Linear regression

Chrysafis Vogiatzis

Lecture 30-31

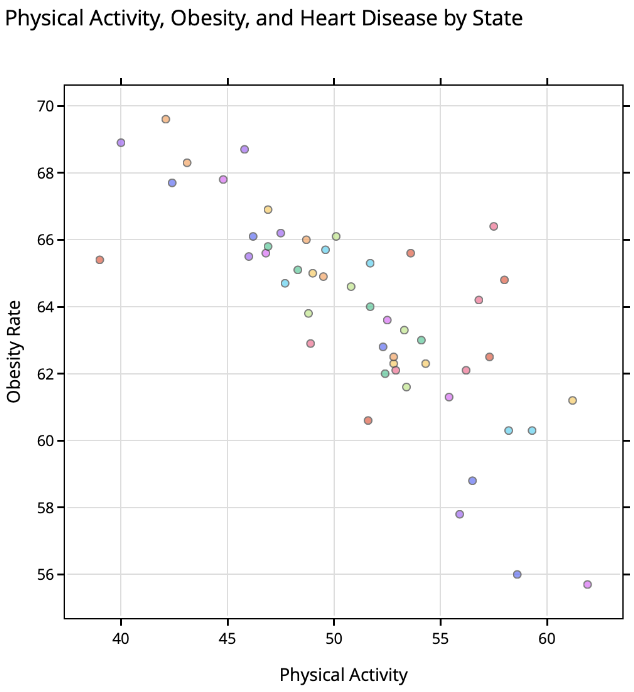
Learning objectives

After lectures 30-31, we will be able to:

- Explain the difference between regression and classification.
- Describe regression and linear regression.
- Derive, use, and interpret the results of the least squares line.
- Check whether a simple linear regression is significant or not.

Motivation: Physical activity and obesity

See below a figure representing the different levels of physical activity in each of the 50 states (x axis) and the resulting obesity rates (y axis). Do we see a relationship between activity and obesity? Is it linear? And, very importantly, is it significant?



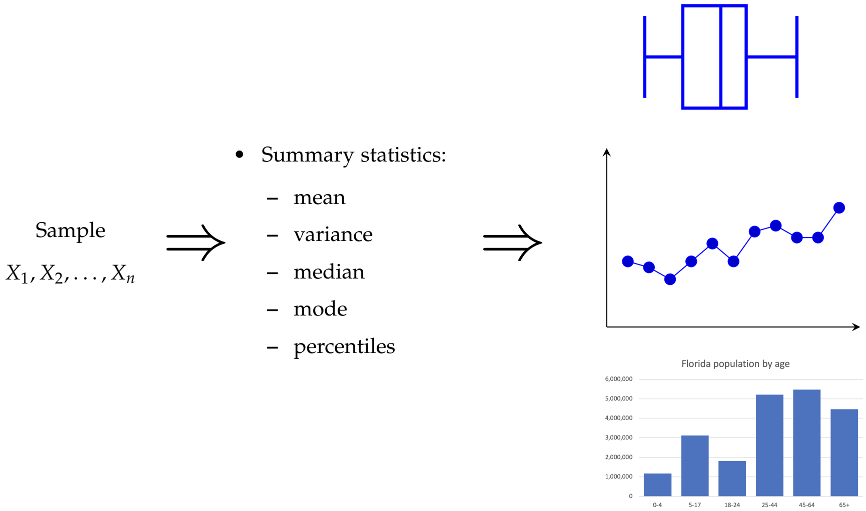
Motivation: Education level and income

Is there a relationship between the annual income of a person and their education level? And, if so, can we predict the income of a person before and after they have obtained a Master's degree?

Model building

In the second part of the class, we saw **descriptive statistics**. Visually, see Figure 1, where from a sample we obtain a series of descriptive information (referred to as summary statistics), that we then present in a pictorial form (for ease of use and understanding).

Figure 1: A visual representation of descriptive statistics.



Sample

$X_1, X_2, \dots, X_n$ $\Longrightarrow$

- Summary statistics:
  - mean
  - variance
  - median
  - mode
  - percentiles

$\Longrightarrow$

Then, in the third part of the class, we moved towards **inferential statistics**. Again, we represent this part visually in Figure 2.

Figure 2: A visual representation of inferential statistics.

From sample: Infer To a population:

$X_1, X_2, \dots, X_n$ $\Longrightarrow$ μ, σ, p

Point estimation:

$\hat{\Theta}$

- bias.
- variance.
- MSE.

Interval estimation:

Confidence interval

- unknown mean, var., proportion.
- confidence level $1 - \alpha$.

Hypothesis test:

H_0 vs. H_1

- one or two populations.
- mean, var, proportion.
- α, β, P-values.

There are three classifications of modern statistical methods:

1. **Descriptive statistics**: techniques to describe and visualize data.

2. **Inferential statistics**: techniques to draw conclusions for a large, unknown population based on observations of a smaller group (sample).

3. **Model building**: techniques to find relationships between data points, measure how strong these relationships are, and build models that can make predictions about the future.

In this last part of the class, we will focus on **model building**. Model building has three goals then:

- Goal #1: **investigate whether a relationship exists** between variables of our model.

  Does a relationship exist?

  - Do students perform better in tests that are in the morning or in the evening? Does time of day affect performance?
  - Does cold weather increase the number of accidents? Does the temperature affect driving patterns? Or do weather conditions, regardless of temperature, affect driving?
  - Does physical activity affect obesity rates? Does income affect obesity rates?

- Goal #2: **measure how strong the relationship is**.

  Strong relationship?

  - Obesity rates have been shown to depend on physical activity, income, age, education, built environment, etc.
  - Physical activity and age have been found to be more important.

- Goal #3: **predict an outcome given data**.

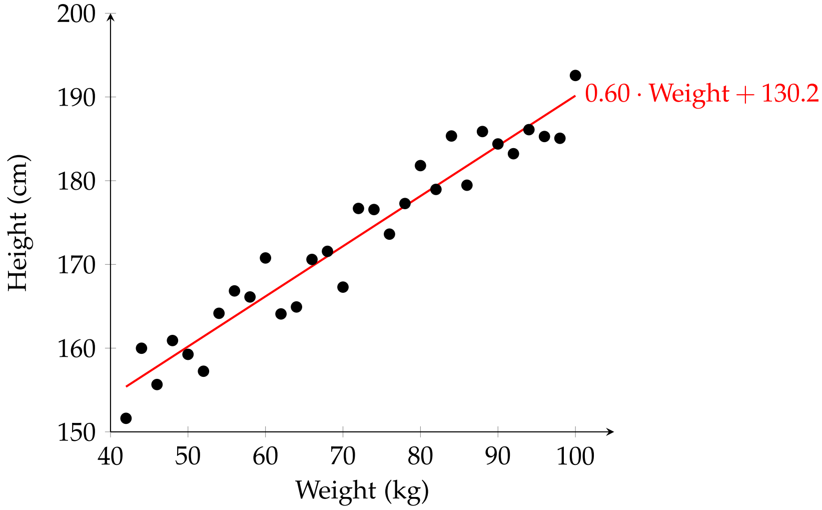
  Predicting

  - Given physical activity levels, predict the obesity rates for a specific state.
  - Given the weather conditions, predict the number of accidents at an interstate.

We define two types of models: **regression** and **classification**. They are visually contrasted in Figures 3 and 4. In the remainder of the semester, we will focus solely on regression.

1. **Regression**: for given values of *independent variables* x_i, predict the value of *dependent variable* y. Typically, regression applies to **continuous** y variables.

Figure 3: A possible regression line helping us predict the height of someone given their weight.



2. **Classification**: for given values of *independent variables* x_i, predict the class where *dependent variable* y belongs to. Typically, classification applies to **discrete** y variables.

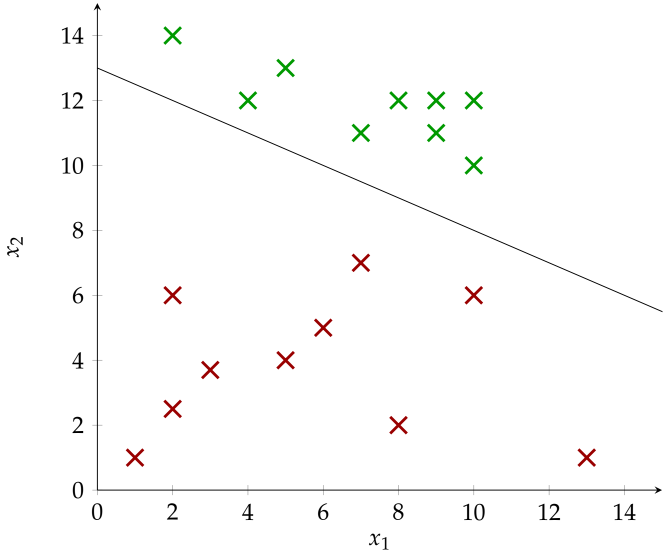
In the remainder of today's lecture, we shall focus on regression models.

Linear regression

Before we begin with linear regression, a really quick overview of some necessary notation. We will assume the existence of two types of variables:

1. independent variables x: these may also be called predictor variables or regressors.
2. dependent variables y: sometimes also referred as response variables, outcome variables, or regressands.

Figure 4: An example of a classification problem. The line here separates our observations in two classes (green and red).



Typically, independent variables are given to us in an attempt to predict the value of a dependent variable. Of course, this depends on the specifics of the problem we are tackling at each time!

Independent vs. dependent variables

- Does the duration of a call (y) depend on the reception signal (x)?
- Does income (y) depend on years of education (x)?
- Does obesity rate (y) depend on income (x_1), days of physical activity per week (x_2), and age (x_3)?

As we note with the earlier example, it is not necessary to only have one independent variable x! Formally, we define regression as follows:

Definition 1 (Regression) *Regression is a statistical technique that is used to model the relationships between the response variable (also called the* ***dependent*** *variable) y and the predictor variables (also called the* ***independent*** *variables) x.*

We may define multiple types of regression:

1. Simple linear regression: one independent and one dependent variables tied together through a linear relationship.
2. Multiple linear regression: multiple independent and one dependent variables tied together using a linear relationship.

3. Polynomial regression: one or more independent and one dependent variables tied together using a polynomial relationship.

4. Logistic regression: one or more independent and one dependent variable tied together using any relationship. However now, the dependent variable takes on two discrete values (true or false, healthy or unhealthy, etc.). This is also called a dichotomous regression.

Simple linear regression

In simple linear regression, we want to express the dependent variable y as a linear function of the independent variable x. In mathematical terms, we are looking for coefficients β_0, β_1 such that:

$$y = \beta_0 + \beta_1 x.$$

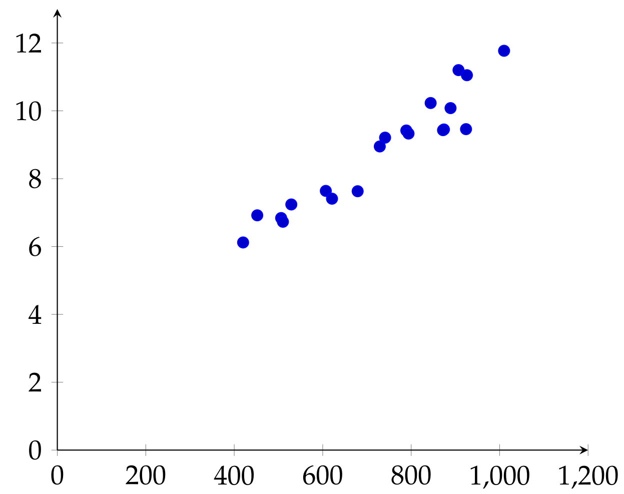
A webstore example

A webstore has collected the following data on the weekly visitors of the website and the profits from the past 20 weeks. They want to investigate that relationship and see whether they can direct more clicks towards their store. The data they have collected is as follows:

n	Visitors	Profit	n	Visitors	Profit
1	907	11.2	2	926	11.05
3	506	6.84	4	741	9.21
5	789	9.42	6	889	10.08
7	874	9.45	8	510	6.73
9	529	7.24	10	420	6.12
11	679	7.63	12	872	9.43
13	924	9.46	14	607	7.64
15	452	6.92	16	729	8.95
17	794	9.33	18	844	10.23
19	1010	11.77	20	621	7.41

What is the relationship between the profit and the number of visitors in their website?

Let's try this again, from a visual perspective...

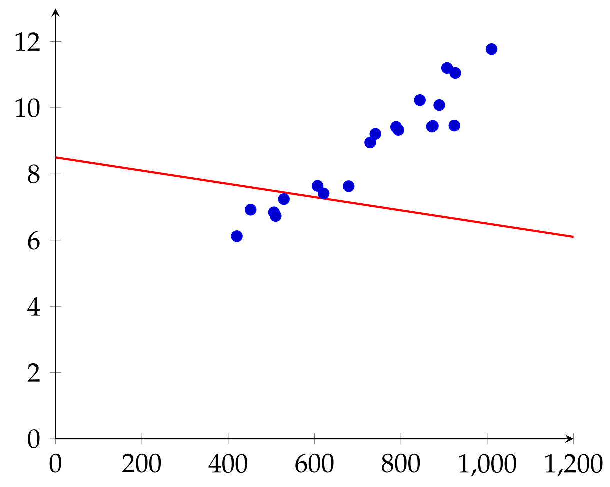


Some of the questions you may have already:

1. Do we see a relationship between profits and visits?
2. Does the relationship appear to be linear?
3. Does the relationship appear to be strong?
4. Can we predict profits based on the number of visitors?

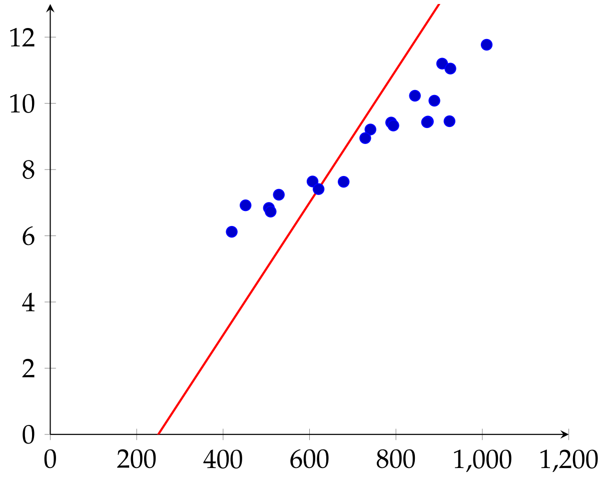
Our answers must have been Yes, Yes, Yes, and We sure hope so. Since there appears to be a linear relationship, what is the **best line** we can come up with to connect the dots? Let us try some and discuss why they work and why they do not work.

Line 1:



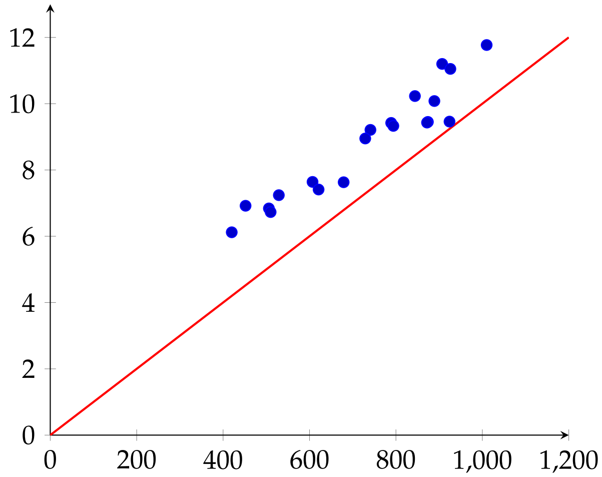
Bad line as it does not seem to capture the data provided.

Line 2:



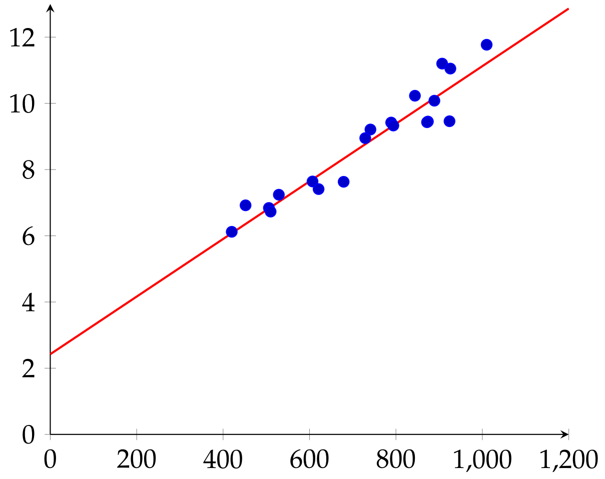
Better than before, but it still seems to **miss the "trend" of the data**, doesn't it?

Line 3:



This one seems to follow the trend, but is **underestimating the outcome** at each point...

Line 4:



The best fit line is the one that **minimizes the deviations of the data from the estimated regression line**.

Let's see what that means from a mathematical point of view. Based on our available data, we have n pairs of independent variables (x_i, y_i), for $i = 1, \ldots, n$. **If our line is correct**, then we should expect $y_i = \beta_0 + \beta_1 x_i$, no?

However, we recall that real life is not modeled exactly and neatly by a model, so maybe we can **incorporate some noise**? In that case, we now should get $y_i \;=\; \beta_0 + \beta_1 x_i + \epsilon_i$. In this last equation,$\beta_0$ is the **intercept**, β_1 is the **slope**; and ϵ_i is the noise related to data point (x_i, y_i).

In order for the quantity referred to as noise to make sense, we need to make some assumptions. Namely, we have for all noises ϵ_i that:

- they are independent normally distributed random variables;
- with zero mean;
- and with the same variance;
- $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$.

Let us consider the total "error". What could this mean? We could potentially define it as:

- the sum of all errors (positive or negative): $L = \sum\limits_{i=1}^{n} \epsilon_i$.
- the sum of all absolute errors: $L = \sum\limits_{i=1}^{n} \left|\epsilon_i\right|$.
- the sum of all squared errors: $L = \sum\limits_{i=1}^{n} \epsilon_i^2$.

The last one is called the **least squares** error. Recall that

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \implies \epsilon_i = y_i - \beta_0 - \beta_1 x_i.$$

Hence, we may derive for the least squares error:

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

A quadratic term! And one that we need to minimize in order to identify the least squares line. What are our unknowns? Those would be the slope and the intercept, β_0 and β_1. And what are our known parameters? Of course all the pairs (x_i, y_i) for all $i = 1, \ldots, n$ known data points.

Finally, how can we minimize $L = \sum\limits_{i=1}^{n} \epsilon_i^2 = \sum\limits_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$? We could take the derivative for each of the unknowns and equate to zero, leading to:

$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0 \implies$$

$$\implies \boxed{\hat{\beta}_1 = \frac{n \sum\limits_{i=1}^{n} x_i y_i - \left(\sum\limits_{i=1}^{n} x_i\right)\left(\sum\limits_{i=1}^{n} y_i\right)}{n \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}}$$

$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \implies$$

$$\implies \boxed{\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}}$$

Practice with least squares

Earlier, we saw a webstore and part of the data they had collected about the number of visitors and their profits. As a reminder, here is the table with the data again:

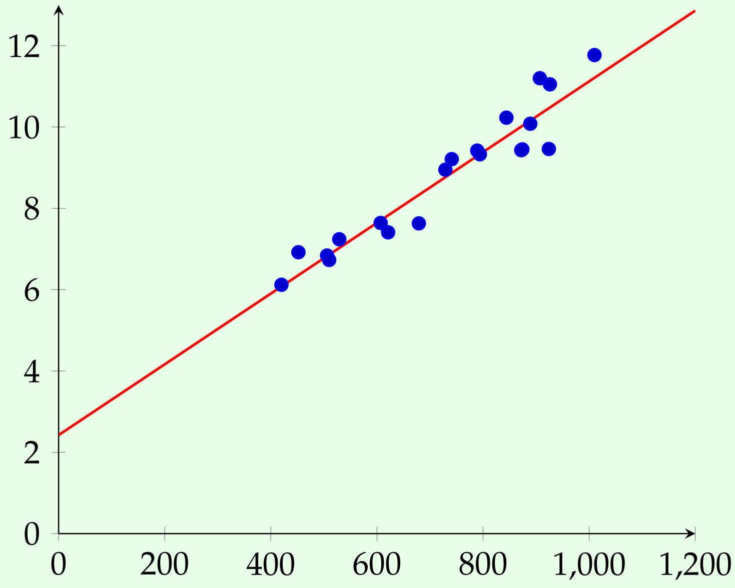
n	Visitors	Profit	n	Visitors	Profit
1	907	11.2	2	926	11.05
3	506	6.84	4	741	9.21
5	789	9.42	6	889	10.08
7	874	9.45	8	510	6.73
9	529	7.24	10	420	6.12
11	679	7.63	12	872	9.43
13	924	9.46	14	607	7.64
15	452	6.92	16	729	8.95
17	794	9.33	18	844	10.23
19	1010	11.77	20	621	7.41

What is the least squares line?

First, calculate $\sum x_i = 907 + 506 + \ldots = 14623, \sum y_i = 11.2 + 6.84 + \ldots = 176.11, \sum x_i y_i = 907 \cdot 11.2 + 506 \cdot 6.84 + \ldots = 134127.9, \sum x_i^2 = 907^2 + 506^2 + \ldots = 11306209$.

- $\hat{\beta}_1 = \frac{n \sum\limits_{i=1}^{n} x_i y_i - \left(\sum\limits_{i=1}^{n} x_i\right)\left(\sum\limits_{i=1}^{n} y_i\right)}{n \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2} = 0.0087.$
- $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 8.8055 - 0.0087 \cdot 731.15 = 2.423.$

Or, visually:



How can we use the regression line to help predict outcomes? Well, for a given value x, we may now predict y by plugging x in the regression line formula..

Using the regression line

For the previous webstore, how many profits should they anticipate on a very good day with 1200 visitors?

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 2.423 + 0.0087 \cdot 1200 = 12.863.$$

From now on, we will use the following terminology:

1. *observed values*:
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$
where (x_i, y_i) are the pairs of independent and dependent variables and ϵ_i the noise for $i = 1, \ldots, n$.

2. *fitted values*:
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$
where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the intercept and slope.

3. *residuals/errors*:
$$e_i = y_i - \hat{y}_i,$$
the difference between the observed and the fitted dependent values.

4. *sum of squares of errors*:
$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

Significance of simple linear regression

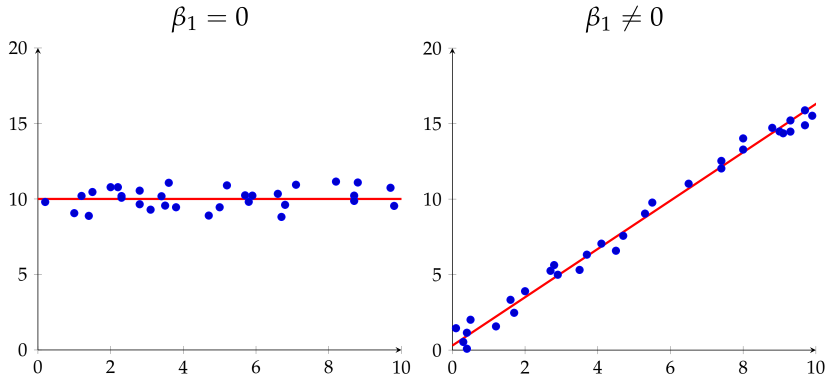
We got our intercept and slope; but is the regression line we got **significant**. What does that mean? What we want to ask is: “is there enough evidence to suggest that x and y are related?” Or does it appear to be just a random phenomenon, a coincidence?

Well, every time we want to check if we have enough evidence to “reject” something, we need *hypothesis testing*. When are x and y unrelated? When $\beta_1 = 0$! So, this is what we will formulate a hypothesis for.

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0.$$

An example of what this looks like is presented below in Figure 5.

Figure 5: An example of an insignificant regression (left), where the slope is 0, and an example of a significant regression (right), where the slope is non-zero.



Before we proceed with this, let us redefine the slope calculations. This will come in handy later. We have:

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

If we define:

- $S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$
- $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$

then we may get that:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}.$$

So, how is $\hat{\beta}_1$ distributed as? Recall that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. We then have that:

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}\right) \to \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

Unfortunately, σ^2 is not known – we will need some way to estimate it. Luckily, there is an easy to calculate estimator. We will need to keep track of the following notions:

- Recall that a sample variance can be calculated as $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$, where $n-1$ are the degrees of freedom as we needed to estimate one parameter in the calculation.

- In our case, we want to compare y_i to the average y value. $SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. However, it comes with $n-2$ degrees of freedom as we needed to estimate two parameters in its calculation $(\hat{\beta}_0, \hat{\beta}_1)$.
- Hence, we may use $\frac{SS_E}{n-2}$ as an estimator for σ^2!

This last quantity is called the **mean square error**:

$$MS_E = \frac{SS_E}{n-2}$$

and we can show that

$$E\left[MSE\right] = \sigma^2,$$

which serves to show that it is an unbiased estimator for our unknown variance:

$$\hat{\sigma}^2 = MS_E.$$

Finally, we are ready to pose the hypothesis test for the significance of our regression.

Simple linear regression significance

Null hypothesis:	Test statistic:	Distribution:
$H_0 : \beta_1 = 0.$	$T_0 = \frac{\hat{\beta}_1}{\sqrt{MS_E / S_{xx}}}.$	$T_0 \sim T_{n-2}.$

H_1	Rejection region	CI region
$\beta_1 \neq 0$	$\lvert T_0 \rvert > t_{\alpha/2,n-2}$	$\left[\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{MS_E}{S_{xx}}}, \right.$ $\left. \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{MS_E}{S_{xx}}}\right]$
$\beta_1 > 0$	$T_0 > t_{\alpha,n-2}$	$\left(\infty, \hat{\beta}_1 + t_{\alpha,n-2}\sqrt{\frac{MS_E}{S_{xx}}}\right]$
$\beta_1 < 0$	$T_0 < -t_{\alpha,n-2}$	$\left[\hat{\beta}_1 - t_{\alpha,n-2}\sqrt{\frac{MS_E}{S_{xx}}}, +\infty\right)$

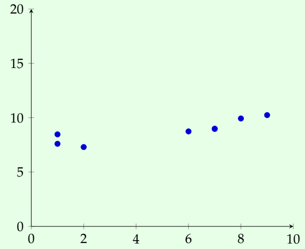
Note that this hypothesis test can be easily adapted to test for any value (not just zero!). How?

Let us put this to the test.

Is the regression significant?

Consider the following points:

x	y
1	7.6
9	10.24
2	7.3
7	8.97
6	8.74
7	8.99
8	9.93
1	8.47



1. Calculate $\hat{\beta}_0, \hat{\beta}_1$.

2. Using $\alpha = 0.10$, is there significant evidence that $\beta_1 \neq 0$?

3. Build a 90% confidence interval around $\hat{\beta}_1$.

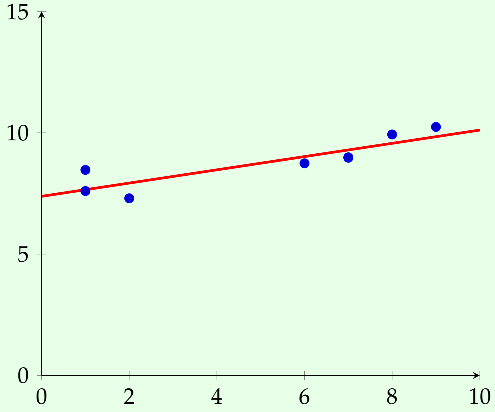
We'll again need to calculate: $n = 8, \sum x_i = 41, \sum y_i = 70.24, \sum x_i y_i = 380.43, \sum x_i^2 = 285$.
First, to calculate $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = 0.273.$$

Now, we can calculate $\hat{\beta}_0$:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = \frac{70.24}{8} - 0.273 \cdot \frac{41}{8} = 7.381.$$

Overall: $\hat{y} = 7.381 + 0.273 \cdot \hat{x}$.



Is the regression significant?

Recall that for our hypothesis test, we will need an estimator of the variance of the error σ^2..

- $\hat{\sigma}^2 = \frac{SS_E}{n-2}$.

To calculate SS_E, consider the original data, and append a new column (called $\hat{y}$). Populate it with the result $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$:

x	y	$\hat{y}$
1	7.6	7.654
9	10.24	9.838
2	7.3	7.927
7	8.97	9.292
6	8.74	9.019
7	8.99	9.292
8	9.93	9.565
1	8.47	7.654

Finally,

$$SS_E = \sum (y_i - \hat{y}_i)^2 = 1.629$$

and hence $\hat{\sigma}^2 = \frac{1.629}{6} = 0.272$.
We finally move to the hypothesis testing part.

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0.$$

- $T_0 = \frac{\hat{\beta}_1 - 0}{\sqrt{MS_E/S_{xx}}} = \frac{0.273}{\sqrt{0.272/74.875}} = 4.529$, where $S_{xx} = \sum (x_i - \overline{x})^2 = 74.875$.
- Compare to $t_{0.05,6} = 1.943$.
- Because $|T_0| > 1.943$, we reject the null hypothesis and deduce that with 90% confidence $\beta_1 \neq 0$.

Also note that

$$\beta_1 \in [0.273 - 1.943 \cdot 0.06, 0.273 + 1.943 \cdot 0.06] = [0.156, 0.390] .$$

Wait.. So does that mean that we can also use hypothesis testing to check whether $\hat{\beta}_1$ (the slope) has a certain value or not? The answer is a resounding yes!

Simple linear regression slope testing

Null hypothesis:	Test statistic:	Distribution:
$H_0 : \beta_1 = \beta_{10}.$	$T_0 = \dfrac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_E / S_{xx}}}.$	$T_0 \sim T_{n-2}.$

H_1	Rejection region	CI region
$\beta_1 \neq \beta_{10}$	$\lvert T_0 \rvert > t_{\alpha/2,n-2}$	$\left[\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{MS_E}{S_{xx}}}, \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{MS_E}{S_{xx}}}\right]$
$\beta_1 > \beta_{10}$	$T_0 > t_{\alpha,n-2}$	$\left(\infty, \hat{\beta}_1 + t_{\alpha,n-2}\sqrt{\frac{MS_E}{S_{xx}}}\right]$
$\beta_1 < \beta_{10}$	$T_0 < -t_{\alpha,n-2}$	$\left[\hat{\beta}_1 - t_{\alpha,n-2}\sqrt{\frac{MS_E}{S_{xx}}}, +\infty\right)$

A different perspective

For the previous example we have hypothesized that the line is $7.381 + 0.273 \cdot \hat{x}$. New data come in and give us the following four points: $(7, 9.97), (2, 7.95), (5, 8.91), (5, 8.14)$.

Using $\alpha = 0.05$, is there enough evidence in the new data to suggest that the slope has changed and we now have $\beta_1 > 0.273$?

Again we may calculate (for the new set of points) that: $n = 4, \sum x_i = 19, \sum y_i = 34.97, (\sum x_i) \cdot (\sum y_i) = 664.43, \sum x_i y_i = 170.94, \sum x_i^2 = 103, (\sum x_i)^2 = 361$. This leads to:

- $\hat{\beta}_1 = \frac{4 \cdot 170.94 - 664.43}{4 \cdot 103 - 361} = \frac{35.61}{51} = 0.379.$
- $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 6.942.$

We then have $(y_i - \hat{y}_i)^2 = (y_i - 6.942 - 0.379 \cdot x_i)^2$, which leads to:

- $(y_1 - \hat{y}_1)^2 = 0.140$
- $(y_2 - \hat{y}_2)^2 = 0.062$
- $(y_3 - \hat{y}_3)^2 = 0.005$
- $(y_4 - \hat{y}_4)^2 = 0.486$

This finally gives $SS_E = 0.694$ and a $\hat{\sigma} = \sqrt{MS_E} = \sqrt{\frac{SS_E}{n-2}} = \sqrt{\frac{0.694}{2}} = 0.589$. We are ready to formulate our hypothesis:

$$\boxed{H_0 : \beta_1 = 0.273 \qquad H_1 : \beta_1 > 0.273}$$

A different perspective

On to our hypothesis testing calculations:

- **the test statistic**: $T_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{0.106}{0.589/\sqrt{12.75}} = 0.643$, where $S_{xx} = \sum (x_i - \overline{x})^2 = 12.75$.
- **the critical value**: $t_{0.05,2} = 2.92$. Recall that the hypothesis is one-sided here.
- **the comparison**: we have that $T_0 < t_{\alpha,n-2}$, which means that we accept the null hypothesis.

Hence, we deduce that with 95% confidence β_1 is still equal to 0.273 (even with the new data suggesting otherwise). Also note that

$$\beta_1 \in (-\infty, 0.273 + 2.92 \cdot 0.165] = (-\infty, 0.755]\,,$$

which further reinforces that the new data should be even more indicative of a change (result in $\hat{\beta}_1 > 0.755$) to accept the change.

So.. this is how it works in simple linear regression with one dependent and one independent variable. How about we generalize this to more than just one independent variable? More on that, next time!