Week 9: Modeling II

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2019

Updated notes are here: https://clas.ucdenver.edu/marcelo-perraillon/teaching/health-services-research-methods-i-hsmp-7607

Outline

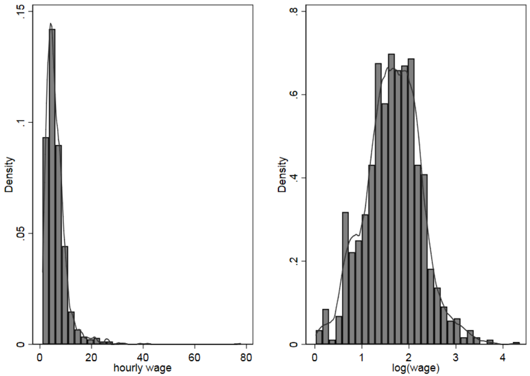
- More on transformations
- Taking the log
- The retransformation problem
- Other transformations

Logarithms

- A very common transformation is to take the log of either the outcome or some of the predictors
- We saw in the homework that taking the $log(wage)$ significantly improved the model fit in the beauty example
- Taking the log of the outcome is often done to make model assumptions fit better; taking the log of predictors is often done for model interpretation
- In particular, the log of skewed data looks more normally distributed
- **A note on notation**: We use *log* and *ln* interchangeable. That's the logarithm with base e
- If in different base, usually noted. Like: $log_{10}(x)$

Compare wages vs log(wages)

```
hist wage, kdensity saving(w1.gph, replace)
hist lwage, kdensity saving(lw1.gph, replace)
graph combine w1.gph lw1.gph
graph export wvslw.png, replace
```



The transformation helps with model fit

- In the beauty dataset we saw that taking the *log*(*wage*) improved the fit by a lot. For simplicity consider just one predictor, experience

```
reg wage exper
est sto m1
reg lwage exper
est sto m2
 ----------------------------------------------
    Variable |      m1              m2
-------------+--------------------------------
       exper |  .09140614***    .01523377***
       _cons |  4.6425183***    1.3814481***
-------------+--------------------------------
          r2 |  .05505228       .09397574
          ll | -3691.0204      -1069.9606
----------------------------------------------
      legend: * p<0.05; ** p<0.01; *** p<0.001
```

- In case you missed it. We just changed the scale of y and now the model fits much better. We have not done anything else!

log(y) interpretation

- One problem though is that now the coefficients are changes in the log(wage) scale but we care about wages, not log(wages)
- The model is $log(wage) = \beta_0 + \beta_1 exper + \epsilon$, where experience is measured in years
- We can of course interpret β_1 as the change in average $log(wage)$ for an extra year of experience
- A **shortcut** for interpretation is that the **percent** change in wage is $100 * \beta_1 \Delta exper$. For a one year change in education: $100 * \beta_1$
- This works because $log(x_1) - log(x_0)$ approximates $\frac{(x_1 - x_0)}{x_0}$ for **small changes** in x (the proof requires using the first order Taylor expansion)
- With other covariates you you'd just need to add the "holding other factors constant" or "taking the other variables into account"

log(y) interpretation

- The model again

```
reg lwage exper

      Source |       SS           df       MS      Number of obs   =     1,260
-------------+----------------------------------   F(1, 1258)      =    130.48
       Model |  41.8173212         1  41.8173212   Prob > F        =    0.0000
    Residual |  403.162651     1,258  .320479055   R-squared       =    0.0940
-------------+----------------------------------   Adj R-squared   =    0.0933
       Total |  444.979972     1,259  .353439215   Root MSE        =    .56611

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0152338   .0013336    11.42   0.000     .0126174    .0178501
       _cons |   1.381448   .0290495    47.55   0.000     1.324457    1.438439
------------------------------------------------------------------------------
```

- An additional year of experience increases average wage by approximately 1.5%

Econ digression: elasticity

- Expressing changes in terms of percentages is near and dear to economists because it is related to the concept of **elasticity**
- What happens to the demand of iWatches when the price increases? What about table salt? Comparing a 1 unit change in price doesn't make much sense because prices are different. Salt is about super cheap; iWatches are expensive
- Instead, use a common metric for both: **percent changes**
  $Elasticity = \epsilon = \frac{\Delta y}{\Delta x}\frac{x}{y} = \frac{\%\Delta y}{\%\Delta x}$
- So the elasticity is the percent change in y for a percent change in x. (By the way, salt is more inelastic than an iWatch)
- What does this have to do with log transformations?

Using logs to get elasticity

- If we take the log of **both** y and x we can **interpret the parameter of x as an elasticity**
- For example, in the model: $log(wage) = \beta_0 + \beta_1 log(educ) + \epsilon$
- A 1% change in years of education changes wages in by $\beta_1 * 100$ percent
- The proof is a bit complicated (you need to take the implicit derivative); only valid for small changes
- These models are not that common in HSR but are much more common in economics

Retransformation problem

- Back to the more common case of taking the log of the outcome y. Sometimes called the **log-level** model
- We just saw the shortcut but we may not care about the percent change in y but rather the effect of x on the average y
- There is a problem that is often called the **retransformation problem** in the health economics literature
- The earliest recognition of this problem was in the RAND health insurance experiment by Duan, Manning and Co
- See Duan (1983) and Manning (2001)

Not an innocent transformation

- A lot more happens when we take the log of the outcome
- Suppose we have $log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$
- We can solve for y by taking the $e()$ on both sides
- We end up with: $y = e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon)}$. We can rewrite as: $y = e^{\beta_0} \times e^{\beta_1 x_1} \times e^{\beta_2 x_2} \times e^{\beta_3 x_3} \times e^{\epsilon}$
- A non-linear model with multiplicative error. The effect of one variable depends on the value of the others. The effect of $X1$ for example, is:
- $\frac{\partial y}{\partial x_1} = exp(\beta_0) \times exp(\beta_1 x_1) \times exp(\beta_2 x_2) \times exp(\beta_3 x_3) \times exp(\epsilon) \times \beta_1$
- The other problem is that $E[log(y)] \neq log(E[y])$. If the we take the exponent of the predicted $log(\hat{y})$ we are not going to get $E[\hat{y}]$

Retransformation problem

- Easy to see the problem with a very simple model

```
reg lwage abvavg
      Source |       SS           df       MS      Number of obs   =     1,260
-------------+----------------------------------   F(1, 1258)      =      0.05
       Model |  .019425671         1  .019425671   Prob > F        =    0.8148
    Residual |  444.960547     1,258  .353704727   R-squared       =    0.0000
-------------+----------------------------------   Adj R-squared   =   -0.0008
       Total |  444.979972     1,259  .353439215   Root MSE        =    .59473

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      abvavg |  -.0085364   .0364256    -0.23   0.815     -.079998    .0629252
       _cons |   1.661394   .0200826    82.73   0.000     1.621995    1.700793
------------------------------------------------------------------------------
```

- The model is $log(wage) = \beta_0 + \beta_1 abvavg + \epsilon$. For those of below average looks it's just $log(wage) = \beta_0$. If we take the exponent of both sides: $wage = exp(\beta_0)$
- But this is actually **NOT** the average wage for those of below average looks: $E[wage|abvavg = 0] \neq exp(\beta_0)$

Retransformation problem

- Verify

```
qui reg lwage abvavg

* below average looks
di exp(_b[_cons])
5.2666493

* above average
 di exp(_b[_cons] + _b[abvavg])
5.2218825

* Actual for below average
 sum wage if e(sample) & abvavg ==0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        wage |        877    6.286306    4.214598       1.02      38.86

* Actual above average
 sum wage if e(sample) & abvavg ==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        wage |        383    6.353368    5.554582       1.16      77.72
```

- **Underestimated** in both cases even in the simplest of models

Retransformation problem

- Again, this happens for a rather simple reason: $E[log(x)] \neq log(E[x])$ or the expected value of log(x) is not the same as the log of the expected value
- So just taking the exponent function doesn't work
- Actually, it turns out that what the log-level model is giving you is the **geometric mean** rather than the arithmetic mean
- The geometric mean is defined as $(\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$
- For example, the geometric mean of 2,3,4 is $\sqrt[3]{2 * 3 * 4}$

Retransformation problem

- Check that this is the case by using the ameans command

```
ameans wage if e(sample) & abvavg ==0

    Variable |    Type             Obs        Mean       [95% Conf. Interval]
-------------+---------------------------------------------------------------
        wage | Arithmetic          877    6.286306        6.006984   6.565627
             |  Geometric          877    5.266649        5.063633   5.477805
             |   Harmonic          877    4.414825        4.233009   4.612961
-----------------------------------------------------------------------------

ameans wage if e(sample) & abvavg ==1

    Variable |    Type             Obs        Mean       [95% Conf. Interval]
-------------+---------------------------------------------------------------
        wage | Arithmetic          383    6.353368        5.795311   6.911425
             |  Geometric          383    5.221882        4.917211   5.545432
             |   Harmonic          383    4.398983        4.142901    4.68881
-----------------------------------------------------------------------------
```

Retransformation problem

- Interestingly enough, this problem was not apparent until the RAND health insurance experiment
- So we know what is the problem, but what is the solution? Remember, we would like to be able to interpret the coefficients in the wage scale, not the $log(wage)$ scale. We want to understand what is the effect of covariates on $E[wage]$
- Duan (1983) proposed a **smearing** factor, which turns out depends on whether the errors are **heteroskedastic** or not
- In the simplest case of homoskedastic errors the smearing factor is the exponent of the of residuals:
- $smearing = \frac{1}{n}\sum_{i=1}^{n} e^{(ly - \hat{ly})} = \sum_{i=1}^{n} e^{\hat{e}}$, where ly is to emphasize that we use $log(y)$ not y
- You will learn more about it in Methods II because modeling costs is a key issue in HSR

Retransformation problem

- Simple smearing factor

```
* Estimate model again
qui reg lwage abvavg

* Residual
predict lres if e(sample), res
* Exponentiate
gen lresexp = exp(lres)
* Smearing
sum lresexp

 Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     lresexp |      1,260     1.20062     .888256   .1936715   14.88352

* Apply factor
* below average looks
di (exp(_b[_cons]))*r(mean)
6.3232467

* above average
di (exp(_b[_cons] + _b[abvavg]))*r(mean)
6.2694987
```

- The actual means are 6.28 and 6.35, not bad at all

Another way

- The **smearing** solution was developed in the 80s but there are other approaches
- There is a type of models called Generalized Linear Models (GLM) which encompass our linear regression model, logistics, Poisson and many more
- You choose a "family" and and "link" function. A GLM with family Normal (sounds kind of funny) or Gaussian and an identity link is the same as the linear model we have covered this semester
- A GLM with a Normal family and a **log link** is like the log-level model except that **it doesn't have the retransformation problem**
- This is so because GLM estimates $log(E[x])$ rather than $E[log(x)]$

GLM

- Convince yourself

```
glm wage abvavg, family(normal) link(log)
...
Iteration 4:   log likelihood =  -3726.667
Generalized linear models                         No. of obs      =      1,260
Optimization     : ML                             Residual df     =      1,258
                                                  Scale parameter =   21.73787
Deviance         =  27346.24025                   (1/df) Deviance =   21.73787
Pearson          =  27346.24025                   (1/df) Pearson  =   21.73787
Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   5.918519
Log likelihood   =  -3726.66697                   BIC             =   18365.55
------------------------------------------------------------------------------
             |                 OIM
        wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      abvavg |   .0106115   .0450922     0.24   0.814    -.0777676    .0989907
       _cons |   1.838374   .0250445    73.40   0.000     1.789287     1.88746
------------------------------------------------------------------------------
. di exp(_b[_cons])
6.2863056
. di exp(_b[_cons] + _b[abvavg])
6.3533681
```

- Matches the actual means: 6.28 and 6.35

Big picture

- In many circumstances taking the log of the outcome is necessary to make the model fit better since it makes the outcome variable more normally distributed
- But you must be careful with the interpretation of parameters since taking the log induces non-linearity of effects and also changes the interpretation of the coefficients
- A GLM model with log link and Gaussian family provides an alternative

Loose ends

- What about zeros? The log of zero is undefined
- Not uncommon to take $log(x + 1)$ when $x = 0$; not much is loss
- There is a large literature on modeling **cost data**. Health care cost data (not all cost data) tend to be skewed, with many zeroes or low values, and a large tail, which means that SEs of cost models are likely to be wrong (although not terribly wrong either)
- There are some tests to diagnose functional form specifications, like **Ramsey's regression specification error test (RESET)**. Super simple idea: no other retransformation of the Xs should be better
- **Box-cox** transformations (some transformations make parameters hard to interpret)
- You will see them next semester in the context of analyzing cost data

Transformation to achieve linearity

- Your textbook has examples about transformations to achieve linearity
- For example, we may want to model exponential growth: $y = \alpha X^{\beta}$, which is not linear on β but can be made linear by taking the log:
- $log(y) = log(\alpha) + \beta x$, which is the log-level model we have just seen
- The last one in the textbook table is $y = \frac{exp(\alpha+\beta x)}{1+exp(\alpha+\beta x)}$. This is the **logit** transformation
- For all values of α, β, and x the outcome y is restricted to be between 0 and 1
- Useful to model probabilities. Can be made linear: $log(\frac{y}{1-y}) = \alpha + \beta x$. That's the **log-odds scale**

Transformation to stabilize variance

- We have briefly covered **heteroskedasticy**, when the variance conditional on explanatory variables is not the same
- This is a common violation of the linear model. By stabilize, we mean **making the variance constant** conditional on Xs
- We have seen this problem in many of the examples we have covered

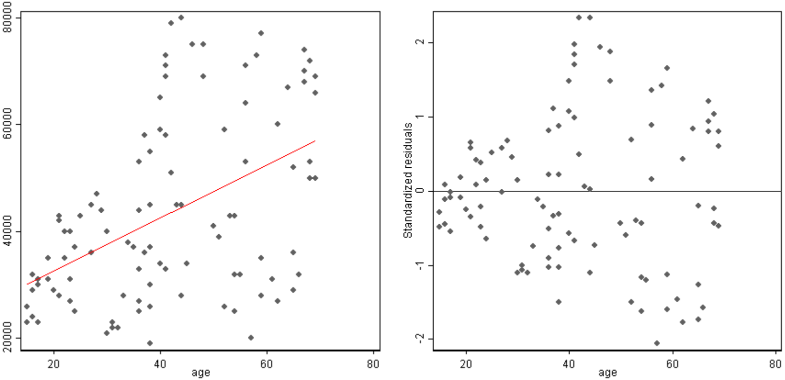
Heteroskedastic errors

- heteroskedastic errors are fairly common

```
webuse mksp1, clear

* Plot income and age
scatter income age || lfit income age, color(red) ///
    saving(inage.gph, replace) legend(off)

* Get standardize residuals
qui reg income age
predict inres, rstandard

* Plot residuals
scatter inres age, yline(0) saving(inres.gph, replace)
graph combine inage.gph inres.gph, ysize(10) xsize(20)
```

- Note the **ysize()** and **xsize()** options

Heteroskedastic erros



- Clearly not that great

Big picture reminder

- When do we need the assumption of equal variance?
- We didn't need it to estimate the parameters of the linear model (with OLS)
- We do need the assumption for statistical inference
- One issue with heteroskeasticity erros is that SEs **tend** to be smaller, so we think that we have more precision
- Some transformations tend to make the assuption of constant variance (conditional on x) more plausible

Common transformations

- Your textbook has some examples of transformations
- It is somewhat outdated and in many cases unnecessary; there are other options
- For example, taking the $\sqrt{y}$ of count data may help make the assumption of constant variance more realistic (assuming that the data comes from a Poisson distribution)
- In Poisson random variable, the mean and the variance are the same
- But if we know that, why not use a Poisson model instead? By now, GLM models are mainstream

Example

- Number of children using income and education as predictors

```
reg children educ incthou

      Source |       SS           df       MS      Number of obs   =     1,189
-------------+----------------------------------   F(2, 1186)      =     21.14
       Model |   95.277586         2   47.638793   Prob > F        =    0.0000
    Residual |  2672.33553     1,186  2.25323401   R-squared       =    0.0344
-------------+----------------------------------   Adj R-squared   =    0.0328
       Total |  2767.61312     1,188  2.32964067   Root MSE        =    1.5011

------------------------------------------------------------------------------
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0997509   .0155237    -6.43   0.000     -.130208   -.0692939
     incthou |   .0014315   .0005881     2.43   0.015     .0002777    .0025853
       _cons |   3.020157   .2155896    14.01   0.000     2.597178    3.443137
------------------------------------------------------------------------------
```

- Do we trust p-values if we know that the outcome is Poisson and not normal?

Example

- Taking the square root

```
gen sqrtc = sqrt(children)

reg sqrtc educ incthou

     Source |       SS           df       MS      Number of obs   =     1,189
-------------+----------------------------------   F(2, 1186)      =     18.95
       Model |  20.9654628         2  10.4827314   Prob > F        =    0.0000
    Residual |  656.024669     1,186  .553140531   R-squared       =    0.0310
-------------+----------------------------------   Adj R-squared   =    0.0293
       Total |  676.990132     1,188  .569857014   Root MSE        =    .74373

------------------------------------------------------------------------------
       sqrtc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0453013   .0076915    -5.89   0.000    -.0603918   -.0302109
     incthou |   .0008992   .0002914     3.09   0.002     .0003276    .0014709
       _cons |   1.654554   .1068175    15.49   0.000     1.444982    1.864126
------------------------------------------------------------------------------
```

- Interpretation changes of course. But are the SEs better? Maybe...

Example

- GLM for a Poisson correcting for overdispersion

```
 glm children educ incthou, family(poisson) link(log) scale(x2)
Generalized linear models                          No. of obs      =      1,189
Optimization     : ML                              Residual df     =      1,186
                                                   Scale parameter =          1
Deviance         =   1853.32038                    (1/df) Deviance =   1.562665
Pearson          =   1586.99291                    (1/df) Pearson  =   1.338105
Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]
                                                   AIC             =   3.448657
Log likelihood   =  -2047.22665                    BIC             =  -6544.589
------------------------------------------------------------------------------
             |                 OIM
    children |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0574921   .0089408    -6.43   0.000    -.0750157   -.0399684
     incthou |   .0007919   .0003182     2.49   0.013     .0001681    .0014156
       _cons |   1.278896   .1203865    10.62   0.000     1.042943    1.514849
------------------------------------------------------------------------------
(Standard errors scaled using square root of Pearson X2-based dispersion.)
```

- The option scale(x2) uses the Pearson's chi-squared correction for overdispersion
- Note that SEs are closer to the model that does NOT use $\sqrt{children}$

Example

- Parameter interpretation is a bit more complicated as usual with non-linear models
- In Poisson models, taking the exponent of the coefficients makes them have a **relative risk** interpretation
- As an alternative, we can numerically take the derivative in the "children" scale rather than the log(children) scale

```
 margins, dydx(educ)

Average marginal effects                        Number of obs     =      1,189
Model VCE    : OIM
Expression   : Predicted mean children, predict()
dy/dx w.r.t. : educ
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0969968   .0132187    -7.34   0.000    -.1229049   -.0710886
------------------------------------------------------------------------------
```

- The margins command is worth the price of Stata. We will see more about the margins command when we cover logistic models

Some transformations are based on theory

- Suppose that you have data on the area and perimeter of old churches and want to predict the area based on the perimeter

  $area_i = \beta_1 + \beta_0 perimeter_i + \epsilon_i$
- You'll probably have a pretty good model but the relationship won't be linear
- Churches are (more or less) squares and the area of a square is s^2, where is s is the length of a side. The perimeter is $4 \times s$, so the relationship between area and perimeter is non-linear
- The fit will be much better if we instead model $\sqrt{area}_i = \gamma_0 + \gamma_1 perimeter_i + \epsilon_i$
- This is a favorite stats question. I have seen it with trees (they are triangles) and circles
- Good didactic way of teaching transformations. Sadly, not that great in the social sciences or HSR

Summary

- Modeling is a key part of analyzing data
- We transform variables for presentation, interpretation or to make the data fit model assumptions
- We will deal with violations of some assumptions next week
- Then logistic regression and more modeling