PLSC 502: "Statistical Methods for Political Research"

Linear Regression, III
November 17, 2016

Introduction: Model Fit

"Goodness of fit" refers to how "well" the model fits the data...

- Does X do a "good job" of accounting for variation in Y? Or, equivalently,
- Are the errors $\hat{u}_i$ generally small or large?

There was a time when people in our discipline were obsessed with model fit, and with R^2 in particular. Thankfully, that time has now passed. But, it's still useful to think and learn about model fit in general, and to be familiar with what to do and not to do about the issue.

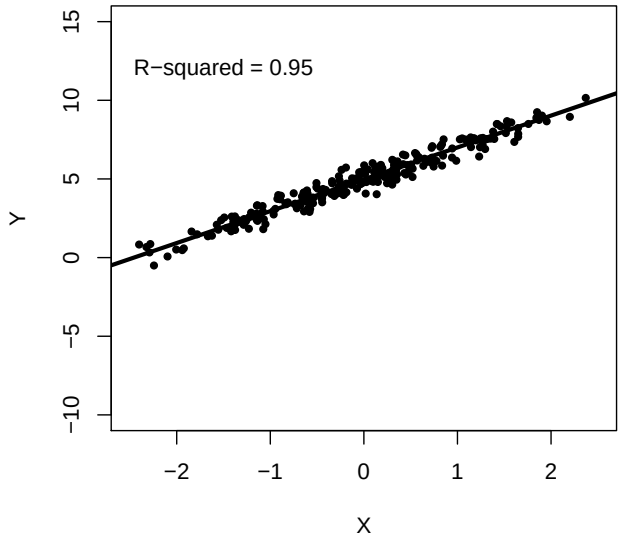
An Illustration

For starters, let's get some intuition going. Consider two regressions on two (simulated) Y variables, Y_1 and Y_2, each with $N = 250$ and each with the same linear relationship between Y and X:

```
> X<-rnorm(250)
> Y1<-5+2*X+rnorm(250,mean=0,sd=sqrt(0.2))
> Y2<-5+2*X+rnorm(250,mean=0,sd=sqrt(20))
> fit<-lm(Y1~X)
> summary(fit)

Call:
lm(formula = Y1 ~ X)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0585     0.0275   184.0   <2e-16 ***
X             1.9695     0.0290    67.9   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.435 on 248 degrees of freedom
Multiple R-squared: 0.949,Adjusted R-squared: 0.949
F-statistic: 4.61e+03 on 1 and 248 DF,  p-value: <2e-16
```

The scatterplot looks like this:

Figure 1: Regression of $Y_{1i} = 5 + 2X_i + u_i$ ($R^2 = 0.95$)

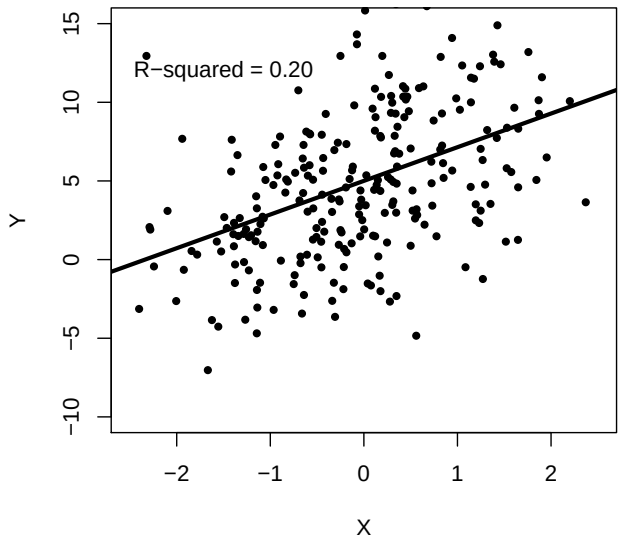


For Y_2, we have:

```
> fit2<-lm(Y2~X)
> summary(fit2)

Call:
lm(formula = Y2 ~ X)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.226      0.242   21.63  < 2e-16 ***
X              2.025      0.255    7.95  6.8e-14 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.82 on 248 degrees of freedom
Multiple R-squared: 0.203,Adjusted R-squared:  0.2
F-statistic: 63.1 on 1 and 248 DF,  p-value: 6.8e-14
```

with a scatterplot like this:

Figure 2: Regression of $Y_{2i} = 5 + 2X_i + u_i$ ($R^2 = 0.20$)



Note a few things:

- The slope and intercept of the regression line is identical in both.
- What is different is the *dispersion* of the points around the line.
  - That is, the *errors* are smaller for Y_1.
  - Put differently, a greater part of the variation of Y_1 is explained by X than is variation in Y_2.

This raises the issue of...

Variation

First, think about the variation we're interested in explaining. If we had no variables at all, we could still make a guess at Y. In that circumstance,

- The "best" guess we could make would be the mean $\bar{Y}$.
- That is, $\bar{Y}$ is the value of Y that minimizes the sum of squared deviations in Y...
- This is equivalent to estimating an OLS regression with a constant only...

Since we can always "guess the mean," what we're really interested in is how much our knowledge of X improves our guess of Y over a guess based on the mean. As a result, we commonly think of the total amount of variability in Y as variation around its mean:

- Remember that Y is a random variable: $Y_i = \mu + u_i$...
- ...and that X influences the systematic part of Y: $\mu_i = \beta_0 + \beta_1 X_i$.

Once we've estimated a regression line, we can say something about the variability of Y that is "due to" (or "explained by") X...

- The total variation in Y can then be written as

$$\begin{aligned} \text{Var}(Y) &= \text{Var}(\hat{Y} + \hat{u}) \\ &= \text{Var}(\hat{Y}) + \text{Var}(\hat{u}) + 2\,\text{Cov}(\hat{Y}, \hat{u}) \\ &= \text{Var}(\hat{Y}) + \text{Var}(\hat{u}) \end{aligned} \tag{1}$$

  where the last equality holds because $2\,\text{Cov}(\hat{Y}, \hat{u}) = 0$ by assumption.

- In other words, the total variation in Y now consists of that part due to X and the part due to the error term.
- The (squared) variation in Y around its mean which can be explained by X and the estimates of is called the "estimated (or "explained") sum of squares."
- Once we've estimated the regression, we're interested in the amount of residual variation in Y around the estimated regression line.
  - This is equal to the sum of the squared errors ($\sum_{i=1}^{N} \hat{u}_i^2$).
  - This is the variation in Y that can't be "explained by"/"accounted for by" X.

We often see this expressed as:

$$\underset{\text{("Total")}}{\textbf{TSS}} = \underset{\text{("Estimated," or "Model")}}{\textbf{MSS}} + \underset{\text{("Residual")}}{\textbf{RSS}}$$

The extent of improvement in our "guess" of Y due to the information contained in X and our estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ is therefore equal to the proportion of the TSS accounted for by the ESS:

$$\begin{aligned} R^2 &= \frac{\text{MSS}}{\text{TSS}} \\ &= \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \end{aligned} \tag{2}$$

which can also be thought of as:

$$\begin{aligned} R^2 &= 1 - \frac{\text{RSS}}{\text{TSS}} \\ &= 1 - \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2} \end{aligned} \tag{3}$$

This is the much-ballyhooed R^2 *statistic*.

R-squared:

- is often talked about as "the proportion of variance explained" by the model.
- is bounded between zero and one:
  - $R^2 = 1.0$ means a "perfect (linear) fit": all the points Y lie exactly on the estimated regression line.
  - $R^2 = 0$ means that X tells us nothing about Y beyond its mean.

Note some *characteristics* of R^2...

- An easy formula for R^2 in the bivariate case is

$$R^2 = \hat{\beta}_1^2 \frac{\sum (X_i - \bar{X})^2}{\sum (Y_i - \bar{Y})^2}. \tag{4}$$

  that is, the estimated slope $\hat{\beta}_1$ squared times the total variability in X divided by the variability in Y. Note that squaring $\hat{\beta}_1$ ensures that R^2 is always positive (as it must be).

- A little algebra can show that, in the bivariate case, R^2 is also equal to the square of the Pearson's correlation r_{XY}.

- As we add additional X variables to our model specification, the R^2 must increase (or at least not decrease).
  - Intuitively, this is because adding additional variables can't make the fit of the line (plane) to Y any worse.
  - Mathematically, this is because we can always calculate R^2 as (3), even when we have more than one explanatory variable. Since (as we'll see later) adding more explanatory variables can only make the fit of the model better (never worse), the R^2 can only go up (or stay the same).
  - This fact has made some people wary of R^2 (since you can make R^2 increase by just adding a bunch of variables), and led to alternative measures of goodness-of-fit (see below...).

Treating R^2 Like A Statistic

As Luskin (1991) notes, we can also think of R^2 as an estimate of a particular quantity. That is, if we denote the total amount of variability in Y as σ_Y^2, then the "population" analogue to R^2 for a particular model specification (what Luskin calls "P^2") is

$$P^2 = 1 - \frac{\sigma^2}{\sigma_Y^2} \tag{5}$$

Seen in this way, R^2 can be thought of as something like $\hat{P}^2$ – the sample-data-based estimate of the (population) coefficient of variation.

This characterization underscores the fact that – like our $\hat{\beta}$s – the R^2 we calculate is itself a random variable, with its own variation. In fact, Wishart showed years ago that

$$\widehat{\text{Var}(R^2)} = \frac{4R^2(1-R^2)^2(N-k)^2}{(N^2-1)(N+3)} \tag{6}$$

and so the corresponding standard error estimate of R^2 is

$$\widehat{\text{s.e.}(R^2)} = \sqrt{\frac{4R^2(1-R^2)^2(N-k)^2}{(N^2-1)(N+3)}}. \tag{7}$$

Note that these quantities are calculable using only the values of R^2, N, and k – we need not know about (e.g.) the total variation in Y or X to calculate them.[1]

Equations (6) and (7) mean that, in principle, we can conduct inference on R^2 in the same way as we do with other estimated quantities: calculating confidence intervals, for example, and doing hypothesis testing. As a practical matter, this might allow us to say that (e.g.) "(W)e can reject the hypothesis that $P^2 = 0.50$ at the 95 percent level of confidence."

As it happens, R^2 is a biased (but consistent) estimator of P^2. An unbiased estimator is...

Adjusted R^2

The fact that R^2 can be made larger simply by adding variables led people to think that it might be wise to have a measure that didn't do so. One way to do this is to "adjust" R^2 according to the number of covariates in the model. This led to *adjusted* R^2:

$$R^2_{adj.} = 1 - \frac{(1-R^2)(N-c)}{(N-k)} \tag{8}$$

[1] Also note that, in R, the `CI.Rsq` routine (in the `psychometric` package) will calculate the standard error of (and confidence intervals around) R^2 for particular values of R^2, N, and k.

where c equals one if there is a constant in the model and zero otherwise, and k is the number of covariates in the model (including the constant). $R^2_{adj.}$ behaves a bit differently than "normal" R^2:

- Asymptotically (that is, as $N \to \infty$), $R^2_{adj.}$ and R^2 are the same.
- Unlike regular R^2, $R^2_{adj.}$ can be > 1, or < 0...
- For the constant-only model, the $R^2_{adj.}$ equals R^2 (that is, both equal zero).
- As model fit (R^2) increases, $R^2_{adj.}$ increases as well, *but*
- The extent of that increase is discounted by a factor proportional to the number of covariates.

As a practical matter, people tend to report $R^2_{adj.}$ rather than plain R^2 when they have lots of variables in a model, and/or when there is a large difference between the two statistics, and/or in any other circumstance in which R^2 might be prone to giving a misleading impression.

R^2: An Example

Consider these data:

	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
	1	1	-1	-2	1	4	2
	2	1	0	-2	0	4	0
	3	7	1	4	1	16	4
$\sum_{i=1}^{3}(\cdot) =$	6	9	0	0	2	24	6

so that $\bar{X} = 2$ and $\bar{Y} = 3$. We can first calculate the $\hat{\beta}$s:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} \\ &= \frac{6}{2} \\ &= \mathbf{3} \end{aligned}$$

and:

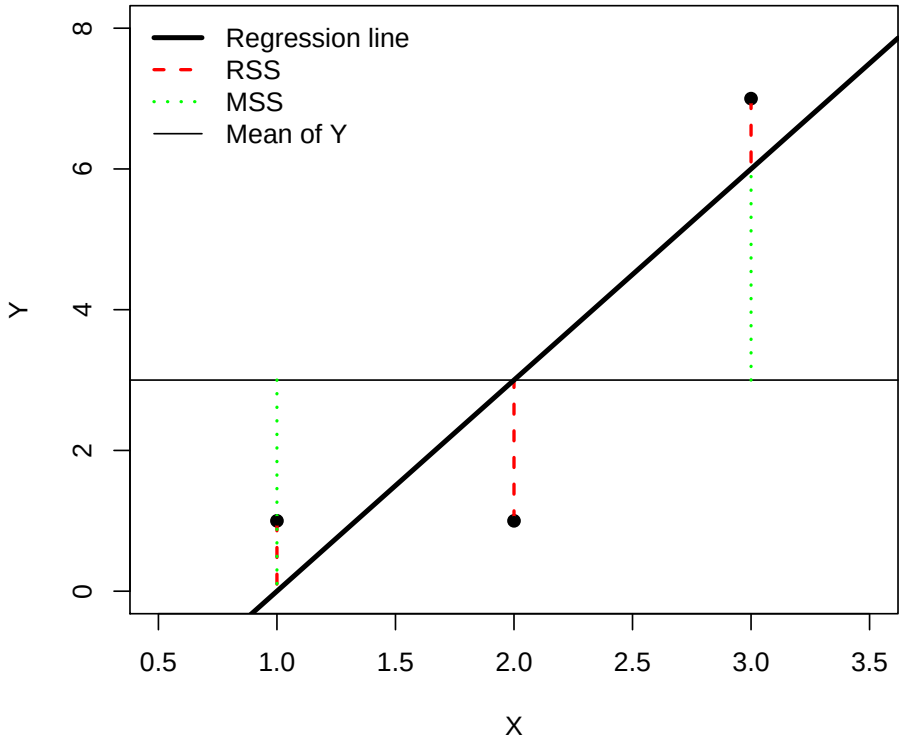
$$\begin{aligned} \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ &= 3 - 3(2) \\ &= \mathbf{-3} \end{aligned}$$

From these, we can calculate predicted values, residuals, and so forth:

	X_i	Y_i	$\hat{Y}$	$(\hat{u}_i)$ $Y_i - \hat{Y}$	(RSS) $(Y_i - \hat{Y})^2$	(MSS) $(\hat{Y}_i - \bar{Y})^2$
	1	1	0	1	1	9
	2	1	3	-2	4	0
	3	7	6	1	1	9
$\sum_{i=1}^{3}(\cdot) =$	6	9	9	0	6	18

Here's a graph:

Figure 3: A Three-Observation Regression



Here,

- The dots are observed values,
- The heavy line is the regression line (i.e., the predicted values of Y given X and the $\hat{\beta}$s),
- The lighter horizontal line is the mean of Y (that is, $\bar{Y} = 3$),

- The long-dashed (red) lines represent the differences between the observed and the predicted values (i.e., the residuals: $\hat{u}_i = Y_i - \hat{Y}_i$), and
- The shorter-dashed (green) lines are the "model improvements" (that is, $\hat{Y}_i - \bar{Y}$).

Now, for R^2, we can calculate it two ways. The simplest way is just:

$$\begin{aligned} R^2 &= \frac{\text{MSS}}{\text{TSS}} \\ &= \frac{18}{24} \\ &= \mathbf{0.75} \end{aligned}$$

But we could also use the formula in (4):

$$\begin{aligned} R^2 &= \hat{\beta}_1^2 \frac{\sum(X_i - \bar{X})^2}{\sum(Y_i - \bar{Y})^2} \\ &= 3^2 \left(\frac{2}{24}\right) \\ &= 9\,(0.08\bar{3}) \\ &= \mathbf{0.75} \end{aligned}$$

The corresponding $R^2_{adj.}$ is then:

$$\begin{aligned} R^2_{adj.} &= 1 - \frac{(1 - R^2)(N - c)}{(N - k)} \\ &= 1 - \frac{(1 - 0.75)(3 - 1)}{3 - 2} \\ &= 1 - \frac{0.5}{1} \\ &= \mathbf{0.50} \end{aligned}$$

Here, the value of R^2 tells us that 75 percent of the variance around the mean of Y is "explained" by X. (The adjusted R^2 doesn't really tell us much here, both because we have only a single covariate and because the N is very small...). One would normally consider this a pretty good-fitting model.

We can check our calculations with R:

```
> summary(lm(Y~X))

Call:
```

```
lm(formula = Y ~ X)

Residuals:
 1  2  3
 1 -2  1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.00       3.74   -0.80     0.57
X               3.00       1.73    1.73     0.33

Residual standard error: 2.45 on 1 degrees of freedom
Multiple R-squared: 0.75,Adjusted R-squared:  0.5
F-statistic:    3 on 1 and 1 DF,  p-value: 0.333
```

Alternatives to R^2

As the King, Luskin, etc. readings suggested, R^2 is not the only game in town when it comes to statistics for assessing model fit...

The Standard Error of the Estimate (SEE)

The SEE (sometimes called the "root mean squared error," or "RMSE") is just

$$\text{SEE} = \sqrt{\frac{\text{RSS}}{N-k}} \tag{9}$$

One can think of it as a kind of "average residual;" of course, since OLS residuals sum to zero, its not really an average per se, but it does represent something like a "typical" error value for the model. SEE:

- Is expressed in units of Y, and so is easily interpretable.
- Effectively tells you the same thing R^2 does – how well the model fits the data.

F-tests

Another summary measure of goodness-of-fit is the F-test. This is a statistical test for H_0: All βs except the constant are equal to zero. We won't go into this in any detail here, since we'll talk about F statistics at length when we do multivariate regression.

Using R^2

It's fair to say that, in general, R^2 has an undeservedly bad rap in political science. A big reason for this is that, not so long ago (30-40 years), a high R^2 was considered the *sine qua*

non of a good model. Now, R^2 is still reported, but analysts don't make a big deal over it the way they did back then. The most important thing(s) to remember fall into two general categories:

What R^2 *Can* Tell You

1. How well the *sample regression* estimate(s) fit the *sample data*.
2. Accordingly, how well you can *predict the sample data* from your estimates.
3. Comparing R^2s for different models *on the same data* can also be useful:
   - Doing so can tell you, for example, if adding variables improves model fit.
   - Of course, other things can also tell you this (F, t-tests, etc.).
4. As we discussed above, one can also think of R^2 as an *estimate of its corresponding population parameter*.
   - Luskin (1991) refers to the latter as "P^2."
   - R^2 is a biased (but consistent) estimator of P^2; $R^2_{adj.}$ is an unbiased estimator.
   - Of course, as with any sample statistic, this may or may not tell you much about what's going on in the population.

What R^2 *Can't* Tell You

1. *Which model is the "right" one.*
   - For one thing, some phenomena may just not be very predictable (i.e., low P^2).
   - It can then be easy to "overfit" the data in the sample – e.g., include variables that, for idiosyncratic reasons, explain some of the variance in Y in the sample, but not in the population.
   - The out-of-sample predictions will then be *worse* for the higher-R^2 model...
   - The strong implication is that one shouldn't use R^2 (or anything else, for that matter) for specification searches.
2. The *relative performance of two completely different models* (that is, using different data, or different dependent variables).
3. How well your model will *predict Y out-of-sample*.
   - Again, there's always the risk of overfitting.
   - Moreover, you could have one really weird sample, such that the estimates are *way* off from the population parameters – R^2 won't tell you if this is the case or not.