6.2: Linear Regression

Navona Calarco

The University of Toronto

Motivation

Throughout this Module we will be making use of the `Boston` dataset in the R package `ISLR2`. We can install the package in R and add it to our library:

```
install.packages("ISLR2")
library(ISLR2)
```

The `Boston` dataset contains housing values in 506 Boston suburbs along with 12 other variables associated with the suburbs. To name a few,

- `rm`: average number of rooms per dwelling
- `nox`: nitrogen oxides concentration (parts per 10 million)
- `lstat`: percent of households with low socioeconomics status

We can take `medv`, the median value of owner-occupied homes in \$1000s, to be the response variable Y and the 12 other variables to be the predictors $X = (X_1, \ldots, X_{12})$.

Motivation

There may be some specific question we'd like to address

- Is there a relationship between the 12 variables and housing price?
    - Does the data provide evidence of an association?
- Are all of the 12 variables associated with housing price?
    - Perhaps only a few of the variables have an effect on housing price.
- How accurate are the predictions for housing prices based on these variables?
- Is the relationship between the variables and housing price linear?
    - Perhaps we can transform some variables to make the relationship linear.

All of these questions can be answered using linear regression!

Simple Linear Regression

Simple linear regression uses a *single* predictor variable X to predict a *quantitative* response Y by assuming the relationship between them is linear.

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

- β_0 and β_1 are the model **parameters** which are unknown.
- β_0 is the intercept term and β_1 is the slope term.

We can use the training data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ and predict future responses

$$\hat{y} \approx \hat{\beta}_0 + \hat{\beta}_1 X$$

Estimating the Coefficients

Suppose we have n observations in our training data which each consists of a measurement for X and Y represented by

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

We want to find estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$ such that for all $i = 1, \ldots, n$

$$y_i \approx \hat{y}_i$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the prediction for y_i given x_i.

The most common method used to measure the difference between y_i and $\hat{y}_i$ is the least squares criterion. The idea being that we want to find the $\hat{\beta}_0$ and $\hat{\beta}_1$ that give us the smallest difference.

Least Squares Criterion

We define the ith **residual** to be the difference between the ith observed response value and the ith predicted response value:

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

The **residual sum of squares** (RSS) is the following

$$\text{RSS} = e_1^2 + \cdots + e_n^2 = \left(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \cdots + \left(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right)^2$$

The RSS is minimized by the estimates below (where $\bar{x}$, $\bar{y}$ are the sample means):

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

So $\hat{\beta}_1$ and $\hat{\beta}_0$ definte the least squares coefficient estimates

Assessing the Accuracy of the Coefficient Estimates

Recall from section 6.1 that we assume the true relationship between the predictor X and the response Y is

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

where f is an unknown function and ϵ is the random error with mean zero. By assuming f is linear, we obtain

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

Now suppose we have the least squares coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, so

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

We would like to assess the how close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true parameter values β_0 and β_1.

Standard Error

We can compute the **standard erorrs** associated with $\hat{\beta}_0$ and $\hat{\beta}_1$ with the following:

$$\mathrm{SE}\left(\hat{\beta}_0\right)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right], \qquad \mathrm{SE}\left(\hat{\beta}_1\right)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

where $\sigma^2 = \mathrm{Var}(\epsilon)$ and is usually unknown. Luckily, σ can be estimated from the data using the **residual standard error** (RSE)

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

The standard errors for $\hat{\beta}_0$ and $\hat{\beta}_1$ can be used to compute confidence intervals of the estimates or perform hypothesis tests on the coefficients.

Hypothesis Tests on the Coefficients

Once we have the standard errors, we can perform a hypothesis test on the coefficients to determine whether there is a relationship between X and Y. The null hypothesis is

$$H_0: \text{ There is no relationship between } X \text{ and } Y$$

and the alternative hypothesis is

$$H_a: \text{ There is some relationship between } X \text{ and } Y$$

Mathematically, this is

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_a : \beta_1 \neq 0$$

since if $\beta_1 = 0$ then $Y = \beta_0 + \epsilon$ so Y is not associated with X.

Hypothesis Tests on the Coefficients

In order to test the null hypothesis, we need to determine whether $\hat{\beta}_1$ is sufficiently far from zero. The **t-statistic**

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

measures the number of standard deviations that $\hat{\beta}_1$ is away from 0. The p-value can be computed from the t-statistic which will allow us to either accept or reject our null hypothesis.

Assessing the Accuracy of the Model

The quality of the linear regression fit is often assessed with the residual standard error (RSE) and the R^2 statistic.

- The RSE gives an absolute measure of lack of fit of the model to the data.
- The $\mathbf{R^2}$ **statistic** measures the proportion of variability in Y that can be explained by X.

We've already seen how the RSE is computed from the RSS and the R^2 statistic can be computed using

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

where $\mathrm{TSS} = \sum (y_i - \bar{y})^2$ is the **total sum of squares** which measures the amount of variability in the responses before regression is performed.

Simple Linear Regression Summary

Simple linear regression uses a single predictor variable X to predict a response Y with

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

- β_0, β_1 are estimated by minimizing the residual sum of squares (RSS)
- The standard error (SE) of the coefficient estimates is a measure of accuracy.
- The residual standard error (RSE) gives a measure of lack of fit of the model to the data.
- The R^2 statistic measures the proportion of variability explained by the regression. - A hypothesis test on β_1 indicates whether there is a relationship between X and Y.
- The `lm()` function in R can be used to perform all of these tasks!

Any Questions?

Exercises: Simple Linear Regression

Open the Linear Regression Exercises R Markdown file.

- Go over the “Simple Linear Regression” section together as a class.

Multiple Linear Regression

Suppose we have n observations in our data each consisting of p predictor values and one response value. That is,

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \text{ where } x_i = (x_{i1}, x_{i2}, \ldots, x_{ip}).$$

We want to fit this data with a linear model. We can extend simple linear regression to accommodate p predictors.

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

We interpret β_j as the average effect on Y of one unit increase in X_j while holding all other predictors fixed.

As with simple linear regression, the coefficients $\beta_0, \ldots, \beta_p$ are unknown and must be estimated.

Estimating the Coefficients

We want to find estimates $\hat{\beta}_0, \ldots, \hat{\beta}_p$, so that predictions for the response can be made using

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

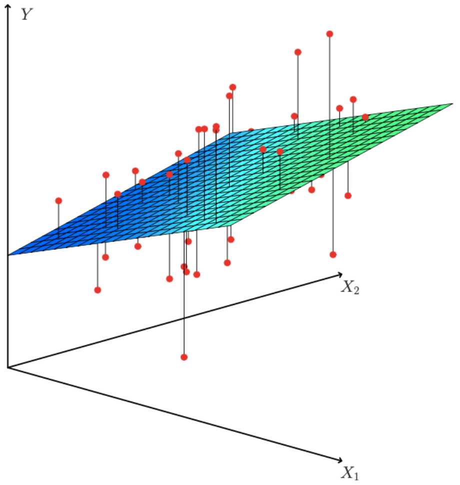
The least squares approach is used again in this case to estimate the p parameters. That is, we choose $\beta_0, \ldots, \beta_p$ to minimize the sum of the squared residuals

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

The equations for $\hat{\beta}_0, \ldots, \hat{\beta}_p$ which minimize the *RSS* are complicated and not entirely important since there are functions that perform the computation for us in R and Python.

Least Squares Regression Plane

The figure shows the relationship between two predictor variables and a response variable. Linear regression in this case gives a plane fit by minimizing the squared vertical distance between the observations and the plane.



Important Questions

When working with multiple linear regression, we are often interested in several important questions.

- Is there a relationship between the response and the predictors?
- How well does the model fit the data?
- Given a set of predictor values, what is the predicted response, and how accurate is our prediction? We will go over the methods for answering each of these questions.

Hypothesis Test for Parameters

One: *Is there a relationship between the response and the predictors?*

We can address this question by testing whether the regression coefficients are far enough from zero.

Our null hypothesis and alternative hypothesis are the following:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \qquad H_a : \beta_j \neq 0 \text{ for some } j$$

The hypothesis can be tested with the **F-statistic**, which is defined by

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

If the F-statistic is much larger than 1 we reject the null hypothesis and conclude there is a relationship between at least on the the predictors and the response. If the F-statistic is close to 1, the p-value can be computed to determine the outcome.

RSE and R^2

Two: *How well does the model fit the data?*

The RSE and R^2 are measures of the model fit. In the multiple linear regression context, R^2 is the square of the correlation between the response and the fitted linear model. That is, $R^2 = \text{Cor}(Y, \hat{Y})^2$.

The RSE is defined by

$$\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{n - p - 1}} \qquad \text{where } n = \#\text{ observations}, \; p = \#\text{ predictors}$$

Important considerations as the number of variables in the model increases:

- R^2 will increase even if the new variables have a weak association with the response.
- *RSS* of the training data will decrease, but not necessarily that of the testing data.
- *RSE* will increase if the decrease in *RSS* is small relative to the increase in p.

Prediction Accuracy

Three: *Given a set of predictor values, what is the predicted response, and how accurate is our prediction?* Once we have fit the multiple regression model, the response Y is predicted by

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

There are three types of uncertainty associated with the prediction $\hat{Y}$

1. The **reducible error** arising from the inaccuracy of the coefficient estimates.
2. The reducible error stemming from the assumption that the relationship between Y and X is linear; **model bias**.
3. The **irreducible error** from the random error associated with the true response $Y = f(x) + \epsilon$.

We can address how much Y will vary from $\hat{Y}$ using prediction intervals.

Exercises: Multiple Linear Regression

Open the Linear Regression Exercises R Markdown file.

- Go over the “Multiple Linear Regression” section together as a class.