, Second Edition, p. 45.

Multiple Regression

Again, we apply the method of least squares by choosing $\beta_1, \ldots, \beta_p$ to minimize the residual-sum-of-squares:

$$R(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

Defining the $n \times (p+1)$ matrix $\mathbf{X}$ to be the input data with each row an input vector with 1 in the first position, and $\mathbf{y}$ the n-vector of outputs in the training data, we can write

$$R(\beta) = (\boldsymbol{y} - X\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta)$$

Proceeding as before,

$$-\frac{1}{2}\frac{\partial R}{\partial \beta} = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta)$$

$$\frac{1}{2}\frac{\partial^2 R}{\partial \beta \partial \beta^T} = \boldsymbol{X}^T\boldsymbol{X}$$

Multiple Regression

Assuming that $\mathbf{X}$ has full column rank, and thus $\mathbf{X}^T\mathbf{X}$ is positive definite, setting $\frac{\partial R}{\partial \beta} = 0$:

$$\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta) = 0$$

yields the unique solution for the β's

$$\hat{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

The predicted values at an input vector are given by $(1, x_1, x_2, \ldots, x_p)^T\hat{\beta}$ and the fitted values at the training inputs are

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

Geometrical Representation of Least Squares*

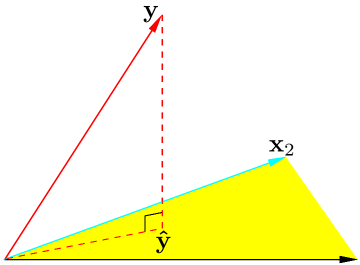


Figure: The vector $\hat{y}$ is the projection of y onto the column space of $\mathbf{X}$.

Denote the column vectors of $\mathbf{X}$ by $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_p\}$ where $\mathbf{x}_0 = \mathbf{1}$, a column of ones. These vectors span the column space of $\mathbf{X}$, a subspace of $\mathscr{R}^n$. We minimize $||\mathbf{y} - \mathbf{X}\hat{\beta}||^2$ by choosing $\hat{\beta}$ so that the residual vector $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to this subspace. The vector $\hat{\mathbf{y}}$ is the orthogonal projection of y onto this subspace.

*The Elements of Statistical Learning**, Second Edition, p. 46.

Polynomial Regression

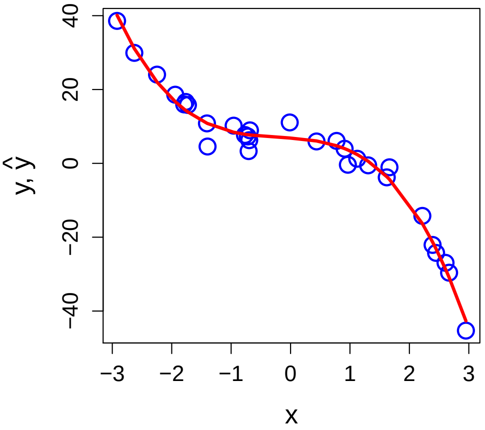
Polynomial regression is a special case of multiple regression. Recall that with $X^T = (1, X_1, X_2, \ldots, X_p)$ [we've included the input $X_0 = 1$], the linear regression model is

$$Y \approx \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

Now let $X^T = (1, X, X^2, \ldots, X^p)$, then

$$Y \approx \beta_0 + \sum_{j=1}^{p} X^j \beta_j$$

Polynomial Curve Fitting



$$Y \sim \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Applied and Computational Mathematics

Data Mining

625.740

Regression