Statistical Learning

Maryclare Griffin

2024-09-12

This material is based on Chapter 3 of Introduction to Statistical Learning (ISL) and parts of Chapter 3 of Elements of Statistical Learning (ESL). We will tend to follow ISL more closely, and look to ESL for occasional additional higher level material.

Simple linear regression relates a quantitative response Y to a single predictor variable Y by assuming

$$Y \approx \beta_0 + \beta_1 X \tag{1}$$

and obtaining estimates of the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$, which can be used to form predictions of Y given $X = x$, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ and which can be used to draw inference on the presence and strength of the relationship between X and Y.

Sometimes, we will refer to fitting the model described by (1) as regressing Y on or onto X.

We often refer to the unknown constant β_0 as the **intercept** and β_1 as the **slope**. The intercept β_0 represents the expected value of Y when $X = 0$, and the slope β_1 represents the average increase in Y associated with a one unit change in X. Together, β_0 and β_1 are often called (regression) coefficients or parameters.

We estimate β_0 and β_1 given n observations $(x_1, y_1), \ldots, (x_n, y_n)$ by making the estimates $\hat{\beta}_0 + \hat{\beta}_1 x$ as close as possible to the observed values y_i. We measure closeness using the least squares criterion,

$$\sum_{i=1}^{n} (y_i - b_0 - x_i b_1)^2,$$

with and minimize it with respect to b_0 and b_1.

When evaluated at the least squares minimizing values $\hat{\beta}_0$ and $\hat{\beta}_1$, this quantity is called the **residual sum of squares (RSS)**,

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

where $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ refers to the i-th **residual**, denoted by e_i.

Letting $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ refer to the sample means of the response and the single predictor, we can show that closed form solutions are available for $\hat{\beta}_1$ and $\hat{\beta}_0$,

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

Going back to the idea that we assume that the true relationship between X and Y satisfies $Y = f(X) + \epsilon$ for an unknown function f and mean-zero error ϵ, then we can think of linear regression as approximating f by a linear function $\beta_0 + \beta_1 X_i$. The error ϵ may reflect measurement error or deviations of the relationship

between X and Y from linearity. The error ϵ are assumed to be independent of the predictor X. In what follows, when we say "the simple linear model holds" we mean that $Y = \beta_0 + \beta_1 X + \epsilon$ for some fixed but unknown β_0 and β_1 and ϵ is a mean zero error that is independent of X. When the simple linear model holds, we may refer to $\beta_0 + \beta_1 X$ as the **population regression line**. The regression line obtained from the least square estimates $\hat{\beta}_0 + \hat{\beta}_1 X$ refers to the **least squares regression line**.

When the simple linear model holds, the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators for β_0 and β_1, i.e. $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$. With some additional assumptions, specifically that the errors ϵ are independent, identically distributed, and have finite variance $\sigma^2 = \text{Var}[\epsilon] < \infty$, the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ under the simple linear model are

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

We can estimate these variances by plugging in an estimate of σ^2, obtained by dividing the residual sum of squares by $n - 2$. The square root of the estimated variances is often called the **standard error**,

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

Standard errors can be used to obtain $100 \times (1 - \alpha)\%$ **confidence intervals**, which describes a range of values which are expected to cover the true value of the corresponding parameter in $100 \times (1 - \alpha)\%$ of repeated samples. An approximate $100 \times (1 - \alpha)\%$ confidence interval is

$$\left(\hat{\beta}_j + z_{\alpha/2} \text{SE}[\hat{\beta}_j], \hat{\beta}_j + z_{1-\alpha/2} \text{SE}[\hat{\beta}_j] \right) \text{ for } j = 0, 1,$$

where z_q is the q-th quantile of a standard normal random variable.

Standard errors can also be used to conduct **hypothesis tests**, which allow us to test if the true value of a parameter is equal to a specific value. The most common hypothesis test tests the **null hypothesis** that there is no relationship between X and Y, denoted by H_0, versus the **alternative hypothesis** that there is some relationship between X and Y, denoted by H_a. This corresponds to the test H_0: $\beta_1 = 0$ versus the alternative H_a: $\beta_1 \neq 0$. We can test this null hypothesis using a t-statistic, $\hat{\beta}_1 / \text{SE}[\hat{\beta}_1]$. This is a quantity that - if our simple linear model holds and the errors are independent and identically distributed with finite variance - should be approximately distributed according to a standard normal distribution. Therefore, a level-α test of the null hypothesis H_0 can be obtained by comparing $\hat{\beta}_1 / \text{SE}[\hat{\beta}_1]$ to the $\alpha/2$ and $1 - \alpha/2$ quantiles of a standard normal distribution $z_{\alpha/2}$ and $z_{1-\alpha/2}$. If our test statistic is outside of the interval $(z_{\alpha/2}, z_{1-\alpha/2})$ we reject the null hypothesis and conclude there is a relationship between X and Y. Otherwise, we fail to reject the null hypothesis and cannot conclude that there is a relationship between X and Y.

We can also compute the ***p*-value** of this test by computing the probability that a standard normal random variable z is more extreme than the t-statistic we observed, $\hat{\beta}_1 / \text{SE}[\hat{\beta}_1]$,

$$2\text{Pr}(z \geq |\hat{\beta}_1 / \text{SE}[\hat{\beta}_1]|).$$

This is the probability of observing a t-statistic as extreme or more extreme than the one we observed relative to the distribution we would expect it to have if the null hypothesis were true. We can (roughly speaking) interpret the p-value as an (inverse) measure of the strength of the evidence of a relationship between X and Y; smaller p-values correspond to stronger evidence of a relationship. For a level-α test of the null hypothesis H_0, we reject the null when the corresponding p-value is smaller than α.

Having fit a simple linear regression model, it is natural to ask how well the model fits the observed data. We will consider two measures of fit: the **residual standard error (RSE)**, which we encountered earlier when we discussed estimating the error variance σ^2, and the R^2 statistic. RSE is the square root of the least squares estimator of σ^2, $\text{RSS}/(n-2)$.

A limitation of using RSE to measure model fit is that it is sensitive to the scale of the response, Y. A scale-free measure of model fit that is very interpretable is R^2, which descibes the proportion of overall

variability of the response that is explained by the fitted model. Letting $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ be a measure of the overall variability of the response, which we refer to as the total sum of squares. The R^2 statistic is

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

We can also interpret R^2 as a measure of the correlation between X and Y, recalling that the sample correlation of X and Y is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

It can be shown that $R^2 = r$.

Most interesting questions involve more than one predictor variable, i.e. $X = (X_1, \ldots, X_p)$ with $p > 1$. This requires a **multiple linear regression model**,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon, \tag{3}$$

where β_j is the unknown average change in the response Y associated with a one unit increase in the j-th predictor X_j holding all other predictors constant.

As in simple linear regression, we estimate $\beta_0, \beta_1, \ldots, \beta_p$ by finding the values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize the least squares criterion

$$\sum_{i=1}^{n}(y_i - b_0 - x_{i1}b_1 - \cdots - x_{ip}b_p)^2,$$

with respect to $b_0, b_1, \ldots, \beta_p$. Importantly, there is only a unique minimizer of the leasts squares criterion when $p < n$, i.e. we have more observations than predictors, and when the predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ aren't too correlated with each other.

The least squares criterion evaluated at the estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ is

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - x_{i1}\hat{\beta}_1 - \cdots - x_{ip}\hat{\beta}_p)^2, \tag{4}$$

where $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}$ continues to refer to the i-th **residual**, denoted by e_i.

It is very helpful to use linear algebra when we are talking about multiple linear regression. Using linear algebra, the least squares criterion is

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}).$$

The closed form expression for the estimate vector of regression coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

Note we need to be able to invert $\boldsymbol{X}^\top \boldsymbol{X}$ to obtain the least squares estimator. This corresponds to the condition for a unique minimizer of the least squares criterion that we referenced earlier - $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible when $p < n$, i.e. we have more observations than predictors, and when the predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ aren't too correlated with each other. Having obtained $\hat{\boldsymbol{\beta}}$, a predicted value of the response $\hat{Y}$ can be obtained from the estimated coefficients as

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p.$$

Accordingly, the residual sum of squares (RSS) is

$$RSS = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^\top \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right).$$

As in simple linear regression, when the (multiple) linear regression model holds, i.e. when $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$ and ϵ is mean 0 and not correlated with $X_1, \ldots, X_p$, the least squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is unbiased. When the linear regression model holds and the errors are independent and identically distributed with finite variance $\text{Var}[\epsilon] = \sigma^2 < \infty$, the least squares estimator's variance is

$$\text{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1}.$$

We can plug in the least squares estimator for σ^2, $RSS/(n-p-1)$, to obtain the estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}$,

$$\left(\frac{RSS}{n-p-1} \right) \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1}$$

Square roots of diagonal elements of this variance-covariance matrix are the standard errors of the corresponding regression coefficent. As in the simple linear regression setting, we can use the estimated standard errors, along with the estimated coefficients, to construct confidence intervals and perform tests of null hypotheses that an individual regression coefficient is equal to a specific value. Furthermore, the estimated variance-covariance matrix can be used to estimate the variance of predictions, construct confidence intervals for predictions, and construct prediction intervals for future values.

When the linear regression model holds and the errors are independent and identically distributed with finite variance $\text{Var}[\epsilon] = \sigma^2 < \infty$, we can test the null hypothesis H_0 that that there is no relationship between the response and the predictors versus the alternative H_a that there is a relationship between at least one of the predictors. The null hypothesis H_0 corresponds to the scenario where $\beta_1 = \cdots = \beta_p = 0$, and the alternative hypothesis H_a corresponds to the scenario where at least one β_j is non-zero. To perform a level-α test of this hypothesis, we can compute the F-statistic

$$\frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)}$$

and compare it to the $1-\alpha$ quantile of an F distribution with p and $n-p-1$ degrees of freedom. We reject the null when the F-statistic is more extreme than the $1-\alpha$ quantile. Alternatively, we can compute a p-value by evaluating the probability that an F-distributed random variable with p and $n-p-1$ degrees of freedom exceeds the F-statistic.

Alternatively, we may want to test the null hypothesis H_0 that that there is no relationship between the response and a specific subset of q predictors versus the alternative H_a that there is a relationship between at least one of the q selected predictors. Without loss of generality, we can imagine that the specific set of q predictors whose relationship with the response we want to test are the last q predictors. Then the null hypothesis H_0 corresponds to the scenario where $\beta_{p-q+1} = \cdots = \beta_p = 0$, and the alternative hypothesis H_a corresponds to the scenario where at least one β_j is non-zero for $p-q+1 \leq j \leq p$. To perform a level-α test of this hypothesis, we can compute the F-statistic

$$\frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p-1)},$$

where RSS_0 refers to the residual sum of squares based on fitting the multiple linear regression model without the last q predictors. We can then compare this F-statistic to the $1-\alpha$ quantile of an F distribution with q and $n-p-1$ degrees of freedom. We reject the null when the F-statistic is more extreme than the $1-\alpha$ quantile. Alternatively, we can compute a p-value by evaluating the probability that an F-distributed random variable with q and $n-p-1$ degrees of freedom exceeds the F-statistic.

This last hypothesis test starts to lead us to the idea of model selection, which is the choice of which variables to include in a linear regression model. We could imagine selecting a model by starting with all p possible predictors, and then performing F-tests to eliminate individual variables or subsets of variables from the model. This type of strategy where we perform model selection by starting with all of the predictors is related to **backward selection**.

This can be tricky or not even feasible when we have measured a lot of variables. If $p > n$, we cannot actually compute any of these F-statistics. This motivates the idea of **forward selection**, which is to start with an empty model with no predictors and add predictors one at a time in a systematic way. Forward selection has the advantage of being feasible in situations where backward selection is not, but it also has the disadvantage of being **greedy**. This is something we'll discuss much more later in the semester.

More broadly, comparing models requires a measure of model fit. It is especially useful to have measures of model fit that do not require that models be **nested**, meaning that the variables in one model are a subset of the variables in another. Naively, we can consider RSE. There are also many other widely used measures of model fit, including **Mallow's C_p**, **Akaike information criterion (AIC)**, **Bayesian information criterion (BIC)**, and **adjusted R^2.** We will talk about these in future chapters. Note that we need to be very careful about using R^2 as a measure of model fit - it will never decrease when we add a variable to a model.

Linear models are much more flexible than one might initially realize, because they only require linearity of $f(X)$ in $\boldsymbol{\beta}$. This means that we can use linear models to relate a response to qualitative predictors/variables and we can use linear models to describe nonlinear relationships between Y and $X_1, \ldots, X_p$.

We can incorporate qualitative predictor that takes on K unique values by constructing $K-1$ binary variables

$$v_{ik} = \begin{cases} 1 & \text{if the qualitative predictor takes on the } k\text{-th unique value} \\ 0 & \text{otherwise} \end{cases}$$

for $k = 1, \ldots, K-1$. These are sometimes called **dummy variables**. Alternatively, this way of coding a qualitative predictor is sometimes called **one-hot encoding**. Note that if we want to test if there is a relationship between a qualitative predictor and the response, we need to test if all of the coefficients associated with all of the corresponding dummy variables are jointly equal to zero.

Incorporating interactions refers to incorporating an additional variable that correspond to the product of a pair of predictors, e.g.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + \epsilon.$$

This allows for the average change in response associated with changes in one variable to depend on the value of the other variable. Note that when interactions have been introduced, the question of whether or not a specific predictor is associated with the response no longer corresponds to the question of whether or not a single regression coefficient is equal to zero, but rather the question of whether or not all of the regression coefficients associated with functions of the specific predictor are jointly equal to zero.

Last, if there is evidence that the relationship between the response and a predictor is non-linear, we can introduce polynomial functions of the predictor as additional variables in the multiple linear model, e.g.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon.$$

Again, when additional polynomial functions of a predictor have been introduced, the question of whether or not a specific predictor is associated with the response no longer corresponds to the question of whether or not a single regression coefficient is equal to zero, but rather the question of whether or not all of the regression coefficients associated with functions of the specific predictor are jointly equal to zero. We will explore this idea more later in the semester.

The arguments we made to justify the value of the multiple least squares estimator assumed that the linear regression model holds. This is unrealistic! There are many ways that it can be violated, some of which have greater consequences than others. Violations of the linear regression model can often be diagnosed by examining the residuals, $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}$. In general, if the linear model holds, the residuals should not contain any meaningful trends or information.

- Trends relating the residuals to the predictors can indicate nonlinearity and motivate the inclusion of interactions and/or polynomial terms. They can also motivate transforming the response, which we haven't talked about yet.

- Residuals that become more variable as the associated values of the response increase or decrease indicate violation of the constant variance assumption. In some cases, this can be addressed by implementing weighted least squares. For instance, if the i-th value of the response y_i reflects an average over n_i units, we could perform weighted least squares with weights equal to n_i.
- If observations are related in some way a priori, e.g. correspond to consecutive points in time or space or related individuals, and residuals corresponding to more related observations tend to be more similar, the assumption of independent errors may be violated.
- Extreme residual values suggest the presence of outliers.

We haven't talked about how to address the last two issues yet.

There can also be issues that arise due to the structure of the data itself. Specifically, certain observations may have high leverage, meaning that they make an especially large contribution to estimation of the linear regression coefficients. These tend to be observations that correspond to extreme values of the predictors. Additionally, predictors can be strongly correlated with each other, which can lead to more variable and difficult to interpret estimates of the regression coefficients. This is often referred to as **collinearity** of the matrix of variables $\boldsymbol{X}$.

K-nearest neighbors regression (KNN) is one of the simplest non-parametric methods for estimating f. KNN estimates $f(x_i)$ by averaging the K observed values of the response corresponding to the K observations with predictor x_j closest to x_i.

Let $\mathcal{N}_i^{(K)}$ refer to the set of K indices of observed values of the predictor that are closest to x_i. Then the KNN estimate of $f(x_i)$ is

$$\hat{f}(x_i) = \frac{1}{K} \sum_{j \in \mathcal{N}_i^{(K)}} y_j.$$

The value of K is chosen by the user and determines the bias and variance of the estimate of f. Smaller K correspond to less biased but more variable estimates, whereas larger K correspond to (potentially) more biased and less variable estimates.

Bias seems quite undesirable, but there are several disadvantages to using KNN. Specifically, KNN specifically and non-parametric more generally, are less interpretable and less amenable to inference, e.g. implementation of hypothesis tests. Furthermore, the performance of KNN deteriorates rapidly as p, the number of predictors and/or dimension of X, increases. This is because it is harder to find neighbors in high dimensions. This is related to the idea of the **curse of dimensionality**.