# Multiple regression & Gauss-Markov Theorem

Luis Castro-de-Araujo[a]

11/07/2022

Virginia Institute for Psychiatric and Behavioral Genetics

[a] Post-doc T32. luis.araujo@vcuhealth.org

Linear regression recap

Multiple regression

Gauss-Markov Theorem

# Linear regression recap

# Regression

- Develop basic concepts of linear regression from a probabilistic framework
- Estimating parameters and hypothesis testing with linear models

# Regression[2]

- Technique used for the modeling and analysis of numerical data
- Exploits the relationship between two or more variables so that we can gain information about one of them through knowing values of the other
- Regression can be used for prediction, estimation, hypothesis testing, and modeling causal relationships

[1] Akey, 2020, “Regression,” (Washington 2020).

[2] Akey, 2020, “Regression,” (Washington 2020).

- It is all about describing relationships between variables

# THE CULPRIT

## Bio

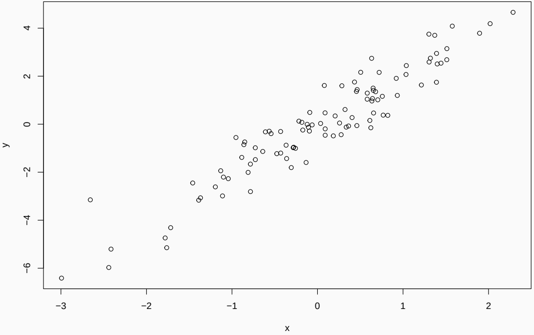
- 30 April 1777 - 23 April 1855
- Worked in the theorem by 1794
- 17 years old![a]

## Carl Friedrich Gauss

[a] "Carl Friedrich Gauss," (2022), *Wikipedia*.

# Before him, description of relationship was not systematic

```r
set.seed(42)
x <- rnorm(100)
y <- 2 * x + rnorm(100, sd = 0.8)
plot(x, y, xlab = "x", ylab = "y")
```



# Regression Lingo[4]

$$Y = X1 + X2 + X3$$

| Left of expression | Right of expression |
|---|---|
| Dependent Variable | Independent Variable |
| Outcome Variable | Predictor Variable |
| Response Variable | Explanatory Variable |

[3] Joshua Akey, "Regression."

[4] Joshua Akey, "Regression."

# Why Linear Regression?[6]

- Suppose we want to model the dependent variable Y in terms of three predictors, X1, X2, X3

$$Y = f(X1, X2, X3)$$

- Typically will not have enough data to try and directly estimate *f*
- Therefore, we usually have to assume that it has some restricted form, such as linear

$$Y = X1 + X2 + X3$$

[5] Ibid.

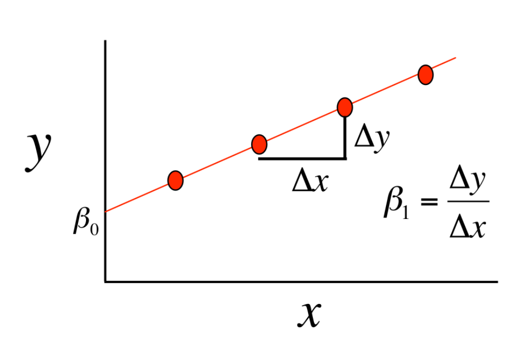
[6] Ibid.

# Linear Regression is a Probabilistic Model

- Much of mathematics is devoted to studying variables that are deterministically related to one another[a]

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

- But we're interested in understanding the relationship between variables related in a nondeterministic fashion.



[a] Ibid.

# A Linear Probabilistic Model

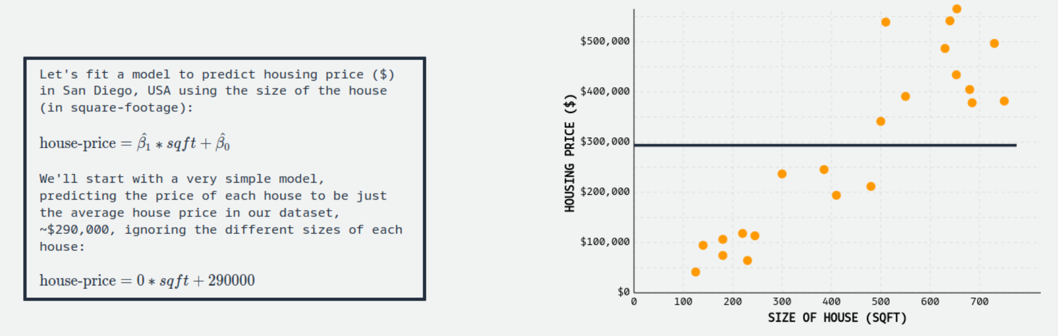
- Definition: There exists parameters , , and , such that for1 2 any fixed value of the independent variable x, the dependent variable is related to x through the model equation[7]

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

- The error term $\epsilon$ is a random variable with mean 0 and constant variance $\sigma^2$ *is a rv assumed to be N(0, 2)*

[7] Ibid.

## From a bad model[8]



Let's fit a model to predict housing price ($) in San Diego, USA using the size of the house (in square-footage):

$$\text{house-price} = \hat{\beta}_1 * sqft + \hat{\beta}_0$$

We'll start with a very simple model, predicting the price of each house to be just the average house price in our dataset, ~$290,000, ignoring the different sizes of each house:

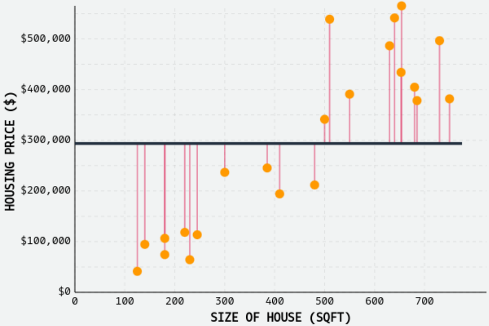
$$\text{house-price} = 0 * sqft + 290000$$

[8] Wilber, “Linear Regression,” *MLU-Explain*.

# IN OTHER WORDS

Of course we know this model is bad - the model doesn't fit the data well at all. But how can do quantify exactly *how* bad?

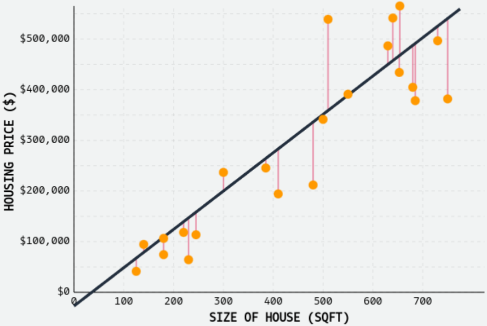
To evaluate our model's performance quantitatively, we plot the error of each observation directly. These errors, or **residuals**, measure the distance between each observation and the predicted value for that observation. We'll make use of these residuals later when we talk about evaluating regression models, but we can clearly see that our model has a lot of error.



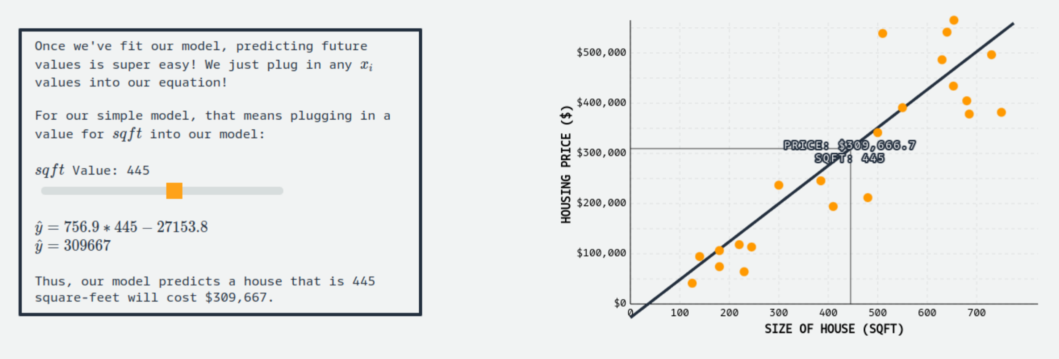
# In other words

The goal of linear regression is reducing this error such that we find a line/surface that 'best' fits our data. For our simple regression problem, that involves estimating the y-intercept and slope of our model, $\hat{\beta}_0$ and $\hat{\beta}_1$.

For our specific problem, the best fit line is shown. There's still error, sure, but the general pattern is captured well. As a result, we can be reasonably confident that if we plug in new values of square-footage, our predicted values of price would be reasonably accurate.



# To the best possible model[9]



Once we've fit our model, predicting future values is super easy! We just plug in any $x_i$ values into our equation!

For our simple model, that means plugging in a value for $sqft$ into our model:

$sqft$ Value: 445

$$\hat{y} = 756.9 * 445 - 27153.8$$
$$\hat{y} = 309667$$

Thus, our model predicts a house that is 445 square-feet will cost $309,667.

[9] Ibid.

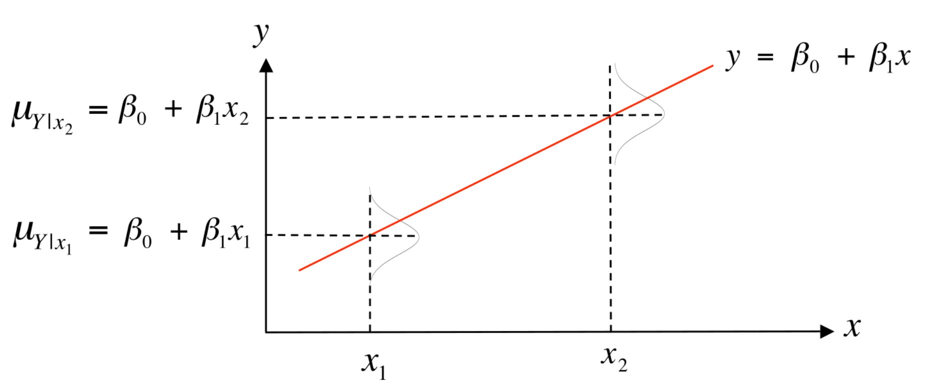
# Implications

- The **expected** value of Y is a linear function of X, but for fixedx, the variable Y differs from its expected value by a random amount
- Formally, let x* denote a particular value of the independent variable x, then our linear probabilistic model says:

$E(Y|x^*) = \mu_{Y|x^*}$ = mean value of Y when x is x*

$V(Y|x^*) = \sigma^2_{Y|x^*}$ = variance of Y when x is x*

# Graphical Interpretation



$y$
$y = \beta_0 + \beta_1 x$
$\mu_{Y|x_2} = \beta_0 + \beta_1 x_2$
$\mu_{Y|x_1} = \beta_0 + \beta_1 x_1$
$x$
$x_1$
$x_2$

- For example, if x = height and y = weight then $\mu_{Y|x^*} = 60$ is the average weight for all individuals 60 inches tall in the population[10]

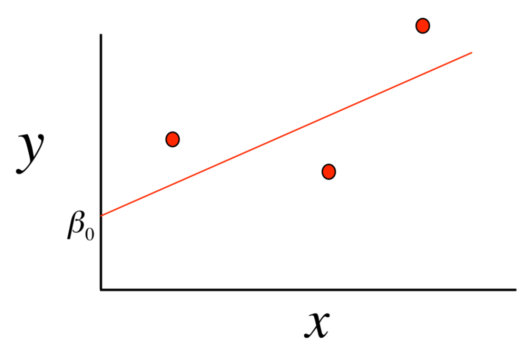
[10] Ibid.

# Estimating Model Parameters

- Point estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained by the principle of least squares

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

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



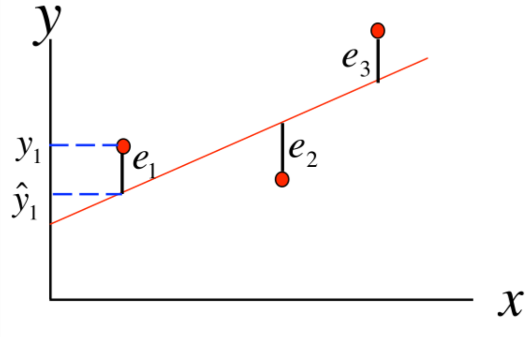
# Predicted and Residual Values

**Predicted** or fitted, values of y predicted by the least-squares regression line obtained by plugging in x1,x2,. . . ,xn into the estimated regression line

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_2$$

**Residuals** are the deviations of observed and predicted values



# Residuals Are Useful!

- They allow us to calculate the error sum of squares (SSE):

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

- Which in turn allows us to estimate $\sigma^2$:

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

- As well as the **coefficient of determination**:

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

# Multiple regression

# Multiple Linear Regression

- *Extension of the simple linear regression model to two or more independent variables*

  *y=0 + 1x1 + 2x2 + ...+ nxn +*

  *Expression = Baseline + Age + Tissue + Sex + Error*

- *Partial Regression Coefficients: ieffect on the dependent variable when increasing the ith independent variable by 1 unit,* ***holding all other predictors constant***

  *Categorical Independent Variables*

- Qualitative variables are easily incorporated in regression *framework through* ***dummy variables***
  - *Simple example: sex can be coded as 0/1*
  - *What if my categorical variable contains three*

# Acknowledgements

## Team

- Joshua Pritkin.
- Rob Kirkpatrick.

- Michael C Neale.
- NIH grant no R01 DA049867 and 5T32MH-020030

## Contact

- Questions?