Regression

January 12, 2014

©2014 by Christopher A. Sims. This document is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

What a regression line looks like

How to get it in R

```
> lmout <- lm(testscr ~ str, data=caschool)
> with(caschool, plot(str, testscr))
> lines(caschool$str, lmout$coefficients[1]
              + caschool$str * lmout$coefficients[2], col="green")
```

What are we going to use the results for?

1. Intervening to change the distribution of `str`, the student teacher ratio. (**causal**)

2. Predicting test scores for a new school, not in our data set but drawn from the same population, for which we have data on student teacher ratio, but not yet on tests scores. **predictive**

What are we going to use the results for?

1. Intervening to change the distribution of `str`, the student teacher ratio. (**causal**)

2. Predicting test scores for a new school, not in our data set but drawn from the same population, for which we have data on student teacher ratio, but not yet on tests scores. **predictive**

3. The first makes a stronger assumption.

Deriving the least squares regression line, univariate case

$$\min_{\beta} \mathsf{SSR} = \sum_j (Y_j - X_j\beta)^2$$

$$\frac{d\mathsf{SSR}}{d\beta} = \sum_j 2(Y_j - X_j\beta)X_j = 0$$

$$\hat{\beta} = \frac{\sum_j Y_j X_j}{\sum_j X_j^2}$$

Multivariate case

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

$$Y : n \times 1 \qquad X : n \times k$$

$$\min_{\beta} \mathsf{SSR} = (Y - X\beta)'(Y - X\beta)$$

$$2X'(Y - X\beta) = 0$$

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

The constant term

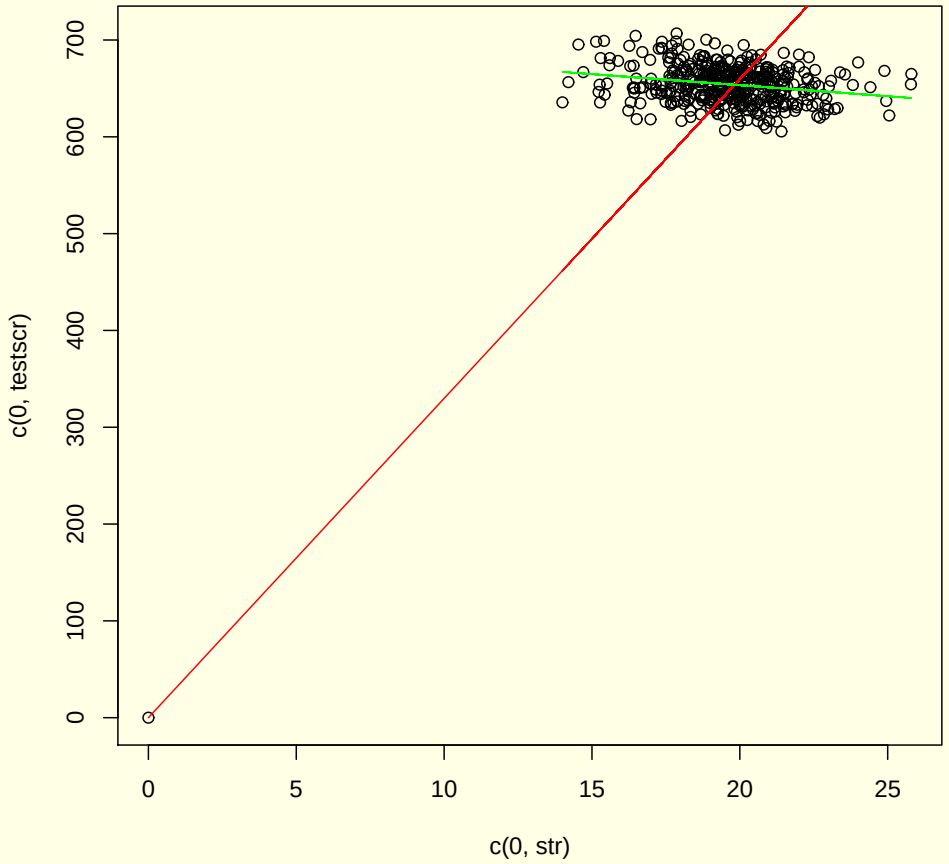
The regression line in our initial plot was for

$$Y_i = \alpha + X_i\beta + \varepsilon_i \, .$$

That is, it includes a "constant term" α.

R includes a constant in a regression by default. If one excludes the constant, one gets a "line through the origin"

Regression line through the origin



Matrix notation for regression with a constant

$$\underset{n\times 1}{\mathbf{1}}$$

is a column vector of 1's. If we include this as a column in the X matrix, its coefficient is the constant term.

To get a regression without a constant in R, write

```
lmout <- lm(testscr ~ str + 0, data=caschool)
```

Properties of least squares estimates

Let

$$\underset{n\times 1}{\hat{\varepsilon}} = \underset{n\times 1}{Y} - \underset{n\times k}{X}\underset{k\times 1}{\hat{\beta}}$$

be the **least squares residuals**, the error terms calculated using the least squares estimate $\hat{\beta}$. Then the least squares residuals have zero sample covariance with X:

$$X'\hat{\varepsilon} = X'(Y - X(X'X)^{-1}X'Y) = X'Y - X'X(X'X)^{-1}X'Y = X'Y - X'Y = 0.$$

A special case of this result is that if X includes a column of ones (i.e. if the regression has a constant term), the sum, and thus the sample mean, of the $\hat{\varepsilon}_j$'s is zero.

Using deviations from sample means

If $\bar{Y}$ is the sample mean of Y and $\bar{X}$ the sample mean of X, and if we write the estimated equation of a regression with constant as

$$Y_j = \hat{\alpha} + X_j\hat{\beta} + \hat{\varepsilon}_j \,,$$

then taking averages of both sides we can conclude that

$$\bar{Y} = \alpha + \bar{X}\beta + \bar{\varepsilon} \,.$$

So $\bar{Y} = \alpha + \bar{X}\beta$.

Using deviations from sample means, continued

Now subtract $\bar{Y}$ from the left hand side of the regression equation and $\alpha + \bar{X}\beta$ (which is the same thing) from the right hand side. This gives us

$$Y_j - \bar{Y} = (X_j - \bar{X})\hat{\beta} + \hat{\varepsilon}_j \ ,$$

i.e. a new equation with the same $\hat{\beta}$ and the same least squares residuals $\hat{\varepsilon}_j$, but no constant term.

Using deviations from sample means, continued

Now subtract $\bar{Y}$ from the left hand side of the regression equation and $\alpha + \bar{X}\beta$ (which is the same thing) from the right hand side. This gives us

$$Y_j - \bar{Y} = (X_j - \bar{X})\hat{\beta} + \hat{\varepsilon}_j \ ,$$

i.e. a new equation with the same $\hat{\beta}$ and the same least squares residuals $\hat{\varepsilon}_j$, but no constant term.

Conclusion: Instead of including a column of 1's in the X matrix, and estimating a two-dimensional β, we can remove the sample averages from the data series and estimate a regression with a single β and no constant. We'll get the same β and residuals.

Assumptions that suggest OLS is a good idea

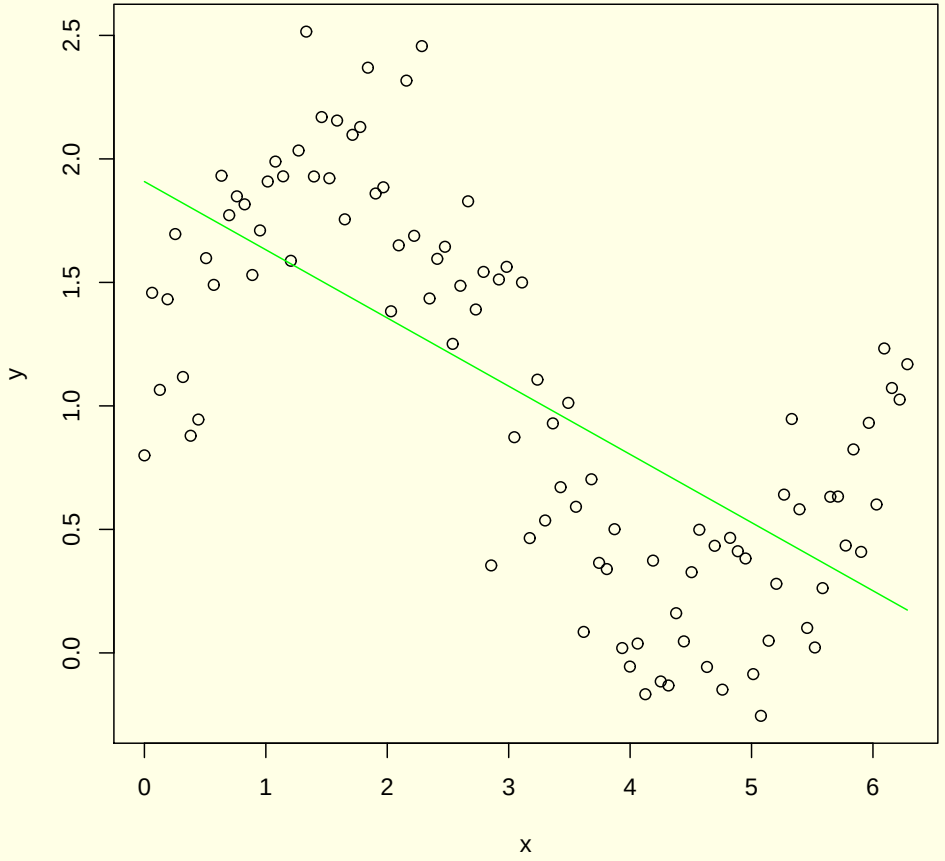
(These are not in the same order that S&W discuss them.)

1. (X_i, Y_i) have finite variances.

2. (X_i, Y_i) are i.i.d.

3. $E[Y_i \mid X_i] = X_i\beta$

x and y distribution that doesn't satisfy (3)



When we need which assumptions

- If we are just predicting Y_i from X_i for an observation newly drawn from the same population as the original data on which we estimated $\hat{\beta}$, and if we simply want the best *linear* predictor of Y_i using X_i, we do not need assumption (3).

- But if we are considering predicting the effects of changing the distribution of the X_i, and we want to use the regression line to do that, we need the stronger assumption (3).

- In most economic applications we want to make this causal (sometimes also called **structural**) interpretation of the regression.

Justifying using least squares without 3.

We wish to minimize $E[(Y_j - X_j\beta)^2$ by choosing β appropriately. We assume finite variance i.i.d. data (i.e. (1-2) above). Let $\Sigma_{YY} = E[Y_j^2]$, $\Sigma_{XY} = E[X_j'Y_j]$, and $\Sigma_{XX} = E[X_j'X_j]$. Then

$$E[(Y_j - X_j\beta)^2] = \Sigma_{YY} - 2\beta'\Sigma_{XY} + \beta'\Sigma_{XX}\beta \, .$$

Taking derivatives of this expression with respect to the elements of the β vector and setting them to zero gives us the vector of equations

$$-2\Sigma_{XY} + 2\Sigma_{XX}\beta = 0 \, ,$$

whose solution is $\beta = \Sigma_{XX}^{-1}\Sigma_{XY}$.

Least squares without 3, continued

What we have derived on the previous slide is the best population value for β, assuming we knew the true population joint distribution. The β we found depends on Σ_{XX} and Σ_{XY}, which we don't observe. But for i.i.d. variables, with finite expectations, time averages converge to expectations, so

$$\frac{1}{n}X'X = \frac{1}{n}\sum_j X_j'X_j \xrightarrow[n\to\infty]{} E[X_j'X_j] = \Sigma_{XX}$$

$$\frac{1}{n}X'Y = \frac{1}{n}\sum_j X_j'Y_j \xrightarrow[n\to\infty]{} E[X_j'Y_j] = \Sigma XY \ .$$

Substituting these estimates of the unknown Σ matrices into our formula

gives an estimate of the population least squares β, and this estimate coincides with the sample $\hat{\beta}$ we derived above.

Using (3) to get further properties of $\hat{\beta}$

Using (3), we can see that

$$E[\hat{\beta} \mid X] = E[(X'X)^{-1}X'Y \mid X] = E[(X'X)^{-1}X'X\beta \mid X] = \beta\,.$$

In other words, least squares provides an **unbiased** estimate of β. This is not true, in general, without assumption (3)

Measures of fit: R^2

Because of the zero sample covariance between least squares residuals and X, we split the sum of squared Y's into two components, "explained" and "unexplained":

$$Y'Y = (X\hat{\beta} + \hat{\varepsilon})'(X\hat{\beta} + \hat{\varepsilon}) = \underbrace{\hat{\beta}'X'X\hat{\beta}}_{ESS} + \underbrace{\hat{\varepsilon}'\hat{\varepsilon}}_{RSS}$$

The right-hand side above leaves out some crossproduct terms that are zero because of the zero covariance of X with $\hat{\varepsilon}$. Stock and Watson use the abbreviations ESS for "explained some of squares", RSS for "residual sum of squares", and TSS for "total sum of squares", which is just $Y'Y$. The

regression R^2 is defined as

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{RSS}{TSS}\,.$$

R^2 defined this way is always between 0 and 1. $R^2 = 0$ implies the residuals have the same sample variance as the original Y data. $R^2 = 1$ occurs only when the residuals are all zero, i.e. when all the data points are exactly on the regression line.

R^2 in terms of deviations from means

As given above, $R^2 > 0$ in a regression with a constant, but in which all other coefficients are zero. Since such a regression predicts Y just using the sample mean of Y, R^2 is usually calculated using deviations from sample means. That is, we use TSS $= (Y - \bar{Y})'(Y - \bar{Y})$ and define

$$R^2 = 1 - \frac{RSS}{TSS}$$

This will make $R^2 = 0$ if the only non-zero least squares coefficient is the constant term.

Note, though, that if there is no constant term in the regression, this way of computing R^2 can give negative values.

Measures of fit: SEE

The standard error of estimate or SEE is defined, when the sample size is n, as

$$\sqrt{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}} : .$$

It can also, and is perhaps more commonly, defined by the same formula with n replaced by $n - k$, where k is the length of an X_j vector.

The multivariate normal distribution

If the $k \times 1$ vector x is multivariate normally distributed with mean μ and covariance matrix Σ (i.e. $x \sim N(\mu, \Sigma)$), it has the pdf

$$(2\pi)^{-k/2} |\Sigma|^{-1/2} e^{-\frac{1}{2}(x-\beta)'\Sigma^{-1}(x-\beta)} .$$

Jointly normal random variables are independent if and only if their covariances are zero.

Properties of multivariate normal variables

If $X \sim N(\mu, \Sigma)$, and if $Z = AX + C$, where A is any non-random matrix whose number of columns matches the number of elements in X and C is any non-random vector of the same length as X, then

$$Z \sim N(A\mu + C, A\Sigma A') \,.$$

The formulas for transforming the mean and covariance matrix apply regardless of the distribution of X. What's special about joint-normal variables is that such linear transformations leave the result in the class of multivariate normals.

Conditional distributions for joint normals

If

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right),$$

then

$$X_1 \mid X_2 \sim N\big(\mu_1 + (X_2 - \mu_2)\Sigma_{22}^{-1}\Sigma_{21}, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\big)$$

Conditional distributions continued

- These formulas are complicated, and you don't need to memorize them.
- The important things about them to remember are
  - Conditional distributions of one set of jointly normal variables given others are normal.
  - The conditional mean is the unconditional mean plus the deviation of the conditioning variable from its mean times the least squares regression coefficients for X_1 on X_2.
  - The conditional covariance matrix is always smaller than the unconditional covariance matrix, in that it differs from it by a p.s.d. (positive semi-definite) matrix.

The SNLM

SNLM stands for "Standard Normal Linear Model". It is

$$\underset{n\times 1}{Y} \mid \underset{n\times k}{X} \sim N(X \underset{k\times 1}{\beta}, \sigma^2 I)\,.$$

Equivalently, we can write

$$Y = X\beta + \underset{n\times 1}{\varepsilon}\,, \qquad \varepsilon \mid X \sim N(0, \sigma^2 I)\,.$$

Least squares in the SNLM

Conditional on X, $\hat{\beta} = (X'X)^{-1}X'Y$ is just a linear transformation of Y, which is normally distributed, so we conclude immediately that

$$\hat{\beta} \mid X \sim N((X'X)^{-1}X'X\beta, \sigma^2(X'X)^{-1}X'X(X'X)^{-1}) = N(\beta, \sigma^2(X'X)^{-1}).$$

$\hat{\beta}$ is the MLE

The log likelihood is the multivariate normal pdf for Y, treated as a function of the unknown β and σ^2.

$$-\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{\sigma^2}(Y - X\beta)'(Y - X\beta)$$

The only place β appears in this expression is in the last crossproduct, which is just the sum of squared residuals. So it is obvious that the likelihood is maximized when this sum of squares is minimized, i.e. at the least squares (sometimes called "ordinary least squares", or OLS) estimator $\hat{\beta}$. Even though we don't know σ^2, since $\hat{\beta}$ maximizes likelihood for any given value of σ^2, our beliefs about σ^2 have no effect on the MLE (Maximum Likelihood Estimator).

The distribution of $\beta \mid Y, X$

If the prior is flat relative to the likelihood, so we can treat the likelihood as the posterior pdf, it is fairly easy to see what is the distribution of $\beta \mid Y, X, \sigma^2$. As we have already noted, with Y, X, and σ^2 held fixed, the log likelihood is quadratic in β. That means that exponentiated, it will be proportional to a multivariate normal distribution.

We can rewrite the last term in the log likelihood as

$$\frac{1}{2\sigma^2}\left((\beta - \hat{\beta})' X' X (\beta - \hat{\beta}) + \hat{\varepsilon}'\hat{\varepsilon}\right).$$

So when we exponentiate this, we get something proportional to a $N(\hat{\beta}, \sigma^2 (X'X)^{-1})$ distribution.

What about the unknown σ^2?

The distribution of $\beta \mid X, \sigma^2$ is not much use to us directly if we don't know σ^2. However, a maximum likelihood estimator of σ^2 is available:

$$s^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}\,.$$

This estimate reliably gets close to σ^2 as $n \to \infty$, so that if n is fairly large, replacing σ^2 with s^2 in the normal distribution conditional on X, σ^2 is a very good approximation. We will discuss what to do if n is not so big a little further on.

Distributions for individual coefficients

Since $\beta \mid X, \sigma^2$ is joint normal, each individual coefficient (as well as any linear combination of coefficients) is normal. The mean can be read off the mean vector of the joint distribution, and the variance can be read off from the corresponding diagonal element of the joint covariance matrix $\sigma^2(X'X)^{-1}$. So, approximating σ^2 with s^2, we get

$$\beta_i \sim N(\hat{\beta}_i, s^2[(X'X)^{-1}]_{ii}) \,,$$

where the notation $[A]_{ii}$ means the $i'th$ diagonal element of A. This allows us to form flat-prior posterior probability intervals for individual coefficients using the normal distribution. For example in R the 95% interval for the j'th coefficient is

```
c(qnorm(.05), qnorm(.95)) * s * sqrt(diag(solve(crossprod(X)))[j,j] + betahat[j]\:,
```

where `betahat[j]` is the j'th element of the least squares coefficient. One does not need to do it this way, though, because if the output of `lm` is stored in `lmout`, the estimated standard errors of the coefficients are printed out as output of `summary(lmout)`.

Why multivariate regression?

- An alternative might be to consider a collection of bivariate regressions.

- But if, say we include an important explanatory variable X, and also another variable Z that is simply a crude approximation to X, both X and Z will generate substantial R^2's in bivariate regressions, while in multivariate regression the coefficient on Z will be negligible, because once X's effect on Y has been accounted for, there is no explanatory power left for Z.

- Also, if there are several important X variables, it can happen that no individual binary regression of Y on the columns of X has a high R^2, but the multiple regression using of all of them has a high R^2. Then, because

the multiple regression has much smaller σ^2, all the coefficients will be more accurately estimated than they would be from binary regressions.

When n is not large

The true distribution for $\beta \mid X$ when σ^2 is unknown depends on the prior. With a flat prior on both β and $\log \sigma^2$, the posterior for β is a multivariate t distribution with scale matrix $s^2(X'X)^{-1}$ and degrees of freedom $n - k$. This joint distribution implies that each individual coefficient, and any linear combination of coefficients, has a $t(n - k)$ distribution. The "significance levels" displayed in regression program outputs are probabilities from this distribution.

Estimated standard errors and t-statistics

A standard regression program will give you estimates of a list of coefficients (the $\hat{\beta}$ vector in the SNLM) and, beside or below them, a list of estimated standard deviations for them. Or sometimes the same information is presented by giving you the estimatated coefficients and their t-statistics — the ratios of the coefficients to their estimated standard errors. As we have discussed, a more useful and natural interpretation of these statistics is that the standard deviations are standard deviations of the unknown coefficient values around the flat-prior posterior means in $\hat{\beta}$.

"Significance" based on coefficient estimates and standard errors

People commonly speak of estimated coefficients as "significant" or "insignificant". Usually they mean, by saying a coefficient is "significant", that its t-statistic is big. What's big, conventionally, is a value of this ratio big enough to put $\beta_i/\sigma_{\beta_i} = 0$ in the 5% or 2.5% tail of its distribution. The 5% tail is used when it's clear that only one sign for the coefficient makes sense. If $\hat{\beta}_i > 0$, for example, and this is the sign we expect when X_i has a non-trivial impact on Y, then we might use the "one-tailed" criterion and say it's "significant" if $P[\beta_i/\sigma_{\beta_i} < 0] \leq .05$ If either sign on the coefficient is plausible, people use a "two-tailed" test and say the coefficient is "significant" if $P[\beta_i/\sigma_{\beta_i} < 0] \leq .025$ or $P[\beta_i/\sigma_{\beta_i} \geq 0] \leq .025$ From a Bayesian perspective, the posterior probabilities are just useful guides to the shape of the likelihood.

Economic vs. statistical significance

It is important to remember that a coefficient can be "insignificant", meaning that it is not large relative to its posterior standard error, yet still be substantively important. Likewise a coefficient can be "significant", yet so small as to be zero for practical purposes. When sample size is small, it is easy for estimated standard errors to be large, so that even coefficients that imply important effects are not "significantly different from zero". This situation, where the data leaves substantively important uncertainty about effects, should be carefully distinguished from cases where the data are highly informative, the coefficients *and* their standard errors are small, and we are thus quite sure that the coefficients are substantively small. Many modern data sets have many thousands or even millions of observations. This can lead to estimated coefficients that are several times bigger than their estimated standard errors, thus highly "significant", but nonetheless so small as to be negligible for practical purposes.