Lecture 38 Simple Linear Regression - The Model

BIO210 Biostatistics

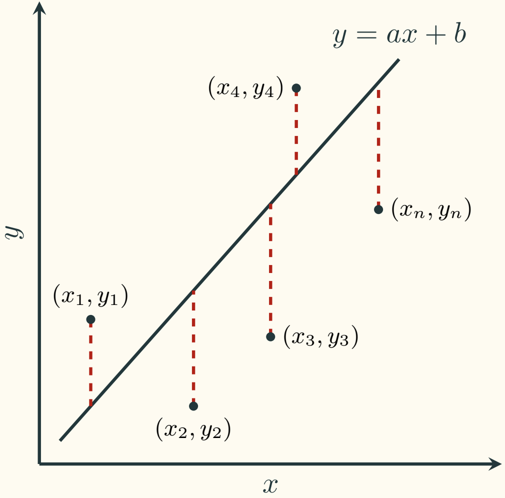
Xi Chen

Spring, 2025

School of Life Sciences
Southern University of Science and Technology

南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

Simple Linear Regression



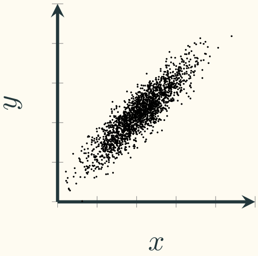
Using OLS regression:

$$SE_{line} = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$

$$\Downarrow \text{ minimise}$$

$$\begin{cases} a = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \\ b = \bar{y} - a \cdot \bar{x} \end{cases}$$

Simple Linear Regression - the model



Simple Linear Regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables.

$\boldsymbol{X}$:	independent variable	explanatory variable	predictor variable
$\boldsymbol{Y}$:	dependent variable	outcome variable	response variable

The Simple Linear Regression Model using OLS:

For the entire population: $\overset{\text{population regression line}}{\boxed{\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X}}} + \boldsymbol{\epsilon}$

For each observation: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

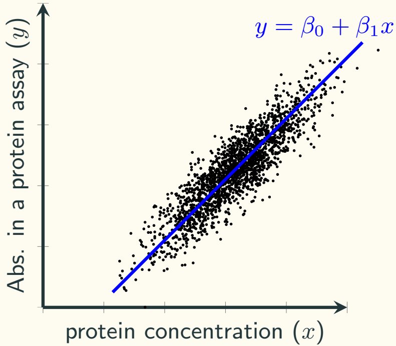
where:

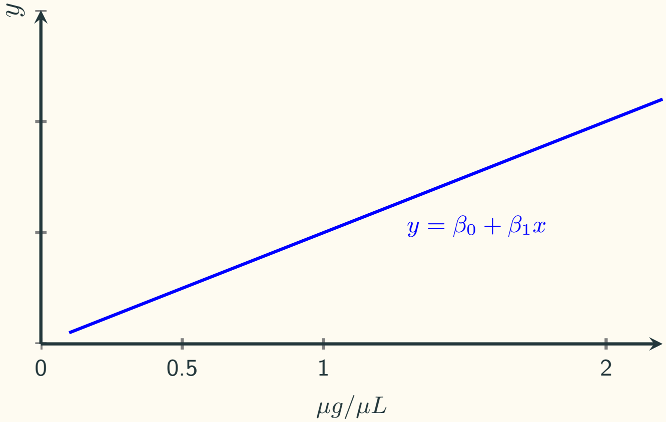
β_0 is the population intercept

β_1 is the population slope

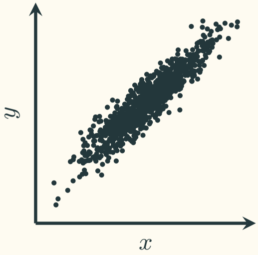
ϵ_i is the error from y_i to the line $\beta_0 + \beta_1 x_i$

Simple Linear Regression - the model



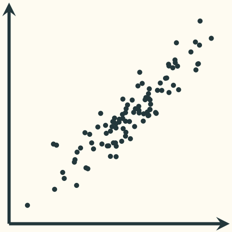


Best Fit Line

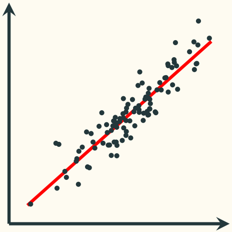


To estimate the population parameters β_0, β_1

Take a sample (size n)



OLS



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\epsilon_i} = y_i - \hat{y_i}$$

$\hat{\beta}_0$: sample intercept

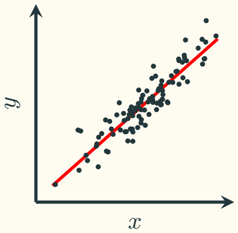
$\hat{\beta}_1$: sample slope

$\hat{\epsilon_i}$: residual

In OLS, $\sum_{i=1}^{n} \hat{\epsilon_i}^2$ is minimised.

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} \end{cases}$$

Evaluation of the model: Coefficient of Determination r^2



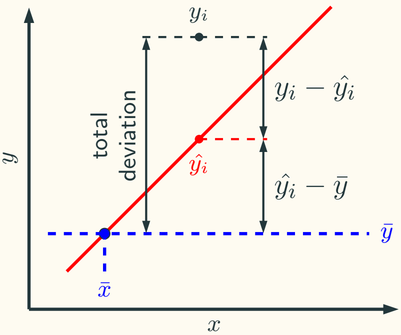
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

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

minimise $\sum_{i=1}^{n} \hat{\epsilon}_i^{\,2}$

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} \end{cases}$$

How useful is the model?



total deviation

deviation explained by the regression line

$$\boxed{y_i - \bar{y}} = \boxed{(y_i - \hat{y}_i)} + \boxed{(\hat{y}_i - \bar{y})}$$

unexplained deviation due to error

Sum of squares total:

$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$

Sum of squares regression:

$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

Sum of squares error/residual:

$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

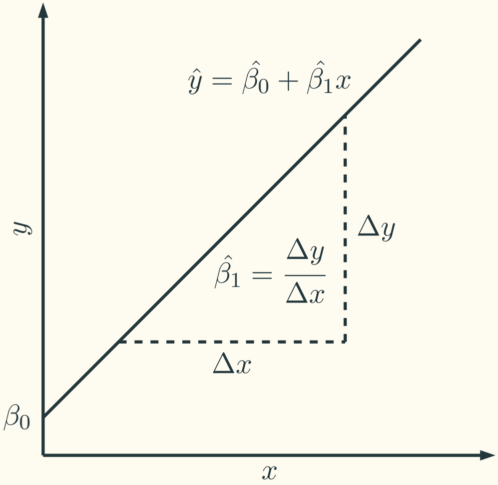
$$SST = SSR + SSE$$

$$r^2 = \frac{\text{explained}}{\text{total}} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

The ANOVA Table For OLS

Source of Variation	SS	d.f.	MS
Regression	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	1	$MSR = \frac{SSR}{1} = SS_R$
Error/Residual	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Total	$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SSR + SSE$	$n-1$	

Interpretation of The Regression Parameters

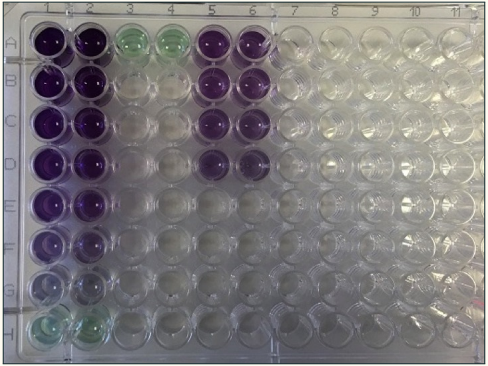


$\hat{\beta}_1$: the predicted change of the dependent variable y when the independent variable x changes one unit

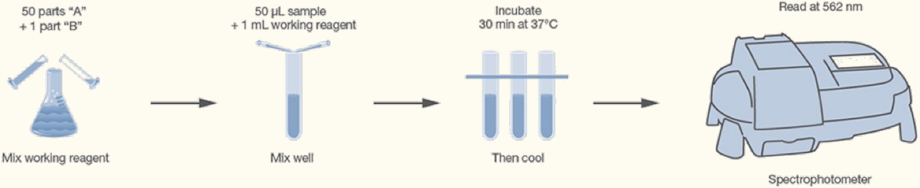
$\hat{\beta}_0$: the predicted value of the dependent variable y when the independent variable x takes the value of 0. It may not have actual meaning.

BCA To Measure Protein Concentration

The BCA Protein Assay combines the well-known reduction of Cu^{2+} to Cu^{1+} by protein in an alkaline medium with the highly sensitive and selective colorimetric detection of the cuprous cation (Cu^{1+}) by bicinchoninic acid (BCA).



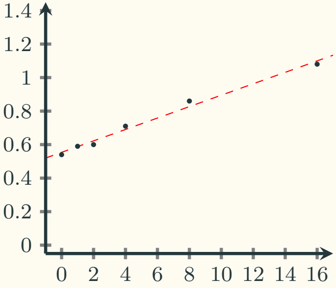
BCA To Measure Protein Concentration



BSA (mg/mL)	Absorb.
0	0.54
1	0.59
2	0.60
4	0.71
8	0.86
16	1.08
$\bar{x} = 5.17$	$\bar{y} = 0.73$

$x_i - \bar{x}$	$y_i - \bar{y}$
-5.17	-0.19
-4.17	-0.14
-3.17	-0.13
-1.17	-0.02
2.83	0.131
10.83	0.351

$(x_i - \bar{x})^2$	prod.
26.73	0.98
17.39	0.57
10.049	0.42
1.37	0.03
8.00	0.37
117.29	3.80



Assumptions For Simple Linear Regression

The "LINE" assumptions must be met when performing a simple linear regression:

- The mean of the dependent variable ($\mathbb{E}\left[\boldsymbol{Y}|\boldsymbol{X}\right]$, $\mu_{y|x}$) is a **L**inear function of $\boldsymbol{X}$
- The errors/residuals $\boldsymbol{\epsilon_i}|\boldsymbol{X} = x_i$ are **I**ndependent
- The errors/residuals $\boldsymbol{\epsilon_i}|\boldsymbol{X} = x_i$ are **N**ormally distributed
- The errors/residuals $\boldsymbol{\epsilon_i}|\boldsymbol{X} = x_i$ have **E**qual variance for all x_i values (homoscedasticity)

$$\boldsymbol{\epsilon} \sim \boldsymbol{\mathcal{N}}(0,\, \sigma^2_{\boldsymbol{\epsilon}})$$

Seaborn Tips Datasets

Food servers' tips in restaurants may be influenced by many factors, including the nature of the restaurant, size of the party, and table locations in the restaurant. Restaurant managers need to know which factors matter when they assign tables to food servers. For the sake of staff morale, they usually want to avoid either the substance or the appearance of unfair treatment of the servers, for whom tips (at least in restaurants in the United States) are a major component of pay. In one restaurant, a food server recorded the following data on all customers they served during an interval of two and a half months in early 1990. The restaurant, located in a suburban shopping mall, was part of a national chain and served a varied menu. In observance of local law, the restaurant offered to seat in a non-smoking section to patrons who requested it. Each record includes a day and time, and taken together, they show the server's work schedule.

https://www.kaggle.com/ranjeetjain3/seaborn-tips-dataset

Tips

Restaurant
Address

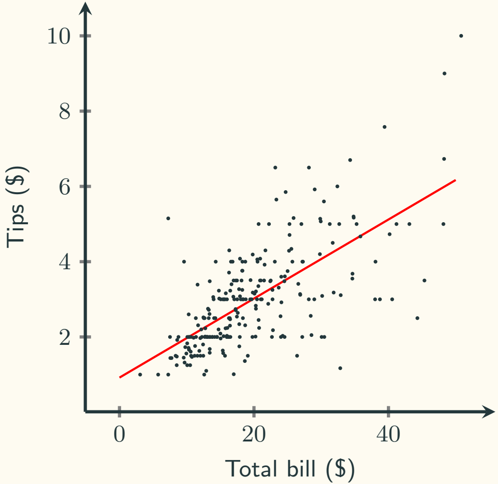
1 Burger	£13.99
1 French fries	£5.99
2 Fish & chips	£11.99
1 Lamb kebab	£10.99
5 Coke	£3.99

AMOUNT: £74.90

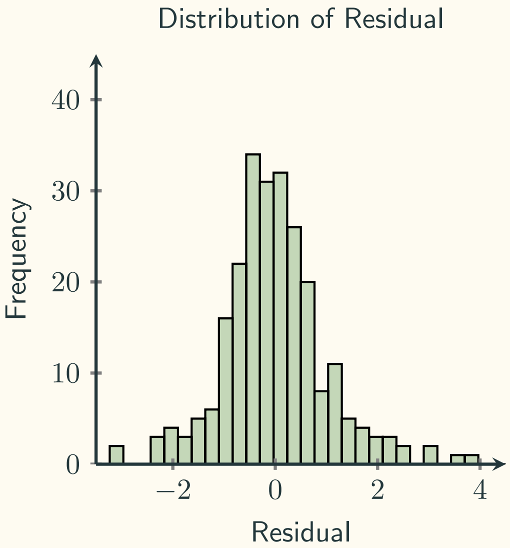
TIP: ______

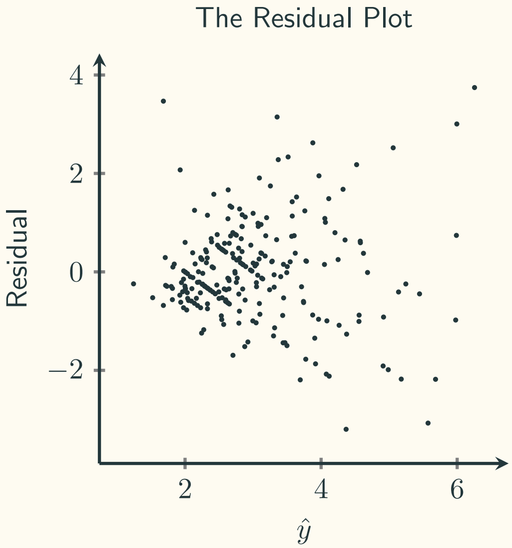
TOTAL: ______

Total bill	Tips
16.99	1.01
10.34	1.66
21.01	3.5
23.68	3.31
24.59	3.61
25.29	4.71
8.77	2
26.88	3.12
15.04	1.96
14.78	3.23
10.27	1.71
⋮	⋮

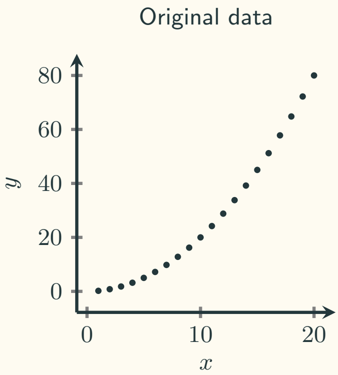


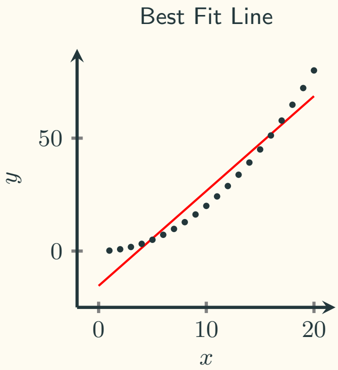
The Residual Plot

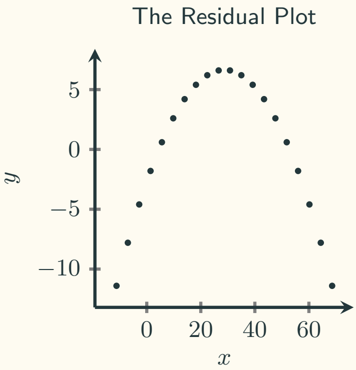




The Residual Plot







Linear Regression

- The Simple Linear Regression Model

- $\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X} + \boldsymbol{\epsilon}$

- The Multiple Linear Regression Model

- $\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X_1} + \beta_2 \boldsymbol{X_2} + \beta_3 \boldsymbol{X_3} + \cdots + \beta_q \boldsymbol{X_q} + \boldsymbol{\epsilon}$

- The Logistic Regression Model ($\boldsymbol{Y}$ is categorical)

- $\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X_1} + \beta_2 \boldsymbol{X_2} + \beta_3 \boldsymbol{X_3} + \cdots + \beta_q \boldsymbol{X_q} + \boldsymbol{\epsilon}$