Lecture Notes 5

Econ B2000, Statistics and Introduction to Econometrics

Kevin R Foster, the City College of New York, CUNY

Fall 2014

Regression in R

To have R do a linear regression, we use the command "lm()" as for example

```
model1 <- lm(Y ~ X1)
summary(model1)
```

This estimates a linear model of $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ and reports estimates of the intercept and slope coefficients.

So for example with the CEX data, create a variable for fraction of income spent on housing then replace the zero-income (thus Inf value) with missing NA values:

```
fraction_housing <- HOUSPQ/FINCATAX
fraction_housing[is.infinite(fraction_housing)] <- NA
```

Then form a regression of this on age of the reference person:

```
model2 <- lm(fraction_housing ~ AGE_REF)
summary(model2)
```

Regression Details

We'll often form hypotheses about regression coefficients: t-stats, p-values, and confidence intervals – so that's the same basic process as before. Usually two-sided (rarely one-sided).

We will commonly test if the coefficients 'are significant' – i.e. is there evidence in the data that the coefficient is different from zero? This goes back to our original example where we looked at the difference between the Hong Kong/Singapore stock returns and the US stock returns/interest rate. A zero slope is evidence against any relationship – this shows that the best guess of the value of Y does not depend on current information about the level of X. So coefficient estimates that are statistically indistinguishable from zero are not evidence that the particular X variable is useful in prediction.

A hypothesis test of some statistical estimate uses this estimator (call it $\hat{X}$) and the estimator's standard error (denote it as $se_{\hat{X}}$) to test against some null hypothesis value, X_{null} . To make the hypothesis test, form $Z = \frac{\hat{X} - X_{null}}{se_{\hat{X}}}$, and – here is the magic! – under certain conditions this Z will have a Standard Normal distribution (or sometimes, if there are few degrees of freedom, a t-distribution; later in more advanced stats courses, some other distribution). The magic happens because if Z has a Standard Normal distribution then this allows me to measure if the estimate of X, $\hat{X}$, is very far away from X_{null} . It's generally tough to specify a common unit that allows me to say sensible things about "how big is big?" without some statistical measure. The p-value of the null hypothesis tells me, "If the null hypothesis were actually true, how likely is it that I would see this $\hat{X}$ value?" A low p-value tells me that it's very unlikely that my hypothesis could be true and yet I'd see the observed values, which is evidence against the null hypothesis.

Often the formula, $Z = \frac{\hat{X} - X_{null}}{se_{\hat{X}}}$, gets simpler when X_{null} is zero, since it is just $Z' = \frac{\hat{X} - 0}{se_{\hat{X}}} = \frac{\hat{X}}{se_{\hat{X}}}$, and this is what SPSS prints out in the regression output labeled as "t". This generally has a t-distribution (with enough degrees of freedom, a Standard Normal) so SPSS calculates the area in the tails beyond this value and labels it "Sig".

This is in Chapter 5 of Stock & Watson.

We know that the standard normal distribution has some important values in it, for example the values that are so extreme, that there is just a 5% chance that we could observe what we saw, yet the true value were actually zero. This 5% critical value is just below 2, at 1.96. So if we find a t-statistic that is bigger than 1.96 (in absolute value) then the slope would be "statistically significant"; if we find a t-statistic that is smaller than 1.96 (in absolute value) then the slope would not be "statistically significant". We can re-write these statements into values of the slope itself instead of the t-statistic.

We know from above that

$$\frac{\hat{\beta}_1 - 0}{se(\beta_1)} = \frac{\hat{\beta}_1}{se(\beta_1)} = t ,$$

and we've just stated that the slope is not statistically significant if:

$$|t| < 1.96 .$$

This latter statement is equivalent to:

$$-1.96 < t < 1.96$$

Which we can re-write as:

$$-1.96 < \frac{\hat{\beta}_1}{se\left(\hat{\beta}_1\right)} < 1.96$$

Which is equivalent to:

$$-1.96\left(se\left(\hat{\beta}_1\right)\right) < \hat{\beta}_1 < 1.96\left(se\left(\hat{\beta}_1\right)\right)$$

So this gives us a "Confidence Interval" – if we observe a slope within 1.96 standard errors of zero, then the slope is not statistically significant; if we observe a slope farther from zero than 1.96 standard errors, then the slope is statistically significant.

This is called a "95% Confidence Interval" because this shows the range within which the observed values would fall, 95% of the time, if the true value were zero. Different confidence intervals can be calculated with different critical values: a 90% Confidence Interval would need the critical value from the standard normal, so that 90% of the probability is within it (this is 1.64).

OLS is nothing particularly special. The Gauss-Markov Theorem tells us that OLS is **BLUE**: **B**est **L**inear **U**nbiased **E**stimator (and need to assume homoskedasticity). Sounds good, right? Among the linear unbiased estimators, OLS is "best" (defined as minimizing the squared error). But this is like being the best-looking economist – best within a very small and very particular group is not worth much! Nonlinear estimators may be good in various situations, or we might even consider biased estimators.

If X is a binary dummy variable

Sometimes the variable X is a binary variable, a dummy, D_i, equal to either one or zero (for example, female). So the model is $Y_i = \beta_0 + \beta_1 D_i + u_i$ can be expressed as

$Y_i = \begin{cases} \beta_0 + \beta_1 + u_i & if\ D_i = 1 \\ \beta_0 + u_i & if\ D_i = 0 \end{cases}$. So this is just saying that Y has mean $\beta_0 + \beta_1$ in some cases and mean β_0 in other cases. So β_1 is interpreted as the difference in mean between the two groups (those with D=1 and those with D=0). Since it is the difference, it doesn't matter which group is specified as 1 and which is 0 – this just allows measurement of the difference between them.

Other 'tricks' of time trends (& functional form)

- If the X-variable is just a linear change [for example, (1,2,3,...25) or (1985, 1986,1987,...2010)] then regressing a Y variable on this is equivalent to taking out a linear trend: the errors are the deviations from this trend.

- If the Y-variable is a log function then the regression is interpreted as explaining percent deviations (since derivative of lnY = dY/Y, the percent change). (So what would a linear trend on a logarithmic form look like?)

- If both Y and X are logs then can interpret the coefficient as the elasticity.

- examine errors to check functional form – e.g. height as a function of age works well for age < 12 but then breaks down

- plots of X vs. both Y and predicted-Y are useful, as are plots of X vs. error.

In addition to the standard errors of the slope and intercept estimators, the regression line itself has a standard error.

A commonly overall assessment of the quality of the regression is the R^2 (displayed by many statistical programs). This is the fraction of the variance in Y that is explained by the model so $0 \leq R^2 \leq 1$. Bigger is usually better, although different models have different expectations (i.e. it's graded on a curve).

Statistical significance for a univariate regression is the same as overall regression significance – if the slope coefficient estimate is statistically significantly different from zero, then this is equivalent to the statement that the overall regression explains a statistically significant part of the data variation.

- Excel calculates OLS both as regression (from Data Analysis TookPak), as just the slope and intercept coefficients (formula values), and from within a chart

Multiple Regression – more than one X variable

Regressing just one variable on another can be helpful and useful (and provides a great graphical intuition) but it doesn't get us very far.

Suppose we wanted to look at a modern version of the classic Engel curve study: what fraction of expenditure goes to food? We can define the fraction spent on food,

```
fraction_food <- FOODPQ/TOTEXPPQ

fraction_food[is.infinite(fraction_food)] <- NA

fraction_food[fraction_food<0] <- NA # 1 reported negative total expenditure?!
```

There are probably lots of factors driving this variation. For example, people who label themselves as white, African-American, Asian, Native American, other race, and Hispanic have different average expenditures. Households where the reference person is African-American spend an average of 19.6% on food, Asians spend 17.5% on food, Native Americans spend

19.1%, other races spend 20.8%, whites spend 17.8%, and Hispanics (who may be of any race) spend 21.7%. (I will leave it as an exercise to determine if these are statistically significantly different.)

There are other differences: people in their 20s average 20.13%, in their 30s spend 18.1%, in their 40s it's down to 17.6%, in 50s 16.8%, then people 60 and up spend 17.8% (somewhat larger). There is a strong relationship with education as well: from those without a high-school diploma who spend 22.9% to those with an advanced degree who spend just 14.4% - suggesting that total income probably is important as well.

So how can we keep all of these different factors straight?

Multiple Regression in R

Chapter 3 of *Applied Econometrics in R* by Kleiber and Zeileis is terrific – gives an enormous amount of detail for how to do lots of different things! Most of this section of notes is based on material from that book. They created a package, `AER`, Applied Econometrics in R, which has lots of useful functions – so load that in.

From the standpoint of just using R, there is little difference for the user between a univariate and multivariate linear regression. Again use "`lm()`" but then add a bunch of variables to the model specification, so "`Y ~ X1 + X2 + X3`".

In formulas, model has k explanatory variables for each of $i = (1, 2, \ldots n)$ observations (must have n > k)

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_k x_{k,i} + \varepsilon_i$$

Each coefficient estimate, notated as $\hat{\beta}_j$, has standardized distribution as t with (n – k) degrees of freedom.

Each coefficient represents the amount by which the y would be expected to change, for a small change in the particular x-variable (i.e. $\beta_j = \frac{\partial y}{\partial x_j}$).

Note that you must be a bit careful specifying the variables. Educational attainment is might be coded with a bunch of numbers from 31 to 46 but these numbers have no inherent meaning. So too race, geography, industry, and occupation. If a person graduates high school then their grade coding changes from 38 to 39 but this must be coded with a dummy variable. If a person moves from New York to North Dakota then this increases their state code from 36 to 38; this is not the same change as would occur for someone moving from North Dakota to Oklahoma (40) nor is it half of the change as would occur for someone moving from New York to North Carolina (37). Each state needs a dummy variable.

A multivariate regression can control for all of the different changes to focus on each item individually. So we might model a household's fraction of expenditure on food as a function of their age, family size, gender of the reference person, race/ethnicity, educational level (high school diploma, some college but no degree, Associate's, a 4-year degree, or advanced degree), if they're married or divorced/widowed/separated, and so forth.

These results are:

```
Call:
lm(formula = fraction_food ~ AGE_REF + FAM_SIZE + female + AfAm +
    Asian + race_oth + Amindian + Hispanic + educ_hs + educ_smcoll +
    educ_as + educ_bach + educ_adv)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22494 -0.06511 -0.01622  0.04491  0.83229

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.282e-01  6.472e-03  35.251  < 2e-16 ***
AGE_REF     -4.059e-04  7.565e-05  -5.366 8.31e-08 ***
FAM_SIZE    -1.140e-03  8.719e-04  -1.308   0.1911
female      -4.303e-04  2.480e-03  -0.174   0.8622
AfAm         1.931e-02  3.771e-03   5.121 3.12e-07 ***
Asian        7.080e-03  5.812e-03   1.218   0.2232
race_oth     2.686e-02  1.067e-02   2.518   0.0118 *
Amindian     7.390e-03  1.370e-02   0.539   0.5896
Hispanic     3.055e-02  3.904e-03   7.824 5.88e-15 ***
educ_hs     -2.076e-02  3.995e-03  -5.197 2.08e-07 ***
educ_smcoll -3.235e-02  4.237e-03  -7.634 2.58e-14 ***
educ_as     -4.113e-02  5.024e-03  -8.187 3.17e-16 ***
educ_bach   -5.292e-02  4.306e-03 -12.292  < 2e-16 ***
educ_adv    -6.481e-02  5.296e-03 -12.238  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1018 on 6823 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.05925,  Adjusted R-squared:  0.05746
F-statistic: 33.06 on 13 and 6823 DF,  p-value: < 2.2e-16
```

Take the output a piece at a time. First it confirms what model you had called (useful when you go back later, after you've run lots of regressions). Next it gives a summary of the residuals,

$$\varepsilon_i = y_i - \hat{y} = y_i - \left(\widehat{\beta_0} + \widehat{\beta_1} x_{1,i} + \widehat{\beta_2} x_{2,i} + \cdots + \widehat{\beta_k} x_{k,i}\right)$$

These can be called at any point with "`residuals(model3)`" so the output is simply from "`summary(residuals(model3))`". The mean is not reported here since the model constrains the mean of the residuals to zero. The fitted values, $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_{1,i} + \widehat{\beta_2} x_{2,i} + \cdots + \widehat{\beta_k} x_{k,i}$, can be called as `fitted.values(model3)`.

Then R reports the coefficients, standard errors, t-statistics, and p-values for each term in the model. The coefficients and standard errors are calculated by the estimation routine.

The t-statistic is the ratio of the coefficient estimate divided by the standard error, $t = \frac{\hat{\beta}}{se\left(\hat{\beta}\right)}$.

The p-value is the area in the tails of a t-distribution (with degrees of freedom as shown on bottom line, here "6823 DF") beyond the t-statistic. The command, "`coefficients(model3)`", accesses the coefficient values.

At the bottom of the R summary it shows the R-squared, the standard error of the residual (which is basically the same as `sd(residuals(model3))`), and the F-statistic, which is another measure of how well the model fits.

Residuals are often used in analyses of productivity. Suppose I am analyzing a chain's stores to figure out which are managed best. I know that there are many reasons for variation in revenues and cost so I can get data on those: how many workers are there and their pay, the location of the store relative to traffic, the rent paid, any sales or promotions going on, etc. A regression on all of those factors delivers an estimate, $\hat{y}$, of what profit would have been expected, given external factors. Then the difference represents the unexplained or residual amount of variation: some stores would have been expected to be profitable and are indeed; some are not living up to potential; some would not have been expected to do so well but something is going on so they're doing much better than expected. But in general it's tricky to assign a name to the residual – unless that name is "ignorance.'

You should be able to calculate the t-statistic and p-value from the coefficient estimates and standard errors by yourself (the next homework will give you some chances to practice that).

You should also be able to calculate confidence intervals, although R can do that for you as well, with for example, `confint(model3,level = 0.95)`.

R will also produce lots of plots, simply with `plot(model3)`, which gives lots of plots in sequence – you can pick off particular ones with `plot(model3, which = 3)` that will give the 3rd plot. (The plots indicate that this might not be a great model.)

You can get an Analysis of Variance (ANOVA) with `anova(model3)`. For now don't worry about the details of the output except to the final row of figures, labeled "`Residuals`". This gives one of the most important bits of information about the model: how big are the residuals? Remember that's the whole point of the OLS estimator – it minimizes the (squared) residuals. So this gives you the value of the sum of squared residuals.

We often want to know particular predictions, for example we might want to know what the model would predict is the fraction of expenditure for a 30-year-old female, without anyone else in the household, who is African-American and has a bachelor's degree. To do this in R, we would first create the data frame then use the predict command:

```
to_be_predicted <- data.frame(AGE_REF = 30, FAM_SIZE = 1, female = 1,
                              AfAm = 1, Asian = 0, race_oth = 0, Amindian =0,
                              Hispanic = 0, educ_hs = 0, educ_smcoll = 0,
```

```
                                    educ_as = 0, educ_bach = 1, educ_adv = 0)
predict(model3, newdata = to_be_predicted, interval = "confidence")
```

There is a final detail, that we use `interval = "confidence"` if the x-values to be predicted are inside the values estimated, and `interval = "prediction"` if the x-values are outside.

Statistical significance of coefficient estimates is the same when we look at individual coefficients but more complicated for multiple coefficients: we can ask whether a group of variables are jointly significant, which takes a more complicated test. We can even ask if all of the slope coefficients together are statistically significant.

For a univariate regression, if the single slope coefficient is statistically significant then the overall regression is as well (the F statistic is the square of the t-stat in that case).

The difference between the overall regression fit and the significance of any particular estimate is that a hypothesis test of one particular coefficient tests if that parameter is zero; is $\beta_i = 0$? This uses the t-statistic $t = \frac{\hat{\beta}}{se\left(\hat{\beta}\right)}$ and compares it to a t distribution. The test of the regression significance tests if ALL of the slope coefficients are simultaneously zero; if $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_K = 0$. The latter is much more restrictive. (See Chapter 7 of Stock & Watson.)

It is often sensible to make joint tests of regression coefficients, for example with a group of dummy variables. If we have a set of dummies for education levels, it is strange to think of omitting just one or two; it is more reasonable to ask whether education measures (overall) are statistically significant. We might also want to know if individual coefficients are equal to each other (e.g. to ask if going to college, without getting any degree, is really different from the estimate for just a high school diploma.

To do this in R, there is a package, `linearHypothesis` (part of the package, `car`, Companion to Applied Regression, which is auto-loaded by `AER` package). But the commands shouldn't obscure the simple basic point: we evaluate variables based on how well they fit in the model.

To consider the question of whether a set of variables is statistically significant, we basically are just looking at how big is the error (the Sum of Squared Errors) with and without those variables. In general adding more variables to the model can never make the errors bigger (can never increase the Sum of Squared Errors) – basically this is a statement that the Marginal Benefit of more variables can never be negative. But profit maximization requires that we balance Marginal Benefit against Marginal Cost – what is the marginal cost of adding more variables? Statistical significance is one measure of profitability in this sense.

If adding new predictors makes the error "a lot" smaller, then those predictors are jointly statistically significant. The essence of statistical testing is just finding a good metric for "a lot".

Note that we can only properly make comparisons within models – it doesn't make much sense to look across models. If I have a model of the fraction of income spent on food, and another model of the level of income, it is difficulty to sensibly pose a question like, "in which model is education more important?" It would be like asking who scored more points per game, Shaq or Jeter? – you can ask the question but it's difficult to interpret in a sensible way.

But within a model we can make comparisons and many of them come down to asking, how much smaller are the errors? (Did the Sum of Squared Errors fall by a lot?) Sometimes it is easiest to just estimate the model twice, with or without the variables of interest, and look at how much the Sum of Squared Errors (from ANOVA in R) fell. But once you get some experience, you'll appreciate `linearHypothesis`.

Why do we always leave out a dummy variable? Multicollinearity. (See Chapter 6 of Stock & Watson.)

- OLS basic assumptions:
  - The conditional distribution of u_i given X_i has a mean of zero. This is a complicated way of saying something very basic: I have no additional information outside of the model, which would allow me to make better guesses. It can also be expressed as implying a zero correlation between X_i and u_i. We will work up to other methods that incorporate additional information.
  - The X and errors are i.i.d. This is often not precisely true; on the other hand it might be roughly right, and it gives us a place to start.
  - X and errors don't have values that are "too extreme." This is technical (about existence of fourth moments) and broadly true, whenever the X and Y data have a limit on the amount of variation, although there might be particular circumstances where it is questionable (sometimes in finance).
- So if these are true then the OLS are unbiased and consistent. So $E\left[\hat{\beta}_0\right] = \beta_0$ and $E\left[\hat{\beta}_1\right] = \beta_1$. The normal distribution, as the sample gets large, allows us to make hypothesis tests about the values of the betas. In particular, if you look back to the "eyeball" data at the beginning, you will recall that a zero value for the slope, β_1, is important. It implies no relationship between the variables. So we will commonly test the estimated values of β against a null hypothesis that they are zero.

Heteroskedasticity-consistent errors

You can choose to use heteroskedasticity-consistent errors as in the textbook.

The Stock and Watson textbook uses heteroskedasticity-consistent errors (sometimes called Eicker-Huber-White errors, after the various authors who figured out how to calculate them).