2: Linear Regression

```
$ echo "Data Science Institute"
```

Motivation

Throughout this Module we will be making use of the `Boston` dataset in the Python package `ISLP`. We can use the terminal to install the Python package and use the `load_data` function from the `ISLP` package to load the `Boston` dataset:

```
from ISLP import load_data
Boston = load_data("Boston")
```

Motivation

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 socioeconomic 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

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

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

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

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

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

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.

Breakout Room: What do you think the Hypothesis Test is?

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.

Any Questions?

Exercises: Simple Linear Regression

Open the Linear Regression Jupyter Notebook file.

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

References

Chapter 3 of the ISLP book:

James, Gareth, et al. "Linear Regression." An Introduction to Statistical Learning: with Applications in Python, Springer, 2023.