

## Contents

<b>1 Linear Methods for Regression</b>	<b>1</b>
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 Linear Methods for Regression

### 1.1 Linear Regression Models and Least Squares

The linear regression model has the form:

$$f(x) = \beta_0 + \sum_{j=1}^p X_j \beta_j + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2)$$

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  $\beta = (\beta_0, \beta_1, \dots, \beta_n)^T$  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  $\sigma^2$
- 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  $\sigma^2$  ( $E[\hat{\sigma}^2] = \sigma^2$ ).

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

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

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

Where

- $\epsilon \sim N(0, \sigma^2)$
- $\hat{\beta} \sim N(\beta, (X^T X)^{-1} \sigma^2)$
- $(N - p - 1) \hat{\sigma}^2 \sim \sigma^2 \chi_{N-p-1}^2$
- $\hat{\beta}$  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:

$$z_j = \frac{\hat{\beta}_j}{\sigma \sqrt{v_j}}$$

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  $\beta_j$  to obtain a  $1 - 2\alpha$  confidence interval for  $\beta_j$ :

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

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

To test for significance of groups of variables simultaneously (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  $\beta$ :

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

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  $\beta$  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  $\beta$  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  $\theta$ :

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

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