Linear Regression: Introduction & Estimation

Connor K. Brubaker

Department of Statistics
Texas A&M University

Equation of a Line

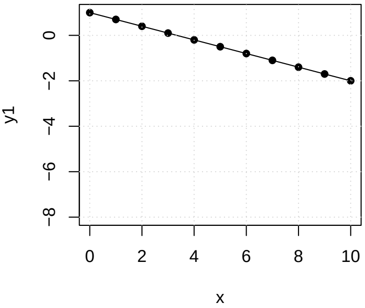
The equation of a straight line is

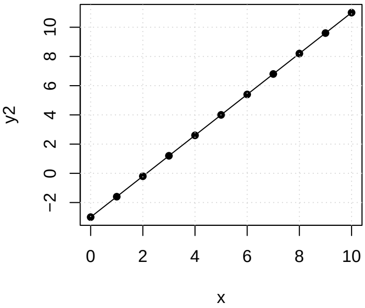
$$y = bx + a$$

where b is the slope of the line and a is the intercept.

Linear Relationships

Below are examples of *perfect* linear relationships.





Interpretation of slope

For any two points (x_1, y_1) and (x_2, y_2),

$$b = \frac{y_2 - y_1}{x_2 - x_1}.$$

The slope is the **exact** rate of change in y for every unit increase in x.

Interpretation of intercept

When $x = 0$,

$$y = b(0) + a = a$$

The intercept is the **exact** value of y when $x = 0$.

Example: Housing Prices

- How much should you pay for a house?

Example: Housing Prices

- How much should you pay for a house?
- Factors that influence price:

Example: Housing Prices

- How much should you pay for a house?
- Factors that influence price:
  - Location

Example: Housing Prices

- How much should you pay for a house?
- Factors that influence price:
  - Location
  - Year built

Example: Housing Prices

- How much should you pay for a house?
- Factors that influence price:
  - Location
  - Year built
  - Amenities

Example: Housing Prices

- How much should you pay for a house?
- Factors that influence price:
  - Location
  - Year built
  - Amenities
  - **Square footage**

Example: Housing Prices

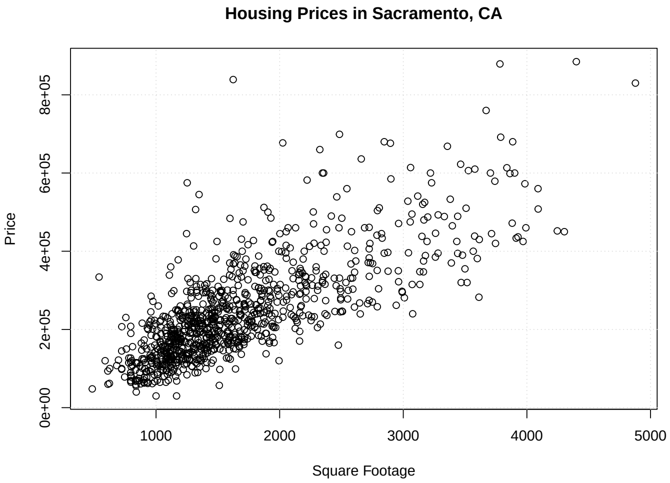
- How much should you pay for a house?
- Factors that influence price:
  - Location
  - Year built
  - Amenities
  - **Square footage**
- To determine a price, we might **model** price as a function of square footage:

$$\text{Price} = f(\text{Square Footage}) + \varepsilon$$

f is called the **regression function.**

Motivating Example: Housing Prices

Scatter plots help determine the functional relationship between two variables.



Simple Linear Regression

The simplest model is where f is a linear function:

$$f(x) = \beta_0 + \beta_1 x.$$

The model becomes

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

$$\text{Price}_i = \beta_0 + \beta_1 \text{Square Footage}_i + \varepsilon_i.$$

Y_i is called the **dependent variable** or the **response** and X_i is called the **independent variable** or **predictor**.

Parameters of SLR

Under SLR, the model is

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

The **parameters** of the model are the intercept β_0 and the slope β_1. These are unknown and must be estimated using data.

The Error Term

Under SLR, the model is

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

The **error term** ε_i accounts for the fact that

- not all the points lie exactly on the regression line and
- Y cannot be perfectly predicted from X alone

The Error Term

Under SLR, the model is

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

We assume the **error term** ε_i satisfies

- $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$
- the error terms are all independent of each other (mutually independent)

Regression Models the Conditional Expectation

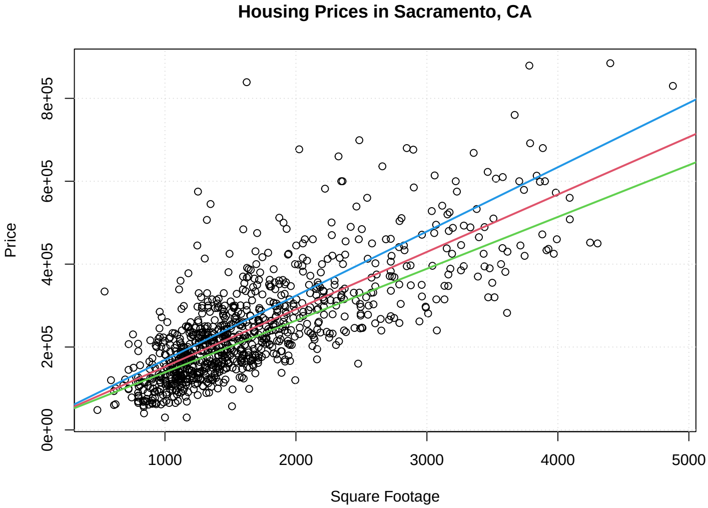
In SLR, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ so that $\mathbb{E}(\varepsilon_i) = 0$. Treating X_i as a given constant, we have

$$\begin{aligned}\mathbb{E}[Y_i|X_i] &= \mathbb{E}(\beta_0 + \beta_1 X_i + \varepsilon_i) \\ &= \beta_0 + \beta_1 X_i + \mathbb{E}(\varepsilon_i) \\ &= \beta_0 + \beta_1 X_i\end{aligned}$$

$\mathbb{E}[Y_i|X_i]$ is the conditional expectation of Y_i given X_i - it *depends* on the value X_i.

Line of Best Fit

Many different lines “fit” the data, which is the best?



RESIDUALS

Given some β_0 and β_1, the predicted value of Y_i at X_i is

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

The *i*th residual is

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \beta_0 - \beta_1 X_i$$

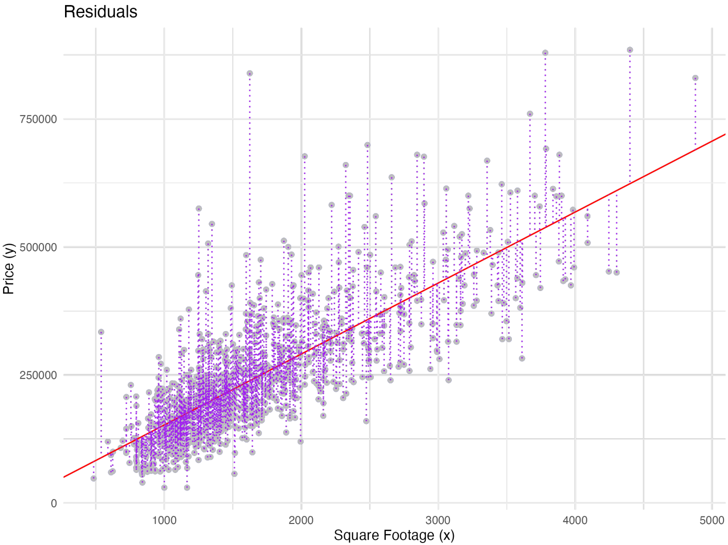
LEAST SQUARES CRITERION

We choose the line with intercept β_0 and slope β_1 that minimizes the sum of squared residuals,

$$RSS = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

The values of β_0 and β_1 that minimize RSS are called the **least squares estimators.**

Residuals of the Housing Data



Least Squares Estimators

Use principles of calculus to find the minimizers of the residual sum of squares,

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Least Squares Estimators

Define

$$SXY = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \text{ and } SXX = \sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The least squares estimators are

$$\hat{\beta}_1 = \frac{SXY}{SXX} \text{ and } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Covariance

The quantity

$$SXY = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

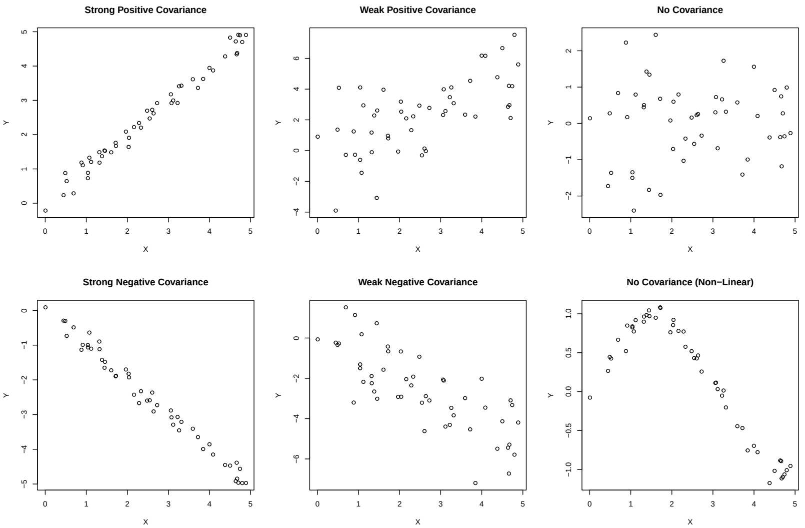
is an estimator of the **covariance** between X and Y.

COVARIANCE

The covariance $Cov(X, Y)$ between X and Y

- quantifies the strength of the linear relationship between X and Y,
- can be positive or negative, and
- could be near zero if the relationship is non-linear.

Covariance



Strong Positive Covariance
Weak Positive Covariance
No Covariance
Strong Negative Covariance
Weak Negative Covariance
No Covariance (Non−Linear)
X
Y

Correlation

Standardizing the covariance gives the correlation ρ:

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

It has the same interpretation but is restricted to lie between -1 and 1. Correlation of 1 is a perfect positive linear relationship.

Interpretation of Slope Parameter

Recall

$$\mathbb{E}[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

- The slope is the change in the *expected value* of Y for every unit increase in X.
- The intercept is the *expected value* of Y when $X = 0$.

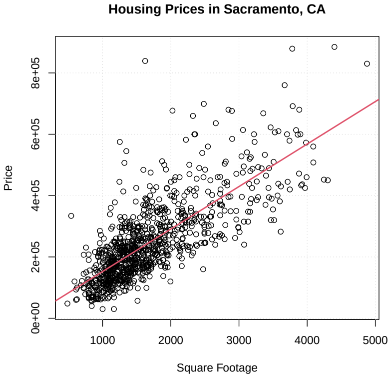
Example: Housing Prices

For the housing data,

$$\hat{\beta}_0 = 13859.393$$

and

$$\hat{\beta}_1 = 138.546$$



No Intercept Model

Sometimes it is appropriate to fit the SLR model with no intercept:

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

Is this appropriate for the housing data?

LINEAR MODEL IN R

A linear model can be fit using

```
> model <- lm(price ~ sqft)
> model <- lm(price ~ sqft - 1) # no intercept
> summary(model)
```

Linear Model in R

```
> summary(model)

Call:
lm(formula = price ~ sqft, data = Sacramento)

Residuals:
    Min      1Q  Median      3Q     Max
-231889  -54717  -11822   38993  600141

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13859.393   6948.714   1.995   0.0464 *
sqft          138.546      3.796  36.495   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 84130 on 930 degrees of freedom
Multiple R-squared:  0.5888,    Adjusted R-squared:  0.5884
F-statistic:  1332 on 1 and 930 DF,  p-value: < 2.2e-16
```

$\hat{\beta}_0$

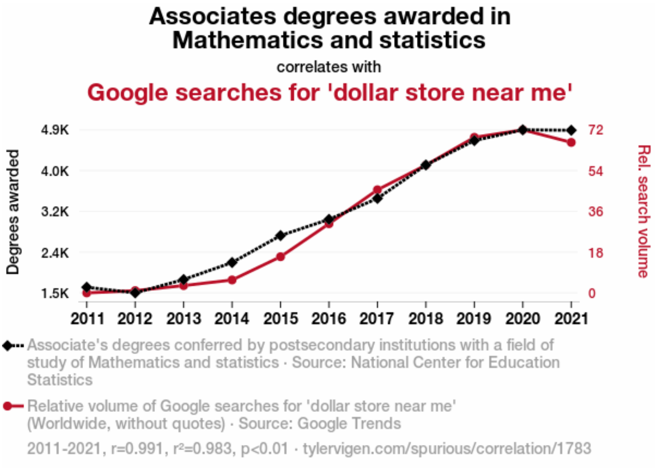
$\hat{\beta}_1$

Association and Causation

A strong relationship between two variables does not always imply a causal relationship.

A strong association between two variables is often due to lurking variables that we are not aware of.

Association and Causation



https://www.tylervigen.com/spurious-correlations

Association and Causation

- The best evidence for causal relationships comes from properly designed randomized experiments.
- Observational studies can show a strong association, but it is not appropriate to conclude causation.

Does Smoking Cause Lung Cancer?

- Unethical to investigate this with an experiment.
- Observational studies have demonstrated an association between lung cancer and smoking.
- Evidence has been collected from many studies.
- It is plausible that smoking causes cancer, but the conclusion is not as strong as evidence from a randomized experiment.