Simple Linear Regression: Modelling Linear Relationship

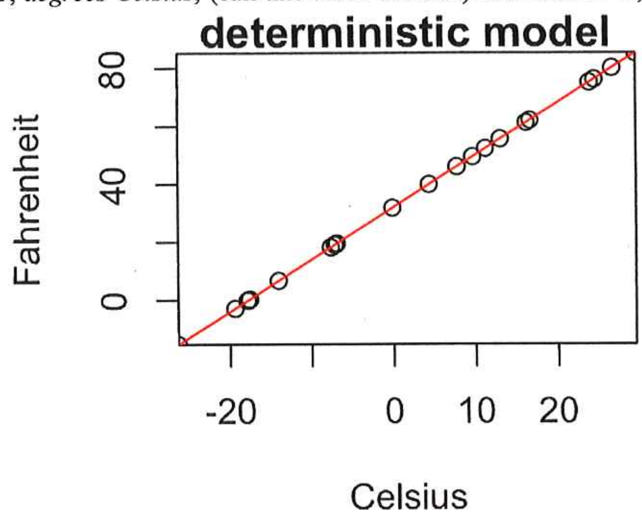
After discovering covariance between bivariate data, you will likely want to know how to describe/express the relationship. Successful descriptions can then be used to model expected results on one of the variables deemed the **response variable, *Y***, based only off the **predictor variable, *X***.

The steps are:

1. Collect bivariate X *and* Y variables from historic events. This data set will serve as "training" to understand the linear relationship that exists between the variables.
2. Develop a mathematical expression/equation to transform an particular/hypothetical X into an **expected/estimated** Y.

Consider, "bivariate" data expressing temperature in $°C$, *degrees Celsius*, (call this the X variable) and then in $°F$, *degrees Fahrenheit*, (call this the Y variable).

	Celsius	Fahrenheit
1	-14.0694802	6.6749356
2	-7.6725660	18.1893812
3	4.3712018	39.8681632
4	24.4924674	76.0864413
5	-17.8990841	-0.2183514
6	23.9033811	75.0260860
7	26.6805161	80.0249290
8	9.6478675	49.3661616
9	7.7468426	45.9443167
10	-26.2928238	-15.3270828
11	-17.6415255	0.2452541
12	-19.4065948	-2.9318707



What do you notice about the scatterplot?
Do you see how all the points fall exactly on the line? This is called a deterministic model; since all points fall exactly on the line we could perfectly predict where no points exist.

The linear equation will follow the **deterministic model's** form:

$$Y_i = \beta_0 + \beta_1 X_i, \quad i = 1, 2, \dots, n$$

where β_0 is the y-intercept (the Y value when $X = 0$) and β_1 is the slope (rate of change in Y with respect to X).

In a deterministic model only two sample points are all that is required to find the model:

$$\beta_1 = \frac{rise}{run} = \frac{y_2 - y_1}{x_2 - x_1}, \text{ then } \beta_0 = y_1 - \beta_1 x_1$$

Of course in the "real" world we often lack the ability to measure variables with deterministic precision. We expect response and or measurement bias in our observations. Additionally, when dealing with random variables we know that our observations may lack consistency not due to any bias at all!

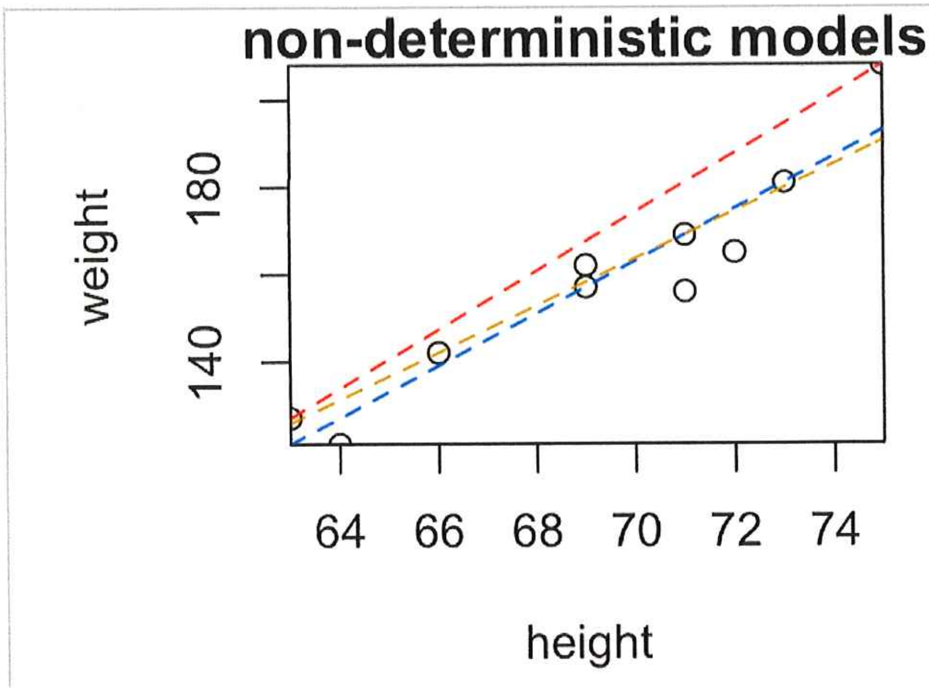
$\beta_1 \cong \frac{6.675 - 18.189}{-14.069 - -7.673} \cong 1.800$ $\beta_0 \cong 6.675 - (1.8)(-14.069) \cong 32.$

$(°F) = 32 + 1.8(°C)$ ⇒ for each Additional $\underbrace{1\,°C}$, °F goes up $\underbrace{1.8}$ degrees ← Slope, β_1

⇒ At 0°C ⇒ 32°F ← β_0, y intercept.

Let's return to our example of ten student's height and weight, and try to select two points from the data set then create linear models…

I selected three combinations of two point set's and tried to use Deterministic Approach… "Floss Example"



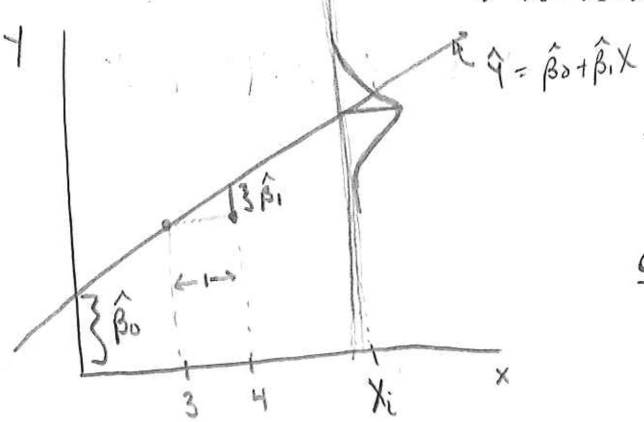
So which model of a non-deterministic data set is best?...

The **probabilistic model** appears to the same as the deterministic model, however, it includes an addition ε_i term:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \qquad where\ \varepsilon_i \sim Norm(\beta_0 + \beta_1 X_i, \sigma^2)$$

The probabilistic model can also be written this way, to "hide" the ε_i term by admitting the following values are estimates:

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_i, \qquad i = 1, 2, \ldots, n$$



$Y_i \sim Norm(\hat{\beta}_0 + \hat{\beta}_1 X_i, \sigma^2)$

(Y is random variable)

OR…

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

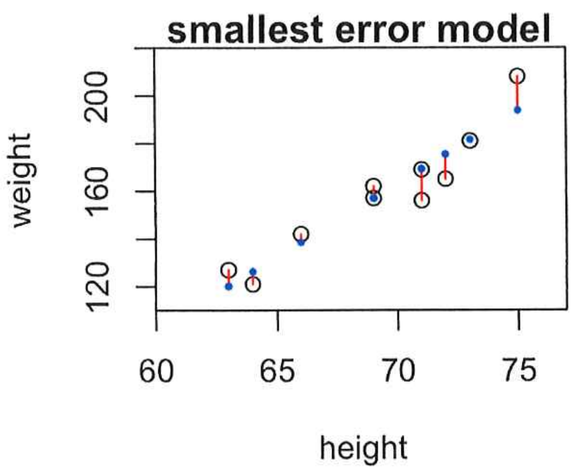
$\varepsilon_i \sim Norm(0, \sigma^2)$

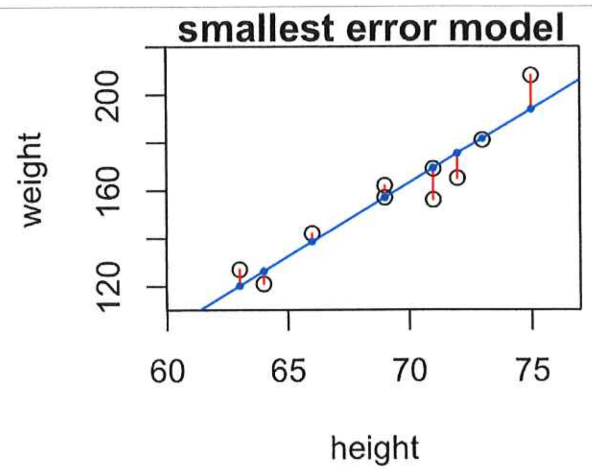
(Y is a conditional constant, that is measured inaccurately)

$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X$; The Least-Squares Estimate Model

Let $\hat{Y}$ be the probabilistic model that is the **closest** "overall" to each sample point, meaning the set of sample points $(X_i, Y_i)'s$ that are respectively **closest** to the $(X_i, \hat{Y}_i)'s$. Then we define the difference between Y_i and $\hat{Y}_i$ to be ε_i.

Then $\varepsilon_i = Y_i - \hat{Y}_i$ (residuals/errors/residual errors), we are setting $\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n}(Y_i - \hat{Y}_i) = 0$ so we can find the model with the **least** overall error!





One complication that comes up due to setting the sum of errors equal to zero is that you have now clearly made the signs of some of the errors negative.

To overcome this we square the individual error terms and then discuss the SUM of SQUARED ERRORS, $SSE = \sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (\widehat{\beta_0} + \widehat{\beta_1} X_i))^2$ $\quad Y_i \rightarrow$ observed Value $\quad \hat{Y}_i \rightarrow$ predicted Value

We wish to estimate the probabilistic model in such a way that the square of these vertical distances is as small as possible, a method that invokes **least-squares estimation**. Consider the sum of the squared distances/errors, SSE, each bivariate data point lies away from the imaginary linear line, $\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X$.

We need to **minimize** SSE with respect to $\widehat{\beta_0}$ then with respect to $\widehat{\beta_1}$ by finding then setting the partial derivatives equal to zero. $\frac{\delta SSE}{\delta \widehat{\beta_i}}, where\ i = 0, 1$

We will see:

The least-squares estimate of the Y -intercept of the model is:

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$$

The least-squares estimate of the slope of the model is:

$$\widehat{\beta_1} = \frac{S_{XY}}{S_{XX}} = \frac{S_{XY}}{S_X S_X} = \frac{S_{XY}}{S_X^2} = r\frac{S_y}{S_X} = \frac{\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i Y_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i Y_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\,\bar{X}^2}$$

$\frac{\delta SSE}{\delta \widehat{\beta_0}} = \frac{\partial}{\partial \hat{\beta}_0}\left[\Sigma\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2\right]$

need to expand $\left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\right]^2$ terms.

However I will be taking the derivative wrt $\hat{\beta}_0$

So... I only care about terms with $\hat{\beta}_0$ in them...

expantion:

$(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$

Only 5 terms matter...

$\propto \frac{\partial}{\partial \hat{\beta}_0}\Sigma\left[-y_i\hat{\beta}_0 + \hat{\beta}_0^2 + \hat{\beta}_0\hat{\beta}_1 x_i - y_i\hat{\beta}_0 + \hat{\beta}_0\hat{\beta}_1 x_i\right] = \frac{\partial}{\partial \hat{\beta}_0}\Sigma\left[-2y_i\hat{\beta}_0 + \hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1 x_i\right]$

Sum of ders = der of sums

$= \Sigma\left[-2y_i + 2\hat{\beta}_0 + 2\hat{\beta}_1 x_i\right] = -2\Sigma y_i + 2n\hat{\beta}_0 + 2\hat{\beta}_1\Sigma x_i$

find min/max set $=0$! Solve for $\hat{\beta}_0$!

$-2\Sigma y_i + 2n\hat{\beta}_0 + 2\hat{\beta}_1\Sigma x_i = 0 \Rightarrow \frac{2n\hat{\beta}_0}{2n} = \frac{2\Sigma y_i}{2n} - \frac{2\hat{\beta}_1\Sigma x_i}{2n}$

$\Rightarrow \hat{\beta}_0 = \frac{\Sigma y_i}{n} - \hat{\beta}_1\frac{\Sigma x_i}{n} = \boxed{\bar{Y} - \hat{\beta}_1\bar{X} = \hat{\beta}_0}$

ensure it is a min (want to min Error to have least-square estimate!)

$\frac{\partial^2(SSE)}{\partial \hat{\beta}_0^2} \propto \frac{\partial}{\partial \hat{\beta}_0}(-2\Sigma y_i + 2n\hat{\beta}_0 + \hat{\beta}_1\Sigma x_i) = 2n > 0$ since $n \geq 1$

positive!

So by 2nd derivative test. $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$ minimizes SSE!

$$\frac{\delta SSE}{\delta \widehat{\beta_1}} = \frac{\partial}{\partial \hat{\beta}_1}\left[\sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2\right]$$

Need to expand $[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$ term...

However, I know I will be taking the derivative wrt. $\hat{\beta}_1$

So... I only care about terms that will have $\hat{\beta}_1$ in them. ignore the rest?...

Consider expantion:

$$(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

Only five terms will matter!

and only three like terms.

$$\frac{\partial SSE}{\partial \hat{\beta}_1} \propto \frac{\partial}{\partial \hat{\beta}_1}\left[\sum (\hat{\beta}_1^2 x_i^2 + 2\hat{\beta}_1 \hat{\beta}_0 x_i - 2\hat{\beta}_1 x_i y_i)\right]$$

Sum of der is same as der of sum

$$= \sum \left[2\hat{\beta}_1 x_i^2 + 2\hat{\beta}_0 x_i - 2x_i y_i\right] = -2 \cdot \sum (x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i))$$

Set $\frac{\partial SSE}{\partial \hat{\beta}_1} = 0$ to find min or max

$$-2\sum (x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)) = 0 \Rightarrow \sum (x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)) = 0$$

$$\Rightarrow \sum (x_i y_i - \hat{\beta}_0(x_i) - \hat{\beta}_1 x_i^2) = \sum x_i y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2 = 0$$

$$\Rightarrow \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i - \hat{\beta}_0 \sum x_i$$

$$\Rightarrow \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum x_i \left(\frac{n}{n}\right) \quad \text{recall} \left(\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}\right) \text{ sub in!}$$

$$\Rightarrow \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i - \bar{x}\bar{y}n - \hat{\beta}_1 \bar{x}^2 n \Rightarrow \hat{\beta}_1(\sum x_i^2 - n\bar{x}^2) = \sum x_i y_i - n\bar{x}\bar{y}$$

$$\Rightarrow \boxed{\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}}$$

ensure $\hat{\beta}_1$ minimizes SSE!

$$\frac{\partial^2 SSE}{\partial \hat{\beta}_1^2} \propto \frac{\partial}{\partial \hat{\beta}_1} \left(\Sigma \left(2\hat{\beta}_1 x_i^2 + 2\hat{\beta}_0 x_i - 2 x_i y_i \right) \right)$$

$$\propto \Sigma \left(2 x_i^2 \right) = 2 \Sigma x_i^2 > 0$$

clearly positive

Sum of squared values doubled...

then by 2nd der. test $\hat{\beta}_1 = \dfrac{\Sigma x_i y_i - n\bar{X}\bar{Y}}{\Sigma x_i^2 - n\bar{X}^2} = \dfrac{\Sigma (y_i - \bar{y})(x_i - \bar{x})}{\Sigma (x_i - \bar{x})^2}$

minimizes SSE!

Example 1: Let's look back at our student height and weight data. Found in data file in D2L, using height as the predictor and weight as the response.

$$\text{Weight}_i = \beta_0 + \beta_1 \text{ height}_i + e_i \iff \widehat{\text{weight}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{ height}_i$$

height = c(63, 64, 66, 69, 69, 71, 71, 72, 73, 75)

weight = c(127, 121, 142, 157, 162, 156, 169, 165, 181, 208)

(i) Find the least-squares estimate of the model.

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{(63\cdot127)+(64\cdot121)+\dots+(75\cdot208) - 10\left[\frac{63+64+\dots+75}{10}\right]\left[\frac{127+121+\dots+208}{10}\right]}{(63^2+64^2+\dots+75^2) - 10\left[\frac{63+64+\dots+75}{10}\right]^2} \qquad \hat{\beta}_1 \cong 6.137581$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} \cong \left(\frac{127+121+\dots+208}{10}\right) - 6.1376\left(\frac{63+64+\dots+75}{10}\right) \qquad \hat{\beta}_0 \cong -266.534395$$

R-command:

fit = lm(Y~X)

fit$coef

$$\Rightarrow \hat{Y}_i = -266.534 + 6.138 X_i$$

$$\Rightarrow \widehat{\text{weight}}_i = -266.534 + 6.138(\text{height}_i)$$

(ii) Interpret the meaning of $\widehat{\beta_1}$ found in part (i).

$\hat{\beta}_1$ for every additional unit increase in X Expect, $\hat{\beta}_1$, unit increases in Y.

Here in Context

for every additional inch in height expect a person to weight and additional ~6.14 lbs...

Causation? Does weight cause height? Does height "cause" weight?

Clearly Not The only factor involved...

Confounding / Additional variables!

(iii) Using the estimate of the model in (i), predict the weight a student with a height of 2.5 feet (recall there are 12 inches in a foot). Interpret the meaning of this predicted value. Try again with a height of 67 inches.

? $\hat{Y}_i = -266.534 + 6.138(2.5)$ $\Rightarrow 2.5(12) = 30$ inches.

$\Rightarrow \hat{Y}_i = -266.534 + 6.138(30) \cong -82.394.$

$\Rightarrow$ We expect people who are 30 inches tall to weigh -82.4 lbs...

why is this crazy? $\Rightarrow$ Scope of Sampling from X range from 63 - 75

No where near value of 30 we don't know the relationship at values like 30 but clearly it is Not the same relationship as it is around 63-75 inch in height. (extrapolation!)

$\Rightarrow \hat{Y}_i = -266.534 + 6.138(67) \cong 144.713.$

We expect people who are 67 inches tall to weigh <u>about</u> 144.7 lbs!

(iv) Find the value of the <u>residual</u> corresponding to the fourth data point: (X4 = 69, Y4 = ~~162~~ 157)

$\hat{Y}_i = -266.534 + 6.138(69) \cong 156.988$

$Y_i = 157.$

$e_i = Y_i - \hat{Y}_i \cong 157 - 156.988 \cong 0.012.$

* See R-code

Example 2: How strong is the linear relationship between the age of a driver and the distance the driver can see? A research firm (Last Resource, Inc., Bellefonte, PA) collected data on a sample of n = 30 drivers. What can you say about the relationship? What is likely to be X? → Easy to know?

Y? → interest!

$\widehat{Distance}_i = 576.6819 - 3.0068\ Age_i$

for each year old you can see 3.0068 (feet?/meters?/yard? ...) <u>less</u>

if you were just born you should be able to see 576.6819 (?)

Cov (Distance, Age) $\doteq$ -1425.862

Cor (Distance, Age) $\doteq$ -0.8012447.