Linear Regression

Supervised Learning

- In data science, many applications involve making predictions about the outcome y based on a number of predictors x.

- Often we assume models of the form

$$y \approx f(x)$$

  where $f(x)$ is a function that maps the predictor(s) to the outcome.

- One example is the linear regression model.

Supervised Learning

- It is called 'supervised learning' because we have data on both the outcome y and the predictor x.

- Therefore, the data can 'teach' us, given a certain predictor value for x, what is the most likely corresponding outcome y.

Linear regression

- Linear regression is an analytical technique used to model the relationship between several input variables and a continuous outcome variable.

- A key assumption is that the relationships between the input variables and the outcome variable are linear.

- For example, in simple linear regression with only one predictor, we assume a model of the form
$$y \approx f(x) = \beta_0 + \beta_1 x.$$

Examples

- **Demand forecasting**: Businesses and governments can use linear regression models to predict demand for goods and services.
- E.g., coffee shops can appropriately prepare for the predicted type and quantity of food that customers will consume based upon the weather, the day of the week, whether an item is offered as a special, the time of day, and the reservation volume.

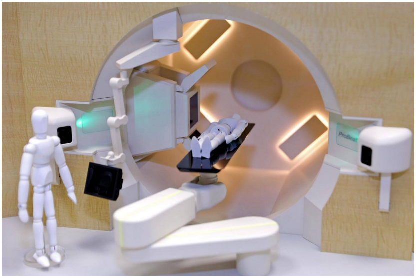
Source: The Business Times

Examples

- Similar forecasting models can be built to predict taxi demand, emergency room visits, and ambulance dispatches.

Source: The Business Times

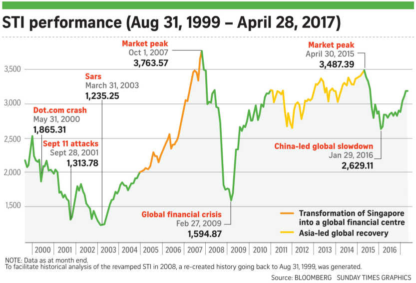
Examples



Source: The Business Times

- **Medical**: Linear regression model can be used to analyze the effect of a proposed radiation treatment on reducing tumor sizes.
- Multiple input variables might include duration of a single radiation treatment, frequency of radiation treatment, and patient attributes such as age or weight.

Examples



Source: The Business Times

- **Finance**: Linear regression is used to model the relationships between stock market prices and other variables such as economic performance, interest rates and geopolitical risks.

Examples

Source: The Business Times

- **Pharmaceutical Industry:**: Linear regression model can be used to analyze the clinical efficacy of drugs.
- Input variables may include age, gender and other patient characteristics such as blood pressure and blood sugar level.

Examples

Source: The Business Times

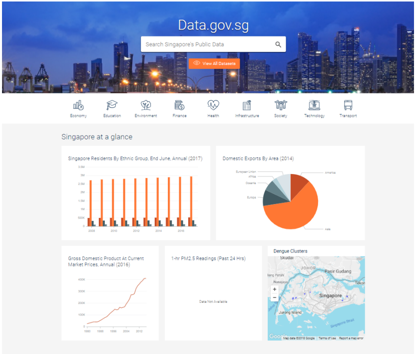
- **Real estate**: Linear regression analysis can be used to model flat's price as a function of the floor area.
- Such a model helps set or evaluate the list price of a flat on the market.
- The model could be further improved by including other input variables such as number of bathrooms, number of bedrooms, district rankings, etc.

Data on HDB Resale Flats

Source: The Business Times

- Data on resale HDB prices based on registration date is publicly available from `https://data.gov.sg/dataset/resale-flat-prices`.
- We have extracted a subset of all the resale records from March 2012 to December 2014 based on registration date.
- Available as the data set `hdbresale_reg.csv` on the course website.

GovTech Datasets



Source: data.gov.sg

- `https://data.gov.sg` hosts many publicly available data sets for data analytics.

HDB Resale Flats

```
> resale = read.csv("C:/Data/hdbresale_reg.csv")
> head(resale[ ,2:7]) # 1st column indicates ID of flats
    month         town flat_type block     street_name storey_range
1 2012-03 CENTRAL AREA    3 ROOM   640       ROWELL RD     01 TO 05
2 2012-03 CENTRAL AREA    3 ROOM   640       ROWELL RD     06 TO 10
3 2012-03 CENTRAL AREA    3 ROOM   668      CHANDER RD     01 TO 05
4 2012-03 CENTRAL AREA    3 ROOM     5 TG PAGAR PLAZA     11 TO 15
5 2012-03 CENTRAL AREA    3 ROOM   271        QUEEN ST     11 TO 15
6 2012-03 CENTRAL AREA    4 ROOM  671A      KLANG LANE     01 TO 05
```

HDB Resale Flats

```
> head(resale[ ,8:11])
  floor_area_sqm flat_model lease_commence_date resale_price
1             74    Model A                1984       380000
2             74    Model A                1984       388000
3             73    Model A                1984       400000
4             59   Improved                1977       460000
5             68   Improved                1979       488000
6             75    Model A                2003       495000
```

- Suppose we are interested to build a linear regression model that estimates a HDB unit's resale price as a function of floor area in square meters.

- How to form such function?

Simple Linear Regression (SLR)

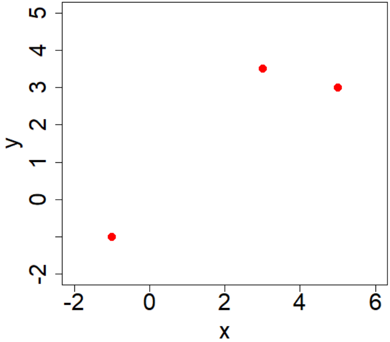
- Suppose we have three observations. Each observation has an outcome y and an input variable x.

- We are interested in the linear relationship

$$y_i \approx \beta_0 + \beta_1 x_i$$

- Since there is only **one input variable**, this is an example of **simple linear model**.

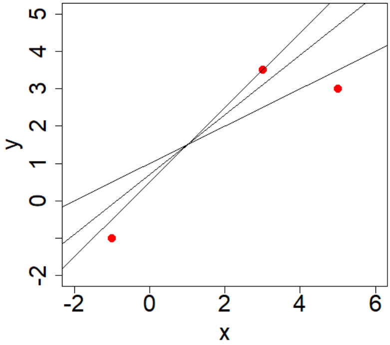
Ordinary Least Squares for SLR



i	x_i	y_i
1	-1	-1
2	3	3.5
3	5	3

Plot of the three data points.

Ordinary Least Squares for SLR

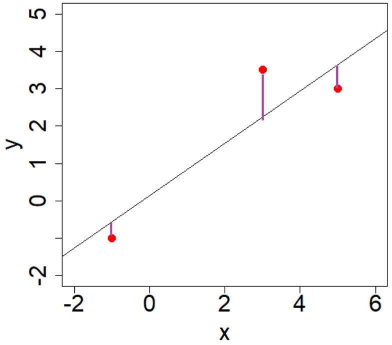


- We are interested in the linear relationship

$$y_i \approx f(x_i) = \beta_0 + \beta_1 x_i$$

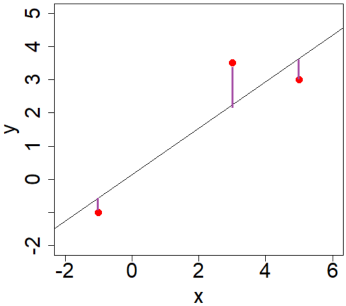
- There are many different straight lines that can be used to model x and y, as shown in the plot.

Ordinary Least Squares for SLR



- Intuitively, we want the line to be as close to the data points as possible.
- This "closeness" can be measured in terms of the vertical distance between each point to the line.
- The line that is closest to the data points is chosen as best-fitting line. The values of intercept and slope of it are picked for β_0 and β_1.

Ordinary Least Squares for SLR



i	x_i	y_i	$\beta_0 + \beta_1 x_i$	residual: $e_i = y_i - (\beta_0 + \beta_1 x_i)$
1	-1	-1	$\beta_0 + (-1)\beta_1$	$-1 - (\beta_0 + (-1)\beta_1) = -1 - \beta_0 + \beta_1$
2	3	3.5	$\beta_0 + (3)\beta_1$	$3.5 - (\beta_0 + (3)\beta_1) = 3.5 - \beta_0 - 3\beta_1$
3	5	3	$\beta_0 + (5)\beta_1$	$3 - (\beta_0 + (5)\beta_1) = 3 - \beta_0 - 5\beta_1$

Ordinary Least Squares for SLR

- The residual for each point may be positive or negative.
- We do not want the residuals to “cancel off” each other, so we square each of them, leading to the squared residuals.

i	residual: $e_i = y_i - (\beta_0 + \beta_1 x_i)$	squared residual: e_i^2
1	$-1 - (\beta_0 + (-1)\beta_1) = -1 - \beta_0 + \beta_1$	$[-1 - \beta_0 + \beta_1]^2$
2	$3.5 - (\beta_0 + (3)\beta_1) = 3.5 - \beta_0 - 3\beta_1$	$[3.5 - \beta_0 - 3\beta_1]^2$
3	$3 - (\beta_0 + (5)\beta_1) = 3 - \beta_0 - 5\beta_1$	$[3 - \beta_0 - 5\beta_1]^2$

Ordinary Least Squares for SLR

- To express the total magnitude of the deviations, we sum up the squared residuals for all the data points, Residual Sum of Squares, abbreviated as **RSS**, some might denote it as SS_{res}, sum of squared residuals.

- For the 3 data points, we have

$$\begin{aligned} RSS &= e_1^2 + e_1^2 + e_3^2 \\ &= [-1 - \beta_0 + \beta_1]^2 + [3.5 - \beta_0 - 3\beta_1]^2 + [3 - \beta_0 - 5\beta_1]^2 . \end{aligned}$$

Ordinary Least Squares for SLR

- We now need to find the values of β_0 and β_1 such that RSS is minimized, where

$$RSS = \left[-1 - \beta_0 + \beta_1\right]^2 + \left[3.5 - \beta_0 - 3\beta_1\right]^2 + \left[3 - \beta_0 - 5\beta_1\right]^2 .$$

- The whole process (from getting each e_i, e_i^2, RSS and minimize it to get the values of β_0 and β_1) is known as the method of ordinary least squares (OLS).

Ordinary Least Squares for SLR

- Consider RSS as a function in terms of β_0 and β_1. Let's call it $h(\beta_0, \beta_1)$.

- To find the minimum of $h(\beta_0, \beta_1)$, we first take the derivative of it w.r.t β_0 while holding β_1 constant, and then take the derivative of it w.r.t β_1 while holding β_0 constant.

$$\frac{\partial h(\beta_0, \beta_1)}{\partial \beta_0} = -11 + 6\beta_0 + 14\beta_1.$$

$$\frac{\partial h(\beta_0, \beta_1)}{\partial \beta_1} = -53 + 14\beta_0 + 70\beta_1.$$

Ordinary Least Squares for SLR

- Lastly, setting both the derivative to zero, we have a system of two equations

$$-11 + 6\beta_0 + 14\beta_1 = 0$$
$$-53 + 14\beta_0 + 70\beta_1 = 0$$

- Solving it, we have the solution which is the *least squares* estimates

$$\beta_0 \approx 0.1250$$
$$\beta_1 \approx 0.7321$$

- We usually add the "hat" sign on top of the parameter to denote **estimated value** of the parameter, so the least squares estimates are

$$\hat{\beta}_0 = 0.1250$$
$$\hat{\beta}_1 = 0.7321.$$

Ordinary Least Squares for SLR in General

- In the previous slides, we had a specific example, a data with 3 points, and had manually practiced the OLS method.
- We now generalize OLS to a data set which has 2 variables X and Y with n observations $(x_1, y_1), ..., (x_n, y_n)$.
- The simple model (straight line) has the form

$$y_i \approx \beta_0 + \beta_1 x_i, \ i = 1, ..., n.$$

- The residual is then

$$e_i = y_i - (\beta_0 + \beta_1 x_i), \ i = 1, ..., n$$

Ordinary Least Squares for SLR in General

- The residuals sum of squares is then

$$RSS = \sum_{i=1}^{n} e_i^2 = \Big[y_i - (\beta_0 + \beta_1 x_i)\Big]^2.$$

- Take derivative of RSS w.r.t β_0 and β_1, one at a time.

$$\frac{\partial RSS}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial RSS}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i$$

- The least square estimate of β_0 and β_1, $\hat{\beta_0}$ and $\hat{\beta_1}$, are the solution when we set the derivative to zero.

$$\hat{\beta_0} + \hat{\beta_1}\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} y_i = 0 \qquad (1)$$

$$\hat{\beta_0}\frac{1}{n}\sum_{i=1}^{n} x_i + \hat{\beta_1}\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n} y_i x_i = 0 \qquad (2)$$

Ordinary Least Squares for SLR in General

- Denote $\bar{y} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i$ and $\bar{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$.

- From (1), we have $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, and replace this $\hat{\beta}_0$ into (2), we have

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

Ordinary Least Squares in R

- We can check that the least squares estimates we computed manually are equivalent to those returned by the lm() function in R:

```
> x = c( -1, 3, 5)
> y = c( -1, 3.5 , 3)
> lm(y~x)
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     0.1250       0.7321
```

- We now can write the **fitted model** as

$$\hat{y} = 0.125 + 0.7321x$$

Ordinary Least Squares in R

- With the fitted model, we now can obtain the fitted outcome (predicted outcome) value $\hat{y}$ given any value of the predictor, x.

- For example, if $x = 2$, then the fitted value for the outcome is

$$\hat{y} = 0.125 + 0.7321 \times 2 = 1.589.$$

- In R, we use function predict().

```
> M = lm(y~x)                # M = name of the fitted model
> new = data.frame(x = 2) # create dataframe of new point
> predict(M, newdata = new)
       1
1.589286
```

HDB Resale Flats Data Set

- We now can answer the question in slide 16 on building a linear regression model that estimates a HDB unit's resale price as a function of floor area in square meters.

```
> price = resale$resale_price
> area = resale$floor_area_sqm
> lm(price~area)$coef # coefficients of the model
(Intercept)        area
 115145.730    3117.212
```

- The fitted model is then

$$\hat{y} = 115145.73 + 3117.21 * \text{area}$$

where y is the resale price of a flat, in SGD.

Goodness-of-fit of a Model

- The goodness-of-fit of a model could be accessed by some measures. In this course, we consider only two **basic** measurements:

    - Residual Standard Error (RSE)

    - Coefficient of determination, R^2.

Residual Standard Error (RSE)

- RSE in **simple** linear regression is defined as

$$RSE = \sqrt{\frac{1}{n-2}RSS} \quad \text{where} \quad RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Larger RSE indicates poorer model fit.

Residual Standard Error (RSE)

- Calculating RSE directly as its formula above:

```
> sqrt(sum((y - M$fitted)^2)/(length(y) - 2))
[1] 1.469937
```

- Or reading it off from the R output, at "Residual standard error".

```
> summary(M)
Call:
lm(formula = y ~ x)

Residuals:
      1       2       3 
-0.3929  1.1786 -0.7857 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1250     1.1621   0.108    0.932
x             0.7321     0.3402   2.152    0.277

Residual standard error: 1.47 on 1 degrees of freedom
Multiple R-squared:  0.8224,	Adjusted R-squared:  0.6448 
F-statistic: 4.631 on 1 and 1 DF,  p-value: 0.2769
```

Coefficient of Determination R^2

- The quantity R^2 is defined as

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

  where $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the *total sum of squares*.

- TSS measures the total variance in the response in the given data, and can be thought of as the amount of variability inherent in the response before the regression is performed.

- For a given data, TSS is fixed, it does not depend on the model.

- In contrast, RSS measures the amount of variability that is left unexplained after performing the regression.

Coefficient of Determination R^2

- R^2 measures the proportion of variability in the response Y that is explained using by the fitted model.

- Larger R^2 indicates better model fit.

- Deriving R^2 directly

```
> RSS =sum ((y- M$fitted )^2)
> TSS = var(y)*(length(y) -1)
> R2 = 1 - RSS/TSS; R2
[1] 0.822407
```

- Or getting the value of R^2 from the model output:

```
> summary(M)$r.squared
[1] 0.822407
```

Settings

- Suppose we have n observations. Each observation has an outcome y and multiple input variables $x^1, ..., x^p$.

- We are interested in the linear relationship

$$y \approx \beta_0 + \beta_1 x^1 + \beta_2 x^2 + ... + \beta_p x^p$$

  or equivalently

$$y_i \approx \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + ... + \beta_p x_i^p, \;\; i = 1, ..., n.$$

- The OLS method minimizes the RSS given by

$$RSS = \sum_{i=1}^{n} \left[y_i - \left(\beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + ... + \beta_p x_i^p \right) \right]^2.$$

- We rely on R to derive the minimizers, $\hat{\beta}_0, ..., \hat{\beta}_p$.

MLR in R

- The least squares estimates, $\beta_0, \beta_1, \beta_2, ..., \beta_p x^p$, are returned by the lm() function in R.

- Consider a simulated data with x_1, x_2 and response y where y is created as $(1 + 2x_1 - 5x_2)$ with some noise added.

```
> set.seed(250)
> x1 = rnorm(100)
> x2 = rnorm(100)
> y = 1 + 2*x1 -5*x2+ rnorm(100)
```

- We now fit a linear model, $y \sim x_1 + x_2$.

```
> lm(y~x1+x2)
Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
     0.9362       1.7649      -4.9560
```

MLR in R

- Visualize the data points in a 3D plot and the fitted plane added.

```
> #instal.packages("rgl")
> library(rgl)
> M.2 = lm(y~x1+x2)
> # 3D plot to illustrate the data points
> plot3d (x1 , x2 , y, xlab = "x1", ylab = "x2", zlab = "y",
+         type = "s", size = 1.5 , col = "red")
> coefs = coef(M.2)
> a <- coefs[2] # coef of x1
> b <- coefs[3] # coef of x2
> c <- -1       # coef of y in: ax1 + bx2 -y + d = 0.
> d <- coefs[1] # intercept
> planes3d (a, b, c, d, alpha = 0.5) # the plane is added.
```

Adjusted R^2 in MLR

- A multiple linear model has R^2 which is defined exactly as in simple linear regression, and its meaning remains the same.
- R^2 can be inflated simply by adding more regressors to the model (even insignificant terms).
- However, for the similar accuracy, a simpler model is preferred, hence we have adjusted R^2, denoted as R^2_{adj}-which penalizes you for adding regressors of too little help to the model.

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}.$$

- When comparing two models of the same response, the model with larger R^2_{adj} is preferred.

MLR for HDB Resale Flats

- Do note that, HDB flats are sold on 99-year leases. Hence, the older lease commence date (it's the date the first owner took the key from HDB), usually the lower resale price, given other conditions are similar.

- Hence, we may consider the number of years from the lease commence date till this year as a quantitative regressor, called "years_left".

```
> years_left = 2022 - resale$lease_commence_date
```

- Can you try to fit a linear model for the resale price with two regressors, floor area and the years left?

- Report the fitted model and the goodness-of-fit of the model. Compared to the simple model (with only floor area as the regressor), which model would you prefer?