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

- 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

Bio

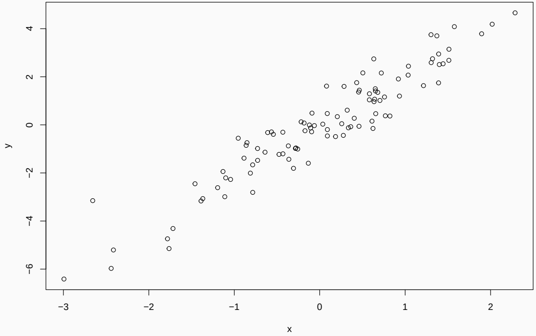
- 30 April 1777 - 23 April 1855
- Worked in the theorem around 1794
- 17 years old![a]

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

Carl Friedrich Gauss

Before him, description of relationship was not systematic

```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.

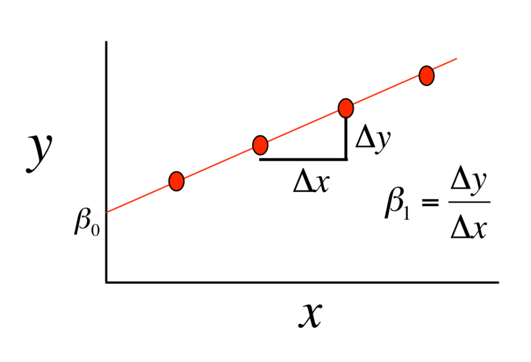
[6] Ibid.

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.

- 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 ϵ is a random variable with mean 0 and constant variance σ^2
  is a rv assumed to be N(0, 2)

[7] Ibid.

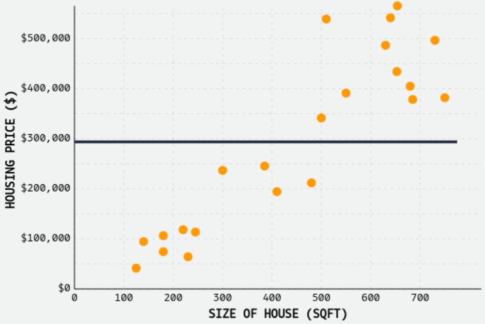
From a bad model[8]

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:

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

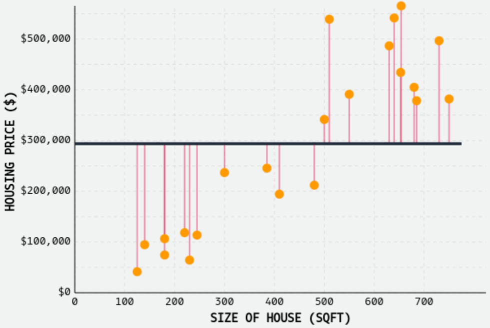


[8] Wilber, "Linear Regression," *MLU-Explain*.

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]

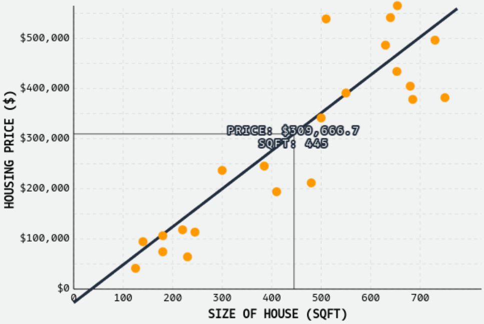
Once we've fit our model, predicting future values is super easy! We just plug in any x_i values into our equation!

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.



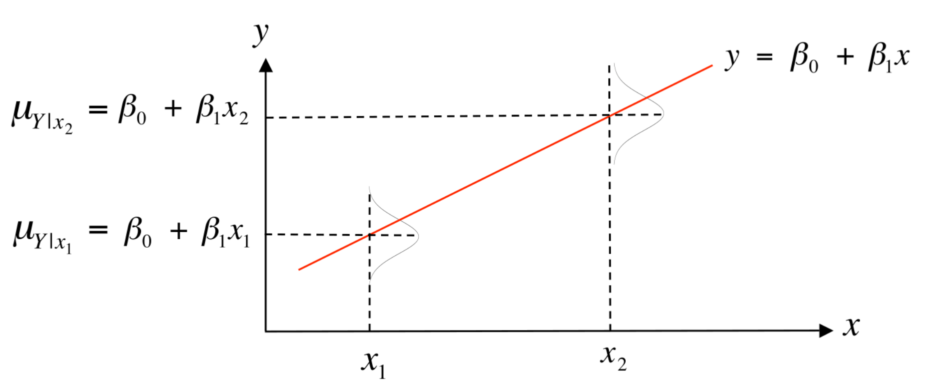
[9] Ibid.

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*



- 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]

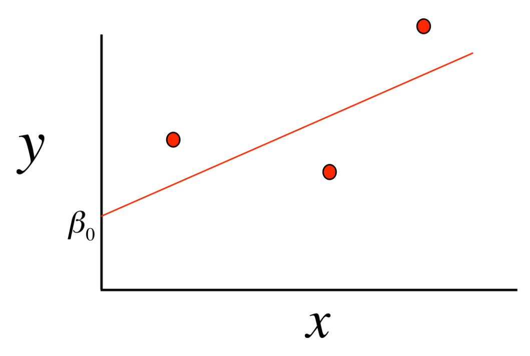
[10] Ibid.

Estimating Model Parameters

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

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

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

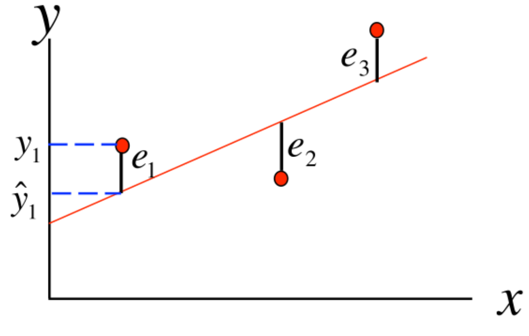


Predicted and Residual Values

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):

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

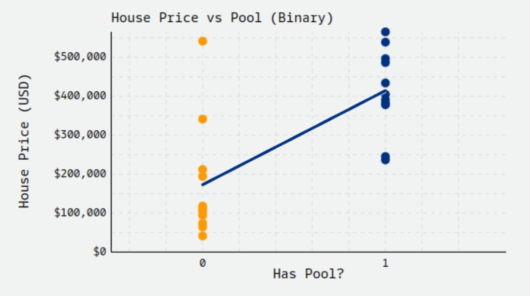
- Which in turn allows us to estimate σ^2:

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

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

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

BINARY VARIABLE

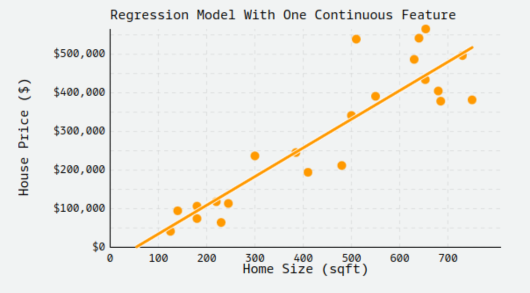


house price = 172893 + 241582 $*$ *pool*

- 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 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.

ONE CONTINUOUS VARIABLE



house price=-39591+742*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 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: ieffect 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:

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

- 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?

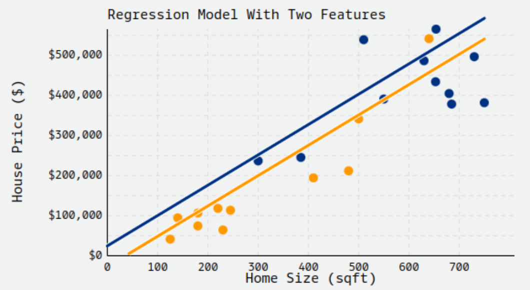
Additivity and Linearity The deterministic component of a regression model is a linear function of the separate predictors: $y = B_0 + B_1x_1 + ... + B_px_p$

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.

Multivariate Regression



house price=-27154+757*sqft+51867*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

- 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.

- Provided all previous assumptions hold
- It is possible to prove that OLS is precise and optimal in a sense
- Which sense?

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*.

$$y = X\beta + \epsilon$$

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);
- β is a $K \times 1$ vector of regression coefficients;
- ϵ is an $N \times 1$ vector of errors.[11]

[11] Taboga, "Gauss Markov Theorem," in, *Lectures on probability theory and mathematical statistics* (2021).

The OLS estimator of β is

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

We assume that:

- X has full-rank (as a consequence, $X'X$ is invertible, and $\widehat{\beta}$ is well-defined);
- ϵ is a random vector with mean zero and covariance matrix $\sigma^2 I$ (where σ^2 is the variance of the errors);
- ϵ is independent of X (i.e., $E(\epsilon|X) = 0$).[12]

Proof not shown[13]

[12] Ibid.

[13] **???**

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

First of all, note that $\widehat{\beta}$ is linear in y. In fact, $\widehat{\beta}$ is the product between the $K \times N$ matrix $(X'X)^{-1}X'$ and y, and matrix multiplication is a linear operation.[14]

[14] Marcos Taboga, "Gauss Markov Theorem."

OLS IS UNBIASED

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

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.

HAVE THE LEAST VARIANCE, BEST

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

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

$$b^* = Ay + (X'X)^{-1}X'y$$

$$= Ay + b$$

where b is the least-squares estimator

Now,

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

$$= \left[A + (X'X)^{-1}X\prime\right] (XB)$$

$$= (AX + I)B$$

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.

Have the least variance, best

Thus,

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

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

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

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

Best Linear Unbiased Estimator (BLUE)

1. The parameters are *linear*

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

2. The parameters are *unbiased*

$$E(\widehat{\beta}|X) = \beta$$

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

$$\widehat{\beta}^{*} = var - cov(\widehat{\beta}) + \sigma^{2}CC'$$

- ALT-TAB to Excel, linear estimator example

Conclusion

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

Acknowledgements

Team

- Joshua Pritkin.
- Rob Kirkpatrick.

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

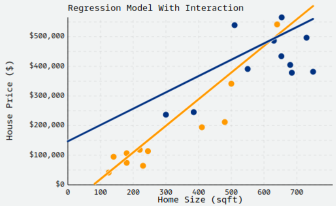
Contact

- Thank you

Interpretation 4

A Binary Feature | A Continuous Feature | Multivariate Regression | Regression with Interaction

A Regression Model With Interaction Terms



Example:
house price $= -70296 + 899 * sqft + 217111 * pool - 347 * (sqft : pool)$

Interpretation: If we believe that the slope for $sqft$ should differ between houses that do have pools and houses that do not, we can add an interaction term to our model, $(sqft : pool)$.

The coefficient of the interaction term $(sqft : pool)$, -\$347, represents the difference in the slope for $sqft$, comparing houses that do and do not have pools. Visually, this represents the difference between the slopes of the two lines, — and —, above.

The intercept, -\$70,296, represents the predicted housing price for houses with no pools and a square-footage of zero.[i]

The coefficient of $pool$, \$217,111, represents the average expected difference in houses of the same size (0 $sqft$) that differed in whether or not they had a pool. (It's not super useful since we don't have houses with 0 square-feet).

The coefficient of $sqft$, \$899, represents the average expected difference in housing price for houses that do not have a pool ($pool = 0$) but differ in size by one square-foot.

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 $\widehat{\beta}$ is a scalar (i.e., there is only one regressor), we consider $\widehat{\beta}$ 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, $\widehat{\beta}$ is the best linear unbiased estimator (BLUE) if and only if

$$Var[\widehat{\beta}|X] \leq Var[\widetilde{\beta}|X]$$

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

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

$$Var[\alpha\widehat{\beta}|X] \leq Var[\alpha\widetilde{\beta}|X]$$

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

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

$$Var[\widetilde{\beta}|X] - Var[\widehat{\beta}|X]$$

is a positive semi-definite matrix.

In the next two sections we will derive $Var[\widehat{\beta}|X]$ (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

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

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

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

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

By using this result, we can also prove that

$$Var[\widehat{\beta}|X] = Var[\widetilde{\beta}|X] + \sigma^2 DD'$$

As a consequence,

$$Var[\widehat{\beta}|X] - Var[\widetilde{\beta}|X] + \sigma^2 DD'$$

is positive semi-definite because [eq28] is positive semi-definite. This is true for any unbiased linear estimator *widetilde*β. Therefore, the OLS estimator is BLUE.