

Simple Linear Regression: Model and Estimation



UNIVERSITY OF
SAN FRANCISCO

James D. Wilson

MSAN 601 - Linear Regression Analysis



Outline

- Measures of association: covariance and correlation
- The Simple Linear Regression Model
- Least Squares Estimation
- The Gauss-Markov Theorem



Setting: Simple Linear Regression

Given:

- Response $y = (y_1, \dots, y_n)^T$ – *continuous-valued*
- A single explanatory variable X

Aim: Fit / estimate the *simple linear regression model* of y on X :

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

Important Considerations:

- β_0 and β_1 are *unknown regression coefficients*
- Estimation = calculating our “best” guess: $\hat{\beta}_0$, and $\hat{\beta}_1$



Defining the SLR Model

- Two variables: the explanatory variable, X , explains—in part or wholly—the response variable Y
- You may encounter other terminology when referring to explanatory and/or response variables:

Y	X
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor variable

Table: Terminology for Simple Linear Regression

Why *Simple*? Why *Linear*? Why *Regression*?



- **Simple**: only dealing with two variables, an *explanatory* variable and a *response* variable
- **Linear**: linear relationship between X and y
- **Regression**: developed by Sir Francis Galton in the late 1800s:
 - when studying the relationship between the heights of parents and their children, he noted “children of both shorter and taller parents regressed to the group mean height”



Features

- Functional relationship, e.g., $y = f(X) = 3X - 9$
 - for each X , function returns a corresponding value of y
- Statistical relationships are imperfect
 - observations for a statistical relationship do not typically fall directly on the curve (line) of the relationship



- Analyze individual variables
 - **numerically:** mean, median, mode, variance, quantiles, skewness, kurtosis, etc.
 - **graphically:** histograms, box plots, etc.
- Analyze relationship(s) between variables
 - **numerically:** covariance, correlation
 - **graphically:** scatter plots



Quantifying Relationships Between X and y

Suppose that (X, Y) are jointly distributed random variables

Covariance between X and Y

Let X and Y be two random variables. Then the covariance of the two random variables is

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Idea: Covariance gives idea of the relationship between X and Y .



Properties of Covariance

Basic Properties

- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- $\text{Cov}(X, X) = \text{Var}(X)$
- $\text{Cov}(aX, Y) = a\text{Cov}(X, Y)$
- If X and Y are independent, then $\text{Cov}(X, Y) = 0$

Important property

Let X_1, \dots, X_n and Y_1, \dots, Y_n be a collection of random variables.

Then,

$$\text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^n Y_j\right) = \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(X_i, Y_j)$$



Properties of Covariance

Important consequence 1

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i < j} \text{Cov}(X_i, X_j)$$

Important consequence 2

If X_i 's are *pairwise independent*, then

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i)$$

Notes:

- These properties can be extended to multivariate X and Y
- Covariances can take on *any* value on the real line..



Correlation

Correlation between X and Y

$$\rho(X, Y) = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Fact 1

$$-1 \leq \rho(X, Y) \leq 1$$

Terminology: X and Y are called **uncorrelated** when $\rho(X, Y) = 0$.



Properties of Correlation

Fact 2

Suppose that $\rho(X, Y) = -1$, then $Y = -aX + b$ with $a > 0$.

... what does this suggest? ...

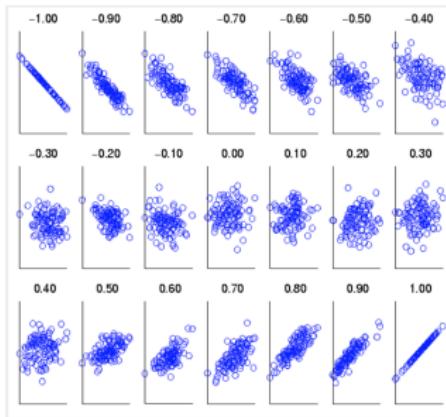
Fact 3

For $a > 0$,

$$\rho(aX + b, Y) = \rho(X, Y)$$



Properties of Correlation



Important Takeaway:

$\rho(X, Y)$ measures the **strength** and **direction** of a **linear** relationship between X and Y (close to 1: strong linear, positive slope; close to -1 : strong linear, negative slope).



Empirical Correlation

Setting: Observe n values (x_1, \dots, x_n) and (y_1, \dots, y_n)

- Empirical Correlation r_{xy} :

$$\begin{aligned} r_{xy} &= \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \end{aligned}$$

where \bar{x}, \bar{y} are sample means, s_x, s_y are sample standard deviations and s_{xy} is the sample covariance between x and y

- This is how we *quantify* the linear relationship between two variables!

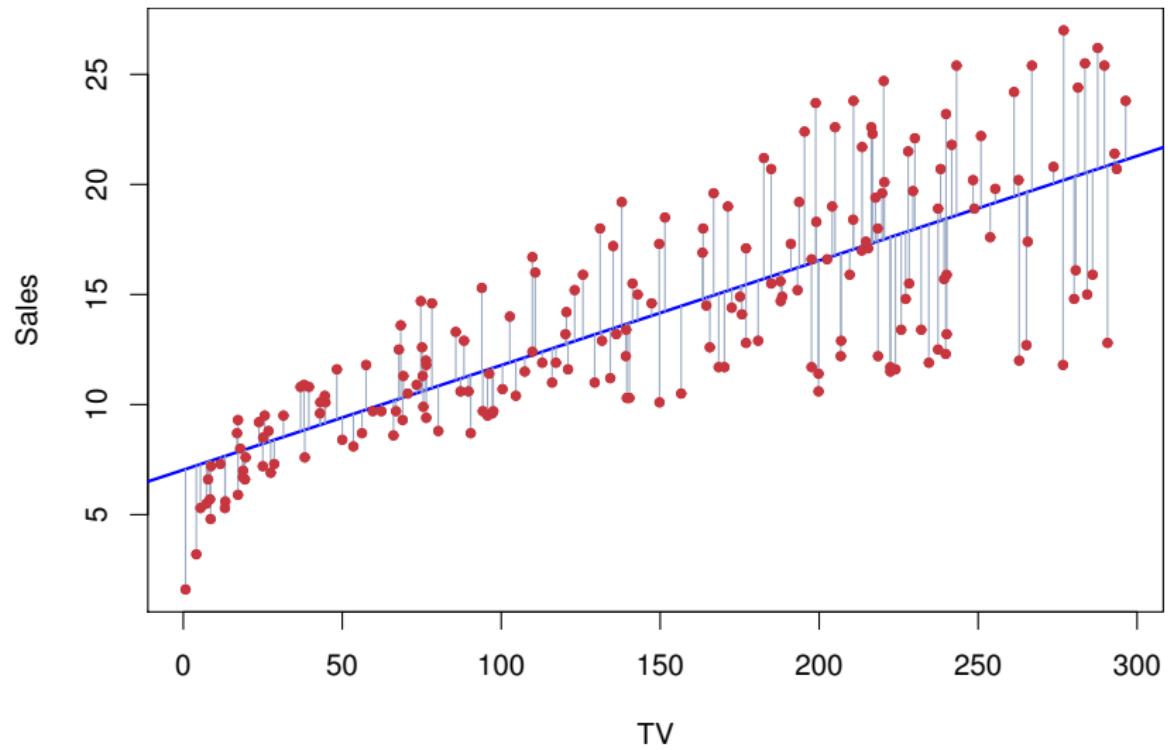


Once we have an idea that a regression model seems reasonable (from correlation values), we can then proceed to fit the model.

- Regression models postulate that:
 - There is a probability distribution of Y for each level of X
 - The means of these probability distributions vary in some systematic fashion with X



Example





Caution: Regression and Causality

- The existence of a statistical relationship between the response variable Y and the explanatory variable X **does not** imply in any way that Y depends *causally* on X
- No matter how strong the relationship between Y and X , no cause and effect pattern is implied by a regression model



The SLR Model

- **Data:**

- Explanatory variable: (x_1, \dots, x_n)
- Response: (y_1, \dots, y_n)

SLR Model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \dots, n$$

- **Coefficients:** β_0 = y -intercept and β_1 = slope of the regression function
- **Error:** ε_i is a random error term for observation i



Possible Assumptions on ε

- ε is a random error term with

- $\mathbb{E}[\varepsilon_i] = 0$
- $\text{Var}(\varepsilon_i) = \sigma^2 \geq 0$
- $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$

or

- $\varepsilon_1, \dots, \varepsilon_n \stackrel{i.i.d}{\sim} N(0, \sigma^2)$

The second assumption is *stronger* than the first. It is required for *inference* but not *estimation*.



Why Simple? Why Linear? [REVISITED]

- **Simple**: only one explanatory variable, X
- **Linear**:
 - linear in the parameters because no parameters appear as exponents or are multiplied or divided by another parameter
 - linear in the explanatory variable because this variable appears only in the first power
 - a model that is linear in both the parameters and the variables is called a *first-order model*



Important Features of the Model

Expectation

Taking expectations we obtain

$$\mathbb{E}[Y_i] = \mathbb{E}[\beta_0 + \beta_1 x_i + \varepsilon_i] = \beta_0 + \beta_1 x_i$$

Variance

Taking the variance of Y_i , we obtain:

$$\text{Var}(Y_i) = \text{Var}(\beta_0 + \beta_1 x_i + \varepsilon_i) = \sigma^2$$

Why are the above two statements true?

Conclusion: The probability distribution of Y have the same variance, regardless of value X



Important Features of the Model

Correlation

Given the assumption that the error terms are *uncorrelated*, i.e.,

$$\text{Corr}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i, j; i \neq j$$

it follows that the responses Y_i and Y_j are also *uncorrelated*

$$\text{Corr}(Y_i, Y_j) = 0 \quad \forall i \neq j$$



Example

- An electrical distributor is studying the relationship between the number of bids requested by construction contractors for basic lighting equipment during a week (x) and the time required to prepare those bids (Y), and the regression model is:

$$Y_i = 9.5 + 2.1x_i + \varepsilon_i$$

and the regression function is

$$\mathbb{E}[Y] = 9.5 + 2.1x_i$$

- Suppose in the i^{th} week, $x_i = 45$ bids are prepared and the actual number of hours required is $Y_i = 108$, which implies that the error term value is $\varepsilon_i = 4$, as $9.5 + 2.1(45) = 104$



Example

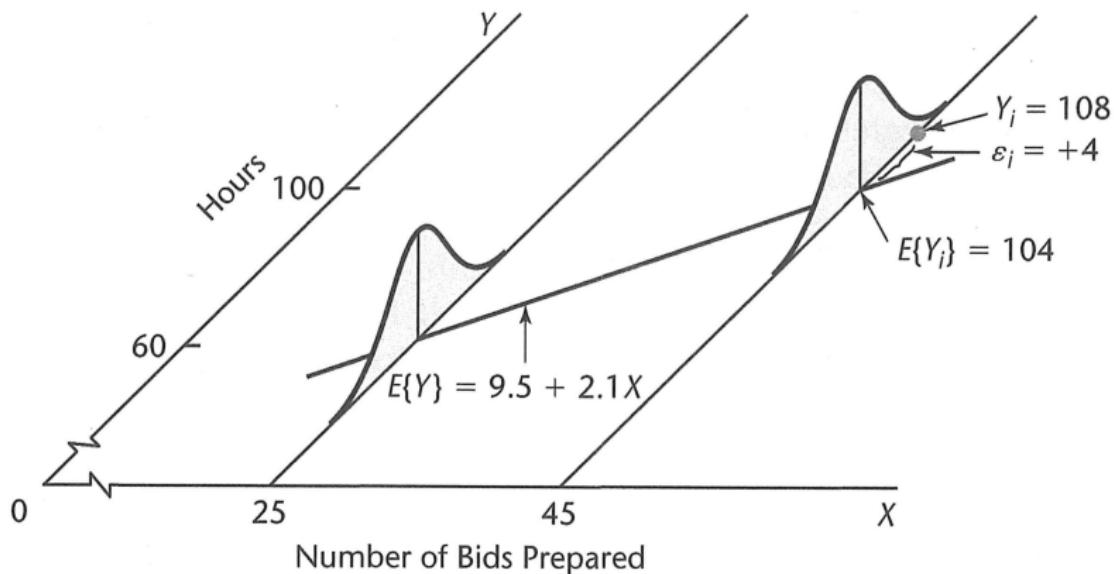


Figure: Illustration of an SLR Model



Example

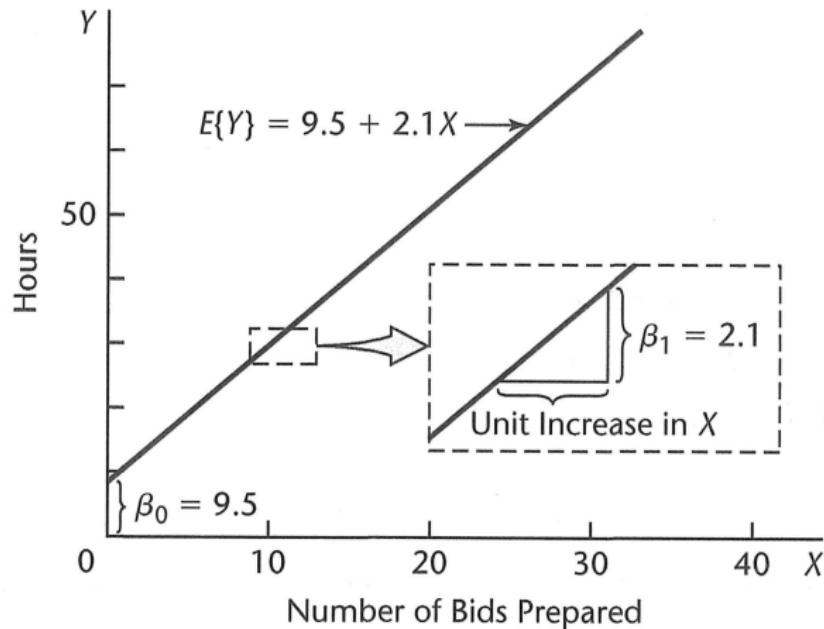


Figure: Meaning of the SLR Parameters



Population Regression Function

The population regression function expresses the mean (but unknown) value of Y given the data x :

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 x$$

Properties

- Linear function of x
- A one-unit increase in explanatory variable x changes the expected value of the response variable y by the amount β_1



Estimating β_0 and β_1

- Recall in general, we would like to minimize the mean squared error:

$$MSE(f) = \frac{1}{n} \sum_{i=1}^n (Y_i - f(x_i))^2$$

- In the SLR setting, we have

$$f(x_i) = \mathbb{E}[Y_i | x_i] = \beta_0 + \beta_1 x_i$$

- Thus, to estimate β_0 and β_1 we find $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimize:

$$S(\beta) = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 x_i))^2$$

- This is called the method of least squares



Method of Least Squares

Goal: Minimize $S(\beta) = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 x_i))^2$ with respect to β

- Setting $\beta = (\beta_0, \beta_1)^T$ and

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

we can write

$$S(\beta) = (Y - X\beta)^T (y - X\beta)$$

- In general, $S(\beta)$ can always be written in the matrix form above



Method of Least Squares

Facts: (test in the 1-dimensional case)

- $\frac{\partial}{\partial \beta} X\beta = X$
- $\frac{\partial}{\partial \beta} \beta^T (X^T X) \beta = 2X^T X\beta$

We can now estimate β by solving the following:

$$\frac{\partial}{\partial \beta} S(\beta) = 0$$

Doing so gives the **normal equations**:

$$X^T y = X^T X\beta$$



Method of Least Squares

- If $X^T X$ is invertible, then the normal equations yield the least squares estimates:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- $\hat{\beta}$ is often referred to as the **ordinary least squares** (OLS) estimates
- In the simple linear regression model, $\hat{\beta}$ simplifies to:

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

- $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$



The Gauss-Markov Theorem

Theorem

Under either assumption of the linear regression model, the least-squares estimators are unbiased and have a minimum variance among all unbiased linear estimators. That is,

- $\mathbb{E}[\hat{\beta}] = \beta$
- *For all linear estimators b such that $\mathbb{E}[b] = \beta$, we have:*

$$\text{Var}(\hat{\beta}) \leq \text{Var}(b)$$

We say that $\hat{\beta}$ is the **best linear unbiased estimator (BLUE)** of β .



Properties of $\hat{\beta}$

- Among other unbiased linear estimators, $\hat{y} = X\hat{\beta}$ will have the smallest MSE (at least for the training set!)
- $\mathbb{E}[\hat{\beta}] = \beta$
- $\text{Var}(\hat{\beta}) = \sigma^2(X^T X)^{-1}$

Question: How do we prove the above 2 statements?



Fitted Values and Residuals

- With $\hat{\beta}$ in hand, we can now estimate y with the **fitted value** $\hat{y} = X\hat{\beta}$
- In SLR, the i th fitted value is given by:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

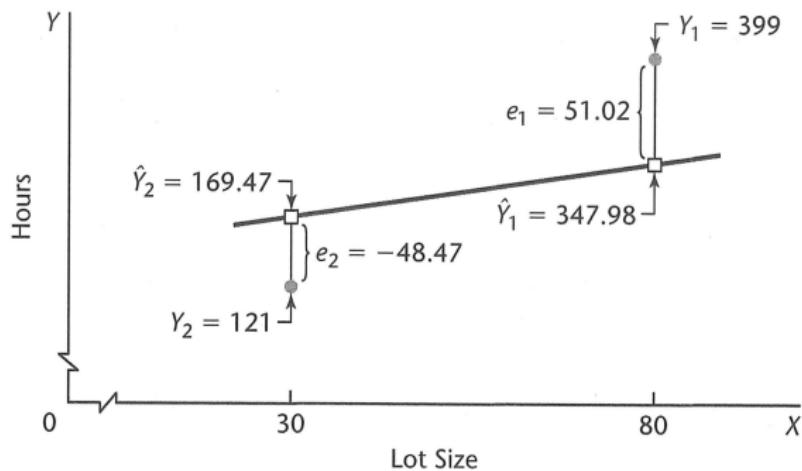
- The **residual** for the i th observation is defined as:

$$e_i = y_i - \hat{y}_i$$



Residuals

The residual e_i is the vertical deviation of y_i from the fitted value \hat{y}_i on the estimated regression line.





Comments on SLR

- Evaluating features of the residuals e_1, \dots, e_n is how we formally *test* the assumptions of the regression model (more on this next time!)
- Estimation *did not* depend on any assumptions of ε !
- Estimation *did* depend on $X^T X$ being invertible, which requires-
 - X must be of rank p
 - This means that the columns of X must be linearly independent (i.e. no multicollinearity!)



Next Time

- Properties of the residuals
- Point estimation for σ^2
- Distributional theory for Normal random variables
- Hypothesis testing and confidence intervals for estimated coefficients