R. Nazim Khan
Department of Mathematics and Statistics
nazim.khan@uwa.edu.au

The University of Western Australia

template=wideslide,tocsection=false

2.1 Introduction

Regression is the study of relationships between variables, and is a very importantm statistical tool because of its wide applicability. **Simple linear regression** involves only two variables:

X = independent or explanatory variable;
Y = dependent or response variable;

and they are related by a straight line.

The observations are $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.

Example 2.1

X = height of person
Y = weight

People of the same height can have different weights.

On average as height increases, weight also increases. For any given height, there is random variation in weight. That is, given the height, the weight has random variations from some mean weight.

The Model

Linear regression assumes that on average, Y is a linear function of X, that is

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X.$$

THE MODEL

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ i = 1, 2, \cdots, n$$

where

$$\begin{aligned} Y_i &= \text{response (or dependent) variable,} \\ X_i &= \text{explanatory (or independent) variable,} \\ \beta_0 &= \text{ intercept,} \\ \beta_1 &= \text{ slope,} \\ \epsilon_i &= \text{ error or random variation} \end{aligned}$$

Model Assumptions

Observed data (X_i, Y_i), and

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

We treat Y_i as *random variables* corresponding to observations X_i

ASSUMPTIONS

1. A linear model is appropriate, that is, $\mathrm{E}\left(Y_i\right) = \beta_0 + \beta_1 X_i$.
2. The error termsϵ_i are normally distributed.
3. The error terms ϵ_i have constant variance.
4. The error terms ϵ_i are uncorrelated.

That is,

$$\epsilon_i \overset{i.i.d.}{\sim} N\left(0, \sigma^2\right)$$

Variance

$$\begin{aligned}\mathrm{Var}(Y_i) &= \mathrm{Var}(\beta_0 + \beta_1 X_i + \epsilon_i)\\ &= \mathrm{Var}(\epsilon_i)\\ &= \sigma^2,\end{aligned}$$

which does not depend on *i*, that is, homogeneous variance, or constant variance, or homoscedacity.

template=wideslide,tocsection=false

2.2 Parameter Estimation — Method of Least Squares

The model has three parameters, β_0, β_1 and σ^2, and these need to be estimated from the data. Denote by B_0 and B_1 respectively the estimates of β_0 and β_1.

$$\begin{aligned}
\text{Fitted value } \ \hat{Y}_i &= B_0 + B_1 X_i \\
\text{Residual } \ r_i &= Y_i - \hat{Y}_i \\
&= Y_i - (B_0 + B_1 X_i)
\end{aligned}$$

We use the method of least squares to estimate B_0 and B_1 so as to minimise $\sum_{i=1}^{n} r_i^2$.

2.2 Parameter Estimation — Method of Least Squares (ctd)

Now

$$\begin{aligned}\sum_{i=1}^{n} r_i^2 &= \sum_{i=1}^{n} [Y_i - (B_0 + B_1 X_i)]^2 \\ &= \sum_{i=1}^{n} [Y_i - B_0 - B_1 X_i]^2\end{aligned}$$

template=wideslide,tocsection=false

2.2 Parameter Estimation — Method of Least Squares (ctd)

2.2 Parameter Estimation — Method of Least Squares (ctd)

Define

$$\mathrm{SS}_Y = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$\mathrm{SS}_X = \sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\mathrm{SS}_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})Y_i.$$

Then

$$\boxed{\sum_{i=1}^{n} r_i^2 = B_1^2\mathrm{SS}_X - 2B_1\mathrm{SS}_{XY} + \mathrm{SS}_Y}$$

template=wideslide,tocsection=false

Parameter estimates

This is a quadratic function in B_1, and has a minimum at $\hat{B}_1 = \dfrac{2\mathrm{SS}_{XY}}{2\mathrm{SS}_X}$. Thus

$$\boxed{\hat{B}_1 = b_1 = \frac{\mathrm{SS}_{XY}}{\mathrm{SS}_X} \qquad \hat{B}_0 = b_0 = \bar{Y} - \hat{B}_1\bar{X}}$$

template=wideslide,tocsection=false

Exercise

Show that

$$1.\quad \mathrm{SS}_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$
$$= \sum_{i=1}^{n}(X_i - \bar{X})Y_i = \sum_{i=1}^{n}(Y_i - \bar{Y})X_i$$
$$= \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$$

$$2.\quad \mathrm{SS}_{X} = \sum_{i=1}^{n}(X_i - \bar{X})^2$$
$$= \sum_{i=1}^{n}(X_i - \bar{X})X_i = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

template=wideslide,tocsection=false

Exercise (cont'd)

Show that

$$3.\quad \mathrm{SS}_Y = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})Y_i = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2.$$

Note

$$B_0 = \bar{Y} - B_1\bar{X} \Rightarrow \bar{Y} = B_0 + B_1\bar{X}.$$

That is, the point $(\bar{X}, \bar{Y})$ satisfies the equation of regression, i.e., $(\bar{X}, \bar{Y})$ ALWAYS lies on the line of regression.

2.3 Parameter Estimation — Least Squares Estimate

$$\sum_{i=1}^{n} r_i^2 = f(B_0, B_1) = \sum_{i=1}^{n} [Y_i - (B_0 + B_1 X_i)]^2$$

We need to minimise the residual sum of squares, and this can be achieved using differential calculus. We consider $r_i^2 = f(B_0, B_1)$. Thus

2.3 Parameter Estimation (ctd)

template=wideslide,tocsection=false

2.3 Parameter Estimation (ctd)

2.3 Parameter Estimation (ctd)

Example 2.2: Pharmex Data

X = Promotional expenditure as a percentage of those of the leading competitor.

Y = Sales as a percentage of those of the leading competitor.

Summary Statistics

$\sum_{i=1}^{50} x_i = 4894$ $\quad$ $\sum_{i=1}^{50} y_i = 4987$ $\quad$ $\sum_{i=1}^{50} x_i^2 = 482,764$

$\sum_{i=1}^{50} y_i^2 = 502,201$ $\quad$ $\sum_{i=1}^{50} x_i y_i = 490,978$

Example 2.2: Pharmex Data (cont'd)

Interpreting Regression Coefficients

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

where

$$\beta_0 = \text{intercept}$$

$$\beta_1 = \text{slope of line}$$

β_0 is the average value of y when $x = 0$. Usually the cost/return of doing nothing. Does not always have a meaningful interpretation.

β_1 is the average **change** in y (increase if $\beta_1 > 0$, decrease if $\beta_1 < 0$) for an increase of 1 unit in x.

For Pharmex, if nothing is spent on promotions, the sales are 25.13% of that of the competitor, on average. For every percent increase in promotions, sales increase on average by 0.7623% of that of the competitor.

2.3 Partitioning the sum of squares

Define

$$\text{Residual SS} = \text{SS}_{\text{Res}} = \sum_{i=1}^{n} r_i^2.$$

$$\begin{aligned}
\text{Now} \sum_{i=1}^{n} r_i^2 &= b_1^2 \text{SS}_X - 2b_1 \text{SS}_{XY} + \text{SS}_Y \\
&= b_1^2 \text{SS}_X - 2b_1 \frac{\text{SS}_{XY}}{\text{SS}_X} \text{SS}_X + \text{SS}_Y \\
\text{SS}_{\text{Res}} &= \text{SS}_Y - b_1^2 \text{SS}_X
\end{aligned}$$

Define

$$\begin{aligned}
\text{SS}_{\text{Regression}} &= b_1^2 \text{SS}_X \\
\text{SS}_{\text{Total}} &= \text{SS}_Y.
\end{aligned}$$

Then

$$\boxed{\text{SS Total} = \text{SS Reg} + \text{SS Res}}$$

2.3 Partitioning the sum of squares

1. **Notes**

$$\begin{aligned}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 &= \sum_{i=1}^{n}(b_0 + b_1 x_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}(\bar{y} - b_1\bar{x} + b_1 x_i - \bar{y})^2 \\ &= b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \mathrm{SS}_{\mathrm{Reg}}.\end{aligned}$$

2. The proportion of variation explained by the regression is

$$R^2 = \frac{\mathrm{SS}_{\mathrm{Reg}}}{\mathrm{SS}_{\mathrm{Total}}}.$$

Note $0 \leq R^2 \leq 1$.

2.4 Hypothesis Test for β_1

To determine if there is a significant linear relationship between x and y, we test the hypotheses:

$$\begin{array}{rll} H_0 : \beta_1 = 0 & & \text{No linear relationship} \\ H_1 : \beta_1 \neq 0 & & \text{Significant linear relationship} \\ \beta_1 > 0 & & \text{Significant positive linear relationship} \\ \beta_1 < 0 & & \text{Significant negative linear relationship} \end{array}$$

If H_0 is true then $\beta_1 = 0$, so

$$\text{SS}_{\text{Reg}} = b_1^2 \text{SS}_X \approx 0 \quad \text{and} \quad \text{SS}_{\text{Tot}} \approx \text{SS}_{\text{Res}}.$$

Thus we can use the ratio $\text{SS}_{\text{Reg}}/\text{SS}_{\text{Res}}$ as a test statistic. However, each of these sums of squares need to be adjusted by their **degrees of freedom**.

2.4 Hypothesis Test for β_1

$$\begin{aligned}\text{Regression df} &= \text{Number of parameters -1}\\ &= k-1=1 \text{ for simple linear regression.}\end{aligned}$$

$$\text{Total df} = n-1$$

$$\begin{aligned}\text{Residual df} &= n-k\\ &= n-2 \text{ for simple linear regression.}\end{aligned}$$

Note that two parameters are estimated from the data in simple linear regression, the intercept and the slope, so $k = 2$. These calculations can be set up as a table.

ANOVA Table

2.4 Hypothesis Test for β_1

The test then proceeds as usual.

If F_{obs} is in the critical region OR if p-value $< \alpha$, then reject H_0, and conclude there is a significant linear relationship between Y and X.

Example 2.3: Pharmex Data

Example 2.3: Pharmex Data (ctd)

2.5 Estimate of σ^2

template=wideslide,tocsection=false

2.6 Distribution of B_1

$$B_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\mathrm{SS}_{XY}}{\mathrm{SS}_X}$$

$$\mathbb{E}(B_1) =$$

so B_1 is an unbiased estimator of β_1.

2.6 Distribution of B_1

$$\mathrm{Var}(B_1) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\mathrm{SS}_X}\right)$$

Put $C_i = \dfrac{X_i - \bar{X}}{\mathrm{SS}_X}$, so $B_1 = \sum_{i=1}^{n} C_i Y_i$. Then

$$\mathrm{Var}(B_1) = \sum_{i=1}^{n} C_i^2 \,\mathrm{Var}(Y_i)$$

so $\mathrm{Var}(B_1) = \dfrac{\sigma^2}{\mathrm{SS}_X}$.

2.6 Distribution of B_1

Further, Y_i are normally distributed, so $B_1 = \sum_{i=1}^{n} C_i Y_i$ is the sum of normal random variables; thus B_1 is also normal.

$$\boxed{B_1 \sim \mathrm{N}\left(\beta_1, \frac{\sigma^2}{\mathrm{SS}_X}\right)} \Rightarrow \boxed{Z = \frac{B_1 - \beta_1}{\sigma/\sqrt{\mathrm{SS}_X}} \sim \mathrm{N}(0,1)}$$

If σ is unknown (usually the case), we replace it by S, and then

$$\boxed{T = \frac{B_1 - \beta_1}{S/\sqrt{\mathrm{SS}_X}} \sim t_{n-2}}$$

We can thus use the t-distribution for hypothesis tests and confidence intervals for β_1.

template=wideslide,tocsection=false

Example 2.4: Pharmex Data

$b_1 = 0.7623, \quad \mathrm{SS}_X = 3739.28, \quad s = \sqrt{54.68}.$

(i) Is there a significant

1. linear relationship
2. positive linear relationship

between Sales and Promotions?

Example 2.4: Solution

Example 2.4: Solution (ctd)

template=wideslide,tocsection=false

Example 2.4:(ctd)

(ii) Find a 95% confidence interval for β_1.