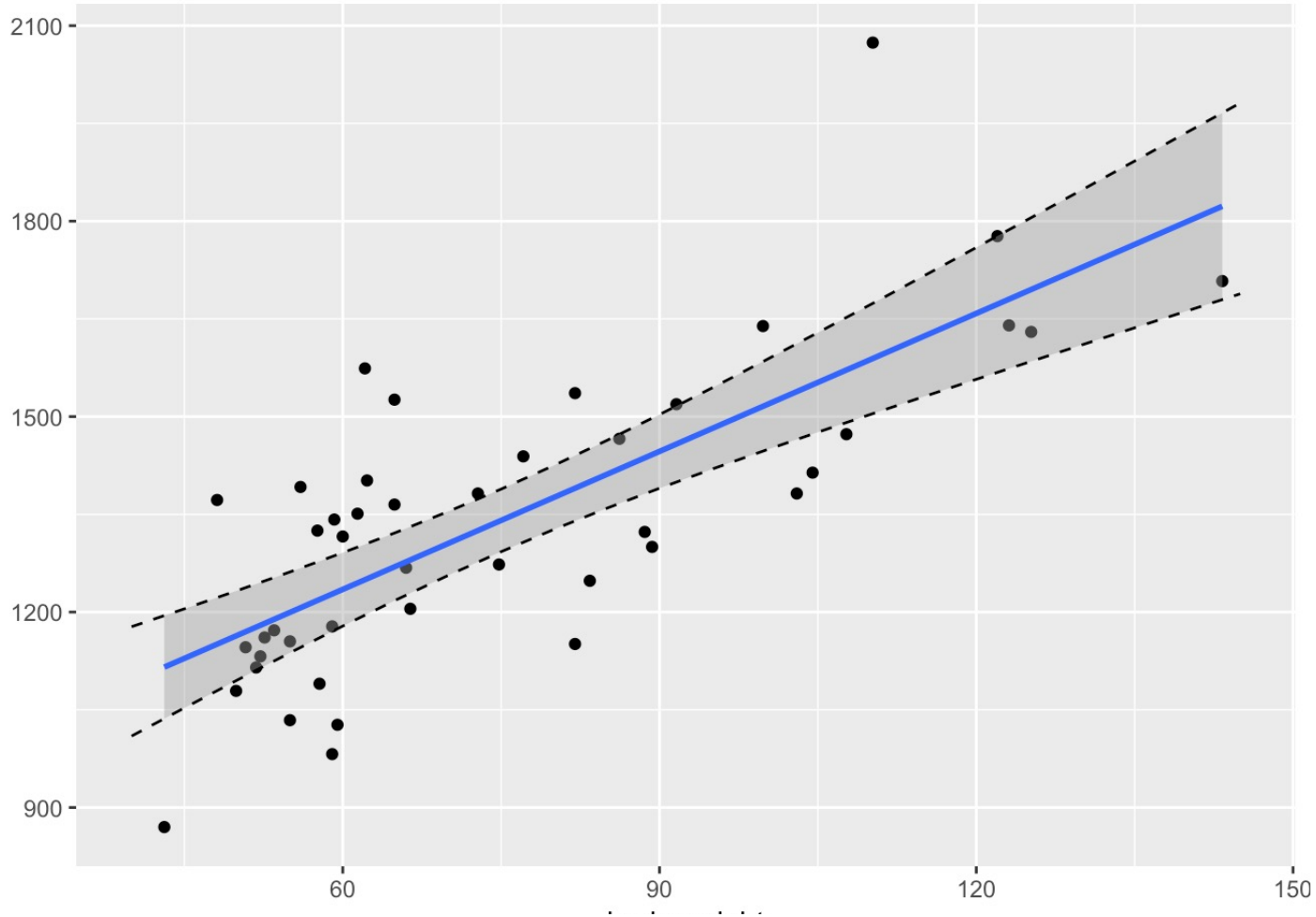
L09: Linear Regression

- Simple linear regression analysis
- Multiple regression analysis



Quiz 5 (next week): simple linear regression analysis

The goal of studying linear regression

The goal of studying linear regression is to understand and model the relationship between two or more variables. Specifically, linear regression aims to:

1. **Identify the Relationship:** Determine whether a linear relationship exists between the independent variable(s) and the dependent variable, and quantify the strength and direction of this relationship.

2. **Predict Outcomes:** Use the linear equation to predict the value of the dependent variable based on the value(s) of the independent variable(s).

3. **Estimate Effects:** Assess the impact of changes in the independent variable(s) on the dependent variable. This is particularly useful in understanding how different factors influence the outcome variable.

4. **Model Evaluation:** Evaluate the goodness of fit of the linear model to the data

5. **Data Analysis and Interpretation:** Analyze and interpret data to make informed decisions and policy planning.

Simple linear regression

WHAT IS A SIMPLE LINEAR REGRESSION?

•Regression is a very useful statistical model used to capture a relationship between related variables of our interest. If the relationship is shown to be "LINEAR", then the regression is said to be a linear regression.

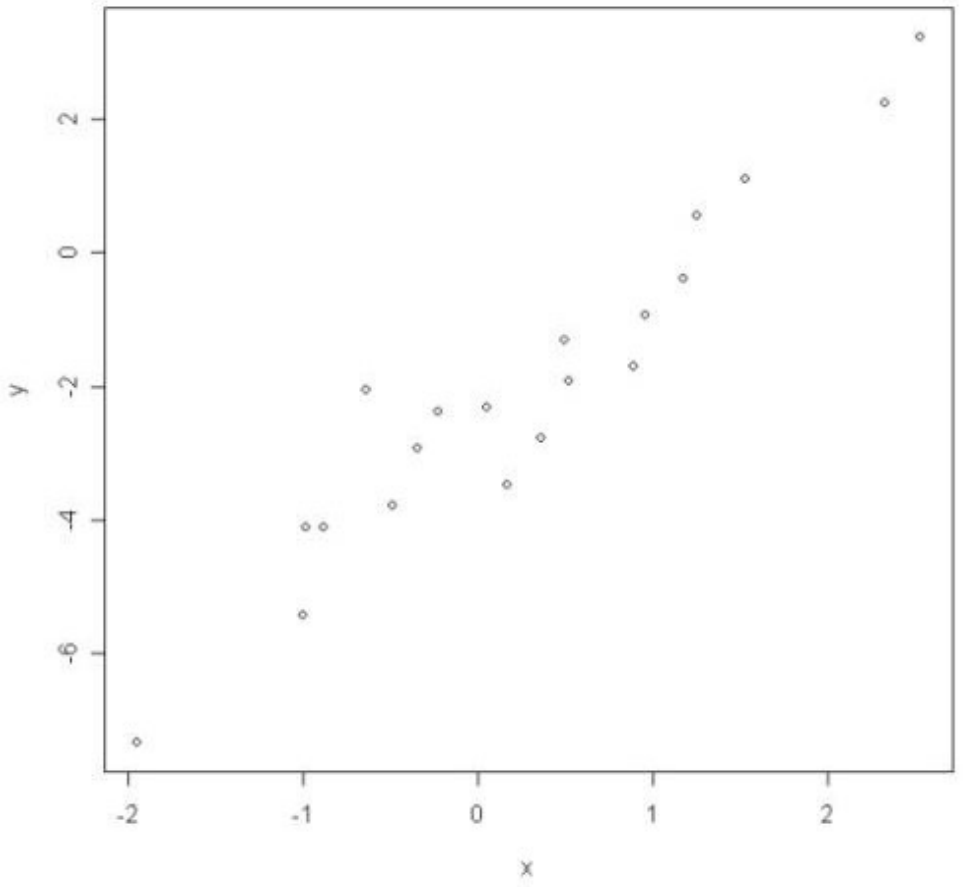
•"SIMPLE" means that there is only ONE variable (called **explanatory variable** labeled by x) used to explain our target variable (called **response variable** labeled by y) --- the variable we want to explain or predict.

A SIMPLE LINEAR REGRESSION

Is a statistical model used to study the relationship between *y* and *x* if they are related LINEARLY

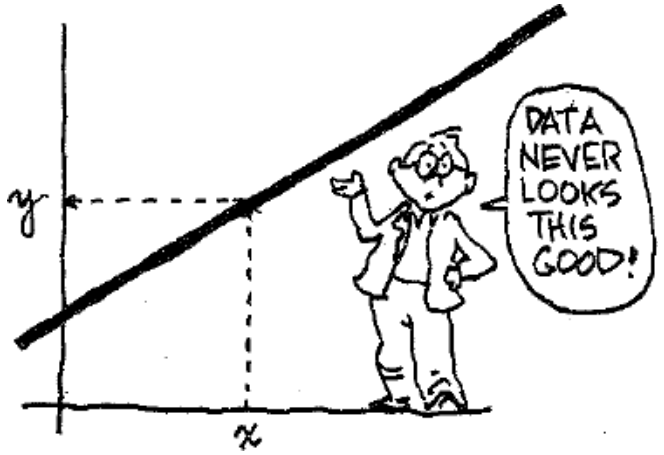
Introduction

- A **scatter plot** is a power graphical method used to visualize the relationship between *y* and *x*.
- To be more precise, we would have a collection of a **PAIRED data of *x* and *y***, denoted by $\{(x_i, y_i), i = 1,\ldots,n\}$, then plot a graph of *y* against *x*, like the picture on the right.

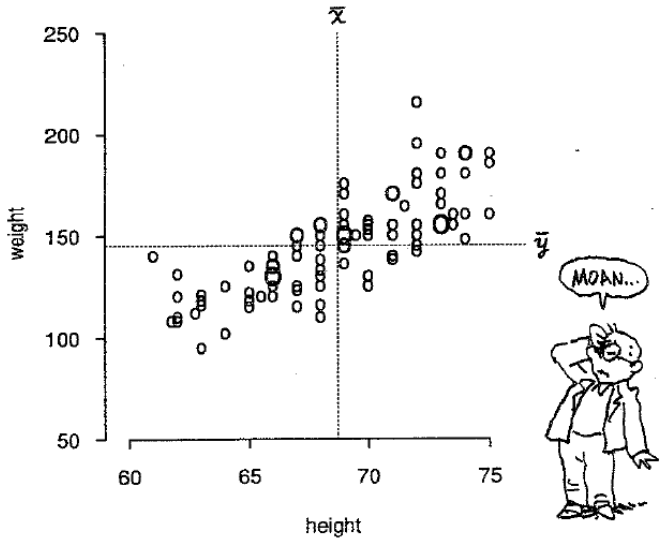


Introduction

- Note that in Math class, we probably learned to see relationships displayed as GRAPHS. Given *x*, we can predict *y*.



- However, in statistics, things are never so clean. Data do <u>not perfectly lie</u> on a line or curve!

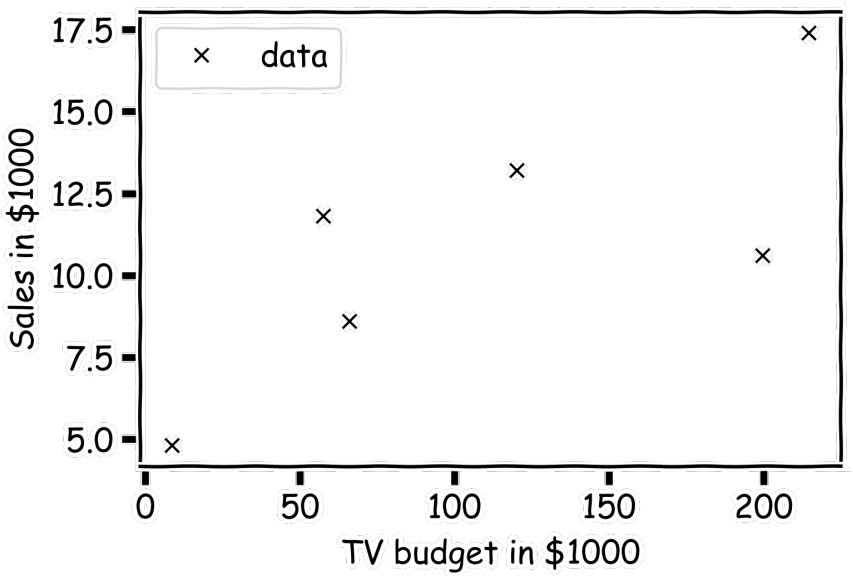


Introduction

- If the scatter plot shows a “linear” pattern, then **we can FIND a straight line to fit the messy data statistically.**
- In the case of two variables (a response variable *y* and explanatory variable *x*) which are related linearly:

$$Y = \beta_0 + \beta_1 x + \mathrm{e}$$

- Note that e is random and is used to measure all uncertainty of the model, like a measurement error, and that the regression coefficients β_0 and β_1 are unknown.



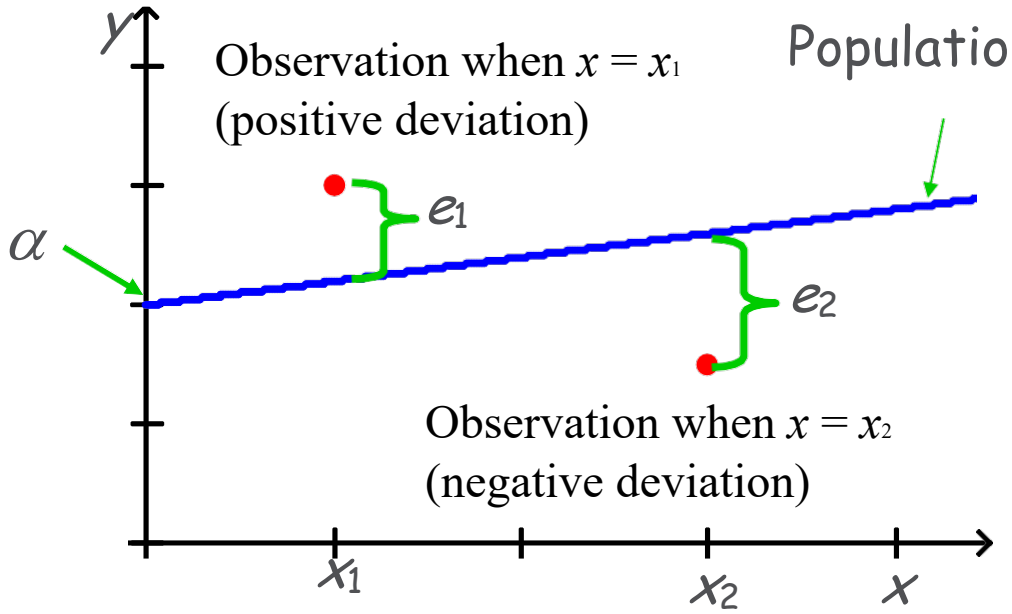
- How do we fit the data appropriately by a straight line? Which straight line should we use?
- Or equivalently, what are the "good" estimates of β_0 and β_1 used to fit the data.

- Intuitively, we would like to find a straight line so that it is "close" to the collected data. In other words, we need a measure of the closeness of the straight line to the data, or equivalently, we need an estimation criterion of finding the "good" estimates of β_0 and β_1.

The simple linear regression model assumes that there is a line with *y*-intercept α and slope β, called the population regression line.

When a value of the independent variable *x* is fixed and an observation on the dependent variable *y* is made,

$$y = \alpha + \beta x + e \quad \text{OR} \quad y = \beta_0 + \beta_1 x + e$$



Without the random deviation *e*, all observed (*x*, *y*) points would fall exactly on the population regression line.

The inclusion of *e* in the model equation recognizes that points will deviate from the line by a random amount.

But we are uncertain about the value of e.

It could be negative, positive, or even 0.

It could be large in magnitude (a point far from the population regression line) or quite small (a point very close to the line).

Basic Assumptions of the Simple Linear Regression Model

The distribution of e at any particular x value has mean value 0. that is, $\mu_e = 0$.

The standard deviation of e is the same for any particular value of x. This standard deviation is denoted by σ .

The distribution of e at any particular value of x is normal.

The random deviations $e_1, e_2, \ldots, e_n$ associated with different observations are independent of one another.

Basic Assumptions of the Simple Linear Regression Model

There is variability in the *y* values observed at any particular value of *x*. Consider *y* when *x* has some fixed value *x**, so that

$$y = \alpha + \beta x^* + e$$

fixed number

normally distributed variable (*e*)

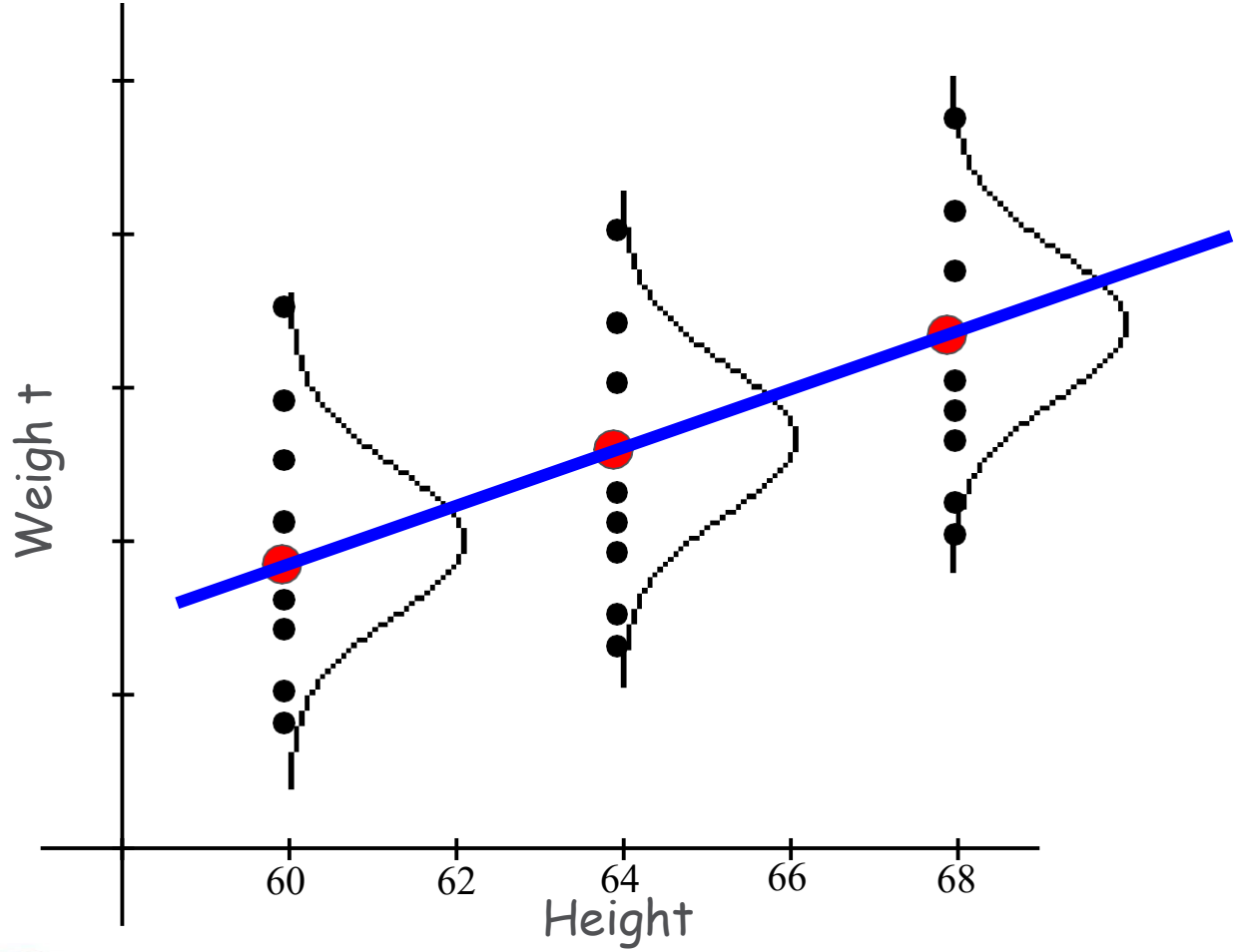
1. y also has a normal distribution

2. Since mean value of e = 0. mean value of *y*= $\alpha + \beta x^*$

3. No variability in the fixed number $\alpha + \beta x^*$, standard deviation of *y* is the same as the standard deviation of *e*.

Example

Let's look at the heights and weights of a population of adult women.



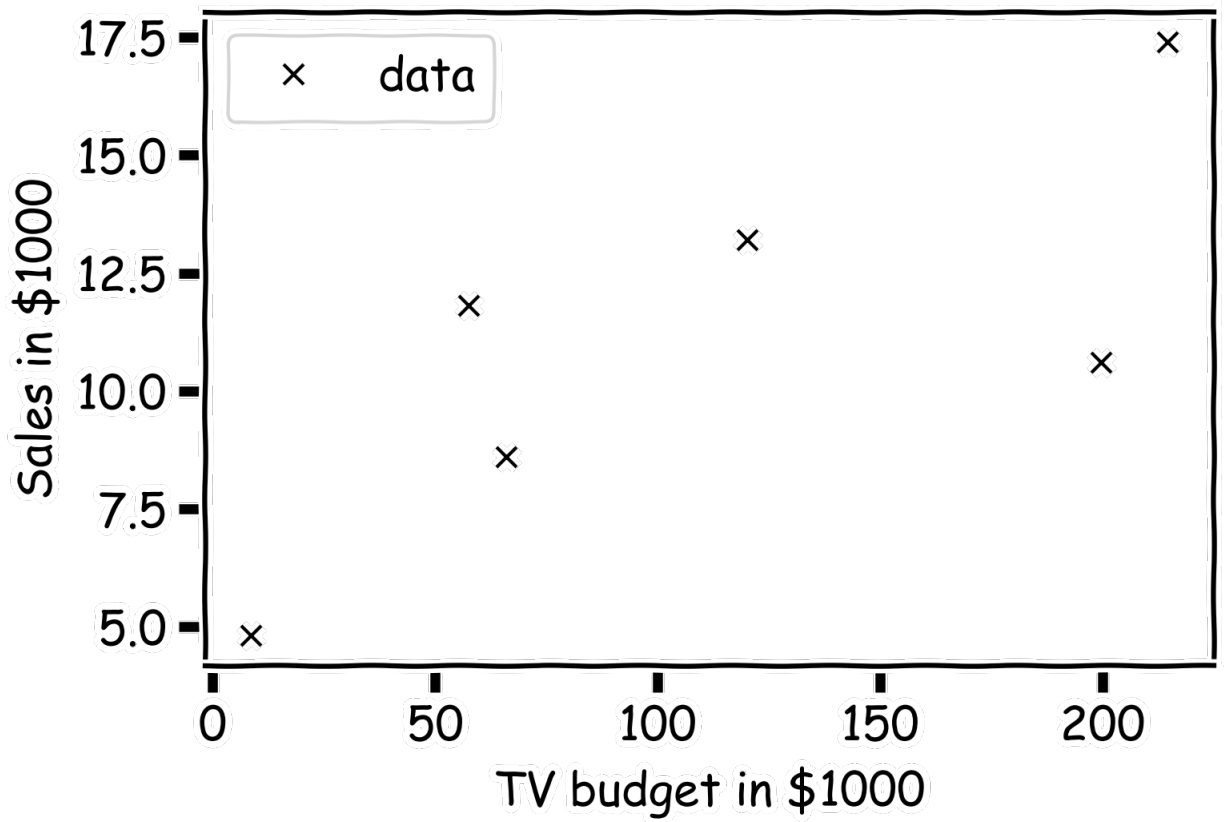
Notice that the three normal curves in Figure have identical spreads.

The standard deviation of e is the same for any particular value of x. This standard deviation is denoted by σ .

Estimating the Regression Line

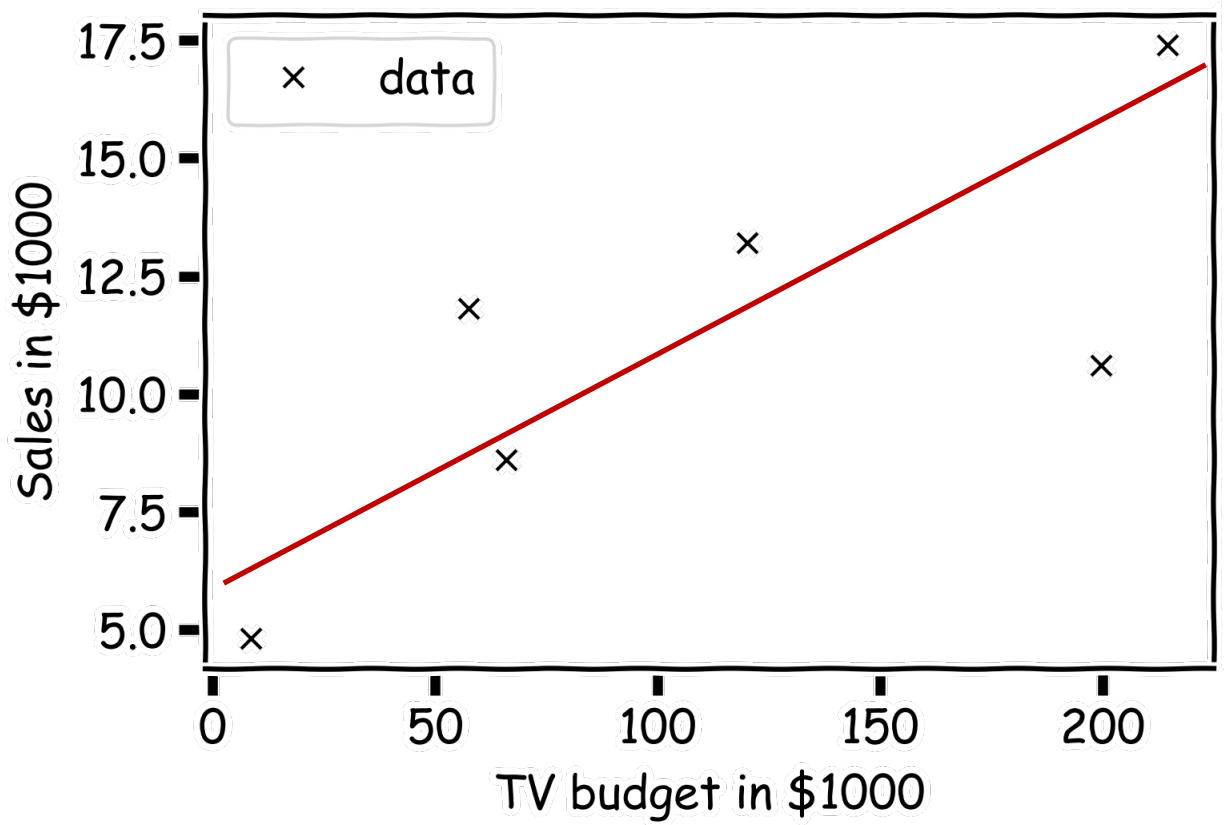
Linear Regression

For a given data set



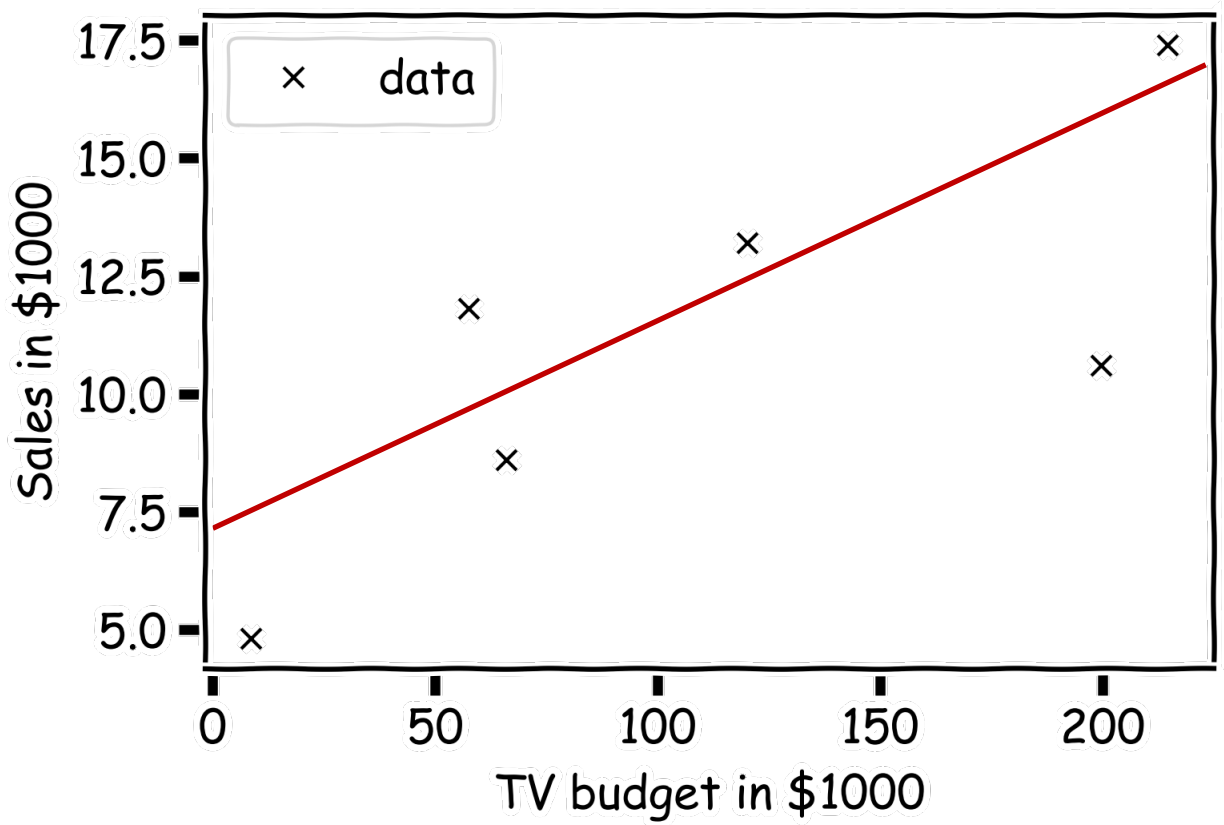
Linear Regression

Is this line good?



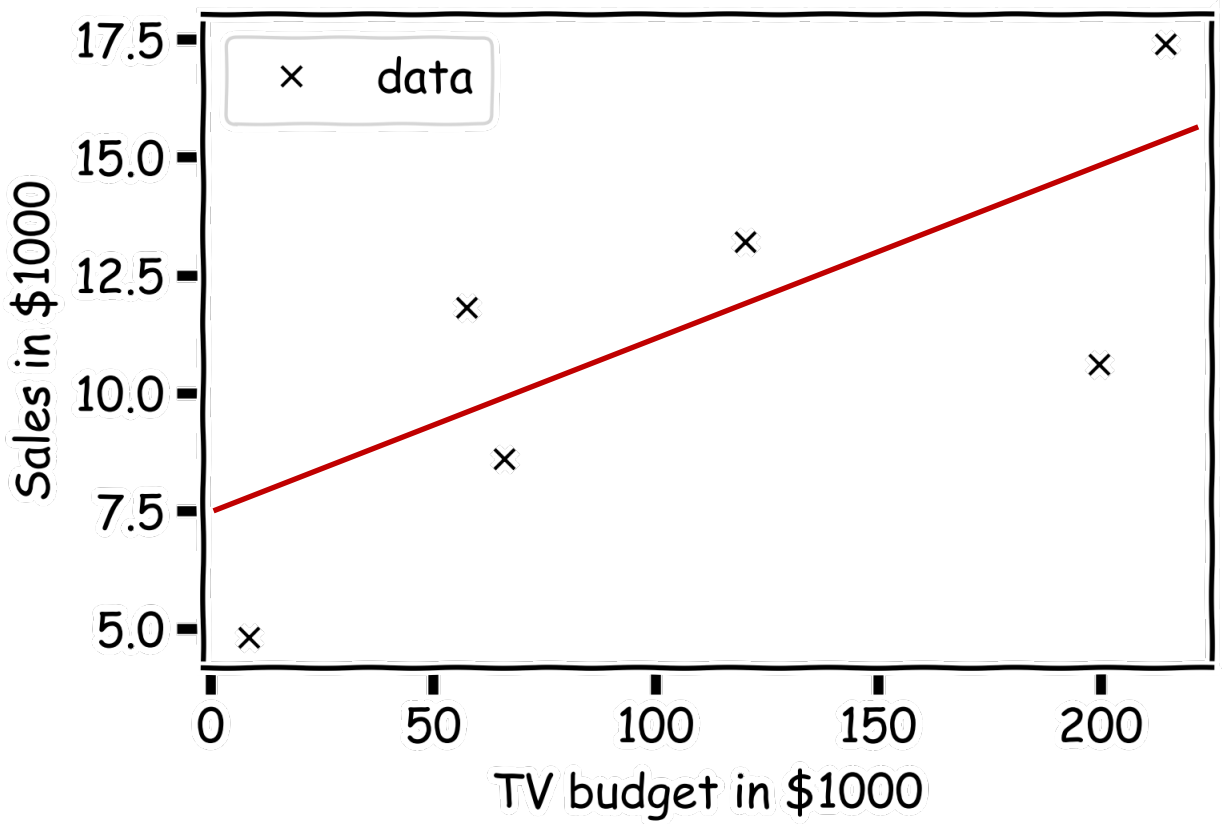
Linear Regression

Maybe this one?



Linear Regression

Or this?



Least Squares Approach

- "LEAST SQUARES approach" is the most commonly used method to find the straight line which is close to the data in statistics.

We estimate the true population regression line.

$$\hat{y} = a + bx$$

b = point estimate of β = $\dfrac{S_{xy}}{S_{xx}}$

a = point estimate of α = $\bar{y} - b\bar{x}$

$$b = \frac{\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n}{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}}$$

$$a = \bar{y} - b\bar{x}$$

$$SXX = \Sigma\,(x - \bar{x})^2 = \Sigma\,x^2 - \frac{(\Sigma\,x)^2}{n} \qquad \text{(Sum Squares X)}$$

$$SYY = \Sigma(y - \bar{y})^2 = \Sigma\,y^2 - \frac{(\Sigma\,y)^2}{n} \qquad \text{(Sum Squares Y)}$$

$$SXY = \Sigma(x - \bar{x})(y - \bar{y}) = \Sigma\,xy - \frac{(\Sigma x)(\Sigma y)}{n} \qquad \text{(Sum Products X,Y)}$$

Fitted Regression Line

- Once we have the least-squares estimates, we can write down the so-called *fitted regression line* (or sometimes called *estimated regression line*):

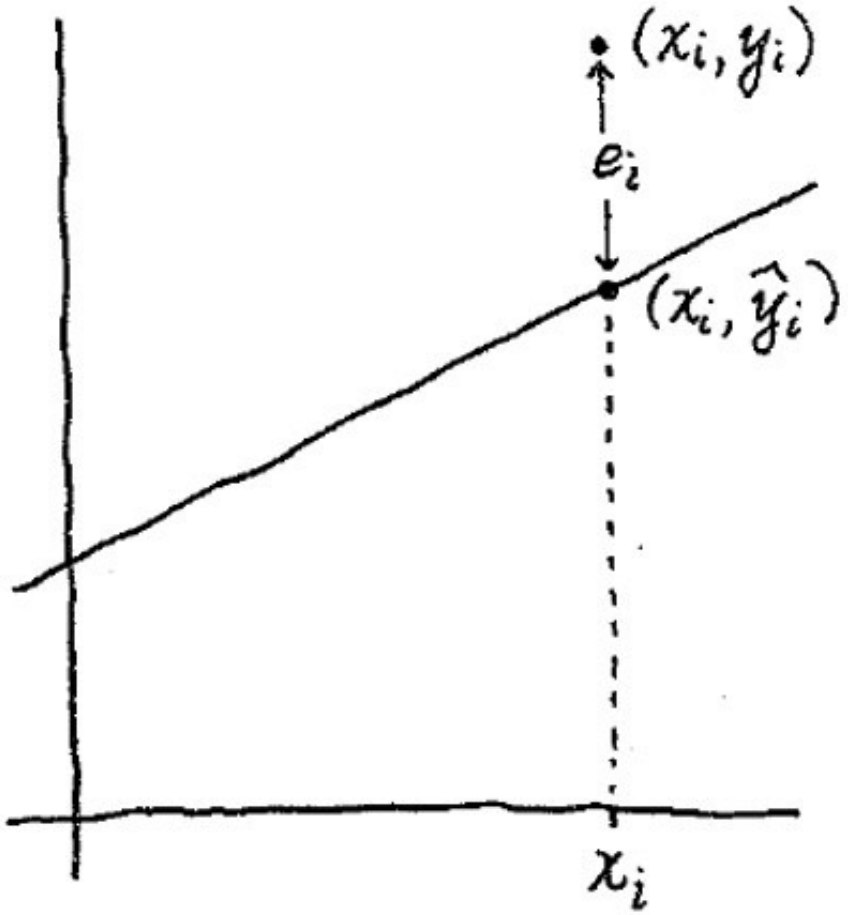
$$\hat{y}=a+bx$$

- If we substitute x_i (for i=1,…,n) to the fitted regression line, then we can get a fitted value, labelled by $\hat{y}_i$ of the ith observation y_i of the response variable, and then have the so- called the residual e_i of y_i, where e_i is defined as

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

Fitted Regression Line

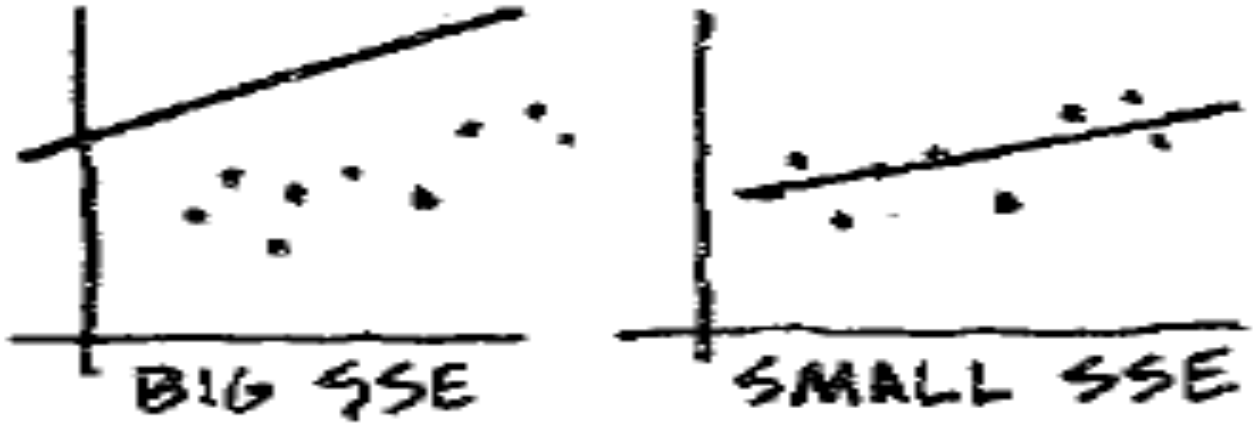
- Residual e_i in practice is regarded as an "actual" value of the unobservable random term ε_i in the simple linear regression, and it plays a very important role in regression analysis because we can use it to do a model diagnostic and quantify the goodness of the regression model.



Residual Sum of Squares (or Sum of Squared Errors)

- Residual sum of squares or called **S**um of **S**quared **E**rrors (hereafter, **SSE**) is one commonly used measure of evaluating the goodness of a simple linear regression model. As its name suggests, this measure is defined as

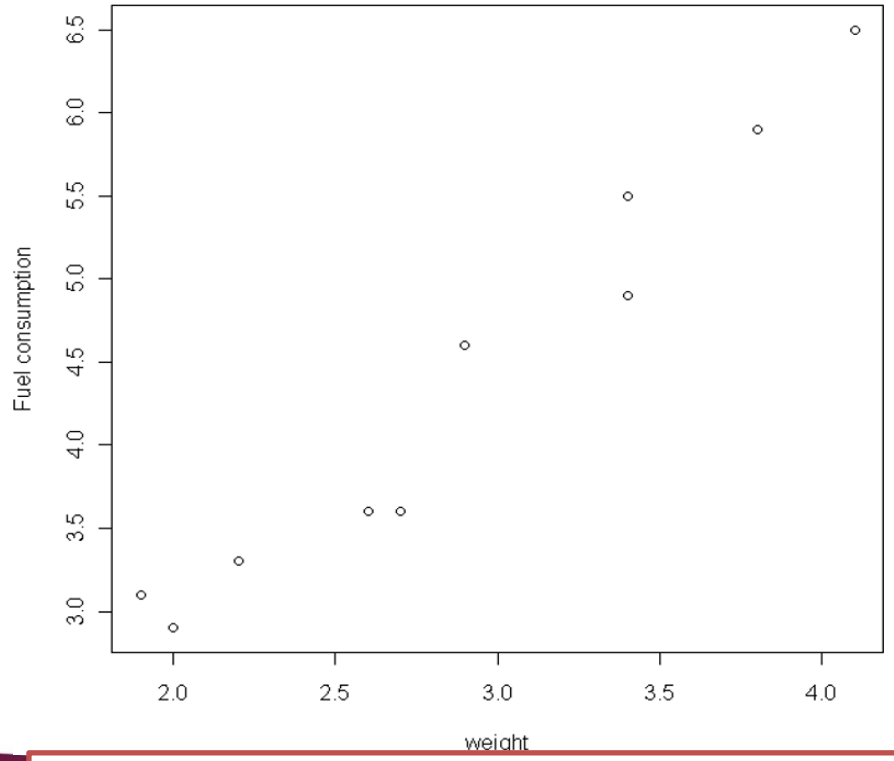
$$SSE = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2$$



Example

Consider the relationship between the **weight (*x*) of an automobile** and **fuel consumption (*y*),** where the latter is measured by gpm --- the amount of fuel (in gallons) that is need to drive 100 miles.

Suppose **we collect paired data of the weight and fuel consumption gpm of 10 cars, and have the following picture to show that there exists a linear relationship between them**.



Here are the raw paired data of *x* and *y*.

Car	Weight (1000 pounds)	Fuel Consumption, gpm (gallons/100 miles)
AMC Concord	3.4	5.5
Chevy Caprice	3.8	5.9
Ford Country Squire Qagon	4.1	6.5
Chevete	2.2	3.3
Toyota Corona	2.6	3.6
Ford Mustang Ghia	2.9	4.6
Mazda GLC	2.0	2.9
AMC Sprint	2.7	3.6
VW Rabbit	1.9	3.1
Buick Century	3.4	4.9

According to the scatter plot, it is reasonable for us to use a simple linear regression to fit the paired data. Thus, let's Determine the slope and the intercept of the regression

$$b = \frac{\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n}{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}}$$

$$a = \bar{y} - b\bar{x}$$

$S_{XY} = \sum_{i=1}^{n}(x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n$

$S_{XX} = \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i^2)/n$

$$\sum_{i=1}^{10} x_i = 29.0, \sum_{i=1}^{10} x_i^2 = 89.28, \sum_{i=1}^{10} y_i = 43.9, \sum_{i=1}^{10} y_i^2 = 207.31, \sum_{i=1}^{10} x_i y_i = 135.80$$

Car	Weight (1000 pounds)	Fuel Consumption, gpm (gallons/100 miles)
AMC Concord	3.4	5.5
Chevy Caprice	3.8	5.9
Ford Country Squire Qagon	4.1	6.5
Chevete	2.2	3.3
Toyota Corona	2.6	3.6
Ford Mustang Ghia	2.9	4.6
Mazda GLC	2.0	2.9
AMC Sprint	2.7	3.6
VW Rabbit	1.9	3.1
Buick Century	3.4	4.9

And then by the least-squares approach we have

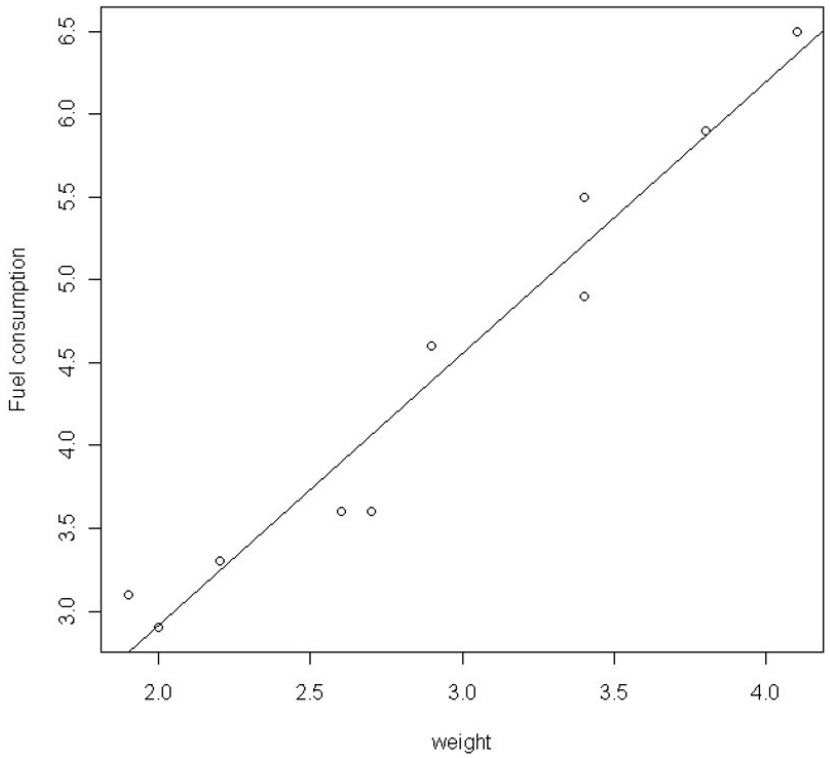
$$b = \frac{135.80 - (29.0)(43.9)/10}{89.28 - (29.0)^2/10} = \frac{8.49}{5.18} = 1.639$$

$a = 43.9/10 - (1.639) * (29.0)/10 = -0.363$

Finally, we can write down the fitted regression line

$\hat{y} = -0.363 + 1.639x$

If we draw this fitted regression line on the scatter plot, then we have



According to the least-squares estimates b = 1.639, we can say that on average each additional unit (1000 pounds) of weight requires an additional 1.639 gallons of fuel to drive 100 miles. **That is, increasing one unit in *x* will increase 1.639 units in *y* on average.**

And then by the least-squares approach we have

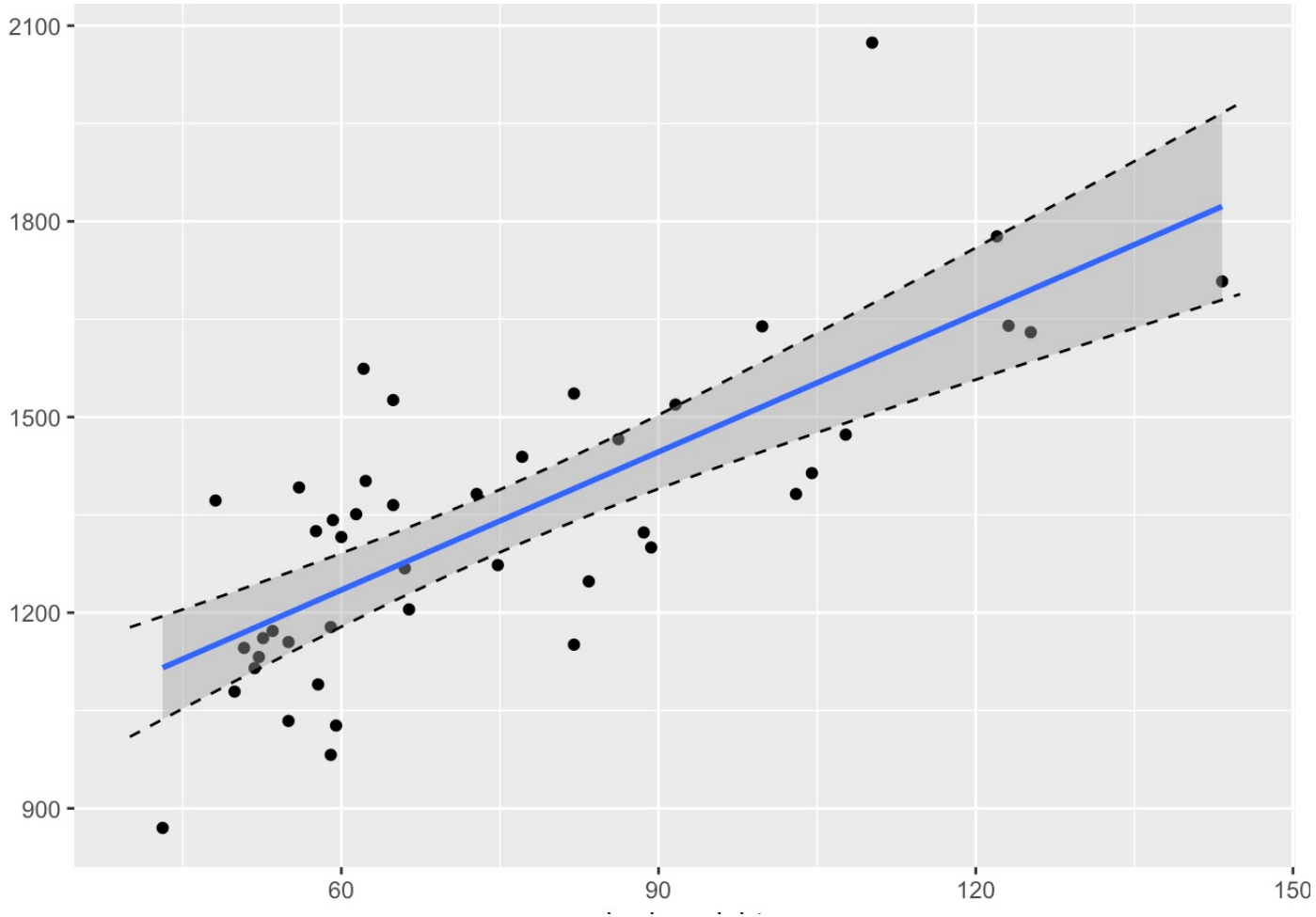
$$b = \frac{135.80 - (29.0)(43.9)/10}{89.28 - (29.0)^2/10} = \frac{8.49}{5.18} = 1.639$$

$$a = 4.39 - (1.639)(2.9) = -0.363$$

Finally, we can write down the fitted regression line

$$\hat{y} = -0.363 + 1.639x$$

Statistical Inference about β_0 and β_1



Statistical Inference about β_0 and β_1

Recall that, according to the least squares approach, we have following *least-squares* ESTIMATES of the unknown true values of β_0 and β_1, respectively.

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n}{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Statistical Inference about β_0 and β_1

So, if we want to study the estimation method for the statistical inference about the true values of β_0 and β_1, then we need their random counterparts. Thus, we have the following respective random variables called *least-squares* ESTIMATORS for the unknown true values of β_0 and β_1:

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

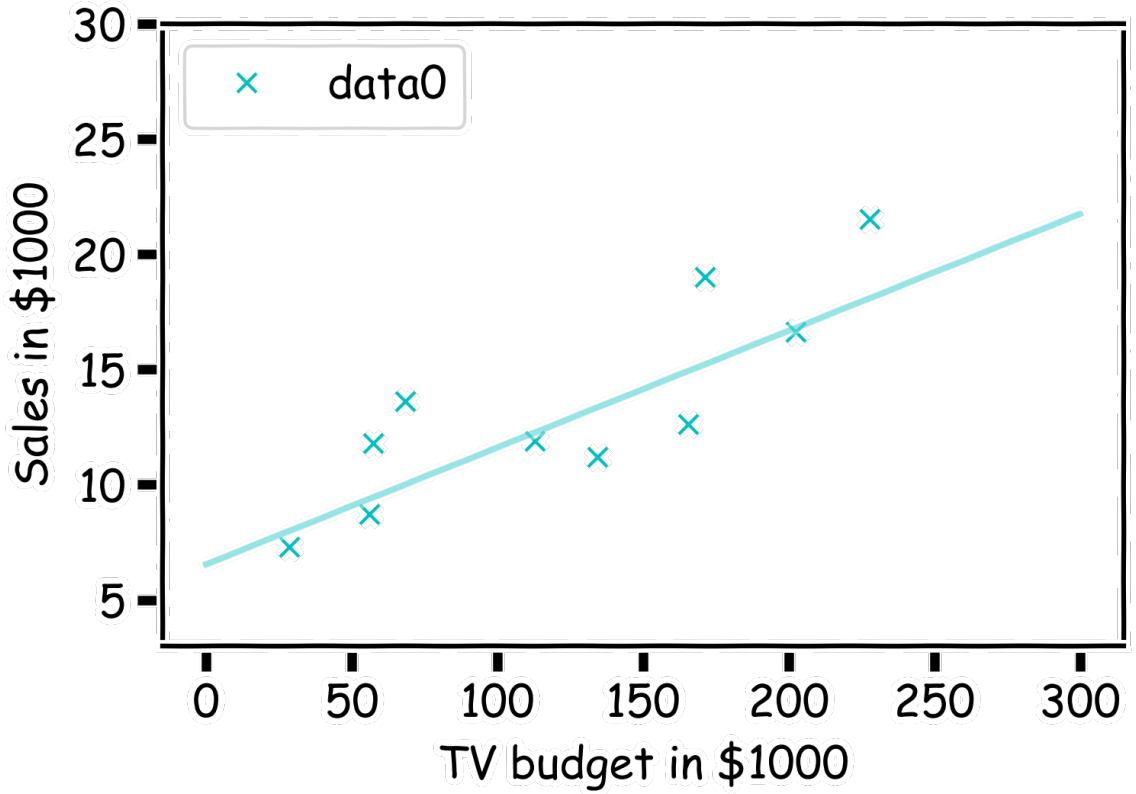
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Statistical Inference about β_0 and β_1

- We would then ask
  - How are $\hat{\beta}_0$ and $\hat{\beta}_1$ distributed around β_0 and β_1, respectively?
  - How do we construct CONFIDENCE INTERVALS and test HYPOTESIS?

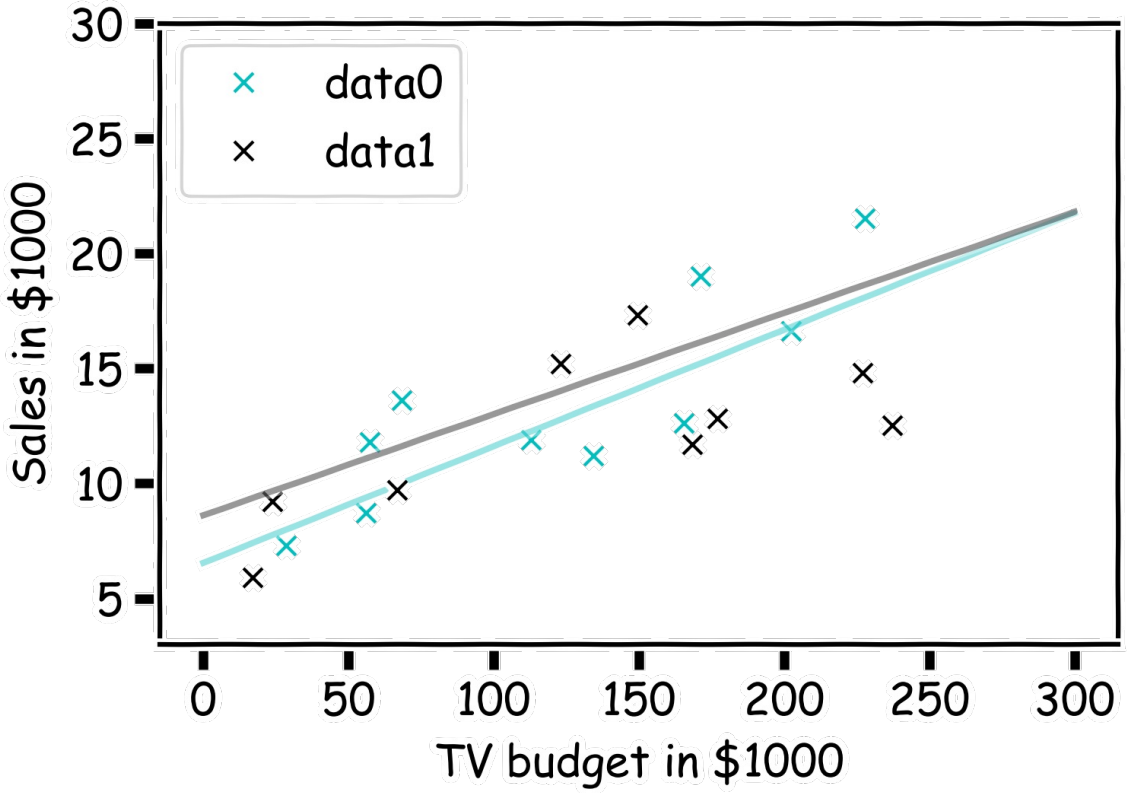
How well do we know $\hat{f}$?

Our confidence in f is directly connected with the confidence in βs. So for each β we can determine the model.



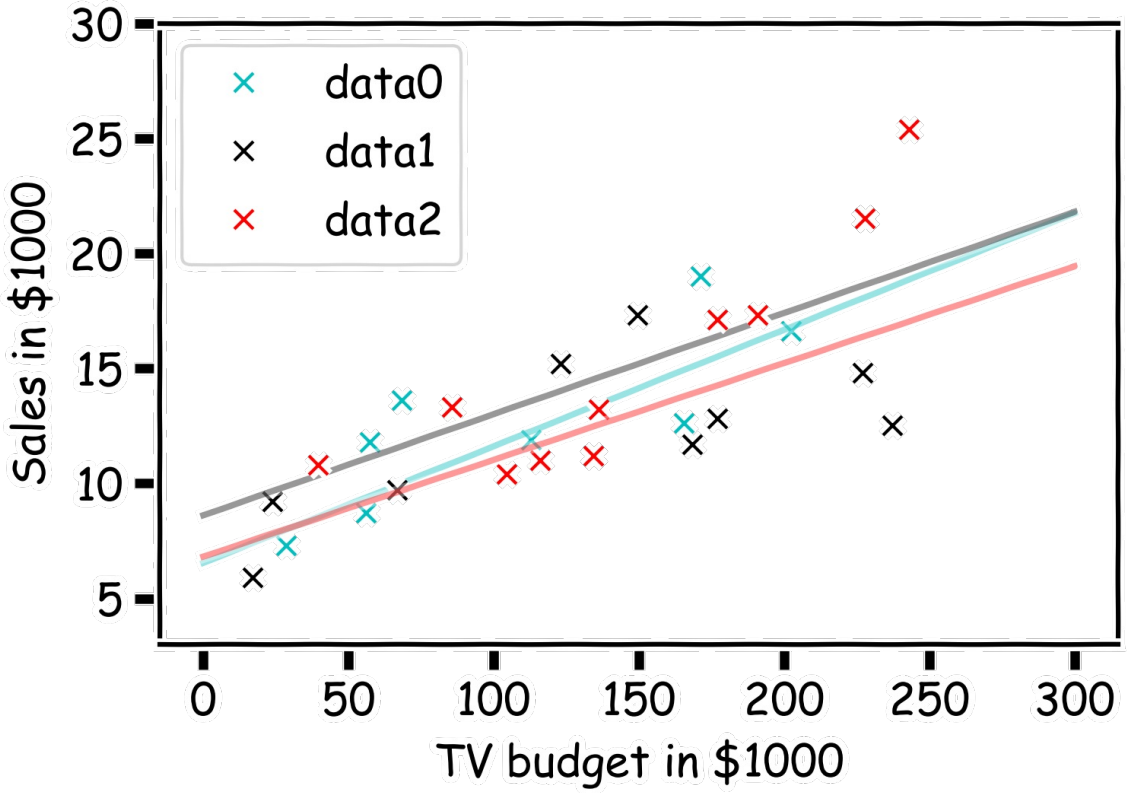
How well do we know $\hat{f}$?

Here we show two difference set of models given the fitted coefficients for a given subsample



How well do we know $\hat{f}$?

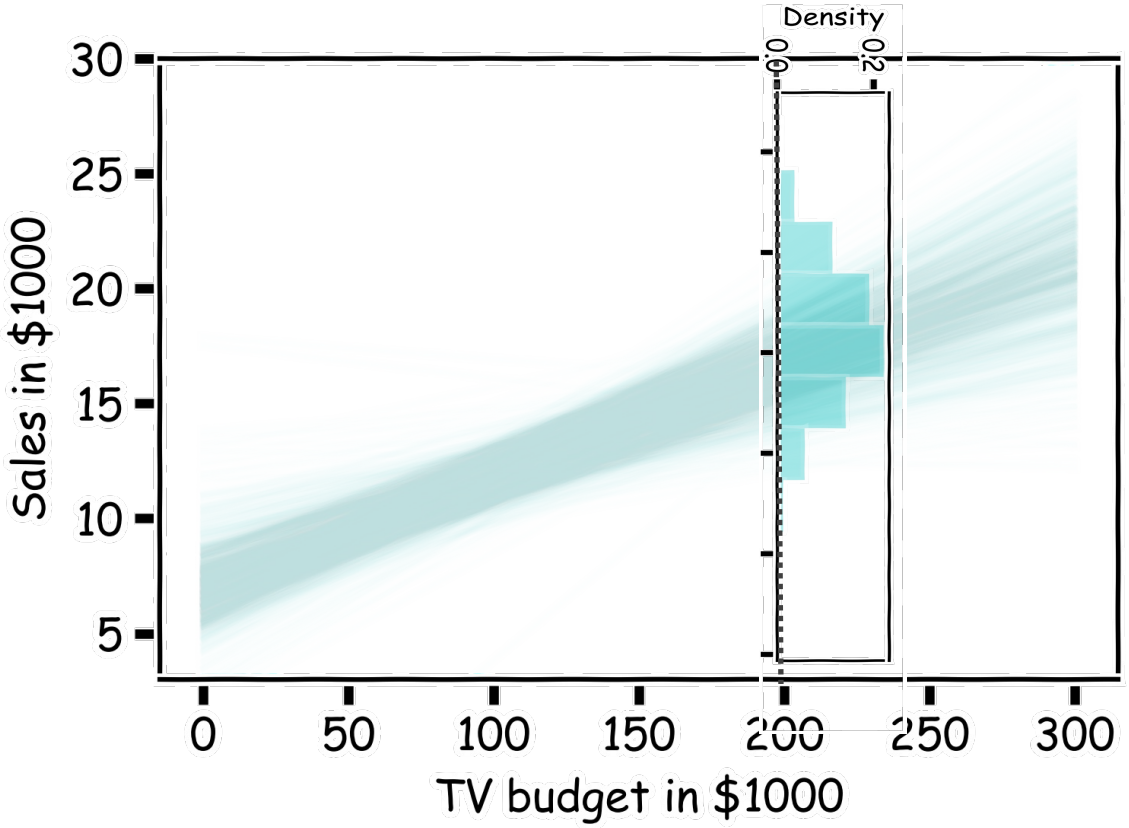
There is one such regression line for every imaginable sub-sample.



How well do we know $\hat{f}$?

Below we show all regression lines for a thousand of such sub-samples.

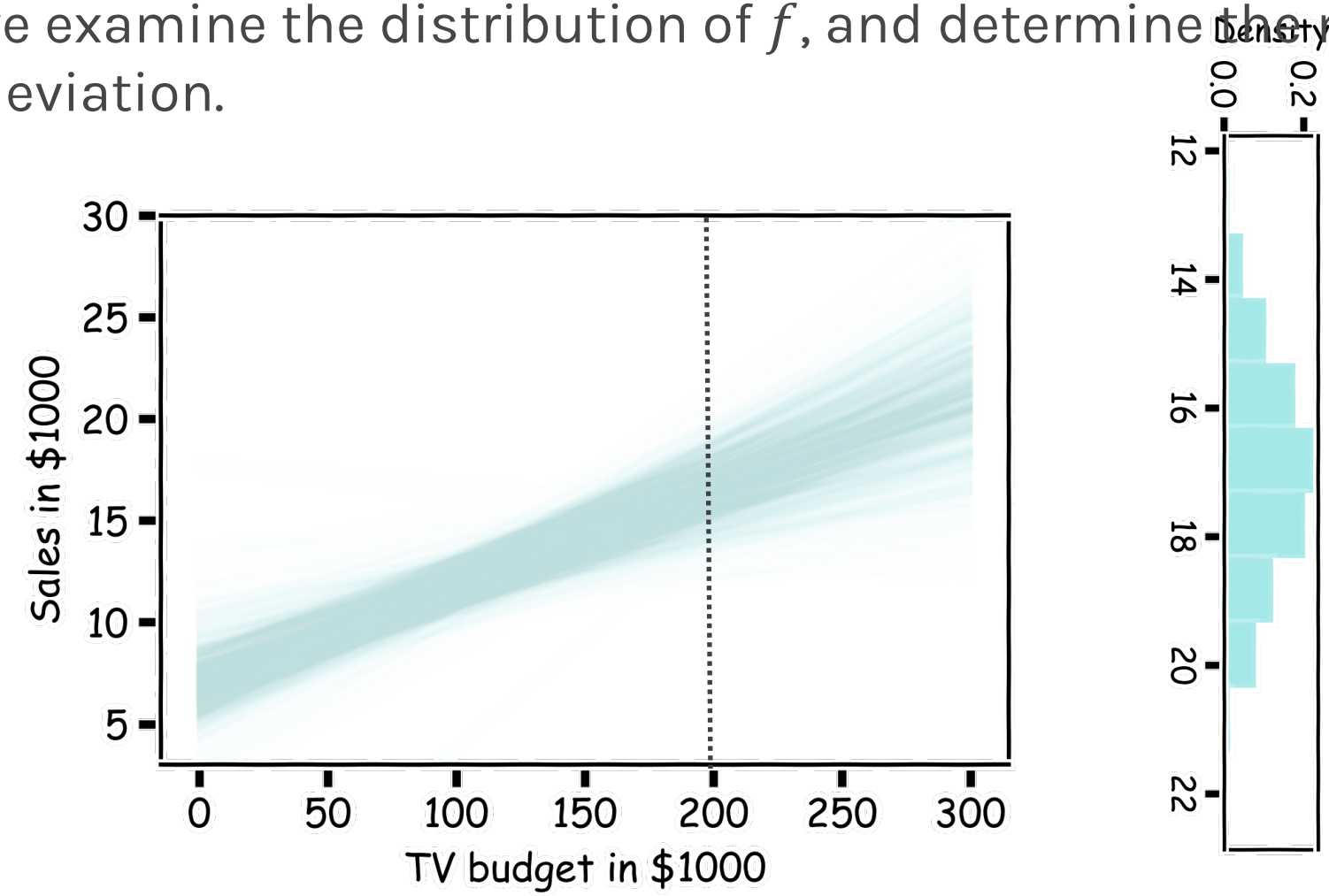
For a given x, we examine the distribution of $\hat{f}$, and determine the mean and standard deviation.



How well do we know $\hat{f}$?

Below we show all regression lines for a thousand of such sub-samples.

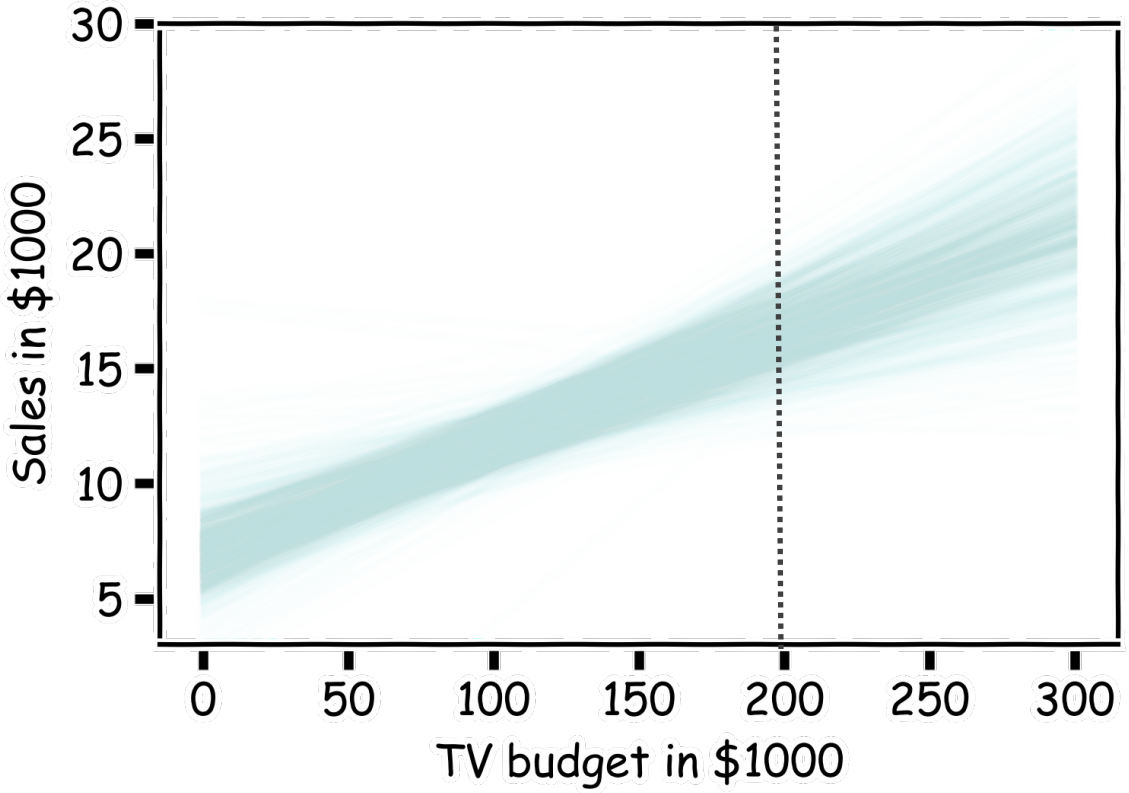
For a given x, we examine the distribution of $\hat{f}$, and determine the mean and standard deviation.

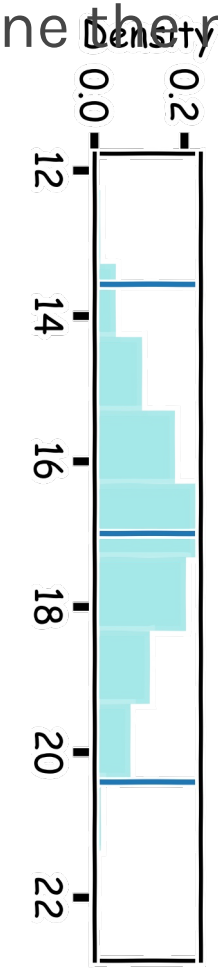


How well do we know $\hat{f}$?

Below we show all regression lines for a thousand of such sub-samples.

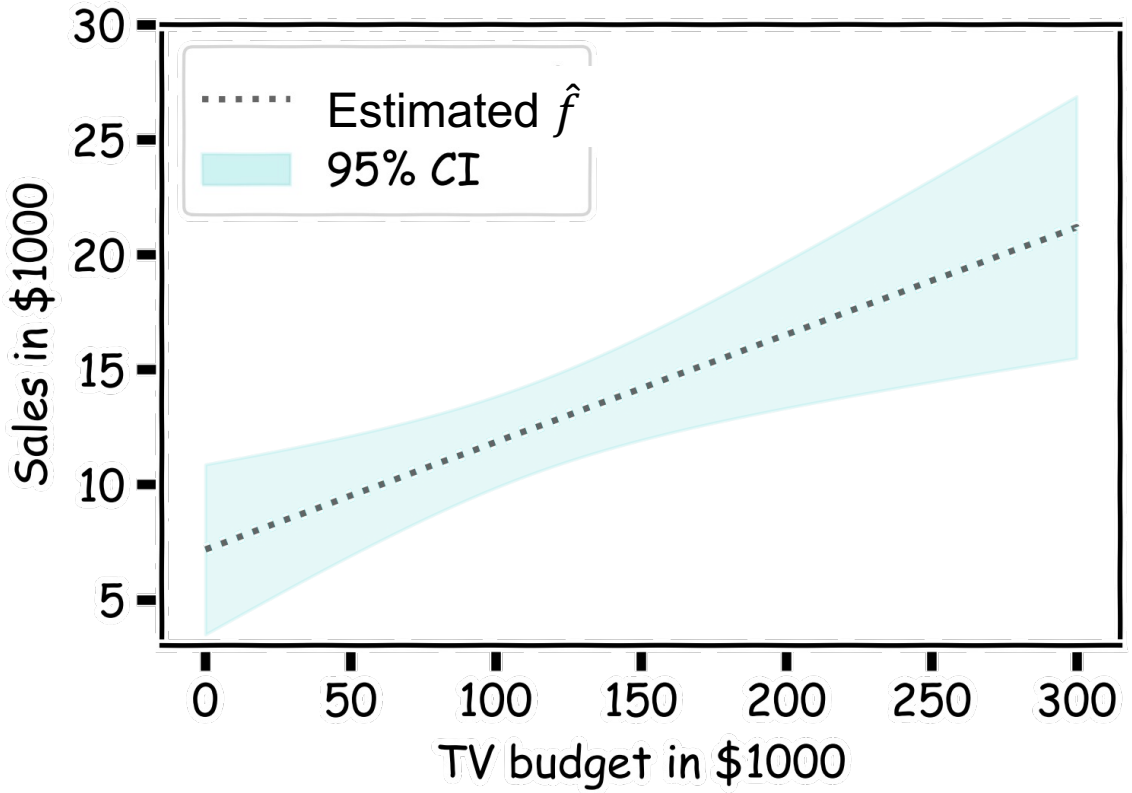
For a given x, we examine the distribution of $\hat{f}$, and determine the mean and standard deviation.





How well do we know $\hat{f}$?

For every x, we calculate the mean of the models, $\hat{f}$ (shown with dotted line) and the 95% CI of those models (shaded area).



Statistical Inference about β_0 and β_1

To make an inference of the true values of $\hat{\beta}_0$ and $\hat{\beta}_1$, say construct their confidence intervals, or do hypothesis test, the four basic assumptions of the simple linear regression model need to be satisfied

1. The distribution of e at any particular x value has mean value 0. that is, $\mu_e = 0$.
2. The standard deviation of e is the same for any particular value of x. This standard deviation is denoted by σ.
3. The distribution of e at any particular value of x is normal.
4. The random deviations e_1, e_2,..., e_n associated with different observations are independent of one another.

Model Assumptions

Under assumptions 1-4, we have

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right) \quad \textbf{and} \quad \hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{XX}}\right)$$

$$SXX = \Sigma\,(x - \bar{x})^2 = \Sigma\, x^2 - \frac{(\Sigma\, x)^2}{n} \qquad \text{(Sum Squares X)}$$

- If σ^2 is known, then we can use these results directly to construct a confidence interval and to formulate a test statement of β_0 and β_1. However, we want to deal with a more practical problem, that is, the problem with Unknown σ^2.

Model Assumptions

- First, we need to know how to estimate the common population variance of the **random error terms**.
- Recall that residual is often regarded as an "actual" value of the unobservable random error term. **Thus, we can use the sample variance of the residual to estimate the unknown population variance σ^2**.

Model Assumptions

The estimate of population standard deviation calculated from a random sample is:

$$s^2 = \frac{\sum_{i-1}^{n}(x_i - \bar{x})^2}{n-1}$$

n-1 is the degrees of freedom

The Mean Squared Error (hereafter, MSE) defined as

$$S^2 = \frac{\sum_{i=1}^{n} E_i^2}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$$

Why n - 2?

Note that the degrees of freedom associated with estimating σ^2 or σ in simple linear regression is df = n - 2

- Its actual value $S^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$ is also called MSE, for simplicity.

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

Model Assumptions - Remarks

- MSE can also be found in the following way:

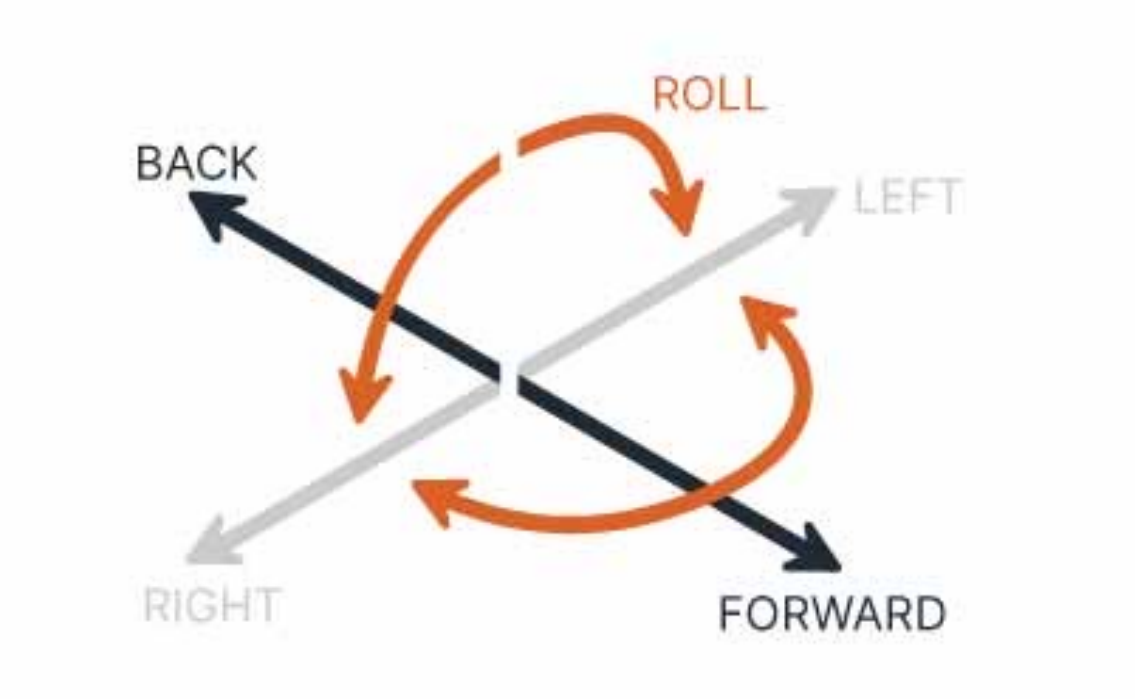
$$S^2 = \frac{S_{YY} - bS_{XY}}{n-2},$$

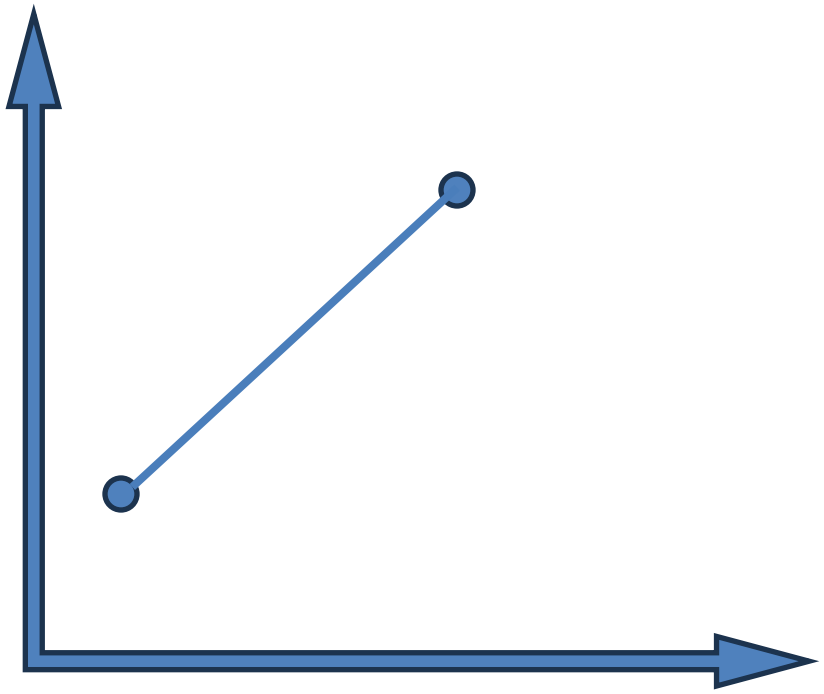
where

$$SYY = \Sigma(y - \bar{y})^2 = \Sigma\, y^2 - \frac{(\Sigma\, y)^2}{n} \qquad \text{(Sum Squares Y)}$$

$$SXY = \Sigma(x - \bar{x})(y - \bar{y}) = \Sigma\, xy - \frac{(\Sigma x)(\Sigma y)}{n} \qquad \text{(Sum Products X,Y)}$$

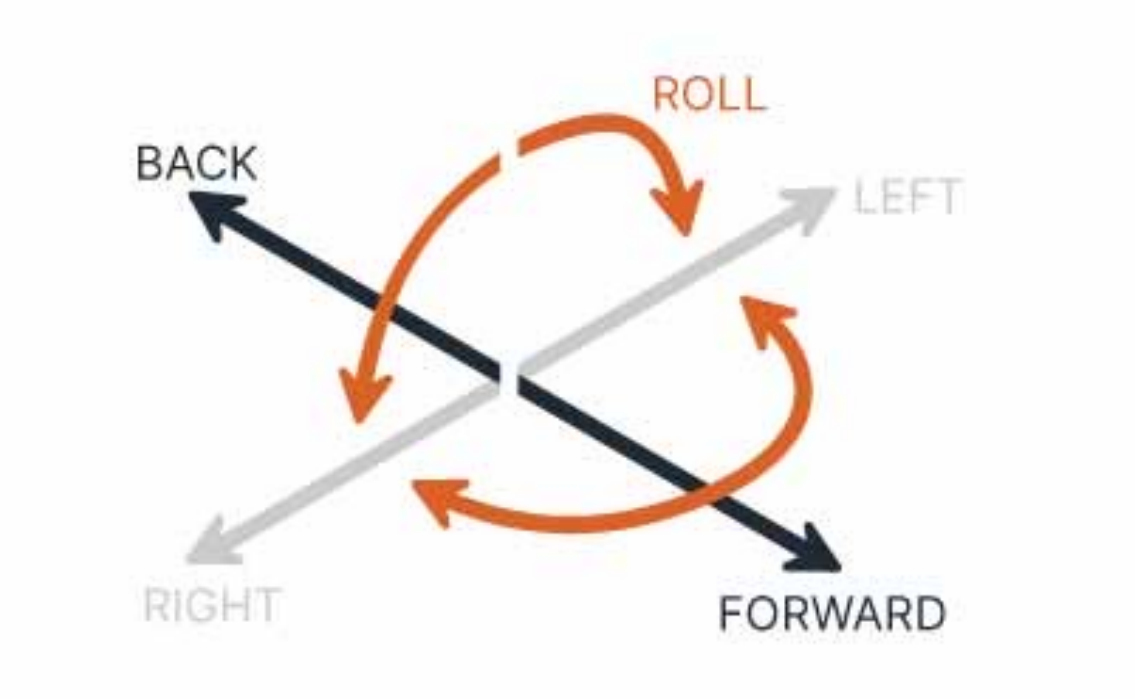
Degrees of freedom in linear regression

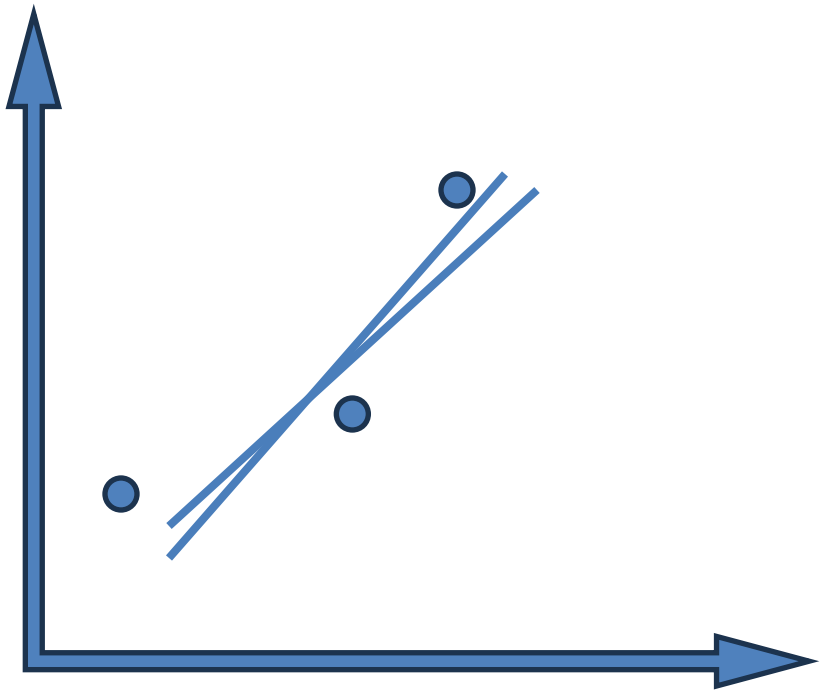




n=2, we can only draw one line

Degrees of freedom in linear regression





n=3, can one degree of freedom

Degrees of freedom in linear regression

Total Degrees of Freedom (df_total)= n - 1.

Model Degrees of Freedom (df_model): For each parameter estimated in the regression model, you lose one degree of freedom. In a simple linear regression with one independent variable and an intercept, you are estimating two parameters: **the slope and the intercept**. Thus, the model degrees of freedom is 2 in this case**. In general, if you have p predictors (not including the intercept), the model degrees of freedom will be p + 1 (including the intercept)**.

Residual Degrees of Freedom (df_residual): This represents the degrees of freedom associated with the residuals of the model, which is the variation in the observations not explained by the model. **df_residual = n- (p+1)**

→For a simple linear regression, this would be n - 2.

Standard Error of the Slope Estimate (SE of β_1)

The Standard Error of the Slope Estimate in regression analysis is a measure of the precision with which the slope coefficient (often denoted as $SE_{\beta 1}$) is estimated.

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

$$= \sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$SE_{\beta 1} = \sqrt{\frac{MSE}{S_{XX}}} = \sqrt{\frac{S^2}{S_{XX}}}$$

- y_i are the observed dependent variable values,
- $\hat{y}_i$ are the predicted values from the regression line,
- x_i are the observed independent variable values,
- $\bar{x}$ is the mean of the independent variable values,
- n is the number of observations

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

$$SXX = \Sigma\,(x - \bar{x})^2 = \Sigma\,x^2 - \frac{(\Sigma\,x)^2}{n} \quad \text{(Sum Squares X)}$$

The Standard Error of β_1 $SE_{\beta 1}$ is used in constructing confidence intervals and hypothesis testing.

CityU

Confidence Intervals for β_0 and β_1

Recall the general equation for the CI:

$$CI \text{ for } \mu = \bar{x} \pm t_{\alpha/2,df} \times SE_{\bar{x}}$$

- $\bar{x}$ is the sample mean,
- $t_{\alpha/2,df}$ is the t-value from the t-distribution for the desired confidence level and degrees of freedom (df),
- $SE_{\bar{x}}$ is the standard error of the mean, which is calculated as the sample standard deviation (s) divided by the square root of the sample size (n):

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Similarly, The confidence interval for β_1, the slope in linear regression, can be determined

$$CI \text{ for } \beta_1 = \hat{\beta}_1 \pm t_{\alpha/2,df} \times SE(\hat{\beta}_1)$$

$$SE_{\beta 1} = \sqrt{\frac{MSE}{S_{XX}}} = \sqrt{\frac{S^2}{S_{XX}}}$$

Consequently, **the 100(1- α)% C.I. for β_1 is given by**

$$b \pm t_{n-2,\frac{\alpha}{2}} \sqrt{\frac{S^2}{S_{XX}}}$$

Confidence Intervals for β_0 and β_1

The confidence interval for β_0, the y-intercept in linear regression, can be determined using a similar approach to that of the slope.

$$CI \text{ for } \beta_0 = \hat{\beta}_0 \pm t_{\alpha/2,df} \times SE(\hat{\beta}_0)$$

$$SE_{\beta 0} = \sqrt{\frac{S^2 \sum_{i=1}^{n} x_i^2}{nS_{XX}}}$$

Consequently, **the 100(1- α)% C.I. for β_0 is given by**

$$(\bar{y} - b\bar{x}) \pm t_{n-2,\frac{\alpha}{2}} \sqrt{\frac{S^2 \sum_{i=1}^{n} x_i^2}{nS_{XX}}}$$

s: is the standard deviation of the residuals, square root of MSE

Hypothesis Testing for β_0 and β_1

- **For the slope β_1 ,**

1. One-sided right test:

Consider $H_0 : \beta_1 = b_1$

$H_1 : \beta_1 > b_1$.

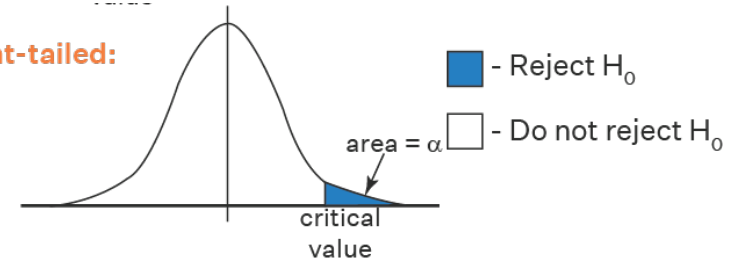
Test statistics

$$t = \frac{\hat{\beta}_j - b_1}{SE(\hat{\beta}_j)}$$

$$SE_{\beta 1} = \sqrt{\frac{MSE}{S_{XX}}} = \sqrt{\frac{S^2}{S_{XX}}}$$

Reject H_0 at a significance level α if

the t value $\frac{b - b_1}{\frac{s}{\sqrt{S_{XX}}}} > t_{n-2,\alpha}$



(when σ_x^2 is UNKNOWN)

Lies in the critical region → reject H_o

Hypothesis Testing for β_0 and β_1

2. One-sided left test:

Consider $H_0: \beta_1 = b_1$
$H_1: \beta_1 < b_1$.

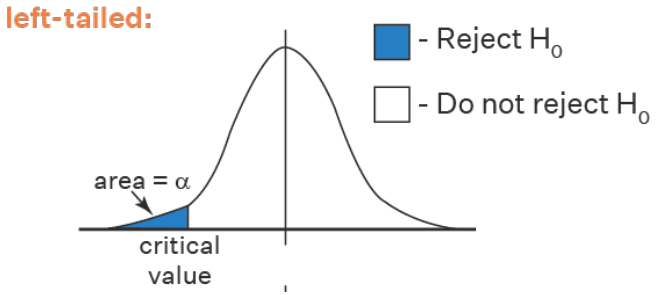
Test statistics

$$t = \frac{\hat{\beta}_j - b_1}{SE(\hat{\beta}_j)}$$

$$SE_{\beta 1} = \sqrt{\frac{MSE}{S_{XX}}} = \sqrt{\frac{S^2}{S_{XX}}}$$

Reject H_0 at a significance level α if

the t value $\frac{b - b_1}{\frac{s}{\sqrt{S_{XX}}}} < -t_{n-2,\alpha}$



(when σ_x^2 is UNKNOWN)

Lies in the critical region → reject H_o

Hypothesis Testing for β_0 and β_1

3. Two-sided test:

Consider $H_0: \beta_1 = b_1$
$H_1: \beta_1 \neq b_1$.

Test statistics

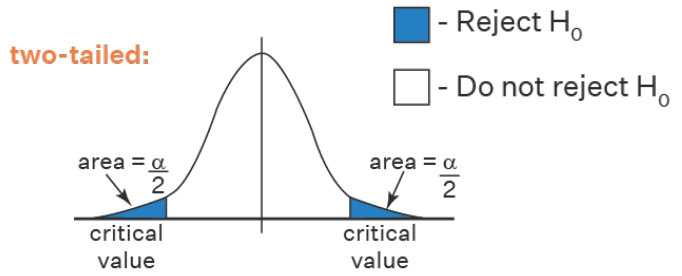
$$t = \frac{\hat{\beta}_j - b_1}{SE(\hat{\beta}_j)}$$

$$SE_{\beta 1} = \sqrt{\frac{MSE}{S_{XX}}} = \sqrt{\frac{S^2}{S_{XX}}}$$

Reject H_0 at a significance level α if

the absolute t value $\left|\frac{b-b_1}{\frac{s}{\sqrt{S_{XX}}}}\right| > t_{n-2,\frac{\alpha}{2}}$

(when σ_x^2 is UNKNOWN)



Lies in the critical region → reject H_o

Hypothesis Testing for β_0 and β_1

- **Similarly, for the intercept β_0 ,**

1. One-sided right test:

Consider $H_0: \beta_0 = b_0$
$H_1: \beta_0 > b_0$.

Test statistics

$$t = \frac{\hat{\beta}_j - b_0}{SE(\hat{\beta}_j)}$$

$$SE_{\beta 0} = \sqrt{\frac{S^2 \sum_{i=1}^{n} x_i^2}{nS_{XX}}}$$

Reject H_0 at a significance level α if

$$\text{the } t \text{ value } \frac{a - b_0}{s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{nS_{XX}}}} > t_{n-2,\alpha}$$

(when σ_x^2 is UNKNOWN)

Lies in the critical region → reject H_o

Hypothesis Testing for β_0 and β_1

2. One-sided left test:

Consider $H_0: \beta_0 = b_0$
$H_1: \beta_0 < b_0$

Test statistics

$$t = \frac{\hat{\beta}_j - b_0}{SE(\hat{\beta}_j)}$$

$$SE_{\beta 0} = \sqrt{\frac{S^2 \sum_{i=1}^{n} x_i^2}{nS_{XX}}}$$

Reject H_0 at a significance level α if

$$\text{the } t \text{ value } \frac{a - b_0}{s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{nS_{XX}}}} < -t_{n-2,\alpha}$$

(when σ_x^2 is UNKNOWN)

Lies in the critical region → reject H_o

Hypothesis Testing for β_0 and β_1

Test statistics

$$t = \frac{\hat{\beta}_j - b_0}{SE(\hat{\beta}_j)}$$

$$SE_{\beta 0} = \sqrt{\frac{S^2 \sum_{i=1}^{n} x_i^2}{nS_{XX}}}$$

3. Two-sided test:

Consider $H_0: \beta_0 = b_0$

$H_1: \beta_0 \neq b_0$.

Reject H_0 at a significance level α if

the absolute t value $\left| \frac{a - b_0}{s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{nS_{XX}}}} \right| > t_{n-2,\frac{\alpha}{2}}$

(when σ_x^2 is UNKNOWN)

Lies in the critical region → reject H_o

Statistical relationship between two variables.

Statistical relationship between two variables

Pearson correlation coefficient, which measures the strength and direction of the linear relationship between two continuous variables.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

- r=1: A perfect positive linear relationship.
- r=−1: A perfect negative linear relationship.
- r=0: No linear relationship.

- 0<r<1: A positive linear relationship.
- −1<r<0: A negative linear relationship.

Short Summary

1. S_{XX} is the sum of the squares of the difference between each *x* and the mean *x* value.

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i^2\right)/n$$

2. S_{YY} is the sum of the squares of the difference between each *y* and the mean *y* value.

$$S_{YY} = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i^2\right)/n$$

3. S_{xy} is sum of the product of the difference between *x* its means and the difference between *y* and its mean.

$$S_{XY} = \sum_{i=1}^{n} (x_i y_i) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n$$

4. **Mean squared error (MSE**) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and what is estimated.

$$MSE = S^2 = \frac{S_{YY} - bS_{XY}}{n-2}$$

5. ***r*** means a statistical relationship between two variables.

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

6. *a* and *b* are called the least-squares ESTIMATES of the unknown true values of β_0 (α) and β_1 (β), respectively.

$$b = \frac{S_{XY}}{S_{XX}}$$

$$a = \bar{y} - b\bar{x}$$

Example

The graduation rate (%) and student's median SAT score for a random sample of the primarily undergraduate public universities and colleges is shown in the table

Median SAT	Expenditure	Graduation Rate
1065	7970	49
950	6401	33
1045	6285	37
990	6792	49
950	4541	22
970	7186	38
980	7736	39
1080	6382	52
1035	7323	53
1010	6531	41
1010	6216	38
930	7375	37
1005	7874	45
1090	6355	57
1085	6261	48

Let's investigate the relationship between graduation rate and median SAT score. With y = graduation rate and x = median SAT score, the summary statistics necessary for a simple linear regression analysis are as follows:

1. Estimate the linear regression coefficient and the correlation coefficient r

$$n = 15; \sum x = 15{,}195; \sum y = 638$$
$$\sum x^2 = 15{,}430{,}725; \sum y^2 = 28{,}294; \sum xy = 651{,}340$$

$$b = \frac{\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n}{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}}$$

$$a = \bar{y} - b\bar{x}$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

$$n = 15; \sum x = 15{,}195; \sum y = 638$$
$$\sum x^2 = 15{,}430{,}725; \sum y^2 = 28{,}294; \sum xy = 651{,}340$$

$$b = \frac{\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n}{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}}$$

$$a = \bar{y} - b\bar{x}$$

$$MSE = S^2 = \frac{S_{YY} - bS_{XY}}{n - 2}$$

$$S_{XY} = \sum_{i=1}^{n}(x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n$$

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i^2)/n$$

$$S_{YY} = \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i^2)/n$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

$b = 0.132$; $a = -91.31$;
$S = 6.146$; $r^2 = 0.576$

Because $r^2 = 0.576$, about 57.6% of observed variation in graduation rates can be explained by the simple linear regression model.

It appears from this that there is a useful linear relation between the two variables, but a confirmation requires a formal model utility test.

We will use a significance level of 0.05 to carry out this test.

1, β_1: the average change in graduation rate associated with an increase of 1 in median SAT score

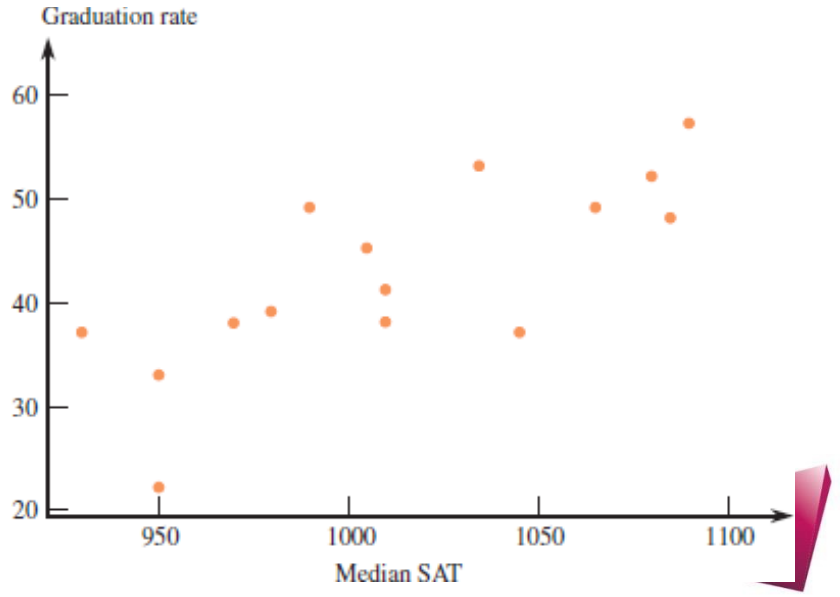
2. H_0: $\beta_1 = 0$ (means that *y* does not change with *x* and there is no relationship between *x* and *y*)

3. H_1: $\beta_1 \neq 0$ (does not equal to 0 means that *y* changes with *x*)

4. $\alpha = 0.05$

5. Assumptions: The data are from a random sample, so the observations are independent.

The accompanying scatterplot of the data shows a linear pattern and the variability of points does not appear to be changing with *x*; Assuming that the distribution of errors at any given *x* value is approximately normal, the assumptions of the simple linear regression model are appropriate.



6. Calculation: $\frac{b-b_1}{\frac{s}{\sqrt{S_{XX}}}} = \frac{0.132-0}{\frac{6.146}{\sqrt{38190}}} = 4.20$

7. From *t*-statistical table, we get: Critical value: $t(13,0.025) = 2.16$

t distribution critical values

	Upper-tail pr					
df	.25	.20	.15	.10	.05	.025
1	1.000	1.376	1.963	3.078	6.314	12.71
2	0.816	1.061	1.386	1.886	2.920	4.303
3	0.765	0.978	1.250	1.638	2.353	3.182
4	0.741	0.941	1.190	1.533	2.132	2.776
5	0.727	0.920	1.156	1.476	2.015	2.571
6	0.718	0.906	1.134	1.440	1.943	2.447
7	0.711	0.896	1.119	1.415	1.895	2.365
8	0.706	0.889	1.108	1.397	1.860	2.306
9	0.703	0.883	1.100	1.383	1.833	2.262
10	0.700	0.879	1.093	1.372	1.812	2.228
11	0.697	0.876	1.088	1.363	1.796	2.201
12	0.695	0.873	1.083	1.356	1.782	2.179
13	0.694	0.870	1.079	1.350	1.771	2.160
14	0.692	0.868	1.076	1.345	1.761	2.145
15	0.691	0.866	1.074	1.341	1.753	2.131

8. Test statistics > *t*(13,0.025)

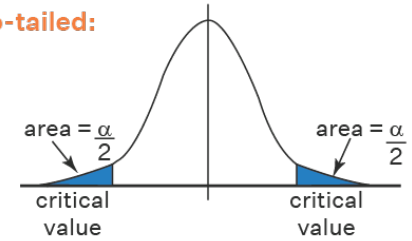
Reject H_0 at a significance level α if

the absolute *t* value $\left|\frac{a-b_0}{s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{nS_{XX}}}}\right| > t_{n-2,\frac{\alpha}{2}}$

(when σ_x^2 is UNKNOWN)

Lies in the critical region → reject H_o

two-tailed:



9. We reject H_0 at a significance level $\alpha = 0.05$. We conclude that there is a useful linear relationship between graduation rate and median SAT score.

Example

The article "Effect of Temperature on the pH of Sea Water" reported on a study involving x = temperature. Under specified experimental conditions and y = sea water pH. The accompanying data (read from a graph) are a representative subset of that which appeared in the article:

x	4	4	24	24	25	38	38	40
y	6.85	6.79	6.63	6.65	6.72	6.62	6.57	6.52
x	45	50	55	56	60	67	70	78
y	6.50	6.48	6.42	6.41	6.38	6.34	6.32	6.34

$\sum x = 678 \quad \sum y = 104.54 \quad \sum x^2 = 36{,}056$

$\sum y^2 = 683.4470 \quad \sum xy = 4376.36$

Do these data strongly suggest that there is a negative linear relationship between temperature and pH? State and test the relevant hypotheses using a significance level of 0.01.

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i^2\right)/n \qquad S_{YY} = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i^2\right)/n$$

$$S_{XY} = \sum_{i=1}^{n} (x_i y_i) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n$$

$$MSE = S^2 = \frac{S_{YY} - bS_{XY}}{n-2}$$

$$b = \frac{S_{XY}}{S_{XX}} \qquad a = \bar{y} - b\bar{x}$$

Example

The article "Effect of Temperature on the pH of Sea Water" reported on a study involving x = temperature. Under specified experimental conditions and y = sea water pH. The accompanying data (read from a graph) are a representative subset of that which appeared in the article:

x	4	4	24	24	25	38	38	40
y	6.85	6.79	6.63	6.65	6.72	6.62	6.57	6.52
x	45	50	55	56	60	67	70	78
y	6.50	6.48	6.42	6.41	6.38	6.34	6.32	6.34

$\sum x = 678 \quad \sum y = 104.54 \quad \sum x^2 = 36{,}056$

$\sum y^2 = 683.4470 \quad \sum xy = 4376.36$

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i^2\right)/n \qquad S_{YY} = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i^2\right)/n$$

$$S_{XY} = \sum_{i=1}^{n} (x_i y_i) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n$$

Do these data strongly suggest that there is a negative linear relationship between temperature and pH? State and test the relevant hypotheses using a significance level of 0.01.

b = -0.0073, S_{xx}= 7325.75, S_{yy}= 0.409, S_{xy}= -53.5225, S^2 = 0.00131, a = 6.843

$$MSE = S^2 = \frac{S_{YY} - bS_{XY}}{n-2}$$

$$b = \frac{S_{XY}}{S_{XX}} \qquad a = \bar{y} - b\bar{x}$$

1. To test whether temperature is a statistically significant negative linear of *pH*, we can set up the following hypotheses:
2. H_0: $\beta_1 = 0$
3. H_1: $\beta_1 < 0$
4. $\alpha = 0.01$

Example

The article "Effect of Temperature on the pH of Sea Water" reported on a study involving x = temperature. Under specified experimental conditions and y = sea water pH. The accompanying data (read from a graph) are a representative subset of that which appeared in the article:

x	4	4	24	24	25	38	38	40
y	6.85	6.79	6.63	6.65	6.72	6.62	6.57	6.52
x	45	50	55	56	60	67	70	78
y	6.50	6.48	6.42	6.41	6.38	6.34	6.32	6.34

$\sum x = 678 \quad \sum y = 104.54 \quad \sum x^2 = 36{,}056$

$\sum y^2 = 683.4470 \quad \sum xy = 4376.36$

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i^2\right)/n \qquad S_{YY} = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i^2\right)/n$$

$$S_{XY} = \sum_{i=1}^{n} (x_i y_i) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n$$

Do these data strongly suggest that there is a negative linear relationship between temperature and pH? State and test the relevant hypotheses using a significance level of 0.01.

b = -0.0073, S_{xx}= 7325.75, S_{yy}= 0.409, S_{xy}= -53.5225, S^2 = 0.00131, a = 6.843

$$MSE = S^2 = \frac{S_{YY} - bS_{XY}}{n-2}$$

$$b = \frac{S_{XY}}{S_{XX}} \qquad a = \bar{y} - b\bar{x}$$

1. To test whether temperature is a statistically significant negative linear of *pH*, we can set up the following hypotheses:
2. H_0: $\beta_1 = 0$
3. H_1: $\beta_1 < 0$
4. α = 0.01
5. Test statistic: $t = \frac{b - b_1}{\frac{s}{\sqrt{S_{XX}}}}$ = (-0.0073-0)/0.0004229 = -17.3

Example

The article "Effect of Temperature on the pH of Sea Water" reported on a study involving x = temperature. Under specified experimental conditions and y = sea water pH. The accompanying data (read from a graph) are a representative subset of that which appeared in the article:

x	4	4	24	24	25	38	38	40
y	6.85	6.79	6.63	6.65	6.72	6.62	6.57	6.52
x	45	50	55	56	60	67	70	78
y	6.50	6.48	6.42	6.41	6.38	6.34	6.32	6.34

$\sum x = 678 \quad \sum y = 104.54 \quad \sum x^2 = 36{,}056$

$\sum y^2 = 683.4470 \quad \sum xy = 4376.36$

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i^2\right)/n \qquad S_{YY} = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i^2\right)/n$$

Do these data strongly suggest that there is a negative linear relationship between temperature and pH? State and test the relevant hypotheses using a significance level of 0.01.

$$S_{XY} = \sum_{i=1}^{n} (x_i y_i) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n$$

b = -0.0073, S_{xx}= 7325.75, S_{yy}= 0.409, S_{xy}= -53.5225, S^2 = 0.00131, a = 6.843

$$MSE = S^2 = \frac{S_{YY} - bS_{XY}}{n-2}$$

$$b = \frac{S_{XY}}{S_{XX}} \qquad a = \bar{y} - b\bar{x}$$

1. To test whether temperature is a statistically significant negative linear of *pH*, we can set up the following hypotheses:
2. H_0: $\beta_1 = 0$
3. H_1: $\beta_1 < 0$
4. α = 0.01
5. Test statistic: $t = \frac{b - b_1}{\frac{s}{\sqrt{S_{XX}}}}$ = (-0.0073-0)/0.0004229 = -17.3

From t-statistical table, we get:
Critical value: $t(14, 0.01) = 2.624$

t distribution critical values

	Upper-tail probability p							
df	.25	.20	.15	.10	.05	.025	.02	.01
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624

Example

The article "Effect of Temperature on the pH of Sea Water" reported on a study involving x = temperature. Under specified experimental conditions and y = sea water pH. The accompanying data (read from a graph) are a representative subset of that which appeared in the article:

x	4	4	24	24	25	38	38	40
y	6.85	6.79	6.63	6.65	6.72	6.62	6.57	6.52
x	45	50	55	56	60	67	70	78
y	6.50	6.48	6.42	6.41	6.38	6.34	6.32	6.34

$\sum x = 678 \quad \sum y = 104.54 \quad \sum x^2 = 36{,}056$

$\sum y^2 = 683.4470 \quad \sum xy = 4376.36$

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i^2\right)/n \qquad S_{YY} = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i^2\right)/n$$

Do these data strongly suggest that there is a negative linear relationship between temperature and pH? State and test the relevant hypotheses using a significance level of 0.01.

$$S_{XY} = \sum_{i=1}^{n} (x_i y_i) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n$$

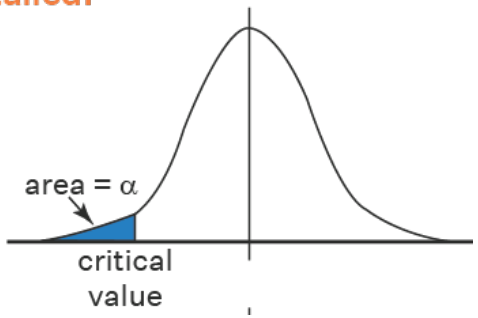
b = -0.0073, S_{xx}= 7325.75, S_{yy}= 0.409, S_{xy}= -53.5225, S^2 = 0.00131, a = 6.843

$$MSE = S^2 = \frac{S_{YY} - bS_{XY}}{n-2}$$

$$b = \frac{S_{XY}}{S_{XX}} \qquad a = \bar{y} - b\bar{x}$$

1. To test whether temperature is a statistically significant negative linear of *pH*, we can set up the following hypotheses:
2. H_0: $\beta_1 = 0$
3. H_1: $\beta_1 < 0$
4. α = 0.01
5. Test statistic: $t = \frac{b - b_1}{\frac{s}{\sqrt{S_{XX}}}}$ = (-0.0073-0)/0.0004229 = -17.3

left-tailed:



From *t*-statistical table, we get:
Critical value: $t(14,0.01)$=2.624

we reject *H0*

t distribution critical values

	Upper-tail probability p							
df	.25	.20	.15	.10	.05	.025	.02	.01
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624

There is a negative linear relationship between temperature and pH

Example

An experiment to study the relationship between x = time spent exercises (minutes) and y = amount of oxygen consumed during the exercise period resulted in the following summary statistics.

$$n = 20, \sum x = 50, \sum y = 16705, \sum x^2 = 150, \sum y^2 = 14194231, \sum xy = 44194$$

a) Estimate the slope and y intercept of the population regression line

b) One sample observation on oxygen usage was 757 for a 2-minute exercise period. What amount of oxygen consumption would you predict for this exercise period, and what is the corresponding residual?

c) Compute a 99% confidence interval for the average change in oxygen consumption associated with 1-minute increase in exercise time.

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i^2\right)/n \qquad S_{YY} = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i^2\right)/n$$

$$S_{XY} = \sum_{i=1}^{n} (x_i y_i) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n$$

$$MSE = S^2 = \frac{S_{YY} - bS_{XY}}{n-2}$$

$$b = \frac{S_{XY}}{S_{XX}} \qquad a = \bar{y} - b\bar{x}$$

Example

An experiment to study the relationship between x = time spent exercises (minutes) and y = amount of oxygen consumed during the exercise period resulted in the following summary statistics.

$$n = 20, \sum x = 50, \sum y = 16705, \sum x^2 = 150, \sum y^2 = 14194231, \sum xy = 44194$$

a) Estimate the slope and y intercept of the population regression line

b) One sample observation on oxygen usage was 757 for a 2-minute exercise period. What amount of oxygen consumption would you predict for this exercise period, and what is the corresponding residual?

c) Compute a 99% confidence interval for the average change in oxygen consumption associated with 1-minute increase in exercise time.

Solution

a)

$$S_{XY} = \sum_{i=1}^{n}(x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n$$

$$= 44194 - \frac{50 \times 16705}{20} = 2431.5$$

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n = 150 - \frac{50 \times 50}{20} = 25$$

$$\bar{x} = 2.5;\ \bar{y} = 835.25$$

$$b = \frac{S_{XY}}{S_{XX}} = \frac{2431.5}{25} = 97.26$$

$$a = \bar{y} - b\bar{x} = 835.25 - 97.26 \times 2.5 = 592.1$$

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i^2)/n \qquad S_{YY} = \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i^2)/n$$

$$S_{XY} = \sum_{i=1}^{n}(x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n$$

$$MSE = S^2 = \frac{S_{YY} - bS_{XY}}{n-2}$$

$$b = \frac{S_{XY}}{S_{XX}} \qquad a = \bar{y} - b\bar{x}$$

Example

An experiment to study the relationship between x = time spent exercises (minutes) and y = amount of oxygen consumed during the exercise period resulted in the following summary statistics.

$$n = 20, \sum x = 50, \sum y = 16705, \sum x^2 = 150, \sum y^2 = 14194231, \sum xy = 44194$$

a) Estimate the slope and y intercept of the population regression line

b) One sample observation on oxygen usage was 757 for a 2-minute exercise period. What amount of oxygen consumption would you predict for this exercise period, and what is the corresponding residual?

c) Compute a 99% confidence interval for the average change in oxygen consumption associated with 1-minute increase in exercise time.

Solution

a)

$$S_{XY} = \sum_{i=1}^{n}(x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n$$

$$= 44194 - \frac{50 \times 16705}{20} = 2431.5$$

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n = 150 - \frac{50 \times 50}{20} = 25$$

$$\bar{x} = 2.5;\ \bar{y} = 835.25$$

$$b = \frac{S_{XY}}{S_{XX}} = \frac{2431.5}{25} = 97.26$$

$$a = \bar{y} - b\bar{x} = 835.25 - 97.26 \times 2.5 = 592.1$$

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i^2)/n \qquad S_{YY} = \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i^2)/n$$

$$S_{XY} = \sum_{i=1}^{n}(x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n$$

$$MSE = S^2 = \frac{S_{YY} - bS_{XY}}{n-2}$$

$$b = \frac{S_{XY}}{S_{XX}} \qquad a = \bar{y} - b\bar{x}$$

Example

An experiment to study the relationship between x = time spent exercises (minutes) and y = amount of oxygen consumed during the exercise period resulted in the following summary statistics.

$$n = 20, \sum x = 50, \sum y = 16705, \sum x^2 = 150, \sum y^2 = 14194231, \sum xy = 44194$$

a) Estimate the slope and y intercept of the population regression line

b) One sample observation on oxygen usage was 757 for a 2-minute exercise period. What amount of oxygen consumption would you predict for this exercise period, and what is the corresponding residual?

c) Compute a 99% confidence interval for the average change in oxygen consumption associated with 1-minute increase in exercise time.

Solution

a)

$$S_{XY} = \sum_{i=1}^{n}(x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n$$

$$= 44194 - \frac{50 \times 16705}{20} = 2431.5$$

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n = 150 - \frac{50 \times 50}{20} = 25$$

$$\bar{x} = 2.5;\ \bar{y} = 835.25$$

$$b = \frac{S_{XY}}{S_{XX}} = \frac{2431.5}{25} = 97.26$$

$$a = \bar{y} - b\bar{x} = 835.25 - 97.26 \times 2.5 = 592.1$$

b)

$$\hat{y} = bx + a = 97.26 \times 2 + 592.1 = 786.62$$

$$y - \hat{y} = 757 - 786.62 = -29.62$$

Example

An experiment to study the relationship between x = time spent exercises (minutes) and y = amount of oxygen consumed during the exercise period resulted in the following summary statistics.

$$n = 20, \sum x = 50, \sum y = 16705, \sum x^2 = 150, \sum y^2 = 14194231, \sum xy = 44194$$

a) Estimate the slope and y intercept of the population regression line

b) One sample observation on oxygen usage was 757 for a 2-minute exercise period. What amount of oxygen consumption would you predict for this exercise period, and what is the corresponding residual?

c) Compute a 99% confidence interval for the average change in oxygen consumption associated with 1-minute increase in exercise time.

c) $r = \dfrac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$ $\quad S_{YY} = \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i^2)/n$

$r^2 = 0.98$

$S^2 = 271.781$

$S = 16.486$

$$MSE = S^2 = \frac{S_{YY} - bS_{XY}}{n-2}$$

Solution

a)

$$S_{XY} = \sum_{i=1}^{n}(x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n$$

$$= 44194 - \frac{50 \times 16705}{20} = 2431.5$$

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n = 150 - \frac{50 \times 50}{20} = 25$$

$$\bar{x} = 2.5;\ \bar{y} = 835.25$$

$$b = \frac{S_{XY}}{S_{XX}} = \frac{2431.5}{25} = 97.26$$

$$a = \bar{y} - b\bar{x} = 835.25 - 97.26 \times 2.5 = 592.1$$

b)

$$\hat{y} = bx + a = 97.26 \times 2 + 592.1 = 786.62$$

$$y - \hat{y} = 757 - 786.62 = -29.62$$

Calculation of the 99% confidence interval for β requires a t critical value based on df = n - 2 = 20 – 2 = 18 → 2.878

$$b \pm (t\ critical\ value) \times \frac{S}{\sqrt{S_{XX}}} = 97.26 \pm 2.878 \times \frac{16.486}{\sqrt{25}} = 106.749\ and\ 87.771$$

Two sides

t distribution critical values

	Upper-tail probability p								
df	.25	.20	.15	.10	.05	.025	.02	.01	.005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878

Multiple regression analysis

In practice, it is unlikely that any response variable Y depends solely on one predictor x. Rather, we expect that is a function of multiple predictors $f(X_1, \ldots, X_J)$.

X
predictors
features
covariates

Y
outcome
response variable
dependent variable

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

n observations

p predictors

Multiple regression can be used to fit models to data with two or more independent variables

Example: Multiple Regression Model: Environment engineers

Consider a construction company in which environment engineers with no prior working experience and no college credits beyond a bachelor's degree start at an annual salary of $38,000. Suppose that for each year of working experience up to 20 years, the engineer receives an additional $800 and that each unit of postgraduate credit up to 75 credits results in an additional $60 per year.

Let:

y = salary of an environment engineer

x_1 = number of years of experience

x_2 = number of postgraduate units

$$y = 38{,}000 + 800x_1 + 60x_2 + e$$

In a simple regression model, x_1 and x_2 represent two observations of a single variable.
In multiple regression, x_1 and x_2 represent two independent variables!

General Additive Multiple Regression Model

A general additive multiple regression model, which relates a dependent variable *y* to *k* predictor variables $x_1, x_2, \ldots, x_k$, is given by the model equation

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

The random deviation *e* is assumed to be normally distributed with mean value 0 and standard deviation *s* for any particular values $x_1, \ldots, x_k$.

This implies that for fixed $x_1, x_2, \ldots, x_k$ values, *y* has a normal distribution with standard deviation *s* and

βs are the population regression coefficients. Each β_i can be interpreted as the mean change in *y* when the predictor x_i increase 1 unit *and* the value of all the other predictors remains fixed.

$$\begin{pmatrix}\text{mean } y \text{ value for fixed} \\ x_1, x_2, \ldots, x_k \text{ values}\end{pmatrix} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Minimize MSE

Again, to fit this model means to compute $\hat{\beta}_0, \ldots, \hat{\beta}_J$ or to minimize the MSE

Given a set of observations,

$$\{(x_{1,1}, \ldots, x_{1,J}, y_1), \ldots (x_{n,1}, \ldots, x_{n,J}, y_n)\},$$

$$Y = f(X_1, \ldots, X_J) + \epsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_J X_J + \epsilon$$

the data and the model can be expressed in vector notation,

$$MSE = \frac{1}{n-(k+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The denominator is n−(k+1) to account for the degrees of freedom used by the *k* independent variables and the intercept term.

Confidence interval

For each coefficient βj (including both slopes and the intercept), the confidence interval is typically given by:

$$\beta_j \pm t^* \cdot SE(\beta_j)$$

- β_j is the estimated coefficient for the jth term (this could be the intercept or one of the slopes).
- t^* is the value from the t-distribution that corresponds to the desired confidence level (for example, 95%) and $n - k - 1$ degrees of freedom, where n is the number of observations and k is the number of independent variables.
- $SE(\beta_j)$ is the standard error of the estimated coefficient β_j.

Standard Error

The standard error of each coefficient is computed from the square root of the diagonal elements of the covariance matrix of the coefficients.

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_y \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,J} \\ 1 & x_{2,1} & \dots & x_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,J} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_J \end{pmatrix},$$

The covariance matrix is often computed as:

$$\mathrm{Cov}(\hat{\beta}) = (X^T X)^{-1} \cdot MSE$$

$$\mathrm{MSE}(\beta) = \frac{1}{n} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

Python's statsmodels can perform these calculations automatically when you fit a multiple regression model. If you're doing this by hand, it is labor-intensive and requires a good understanding of matrix algebra.

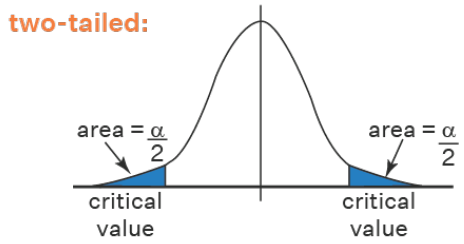
Hypothesis Testing

1. Formulate the null hypothesis and alternative hypothesis:
   - The null hypothesis states that there is no effect or no relationship. In regression, this is usually stated as the coefficient is equal to zero, i.e., H0: βj=0, where βj is the coefficient for the *j*th independent variable.
   - The alternative hypothesis, i.e., HA: βj≠0, suggesting that there is an effect and the coefficient is not zero.

2. Calculate the test statistic

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

- $\hat{\beta}_j$ is the estimated coefficient,
- $SE(\hat{\beta}_j)$ is the standard error of the estimated coefficient.



5. Report the findings

Qualitative Predictors

So far, we have assumed that all variables are quantitative. But in practice, often some predictors are **qualitative**.

Example: The Credit data set contains information about balance, age, cards, education, income, limit , and rating for a number of potential customers.

Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
14.890	3606	283	2	34	11	Male	No	Yes	Caucasian	333
106.02	6645	483	3	82	15	Female	Yes	Yes	Asian	903
104.59	7075	514	4	71	11	Male	No	No	Asian	580
148.92	9504	681	3	36	11	Female	No	No	Asian	964
55.882	4897	357	2	68	16	Male	No	Yes	Caucasian	331

Qualitative Predictors

If the predictor takes only two values, then we create an **indicator** or **dummy variable** that takes on two possible numerical values.

For example for the gender, we create a new variable:

$$x_i = \begin{cases} 1 & \text{if } i\,\text{th person is female} \\ 0 & \text{if } i\,\text{th person is male} \end{cases}$$

We then use this variable as a predictor in the regression equation.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\,\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\,\text{th person is male} \end{cases}$$

Qualitative Predictors

Question: What is interpretation of β_0 and β_1?

Qualitative Predictors

Question: What is interpretation of β_0 and β_1?

- β_0 is the average credit card balance among males,
- $\beta_0 + \beta_1$ is the average credit card balance among females,
- and β_1 the average difference in credit card balance between females and males.

More than two levels: One hot encoding

Often, the qualitative predictor takes more than two values

In this situation, a single dummy variable cannot represent all possible values.

We create additional dummy variable as:

$$x_{i,1} = \begin{cases} 1 & \text{if } i\,\text{th person is Asian} \\ 0 & \text{if } i\,\text{th person is not Asian} \end{cases}$$

$$x_{i,2} = \begin{cases} 1 & \text{if } i\,\text{th person is Caucasian} \\ 0 & \text{if } i\,\text{th person is not Caucasian} \end{cases}$$

More than two levels: One hot encoding

We then use these variables as predictors, the regression equation becomes:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\,\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\,\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\,\text{th person is AfricanAmerican} \end{cases}$$

Beyond linearity

In the examples, we assumed that the effect on predictors is independent of each other.

Synergy effect or **interaction effect** can occur.

For example, if you have two independent variables *X*1 and *X*2, and you suspect that the effect of *X*1 on the dependent variable *Y* depends on the level of *X*2, you will include an interaction term X1×*X*2 in your model:

We change

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

To

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \boxed{\beta_3 X_1 X_2} + \epsilon$$

Captures the interaction effect between X_1 and X_2

Beyond linearity

Testing for Interaction: Just like testing other coefficients, you would use a t-test to determine if the interaction coefficient β_3 is significantly different from zero.

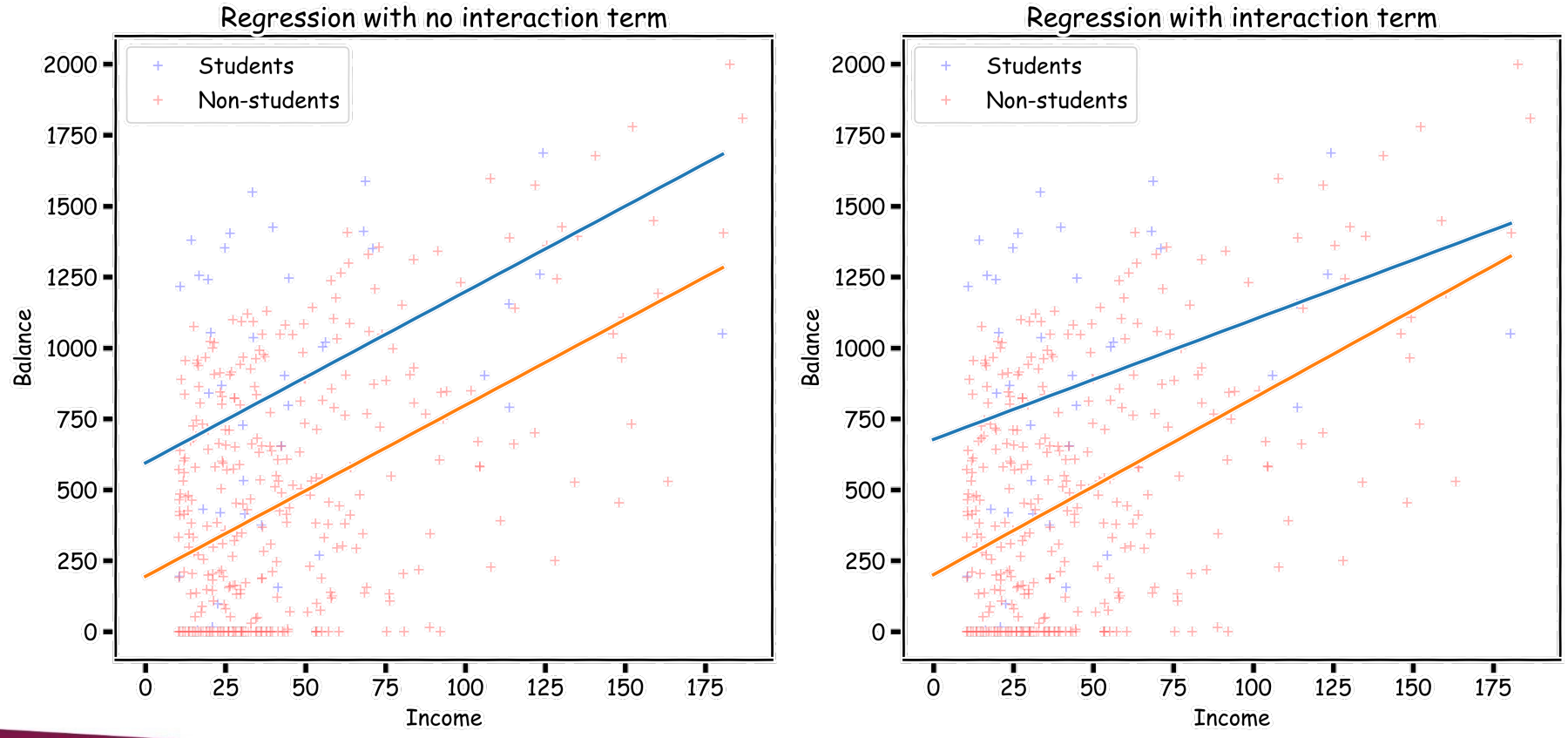
If it is, you have evidence of an interaction effect.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \boxed{\beta_3 X_1 X_2} + \epsilon$$

Captures the interaction effect between X_1 and X_2

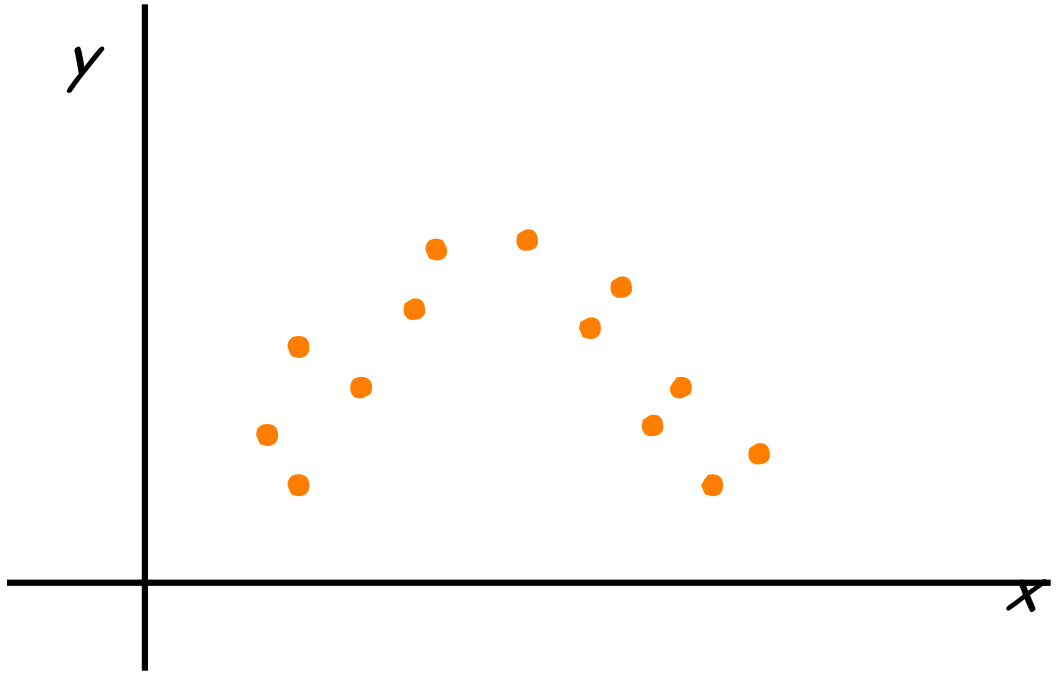
Beyond linearity

Plotting Interactions: To visualize interaction effects, interaction plots are often used where the slopes of lines representing the **relationship between one predictor and the outcome differ at various levels** of another predictor.



Polynomial Regression

Suppose a scatterplot has the following appearance:



Polynomial Regression

The n^{th} degree polynomial regression model is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_n x^n + \epsilon$$

This is the population regression function (mean y value for fixed values of the predictors).

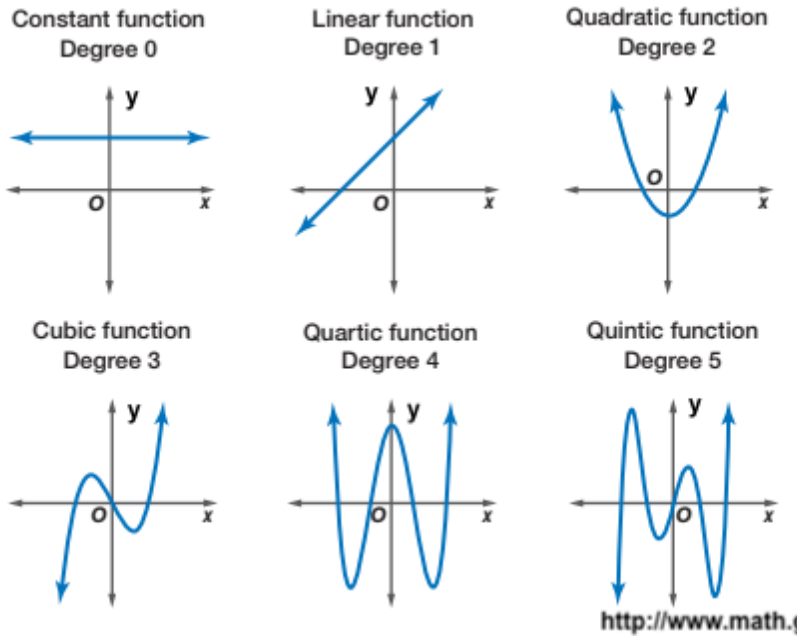
is a special case of the general multiple regression model with

$x_1 = x,\ x_2 = x^2,\ x_3 = x^3,\ \ldots,\ x_n = x^n$

Degree of the Polynomial: The degree *n* determines the flexibility of the curve.

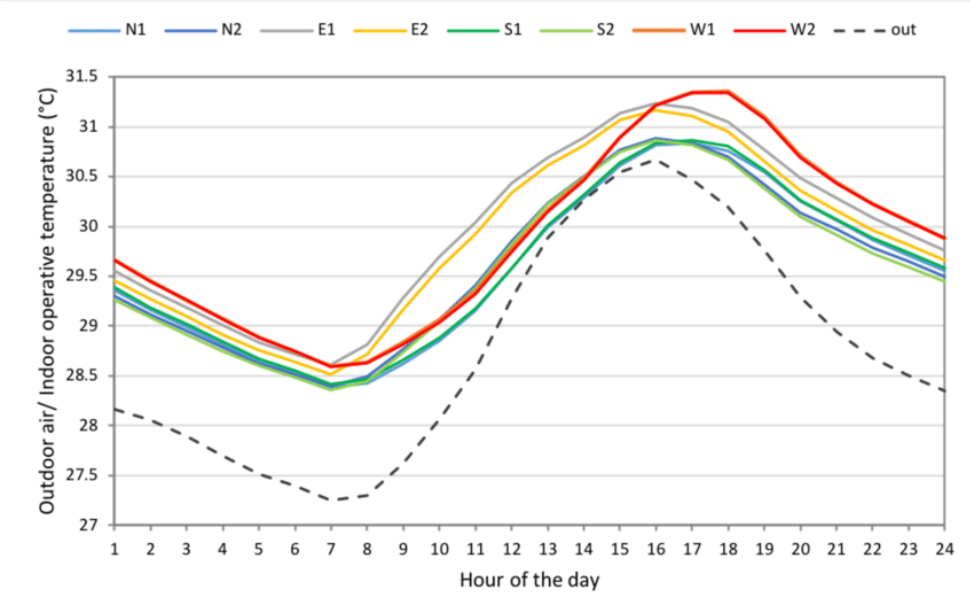
For instance:

- A degree of 1 models a straight line (simple linear regression).
- A degree of 2 models a parabola, which can accommodate one bend.
- A degree of 3 models a cubic curve, which can accommodate two bends, and so on.



Polynomial Regression example

Suppose researchers are studying the relationship between the number of hours after sunrise and the temperature of a particular location.
The relationship is **not linear** since the temperature rise slows down as it approaches midday and then falls again in the evening. A **polynomial regression can model this curved relationship**



$$T(h) = \beta_0 + \beta_1 h + \beta_2 h^2 + \beta_3 h^3 + \epsilon$$

- $T(h)$ is the predicted temperature at hour h after sunrise.
- β_0 is the intercept term; it represents the starting temperature at sunrise.
- β_1, β_2, and β_3 are the coefficients that the model will estimate; they capture the effect of the passing hours on the temperature change.
- h is the number of hours after sunrise, the independent variable.
- h^2 is the squared term of hours, representing the accelerating or decelerating change in temperature.
- h^3 is the cubic term of hours, capturing the curvature in the relationship (for example, the rise and fall pattern of temperature throughout the day).
- ϵ is the error term that accounts for randomness and noise in the data.