Linear Regression

1 Brief Review of Regression

Recall that linear regression is a model for predicting a response or dependent variable (Y, also called an output) from one or more covariates or independent variables (X, also called explanatory variables, inputs, or features). For a given value of a single x, the expected value of y is

$$E[y] = \beta_0 + \beta_1 x$$

or we could say that $Y \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2)$. For data $(x_1, y_1), \ldots, (x_n, y_n)$, the fitted values for the coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$ are those that minimize the sum of squared errors $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, where the predicted values for the response are $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. We can get these values from R or Excel. These fitted coefficients give the least-squares line for the data.

This model extends to multiple covariates, with one β_j for each of the k covariates:

$$E\left[y_i\right] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}\,.$$

Optionally, we can represent the multivariate case using vector-matrix notation.

2 Conjugate Modeling

In the Bayesian framework, we treat the β parameters as unknown, put a prior on them, and then find the posterior. We might treat σ^2 as fixed and known, or we might treat it as unknown and also put a prior on it. Because the underlying assumption of a regression model is that the errors are independent and identically normally distributed with mean zero and variance σ^2, this defines a normal likelihood.

2.1 σ^2 Known

Sometimes we may know the value of the error variance σ^2. This simplifies the calculations. The conjugate prior for the β's is a normal prior. In practice, people typically use a non-informative prior, i.e., the limit as the variance of the normal prior goes to infinity, which is a completely flat prior, and is also the Jeffreys prior. Using this prior gives a posterior distribution for $\boldsymbol{\beta}$ which has the same mean as the standard least-squares estimates. If we are only estimating $\boldsymbol{\beta}$ and treating σ^2 as known, then the posterior for $\boldsymbol{\beta}$ is a (multivariate) normal distribution. If we just have a single covariate, then the posterior for the slope is

$$\beta_1|\boldsymbol{y} \sim \mathrm{N}\left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

If we have multiple covariates, then using matrix-vector notation, the posterior for the vector of coefficients is

$$\boldsymbol{\beta}|\boldsymbol{y} \sim \mathrm{N}\left((X^tX)^{-1}X^t\boldsymbol{y}, (X^tX)^{-1}\sigma^2\right),$$

where X denotes the design matrix and X^t is the transpose of X. The intercept is typically included in X as a column of 1's. Using an improper prior requires us to have at least as many data points as we have parameters to ensure the the posterior is proper.

2.2 σ^2 Unknown

If we treat both $\boldsymbol{\beta}$ and σ^2 as unknown, the standard prior is the non-informative Jeffreys prior, $f(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}$. Again, the posterior mean for $\boldsymbol{\beta}$ will be the same as the standard least-squares estimates. The posterior for $\boldsymbol{\beta}$ conditional on σ^2 is the same normal distribution as when σ^2 is known, but the marginal posterior distribution for $\boldsymbol{\beta}$, with σ^2 integrated out is a t distribution, analogous to the t tests for significance in standard linear regression. The posterior t distribution has mean $(X^tX)^{-1}X^t\boldsymbol{y}$ and scale matrix (related to the variance matrix) $s^2(X^tX)^{-1}$, where $s^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-k-1)$. The posterior distribution for σ^2 is an inverse gamma distribution

$$\sigma^2|y \sim IG\left(\frac{n-k-1}{2}, \frac{n-k-1}{2}s^2\right).$$

In the simple linear regression case (single variable), the marginal posterior for $\boldsymbol{\beta}$ is a t distribution with mean $\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ and scale $s^2 / \sum_{i=1}^n (x_i - \bar{x})^2$. If we are trying to predict a new observation at a specified input x^*, that predicted value has a marginal posterior predictive distribution that is a t distribution, with mean $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ and scale $se_r\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$. se_r is the residual standard error of the regression, which can be found easily in R or Excel. s_x^2 is the sample variance of x. Recall that the predictive distribution for a new observation has more variability than the posterior distribution for $\hat{y}$, because individual observations are more variable than the mean.