Linear Regression with R/Stata

Day 2: Simple and Multiple Regression

Constantin Manuel Bosancianu

February 26, 2019

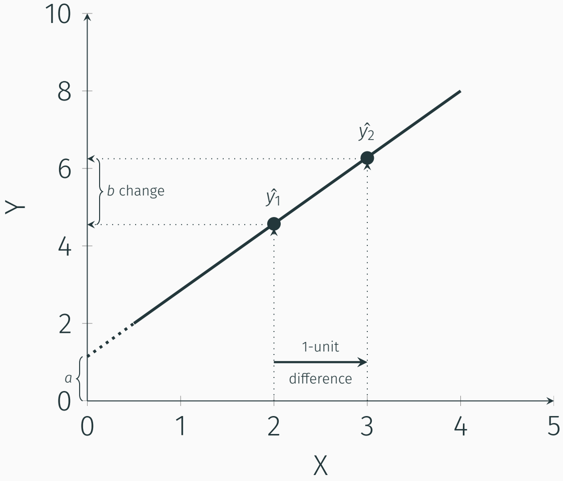
Wissenschaftszentrum Berlin
Institutions and Political Inequality
bosancianu@icloud.com

Preamble

Recap from yesterday

- ✓ Simple regression can be used to summarize a linear relationship between two variables: outcome (Y) and predictor (X);
- ✓ OLS (*ordinary least squares*) is based on the attempt to minimize the sum of the squared distances between line and points (*sum of squared errors*, SSE);
- ✓ The position of the line is given by 2 quantities:
  - ✓ a: the intercept, or the value of the outcome when the predictor is 0;
  - ✓ b: the slope, or the increase in the outcome when predictor increases by 1.

Recap from yesterday



Slope interpretation (the predicted value of Y for a specific value of X, x_i, is $\hat{y}_i$)

Clarification about “increase”

It’s not a temporal increase, e.g. “country X’s GDP increased by 2.3% between 2015 and 2016”.

In our case, it simply means a static comparison of levels.

Think back to yesterday: income inequality and infant mortality. Countries with level $c + 1$ of inequality have, on average, 0.78 more infant deaths per 1000 live births, than countries with c level of inequality.

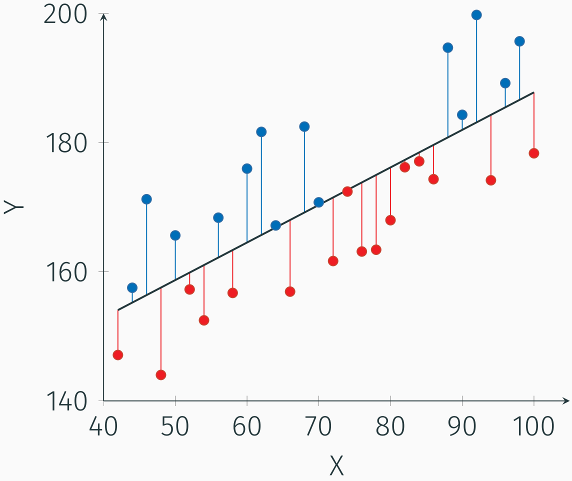
Outline

Today's topics:

- ✓ Assessing how well the line fits the data;
- ✓ From simple to multiple regression: multiple predictors;
- ✓ (if we have time) Inference from sample to population for simple regression.

Model fit

Assessing model fit



Residuals from yesterday (blue is positive, red is negative)

Using the residuals

We can think of model fit as the extent to which the line fits the cloud of points that represents the data.

The residuals are a clear instrument for that.

- ✓ a good fit $\Rightarrow$ the points huddle close around the line $\Rightarrow$ the residuals tend to be small;
- ✓ a bad fit $\Rightarrow$ the points are far away from the line $\Rightarrow$ the residuals tend to be large;

The standard deviation of the residuals is a good measure (a.k.a. *residual standard error* or the *standard error of the regression*).

Residual standard error

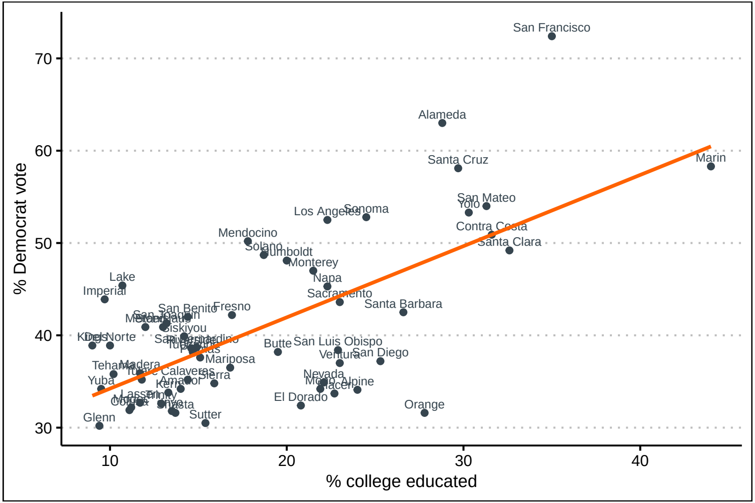
$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-k-1}} \tag{1}$$

e_i are the residuals, n is the sample size, and k is the number of predictors in the regression model. In our case now, $k = 1$.

The e_i are on the same scale as Y. Why? If we denote $a + bY$ by $\hat{Y}$, then $e_i = Y_i - \hat{Y}_i$. Every point on the regression line has coordinates $(X; a + bY)$. $\sqrt{e^2}$ maintains the metric of e.

σ_e can be interpreed as a sort of “average residual”.

Example: California 1992



Education and Democratic support in California, 1992 (county-level data)

Example: California 1992

Residual standard error is 6.76. This means that the “average residual” is about 6.8 percentage points.

In our context, it’s quite a lot of error to have. However, so far we only have one predictor in the model.

It’s an intuitive measure of model fit, that doesn’t get mentioned very often.

The coefficient of determination, R^2

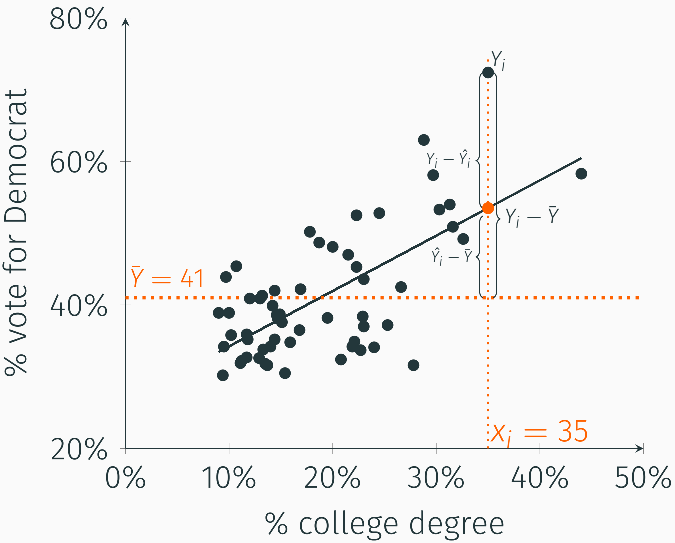
The “standard” measure of fit for OLS regression: R^2.

Has the name R^2 because for simple regression it’s value is the square of Pearson’s r, r^2_{XY}.

It’s (almost) always positive, and ranging between 0 and 1, with higher values meaning a better model fit.

Unfortunately, it’s not very intuitive.

R^2—components



R^2—components

- ✓ $Y_i - \bar{Y}$ = *total* deviation from the mean;
- ✓ $\hat{Y}_i - \bar{Y}$ = *explained* deviation from the mean;
- ✓ $Y_i - \hat{Y}_i$ = *unexplained* deviation out of the total deviation.

Based on these, we can define 3 quantities:

- ✓ $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ = total sum of squared deviations (TSS);
- ✓ $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ = regression (explained) sum of squared deviations (RegSS);
- ✓ $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ = residual (unexplained) sum of squared deviations (RSS).

R^2—components

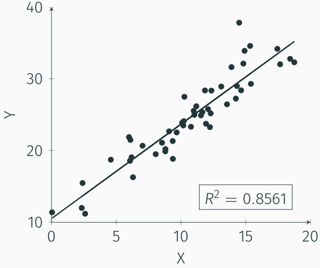
$TSS = RegSS + RSS$

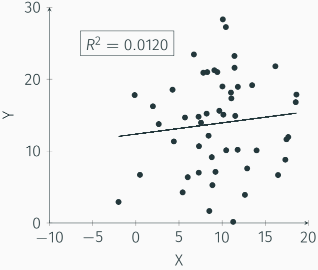
$R^2 = \frac{RegSS}{TSS}$. You can see it as the percentage of the TSS that is explained by our regression model.

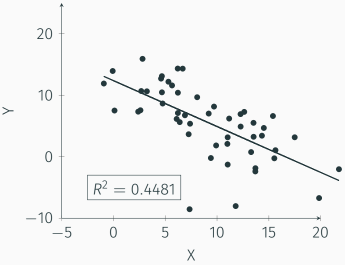
You'll also find it as the share of variance in the outcome explained by our model.

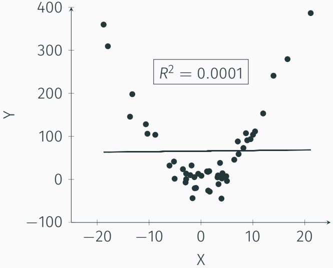
For our regression, R^2 = 0.4398. About 44% of the variance in Democratic vote share is explained by education.

R^2—examples


$R^2 = 0.8561$
Y
X
10
20
30
40
0
5
10
15
20


$R^2 = 0.0120$
Y
X
0
10
20
30
−10
−5
0
5
10
15
20


$R^2 = 0.4481$
Y
X
−10
0
10
20
−5
0
5
10
15
20


$R^2 = 0.0001$
Y
X
−100
0
100
200
300
400
−20
−10
0
10
20

R^2—final considerations

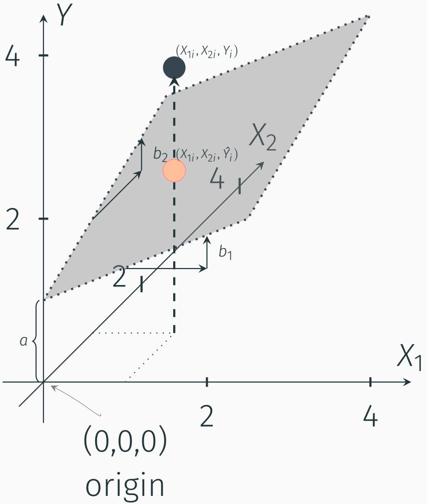
A high R^2 is not the final measure of a model's worth, nor the "be-all and end-all" of regression.

Predict vote in this election with vote in past election, or GDP at time t with GDP at time $t-1$.

R^2 depends on the variation in Y found in the sample $\Rightarrow$ R^2 from different samples can't be compared with each other.

Multiple regression

From line to plane



Two predictors (adapted from Fox, 2008)

Coefficients in multiple regression

$$Y = \underbrace{a}_{\text{intercept}} + \underbrace{b_1}_{\text{slope}} X_1 + \underbrace{b_2}_{\text{slope}} X_2 + \underbrace{e}_{\text{residual}} \tag{2}$$

They are now called *partial regression coefficients*. Interpretation: the effect of a particular variable, *while holding the other variables in the model constant*.

In our example, b_1 is the effect of X_1 on Y, after holding X_2 constant.

Substantively, the interpretation of a and the bs is the same as for simple regression.

Formulas for coefficients

Much more complex than for simple regression.

I denote $(x_1 - \bar{x_1})$ as x_1^*, $(x_2 - \bar{x_2})$ as x_2^*, and $(y - \bar{y})$ as y^*.

$$\begin{aligned} b_1 =& \frac{\sum_{i=1}^{n} x_1^* y^* \sum_{i=1}^{n} x_2^{*2} - \sum_{i=1}^{n} x_2^* y^* \sum_{i=1}^{n} x_1^* x_2^*}{\sum_{i=1}^{n} x_1^{*2} \sum_{i=1}^{n} x_2^{*2} - (\sum_{i=1}^{n} x_1^* x_2^*)^2} \\ b_2 =& \frac{\sum_{i=1}^{n} x_2^* y^* \sum_{i=1}^{n} x_1^{*2} - \sum_{i=1}^{n} x_1^* y^* \sum_{i=1}^{n} x_1^* x_2^*}{\sum_{i=1}^{n} x_1^{*2} \sum_{i=1}^{n} x_2^{*2} - (\sum_{i=1}^{n} x_1^* x_2^*)^2} \\ a =& \bar{y} - b_1 \bar{x_1} - b_2 \bar{x_2} \end{aligned} \tag{3}$$

Model fit—residual standard error

Computed in the same way as for simple regression.

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{\underbrace{n}_{\text{sample size}} - \underbrace{k}_{\text{\# predictors}} - 1}} \tag{4}$$

Of course, now $k > 1$ but the interpretation is identical as before: a sort of “average” residual.

Model fit—R^2

The formula is now more complex.

The interpretation is the same: the share of the variance in Y which is explained by the predictors $X_1, X_2, \ldots,$ considered together.

With every X added to the model, R^2 increases, though. That is not very desirable.

Model fit—adjusted R^2

Applies a correction to the R^2 based on the number of variables (k) in the model. There are multiple types of adjusted R^2 proposed. $\mathcal{R}$ uses what is called the “Wherry Formula -1”.

$$\tilde{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1} \tag{5}$$

As before, n is the sample size.

California 1992

A plausible predictor might be ethnicity (FDR in the 1930s, LBJ in the 1960s).

	Simple	Multiple
(Intercept)	26.516***	25.436***
	(2.358)	(2.083)
% college educated	0.771***	0.657***
	(0.116)	(0.106)
% African-Americans		0.920***
		(0.218)
R^2	0.440	0.577
Adj. R^2	0.430	0.562
Num. obs.	58	58

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$. Standard errors in brackets.
DV (outcome) is percent vote for Democrats in county.

Comparison of simple and multiple regression

Thank you for the kind attention!

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models* (2nd ed.). Thousand Oaks, CA: Sage.