**Lab Assignment Week 09**

**I. Objectives**: The primary objective is to compare the linearmodel to *more* nonlinear models, i.e., models that fit curves other than a straight line through the mean of the data. Our focus will be on the quadratic model, with some attention also paid to reciprocal models. Interpretation of regression coefficients is complicated by the fact that any deviation from straight lines means that a variable’s effects will be conditional.

**II. Dataset**: *oecd.dta*

**III. Packages**: *foreign*, *fBasics*, *lmtest*, ***white-test.R***

**IV. Preparation**

1) Open RStudio by double-clicking the icon or selecting RStudio from the Windows Start menu.

2) Open the POLS6481-Spring2021-UH-lab Project and perform a Git pull.

3) If not using Projects, Git, and *here* download files, place in working directory, and make changes as needed.

6) Open the R script by typing *Ctrl*+*O* or by clicking on File in the upper-left corner, using the dropdown menu, and navigating to the script in your working directory.

7) Run lines 1-5 in the R script to load three packages that you will need. You can install the *fBasics* and *lmtest* packages using the **Packages** tab or by uncommenting and running line 3 prior to running lines 4-5.

**V. Instructions for Lab Week 09**

You will use data on employment and economic growth in 25 OECD countries. The dependent variable in the analysis will be average annual percentage rate of growth in employment between 1988 and 1997 (*EMPLOY*) and the main independent variable will be average annual percentage rate of real GDP growth over the same time period (*GDP*). This is one of the major examples used in chapter 4 of Christopher Dougherty’s *Introduction to Econometrics*, 4th ed. (Oxford University Press, 2011).

Begin by loading data frame that you will work with, named oecd, by running line 7 or using the following code changing the directory as needed:

> oecd <- read.dta("~:/oecd.dta")

Run lines 9-10 to examine the normality of the dependent variable (*EMPLOY*) visually. Line 9 generates the normal quantile plot of *EMPLOY* against the normal distribution; any deviation from a straight line indicates a non-normal distribution. Line 10 generates a histogram of *EMPLOY*; any deviation from a bell curve would indicate a non-normal distribution.

> qqnorm(oecd$EMPLOY)

> hist(oecd$EMPLOY, freq = FALSE)

Run line 11 for the summary statistics of *EMPLOY*. A normally distributed variable has skewness equal to 0 and kurtosis equal to 3; the *kurtosis* command in R automatically subtracts 3, so look for deviations from 0. Based on your inspection, is *EMPLOY* approximately normally distributed?

> skewness(oecd$EMPLOY)

> kurtosis(oecd$EMPLOY)

For comparison, run line 13 for the normal quantile plot of our main independent variable (*GDP*), which is not normally distributed.

> qqnorm(oecd$GDP)

We saw in lab 7 that there might be justification for taking the natural log of an independent variable when its distribution is abnormal. Run line 14 for the normal quantile plot of *log*(*GDP*):

> qqnorm(log(oecd$GDP))

Line 17 plots *EMPLOY* against *GDP*, with a horizontal line for 0 growth in employment. You should observe a nonlinear pattern.

Line 18 and line 19 show two ways to plot *y* against *ln*(*x*). Line 18 implicitly creates a new variable, *log*(*GDP*), and uses these values for *x*. Line 19 does not re-code the variable, but instead tells R to use a logarithmic scale for the *x* axis. It turns out, this is as simple as including the code log="x" inside the parentheses!

Now our attention moves to three methods of modeling the relationship between growth and employment growth. Run lines 21–23 to create three new variables representing average GDP growth, average GDP growth squared, and the reciprocal of average GDP growth.

**A. Linear**

Run line 26 to estimate the model in which employment growth is linearly related to GDP growth.

> linmod<-lm(EMPLOY~growth1, oecd); summary(linmod)

On page 6, I provided some space for you to interpret the effect of average growth on employment growth. Use the results of this estimation to fill in the “Linear” section.

The next five lines of R code provide different ways of considering the linearity of the relationship.

1. Run line 27 to plot a scatterplot of the raw values of employment growth and the fitted values. In the R script file, I included extra codes to make the dots 25% smaller (cex = .75) and the fitted line red (col = “red”) and twice as thick (lwd = 2).

> plot(oecd$growth1, oecd$EMPLOY, pch=19, cex=.75); abline(linmod)

1. The residuals ought to be normally distributed. Run line 28, which generates a straight line plot only if residuals are normal.

> qqnorm(linmod$residuals)

1. Run line 29 to show the residuals-versus-predictor (growth1) plot. Do the residuals fit our expectations of zero expectation given growth and constant variance?

> plot(oecd$growth1, linmod$residuals, cex=.5); abline(h=0)

1. Run line 30 to load the white test function into R. Alternately, if you are not using *here* or you receive an error, you can open and run the ***white-test.R***script. Then run line 31 to perform White’s test of homoscedasticity. Do you reject the null hypothesis of constant error variance?
2. Run line 32 to perform “Ramsey’s RESET” (regression specification error test). If this test yields a statistically significant *p* value, then it suggests we should reject the null hypothesis that the model is correctly specified. Details on this test are in Wooldridge’s section 9.1 (in fifth edition) as well as Dougherty’s section 4.3.

All in all, the numerical results from the linear model suggest that heteroskedasticity is not a problem; this should be confirmed by the residuals-versus-predictor or a residuals-versus-fitted plot. However, this is why a visual approach to regression is necessary: the scatterplot reveals a probable nonlinearity problem. Growth levels appear to have a large effect at low levels and a small effect at high levels, i.e., diminishing marginal returns, but a linear model cannot accommodate this pattern.

**B. Quadratic**

Run line 35 to estimate the model in which employment growth is a quadratic function of growth. (You created the variables growth1 and growth2 in line 21 and line 22, respectively.)

> quadmod<-lm(EMPLOY~growth1+growth2, data=oecd); summary(quadmod)

The next five lines of R code provide some methods to assess whether we have adequately addressed the nonlinearity problem using the quadratic model, and to re-evaluate the homoscedasticity assumption.

1. First, run line 36 to graph a scatterplot of the raw values and the fitted values of employment growth against economic growth.
2. Run line 37 to examine the normal quantile plot; are the residuals approximately normal?
3. Run line 38 to examine the residuals-versus-predictor plot. Is heteroskedasticity still a problem?
4. Having already run the ***white-test.R*** script earlier, run line 39 to perform White’s test to test the null hypothesis that our residuals are free from heteroskedasticity.
5. Run line 40 to perform Ramsey’s RESET to see if the quadratic model is correctly specified.

Compare the significance of an unrestricted model (which contains both growth and growth-squared) and a restricted model (which contains just growth). You can use the anova command in line 42 to do so.

Finally, before we move into the issue of interpreting marginal effects in a nonlinear model, run line 43 to assess whether we introduced multicollinearity. Why is this important to do when using a quadratic model?

Let’s do some interpretation of the quadratic model. On page 6, I provided space for you to interpret the effect of average growth on average employment growth. Use the results of this second estimation to fill in the “Quadratic” section. Note that the code I am showing below is slightly simpler than what is in the R script; using the code shown below correctly requires you to remember that the coefficient for growth is *second* – the constant is the first – and the coefficient for squared-growth is *third*.

1. Run lines 46-49 to preserve some summary statistics of the key independent variable, average growth. I am saving the average value of growth (you might choose the median instead, after looking at line 46’s results), as well as the minimum and maximum values.

> summary(oecd$growth1)

> meangdp <- mean(oecd$growth1)

> mingdp <- min(oecd$growth1)

> maxgdp <- max(oecd$growth1)

1. Run line 50 to compute the marginal effect of growth on employment growth at the mean level of growth in the sample. Fill this in on page 6.

> quadmod$coef[2] + 2\*quadmod$coef[3]\*meangdp

1. To examine how the relationship between economic growth and employment growth changes as GDP growth’s value changes, run line 51 and line 52 to compute the marginal effect of economic growth on employment growth at the minimum and maximum values of GDP growth, respectively. Fill in the marginal effects on page 6.

> quadmod$coef[2] + 2\*quadmod$coef[3]\*mingdp

> quadmod$coef[2] + 2\*quadmod$coef[3]\*maxgdp

1. *What happens to the effect of growth on employment growth as growth increases?* Run lines 53-55 to plot the two fitted lines against each other. Remember, most countries in our sample are clustered on the left part of the scale. The quadratic fitted line’s slope is steeper than the linear fitted line’s slope in the range of most countries. However, the quadratic curve bends so that its slope is shallower, and then its slope turns negative, so that higher levels of growth would correspond to lower employment growth!

Take the time to examine closely the equations for creating these plots. Students who have learned calculus will have an easier time understanding that the marginal effect is the first derivative of the function *y* = *f*(*x*), where *f*(*x*) = *b*0 + *b*1*x + b*2*x*2; using the **power rule** and **sum rule** yields *f’*(*x*) = *b*1 + 2*b*2*x*.

So, we must add the coefficient for growth and the coefficient for growth-squared times 2 times *x*!

The next five lines of code suggest one way to plot marginal effects and confidence intervals around those marginal effects. Recall that in a linear model, the marginal effect equals the regression coefficient, and there is a single confidence interval around that effect. For a nonlinear model, however, the marginal effect is the first derivative, which will depend on the value of *x*. As we will see in lecture 18, the width of the confidence interval also varies depending on the value of *x*.

1. Run line 57 to create the equation for the slope, using the equation for *f’*(*x*)shown above.

> slope = quadmod$coefficients[2] + (2\*quadmod$coefficients[3]\*oecd$growth1)

1. Run line 58 to create the variance-covariance matrix of the estimators, which we then use to generate the standard errors of the slope. Notice that the standard error depends on the variance of the coefficient for growth, the variance of the coefficient for squared growth, and the covariance of these coefficients. Notice also that *growth1* (growth) and *growth2* (squared growth) enter into the equation.

> vce = vcov(quadmod)

> sequad = sqrt(vce[2,2] + 4\*oecd$growth2\*vce[3,3] + 4\*oecd$growth1\*vce[2,3])

1. Run line 59 to create the upper and lower bounds for the confidence interval for the slope; you are adding and subtracting 2.07 times the standard error (computed in line 57) to/from the slope.

> cimax = slope + 2.07\*sequad

> cimin = slope – 2.07\*sequad

The *t* critical value for 23 degrees of freedom is roughly 2.07.

1. Run line 60 to plot the marginal effect and its confidence interval, with a horizontal line for zero. The curve shown by running line 60 is not the predicted value of employment growth given average growth; rather the curve shows the slope of the function relating employment growth to average growth.

When the marginal effect is above zero, it indicates that in this range, growth has a positive effect on text scores; when the curve is below zero, it indicates that in this range, growth has a negative effect on employment growth. Thus, a downward sloping marginal effect plot does not indicate that growth reduces employment growth; instead it indicates that the positive impact declines and ultimately becomes indistinguishable from zero.

1. Run line 61 to plot the values of growth (as a box-and-whisker plot). By combining these plots in a single figure, you can see that most countries have values for growth that fall in a range where growth has a positive effect on employment growth. Only a few outliers are in the range of values of GDP growth where it has an effect on employment growth that is indistinguishable from zero.

[Note: this only looks right if the **Plots** window is close to a square; if it is a rectangle then the box-and-whisker plot might overlap the marginal effect plot.]

Run line 62 to restore the defaults for plotting. There are lots of resources for learning how to combine graphs; I used QuickR: *http://www.statmethods.net/advgraphs/layout.html*.

**C. Reciprocal**

It is hard to imagine a situation in which higher average growth should lead to lower average employment growth. On this basis, you might prefer a model that asymptotes to a flat line, such as a reciprocal model.

Run line 65 to estimate the model in which employment growth are a function of 1/growth (defined way back in line 24). Notice the constant: this is the value to which the function will converge as economic growth approaches infinity (i.e., as 1/growth goes to zero). The next five lines of code allow you to consider the model’s adequacy.

1. Run lines 66-67 to graph a scatterplot of the values of employment growth and their fitted values (i.e., predicted employment growth) against average growth.
2. Run line 68 to examine the normal quantile plot; run line 69 to generate the residuals-versus-fitted-values plot. Have we resolved the heteroskedasticity problem to your satisfaction?
3. Run line 70 to perform White’s test; do you retain or reject the null hypothesis that the residuals are homoscedastic?

Turn your attention now to interpretation; use the following to fill in the “Reciprocal” section on page 6.

1. Run line 73 to compute the marginal effect of growth on employment growth at the mean level of growth.
2. To examine how the relationship between growth and employment growth changes as growth’s value changes, run line 74 and line 75 to compute the marginal effect of growth on employment growth at the minimum and maximum values of economic growth, respectively.
3. Run line 77 to generate the marginal effects, and then run line 78 to graph this curve against a histogram showing the values of average district growth.

The curve shown by running line 77 is not the predicted value of employment growth given average growth; that curve would be downward sloping, due to diminishing returns, but never reaching zero. Rather, the curve shown by running line 77 is the slope of the function relating employment growth to average growth, across all levels of average growth.

Be sure to take some time to inspect the equations for creating the marginal effects. Students who have learned calculus will have an easier time understanding that what we care about is the first derivative of the function *y* = *f*(*x*), where *f*(*x*) = *b*0 + *b*1*x*−1. Using the ‘power rule’ one finds that *f*’(*x*) = –*b*1*x*−2.

Run lines 80-83 to plot employment growth against growth, including predicted employment growth using all three models:

* the linear model’s fitted values are displayed with a **purple** line;
* the quadratic model’s fitted values are displayed with a **neon green** line;
* the reciprocal model’s fitted values are displayed with a **royal blue** line.

You have estimated three models this week: linear, quadratic, and reciprocal. Which seems to fit the data best? Which is the easiest to interpret? Based on these judgments, which model do you prefer for this application?

**Linear Model (R2 = \_\_\_\_\_\_)**

Marginal effect of growth at **all values** of growth: a 1-unit increase in economic growth results in a \_\_\_\_\_ unit change in employment growth

**Quadratic Model (R2 = \_\_\_\_\_\_)**

Marginal effect of growth at **minimum** growth: a 1-unit increase in economic growth results in a \_\_\_\_\_ unit change in employment growth

Marginal effect of growth at **mean** growth: a 1-unit increase in economic growth results in a \_\_\_\_\_ unit change in employment growth

Marginal effect of growth at **maximum** growth: a 1-unit increase in economic growth results in a \_\_\_\_\_ unit change in employment growth

**Reciprocal Model (R2 = \_\_\_\_\_\_)**

Marginal effect of growth at **minimum** growth: a 1-unit increase in economic growth results in a \_\_\_\_\_ unit change in employment growth

Marginal effect of growth at **mean** growth: a 1-unit increase in economic growth results in a \_\_\_\_\_ unit change in employment growth

Marginal effect of growth at **maximum** growth: a 1-unit increase in economic growth results in a \_\_\_\_\_ unit change in employment growth

**For more information on polynomial regression, see**

**https://bookdown.org/tpinto\_home/Beyond-Linearity/polynomial-regression.html**

I have uploaded a polynomial functions graph from this site in the “Additional Resources Folder” on GitHub”