

Week 9 – Problem Set

1.

Consider the population model $y_i = \mathbf{x}_i\beta + u_i$ in which $y_i > 0$. In order to restrict the domain of y_i to be positive, some researchers have proposed transforming y_i and estimating the model

$$\log(1 + y_i) = \mathbf{x}_i\gamma + r_i$$

- (a) Using this transformation, what is $E[y_i|\mathbf{x}_i]$?
- (b) Using this transformation, what is the partial effect of a change in x_{ji} on $E[y_i|\mathbf{x}_i]$?
- (c) Given your answers to (a) and (b), is this transformation a good idea? Why or why not?
- (d) Explain how to estimate the partial effect in (b) assuming r is independent of \mathbf{x} .
- (e) Explain how to estimate the partial effect in (b) assuming a distribution $f(r_i|\mathbf{x}_i)$.

2.

Consider the population model

$$E[y_i|\mathbf{x}_i] = \exp\{\beta_1 + \beta_2\log(x_{1i}) + \beta_3x_{2i}\}$$

For this problem you may want to refer to the formulas for partial effects, partial elasticities, and exact percentage change used in Problem Set 1.

- (a) What is the elasticity of $E[y_i|\mathbf{x}_i]$ with respect to a change in x_1 ?
- (b) What is the *exact* percentage change in $E[y_i|\mathbf{x}_i]$ for $\Delta x_2 = 1$?
- (c) Assuming $E[y_i|\mathbf{x}_i] > 0$, provide a formula for the *approximate* percentage change in $E[y_i|\mathbf{x}_i]$ in response to a change in x_2 .
- (d) Let $E[y_i|\mathbf{x}_i] = \exp\{\beta_1 + \beta_2\log(x_{1i}) + \beta_3x_{2i} + \beta_4x_{2i}^2\}$ where $\beta_3 > 0$ and $\beta_4 < 0$. At which value of x_2 does the partial effect of a change in x_2 on $E[y_i|\mathbf{x}_i]$ become negative?

3.

In Lecture 9 we developed some general results for the Nonlinear Least Squares estimator with a generic conditional mean function $m(\mathbf{x}_i, \beta)$ by using the appropriate objective function $q(\mathbf{w}_i, \theta)$, Score vector $\mathbf{s}(\mathbf{w}_i, \theta)$, and Hessian matrix $\mathbf{H}(\mathbf{w}_i, \theta)$. This problem asks you to apply those results to one of the two most common choices for $m(\mathbf{x}_i, \beta)$:

the exponential regression model

$$y_i = \exp\{\mathbf{x}_i\beta\} + u_i$$

When working with the exponential model it is useful to note that

$$\mathcal{M}_i = \nabla_{\beta} \exp\{\mathbf{x}_i\beta\} = \exp\{\mathbf{x}_i\beta\} \cdot \mathbf{x}_i$$

(a) Assume throughout this problem that $E[u_i|\mathbf{x}_i] = 0$. What is the significance of this condition?

(b) What is the objective function $q(\mathbf{w}_i, \boldsymbol{\theta})$ to be minimized? Is it continuous in $\boldsymbol{\theta}$? Why or why not? Why does continuity matter?

(c) Derive the Score vector $\mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta})$ using the exponential regression

(d) Does $E[\mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_0)] = \mathbf{0}$? (Hint: use the Law of Iterated Expectations.) What is the significance of this condition?

(e) Derive the Hessian matrix $\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta})$ using exponential $m(\mathbf{x}_i, \boldsymbol{\beta})$. What does your result imply about the continuous differentiability in $\boldsymbol{\theta}$ of the Score vector? What is the significance of this condition?

[NOTE: skip part (f)]

(g) Derive the semi-robust estimator $\widehat{Avar}[\hat{\boldsymbol{\theta}}_{NLS}]$ for the exponential regression

(h) Express the Generalized Information Matrix Equality (GIME) for the exponential regression

(i) Assuming the GIME holds, derive the non-robust estimator $\widehat{Avar}[\hat{\boldsymbol{\theta}}_{NLS}]$ for the exponential regression

4.

This problem uses data from 401kpart.txt which contains observations about employee's participation in their employer's 401(k) retirement plan. The variables include:

<u>Variable Name</u>	<u>Variable Label</u>
partic	number of employees participating in the 401(k) plan
totemp	number of firm employees worldwide
employ	number of employees eligible for the 401(k) plan
mrte	plan match rate (per dollar)
age	age of the plan
sole	=1 if the 401(k) is the only pension plan

Define $prate_i = \frac{partic_i}{employ_i}$ and define \mathbf{x} to include a constant, $mrte$, $sole$, and age . Note that $mrte$ is continuous, $sole$ is discrete binary, and age is discrete non-binary.

(a) Is the choice of a logistic function for $m(\mathbf{x}_i, \boldsymbol{\beta})$ sensible for a model that uses $prate$ as the dependent variable? Why or why not?

(b) Using the %LeastSquares macro, estimate a logistic regression by NLS using $prate$ as the dependent variable. Do the average partial effects have their expected signs?

(c) Using the %LeastSquares macro, estimate twice the logistic regression in (b) by WNLS, first using $\hat{u}_i^2 = (y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{NLS}))^2$ as an estimate of $var[y_i|\mathbf{x}_i]$, and then using $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{NLS}) \cdot (1 - m(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{NLS}))$ as an estimate of $var[y_i|\mathbf{x}_i]$. Compare and contrast the estimated average partial effects. What do you conclude about the impact of the choice of weight on the estimates? Which is a more sensible choice for the weight?

5.

In Problem 5 of Problem Set 2 we used the `nls80.txt` data set to estimate a linear equation that explains $\log(wage)$ as a function of explanatory variables. In this problem we consider nonlinear estimation of an exponential *wage* regression using *exper*, *tenure*, *married*, *south*, *urban*, *black*, and *educ* as additional explanatory variables.

(a) Using the `%LeastSquares` macro (include the option to print the variance-covariance matrix estimates), estimate an exponential regression by NLS using *wage* as the dependent variable. Are the estimates of the expected signs?

(b) Using the estimation in (a), perform (by hand) a robust Wald test of the joint significance of *tenure* and *south*. (You will want to form the vectors and matrices for the Wald test statistic using the appropriate values in the variance-covariance matrix estimates.)

Next, confirm and compare your (by hand) estimate with SAS code using the “test” command to test the joint hypothesis above.

[NOTE: skip part (c)]

(d) You have just submitted a paper to be published in a labor economics journal in which you report the estimation in (a). You receive three referee reports, each telling you to use weights. One referee suggests using \hat{u}_i^2 to form your weights; a second referee suggests using $m(\mathbf{x}_i, \hat{\beta}_{NLS})$, and a third referee suggests using $m(\mathbf{x}_i, \hat{\beta}_{NLS})^2$. Which referee is correct? Does the choice affect the conclusion from any of the test results reported in (b)

Using the `%LeastSquares` macro, reset the `weight_type` command to explore these three options, share parameter and variance-covariance estimates, and compute the Wald statistic in each scenario, reflecting on any importance differences.

(e) The regressions above exclude *ability*, which is very difficult to measure. What would you need to assume about *ability* in order to be assured that the APEs resulting from the estimation in (a) are consistently estimated?

[NOTE: next part is extra credit for this problem set]

(f) Suppose that you are interested in the APEs for the subpopulation that lives in cities (*urban* = 1). Which is the appropriate way to proceed: by estimating the model using the observations for which *urban* = 1 and then calculating the APEs, or by estimating the model using all observations and then calculating the APEs by setting *urban* = 1 for each observation? Explain your choice, and calculate both sets of APEs.