3.2 Multiple Linear Regression

Zongyi Liu

2023-05-13

3.2 Multiple Linear Regression

In reality, we typically have more than one predictors in the setting.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

By adding numbers of predictors, we can increase the accuracy of the model

Simple regression of `sales` on `radio`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

Simple regression of `sales` on `newspaper`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	0.00115

We can also have another form:

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

3.3.1 Estimating the Regression Coefficients

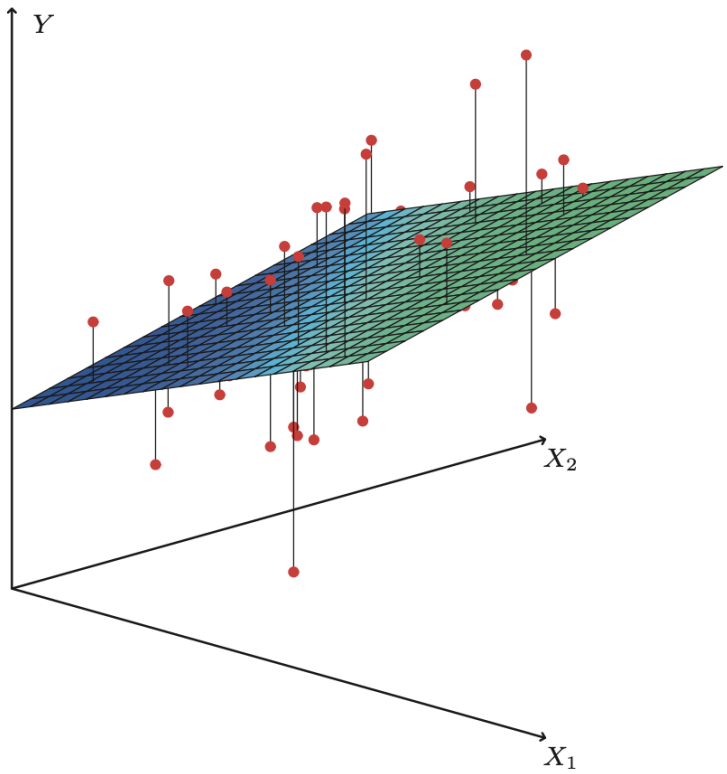
Given $\hat{\beta_0}, \hat{\beta_1}, ...\hat{\beta_p}$, we can make predictions using the formula:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

We then will try to minimize the sum of squared residuals:

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \end{aligned}$$

In plot, we can view it vividly:



3.2.2 Some Important Questions

- Is at least one of the predictors X1, X2,. . . , Xp useful in predicting the response?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

First: Is There a Relationship Between the Response and Predictors?

In the multiple regression setting with p predictors, we need to check whether all of the regression coefficients are zero, i.e. whether $\beta_1 = \beta_2 = ... = \beta_p = 0$. So the null hypothesis becomes:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

and the alternative:

$$H_a : \text{ at least one } \beta_j \text{ is non-zero.}$$

The hypothesis test is performed by computing the F-statistic:

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

We can also show that

$$E\{\text{RSS}/(n-p-1)\} = \sigma^2$$

and that, provided H_0 is true, we have

$$E\{(\text{TSS}-\text{RSS})/p\} = \sigma^2.$$

Therefore, if there is no relationship between the response and predictors, we would expect the F-stats to have a value close to 1. On the other hand, if H_a is true, we would expect the F-stats to be greater than 1.

Second: Deciding on Important Variables

It is important to know which variable is important in the model if all of them satisfied the p-value examination.

There are three approaches for this task:

- **Forward selection**. We begin with the null model——a model that contains an intercept but no predictors.
- **Backward selection**. We start with all variables in the model, and remove the variable with the largest p-value (the variable that is the least statistically significant).
- **Mixed selection**. This is a combination of forward and backward selection. We start with no variables in the model, and as with forward selection, we add the variable that provides the best fit. We continue to add variables one-by-one.

Third: Model Fit

There are two most common numerical measures of model fit: RSE and R^2.

First, for R^2,in simple regression, it equals to the correlation of the response and the variable. In multiple regression, it equals to $Cor(Y, \hat{Y})^2$. If R^2 is close to 1, then we can say that it can be largely explained by the responses.

In general RSE is defined as

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}},$$

Fourth: Predictors

Once we have fit the multiple regression model, it is straightforward to apply in order to predict the response Y on the basis of a set of values for predictors $X_1, X_2, ..., X_p$, but there are still three uncertainties:

- **Inaccuracy in the coefficient estimates**. The coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, ...\hat{\beta}_p$ are estimates for $\beta_0, \beta_1, ...\beta_p$. In other words, there is an inaccuracy in the coefficient estimates. We can compute a confidence interval in order to determine how close $\hat{Y}$ will be to $f(X)$.
- **Model bias**. when we use the linear model to approximate, we are assuming that the linear model is correct, but there are possibilities that it's not.
- **The existence of random error** ϵ. Even if we know $f(X)$, we might still not predict the true value accurately. We will use the prediction interval to solve this problem.