Contents

1 Linear Methods for Regression **1**
1.1 Linear Regression Models and Least Squares 1
1.2 The Gauss-Markov Theorem 3
1.3 Multiple Regression from Simple Univariate Regression 3
1.4 Multiple Outputs 4

1 Linear Methods for Regression

1.1 Linear Regression Models and Least Squares

The linear regression model has the form:
$f(x) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$
The variables X_j can come from different sources:

- Quantitative Inputs
- Transformations of quantitative inputs
- Basis expansions, such as $X_2 = X_1^2$
- Numeric, or dummy coding of the levels of qualitative inputs
- Interactions between variables, e.g. $X_3 = X_1 * X_2$

The most popular estimation method for our beta coefficients is **least squares**, in which we pick the coefficients $\beta = (\beta_0, \beta_1, ..., \beta_n)^T$ to minimize the residual sum of squares:
$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$
The unique solution to minimizing above is:
$\hat{\beta} = (X^T X)^{-1} X^T y$.
In order to pin down the sampling properties of $\hat{\beta}$, we make the following assumptions:

- the observations y_i are uncorrelated and have constant variance σ^2
- the x_i are fixed (non-random)

The variance-covariance matrix of least squares parameter estimates is given by:
$Var(\hat{\beta}) = (X^T X)^{-1} \sigma^2$.

Typically we estimate variance by $\hat{\sigma^2} = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$, where the $N - p - 1$ in the denominator makes $\hat{\sigma^2}$ an unbiased estimate of σ^2 ($E[\hat{\sigma^2}] = \sigma^2$).

To draw inferences about the parameters and the model, we assume:

- That the conditional expectation of Y is linear in $X_1, ..., X_p$
- The deviations of Y around its expectation are additive and Gaussian

$Y = E(Y|X_1, ..., X_p) + \epsilon$
Where

- $\epsilon \sim N(0, \sigma^2)$
- $\hat{\beta} \sim N(\beta, (X^T X)^{-1}\sigma^2)$
- $(N - p - 1)\hat{\sigma^2} \sim \sigma^2 \chi^2_{N-p-1}$
- $\hat{\beta}$ and $\hat{\sigma^2}$ are statistically independent

To test the hypothesis that a particular coefficient $\beta_j = 0$, we form the standardized coefficient, or z-score:

$z_j = \frac{\hat{\beta_j}}{\sigma\sqrt{\hat{v_j}}}$

where v_j is the jth diagonal element of $(X^T X)^{-1}$. Under $H_0 : \beta_j = 0$, $z_j \sim t_{N-p=1}$ and hence a large absolute value of z_j will lead to rejection of the null hypothesis.

We can isolate β_j to obtain a $1 - 2\alpha$ confidence interval for β_j:

$(\hat{\beta_j} \pm z^{1-\alpha} v_j^{\frac{1}{2}} \hat{\sigma^2})$

For example, $z^{(1-0.025)} = 1.96$

To test for significance of groups of variables simulataneously (e.g., testing a categorical variable with k levels), we can use the F statistic:

$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{(RSS_1/(N - p_1 - 1))}$ where RSS_1 is the residual sum of squares for the least squares fit of the bigger model and RSS_0 the same for the smaller model, having $p_1 - p_0$ parameters constrained to be zero.

We can also obtain an approximate confidence set for the entire parameter vector β:

$C_\beta = \{\beta | (\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta) \leq (\hat{\sigma^2} \chi^2_{p+1})^{1-\alpha}\}$

where $\chi_l^{2(1-\alpha)}$ is the $1 - \alpha$ percentile of the chi squared distribution on l degrees of freedom.

This confidence set for β generates a corresponding confidence set for the true function $f(x) = x^T \beta$

1.2 The Gauss-Markov Theorem

The Gauss-Markov theorem states that the least squares estimates of the parameters β have the smallest variance among all linear unbiased estimates.

If we have any other linear estimator $\tilde{\theta} = c^T y$ that is unbiased for $a^T\beta$ ($E(c^T y) = a^T\beta$), then

$Var(a^T\hat{\beta}) \leq Var(c^T y)$

Consider the mean squared estimate $\tilde{\theta}$ in estimating θ:

$MSE(\tilde{\theta}) = E(\tilde{\theta} - \theta)^2 = Var(\tilde{\theta}) + [E(\tilde{\theta}) - \theta]^2$

The first term is variance, the second is squared bias.

The Gauss-Markov theorem implies that the least squares estimator has the smallest mean squared error of all linear estimators with no bias. We can still get models with less MSE but some bias through regularization. These estimators trade bias for a reduction in variance. Picking the right model amounts to creating the right balance between bias and variance.

1.3 Multiple Regression from Simple Univariate Regression

Consider the univariate model with no intercept $Y = X\beta + \epsilon$. The least squares estimates for this model would be $\hat{\beta} = \frac{\sum_1^N x_i y_i}{\sum_1^N x_i^2}$, with residuals $r_i = y_i - x_i\hat{\beta}$.

Let $y = (y_1, ..., y_n)^T$, $x = (x_1, ..., x_N)^T$, and define $< x, y >= \sum_{i=1}^N x_i y_i = x^T y$ (the inner product). Then we can write

$\hat{\beta} = \frac{<x,y>}{<x,x>}$ $r = y - x\hat{\beta}$

Suppose the inputs $x_1, x_2, ..., x_p$ are orthogonal ($< x_j, x_k >= 0$) for all $j \neq k$. Then the multivariate least squares estimates $\hat{\beta}_j$ are equal to $\frac{<x_j,y>}{<x_j,x_j>}$ – the univariate estimates. When the inputs are orthogonal, they have no effect on each other's parameter estimates in the model. Orthogonal inputs occur most often with balanced, designed experiments where orthogonality is enforced.

Suppose we have an intercept with a single input x. Then the least squares coefficient of x has the form

$\hat{\beta}_1 = \frac{<x-\bar{x}1,y>}{<x-\bar{x}1,x-\bar{x}1>}$

where $\bar{x} = \sum_i \frac{x_i}{N}$ and $1 = x_0$, the vector of N ones. We can view the estimate $\hat{\beta}_1$ as the result of two applications of the simple regression.

The steps are:

1. Regress x on 1 to produce the residual $z = x - \bar{x}1$
2. Regress y on the residual z to give the coefficient $\hat{\beta}_1$

In this procedure, regress b on a means a simple univariate regression of b on a with no intercept, producing coefficient $\hat{\gamma} =< a, b > / < a, a >$ and residual vector $b - \hat{\gamma}a$. We say that b is adjusted for a, or is orthogonalized with respect to a.

Generalizing to p predictors we get

$\hat{\beta}_p = \frac{<z_p, y>}{<z_p, z_p>}$

The multiple regression coefficient $\hat{\beta}_j$ represents the additional contribution of x_j on y, after x_j has been adjusted for $x_0, x_1, ..., x_{j-1}, x_{j+1}, ..., x_p$.

If x_p is highly correlated with some of the other x_k's, the residual vector z_p will be close to zero and the coefficient $\hat{\beta}_p$ will be very unstable.

We also can get an alternative formula for the variance estimates:

$Var(\hat{\beta}_p) = \frac{\sigma^2}{<z_p, z_p>}, \frac{\sigma^2}{||z_p||^2}$

In other words, the precision with which we can estimate $\hat{\beta}_p$ depends on the length of the residual vector z_p; this represents how much of x_p is unexplained by the other x_ks.

The algorithm described here is the **Gram Schmidt** procedure for multiple regression. It can be used to obtain the entire multiple least squares fit (as opposed to just $\hat{\beta}_p$).

We can represent the second step of the algorithm in matrix form as the following:

$x = Z\Gamma$

where Z has as columns the z_j in order and Γ is the upper triangular matrix with entries $\hat{\gamma_{kj}}$.

Introducing the diagonal matrix D with the jth diagonal entry $D_{jj} = ||z_j||$, we get

$X = ZD^{-1}D\Gamma = QR$

the QR decomposition of X. Here Q is an Nx(p + 1) orthogonal matrix, $Q^TQ = I$, and R is a $(p + 1)x(p + 1)$ upper triangular matrix.

The QR decomposition represents a convenient orthogonal basis for the column space of X. With it, the least squares solution is given by

$\hat{\beta} = R^{-1}Q^Ty$ $\hat{y} = QQ^Ty$

1.4 Multiple Outputs