14.32 Recitation 5

Nina Wang

MIT Department of Economics

Lecture 10

Table of Contents

1. Nonlinear Regressions
2. Practice Problems

Table of Contents

1 Nonlinear Regressions

2 Practice Problems

Polynomial Regressions

$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + ... + \beta_r X_i^r + \gamma Z_i + e_i$$

- We use a polynomial regression when we suspect that there might be a polynomial relationship between independent and dependent variables.
    - Example: We might suspect that the returns to schooling on earnings increase over time (ie difference between 4-5 years of schooling is different from difference between 11-12 years of schooling).

Polynomial Regressions

$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + ... + \beta_r X_i^r + \gamma Z_i + e_i$$

- What if we're unsure of whether to use polynomial terms or not?
  - We can test the coefficient on the polynomial term. If it is significant, then the polynomial model has a better fit than the linear model.

Polynomial Regressions

$educ^2$

```
. reg earnings height educ educ2 sex
```

Source	SS	df	MS			
Model	2.0446e+12	4	5.1114e+11	Number of obs	=	17,870
Residual	1.0908e+13	17,865	610581095	F(4, 17865)	=	837.14
				Prob > F	=	0.0000
				R-squared	=	0.1578
				Adj R-squared	=	0.1577
Total	1.2953e+13	17,869	724863544	Root MSE	=	24710

earnings	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
height	389.3636	66.30547	5.87	0.000	259.3984	519.3287
educ	555.2525	386.5061	1.44	0.151	-202.3368	1312.842
educ2	128.2131	14.58987	8.79	0.000	99.61554	156.8107
sex	321.4227	526.9841	0.61	0.542	-711.5171	1354.363
_cons	-11247.46	4669.168	-2.41	0.016	-20399.48	-2095.442

> 1.96

Log Regressions

- Log transformations alter the way we interpret regressions
- There are 3 types of log transformed regressions
  - Log-linear
  - Linear-Log
  - Log-Log

Log-Linear Regressions

$$\ln(Y_i) = \alpha + \beta X_i + \gamma Z_i + e_i$$

- The dependent variable is log transformed.
- One unit change in X increases/decreases Y by $\beta \cdot 100\%$

Log-Linear Regressions

$$\ln(Earnings_i) = \alpha + \beta_1 height_i + \beta_2 educ_i + e_i$$

```
. reg logearnings height educ sex
```

Source	SS	df	MS
Model	1267.86936	3	422.623121
Residual	6562.55097	17,866	.367320664
Total	7830.42034	17,869	.438212566

Number of obs	=	17,870
F(3, 17866)	=	1150.56
Prob > F	=	0.0000
R-squared	=	0.1619
Adj R-squared	=	0.1618
Root MSE	=	.60607

logearnings	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
height	.0107823	.0016195	6.66	0.000	.0076078	.0139567
educ	.0973644	.0017373	56.04	0.000	.0939592	.1007696
sex	.0119908	.0128519	0.93	0.351	-.0132001	.0371818
_cons	8.515671	.1036177	82.18	0.000	8.31257	8.718772

Linear-Log Regressions

$$Y_i = \alpha + \beta \ln(X_i) + \gamma Z_i + e_i$$

- The independent variable is log transformed
- One percent change in X is associated with change in Y of 0.01β units.

Example

$$Y_i = 450.2 + 62.35 \ln(X_i) + e_i$$

- Where Y_i is math SAT score, X_i is the expenditure per student
- How do we interpret the 62.35?

Example

$$Y_i = 450.2 + 62.35 \ln(X_i) + e_i$$

- Where Y_i is math SAT score, X_i is the expenditure per student
- How do we interpret the 62.35?
- With every 1% increase in expenditure, expected SAT score increases by 0.62 points.

Log-Log Regressions

$$\ln(Y_i) = \alpha + \beta \ln(X_i) + \gamma Z_i + e_i$$

- The independent variable and dependent variable are log transformed
- One percent change in X is associated with a $\beta\%$ change in Y.
- Also known as elasticity

Example

$$\ln(Y_i) = 0.4 + 2.3\ln(X_i) + e_i$$

- Where X_i is price of good p and Y_i is demand (quantity sold in units).
- How do we interpret 2.3?

Example

$$Y_i = X_1^{\beta_1} \cdot X_2^{\beta_2} \dots$$

$$\log Y_i = \beta_1 \log X_1 + \beta_2 \log X_2$$

$$\ln(Y_i) = 0.4 + 2.3 \ln(X_i) + e_i$$

- Where X_i is price of good p and Y_i is demand (quantity sold in units).
- How do we interpret 2.3?
- A 1% increase in price of good p is associated with a 2.3% increase in quantity of goods sold.

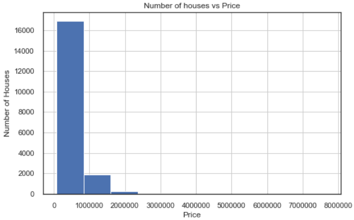
Why do we interpret in percents?

$$\ln(x+c) - \ln(x) = \ln\left(\frac{x+c}{x}\right) = \ln\left(1 + \frac{c}{x}\right) \approx \frac{c}{x}$$

- This approximation holds for small $\frac{c}{x}$, works for changes of up to 10%
- In reality, change is small enough that we are almost always able to interpret log transformation in percentages.

Why do we use logs?

- A logarithmic transformation is useful for transforming highly skewed variables into a more normalized dataset.



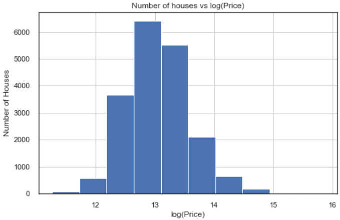


Table of Contents

1 Nonlinear Regressions

2 Practice Problems

Practice

Dependent variable: *"Course Overall" evaluation score*

	(1)	(2)	(3)	(4)	(5)	(6)
Data subset:	All instructors	All instructors	All instructors	All instructors	Male instructors	Female instructors
Regressor						
Beauty	.410 (.081)	.275 (.059)	.229 (.047)	.237 (.096)	.384 (.076)	.128 (.064)
Female	-.166 (.098)	-.239 (.085)	-.210 (.075)	-.255 (.088)	–	–
Minority	-.284 (.015)	-.249 (.012)	-.206 (.014)	-.221 (.012)	.060 (.101)	-.260 (.139)
Non-native English	-.344 (.152)	-.253 (.134)	-.288 (.112)	-.251 (.132)	-.427 (.143)	-.262 (.151)
tenure track	-.150 (.114)	-.136 (.094)	-.156 (.110)	-.131 (.092)	-.056 (.089)	-.041 (.133)
intro course	-.071 (.134)	-.046 (.111)	-.079 (.102)	-.052 (.110)	.005 (.129)	-.228 (.164)
one-credit course (yoga, aerobics, dance, short electives)	–	.687 (.166)	.823 (.129)	.694 (.170)	.768 (.119)	.517 (.232)
dresses well	–	–	.243 (.088)	–	–	–
$Beauty \times D_{Beauty>0}$	–	–	–	.081 (.135)	–	–
Intercept	4.27 (.071)	4.25 (0.56)	4.22 (.054)	4.21 (.054)	4.35 (.081)	4.08 (.088)
Summary statistics						
R^2	.224	.279	.302	.285	.359	.162
n	463	463	463	463	268	195

Practice 1a

- The following variables are not included in regression 2:
  - Amount of time instructor spends in class
  - Marital status of instructor
- For each, explain whether omission of the variablße will result in OVB for the estimated effect of Beauty.

Practice 1b

- Suppose you have data on years of teaching experience (*Experience*) of the instructor, and you are considering choosing among three possible specifications:
  - regression (2) plus *Experience* ✓
  - regression (2) plus *Experience*, $Experience^2$, and $Experience^3$
  - regression (2) plus log(*Experience*)
- In your judgment (before you know the results of these regressions), which specification, (i), (ii), or (iii), is the most appropriate? Explain.
- Suppose you estimated regressions for specifications (i) and (ii). How would you decide, based on the empirical evidence, whether (i) or (ii) is more appropriate?

Practice 2

- In a given population of two-earner male-female couples, male earnings have a mean of \$40,000 per year and a standard deviation of \$12,000. Female earnings have a mean of \$45,000 per year and a standard deviation of \$18,000. The correlation between male and female earning for a couple is 0.8. Let C denote the combined earnings for a randomly selected couple. What is the mean and the standard deviation of C? $C = M + F$

$$Corr(M,F) = \frac{Cov(M,F)}{SD(M)SD(F)}$$

$$E[C] = E[M] + E[F]$$

$$Var[C] = Var[M] + Var[F] + 2Cov(M,F)$$

Practice 2

- In a given population of two-earner male-female couples, male earnings have a mean of \$40,000 per year and a standard deviation of \$12,000. Female earnings have a mean of \$45,000 per year and a standard deviation of \$18,000. The correlation between male and female earning for a couple is 0.8. Let C denote the combined earnings for a randomly selected couple. What is the mean and the standard deviation of C?
- $E[C] = E[M] + E[F] = 40 + 45 = 85$
- $V[C] = V[M + F] = V[M] + V[F] + Cov[M, F] = 28.52$
  - $Cov[M, F] = Corr[M, F] * SD[M] * SD[F]$

Practice 3

T, F, Uncertain

- One runs a regression $Y_i = \gamma_0 + \gamma_1 X_i + e_i$ and gets $\hat{\gamma}_1 = 0$, then it implies that $R^2 = 0$

$\hat{Y} = \gamma_0$

$R^2 = \frac{ESS}{TSS} \leftarrow \sum_{i=1} (\hat{Y}_i - \bar{Y})^2$

$= 0$

$\alpha = \bar{Y} - \beta \bar{X}$

$\gamma_0 = \bar{Y} - \gamma_1 \bar{X}$

$\gamma_0 = \bar{Y}$

True

Practice 3

- One runs a regression $Y_i = \gamma_0 + \gamma_1 X_i + e_i$ and gets $\hat{\gamma}_1 = 0$, then it implies that $R^2 = 0$

- $R^2 = \frac{ESS}{TSS}$
- ESS (Explained Sum of Squares) $\Sigma(\hat{Y}_i - \bar{Y})^2$
- TSS (Total Sum of Squares) $\Sigma(Y_i - \bar{Y})^2$

Practice 3

- One runs a regression $Y_i = \gamma_0 + \gamma_1 X_i + e_i$ and gets $\hat{\gamma}_1 = 0$, then it implies that $R^2 = 0$

- $R^2 = \frac{ESS}{TSS}$
- ESS (Explained Sum of Squares) $\Sigma(\hat{Y}_i - \bar{Y})^2$
- TSS (Total Sum of Squares) $\Sigma(Y_i - \bar{Y})^2$
- Answer: True
  - $\gamma_0 = \bar{Y} - \gamma_1 \bar{X} = \bar{Y}$
  - $\hat{Y} = \bar{Y}$, for all i, thus ESS $= 0$

Practice 4

- A Cobb-Douglas production function relates production Q to factors of production such as capital K, labor L, and raw materials M and an error term using the equation $Q = \gamma K^{\beta_1} L^{\beta_2} M^{\beta_3} e^{i}$, where $\gamma, \beta_1, \beta_2, \beta_3$ are unknown production parameters. Suppose that you have data on production and the factors of production for a random sample of firms. By transforming the data you can use OLS regression to estimate $\beta_1, \beta_2, \beta_3$

Practice 4

T/F/Uncertain

- A Cobb-Douglas production function relates production Q to factors of production such as capital K, labor L, and raw materials M and an error term using the equation $Q = \gamma K^{\beta_1} L^{\beta_2} M^{\beta_3} e^{i}$, where $\gamma, \beta_1, \beta_2, \beta_3$ are unknown production parameters. Suppose that you have data on production and the factors of production for a random sample of firms. By transforming the data you can use OLS regression to estimate $\beta_1, \beta_2, \beta_3$

- We can regress $\log Q$ on $\log K, \log L, \log M$

$$\log Q = \log \gamma + \beta_1 \log K + \dots$$