Linear Regression

Shaun Ho

February 2024

1 Defining the optimization problem

We define the objective function as the sum of the squared divergence between each label y_i and predicted value μ_i:

$$S = \sum_{i=1}^{n} w_i (y_i - \mu_i)^2$$

In matrix notation,

$$S = (\mathbf{y} - \boldsymbol{\mu})^\top W (\mathbf{y} - \boldsymbol{\mu})$$
$$S = (\mathbf{y} - X\beta)^\top W (\mathbf{y} - X\beta)$$

Thus,

$$\min_{\beta} \mathrm{S} = \min_{\beta} \sum_{i=1}^{n} w_i (y_i - \mu_i)^2$$
$$\min_{\beta} \mathrm{S} = \min_{\beta} (\mathbf{y} - X\beta)^\top W (\mathbf{y} - X\beta)$$

$\mathbf{y}$ is of shape $n \times 1$.

X is a matrix of shape $n \times p'$ where n is the number of observations and p' is the number of coefficients, often $(p + 1)$ to account for the intercept term.

β is of shape $p' \times 1$ and the weight matrix W is of shape $n \times n$.

2 Fitting the model

Using the partial derivatives of S for each β_j and equating to zero, the optimal $\hat{\beta}$ is found by solving the set of p' simultaneous equations.

Rearranging the equation that minimizes S such that $\hat{\beta}$ is found by the sum

of cross-products of x and y divided by the sum of squares of x, we obtain in matrix notation:

$$\hat{\beta} = (X^\top W X)^{-1} X^\top W \mathbf{y}$$

Which relies on the assumption that $(X^\top W X)$ admits an inverse.

Then, the fitted values are:

$$\hat{\mu} = X\hat{\beta}$$

For reasons of computational efficiency it is actually preferable to compute $\hat{\beta}$ using the solution to a linear system of equations rather than by explicitly inverting $(X^\top W X)$. In fact, most built-in statistical packages avoid computing $(X^\top W X)$ entirely by using sophisticated methods involving decompositions of X.

Crucially, all of this can be done without knowing the value of σ^2.

2.1 Estimating the variance at the point of the optimal solution

As seen above,

$$S = \sum_{i=1}^{n} w_i (y_i - \mu_i)^2$$

When S is minimized this is known as the RSS, which takes residual degrees of freedom equal to $n - p'$. This gives us the following unbiased estimator for σ^2:

$$s^2 = \frac{S}{\mathrm{df}} = \frac{RSS}{n - p'}$$

2.2 Estimating the variance of $\hat{\beta}$

The variance of $\hat{\beta}$ is given by:

$$\frac{MSE}{SS_x}$$

Expanding,

$$SE(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 / (n - p')}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Crucially, the numerator applies a single scalar estimate of variance (the MSE) across all observations. This implies that the variance of the errors $(y_i - \mu_i)$ is assumed to be constant across all x. This will become important later.

2.2.1 Matrix notation

Using $\hat{\beta} = (X^\top W X)^{-1} X^\top W \mathbf{y}$ it can be shown that the covariance matrix for $\hat{\beta}$ is:

$$\text{var}(\hat{\beta}) = \sigma^2 (X^T W X)^{-1}$$

Since σ^2 is unknown we obtain an estimate as follows:

$$\hat{\text{var}}(\hat{\beta}) = s^2 (X^T W X)^{-1}$$

The diagonal elements are the values of each $\hat{\text{var}}(\hat{\beta_j})$.

3 Performing inference on model outputs

Up to now, no specific statistical distribution has been assumed for the responses in the regression. The responses have simply been assumed to be independent and to have constant variance.

If we assume that the responses are normally distributed, either with constant variance or with variances that are proportional to some known weights w_i, such that:

$$y_i \sim \mathcal{N}(\mu_i, \sigma^2 / w_i)$$
$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}$$

then hypothesis tests and confidence intervals can be developed, since the distributions of the regression parameters would then be known by virtue of the fact that those parameters are simply a linear combination of the responses.

Notably, since β_0 and β_1 are asymptotically normally distributed, the assumption is only needed to validate the following tests in the event of small sample sizes.

3.1 The distribution of $\hat{\beta}$

If y_i is normal, then since $\hat{\beta}$ is a linear combination of y_i we obtain:

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \text{var}(\hat{\beta}_j)).$$

Thus, we have the Z-stat:

$$Z = \frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)},$$

Given that σ^2 is unknown, we estimate each $\text{se}(\hat{\beta}_j)$ with each $\sqrt{\hat{\text{var}}(\hat{\beta}_j)}$. We also use the T-test instead:

$$T = \frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)}$$

Given by a Student's t-distribution with $(n - p')$ degrees of freedom. Notably, with sufficient degrees of freedom, this converges to the normal distribution.

3.2 Testing hypotheses about $\hat{\beta}$

We can directly use the above T-test to obtain t-statistics for testing against any null hypothesis. Crucially, this tests significant differences from the null hypothesis in the presence of all other predictors present in the model.

3.3 Confidence intervals for $\hat{\beta}$

$$\hat{\beta}_j \pm t^*_{\alpha/2, n-p'} \operatorname{se}(\hat{\beta}_j)$$

3.4 Prediction intervals for $\hat{\mu}$

Here we use the the known expression for $\operatorname{var}(\hat{\mu})$:

$$\operatorname{var}[\hat{\mu}_g] = \sigma^2 \left\{ \frac{1}{\sum w_i} + \frac{(x_g - \bar{x})^2}{SS_x} \right\}.$$

Like in the above case for $\operatorname{var}(\hat{\beta})$ we use s^2 as the estimate for σ^2, and thus once again use the t-statistic as follows:

$$\hat{\mu}_g \pm t^*_{\alpha/2, n-p'} \operatorname{se}(\hat{\mu}_g)$$

4 Making sure the features we selected are useful predictors of the responses

4.1 F-stat

Given that the total error TSS is given by $SSR + RSS$, the question of whether our features are useful can be answered by testing whether the regression sum of squares SSR is larger than would be expected due to random variation.

To that end, we first note that under the null hypothesis $\beta = 0$ the ratio of $\frac{RSS}{\sigma^2}$ would have a chi-square distribution with $n - p'$ degrees of freedom. Furthermore, $\frac{SSR}{\sigma^2}$ would have a chi-square distribution with $p' - 1$ degrees of freedom.

As such, the ratio $\frac{MSR}{MSE = s^2} = \frac{SSR/(p'-1)}{RSS/(n-p')}$ would follow an F-distribution with $(p' - 1, n - p')$ degrees of freedom.

A large value for F implies that the proportion of the variation explained by the systematic component MSR is large relative to the contribution of the random component $MSE = s^2$.

If the test statistic is large enough, we reject the null hypothesis that $\beta = 0$.

4.2 $\mathbf{R^2}$

Given the coefficient of determination,

$$R^2 = \frac{SSR}{TSS} = 1 - \frac{RSS}{TSS}$$

adding more features will not increase RSS even if they do not hold real explanatory power. One alternative is to penalize R^2 for the number of features used, as such:

$$\bar{R^2} = 1 - (1 - R^2)\frac{n-1}{n-p'}$$

5 Making sure the inference is robust

The robustness of the inferences we make above depends on a number of assumptions. For instance, we saw that the construction of the confidence interval for $\hat{\beta}$ relies on a single estimate of variance that was computed across all observations. If variance is not actually constant across the observations, our estimate for $SE(\hat{\beta})$ is no longer robust. Below are some tests to assess whether some of these assumptions are obviously violated.

5.1 Linearity

We test the assumption of linearity by plotting each residual against X, $e_i = Y_i - \beta X_i$, standardized by the standard error $\sqrt{s^2(1-h_i)}$ where h_i is the leverage of Y_i.

5.2 Constant Variance

Here we plot the standardized residuals against the **fitted values**.

5.3 Normality

Here, we plot theoretical quantiles vs. observed quantiles of the standardized residuals. If points deviate too far from the Q-Q line, then the assumption that residuals are normally distributed needs to be examined closely.

5.4 Independence

Here, we plot each residual e_i against its lagged residual e_{i-1}. If patterns are visible suggesting correlations, then the error terms may not be independent.

6 What to do when the assumptions are violated

There are some transformations and adjustments that can improve the model's compliance with the above assumptions.

6.1 Examine outliers

6.1.1 Identifying outliers

These observations can be spotted because of the large magnitude of the model residuals that they generate. The propensity of a given large residual to influence the model fit depends on the leverage of the associated observation.

The leverage h_i of each X_i is obtained by taking the diagonals of the hat matrix $H = X(X^\top WX)^{-1}X^\top$. In the case of simple linear regression it can be shown that each h_i is a function of the distance of X_i from the mean:

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SXX}$$

A heuristic that is sometimes used is to flag an observation as high-leverage if:

$$h_i > \frac{3(p+1)}{n}$$

An observation with a large residual may have a very strong influence on the model fit if it is also high-leverage. Notably, both conditions must be satisfied for one to conclude that the model fit is being substantially affected by said observation.

6.1.2 Choosing the right measure for the residual

Before exploring the various methods available for assessing the influence of an outlier, it must first be noted that some of them rely on a more robust method of measuring the residual (as compared to standardized residuals). The **studentized residual** accounts for the possibility that outliers may be hard to detect if the residuals they generate are so large as to significantly influence s^2, which is used in the calculation of the standard error of the standardized residual. The computation is the same as that for the standardized residual, except that the estimate fot s^2 is calculated after dropping observation i from the sample:

$$e_i'' = \frac{\hat{Y}_i - \hat{\beta}X_i}{\sqrt{s^2_{-i}(1-h_i)}}$$

6.1.3 Heuristics for assessing influence of outliers

The **Cook's Distance** compares the extent to which model predictions would differ (globally) if the i-th observation was dropped from the model before fitting:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{-i,j})^2}{(p+1)\hat{\sigma}^2} = \frac{(e_i'')^2}{p+1-h_i}$$

Flag as influential if $D_i > 1$.

DFFITS does the same thing but locally:

$$\text{DFFITS}_i = \frac{\hat{\beta}X_i - \hat{\beta}_{-i}X_i}{\sqrt{\hat{\sigma}^2_{-i}h_i}} = e_i''\sqrt{\frac{h_i}{1-h_i}}$$

Flag as influential if $|\text{DFFITS}_i| > 3\sqrt{\frac{(p+1)}{(n-(p+1))}}$.

DFBETAS measures the degree to which observation i influences the coefficient of $\hat{\beta}$, measured in number of standard deviations of $\hat{\beta}$ generated while dropping observation i from the model fitting process:

$$\text{DFBETAS}_{i,j} = \frac{\hat{\beta}_j - \hat{\beta}_{-i,j}}{\sqrt{V_{-i}(\hat{\beta}_{-i,j})}}$$

Flag as influential if $|\text{DFBETAS}_{i,j}| > 1$.

The **Covariance Ratio** is the average factor by which the confidence intervals for the regression coefficients change when observation i is dropped:

$$CR_i = \frac{1}{1-h_i}\left(\frac{n-p}{n-(p+1)+(e_i'')^2}\right)^p$$

Flag as influential if $CR_i > 3\frac{(p+1)}{(n-(p+1))}$.

6.2 Transformations for linearity

If X and Y turn out to be not obviously linear, sometimes it is possible to transform X and/or Y using the following operations so that the relationship between them is approximately linear.

6.2.1 Scaled power family

$$\psi(t,\lambda) = \begin{cases} t^\lambda, & \text{if } \lambda \neq 0 \\ \log t, & \text{if } \lambda = 0. \end{cases}$$

6.2.2 Modified power family

$$\psi_S(t,\lambda) = \begin{cases} \frac{(t^\lambda - 1)}{\lambda}, & \text{if } \lambda \neq 0 \\ \log t, & \text{if } \lambda = 0. \end{cases}$$

This modification guarantees that the direction of association between X and Y is preserved.

6.2.3 An algorithm for optimal scaling of the Y-variable

The modified power family is used as follows to transform Y:

$$\psi_M(t,\lambda) = \text{gm}(Y)^{1-\lambda}\psi_S(t,\lambda)$$

Where $\text{gm}(Y)$ is the geometric mean of Y. The Box-Cox method consists of selecting the value of λ that minimizes the Residual Sum of Squares (RSS). λ is found where the the "normality" of $\psi_M(t,\lambda)$ is maximized.

6.3 Weighting the model for homoskedasticity

The variance of the error terms can be stabilized by weighting each observation in a strictly positive manner as such:

$$Y = X\beta + \epsilon$$

$$V(\epsilon) = \sigma^2 W^{-1}$$

$$V(Y_i|X_i = x_i) = V(\epsilon_i|X_i = x_i) = \frac{\sigma^2}{w_i}$$

This allows us to control for a number of situations where there are dependencies in the error term, such as:

1. When the i-th observation of Y in the sample is an average of n_i; equally variable observations, $V(Y_i) = \sigma^2/n_i$; and therefore $w_i = n_i$.

2. When the i-th observation of Y in the total sum of n_i; equally variable observations, $V(Y_i) = n_i\sigma^2$ and therefore $w_i = 1/n_i$.

3. When the variance of Y is known to be proportional to some predictor X, $V(Y_i|X = x_i) = \sigma^2 x_i$; and therefore $w_i = 1/x_i$.

4. Finding some other optimal set of weights W that ensures homoskedasticity (subject to considerations pertaining to overfitting). In practice, finding such weights can be challenging. The optimal weights might not be immediately apparent, and different methods, such as iteratively reweighted least squares (IRLS), might be used to approximate them.

6.4 Use nonlinear regression

It may be possible that domain knowledge or other specifications hold that the function is nonlinear, such as functions where there is a floor in Y for all X (which can be represented by $Y = \theta_0 = \theta_1 \max(0, X - \theta_2)$). Here, we wish to estimate the parameters θ that minimize the RSS. However, the key difference is that a closed-form solution for all $\hat{\theta}$ does not exist. Alternative algorithms like gradient descent may be considered here.

7 Practical issues

7.1 Collinearity

Collinearity occurs where there are strong correlations between the predictors in a linear regression model. The variance of the estimated coefficient $\hat{\beta}_j$ is given by:

$$V(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \frac{\sigma^2}{S_{X_j X_j}}$$

where $S_{X_j X_j} = \sum_{i=1}^{n} (X_{ji} - \bar{X}_j)^2$ and R_j^2 is the coefficient of determination associated with the regression of X_j on all other predictors. Note that the stronger the correlation between X_j and the other predictors, the larger and more unstable the variance of $\hat{\beta}_j$ becomes.

The factor

$$VIF_j = \frac{1}{1 - R_j^2}$$

is called the Variance Inflation Factor of the predictor X_j. VIFs can be estimated from the data. Large VIFs (e.g., $VIF > 5$) among predictors are cause of concern. For instance, $VIF_j = 10$ means that the variance of $\hat{\beta}_j$ is 10 times larger than it would be if X_j was uncorrelated with all other predictors.

From the perspective of model fitting in matrix form, high collinearity makes the problem nearly intractable because the design matrix (matrix of predictors) approaches singularity, undermining invertibility which is crucial for computing the least-squares estimate. The rank of a matrix is defined as the maximum number of linearly independent column vectors in the matrix or equivalently the maximum number of linearly independent row vectors in the matrix. If at least one vector can be expressed as a combination of others, those vectors are linearly dependent and the matrix is rank-deficient.

7.2 Categorical variables

The above discussion on rank also explains why, after encoding categorical variables in a one-hot manner, one category must be selected as the baseline feature and dropped from the model before fitting. Otherwise, any variable in the entire

set of categorical variables may be expressed as a linear combination of all other sets, resulting in rank-deficiency.

Note that during the process of model fitting, only one dummy variable was active at each observation, consequently, during prediction, only one of the dummy variables associated with a categorical variable can take a nonzero value. In other words, an observation or prediction cannot simultaneously belong to multiple categories of a single categorical variable at the same time. As such, the effects of each dummy variable are not additive in the sense that one can't simply sum the coefficients to get a total effect.

7.3 Feature selection in high-dimensional space

What are some strategies for finding the best subset C of $p_C \leq p$ predictors for maximizing performance?

7.3.1 Objective functions for penalizing complexity

1. AIC (Akaike Information Criterion):

$$AIC = n \log\left(\frac{RSS_c}{n}\right) + 2p_c$$

   AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model. In estimating the amount of information lost by a model, AIC deals with the trade-off between the goodness of fit of the model and the simplicity of the model. In other words, AIC deals with both the risk of overfitting and the risk of underfitting. Generally, a higher AIC indicates higher probability of information loss.

2. BIC (Bayesian Information Criterion):

$$BIC = n \log\left(\frac{RSS_c}{n}\right) + p_c \log(n)$$

   The penalty term is larger in BIC than in AIC for sample sizes greater than 7.

3. Mallow's C_p:

$$C_p = \frac{RSS_c}{\sigma^2} + 2p_c - n$$

   The optimum model under this criterion is a compromise influenced by the sample size, the effect sizes of the different predictors, and the degree of collinearity between them. It is only valid for large sample sizes and is subject to a number of caveats.

4. Adjusted R^2:

$$\text{Adjusted } R^2 = 1 - \frac{RSS_c/(n - (p_c + 1))}{SYY/(n-1)}$$

7.3.2 Greedy algorithms

Best Subset Selection (Brute Force)

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots, p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here best is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Forward Stepwise Selection

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the best among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here best is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Backward Stepwise Selection

1. Let $\mathcal{M}_p$ denote the full model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k - 1$ predictors.
   (b) Choose the best among these k models, and call it $\mathcal{M}_{k-1}$. Here best is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

8 References

Dunn, P., & Smyth, G. (2018). Generalized Linear Models With Examples in R. Springer-Verlag. https://doi.org/10.1007/978-1-4419-0118-7
Ciollaro, M. (2023). Statistical Foundations of Machine Learning (Lecture Notes).