Nonlinear Regression Functions

SW Chapter 8

Overview of nonlinear regression models

Polynomial regression

Logarithmic functions

Interaction terms

- Two binary variables
- One binary, one continuous variables
- Two continuous variables

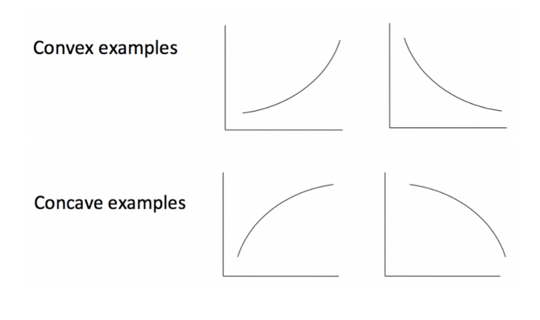
Learning objectives

- Estimate and interpret linear regressions that are functions of one variable
  - Polynomials
  - Logarithms
- Estimate and interpret linear regressions with non-linear functions of two variables: interaction terms!

Overview of nonlinear regression models

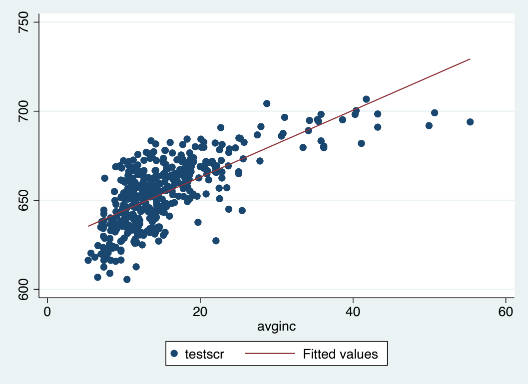
Three types of tests

- So far, we have assumed a linear relationship between Y_i and X_i
- In reality, the relationship between variables is typically non-linear
- Could be convex, concave, or something more complicated!



Income and test scores

Effect of average per-capita income in a school district on test scores



General nonlinear population regression function

$$Y_i = f(X_{1i}, X_{2i}, ..., X_{ki}) + u_i, i = 1, 2, ..., n$$

Assumptions (same):

1. $E[u_i| = X_{1i}, X_{2i}, ..., X_{ki}) = 0$
2. $(X_{1i}, X_{2i}, ..., X_{ki})$ are i.i.d.
3. Big outliers are rare
4. No perfect multicollinearity

The change in Y associated with a change in X_{1i}, holding $X_2, ..., X_k$ constant is:

$$\Delta Y = f(X_1 + \Delta X_1, X_2, ..., X_k) - f(X_1, X_2, ..., X_K)$$

Polynomial regression

Quadratic regression

$$Y_I = \beta_0 + \beta_1 X_i + \beta_2 X^2 + u_i$$

- X_i and Y_i have a non-linear relationship
- β_1 does not measure the effect of a one-unit change in X_i on Y_i: if X_i changes, it is necessarily true that X_i^2 also changes
- The effect of a one-unit change in Y_i depends on *both* β_1 and β_2

A mathematical explanation:

Old model: $Y_i = \beta_0 + \beta_1 X_i + u_i$

- $\dfrac{\partial Y_i}{\partial X_i} = \beta_1$

New Model: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$

- $\dfrac{\partial Y_i}{\partial X_i} = \beta_1 + \beta_2 X_i$
- The effect of a one-unit change in X_i depends on X_i
- If $\beta_2 > 0$, then the effect grows with X_i
- If $\beta_2 < 0$, then the effect diminishes with X_i

Test scores and income

$$TestScore_i = \beta + \beta_1 Income_i + \beta_2 Income_i^2 + u_i$$

- Use data on average income in a school district
- Allow a nonlinear relationship between test score and income
- In Stata, generate the $Income^2$ variable before you include it:

```
gen income2 = income^2
```

Test scores and income

```
. regress testscr avginc, robust

Linear regression                               Number of obs     =        420
                                                F(1, 418)         =     273.29
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5076
                                                Root MSE          =     13.387

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |    1.87855   .1136349    16.53   0.000     1.655183    2.101917
       _cons |   625.3836   1.867872   334.81   0.000      621.712    629.0552
------------------------------------------------------------------------------
```

Test scores and income

```
. gen avginc2 = avginc^2

.  regress testscr avginc avginc2, robust

Linear regression                               Number of obs     =        420
                                                F(2, 417)         =     428.52
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5562
                                                Root MSE          =     12.724

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |   3.850995   .2680941    14.36   0.000      3.32401    4.377979
     avginc2 |  -.0423085   .0047803    -8.85   0.000     -.051705   -.0329119
       _cons |   607.3017   2.901754   209.29   0.000     601.5978    613.0056
------------------------------------------------------------------------------
```

Test scores and income

$$\widehat{TestScore}_i = 607.3 + 3.85 Income_i - 0.042 Income_i^2$$

Is income positively or negatively associated with test scores?

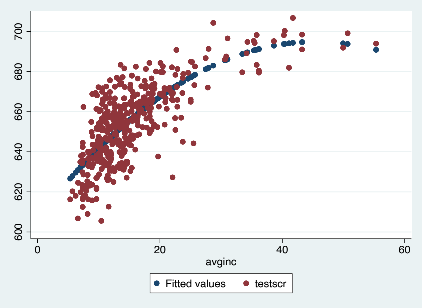
- You could plug in values to see what happens as X changes
  - $Income = 10$, $TestScore = 641.6$
  - $Income = 11$, $TestScore = 644.6$ (3-point increase)
  - $Income = 30$, $TestScore = 685$
  - $Income = 31$, $TestScore = 686.3$ (1.3-point increase)
- Relationship between test scores and income is **concave**

Test scores and income

You could plot the predictions and the data to see what the relationship looks like

```
predict yhat

scatter yhat testscr avginc
```



Test scores and income

$$\widehat{TestScore}_i = 607.3 + 3.85 Income_i - 0.042 Income_i^2$$

- Finally, you could use calculus!
- First derivative is *slope* of regression line at any given value of income
- If first derivative is positive, then increasing *income* increases expected test scores
- If first derivative is negative, then increasing *income* decreases expected test scores

Test scores and income

$$\frac{d^2\,\widehat{TestScore_i}}{dIncome_i^2} = -0.084$$

- If the second derivative is positive, the function is **convex**
- If the second derivative is negative, the function is **concave**
- Relationship between test scores and income is negative and therefore concave for all values of income

How to calculate predicted changes

1. The predicted change in Y must be computed for specific values of X (that's the point!)
    - Predict Y at $X = x$
    - Predict Y at $X = x + \Delta x$
    - Take the difference
2. Rely on the derivative (*approximate* because the slope changes)

$$testscr = \beta_0 + \beta_1 avginc + \beta_2 avginc^2 + u$$

$$\frac{\partial testscr}{\partial avginc} = \beta_1 + 2\beta_2 avginc$$

$$\partial testscr = (\beta_1 + 2\beta_2 avginc)\partial avginc$$

Hypothesis test of linear effect

Test whether the relationship is non-linear: $H_0 : \beta_2 = 0$

```
. gen avginc2 = avginc^2

.  regress testscr avginc avginc2, robust

Linear regression                               Number of obs     =        420
                                                F(2, 417)         =     428.52
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5562
                                                Root MSE          =     12.724
```

testscr	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
avginc	3.850995	.2680941	14.36	0.000	3.32401	4.377979
avginc2	-.0423085	.0047803	-8.85	0.000	-.051705	-.0329119
_cons	607.3017	2.901754	209.29	0.000	601.5978	613.0056

Hypothesis test NO effect

Test whether the relationship is non-linear: $H_0 : \beta_1 = \beta_2 = 0$

```
. test avginc=avginc2 =0

 ( 1)  avginc - avginc2 = 0
 ( 2)  avginc = 0

       F(  2,   417) =  428.52
            Prob > F =    0.0000
```

More than two polynomial terms

Generalize to k polynomial terms (more flexible specification)

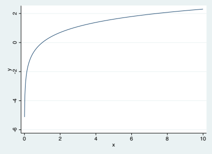
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + ... + \beta_k X_i^k + u_i$$

- Given enough terms, a polynomial can represent any relationship of Y and X as any continuous shape
- This is a simple example of an advanced topic: nonparametric estimation

Logarithmic functions

Logarithmic functions

$ln()$ is a special function: the inverse of the exponential function $x = ln(e^x)$



- Large slope for small x, approaches zero for large x
- Defined only for positive values of x
- Log of zero or a negative number is undefined

In this class, we are ALWAYS referring to NATURAL LOG

Functional forms: logarithmic

- Advantages
  - Convenient percentage/elasticity interpretation
  - Slope coefficients of logged variables are invariant to rescalings
  - Taking logs often eliminates/mitigates problems with outliers
  - Taking logs often helps to secure normality and homoskedasticity
- Caveats
  - Variables measured in units such as years should not be logged
  - Variables measured in percentage points should also not be logged
  - Logs must not be used if variables take on zero or negative values
  - It is hard to reverse the log-operation when constructing predictions

Small changes and logarithms

For *small* changes in x...

$$\boxed{100\Delta log(x) \approx \%\Delta x}$$

Based on insight that $ln(1 + r) \approx r$

Examples of differencing logarithms

Log approximation		Exact percent change	
ln(51)-ln(50)	0.019802	(51-50)/50	0.02
ln(50.5)-ln(50)	0.009950	(50.5-50)/50	0.01
ln(60)-ln(50)	0.182322	(60-50)/50	0.20
ln(80)-ln(50)	0.470004	(80-50)/50	0.6 0

Large changes and log dependent variables

- Are logs still useful with "large" changes? YES!
- "Large" is roughly when a unit change in X is associated with more than a 10% change in Y
- If so, calculate the exact percentage difference by exponentiating the coefficient:

$$\%\Delta \hat{Y} = 100[e^{\hat{\beta}_j} - 1]$$

Make sure you preserve the sign of the coefficient!

Using logs to compute percentage changes

Suppose we want to model hourly wages (wage) as a function of years of education (educ)

$$wage = 10.5 + 3educ$$

Level-level: A 1-year increase in years of education is associated with a \$3 increase in wages

$$log(wage) = 10.5 + 3log(educ)$$

Log-log (elasticity): A 1% increase in years of education is associated with a 3% increase in wages

Using logs to compute percentage changes

Suppose we want to model hourly wages (wage) as a function of years of education (educ)

$$log(wage) = 10.5 + 3educ$$

Log-level (semi-elasticity): A 1-year increase in years of education is associated with a 300% increase in wages *(approximation)*

$$wage = 10.5 + 3log(educ)$$

Level-log: A 1% increase in years of education is associated with a $3/100 = \$0.03$ increase in wages *(approximation)*

Where do these last two come from?

$$log(wage) = 10.5 + 3educ$$

Log-level (semi-elasticity): A 1-year increase in years of education is associated with a 300% increase in wages *(approximation)*

1. Take partial derivative of both sides: $\Delta log(wage) = 3\Delta educ$
2. Multiply by 100: $100\Delta log(wage) = 3 * 100\Delta educ$
3. Recall that $100\Delta log(x) \approx \%\Delta x$
4. $\%\Delta wage = 3 * 100(\Delta educ)$

Where do these last two come from?

$$wage = 10.5 + 3log(educ)$$

Level-log: A 1% increase in years of education is associated with a $3/100 = \$0.03$ increase in wages *approximation*

1. Take partial derivative of both sides: $\Delta wage = 3\Delta log(educ)$
2. Multiply/divide by 100: $\Delta wage = (3/100)\Delta log(educ)$
3. Recall that $100\Delta log(x) \approx \%\Delta x$
4. $\Delta wage \approx 0.03(\%\Delta educ)$

Summary

Type	Population model	Interpretation
Level-level	$y = \beta_0 + \beta_1 x_1 + u$	A 1-unit increase in x_1 is associated with a β_1-unit change in y.
Log-log	$ln(y) = \beta_0 + \beta_1 ln(x_1) + u$	A 1% increase in x_1 is associated with a β_1% unit change in y.
Log-level	$ln(y) = \beta_0 + \beta_1 x_1 + u$	A 1-unit increase in x_1 is associated with a $100\beta_1$% unit change in y.
Level-log	$y = \beta_0 + \beta_1 ln(x_1) + u$	A 1% increase in x_1 is associated with a $0.01\beta_1$-unit change in y.

When to use logarithms?

- For a variable Z, think about which are more meaningful?
  1. Absolute changes in $Z \Rightarrow$ use levels
  2. Percent changes in $Z \Rightarrow$ use logs

  Note that you do not need to transform all variables!

Interaction terms

Interaction terms?

```
. regress salary hispan black nl, robust

Linear regression                               Number of obs     =        353
                                                F(3, 349)         =       3.27
                                                Prob > F          =     0.0214
                                                R-squared         =     0.0276
                                                Root MSE          =     1.4e+06

------------------------------------------------------------------------------
             |               Robust
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hispan |  -212538.6   176821.5    -1.20   0.230    -560308.5    135231.2
       black |   399066.6   184907.6     2.16   0.032      35393.2      762740
          nl |  -160296.4     148307    -1.08   0.281    -451984.2    131391.5
       _cons |    1338400   118547.8    11.29   0.000      1105243     1571558
------------------------------------------------------------------------------
```

- Are Hispanic players paid more or less?
- Are players in the NL paid more or less?
- Is there a differential relationship between being Hispanic and pay in the NL vs AL?

Three types of interactions

1. Interaction between binary variables
   - Gives you finder control over measuring group estimates
2. Interactions between a binary and a continuous variable
   - *example: hits and NL*
3. Interactions between two continuous variables
   - *hits and RBIs*

Interactions with two binary variables

- Let's interact `NL` with both demographic characteristics
  - $genhNL = hispan * NL$: 1 for Hispanic players in the national league
  - $genbNL = black * NL$: 1 for Black players in the national league

Interactions with two binary variables

```
. gen hNL = hispan*nl

. gen bNL = black*nl

. reg salary hispan black nl hNL bNL, robust
```

Linear regression

Number of obs	=	353
F(5, 347)	=	4.13
Prob > F	=	0.0012
R-squared	=	0.0358
Root MSE	=	1.4e+06

salary	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
hispan	110714.8	270045.6	0.41	0.682	-420417.4	641847.1
black	462190.2	264680.3	1.75	0.082	-58389.42	982769.8
nl	10001.68	195271.8	0.05	0.959	-374063.5	394066.9
hNL	-695315.4	341685.5	-2.03	0.043	-1367351	-23280.25
bNL	-143244	370724.2	-0.39	0.699	-872393.4	585905.3
_cons	1261249	132198.2	9.54	0.000	1001238	1521259

Interactions with two binary variables

- Who is in the left-out group? White players in the American League, with average salary of $1,261,249 in 1993
- Effect on salary of being in the National League for White Players? ($nl = 1$) A $10,002 increase in salary
- What is the average salary for Hispanic players in the American League? ($hispan = 1$) 1,261,249 + 110,715= 1,371,964
- What is the average salary for Hispanic players in the National League? ($hispan = 1, nl = 1, hNL = 1$) 1,261,249+ 110,715 + 10,002 - 695,315 = $686,650

Interactions, one binary and one continuous

Effect of change in continuous $X_{1,i}$ and binary $D_{2,i}$ on Y_i:

$$Y_I = \beta_0 + \beta_1 X_{1,i} + \beta_2 D_{2,i} + \beta_3 X_{1,i} D_{2,i} + u_i$$

Effect of a 1-unit change in $X_{1,i}$ when $D_{2,i} = 0$?	β_1
Effect of a 1-unit change in $X_{1,i}$ when $D_{2,i} = 1$?	$\beta_1 + \beta_3$
Effect of change in $D_{2,i} = 0$ from 0 to 1?	$\beta_2 + \beta_3 X_i$

Interactions, one binary and one continuous

Does the relationship between salary and career hits differ if you are in the NL or AL?

```
. gen hitsNL = hits*nl

. reg salary hispan black nl hits hitsNL, robust

Linear regression                               Number of obs     =        353
                                                F(5, 347)         =      25.93
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4017
                                                Root MSE          =     1.1e+06
```

salary	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
hispan	-41753.53	134709.3	-0.31	0.757	-306703	223195.9
black	190197.2	148006.2	1.29	0.200	-100905	481299.3
nl	-116998.6	136614.4	-0.86	0.392	-385695.1	151697.9
hits	1411.073	154.9505	9.11	0.000	1106.313	1715.834
hitsNL	291.2646	310.123	0.94	0.348	-318.6929	901.222
_cons	453947.5	97562.33	4.65	0.000	262059.6	645835.5

Interactions, one binary and one continuous

$$\widehat{salary_i} = 453948 - 41754 hispan_i + 190197 black_i - 116999 nl_i + 1411 hits_i + 291 hitsNL_i$$

- In the AL, the effect of one more career hit: \$1,411 increase in salary
- In the NL, the effect of one more career hit: \$1,411+\$291 =\$1,702 increase in salary
- Effect of being in the NL on salary: -\$116,996+\$291**$Hits_i$*
- If $hits = 500$, effect of being in NL on salary:-\$116,996+\$291(500)=\$28,504

Interactions, two continuous variables

Effect of change in continuous $X_{1,i}$ and continuous $X_{2,i}$ on Y_i:

$$Y_I = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{1,i} X_{2,i} + u_i$$

Effect of a 1-unit change in $X_{1,i}$? $\beta_1 + \beta_3 X_{2,i}$

Effect of a 1-unit change in $X_{2,i}$? $\beta_2 + \beta_3 X_{1,i}$

Interactions, two continuous variables

```
. gen hitsXRBI = hits*rbis

. reg salary hispan black nl hits rbis hitsXRBI, robust

Linear regression                               Number of obs     =        353
                                                F(6, 346)         =      42.78
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5288
                                                Root MSE          =    9.7e+05
```

salary	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
hispan	26130.34	115023	0.23	0.820	-200102	252362.7
black	193761.6	132724.3	1.46	0.145	-67286.29	454809.5
nl	70225.2	107150.8	0.66	0.513	-140523.7	280974.1
hits	662.4478	343.1079	1.93	0.054	-12.39187	1337.287
rbis	5649.045	853.5365	6.62	0.000	3970.272	7327.818
hitsXRBI	-1.800353	.2123612	-8.48	0.000	-2.218034	-1.382671
_cons	-80493.96	87751.74	-0.92	0.360	-253087.9	92100.01

Interactions, two continuous variables

$$\widehat{salary}_i = -80494 + 26130 hispan_i + 193762 black_i + 70225 nl_i + 662 hits_i + +5659 rbis_i - 1.80 hitsXRBI_i$$

- Effect of 100 increase in career hits: $\$662 * 100 - \$1.80 * 100 * rbis$
- Effect of 100 increase in career RBIs: $\$45649 * 100 - \$1.80 * 100 * hits$

Remember economic significance for interpreting results

Choosing interaction terms

- Is there a compelling reason that the effect of changing one regressor might depend on another? If so, interact the two!
- Test whether the interaction term is statistically significant. If not, you still may want to include if the economic indicates it should be there
- Can use the adjusted R^2 ($\bar{R}^2$) - if increases when you add a variable, provides support for keeping it

Conclusion

Overview of nonlinear regression models

Polynomial regression

Logarithmic functions

Interaction terms

- Two binary variables
- One binary, one continuous variables
- Two continuous variables