Linear Regression and Regularization[1]

Germán G. Creamer

June 16, 2022

[1] Sources: Introduction to Statistical Learning with R, and Thomas Lonon

Simple Linear Regression: A very straightforward approach for predicting a quantitative response Y on the basis of a single predictor variable X that assumes there is an approximately linear relationship between X and Y. This is expressed as:

$$Y \approx \beta_0 + \beta_1 X$$

Once we have determined our estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we can predict future sales

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Let

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

represent n observation pairs. We are looking for coeffecient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that represent the data well.

In other words, we are looking for these parameters such that:

$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$$

This is done through **least squares**

Let

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

be the predictor for Y based on the i^{th} value of X. Then the i^{th} **residual**, the difference between the observed and the predicted response, is represented as

$$e_i = y_i - \hat{y}_i$$

The **Residual Sum of Squares (RSS)** is given as:

$$RSS = \sum_{i=1}^{n} e_i^2$$

which can also be represented as

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

For the second representation of RSS given, we can determine the parameters for $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize this value. We do this by taking the derivatives with respect to these parameters and setting them equal to 0 (standard approach). We get:

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0$$

This system of equations is solved as:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

and

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

We have the assumption that the *true* relationship between X and Y takes the form

$$Y = f(X) + \epsilon$$

for some function f. If this function is a linear function, then we have:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where β_0 is the intercept term and β_1 is the slope. This expression is the **population regression line**, the best linear approximation to the true relationship between X and Y.

For a set of i.i.d. random variable $\{x_i\}, i \in 1, \ldots, n$, what can we say about the average($\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$)?

- The expected value is:

$$\mathbb{E}[\bar{x}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i] = \mathbb{E}[x_i]$$

- The variance is:

$$\begin{aligned}
\mathbb{V}(\bar{x}) &= \mathbb{E}[(\bar{x} - \mathbb{E}[\bar{x}])^2] \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i]\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n}(x_i - \mathbb{E}[x_i])\right)^2\right] \\
&= \frac{1}{n}\mathbb{V}(x_i)
\end{aligned}$$

This estimator for $\bar{x}$ is an example of an **unbiased** estimator.

Based on this definition of the variance of a sample mean, we can also have the **standard error of the estimate (SE)** given by:

$$SE(\hat{\mu}) = \frac{1}{\sqrt{n}}\sigma$$

where σ is the standard deviation of each of the realizations y_i of Y.

Using this approach, we can get the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $\sigma^2 = \mathbb{V}(\epsilon)$. If σ isn't known, we use an estimate for σ known as the **residual standard error (RSE)** given by the formula:

$$RSE = \sqrt{\frac{RSS}{n-2}}$$

These standard errors can be used to calculate our **confidence intervals.** For linear regression, the 95% confidence intervals are given by:

$$\hat{\beta}_0 \pm 2SE(\hat{\beta}_0)$$
$$\hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$$

That is, there is approximately a 95% chance that the true value of β_0 is contained within

$$[\hat{\beta}_0 - 2SE(\hat{\beta}_0), \hat{\beta}_0 + 2SE(\hat{\beta}_0)]$$

Hypothesis Testing

The most common approach is to test the null hypothesis (H_0) versus the alternate hypothesis (H_A). For linear regression an example of this test would be to check whether there is a relationship between X and Y.

- : H_0: There is no relationship between X and Y

$$H_0 : \beta_1 = 0$$

- : H_A: There is some relationship between X and Y

$$H_A : \beta_1 \neq 0$$

Note that if $\beta_1 = 0$, then $Y = \beta_0 + \epsilon$

To test these hypotheses, we compute a **t-statistic** given by

$$t = \frac{\hat{\beta}_1 - 0}{\widehat{SE}(\hat{\beta}_1)}$$

which measures the number of standard deviations that $\hat{\beta}_1$ is from 0. If there is no relationship between X and Y, then the value of t will have a t-distribution with $n-2$ degrees of freedom. The **p-value** is the probability of observing $|t|$ or larger with this distribution. If this p-value is small enough, we **reject the null hypothesis**

If we reject the null hypothesis, we will want to know the extent in which the model fits the data. This is assessed using the RSE and the **R^2 statistic.**

This RSE is considered a measure of the **lack of fit** of the model to the data.

To calculate R^2 we use:

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

where

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

is the **total sum of squares (TSS)**

R^2

This R^2 statistic is a relative measure of fit, and so is easier to interpret than the RSE. It measures the proportion of variability in Y that can be expressed using X.

- a R^2 close to 1 indicates a large proportion of variability in the response has been explained by the regression
- R^2 close to 0 indicates the the regression did not explain much of the variability in the response.

There is still some leeway as to what constitutes a "good" R^2 value.

Ridge Regression

Previously defined we have the least squares fitting procedure that estimates $\beta_0, \beta_1, \ldots, \beta_p$ and minimizes

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

Ridge Regression is similar to least squares, but it finds the estimates that minimize:

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

where $\lambda \geq 0$ is a *tuning parameter*

This ridge regression can be alternately formulated as either:

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

subject to $\sum_{j=1}^{p} \beta_j^2 \leq t$

or in matrix notation:

$$RSS(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta^T \beta$$

$$\hat{\beta}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

The standard least squares coefficient estimates are **scale equivariant**

This ridge regression is not scale equivariant so we apply the ridge regression after **standardizing the predictors**

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=r}^{n}(x_{ij} - \bar{x}_j)^2}}$$

[2]

Shrinkage: Lasso

An alternative to ridge regression that picks the coefficients $\hat{\beta}_\lambda^L$ that minimizes:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

This approach produces **sparse** models.

Alternate Formulation

$$\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\}, \text{ subject to } \sum_{j=1}^{p}|\beta_j|\le s$$

$$\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\}, \text{ subject to } \sum_{j=1}^{p}\beta_j^2\le s$$

[1]

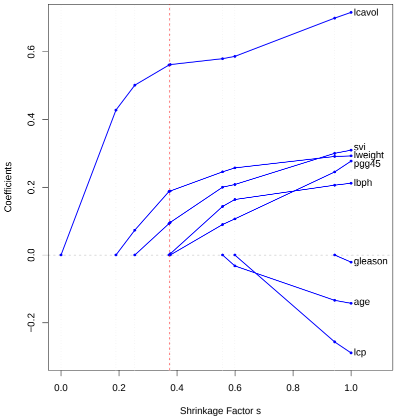

FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 9; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed;*

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: Data mining, inference and prediction*. Springer-Verlag, New York, 2 edition, 2008.

[2] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer-Verlag, New York, 2 edition, 2021.