CHAPTER 17.4 & 22

Simple Linear Regression

We have talked about Descriptive Statistics:

"What does my data look like" (Histograms, box and whisker plots, …)

We have talked about Inferential Statistics:

"This is what I conclude from my sample with $100(1-\alpha)\%$ confidence."

Now, we talk about Predictive Statistics:

"This is what will likely happen tomorrow."

Predictive anlaysis will allow us to answer questions like:

"What will the price of gold be in 6 months?"

"how much additional sales income do I get for each additional $1000 spent on marketing?"

"What is the strength of the relationship between dose and effect ?"

"What is the strength of the relationship between age and income ?"

Think of an input causing an output.
If we were to add 10% more salt, then the horse would be 6% more thirsty.
If we add 5 minutes to the assessment time, then the scores will increase by 7 points.

Three major uses for predictive analysis (aka regression analysis) are:

(1) determining the strength of predictors
(2) forecasting an effect
(3) trend forecasting.

Prediction also requires calling upon our old algebra skills

What is the equation of a line that goes thru the points $(2,1)$ and $(4,3)$?

$$y - y_1 = m(x - x_1)$$

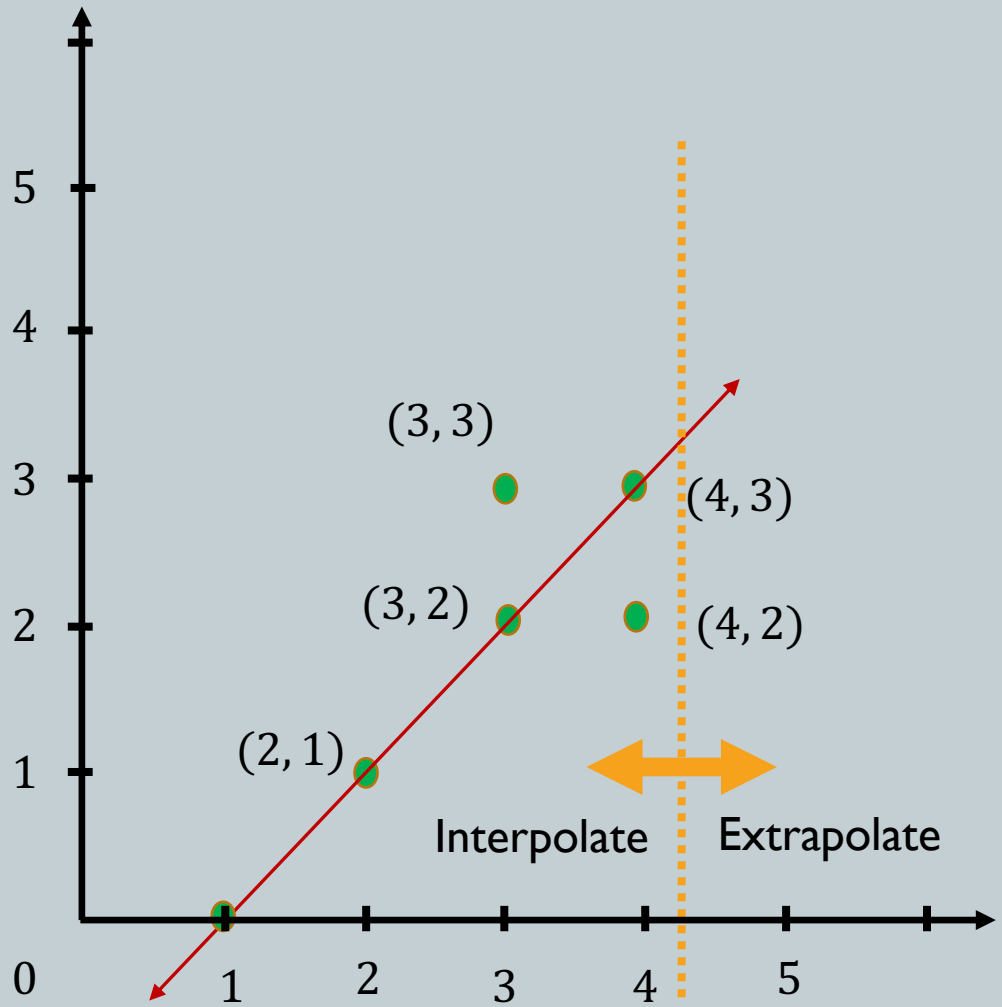
$$m = \frac{rise}{run} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{3-1}{4-2} = 1$$

$$y - 1 = 1(x - 2)$$

$y = 1 \cdot x - 1$ or $y = mx + B$ or

$$f(x) = x - 1$$

"If we put 4 in, then will we get 3 out ?"
"What input is required to get out a 5 ?"



$y = x - 1$

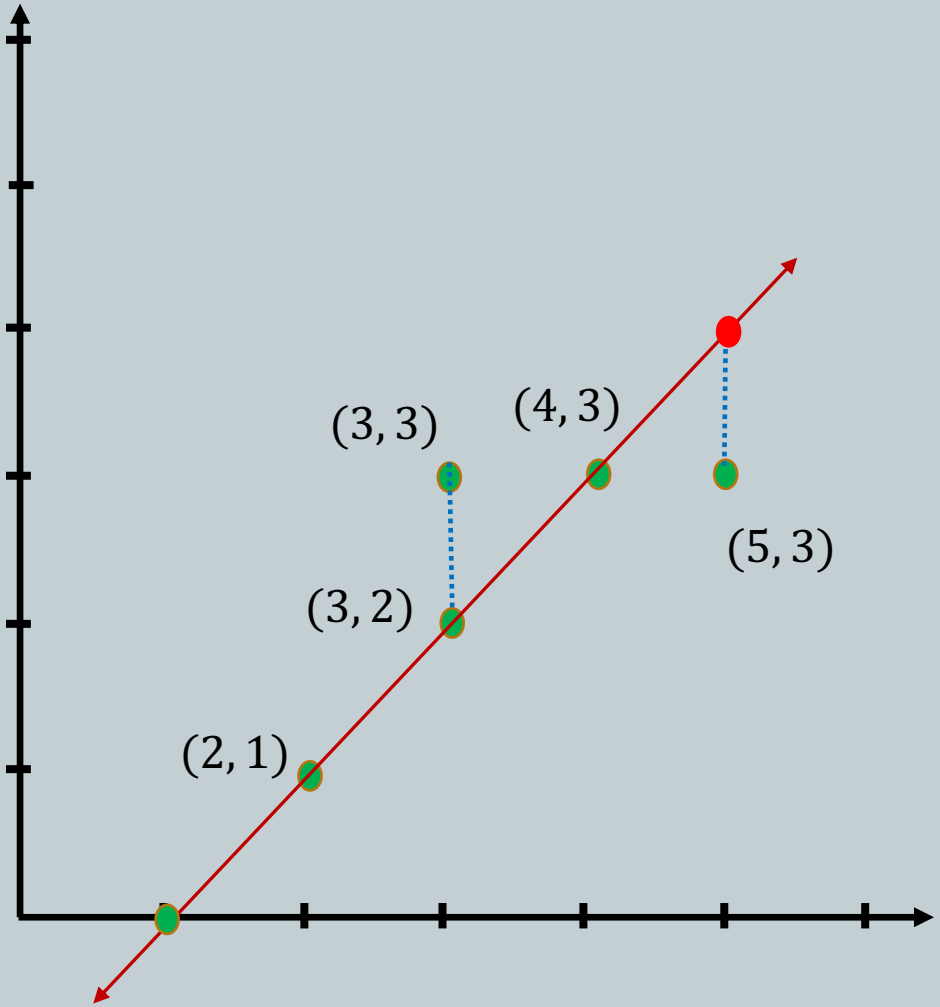
Is the point $(3,2)$ on this line?
Is the point $(3,3)$ on this line?
Is the point $(0,-1)$ on this line?

How far off is the point $(3,3)$?

How far off is the point $(5,3)$?

Are the distances positive or negative?

Sum of the Squared Estimate of Errors (SSE), or
Residual Sum of Squares (RSS), or
Sum of Squared Residuals (SSR)
is the sum of the squares of residuals, of deviations predicted from actual empirical values of data.



Why are we talking about linear equations in data analysis?

First, regression lines/graphs might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable.

Typical questions are, "What is the strength of relationship between…

…dose and effect,

…sales & marketing and spending,

…age and income.

Second, lines/graphs can be used to forecast effects or the impact of changes.
That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables. Slope indicates magnitude of change.

There are a number of types of linear regression.

Simple linear regression

1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

Multiple linear regression

1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

Logistic regression

1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

Ordinal regression

1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

Multinomial regression

1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

Discriminant analysis

1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

When selecting the model for the analysis, an important consideration is model fitting. Adding independent variables to a linear regression model will always increase the explained variance of the model (typically expressed as R^2).

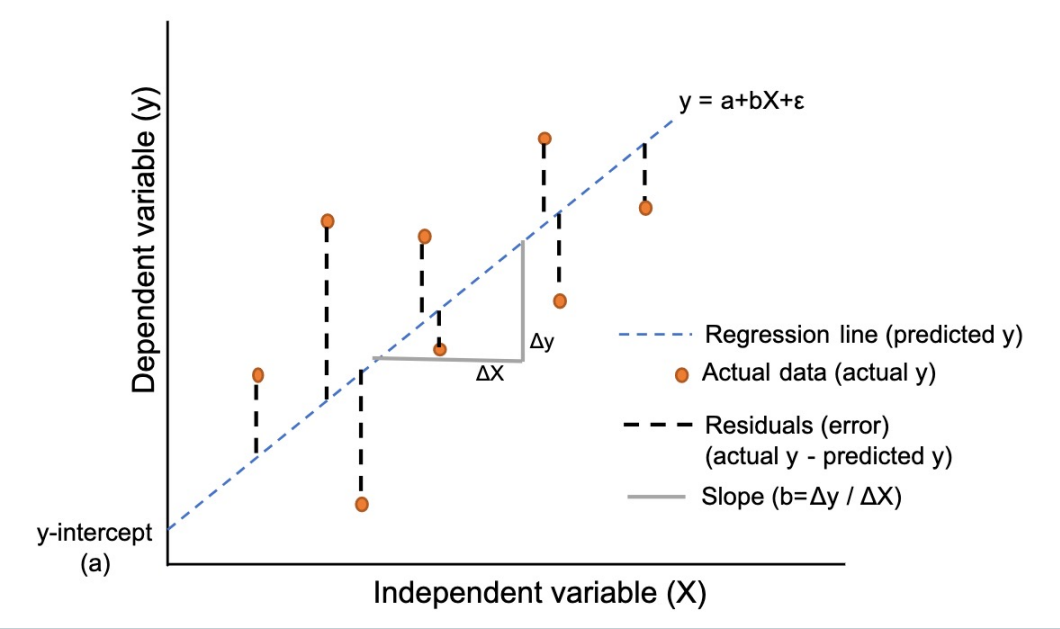
However, overfitting can occur by adding too many variables to the model, which reduces model generalizability.

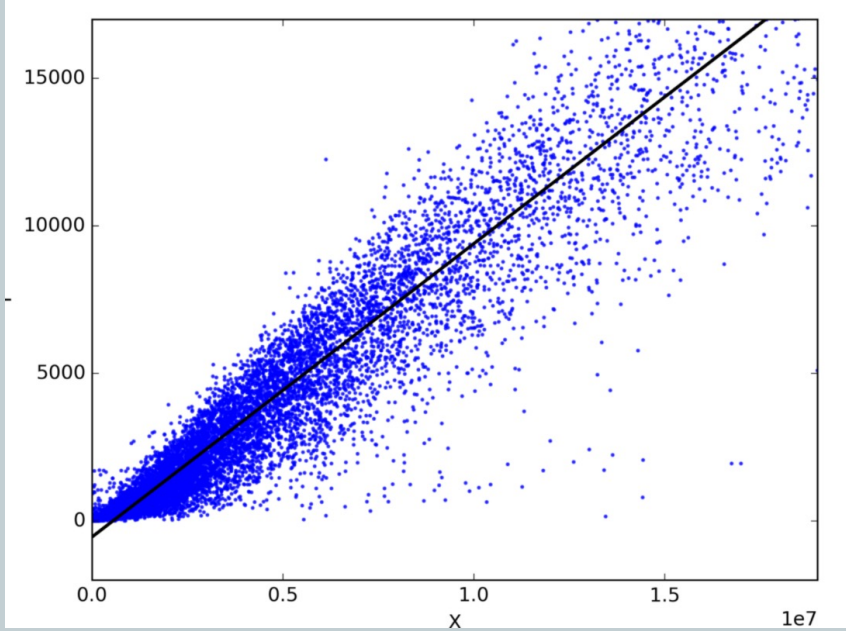
Occam's razor describes the problem extremely well – a simple model is usually preferable to a more complex model. Statistically, if a model includes a large number of variables, some of the variables will be statistically significant due to chance alone.

In statistics linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.





Linear regression is a basic and commonly used type of predictive analysis.
The overall idea of regression is to examine two things:

(1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

(2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

Grades: study time, HS GPA, years at CU, own textbook, number of extra curriculars
Effect of drug: gender, weight, metabolic rate, time of last meal

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

The simplest form of the regression equation with one dependent and one independent variable is defined by the formula:

$$\hat{y} = b_0 + b_1 \cdot x + \epsilon$$

$\hat{y}$ = estimated dependent variable score,
b_0 = constant,
b_1 = regression coefficient, and
x = score on the independent variable.
ϵ = error

Naming the Variables.

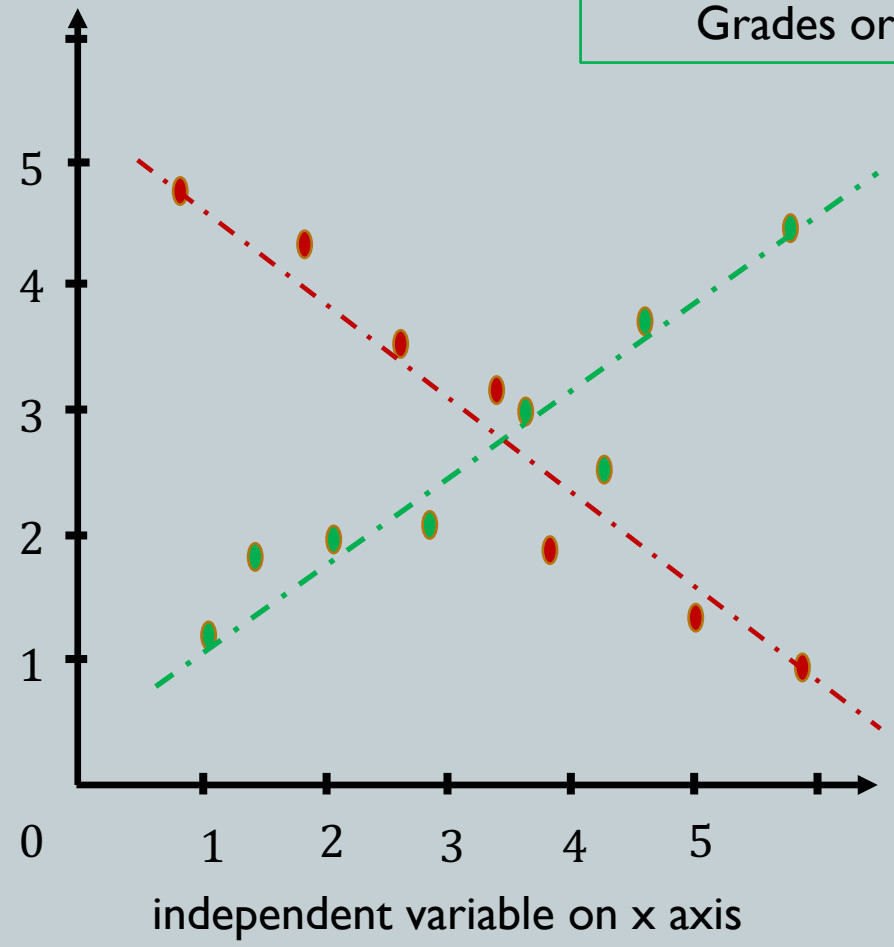
Dependent variable, y, outcome variable, criterion variable, endogenous variable, regressand

Independent variable, x, exogenous variables, predictor variable, regressor

Positive Relationship
Independent variable:
study time.
Dependent:
Grades or GPA

Negative Relationship
Independent variable:
Time on FaceBook (Meta)
Dependent:
Grades or GPA

dependent variable on y axis



We control/manipulate the dependent variable.

We want to minimize the error between the estimated value and the actual value.

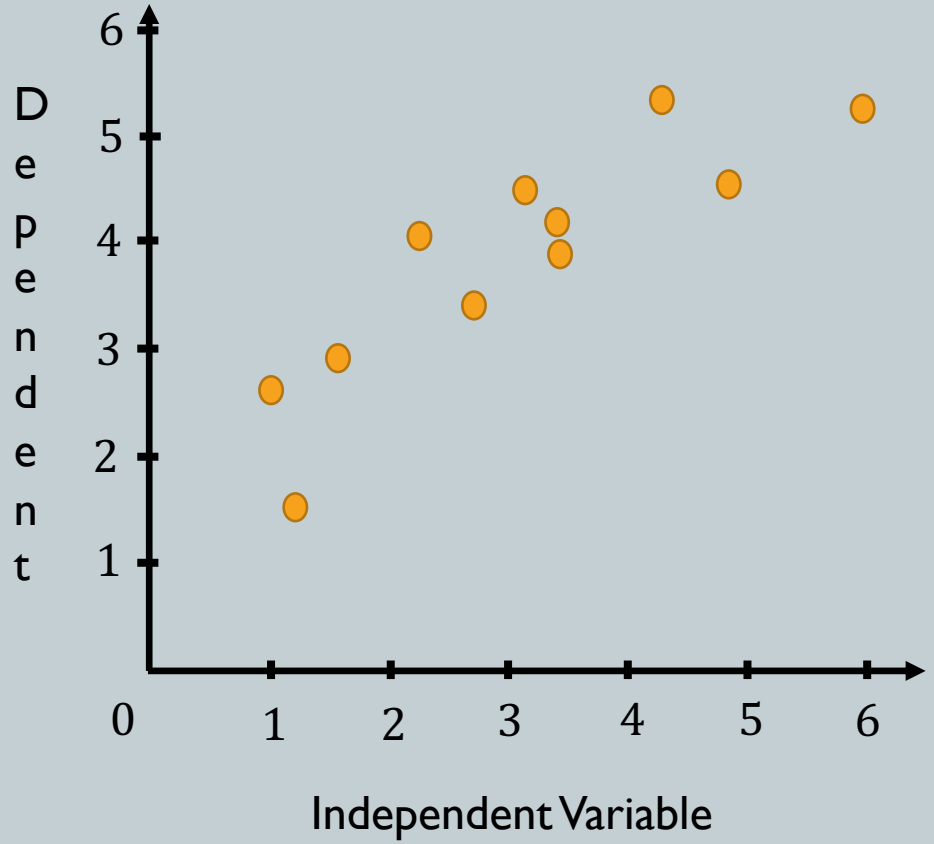
which line best fits the observations?

The best fit line is called the regression line.

$$\hat{y} = b_0 + b_1 \cdot x + \epsilon$$

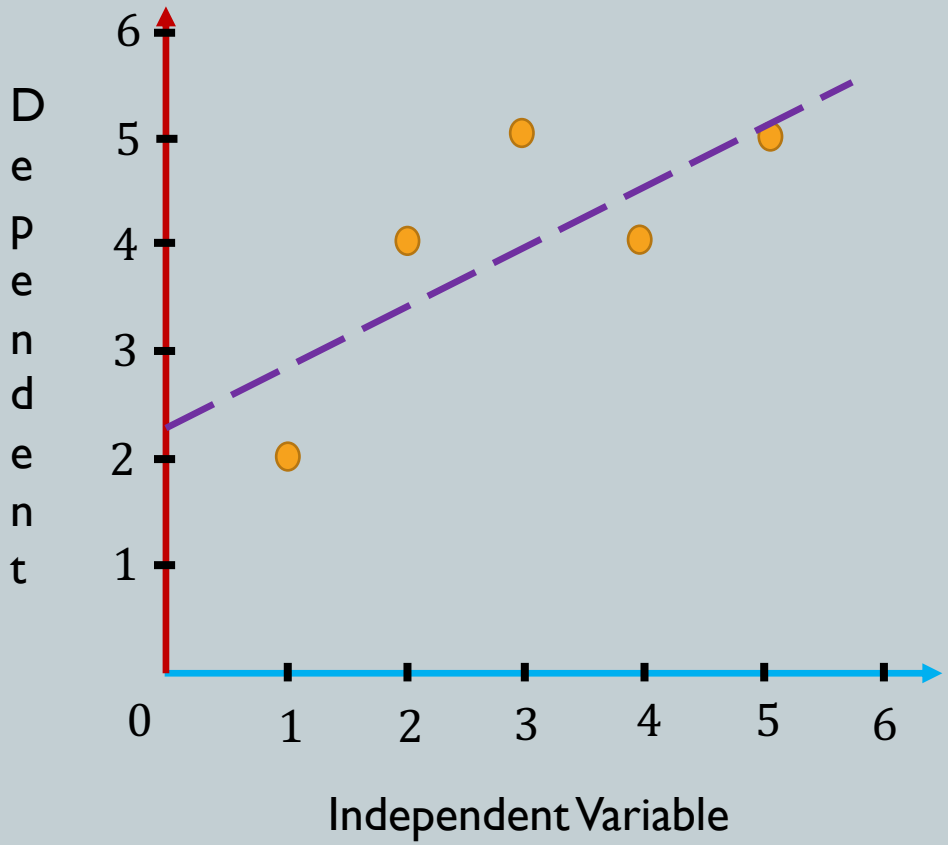
Which line is the best fit for estimation?
There seems to be a general positive trend..

The decision as to which line to use depends on minimizing error.



Q: How do we come up with the regression line?

A: The regression line is created with the least squares method.



Q: What is the least squares method ?

A: $y = \alpha + \beta x + \epsilon$

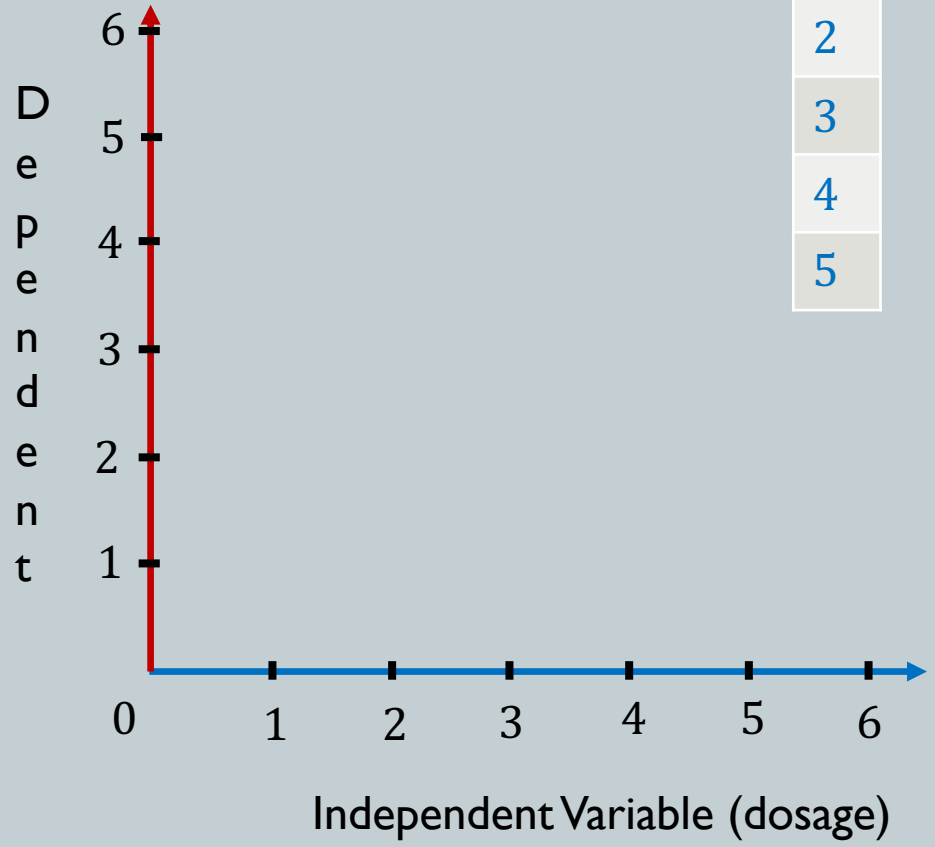
$$\alpha = \bar{y} - \beta \bar{x}$$

$$\beta = \frac{\bar{x} \cdot \bar{y} - \overline{x \cdot y}}{(\bar{x})^2 - \overline{x^2}}$$

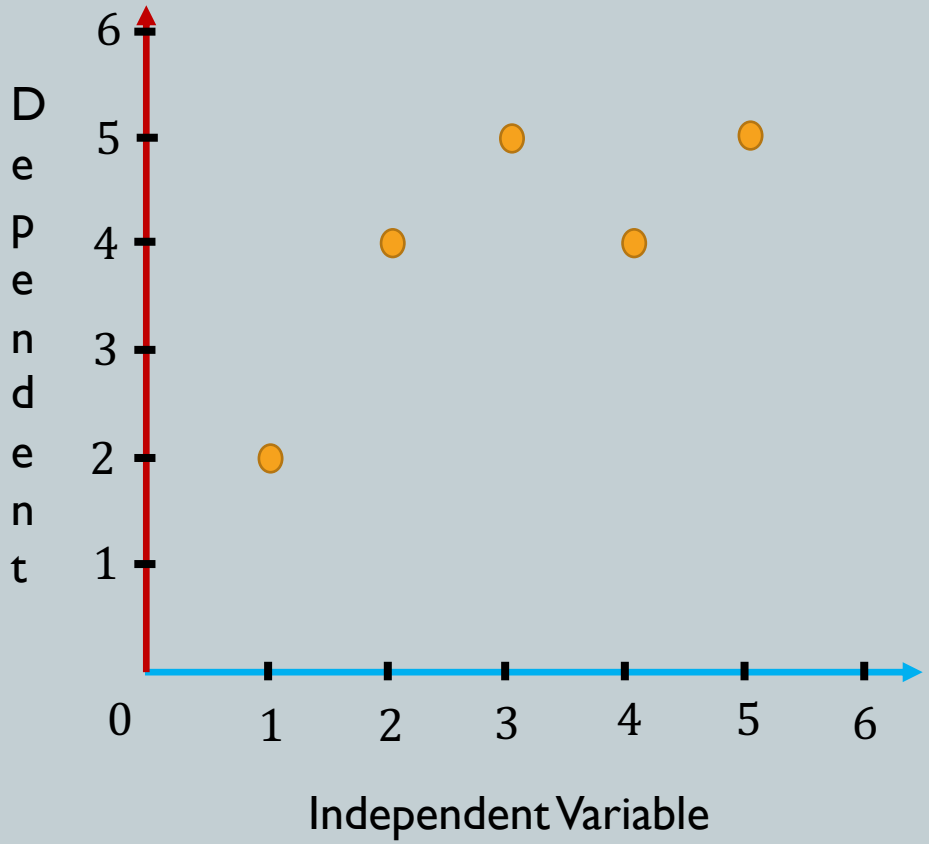
Suppose we test length of reaction (dependent variable, y) after taking a specified dose of a drug (independent variable, x).

So, we try 5 different doses, x.

x
1
2
3
4
5



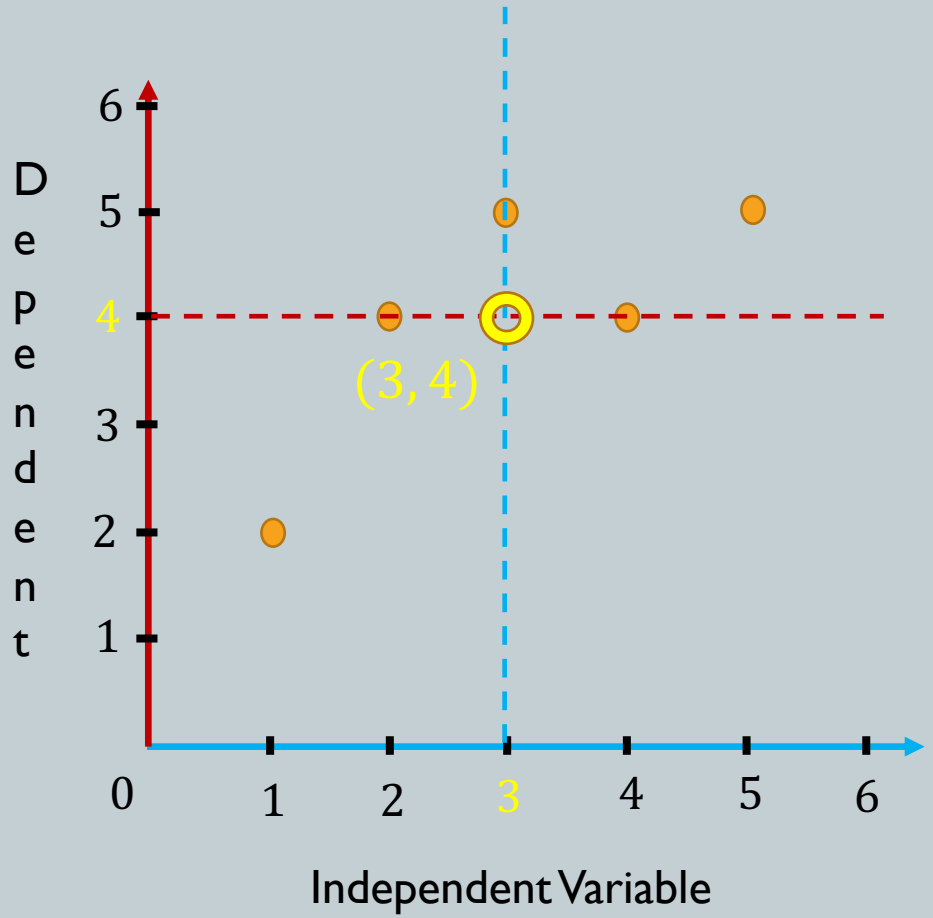
We then record the reaction times and plot the x's and y's in an attempt to visualize a relationship. Is the relationship positive or negative ?



x	y
1	2
2	4
3	5
4	4
5	5

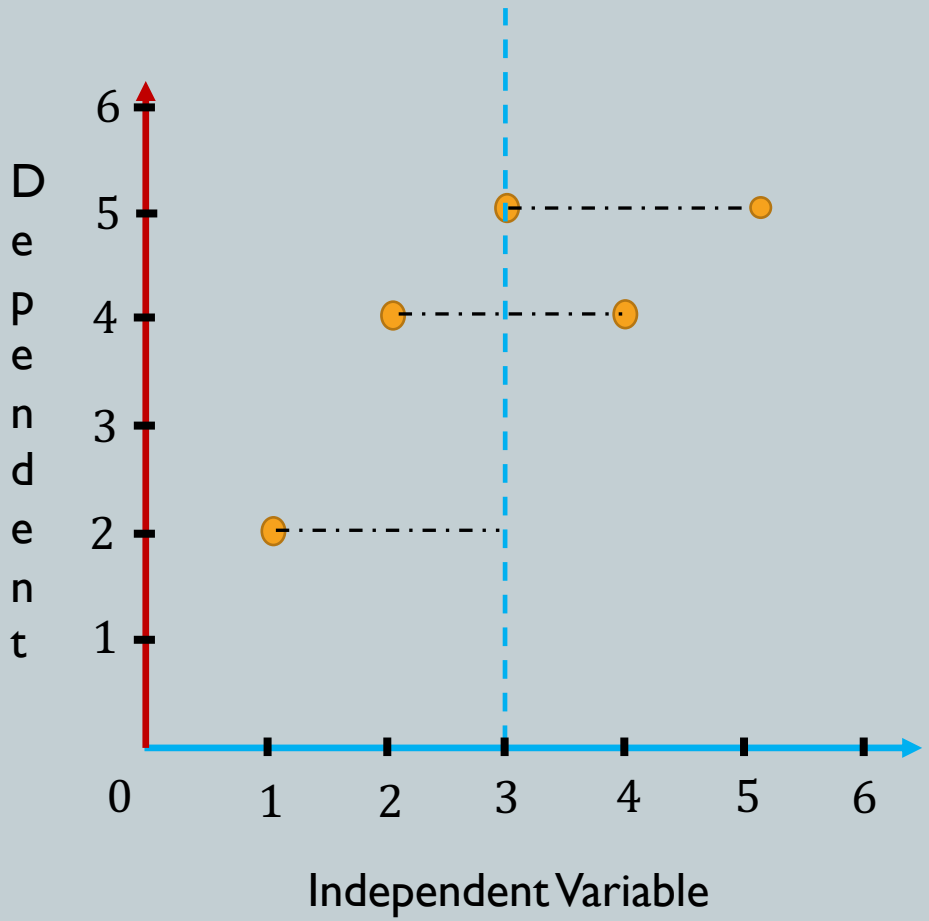
Now, how do we start to create the regression line for this data?

Mathematically, the regression line will go thru the intersection point of the means.



x	y
1	2
2	4
3	5
4	4
5	5
Mean	**3**	**4**

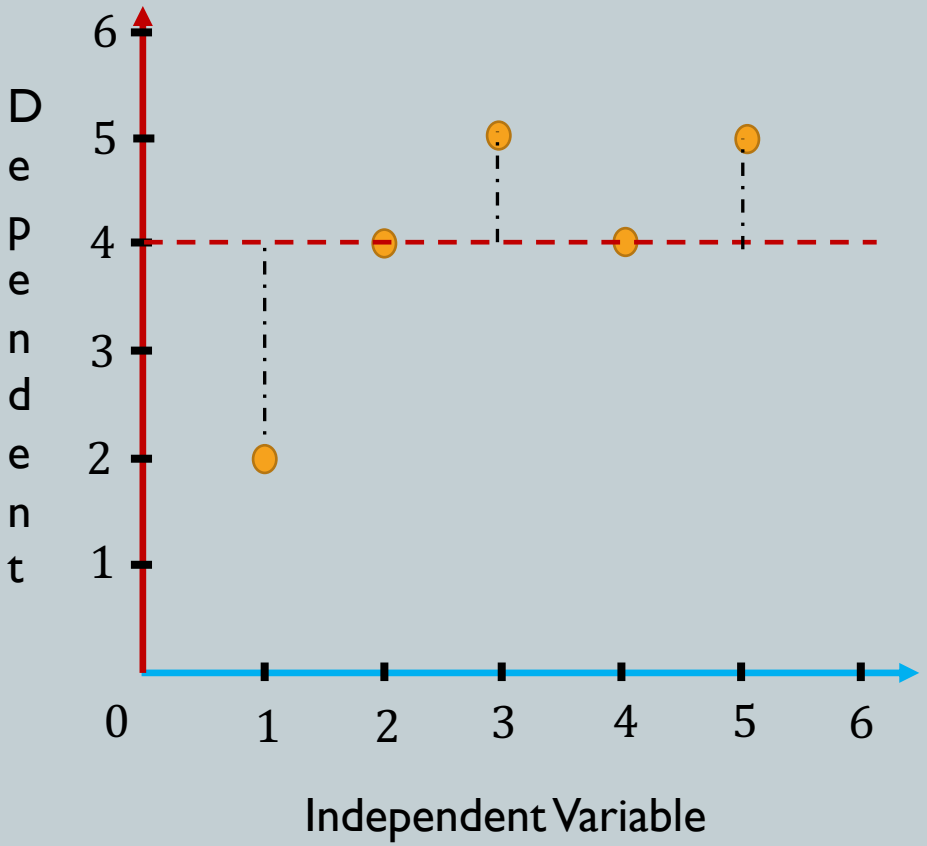
Start by measuring the x-distance of each data point from the mean of the $x's$



x	y	$x - \bar{x}$
1	2	$1 - 3 = -2$
2	4	$2 - 3 = -1$
3	5	$3 - 3 = 0$
4	4	$4 - 3 = 1$
5	5	$5 - 3 = 2$

Mean	**3**	**4**

Then measure the y-distance of each data point from the mean of the $y's$.

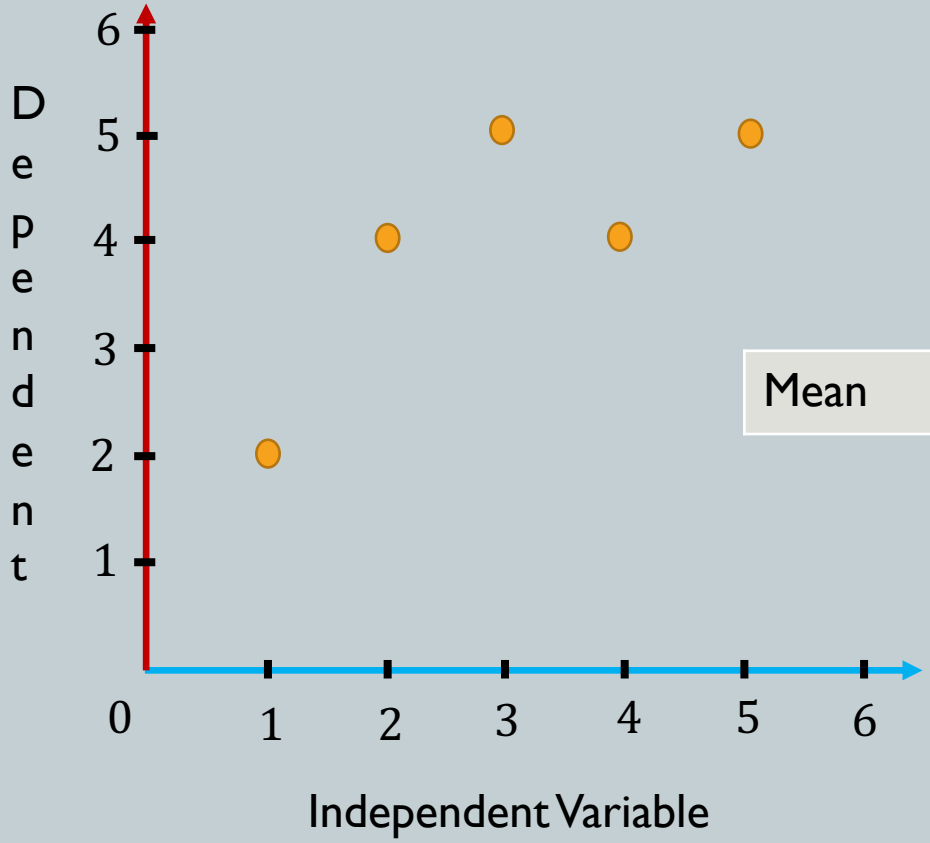


x	y	$x - \bar{x}$	$y - \bar{y}$
1	2	$1 - 3 = -2$	$2 - 4 = -2$
2	4	$2 - 3 = -1$	$4 - 4 = 0$
3	5	$3 - 3 = 0$	$5 - 4 = 1$
4	4	$4 - 3 = 1$	$4 - 4 = 0$
5	5	$5 - 3 = 2$	$5 - 4 = 1$

Mean	**3**	**4**

We want the equation of the regression line $\hat{y} = b_0 + b_1 \cdot x$

$$b_1 = \frac{\sum(x-\bar{x})\cdot(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{6}{10} = 0.6$$



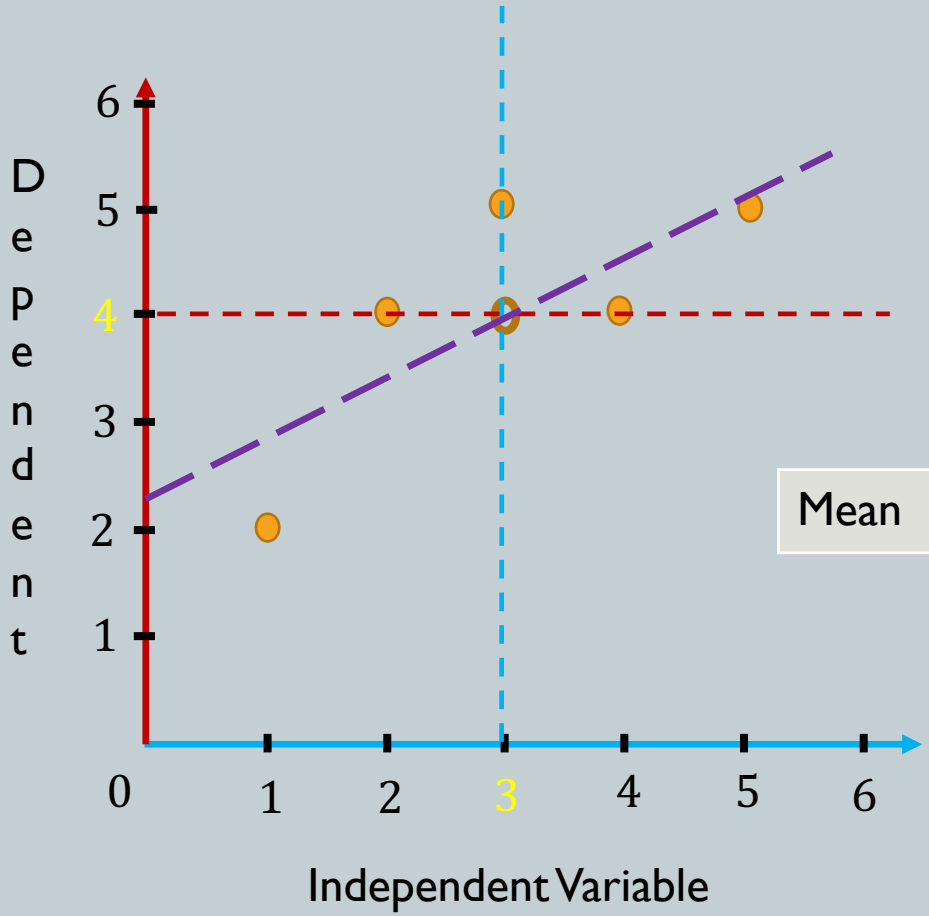
x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})\cdot(y-\bar{y})$
1	2	$1-3=-2$	$2-4=-2$	4	4
2	4	$2-3=-1$	$4-4=0$	1	0
3	5	$3-3=0$	$5-4=1$	0	0
4	4	$4-3=1$	$4-4=0$	1	0
5	5	$5-3=2$	$5-4=1$	4	2
Mean	**3**	**4**		Sum	**10**	**6**

Sum of the Squared Errors?

Now, we want the y-intercept (b_0) of the line $\hat{y} = b_0 + b_1 \cdot x$

We just calculated (slope): $b_1 = 0.6$

$\hat{y} = b_0 + b_1 \cdot x$ implies $4 = b_0 + 0.6 \cdot 3$ which means $b_0 = 2.2$



x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x}) \cdot (y - \bar{y})$
1	2	$1 - 3 = -2$	$2 - 4 = -2$	4	4
2	4	$2 - 3 = -1$	$4 - 4 = 0$	1	0
3	5	$3 - 3 = 0$	$5 - 4 = 1$	0	0
4	4	$4 - 3 = 1$	$4 - 4 = 0$	1	0
5	5	$5 - 3 = 2$	$5 - 4 = 1$	4	2
Mean	3	4			
			Sum	**10**	**6**

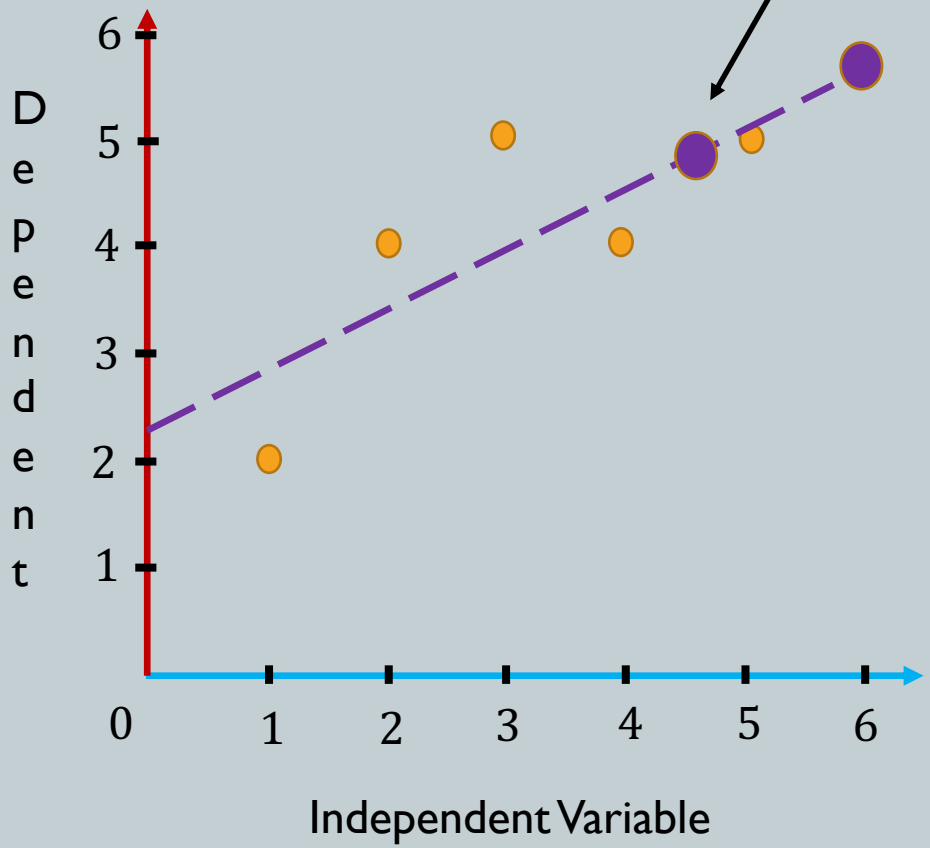
$$\hat{y} = 2.2 + 0.6x$$

So, theoretically, a drug dose of 4.5 would have a reaction that lasts how long?

$f(4.5) = 2.2 + 0.6(4.5) = \mathbf{4.9}$ minutes (interpolate)

$f(6) = 2.2 + 0.6(6) = \mathbf{5.8}$ minutes (extrapolate)

Because $\hat{y} = b_0 + b_1 \cdot x$ $\hat{y} = 2.2 + 0.6x$

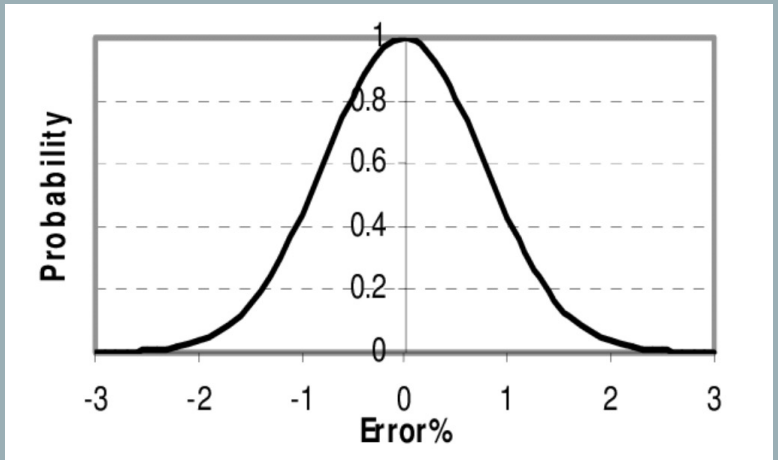


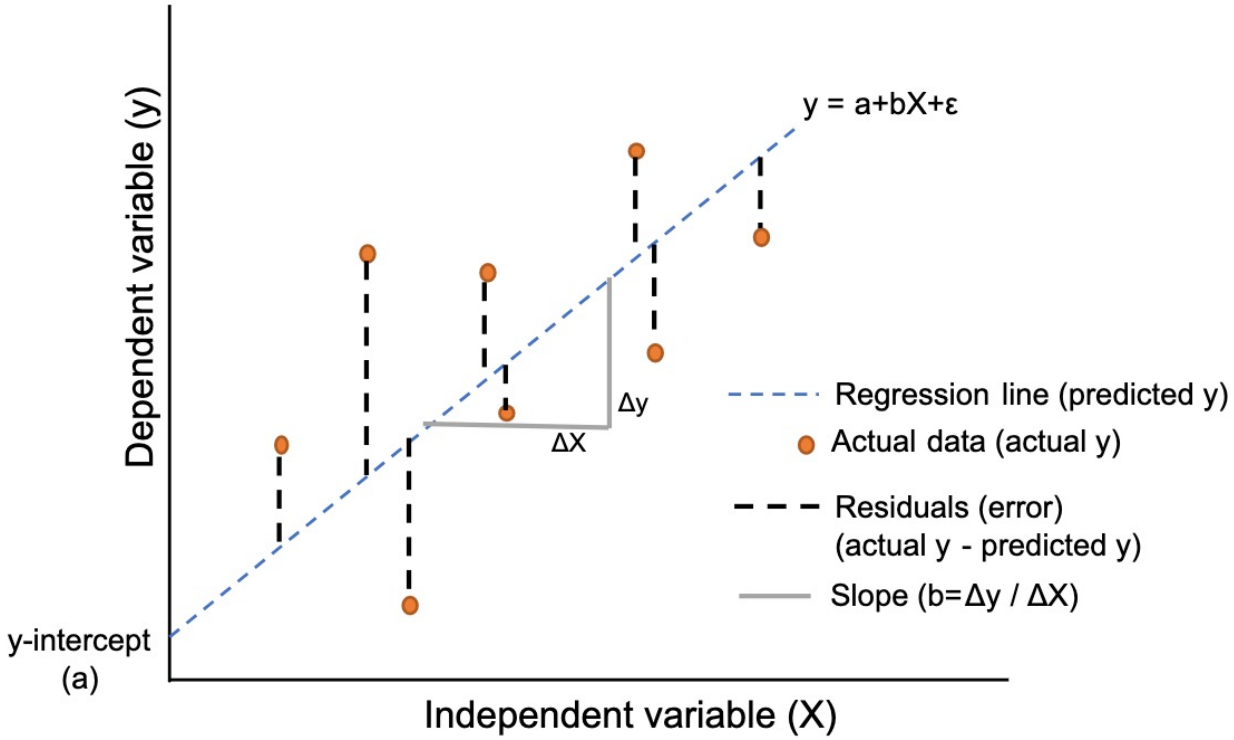
Why did that just work?

The goal is to find the equation of the line of best fit.

There are uncertainties, ϵ, around the model.

We are assuming $\epsilon \sim N(0, \sigma^2)$.





$$y = \alpha + \beta x + \epsilon$$

Y is a random variable.

$$E[Y] = E[\alpha + \beta x + \epsilon] = E[\alpha] + E[\beta x] + E[\epsilon] = \alpha + \beta x + 0 = \alpha + \beta x$$

The variance of our model σ^2 will be smallest if the difference between the estimate and each actual point is the smallest. Therefore, minimize **S**um of the **S**quared **E**rrors.

$$SSE = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

Minimize SSE with respect to α and β: $\frac{\delta SSE}{\delta \alpha} = 0$ and $\frac{\delta SSE}{\delta \beta} = 0$.

This produces: $y = \alpha + \beta x + \epsilon$ with $\alpha = \bar{y} - \beta \bar{x}$ $\beta = \frac{\bar{x} \cdot \bar{y} - \overline{x \cdot y}}{(\bar{x})^2 - \overline{x^2}}$

This is how Python calculates α and β, it is slightly different than our drug dose example.

Simple hand calculated example:

Consider the data: $(2,1),(3,3),(4,3)$

note:

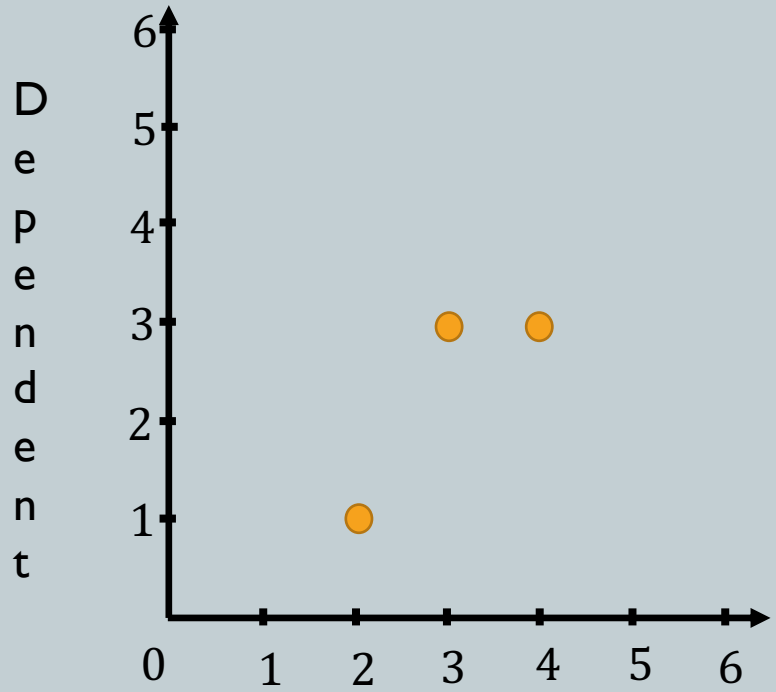
$\bar{x} = 3$ and $\bar{y} = \frac{7}{3}$

Therefore: $(\bar{x})^2 = 9$

$$\overline{xy} = \frac{2\cdot1+3\cdot3+4\cdot3}{3} = \frac{23}{3} \text{ and } \overline{x^2} = \frac{2^2+3^2+4^2}{3} = \frac{29}{3}$$

$$\alpha = \bar{y} - \beta\bar{x} \qquad \beta = \frac{\bar{x}\cdot\bar{y} - \overline{x\cdot y}}{(\bar{x})^2 - \overline{x^2}}$$

So, $\alpha = \frac{7}{3} - \beta \cdot 3$ and $\beta = \frac{3 \cdot \frac{7}{3} - \frac{23}{3}}{9 - \frac{29}{3}} = 1$



$$y = \alpha + \beta x$$

$$\hat{y} = -\frac{2}{3} + 1 \cdot x$$

Simple hand calculated example:

Consider the data: $(2,1),(3,3),(4,3)$

note:

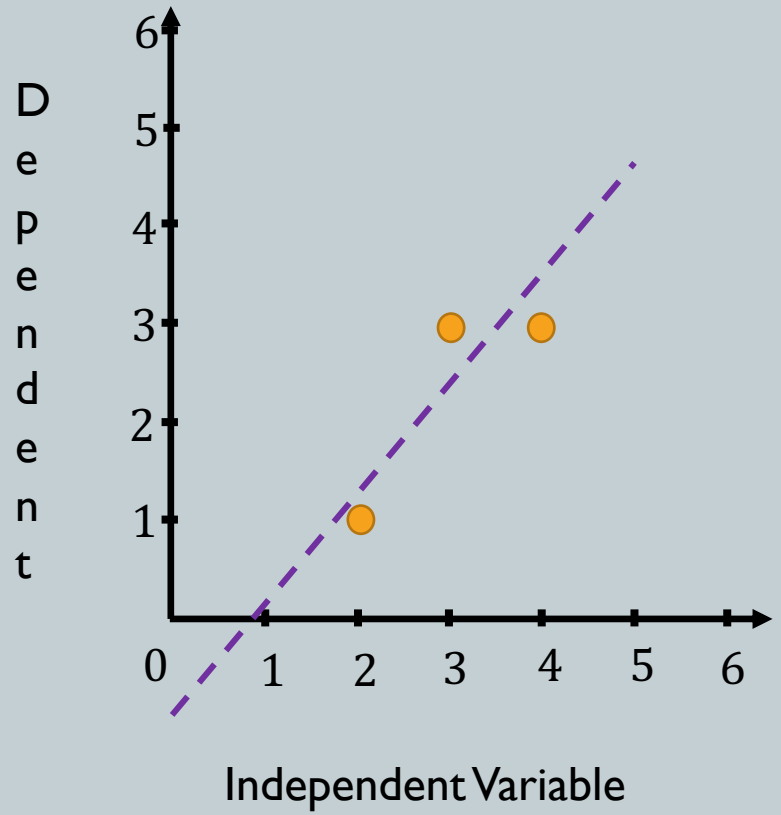
$\bar{x} = 3$ and $\bar{y} = \frac{7}{3}$

Therefore: $(\bar{x})^2 = 9$

$\overline{xy} = \frac{2\cdot1+3\cdot3+4\cdot3}{3} = \frac{23}{3}$ and $\overline{x^2} = \frac{2^2+3^2+4^2}{3} = \frac{29}{3}$

$\alpha = \bar{y} - \beta\bar{x}$ $\qquad$ $\beta = \frac{\bar{x}\cdot\bar{y} - \overline{x\cdot y}}{(\bar{x})^2 - \overline{x^2}}$

So, $\alpha = \frac{7}{3} - \beta \cdot 3$ and $\beta = \frac{3\cdot\frac{7}{3} - \frac{23}{3}}{9 - \frac{29}{3}} = 1$



$$\hat{y} = -\frac{2}{3} + 1 \cdot x$$

Linear Regression Assumptions:

The relationship between the x and y variables should be linear

Errors (residuals) should be independent of each other

Errors (residuals) should be normally distributed with a mean of 0.

Errors (residuals) should have equal variance (Homoscedasticity)

Next time: Inference in Simple Linear Regression

k