Statistics---inference in the linear model

I have barely even mentioned the t-test. Where does that come in? It's printed out every time we run a regression, so it must be useful.

Statistics---inference in the linear model

I have barely even mentioned the t-test. Where does that come in? It's printed out every time we run a regression, so it must be useful.

Before we talk about how to use it, let me remind you of the mathematical basis. Recall that, if the errors have a normal distribution, then so do the $\hat{\beta}$s. But their variances (and covariances) depend on the error variance, which we typically will not know. So when we substitute in $\hat{\sigma}^2$ for σ^2, the standardized version of $\hat{\beta}$ now has a t distribution, not a normal distribution any more.

Statistics---inference in the linear model

That's the mathematical justification for the t-test, but we often don't have or want to assume a normal distribution of the errors. We still use the t-test in that case, essentially as a way to make the hypothesis test a little more conservative than one based on a normal distribution, at least for small samples.

Statistics---inference in the linear model

So here's what it looks like for H_0: $\beta_i = c$:

$$T = (\hat{\beta}_i - c)/SE(\hat{\beta}_i) \text{ where } SE(\hat{\beta}_i) = (\sigma^2(X^TX)^{-1})_{ii}^{1/2}$$

Statistics---inference in the linear model

So here's what it looks like for H_0: $\beta_i = c$:

$$T = (\hat{\beta}_i - c)/SE(\hat{\beta}_i) \text{ where } SE(\hat{\beta}_i) = (\sigma^2(X^TX)^{-1})_{ii}^{1/2}$$

This picks out the ith diagonal element of the variance-covariance matrix.

Statistics---inference in the linear model

So here's what it looks like for H_0: $R\beta = c$:

$$T = (R\hat{\beta} - c)/SE(R\hat{\beta})$$

where $SE(R\hat{\beta}) = (\sigma^2 R(X^T X)^{-1} R^T)^{1/2}$

Statistics---inference in the linear model

So here's what it looks like for H_0: $R\beta = c$:

$$T = (R\hat{\beta} - c)/SE(R\hat{\beta})$$

where $SE(R\hat{\beta}) = (\sigma^2 R(X^T X)^{-1} R^T)^{1/2}$

Since this is a t-test, and we can only test one hypothesis (potentially involving multiple parameters), R is a 1x(k+1) matrix and c is a scalar here.

Statistics---inference in the linear model

Back to the question of when and how it's useful:

Well, for the hypothesis $H_0: \beta_j = c$ versus $H_A: \beta_j \neq c$, the F-test is equivalent to the t-test. (The t-test statistic and critical values are the square root of those for the F-test.)

So, you can use either, but it's easier to use the t-test for a single estimated coefficient if $H_0: \beta_j = 0$ since it's printed out right there for you.

One case where you *need* a t-test: if you want to carry out a one-sided test, like $H_0: \beta_j > 0$ versus $H_A: \beta_j < 0$.

Statistics---inference in the linear model

The F-test always given to us for free is the test of *all* coefficients (but not the intercept) being 0. The t-tests always given to us for free are the tests that each coefficient is 0. So, here, the F-test should be equivalent to the t-test for the coefficient on gss_data$year. Let's check: $(3.911)^2 = 15.296$. (They don't give us the critical values, but we could check that the t critical value squared is equal to the F critical value.)

```
> fit<-lm(gss_data$any_reason~gss_data$year)
> summary(fit)

Call:
lm(formula = gss_data$any_reason ~ gss_data$year)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3595 -2.1089 -0.1308  0.9966  5.4378 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -362.02694  102.99766  -3.515 0.001953 ** 
gss_data$year    0.20204    0.05166   3.911 0.000749 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.764 on 22 degrees of freedom
Multiple R-squared:  0.4101,	Adjusted R-squared:  0.3833 
F-statistic:  15.3 on 1 and 22 DF,  p-value: 0.000749
```

Lecture 19: Practical Issues in Running Regressions

Prof. Esther Duflo

14.310x

Practical issues with regression

- Dummy Variables
- Other Functional Form issues
- On example of Putting things together : Regression discontinuity Design

Dummy Variables

$$Y_i = \alpha + \beta D_i + \epsilon_i$$

D_i is a dummy variable , or an indicator variable, if it takes the value 1 if the observation is in group A, and 0 if in group B. Example:

- RCT: 1 if in treatment group , 0 otherwise
- 1 if male, 0 if female
- 1 before great depression, 0 after
- 1 before generic substitution act passed, 0 otherwise,
- 1 if the house has a deck in the backyard, 0 otherwise,

Interpretation

$$Y_i = \alpha + \beta D_i + \epsilon_i$$

Without any control variables, it is easy to verify that $\widehat{\beta} = \overline{Y_A} - \overline{Y_B}$.
So you can always estimate the difference between the treatment and control group for an RCT using an OLS regression framework. The standard errors will be slightly different from the Neyman standard errors we computed before (because the Neyman standard errors adjust for sample size of EACH group, whereas the OLS standard errors adjust for the size of the overall sample), but it won't matter that much if the samples are large enough, and similar in treatment and control groups.

From a categorical variable to dummy variables

- What if you don't have two groups, but, say, 50 (e.g. 50 states): Your original variable is takes discrete values 1 to 50.
- It usually does not make much sense to include it directly as a regressor
- Transform it into 50 dummy variables: for each state, the dummy $= 1$ if the observation is from that state, and 0 otherwise.
- Careful, what happens if you introduce all of them and the constant?

From a categorical variable to dummy variables

- What if you don't have two groups, but, say, 50 (e.g. 50 states): Your original variable is takes discrete values 1 to 50.
- It usually does not make much sense to include it directly as a regressor
- Transform it into 50 dummy variables: for each state, the dummy $= 1$ if the observation is from that state, and 0 otherwise.
- Careful, what happens if you introduce all of them and the constant?
- R will complain about multi-colinearity. We typically omit ONE of the categories
- So what do we do?

From a categorical variable to dummy variables

- What if you don't have two groups, but, say, 50 (e.g. 50 states): Your original variable is takes discrete values 1 to 50.
- It usually does not make much sense to include it directly as a regressor
- Transform it into 50 dummy variables: for each state, the dummy $= 1$ if the observation is from that state, and 0 otherwise.
- Careful, what happens if you introduce all of them and the constant?
- R will complain about multi-colinearity. We typically omit ONE of the categories
- So what do we do?
- We typically omit ONE group (if we don't do it, R may do it for us), and then what is the interpretation of each coefficient?

From a categorical variable to dummy variables

- What if you don't have two groups, but, say, 50 (e.g. 50 states): Your original variable is takes discrete values 1 to 50.
- It usually does not make much sense to include it directly as a regressor
- Transform it into 50 dummy variables: for each state, the dummy $= 1$ if the observation is from that state, and 0 otherwise.
- Careful, what happens if you introduce all of them and the constant?
- R will complain about multi-colinearity. We typically omit ONE of the categories
- So what do we do?
- We typically omit ONE group (if we don't do it, R may do it for us), and then what is the interpretation of each coefficient?
- It is the difference between the value of this group and the value for the omitted (reference) group.

with other variables in the regression

With other variables in the regression

$$Y_i = \alpha + \beta D_i + X_i \gamma + \epsilon_i$$

In that case β is the difference in intercept between group A and group B. This is the most frequent way that RCT are analyzed: the matrix X are "control" variables: things that did not affect the assignment but may have been different at baseline.

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$:

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group
- $\widehat{\beta}$:

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group
- $\widehat{\beta}$: An estimate of the difference between the treatment and control group means for women [we call this the treatment main effect]

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group
- $\widehat{\beta}$: An estimate of the difference between the treatment and control group means for women [we call this the treatment main effect]
- $\widehat{\gamma}$:

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group
- $\widehat{\beta}$: An estimate of the difference between the treatment and control group means for women [we call this the treatment main effect]
- $\widehat{\gamma}$: An estimate of the difference between Males and Females. [we call this the gender main effect]

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group
- $\widehat{\beta}$: An estimate of the difference between the treatment and control group means for women [we call this the treatment main effect]
- $\widehat{\gamma}$: An estimate of the difference between Males and Females. [we call this the gender main effect]
- $\widehat{\delta}$

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group
- $\widehat{\beta}$: An estimate of the difference between the treatment and control group means for women [we call this the treatment main effect]
- $\widehat{\gamma}$: An estimate of the difference between Males and Females. [we call this the gender main effect]
- $\widehat{\delta}$ An estimate of the difference between the treatment effect for males and for female. [we call this the interaction effect]

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group
- $\widehat{\beta}$: An estimate of the difference between the treatment and control group means for women [we call this the treatment main effect]
- $\widehat{\gamma}$: An estimate of the difference between Males and Females. [we call this the gender main effect]
- $\widehat{\delta}$ An estimate of the difference between the treatment effect for males and for female. [we call this the interaction effect]

How do you obtain, for example, an estimate of the mean for males?

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group
- $\widehat{\beta}$: An estimate of the difference between the treatment and control group means for women [we call this the treatment main effect]
- $\widehat{\gamma}$: An estimate of the difference between Males and Females. [we call this the gender main effect]
- $\widehat{\delta}$ An estimate of the difference between the treatment effect for males and for female. [we call this the interaction effect]

How do you obtain, for example, an estimate of the mean for males?
How do you obtain an estimate of the treatment effect for males?

Difference-in-Differences

- This is the basic "difference in differences" model which is often used by empirical researchers in a situation where there was a change in the law (or an event) affecting one group but not the other, and you are willing to assume that in the absence of the law, the difference between the two group would have remained stable over time
- In this case you have $D_i = 1$ if post law, 0 otherwise, and $G_i = 1$ if pre law, 0 otherwise.
- Famous examples: Mariel Boatlift experiment (David Card) ; New Jersey -Pennsylvania experiment (Card and Krueger)

Example : INPRES school construction program in Indonesia

Second five year plan (1974-79)-Oil shock.

- A large program:
  - 61,807 primary schools constructed from to 1973/74 to 1978/79.
    Number of schools multiplied by 2. 1 school for every 500 children.
  - A *change* in policy: Before 1973, no construction, ban on recruiting for public service positions.
- A program meant to favor low-enrollment regions.
  Allocation rule: number of schools constructed in a district was proportional to the number of children (ages 7 to 12) *not enrolled in primary school*.

Data Available

SUPAS 95: A survey done in 1995: after the children educated in these schools have completed their schooling, and have started working.

- 150,000 men born 1950-1972

- Variables: education, year and region of birth, wages.

Sources of variation

Two factors affect the intensity of the program.

- *Year of birth* :
- *Region of birth* The government was targeting low enrollment regions $\Rightarrow$ substantial variation in program intensity across districts.

Difference in difference

	Years of education			Log(wages)		
	Level of program in Region of birth			Level of program in Region of birth		
	High	Low	Difference	High	Low	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Experiment of Interest						
Aged 2 to 6 in 1974	8.49	9.76	-1.27	6.61	6.73	-0.12
	(0.043)	(0.037)	(0.057)	(0.0078)	(0.0064)	(0.010)
Aged 12 to 17 in 1974	8.02	9.40	-1.39	6.87	7.02	-0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Difference	0.47	0.36	0.12	-0.26	-0.29	0.026
	(0.070)	(0.038)	(0.089)	(0.011)	(0.0096)	(0.015)
Panel B: Control Experiment						
	8.00	9.41	-1.41	6.87	7.02	-0.15
Aged 12 to 17 in 1974	(0.054)	(0.042)	(0.078)	(0.0085)	(0.0069)	(0.011)
	7.70	9.12	-1.42	6.92	7.08	-0.16
Aged 18 to 24 in 1974	(0.059)	(0.044)	(0.072)	(0.0097)	(0.0076)	(0.012)
	0.30	0.29	0.013	0.056	0.063	0.0070
Difference	(0.080)	(0.061)	(0.098)	(0.013)	(0.010)	(0.016)

Note: The sample is made of the individuals who earn a wage. Standard errors are in parentheses

source: Duflo, 2001 "Schooling and Labor market consequence of school constructions in Indonesia: Evidence from an Unusual Experiment" American economic review.

More generally: Interactions

More generally, the coefficient on the interaction between dummy variable and some variable X tells us the extent to which the dummy variable changes the regression function for that regressor.

$$Y_i = \beta_0 + \beta_0^* D_i + \beta_1 X_{1i} + \beta^* D_i X_{1i} + \cdots + \epsilon_i$$

INPRES example: use variation across cohorts

$$S_{ijk} = c_1 + \alpha_{1j} + \beta_{1k} + (P_j * T_i)\gamma_1 + \epsilon_{ijk} \ , \tag{1}$$

where

- S_{ijk} is the education of individual i born in region j in year k,
- T_i is a dummy indicating whether the individual belongs to the "young" cohort in the subsample,
- P_j denotes the intensity of the program in the region of birth (number of school built)
- c_1 is a constant,
- β_{1k} is a set of cohort-of-birth fixed effects [in practice, a series of dummies=1 for each year of birth, omit 1]
- α_{1j} is a set of district-of-birth fixed effects [in practice, a series of dummies=1 for each district of birth, omit 1]

Table

	Observations	(1)	(2)	(3)	(4)	(5)	(6)
		Dependent variable					
		Years of education			Log(hourly wage)		
PANEL A: Experiment of Interest: Individuals Aged 2 to 6 or 12 to 17 in 1974 (Youngest Cohort: Individuals Ages 2 to 6 in 1974)							
Whole sample	78,470	0.124	0.15	0.188			
		(0.0250)	(0.0260)	(0.0289)			
Sample of wage earners	31,061	0.196	0.199	0.259	0.0147	0.0172	0.0270
		(0.0424)	(0.0429)	(0.0499)	(0.00729)	(0.00737)	(0.00850)
PANEL B: Control Experiment : Individuals Aged 12 to 24 in 1974 (Youngest Cohort: Individuals Ages 12 to 17 in 1974)							
Whole sample	78,488	0.0093	0.0176	0.0075			
		(0.0260)	(0.0271)	(0.0297)			
Sample of wage earners	30,225	0.012	0.024	0.079	0.0031	0.00399	0.0144
		(0.0474)	(0.0481)	(0.0555)	(0.00798)	(0.00809)	(0.00915)
Control variables:							
Year of birth*enrollment rate in 1971		No	Yes	Yes	No	Yes	Yes
Year of birth* water and sanitation program		No	No	Yes	No	No	Yes

The coefficient γ tells us that the difference in education between the young cohort and the old cohort is 0.124 year larger for each school built per 1000 kids.

Figure

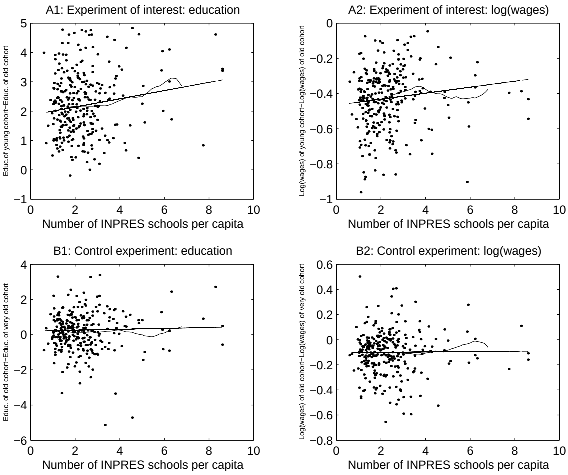

Figure 1: Regional growth in education and log wages accross cohort and program intensity

(Per capita denotes per 1000 children)

Practical issues with regression

- Dummy Variables
- Other Functional Form issues
- One example of putting things together: Regression discontinuity design

Other functional form issues

- Transforming the dependent variable
- Non linear transformations of the independent variables

Transformations of the dependent variable

- Suppose $Y_i = AX_{1i}^{\beta_1} X_{2i}^{\beta_2} e^{\epsilon_i}$ then run linear regression

$$log(Y_i) = \beta_0 + \beta_1 logX_{1i} + \beta_2 logX_{12} + \epsilon_i$$

  to estimate β_1 and β_2. Note that β_1 and β_2 are *elasticities*: when X_1 changes by 1%, Y changes by β_1%.

- Returns to education formulation

$$logY_i = \beta_0 + \beta_1 S_i + \epsilon_i$$

  When education increases by 1 year, wages increase by $\beta_1 \times 100\%$.

Transformations of the dependent variable

- Box Cox Transformation
  Suppose $Y_i = \frac{1}{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\epsilon_i}$
  then run regression

$$\frac{1}{Y_i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

- Discrete choice model
  Suppose

$$P_i = \frac{e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\epsilon_i}}{1+e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\epsilon_i}}$$

  P_i is the percentage of individuals choosing a particular option (e.g. buying a particular car)
  then run regression:

$$Y_i = log(\frac{Pi}{1-P_i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$