Limited Dependent Variables & Selection: PS #1

Francis DiTraglia

HT 2021

This problem set is due on *Friday in Week 2 of HT 2021*. You need only submit solutions to questions 1–3, as question #4 will not be marked. See the explanation immediately preceding question #4 for further information.

1. Let $y \sim$ Poisson(θ).

   (a) Using steps similar to the derivation of $\mathbb{E}[y]$ from the lecture slides, show that $\mathbb{E}[y(y-1)] = \theta^2$.

   Solution:

$$\begin{aligned}\mathbb{E}\left[y(y-1)\right] &= \sum_{y=0}^{\infty} y(y-1)\left(\frac{e^{-\theta}\theta^y}{y!}\right) = \sum_{y=2}^{\infty} y(y-1)\left(\frac{e^{-\theta}\theta^y}{y!}\right)\\ &= \theta^2 \sum_{y=2}^{\infty} \frac{e^{-\theta}\theta^{y-2}}{(y-2)!} = \theta^2 \sum_{y=0}^{\infty} \frac{e^{-\theta}\theta^y}{y!} = \theta^2\end{aligned}$$

   The first equality is the definition of $\mathbb{E}[y(y-1)]$ for a Poisson RV. The second uses the fact that $y(y-1) = 0$ for $y = 0$ and $y = 1$ so the first two terms of the infinite sum are zero. The third factors θ^2 out of the infinite sum (we can always do this provided that the sum converges) and cancels $y(y-1)$ from $y!$ in the denominator. The fourth shifts the index of summation, and the final recognizes that the infinite sum is now a Poisson pmf summed over all possible values of y and hence equals one.

   (b) Use your answer to the preceding part, along with the result $\mathbb{E}[y] = \theta$, to show that $\text{Var}(y) = \theta$.

   Solution: Recall that $\text{Var}(y) = \mathbb{E}(y^2) - \mathbb{E}(y)^2$. Hence,

$$\begin{aligned}\mathbb{E}\left[y(y-1)\right] &= \mathbb{E}(y^2) - \mathbb{E}(y)\\ &= \mathbb{E}(y^2) - \mathbb{E}(y)^2 + \left[\mathbb{E}(y)^2 - \mathbb{E}(y)\right]\\ &= \text{Var}(y) + \left[\mathbb{E}(y)^2 - \mathbb{E}(y)\right]\end{aligned}$$

   and solving for $\text{Var}(y)$,

$$\text{Var}(y) = \mathbb{E}\left[y(y-1)\right] + \mathbb{E}(y) - \mathbb{E}(y)^2.$$

From the preceding part we know that $\mathbb{E}\left[y(y-1)\right] = \theta$ and from the lecture slides we know that $\mathbb{E}(y) = \theta$. Therefore, $\text{Var}(y) = \theta^2 + \theta - \theta^2 = \theta^2$.

2. Suppose that we observe count data $y_1, \ldots, y_N \sim$ iid p_o and our model $f(y_i|\theta)$ is a Poisson(θ) probability mass function. Show that $\widehat{K} = s_y^2/(\bar{y})^2$ where we define $s_y^2 = \frac{1}{N}\sum_{i=1}^N (y_i - \bar{y})^2$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^N y_i$.

Solution: Because θ is a scalar, by definition

$$\widehat{K} \equiv \frac{1}{N}\sum_{i=1}^N \left[\frac{d}{d\theta}\log f(y_i|\widehat{\theta})\right]^2$$

Here $\log f(y_i|\theta) = y_i \log(\theta) - \theta - \log(y_i!)$ and, as derived in the lecture slides, $\widehat{\theta} = \bar{y}$. Differentiating with respect to θ and substituting into the expression for $\widehat{K}$ given above, we have

$$\begin{aligned}
\widehat{K} &= \frac{1}{N}\sum_{i=1}^N \left[y_i/\bar{y} - 1\right]^2 = \frac{1}{N}\sum_{i=1}^N \left[y_i^2/(\bar{y})^2 - 2y_i/\bar{y} + 1\right] \\
&= \frac{1}{(\bar{y})^2}\left[\frac{1}{N}\sum_{i=1}^N y_i^2\right] - \frac{2}{\bar{y}}\left[\frac{1}{N}\sum_{i=1}^N y_i\right] + \left[\frac{1}{N}\sum_{i=1}^N 1\right] \\
&= \frac{1}{(\bar{y})^2}\left[\frac{1}{N}\sum_{i=1}^N y_i^2\right] - \frac{2}{\bar{y}}\cdot\bar{y} + 1 = \frac{1}{(\bar{y})^2}\left[\frac{1}{N}\sum_{i=1}^N y_i^2\right] - 1 \\
&= \frac{1}{(\bar{y})^2}\left\{\left[\frac{1}{N}\sum_{i=1}^N y_i^2\right] - (\bar{y})^2\right\}.
\end{aligned}$$

It remains to show that the term in the curly braces equals s_y^2. Expanding,

$$\begin{aligned}
s_y^2 &\equiv \frac{1}{N}\sum_{i=1}^N (y_i - \bar{y})^2 = \frac{1}{N}\sum_{i=1}^N \left(y_i^2 - 2y_i\bar{y} + \bar{y}^2\right) \\
&= \left[\frac{1}{N}\sum_{i=1}^N y_i^2\right] - 2\bar{y}\left[\frac{1}{N}\sum_{i=1}^N y_i\right] + \bar{y}^2\left[\frac{1}{N}\sum_{i=1}^N 1\right] \\
&= \left[\frac{1}{N}\sum_{i=1}^N y_i^2\right] - 2(\bar{y})^2 + (\bar{y})^2 \\
&= \left[\frac{1}{N}\sum_{i=1}^N y_i^2\right] - (\bar{y})^2.
\end{aligned}$$

3. Let $\widehat{\boldsymbol{\beta}}$ be the conditional maximum likelihood estimator of $\boldsymbol{\beta}_o$ in a Poisson regression model with conditional mean function $\mathbb{E}(y_i|\mathbf{x}_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta}_o)$, based on a sample of

iid observations $(y_1, \mathbf{x}_1), \ldots, (y_N, \mathbf{x}_N)$.

(a) Derive the first-order conditions for $\widehat{\boldsymbol{\beta}}$.

Solution: The log-likelihood of the i^{th} observation is given by

$$\begin{aligned}\ell_i(\boldsymbol{\beta}) &\equiv \log f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = y_i \log\left[\exp\left\{\mathbf{x}_i'\boldsymbol{\beta}\right\}\right] - \exp(\mathbf{x}_i\boldsymbol{\beta}) - \log\left(y_i!\right) \\ &= y_i\mathbf{x}_i'\boldsymbol{\beta} - \exp(\mathbf{x}_i'\boldsymbol{\beta}) - \log\left(y_i!\right)\end{aligned}$$

and hence the score vector is

$$\mathbf{s}_i(\boldsymbol{\beta}) \equiv \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = y_i\mathbf{x}_i - \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\mathbf{x}_i = \mathbf{x}_i\left[y_i - \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right].$$

Therefore, $\widehat{\boldsymbol{\beta}}$ solves the first order condition

$$\frac{1}{N}\sum_{i=1}^N \mathbf{x}_i\left[y_i - \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right].$$

In other words,

$$\frac{1}{N}\sum_{i=1}^N \mathbf{x}_i\left[y_i - \exp\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)\right] = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i\widehat{u}_i = \mathbf{0}.$$

Notice that we are free to include or exclude the $1/N$ factor since multiplying both sides by N gives

$$\sum_{i=1}^N \mathbf{x}_i\left[y_i - \exp\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)\right] = \sum_{i=1}^N \mathbf{x}_i\widehat{u}_i = \mathbf{0}.$$

(b) Using your answer to the previous part show that, so long as $\mathbf{x}_i$ includes a constant, the residuals $\widehat{u}_i \equiv y_i - \exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})$ sum to zero, as in OLS regression.

Solution: The first order conditions derived in the preceding part are a *collection* of equations: one for each regressor x_j. If $\mathbf{x}$ contains a constant, then one of the x_j is simply equal to one. Substituting, the first-order condition for this regressor is

$$\frac{1}{N}\sum_{i=1}^N 1\cdot\left[y_i - \exp\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)\right] = \frac{1}{N}\sum_{i=1}^N \widehat{u}_i = 0.$$

Multiplying through by N gives $\sum_{i=1}^N \widehat{u}_i = 0$.

(c) Using your answer to the preceding part, show that $\left[\frac{1}{N}\sum_{i=1}^N \exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})\right] = \bar{y}$, where $\bar{y}$ is the sample mean of y, so that $\bar{y}\widehat{\beta}_j$ equals the estimated average partial effect of x_j in this model.

Solution: Since $\widehat{u}_i \equiv y_i - \exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})$, we have $\exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}) = y_i - \widehat{u}_i$. Hence,

$$\frac{1}{N}\sum_{i=1}^{N} \exp\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{u}_i) = \frac{1}{N}\sum_{i=1}^{N} y_i - \frac{1}{N}\sum_{i=1}^{N}\widehat{u}_i = \bar{y} - 0 = \bar{y}.$$

(d) Explain why multiplying the estimated coefficients from this model by $\bar{y}$ makes them roughly comparable to the corresponding OLS estimates from the model $y_i = \mathbf{x}_i'\boldsymbol{\theta} + \varepsilon_i$.

Solution: The result of the preceding part implies that the estimated average partial effect of x_j in a Poisson regression model equals $\bar{y}\widehat{\beta}_j$. In a linear regression model, the partial effects do not vary with $\mathbf{x}$. Hence the estimated average partial effect of x_j is simply $\widehat{\theta}_j$. In other words: the estimated *coefficients* in a linear regression are APEs, while the estimated coefficients in a Poisson regression must be rescaled by $\bar{y}$ to convert them to APEs. After carrying out this conversion we are comparing apples-to-apples, albeit from different models. Accordingly we should expect $\widehat{\theta}_j$ and $\bar{y}\widehat{\beta}_j$ to be more comparable in magnitude that $\widehat{\theta}_j$ and $\widehat{\beta}_j$.

The following applied question will *not be marked*, but you encouraged to complete it nonetheless as it will build your understanding of the material from the lectures. Solving this problem will requires some of the R material from Lecture #6.

4. *This question is adapted from Wooldridge (2010).* To answer it you will need to use the dataset `SMOKE.RAW`, which can either be downloaded from the MIT Press website for the text, or loaded directly into R using the package `Wooldridge`. Documentation for the dataset is available in the R package or alternatively at `http://fmwww.bc.edu/ec-p/data/wooldridge/smoke.des`

Solution: See attached pdf document.

(a) Use a linear regression to predict *cigs*, the number of cigarettes smoked each day, using the regressors log(*cigpric*), log(*income*), *restaurn*, *white*, *educ*, *age*, and age^2. Interpret your findings. In particular: are cigarette prices and income statistically significant predictors? Does this depend on whether you use robust standard errors?

(b) Repeat the preceding part but estimate a *Poisson* regression with an exponential conditional mean function rather than a linear regression. Calculate the APEs for the Poisson model and compare them to the OLS estimates.

(c) If you calculated standard errors using the Poisson variance assumption, are cigarette prices and income statistically significant? Compare to your OLS results from above.

(d) Calculate $\widehat{\sigma}^2$. Does your estimate suggest evidence of overdispersion? If you use the Quasi-Poisson Variance assumption, how do your results compare to those of the preceding part?

(e) How do your answers to the preceding two parts change if you instead use the fully-robust "sandwich" standard errors?

Problem Set #1 - Question 4 Solution

Francis J. DiTraglia

Count Data: Smoking Example

Mullahy (1997), Review of Economics and Statistics, 79, 596-593

```r
library(wooldridge)
library(lmtest)
library(sandwich)

names(smoke)
```

```
##  [1] "educ"     "cigpric"  "white"    "age"      "income"   "cigs"
##  [7] "restaurn" "lincome"  "agesq"    "lcigpric"
```

Variable Descriptions: `smoke`

```r
# Specify x'beta
smoking_model <- cigs ~ lcigpric + lincome + restaurn + white + educ + age + agesq
```

- `cigs` number of cigarettes smoked per day
- `lcigpric` log of state cigarette price (cents/pack)
- `lincome` log of annual income (US Dollars)
- `restaurn` equals 1 if restaurant has smoking restrictions
- `white` equals 1 if white
- `educ` years of schooling
- `age` age in years
- `agesq` age squared

OLS with plain-vanilla SEs

```r
ols <- lm(smoking_model, data = smoke)
coeftest(ols)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -2.6824347 24.2207299 -0.1107   0.91184
## lcigpric    -0.8509044  5.7823214 -0.1472   0.88305
## lincome      0.8690144  0.7287636  1.1925   0.23344
## restaurn    -2.8656213  1.1174059 -2.5645   0.01051 *
## white       -0.5592363  1.4594610 -0.3832   0.70169
## educ        -0.5017533  0.1671677 -3.0015   0.00277 **
## age          0.7745021  0.1605158  4.8251 1.676e-06 ***
## agesq       -0.0090686  0.0017481 -5.1878 2.699e-07 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

OLS with Robust SEs

```r
coeftest(ols, vcov. = vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -2.6824347 26.1562565 -0.1026  0.918343
## lcigpric    -0.8509044  6.1132231 -0.1392  0.889334
## lincome      0.8690144  0.6035374  1.4399  0.150296
## restaurn    -2.8656213  1.0215479 -2.8052  0.005151 **
## white       -0.5592363  1.3943263 -0.4011  0.688468
## educ        -0.5017533  0.1632410 -3.0737  0.002186 **
## age          0.7745021  0.1393883  5.5564 3.752e-08 ***
## agesq       -0.0090686  0.0014754 -6.1464 1.250e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Poisson Regression

```r
pois_reg <- glm(smoking_model, family = poisson(link = 'log'), data = smoke)
coeftest(pois_reg)
```

```
##
## z test of coefficients:
##
##                Estimate  Std. Error  z value  Pr(>|z|)
## (Intercept)  0.39644936  0.61394730   0.6457    0.5184
## lcigpric    -0.10596071  0.14339006  -0.7390    0.4599
## lincome      0.10372755  0.02028060   5.1146 3.144e-07 ***
## restaurn    -0.36360594  0.03122216 -11.6458 < 2.2e-16 ***
## white       -0.05520115  0.03741971  -1.4752    0.1402
## educ        -0.05942253  0.00425626 -13.9612 < 2.2e-16 ***
## age          0.11425708  0.00496904  22.9938 < 2.2e-16 ***
## agesq       -0.00137082  0.00005695 -24.0704 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Quasi-Poisson Regression

```r
quasipois <- glm(smoking_model, family = quasipoisson(link = 'log'), data = smoke)
coeftest(quasipois)
```

```
##
## z test of coefficients:
##
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)  0.39644936  2.76738512  0.1433  0.886087
## lcigpric    -0.10596071  0.64633482 -0.1639  0.869778
## lincome      0.10372755  0.09141539  1.1347  0.256508
## restaurn    -0.36360594  0.14073479 -2.5836  0.009777 **
```

```
## white       -0.05520115  0.16867044 -0.3273  0.743462
## educ        -0.05942253  0.01918520 -3.0973  0.001953 **
## age          0.11425708  0.02239807  5.1012 3.375e-07 ***
## agesq       -0.00137082  0.00025671 -5.3400 9.292e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Overdispersion or Underdispersion?

```r
# Extract the estimate of sigma-squared
summary(quasipois)$dispersion
```

```
## [1] 20.31782
```

```r
# Now do it "by hand"
yhat <- predict(pois_reg, type = 'response')
uhat <- residuals(pois_reg, type = 'response')
mean(uhat^2 / yhat)
```

```
## [1] 20.11488
```

Robust "Sandwich" SEs for Poisson Regression

```r
coeftest(pois_reg, vcov. = vcovHC)
```

```
##
## z test of coefficients:
##
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  0.3964494  3.0227099  0.1312   0.89565
## lcigpric    -0.1059607  0.6787042 -0.1561   0.87594
## lincome      0.1037276  0.0844975  1.2276   0.21960
## restaurn    -0.3636059  0.1417068 -2.5659   0.01029 *
## white       -0.0552011  0.1662139 -0.3321   0.73981
## educ        -0.0594225  0.0194336 -3.0577   0.00223 **
## age          0.1142571  0.0214898  5.3168 1.056e-07 ***
## agesq       -0.0013708  0.0002476 -5.5365 3.086e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing OLS and Poisson Estimates

```r
ybar <- mean(smoke$cigs)
OLS_est <- coefficients(ols)[-1]
pois_est <- coefficients(pois_reg)[-1]
cbind(OLS = OLS_est, Poisson_APE = ybar * pois_est, Poisson = pois_est)
```

```
##                   OLS Poisson_APE      Poisson
## lcigpric -0.850904380 -0.92042698 -0.105960710
## lincome   0.869014392  0.90102862  0.103727546
## restaurn -2.865621339 -3.15846053 -0.363605941
## white    -0.559236320 -0.47950441 -0.055201150
## educ     -0.501753267 -0.51617344 -0.059422535
## age       0.774502141  0.99249336  0.114257081
## agesq    -0.009068603 -0.01190761 -0.001370819
```

Note the `age` and `agesq` entries are not partial effects. Why not?