Practical Regression Hints

Po-Chun Huang
NCCU, Spring 2021

Outline

- Units of Measurement and Functional Form
- Dummy Variables and Interactions
- Practical Regression Hints

- Units of Measurement and Functional Form

1. Units of Measurement and Functional Form

TABLE 2.3 Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

© Cengage Learning, 2013

Example Using NLSY

- On average, one year increase in years of education is associated with 0.89 increase in hourly wage

```
. reg wage school
```

Source	SS	df	MS		
Model	26626.3053	1	26626.3053	Number of obs	= 5,898
Residual	247524950	5,896	41981.8436	F(1, 5896)	= 0.63
				Prob > F	= 0.4258
Total	247551576	5,897	41979.2396	R-squared	= 0.0001
				Adj R-squared	= -0.0001
				Root MSE	= 204.89

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
school	.8953485	1.124262	0.80	0.426	-1.308616	3.099313
_cons	.0085636	14.54931	0.00	1.000	-28.51341	28.53053

Example Using NLSY

- On average, one year increase in years of education is associated with 7.68% increase in hourly wage

```
. gen lnwage=log(wage)

. reg lnwage school
```

Source	SS	df	MS	Number of obs	=	5,898
				F(1, 5896)	=	435.38
Model	196.236742	1	196.236742	Prob > F	=	0.0000
Residual	2657.48287	5,896	.450726402	R-squared	=	0.0688
				Adj R-squared	=	0.0686
Total	2853.71961	5,897	.483927354	Root MSE	=	.67136

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
school	.0768647	.0036838	20.87	0.000	.0696431	.0840862
_cons	.8257507	.0476725	17.32	0.000	.7322951	.9192062

Example Using NLSY

- On average, one hundred percent increase in years of schooling is associated with 10.95 increase in hourly wage

```
. gen lnschool=log(school)
(2 missing values generated)

. reg wage lnschool
```

Source	SS	df	MS		
Model	27334.8591	1	27334.8591	Number of obs	= 5,896
Residual	247524105	5,894	41995.9459	F(1, 5894)	= 0.65
				Prob > F	= 0.4198
Total	247551440	5,895	41993.4588	R-squared	= 0.0001
				Adj R-squared	= -0.0001
				Root MSE	= 204.93

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnschool	10.95816	13.5826	0.81	0.420	-15.66871	37.58502
_cons	-16.27237	34.40568	-0.47	0.636	-83.72011	51.17537

Example Using NLSY

- On average, one hundred percent increase in years of schooling is associated with 88 percent increase in hourly wage

```
. reg lnwage lnschool
```

Source	SS	df	MS		
Model	176.300231	1	176.300231	Number of obs	= 5,896
Residual	2676.52194	5,894	.454109594	F(1, 5894)	= 388.23
				Prob > F	= 0.0000
Total	2852.82217	5,895	.4839393	R-squared	= 0.0618
				Adj R-squared	= 0.0616
				Root MSE	= .67388

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnschool	.8800464	.0446642	19.70	0.000	.7924882	.9676045
_cons	-.4186734	.1131376	-3.70	0.000	-.6404645	-.1968824

- Dummy Variables and Interactions

Dummy Variables

- Use regression to estimate gender earning differential

$$\ln wage = \beta_0 + \beta_1 male + u$$

- male: a variable that's equal to 1 for males and 0 for females

Dummy Variables

```
. reg lnwage male
```

Source	SS	df	MS	Number of obs	=	5,898
				F(1, 5896)	=	146.87
Model	69.3584016	1	69.3584016	Prob > F	=	0.0000
Residual	2784.36121	5,896	.472245795	R-squared	=	0.0243
				Adj R-squared	=	0.0241
Total	2853.71961	5,897	.483927354	Root MSE	=	.6872

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	.2171958	.017922	12.12	0.000	.1820621	.2523294
_cons	1.689189	.0130078	129.86	0.000	1.663689	1.714689

Dummy Variables

```
. reg lnwage male school
```

Source	SS	df	MS
Model	289.496622	2	144.748311
Residual	2564.22299	5,895	.434982695
Total	2853.71961	5,897	.483927354

Number of obs	=	5,898	
F(2, 5895)	=	332.77	
Prob > F	=	0.0000	
R-squared	=	0.1014	
Adj R-squared	=	0.1011	
Root MSE	=	.65953	

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	.2529257	.0172735	14.64	0.000	.2190632	.2867881
school	.0817576	.0036343	22.50	0.000	.0746331	.0888821
_cons	.6302659	.0486983	12.94	0.000	.5347995	.7257324

Dummy Variables and Interactions

```
. gen male_school=male*school

. reg lnwage male school male_school
```

Source	SS	df	MS
Model	294.141681	3	98.047227
Residual	2559.57793	5,894	.434268396
Total	2853.71961	5,897	.483927354

Number of obs	=	5,898	
F(3, 5894)	=	225.78	
Prob > F	=	0.0000	
R-squared	=	0.1031	
Adj R-squared	=	0.1026	
Root MSE	=	.65899	

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	.5576785	.0947668	5.88	0.000	.3719009	.7434561
school	.0949798	.0054342	17.48	0.000	.0843267	.1056329
male_school	-.0238893	.0073045	-3.27	0.001	-.0382088	-.0095699
_cons	.4590116	.0714809	6.42	0.000	.3188828	.5991404

Dummy Variables and Interactions

- Test the return to education are the same

  H0: $\beta_3 = 0$

- Test expected earnings conditional on education are the same:

  H0: $\beta_1 = \beta_3 = 0$

```
. test male=male_school=0

 ( 1)  male - male_school = 0
 ( 2)  male = 0

       F(  2,  5894) =  112.72
            Prob > F =    0.0000
```

Dummy Variables for Multiple Categories

- If there are *n* categories, you have to include *n-1* dummies average earning of each category
- Example: race = 1 for non-Hispanic African-Americans, 2 for Hispanics, and 3 for non-Hispanic Caucasians

  - Include two dummies to estimate the average earnings of each race group

$$\ln wage = \beta_0 + \beta_1 H + \beta_2 W + u$$

  - H=1 if Hispanics
  - W=1 if non-Hispanic Caucasians

Dummy Variables for Multiple Categories

```
. reg wage i.race
```

Source	SS	df	MS
Model	74127.2826	2	37063.6413
Residual	247477449	5,895	41980.9073
Total	247551576	5,897	41979.2396

Number of obs	=	5,898	
F(2, 5895)	=	0.88	
Prob > F	=	0.4136	
R-squared	=	0.0003	
Adj R-squared	=	-0.0000	
Root MSE	=	204.89	

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
race						
2	.8503192	7.848541	0.11	0.914	-14.5357	16.23634
3	7.396608	6.135244	1.21	0.228	-4.630719	19.42393
_cons	7.450497	4.872869	1.53	0.126	-2.102113	17.00311

Interactions

- You want to measure the relationship between earnings and marital status, and whether this relationship differs by gender. You estimate the following regression:

$$Y = \beta_0 + \beta_1 MARRIED + \beta_2 FEMALE + \beta_3 FEMXMARRIED + u$$

- Y = log wages, MARRIED = 1 if married and 0 otherwise, FEMALE = 1 if female and 0 otherwise, and FEMXMARR is the interaction FEMALE×MARRIED

Interactions

```
. reg Y MARRIED FEMALE FEMXMARR
```

Source	SS	df	MS		
Model	43.1239872	3	14.3746624	Prob > F	0.0000
Residual	168.323763	528	.318795006	R-squared	0.2039
Total	211.447751	531	.398206687	Root MSE	.56462

	Number of obs =	532
	F(3, 528) =	45.09
	Prob > F =	0.0000
	R-squared =	0.2039
	Adj R-squared =	0.1994
	Root MSE =	.56462

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
MARRIED	-.0028192	.0993045	-0.03	0.977	-.1978995	.1922612
FEMALE	-.5405201	.1180226	-4.58	0.000	-.7723716	-.3086686
FEMXMARR	-.0421958	.1302137	-0.32	0.746	-.2979963	.2136047
_cons	1.690307	.0928228	18.21	0.000	1.50796	1.872655

Interactions

- What is the estimated average of log wages for unmarried men? What about unmarried women?

- What is the estimated effect of being married on log wages for men? What is the estimated effect of being married on log wages for women?

- How would you test the null hypothesis that the effect of being married on log wages does not vary with gender? Write down the null hypothesis in terms of the regression coefficients. Do you reject the null hypothesis?

- Practical Regression Hints

2. Practical Regression Hints

1. Do not always attempt to maximize *R*-squared, adjusted *R*-squared, or some other goodness-of-fit measure. Might inadvertently include in **x** factors that should not be held fixed.

EXAMPLE: y is individual or family demand for a product, $x_1, \ldots, x_{k-1}$ include various product prices, income, and demographics. Should we include the demand for a competing product as x_k? Usually does not make sense to hold a quantity demanded fixed and change the price of any good (bad control problem)

- It is possible to obtain a convincing estimate of a causal effect with a low *R*-squared. For example, under random assignment, a simple regression estimate consistently estimates the causal effect, but the "treatment" may not explain much of the variation in *y*.

- More precisely, in the equation

$$y = \beta_0 + \beta_1 x + u,$$

the question of whether *u* is correlated with *w* is very different from the relative sizes of *Var(y)* and *Var(u)*.

- There are many other factors that explain the health indicator and health expenditures other than the treatment, but there can be no “omitted variable bias” because of random assignment.

2. Be careful in interpreting models nonlinear in explanatory variables, especially with interactions. Coefficients on level terms may become essentially meaningless.

Example: the effect of log hourly wage on minutes of sleep..

```
. reg sleep lhrwage
```

Source	SS	df	MS		
Model	445336.018	1	445336.018	Prob > F	0.1216
Residual	98181344.4	530	185247.82	R-squared	0.0045
				Adj R-squared	0.0026
Total	98626680.4	531	185737.628	Root MSE	430.4

Number of obs	=	532
F(1, 530)	=	2.40

sleep	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lhrwage	-45.89257	29.59886	-1.55	0.122	-104.0381	12.25292
_cons	3325.137	46.2837	71.84	0.000	3234.215	3416.059

```
. reg sleep lhrwage educ
```

Source	SS	df	MS		
Model	668573.104	2	334286.552	Prob > F	0.1654
Residual	97958107.3	529	185176.006	R-squared	0.0068
Total	98626680.4	531	185737.628	Root MSE	430.32

	Number of obs	=	532
	F(2, 529)	=	1.81
	Prob > F	=	0.1654
	R-squared	=	0.0068
	Adj R-squared	=	0.0030
	Root MSE	=	430.32

sleep	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lhrwage	-36.61547	30.7757	-1.19	0.235	-97.07306	23.84212
educ	-7.91636	7.209991	-1.10	0.273	-22.08009	6.247368
_cons	3412.647	92.1608	37.03	0.000	3231.601	3593.693

```
. gen lhrwage_educ=lhrwage*educ
(174 missing values generated)

. reg sleep lhrwage educ lhrwage_educ
```

Source	SS	df	MS		
Model	691238.925	3	230412.975	Prob > F	0.2937
Residual	97935441.5	528	185483.791	R-squared	0.0070
				Adj R-squared	0.0014
Total	98626680.4	531	185737.628	Root MSE	430.68

Number of obs	=	532
F(3, 528)	=	1.24

sleep	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lhrwage	5.036041	123.0679	0.04	0.967	-236.7267	246.7988
educ	-3.395601	14.80935	-0.23	0.819	-32.48808	25.69688
lhrwage_educ	-3.262565	9.33311	-0.35	0.727	-21.59715	15.07202
_cons	3356.447	185.3494	18.11	0.000	2992.334	3720.56

```
. reg sleep lhrwage educ c.lhrwage#c.educ
```

Source	SS	df	MS
Model	691238.949	3	230412.983
Residual	97935441.4	528	185483.791
Total	98626680.4	531	185737.628

Number of obs	= 532
F(3, 528)	= 1.24
Prob > F	= 0.2937
R-squared	= 0.0070
Adj R-squared	= 0.0014
Root MSE	= 430.68

sleep	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lhrwage	5.036064	123.0679	0.04	0.967	-236.7267	246.7988
educ	-3.395599	14.80935	-0.23	0.819	-32.48808	25.69688
c.lhrwage#c.educ	-3.262566	9.33311	-0.35	0.727	-21.59715	15.07202
_cons	3356.447	185.3494	18.11	0.000	2992.334	3720.56

- Little variation in education around zero.

```
. sum sleep lhrwage educ
```

Variable	Obs	Mean	Std. Dev.	Min	Max
sleep	706	3266.356	444.4134	755	4695
lhrwage	532	1.430977	.6310362	-1.049822	3.569814
educ	706	12.78045	2.784702	1	17

- The coefficient on *lhrwage* measures the effect of one hundred percent increase in hourly wage on minutes of sleep for individuals with “zero year of education”.
- The variables *lhrwage* and l*hrwage_educ* are probably highly collinear.
- The collinearity is a result of trying to estimate a poorly identified parameter: the partial effect of *lhrwage* at *educ*=0 is poorly identified and uninteresting.

Solution I: centering

```
. qui sum educ

. gen educ_mean=r(mean)

. gen c_educ=educ-educ_mean

. gen lhrwage_ceduc=lhrwage*c_educ
(174 missing values generated)

. reg sleep lhrwage educ lhrwage_ceduc
```

Source	SS	df	MS	Number of obs	=	532
				F(3, 528)	=	1.24
Model	691238.947	3	230412.982	Prob > F	=	0.2937
Residual	97935441.4	528	185483.791	R-squared	=	0.0070
				Adj R-squared	=	0.0014
Total	98626680.4	531	185737.628	Root MSE	=	430.68

sleep	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lhrwage	-36.66102	30.80154	-1.19	0.234	-97.16964	23.8476
educ	-3.395599	14.80935	-0.23	0.819	-32.48808	25.69688
lhrwage_ceduc	-3.262566	9.33311	-0.35	0.727	-21.59715	15.07202
_cons	3356.447	185.3494	18.11	0.000	2992.334	3720.56

Solution II: use margins command

- This table is the same as the table on page 27.

```
. reg sleep lhrwage educ c.lhrwage#c.educ
```

Source	SS	df	MS
Model	691238.949	3	230412.983
Residual	97935441.4	528	185483.791
Total	98626680.4	531	185737.628

Number of obs	= 532
F(3, 528)	= 1.24
Prob > F	= 0.2937
R-squared	= 0.0070
Adj R-squared	= 0.0014
Root MSE	= 430.68

sleep	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lhrwage	5.036064	123.0679	0.04	0.967	-236.7267	246.7988
educ	-3.395599	14.80935	-0.23	0.819	-32.48808	25.69688
c.lhrwage#c.educ	-3.262566	9.33311	-0.35	0.727	-21.59715	15.07202
_cons	3356.447	185.3494	18.11	0.000	2992.334	3720.56

- Can evaluate the effect at the mean of education using stata margins command

```
. margins, dydx(lhrwage) atmeans

Conditional marginal effects                    Number of obs     =        532
Model VCE    : OLS

Expression   : Linear prediction, predict()
dy/dx w.r.t. : lhrwage
at           : lhrwage         =    1.430977 (mean)
               educ            =     12.7312 (mean)
```

	Delta-method dy/dx	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lhrwage	-36.50033	30.80303	-1.18	0.237	-97.01187	24.0112