



MULTIPLE REGRESSION & GAUSS-MARKOV THEOREM

11/07/2022

Virginia Institute for Psychiatric and Behavioral Genetics

Linear regression recap

Multiple regression

Gauss-Markov Theorem

LINEAR REGRESSION RECAP



- Technique used for the modeling and analysis of numerical data¹
- 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

¹Akey, 2020, "Regression," (Washington 2020).

- **IT IS ALL ABOUT DESCRIBING RELATIONSHIPS BETWEEN VARIABLES**



Bio

- 30 April 1777 - 23 April 1855
- Worked in the theorem around 1794
- 17 years old!^a

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

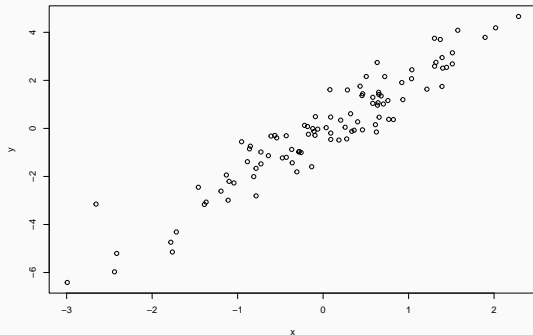
Carl Friedrich Gauss



BEFORE HIM, DESCRIPTION OF RELATIONSHIP WAS NOT SYSTEMATIC



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





$$Y = X1 + X2 + X3$$

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

² Joshua Akey, "Regression."

³ Joshua Akey, "Regression."



- Suppose we want to model the dependent variable Y in terms of three predictors, X_1, X_2, X_3 ⁴

$$Y = f(X_1, X_2, X_3)$$

- We usually have to assume that it has some restricted form, such as linear + error

$$Y = X_1 + X_2 + X_3 + \epsilon$$

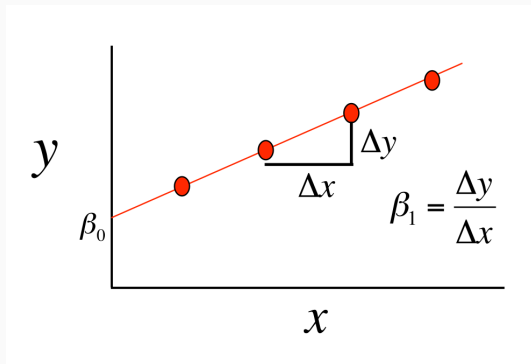
⁴Ibid.



- 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.



^aIbid.



- There exists parameters β_0 , β_1 , and σ^2 , such that for any fixed value of the independent variable x , the dependent variable is related to x through the model equation^a

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

- The error term ϵ is a random variable with mean 0 and constant variance σ^2

^aIbid.

Or, more succinctly,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

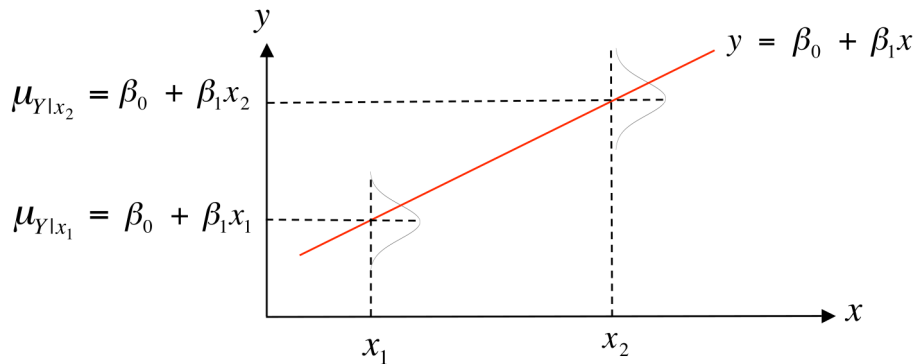


- The **expected** value of Y is a linear function of X , but for fixed x , the variable Y differs from its expected value by a random amount⁵
- Formally, let x^* denote a particular value of the independent variable x , then our linear probabilistic model says:

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

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

⁵Wilber, "Linear Regression," *MLU-Explain*.

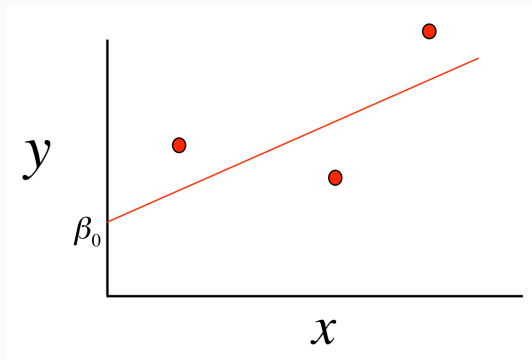


- For example, if x = height and y = weight then $\mu_{Y|x^*} = 60$ is the average weight for all individuals 60 inches tall in the population



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

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





- **Predicted** or fitted, values of y predicted by the least-squares regression line obtained by plugging in x_1, x_2, \dots, x_n into the estimated regression line

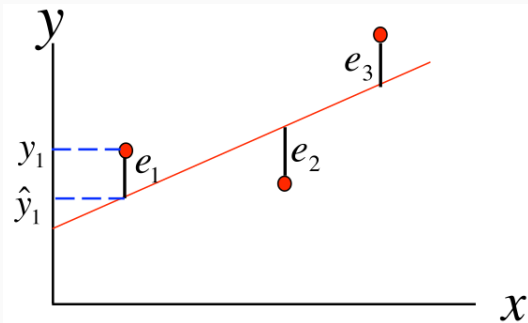
$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

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

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

$$e_1 = y_1 - \hat{y}_1$$

$$e_2 = y_2 - \hat{y}_2$$





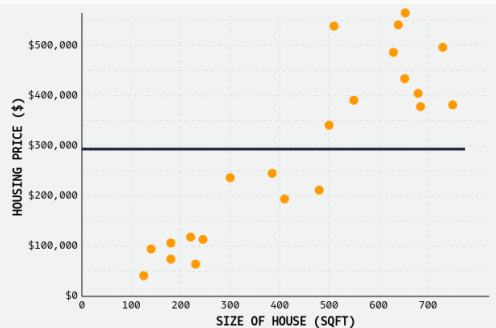
From a bad model⁶

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

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

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:

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



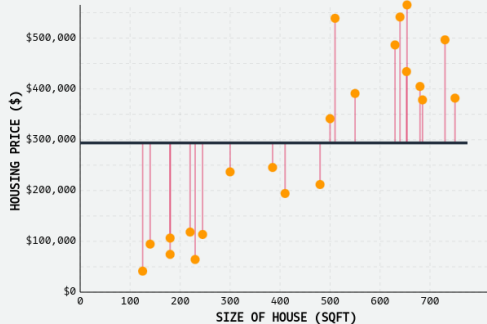
⁶Ibid.

A SUBSTANTIVE EXAMPLE



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

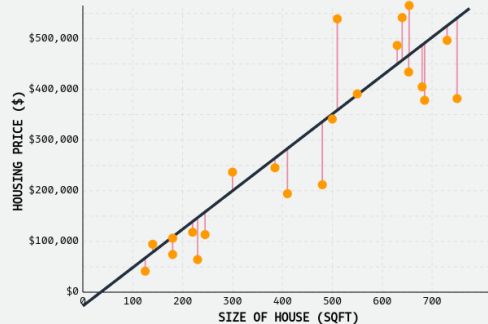


A SUBSTANTIVE EXAMPLE



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.



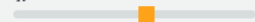


To the best possible model⁷

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:

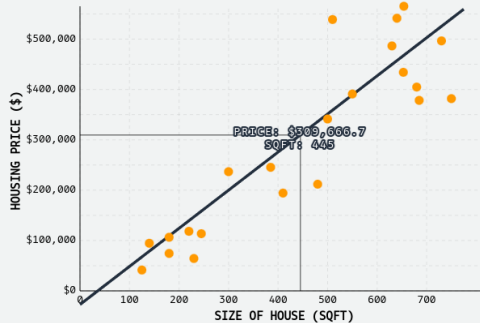
sqft Value: 445



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

$$\hat{y} = 309667$$

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



⁷ Ibid.



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

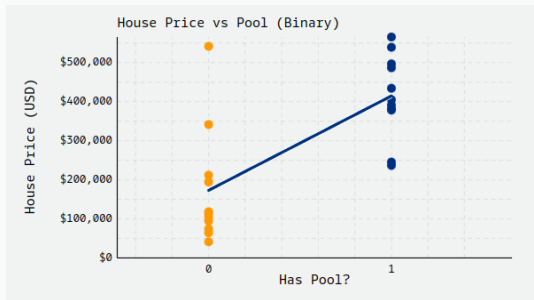
$$SSE = \sum_{i=1}^n (e_i)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Which in turn allows us to estimate σ^2 :

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

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

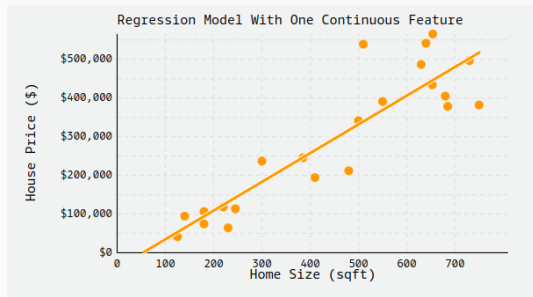
$$R^2 = 1 - \frac{SSE}{SST}; SST = \sum_{i=1}^n (y_i - \bar{y})^2$$



$$\text{house price} = 172893 + 241582 * \text{pool}$$

- summarizes the difference in average housing prices between houses with and without pools

- The intercept, \$172,893, is the average predicted price for houses that do not have swimming pools.
- To find the average price predicted price for houses with pools, we simply plug in $\text{pool}=1$ to obtain $\$172,893 + \$241,582 * 1 = \$414,475$.
- The difference between these two subpopulation means is equal to the coefficient on pool. Houses with pools cost \$241,582 more on average than houses that do not have pools.



house price = $-39591 + 742 \times \text{sqft}$

- summarizes the average house prices across differently sized houses as measured in square feet.

- The coefficient, \$742, represents the average difference in housing price for one-unit difference in the square-footage of the house. In other words, we expect each additional square-foot, on average, to raise the price of a house by \$742.
- The intercept, -\$39,591, represents the predicted housing price for houses with $\text{sqft} = 0$, that is, it represents the average price of a zero square-foot house. Because this value doesn't make much intuitive sense, it's common for models to be transformed and standardized before carrying out a regression model.

MULTIPLE REGRESSION



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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

- Expression = Baseline + Age + Tissue + Sex + Error
- Partial Regression Coefficients: effect on the dependent variable when increasing the i^{th} independent variable by 1 unit, **holding all other predictors constant**



- 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 levels:

$$x_1 = \begin{cases} 0 & \text{if AA} \\ 1 & \text{if AG} \\ 2 & \text{if GG} \end{cases}$$



- Previous coding would result in **colinearity**
- Solution is to set up a series of dummy variable.
- for k levels you need k-1 dummy variables

	x1	x2
AA	1	0
AG	0	1
GG	0	0

$$x_1 = \begin{cases} 1 & \text{if AA} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if AG} \\ 0 & \text{otherwise} \end{cases}$$



Validity Does the data we're modeling matches to the problem we're actually trying to solve?

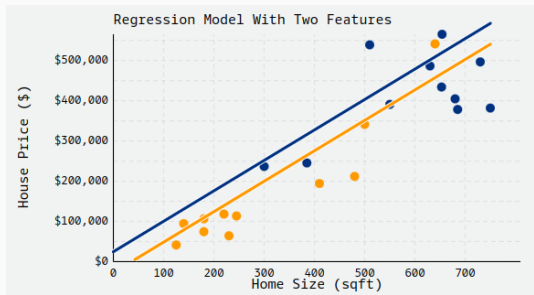
Representativeness Is the sample data used to train the regression model representative of the population to which it will be applied?

Additivity and Linearity The deterministic component of a regression model is a linear function of the separate predictors: $y = B_0 + B_1x_1 + \dots + B_px_p$

Independence of Errors The errors from our model are independent.

Homoscedasticity The errors from our model have equal variance.

Normality of Errors The errors from our model are normally distributed.



- intercept: $-\$27,154$, the predicted average housing price for houses with all $x_i = 0$. Or the cost of houses with no pools and a square-footage of zero.
- coefficient of pool: $\$51,867$, average expected price difference in houses of the same size (in sqft) if they do or do not have a pool. In other words, we expect, on average, houses of the same size to cost $\$51,867$ more if they have a pool than if they do not.
- coefficient of sqft: $\$757$, average expected price difference in housing price for houses that have the same value of pool but differ in size by one square-foot.
- We assume the same slope for sqft. Hence, two lines. This isn't always a valid assumption to make.

house price = $-27154 + 757 \text{sqft} + 51867 \text{pool}$

- In our example, we model home prices as a function of both the size of the house (sqft) and whether or not it has a pool



- It is possible to prove that OLS is precise and optimal (BLUE)

Best Linear Unbiased Estimator (BLUE)

1. The parameters are *linear*
2. The parameters are *unbiased*
3. The parameters are *efficient*. In other words, they have the least variance of all unbiased linear estimators, *best*.



$$Y_1 = B_1 + B_2 X_{21} + B_3 X_{31} + \dots + B_k X_{k1} + u_1$$

$$Y_2 = B_1 + B_2 X_{22} + B_3 X_{32} + \dots + B_k X_{k2} + u_2$$

$$\vdots$$

$$Y_n = B_1 + B_2 X_{2n} + B_3 X_{3n} + \dots + B_k X_{kn} + u_n$$

where:

- y is an $N \times 1$ vector of observations of the output variable (N is the sample size);
- X is an $N \times K$ matrix of inputs (K is the number of inputs for each observation);
- B is a $K \times 1$ vector of regression coefficients;
- u is an $N \times 1$ vector of errors.^a

^aTaboga, "Gauss Markov Theorem," in, *Lectures on probability theory and mathematical statistics* (2021).

Or more succinctly:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{pmatrix} 1 & X_{21} X_{31} \cdots & X_{k1} \\ \vdots & \ddots & \vdots \\ 1 & X_{2n} X_{3n} \cdots & X_{kn} \end{pmatrix} \begin{bmatrix} B_1 \\ B_2 \\ B_3 \\ \vdots \\ B_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$



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

Arriving at this expression⁸

OLS is BLUE if it satisfies the assumptions

- Assumption 1: Linearity. The regression model is linear in the parameters; it may or may not be linear in the variables, the Ys and Xs.

⁸Gujarati, "The Classical Linear Regression Model (CLRM)," in, *Linear Regression: A Mathematical Introduction* (2019).



- The regressors are assumed fixed, or nonstochastic, in the sense that their values are fixed in repeated sampling. However, if the regressors are stochastic, we assume that each regressor is independent of the error term or at least uncorrelated with it.
- X is fixed and exogenous
- X is measured without error



- Given the values of the X variables, the expected, or mean, value of the error term u is 0.

$$E(\epsilon|X) = 0$$

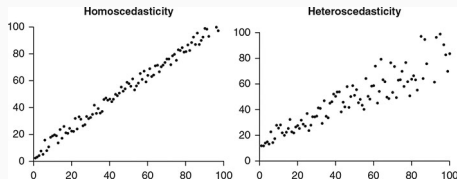
$$E \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} E(u_1) \\ E(u_2) \\ E(u_3) \\ \vdots \\ E(u_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

ASSUMPTION 4: HOMOSCEDASTICITY



- The variance of each u , given the values of X , is constant or homoscedastic (i.e., of equal variance).

$$\text{var}(u_i|X) = E(uu') = \sigma^2 I$$





- There is no correlation between error terms belonging to two different observations.

$$\text{cov}(u_i, u_j | X) = 0$$

Assumption 6: There is no perfect linear relationship among the X variables. This is the assumption of no multicollinearity.



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

First of all, note that $\hat{\beta}$ is linear in y . In fact, $\hat{\beta}$ is the product between the $K \times N$ matrix $(X'X)^{-1}X'$ and y , and matrix multiplication is a linear operation.⁹

⁹Marcos Taboga, "Gauss Markov Theorem."



- The true relationship between x and y is: $y = x\beta + \epsilon$.
- Substituting \mathbf{y} into the equation for $\hat{\beta}$, we get:

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}X'(X\beta + \epsilon) \\ &= (X'X)^{-1}X'X'(X\beta + \epsilon) \\ &= (X'X)^{-1}X'X'X\beta + (X'X)^{-1}X'\epsilon \\ &= I\beta + (X'X)^{-1}X'\epsilon\end{aligned}$$

$$\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon$$

Take the expected value of the equation:

$$\begin{aligned}E(\hat{\beta}) &= E(\beta + (X'X)^{-1}X'\epsilon) \\ &= \beta + (X'X)^{-1}X'E(\epsilon); \text{ since } E(u) = 0 \\ &= \beta\end{aligned}$$

- Because the expected value of the estimate of β is equal to the true value of β , the estimator is unbiased



$$b^* = \left[A + (X'X)^{-1}X' \right] y$$

where A is some nonstochastic $k \times n$ matrix, similar to X . Simplifying, we obtain

$$\begin{aligned} b^* &= Ay + (X'X)^{-1}X'y \\ &= Ay + b \end{aligned}$$

where b is the least-squares estimator

Now,

$$\begin{aligned} E(b^*) &= \left[A + (X'X)^{-1}X' \right] E(y) \\ &= \left[A + (X'X)^{-1}X' \right] (XB) \\ &= (AX + I)B \end{aligned}$$

Now $E(b^*) = B$ if and only if $AX = 0$. In other words, for the linear estimator b^* to be unbiased, AX must be 0.



thus,

$$\begin{aligned} b^* &= [a + (x'x)^{-1}x'] [xb + u], \text{ substituting for } (y) \\ &= b + [a + (x'x)x - 1]u, \text{ because } ax = 0 \end{aligned}$$

given that u has zero mean and constant variance ($= \sigma^2 i$), we can now find the variance of b^* as follows:

$$\begin{aligned} \text{cov}(b^*) &= e [a + (x'x)^{-1}x'] uu' [a + (x'x)^{-1}x']' \\ &= [a + (x'x)^{-1}x'] e(uu') [a + (x'x)^{-1}x']' \\ &= \sigma^2 [aa' + (x'x)^{-1}] \\ &= \sigma^2 (x'x)^{-1} + aa'\sigma^2 \\ &= \text{var}(b) + aa'\sigma^2 \end{aligned}$$

- shows that the covariance matrix of b^* is equal to the covariance matrix of b plus a positive semidefinite matrix



1. The parameters are *linear*

$$(X'X)^{-1}X'$$

2. The parameters are *unbiased*

$$E(b) = B$$

3. The parameters have the least variance of all unbiased linear estimators, *best*.

$$\hat{\beta}^* = \text{var} - \text{cov}(\hat{\beta}) + \sigma^2 AA'$$

- **ALT-TAB TO EXCEL, LINEAR ESTIMATOR EXAMPLE**



- Let's quickly check the properties of OLS, Ridge regression and Quantile regression^a
- Our data generating process to the right

^aPytko, 2020, "Statistical Learning and Big Data in R (Lecture 2)," *Gist*, (2020).

```
#### define our Data Generating Process
alpha <- 25
beta <- 12
N <- 25

n_simulations <- 100000

# expanding the vectors to the needed size
slope_ols <- numeric(n_simulations)
slope_ols2 <- numeric(n_simulations)
slope_orr <- numeric(n_simulations)
slope_LAD <- numeric(n_simulations)

# simulate with uniform
x <- runif(N, min=0, max=10)

# including intercepts
X <- cbind(1, x)

# our true model
y <- alpha + beta*x + rnorm(N, mean=0, sd=5)

# What is the true relationship?
plot(y ~ x)
abline(a=alpha, b=beta, col="orange")
```

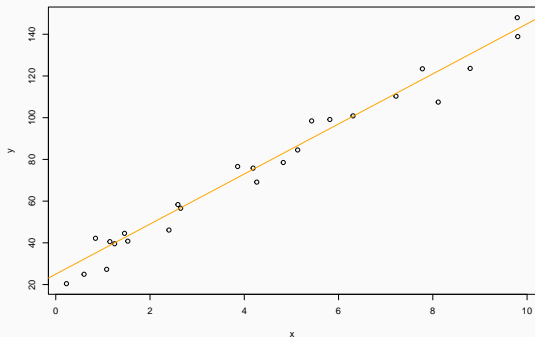


Figure 1: True relationship between x and y

```
#### define our Data Generating Process
alpha <- 25
beta <- 12
N <- 25

n_simulations <- 100000

# expanding the vectors to the needed size
slope_ols <- numeric(n_simulations)
slope_ols2 <- numeric(n_simulations)
slope_orr <- numeric(n_simulations)
slope_LAD <- numeric(n_simulations)

# simulate with uniform
x <- runif(N, min=0, max=10)

# including intercepts
X <- cbind(1, x)

# our true model
y <- alpha + beta*x + rnorm(N, mean=0, sd=5)

# What is the true relationship?
plot(y ~ x)
abline(a=alpha, b=beta, col="orange")
```



Any familiar?

```
unnname(coef(lm(y~x))[2])
```

[1] 12.1

```
(ginv(t(X)%*%X)%*%(t(X)%*%y))[2,1]
```

[1] 12.1

```
(ginv(t(X)%*%X+1.5*diag(2))%*%(t(X)%*%y))[2,1]
```

[1] 12.6

```
unnname(coef(rq(y~x))[2])
```

[1] 11.9



$$(X'X)^{-1}X'y$$

```
ginv(t(X)%*%X)%*%(t(X)%*%y)
```

Ridge regression estimator

$$(X'X + \lambda I)^{-1}X'$$

```
ginv(t(X)%*%X+1.5*diag(2))%*%(t(X)%*%y)
```

- Lambda arbitrary for the example



```
for(i in 2:n_simulations)
{
  y <- alpha + beta*x + rnorm(N, mean=0, sd=5)

  slope_ols[i] <- unname(coef(lm(y~x))[2])
  slope_ols2[i] <- (ginv(t(X)%*%X)%*%(t(X)%*%y))[2,1]
  slope_orr[i] <- (ginv(t(X)%*%X+1.5*diag(2))%*%(t(X)%*%y))[2,1]
  slope_LAD[i] <- unname(coef(rq(y~x))[2])
}
```



```
mean(slope_ols)
```

```
[1] 12
```

```
mean(slope_ols2)
```

```
[1] 12
```

```
mean(slope_orr)
```

```
mean(slope_LAD)
```

```
[1] 13.3
```

```
[1] 12
```



```
A <- qplot(slope_olsP$slope_ols, geom="histogram",
           fill=I("mediumpurple"), xlim=c(10.5, 13.5)) +
  theme_luis()
B <- qplot(slope_LADP$slope_LAD, geom="histogram",
           fill=I("mediumpurple"), xlim=c(10.5, 13.5)) +
  theme_luis()
C <- qplot(slope_orrP$slope_orr, geom="histogram",
           fill=I("mediumpurple"), xlim=c(10.5, 13.5)) +
  theme_luis()
```

```
var(slope_LADP$slope_LAD)/var(slope_olsP$slope_ols)
```

```
[1] 1.53
```

```
var(slope_orrP$slope_orr)/var(slope_olsP$slope_ols)
```

```
[1] 0.573
```

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

Warning: Removed 1 rows containing non values (stat_bin).

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

Warning: Removed 34 rows containing no values (stat_bin).

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.



- We saw a review of linear regression
- How multivariate regression works
- And how OLS is the best of the linear estimators



Team

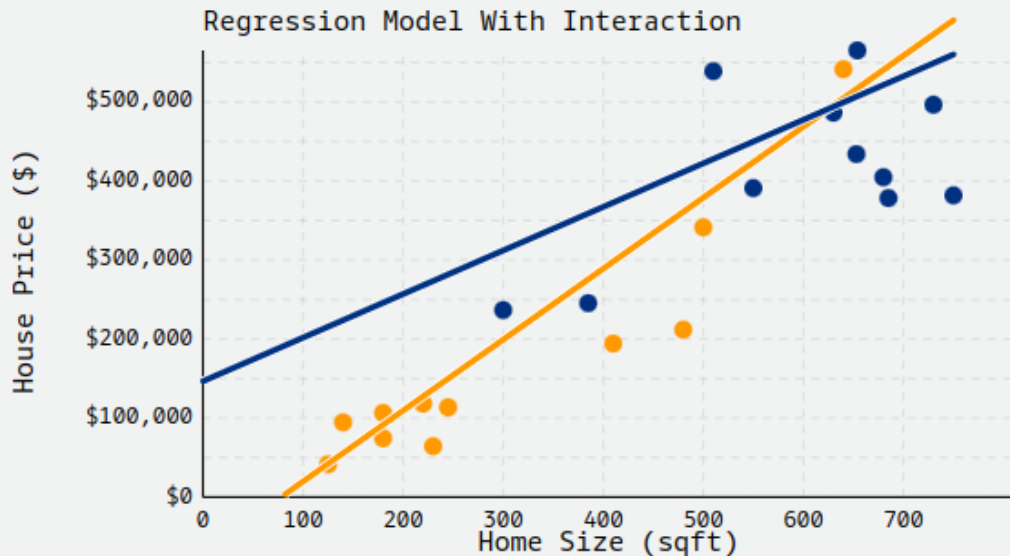
- Joshua Pritkin.
- Rob Kirkpatrick.
- Michael C Neale.
- NIH grant no R01 DA049867 and 5T32MH-020030

Contact



- **THANK YOU**

INTERPRETATION 4





Now that we have shown that the OLS estimator is linear and unbiased, we need to prove that it is also the best linear unbiased estimator.

What exactly do we mean by best?

When $\hat{\beta}$ is a scalar (i.e., there is only one regressor), we consider $\hat{\beta}$ to be the best among those we are considering (i.e., among all the linear unbiased estimators) if and only if it has the smallest possible variance, that is, if its deviations from the true value β tend to be the smallest on average. Thus, $\hat{\beta}$ is the best linear unbiased estimator (BLUE) if and only if

$$\text{Var}[\hat{\beta}|X] \leq \text{Var}[\tilde{\beta}|X]$$

for any other linear unbiased estimator $\tilde{\beta}$.



Since we often deal with more than one regressor, we have to extend this definition to a multivariate context. We do this by requiring that

$$\text{Var}[\alpha\hat{\beta}|X] \leq \text{Var}[\alpha\tilde{\beta}|X]$$

for any $1 \times K$ constant vector α , any other linear unbiased estimator $\tilde{\beta}$.

In other words, OLS is BLUE if and only if any linear combination of the regression coefficients is estimated more precisely by OLS than by any other linear unbiased estimator.



Condition (1, previous) is satisfied if and only if

$$\text{Var}[\tilde{\beta}|X] - \text{Var}[\hat{\beta}|X]$$

is a positive semi-definite matrix.

In the next two sections we will derive $\text{Var}[\hat{\beta}|X]$ (the covariance matrix of the OLS estimator), and then we will prove that (2, above) is positive-semidefinite, so that OLS is BLUE.



The conditional covariance matrix of the OLS estimator is

$$\text{Var}[\hat{\beta}|X] = \sigma^2(X'X)^{-1}$$



Since we are considering the set of linear estimators, we can write any estimator in this set as

$$\tilde{\beta} = Cy$$

where C is a $K \times N$ matrix.

Furthermore, if we define

$$D = C - (X'X)^{-1}X'$$



then we can write

$$\begin{aligned}\tilde{\beta} &= Cy \\ &= Dy + (X'X)^{-1}X'y \\ &= Dy + \hat{\beta}\end{aligned}$$

It is possible to prove that $DX = 0$ if $\tilde{\beta}$ is unbiased.



By using this result, we can also prove that

$$\text{Var}[\hat{\beta}|X] = \text{Var}[\tilde{\beta}|X] + \sigma^2 DD'$$

As a consequence,

$$\text{Var}[\hat{\beta}|X] - \text{Var}[\tilde{\beta}|X] + \sigma^2 DD'$$

is positive semi-definite because [eq28] is positive semi-definite. This is true for any unbiased linear estimator $\tilde{\beta}$. Therefore, the OLS estimator is BLUE.



Thus,

$$\begin{aligned} b^* &= [A + (X'X)^{-1}X'] [XB + u], \text{ substituting for } (y) \\ &= B + [A + (X'X)^{-1}X']u, \text{ because } AX = 0 \end{aligned}$$

Given that u has zero mean and constant variance ($= \sigma^2 I$), we can now find the variance of b^* as follows:

$$\begin{aligned} \text{cov}(b^*) &= E [A + (X'X)^{-1}X'] uu' [A + (X'X)^{-1}X']' \\ &= [A + (X'X)^{-1}X'] E(uu') [A + (X'X)^{-1}X']' \\ &= \sigma^2 [AA' + (X'X)^{-1}] \\ &= \sigma^2 (X'X)^{-1} + AA'\sigma^2 \\ &= \text{var}(b) + AA'\sigma^2 \end{aligned}$$

- shows that the covariance matrix of b^* is equal to the covariance matrix of b plus a positive semidefinite matrix



$$\begin{aligned}b &= (X'X)^{-1}X'y \\&= (X'X)^{-1}X'[XB + u], \text{ substituting for } y \\&= (X'X)^{-1}X'XB + (X'X)^{-1}X'u \\&= B + (X'X)^{-1}X'u\end{aligned}$$

Now,

$$\begin{aligned}E(b) &= B + (X'X)^{-1}X'E(u) \\&= B\end{aligned}$$

In words, the expected value of b is equal to B , thus proving that b is unbiased.

(Recall the definition of unbiased estimator.) Note that $E(u|X) = 0$ by assumption.