Eric Doviak
23 April 2012

Statistics 3400 – Notes on Regression

I drafted these notes to help you understand the motivation behind regression analysis and to show you what you should look for as you read chaps. 11, 12 and 14 in the Schmidt textbook.

After reading these notes, you should have a better understanding of the conditions under which ordinary least squares yields unbiased estimates of the regression coefficients. You should also have a better understanding of variance and covariance and the role they play in the estimation of regression coefficients, testing the statistical significance of those coefficients and testing the overall fit of the model.

The Regression Equation

Suppose that there is a linear relation between two variables, x and y:

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i$$

and we would like to estimate α and β by minimizing the sum of squared errors:

$$\min_{\alpha,\beta} \sum \varepsilon_i^2 = \sum (y_i - \alpha - \beta \cdot x_i)^2$$

To find the minimum sum of squared errors, we must find the values of α and β that satisfy the following conditions:

$$\begin{aligned} \frac{\partial \sum \varepsilon_i^2}{\partial \alpha} &= -2 \cdot \sum (y_i - \alpha - \beta \cdot x_i) &&= 0 \\ \frac{\partial \sum \varepsilon_i^2}{\partial \beta} &= -2 \cdot \sum (y_i - \alpha - \beta \cdot x_i) \cdot x_i &&= 0 \end{aligned}$$

These conditions imply that:

$$\begin{aligned} \frac{1}{n} \sum \varepsilon_i &= E[\varepsilon] &&= 0 \\ \frac{1}{n} \sum \varepsilon_i \cdot x_i &= cov(\varepsilon, x) &&= 0 \end{aligned}$$

the expected value of the error term is zero and there is no covariance between the explanatory variables and the error terms. When those two conditions are satisfied:

$$\begin{aligned} \hat{\alpha} &= \bar{y} - \hat{\beta} \cdot \bar{x} \\ \hat{\beta} &= \frac{\sum (y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{cov(x,y)}{var(x)} \end{aligned}$$

The Standard Error of the Regression

The standard error of the regression is:

$$\begin{aligned}\sigma_\varepsilon^2 &= \frac{1}{n}\sum \varepsilon_i^2 \\ &= \frac{1}{n}\sum \left(y_i - \hat{\alpha} - \hat{\beta}\cdot x_i\right)^2\end{aligned}$$

the mean of the squared errors. Note that the value of the standard error of the regression depends on the estimates: $\hat{\alpha}$ and $\hat{\beta}$, so a change in $\hat{\alpha}$ or $\hat{\beta}$ would affect our estimate of σ_ε^2. For this reason, our estimate of the standard error also has a standard error.

R^2 and the F-statistic

As a measure of "goodness of fit" (i.e. the *coefficient of determination*), we can compare the squared errors to the variance of y.

$$R^2 = 1 - \frac{\sum \varepsilon_i^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$$

where $RSS \equiv \sum \varepsilon_i^2$ is the "residual sum of squares" and $TSS \equiv \sum (y_i - \bar{y})^2$ is the "total sum of squares." In this form, the coefficient of determination is computed by determining what percentage of the total variation in y (TSS) appears in the squared residuals (RSS). The remainder is explained by the regression.

An alternative way of calculating the coefficient of determination is to use the predicted value of y, i.e. $\hat{y}_i = \hat{\alpha} + \hat{\beta}\cdot x_i$ to calculate the "explained sum of squares:" $ESS \equiv \sum (\hat{y}_i - \bar{y})^2$ by the total sum of squares (TSS).

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{ESS}{TSS}$$

In this form, the coefficient of determination tells us what percentage of the total variation in y is explained by x (which is used to calculate $\hat{y}$).

Regardless of which form is used to calculate R^2, we can see that a high R^2 indicates that a large portion of the variance in y is explained by the model. For this reason, researchers frequently use R^2 to measure the overall fit of the model. As we'll see, the R^2 is closely related to the F-statistic, which can be used in formal hypothesis testing.

Like R^2, the F-statistic is a ratio of two variances:

$$F = \frac{ESS/(k-1)}{RSS/(n-k)}$$

but it also accounts for k, the number of regression coefficients in the model. Note that the F-statistic is lower (i.e. a penalty) when the number of explanatory variables is higher (because more explanatory variables increase ESS). The F-statistic is also lower when the residual sum of squares (RSS) is higher. The F-statistic is higher when the explained sum of squares (ESS) is higher.

$(k-1)$ is referred to as the "numerator degrees of freedom" because it accounts for the number of explanatory variables used "in the explanation" (i.e. ESS). Similarly, $(n-k)$ is referred to as the "denominator degrees of freedom" because it accounts for the number of observations, while subtracting off one degree for each explanatory variable (because adding explanatory variables reduces RSS).

Because the F-statistic and R^2 are both defined in terms of TSS, RSS and ESS, we can write the F-statistic in terms of R^2:

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

In hypothesis testing, we can use the F-statistic to test the hypothesis that the model explains a significant portion of the variation in y against the null hypothesis that it does not.

We can also use the F-statistic to test the contribution that individual explanatory variables make towards explaining the variation in y. Specifically, we could run a "restricted regression" where a coefficient (or set of coefficients) is constrained to equal a certain value (e.g. zero) and then compare the fit to the "unrestricted regression." In that case, the F-statistic would be given by:

$$F = \frac{\left(R^2_{UR} - R^2_R\right)/m}{\left(1-R^2_{UR}\right)/(n-k)}$$

where: m is the number of restrictions (i.e. the number of constrained coefficients), R^2_{UR} is the R^2 of the "unrestricted regression" and R^2_R is the R^2 of the "restricted regression."

The Standard Error of the Regression Coefficients

More commonly however, we compare the regression coefficient to its standard error – the most common application of the t-test. So let's first examine the standard errors of the regression coefficients, which are obtained by maximizing the *log-likelihood function*.

Maximum likelihood estimation (MLE) estimates the regression coefficients and the regression's standard error just like minimizing the sum of squared errors, but with the added assumption that the residuals are distributed normally. As before, the standard error of the regression is square root of the mean of the squared errors, $\sigma_\varepsilon = \sqrt{\frac{1}{n}\sum \varepsilon_i^2}$, and the standard errors of the regression coefficients are:

$$\begin{aligned}
\sigma_\alpha &= \sigma_\varepsilon \sqrt{\frac{\frac{1}{n}\sum x_i^2}{\sum (x_i - \bar{x})^2}} \\
\sigma_\beta &= \sigma_\varepsilon \sqrt{\frac{1}{\sum (x_i - \bar{x})^2}}
\end{aligned}$$

Note that the standard errors of the regression coefficients are higher when the regression standard error is higher. They're also lower when the variance of x is lower because we can be more confident in our estimates when we observe more variance in x. Finally, note that the the intercept's standard error, σ_α, is larger when the x_i are farther from zero. Consequently, we have to be less confident in our estimate of the intercept term.

t-tests of the Regression Coefficients

As alluded to in the previous section, we can have more confidence in our estimated regression coefficients when the ratio of the coefficient to the standard error is larger. In particular, we say that the effect of an explanatory variable on the dependent variable is statistically significant from zero if the variable's estimated regression coefficient is at least twice as large as its standard error. This is the most common form of the t-test.

The more general case of the t-statistic compares the distance between the estimated coefficient, $\hat{\beta}$, and a hypothesized value of the coefficient, β_{null} to the standard error of the regression coefficient, σ_ε:

$$t = \frac{\hat{\beta} - \beta_{null}}{\sigma_\varepsilon}$$

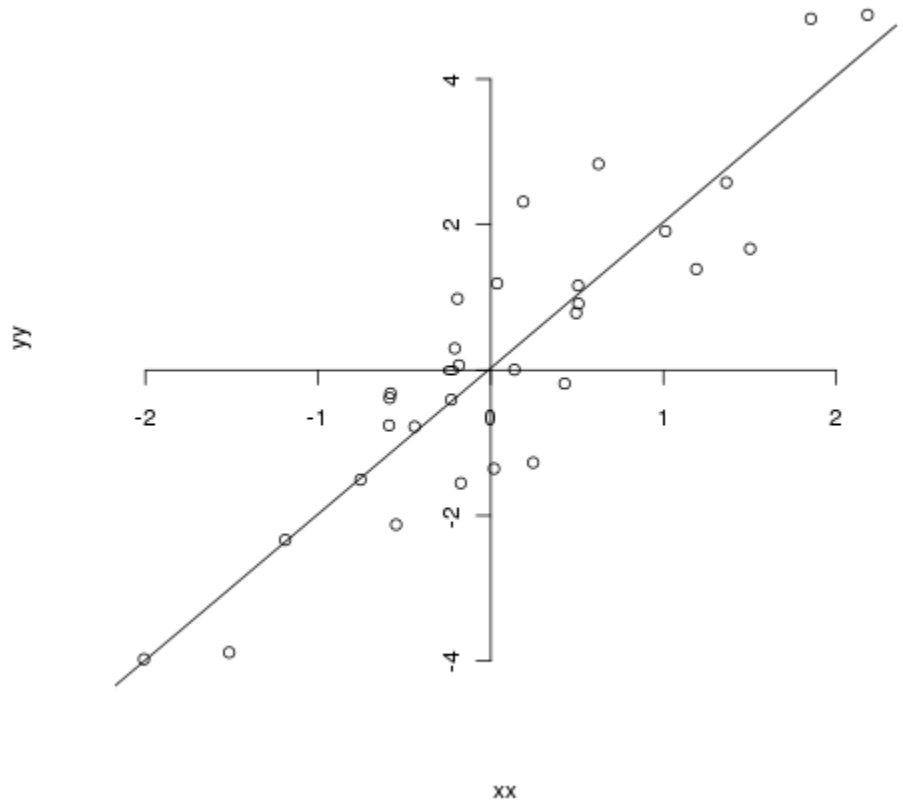
We then compare the t-statistic to a *critical value*. The most commonly used critical value is 1.96 because (in a large sample) 95 percent of the estimates would lie within 1.96 standard errors from the hypothesized value, β_{null}.

In the common form of the t-test, $\beta_{null} = 0$. In other words, the null hypothesis is that the explanatory variable does not have a statistically significant effect on the dependent variable. In such a case, we can reject the null hypothesis if the estimated coefficient, $\hat{\beta}$, is at least twice as large (in absolute value) as its standard error, σ_ε, because it would be very rare to observe such a large coefficient estimate if β were truly zero.

```
R version 2.11.1 (2010-05-31)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> ## quick R script to illustrate correlation and regression
>
> ## randomly generate X and Y with positive relation
> xx <- rnorm(30)
> yy <- 2*xx + rnorm(30)
>
> ## correlation
> cor(xx,yy)
[1] 0.8966973
>
> ## regression
> ols <- lm(yy~xx)
>
> ## plot the relation and
> ## insert regression line into plot
> png( "scatterplot.png" )
> plot( xx ,yy , axes = FALSE )
> axis(1 , pos = 0 )
> axis(2 , pos = 0 )
> abline( a = coef(ols)[1] , b = coef(ols)[2] )
> dev.off()
null device
          1
>
> ## how did we get coefficients?
> ## we minimized the sum of squares
> beta <- seq( 0.5 , 3.5 , 0.01 )
> sqerr <- NA
> for (i in 1:length(beta)) {
+   alfa <- mean(yy) - beta[i]*mean(xx)
+   sqerr[i] <- sum((yy - alfa - beta[i] * xx)^2)
+ }
>
> ## show sum of squares plotted over different possible values of beta
> png( "min-sum-squares.png" )
> plot( beta , sqerr , type = "l" , axes = FALSE )
> axis(1)
> axis(2)
> dev.off()
null device
          1
>
```

```
> ## find beta which minimizes sum of squares
> betahat <- beta[which(sqerr == min(sqerr))]
>
> ## corresponding alpha
> alfahat <- mean(yy) - betahat*mean(xx)
>
> ## compare coefficients with regression output
> alfahat
[1] 0.02580671
> betahat
[1] 2.01
>
> ## summarize regression
> summary(ols)

Call:
lm(formula = yy ~ xx)

Residuals:
     Min       1Q   Median       3Q      Max
-1.79051 -0.72444  0.02788  0.63851  1.91080

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02611    0.17507   0.149    0.883
xx           2.00706    0.18724  10.719 2.03e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.953 on 28 degrees of freedom
Multiple R-squared: 0.8041,     Adjusted R-squared: 0.7971
F-statistic: 114.9 on 1 and 28 DF,  p-value: 2.032e-11

>
> ## the R-squared:
> r2 <- 1 - (var(ols$residuals)/var(yy))
> r2
[1] 0.804066
>
> ## square root of R-squared is equal to the correlation coefficient
> sqrt(r2)
[1] 0.8966973
> cor(xx,yy)
[1] 0.8966973
>
> proc.time()
   user  system elapsed
  0.308   0.040   0.333
```

scatterplot and regression line



sum of squares as a function of “beta