Week 7: Cost data and Generalized Linear Models

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2020

Outline

- Medical care cost data characteristics
- Linear/OLS models
- log-level models and the retransformation model
- GLM models
- GLM with log link and Gaussian family
- GLM with Gamma family
- Interpreting parameters: marginal effects and nonlinear, nonadditive effects
- Dealing large proportion of zeroes: two-part models

Medical cost data

- We already saw that medical cost data have some unique characteristics that have consequences for statistical modeling
- Cost are non-negative and tend to be skewed to the right, with a large portion of observations having low expenditures but a fraction having very large expenditures
- Depending on the type of cost (e.g. outpatient vs inpatient) and population (e.g. elderly vs young), there could be a large proportion of observations with zero costs
- This shouldn't be surprising. Medical costs are related to illness, and illness doesn't hit everybody at the same time – even with chronic conditions
- Most of medical expenditures in a year are incurred by a small portion of people
- Be mindful that we are talking about medical costs, not prices – prices tend to be closer to normally distributed, but of course they can't be negative

Data

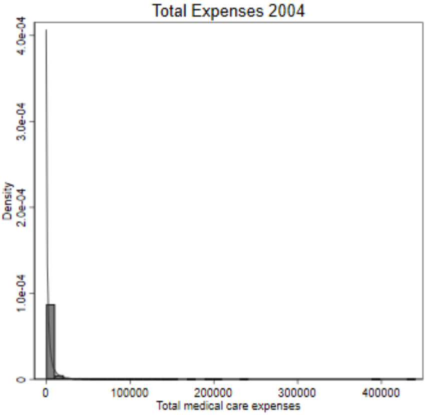
- MEPS 2004 data from Deb, Norton, and Manning (2017)

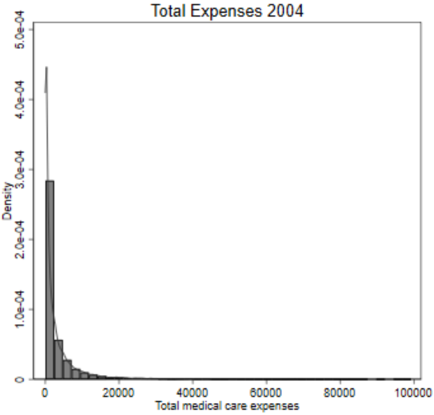
```
use http://www.stata-press.com/data/heus/heus_mepssample, clear
desc exp_* age female pcs race*
              storage   display    value
variable name   type    format     label      variable label
----------------------------------------------------------------------------------------------------
exp_tot         long    %12.0g                Total medical care expenses
exp_ip          float   %9.0g                 Inpatient expenses = exp_ip_fac + exp_ip_md
exp_ip_fac      long    %12.0g                Inpatient facility expenses
exp_ip_md       int     %8.0g                 Inpatient md expenses
exp_er          int     %9.0g                 ER expenses = exp_er_fac + exp_er_md
exp_er_fac      int     %12.0g                ER facility expenses
exp_er_md       int     %8.0g                 ER md expenses
exp_dent        int     %8.0g                 Dental care expenses
exp_self        long    %12.0g                Total expenses paid by self or family
age             byte    %8.0g                 Age
female          byte    %9.0g      lb_female
                                              Female
pcs12           double  %10.0g                Physical health component of SF12
race_bl         byte    %14.0g     lb_race_bl
                                              Black
race_oth        byte    %14.0g     lb_race_oth
                                              Other race, non-white and non-black
```

```
sum exp_tot exp_ip exp_er exp_dent exp_self

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     exp_tot |     19,386     3685.25    9768.475          0     440524
      exp_ip |     19,386    1122.972     7283.09          0     376987
      exp_er |     19,386    130.1588    685.5471          0      20545
    exp_dent |     19,386    211.2738    657.1742          0      16275
    exp_self |     19,386    685.2889    1468.705          0      50850
```

Total expenditures in 2014

```
hist exp_tot, kdensity title("Total Expenses 2004") saving(thist1.gph, replace)
hist exp_tot if exp_tot < 100000, kdensity title("Total Expenses 2004") saving(thist2.gph, replace)
graph combine thist1.gph thist2.gph, ysize(10) xsize(20)
graph export histc.png, replace
```





Exploring a bit more

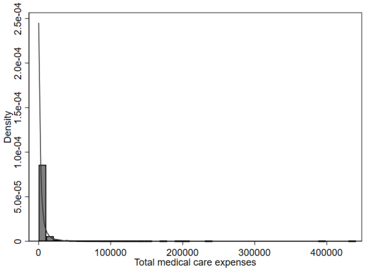
- Check percentiles. It happens at all ages

```
* all ages
tabstat exp_tot, stats(N mean p5 p10 p50 p75 p90 p99)

    variable |         N      mean        p5       p10       p50       p75       p90       p99
-------------+--------------------------------------------------------------------------------
     exp_tot |     19386   3685.25         0         0       952      3507      8940     41373
----------------------------------------------------------------------------------------------


* older than 75
tabstat exp_tot if age >75, stats(N mean p5 p10 p50 p75 p90 p99)

    variable |         N      mean        p5       p10       p50       p75       p90       p99
-------------+--------------------------------------------------------------------------------
     exp_tot |      1285  8900.486       374       764      4159      9594     22161     71343
----------------------------------------------------------------------------------------------

gen     zero = 0
replace zero = 1 if exp_tot ==0

tab zero
       zero |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |     15,946       82.26       82.26
          1 |      3,440       17.74      100.00
------------+-----------------------------------
      Total |     19,386      100.00

tab zero if age > 75

       zero |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,267       98.60       98.60
          1 |         18        1.40      100.00
------------+-----------------------------------
      Total |      1,285      100.00
```

It's not just the zeroes

- The "excess" zeroes pose a statistical problem, but the distribution is skewed without the zeroes as well

```
tabstat exp_tot if exp_tot >0, stats(N mean sd p5 p10 p50 p75 p90 p99 min max)

    variable |         N      mean        sd        p5       p10       p50       p75       p90
-------------+--------------------------------------------------------------------------------
     exp_tot |     15946  4480.262  10604.14        83       153      1537      4482     10476
----------------------------------------------------------------------------------------------

    variable |       p99       min       max
-------------+------------------------------
     exp_tot |     44065         2    440524
--------------------------------------------

hist exp_tot if exp_tot >0, kdensity
graph export noz.png, replace
```



Modeling cost data

- Say that we want to estimate a model like this with total expenditure during the year as the dependent/outcome variable:

  $exp\_tot_i =$
  $\beta_0 + \beta_1 age_i + \beta_2 female_i + \beta_3 pcs_i + \beta_4 race\_bl_i + \beta_5 race\_oth_i + \beta_6 eth\_hisp_i + \epsilon_i$
- We want to understand factors that affect $E[exp\_tot|\mathbf{X}]$ as a function of age, sex, physical functioning, and race/ethnicity
- We could use our trusty linear/OLS model since we know that it's an unbiased conditional expectation function
- But we know that SEs are not correct since costs do not distribute normal and there are likely heteroskedastic problems
- At the very least we need to use robust SEs (robust option in reg command)

Linear/OLS model

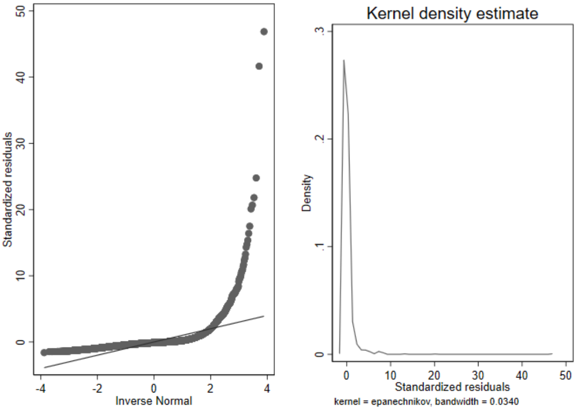
- Interpretation is straightforward. We can check the residuals and predictions

```
reg exp_tot age i.female pcs race* eth_hisp, robust

Linear regression                               Number of obs     =     19,386
                                                F(6, 19379)       =     198.97
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1283
                                                Root MSE          =     9121.6

------------------------------------------------------------------------------
             |               Robust
     exp_tot |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   53.67021   5.448849     9.85   0.000     42.98999    64.35042
             |
      female |
     Female  |   545.4941   138.9665     3.93   0.000     273.1078    817.8804
       pcs12 |   -255.709   13.96654   -18.31   0.000    -283.0846   -228.3334
     race_bl |  -1208.192   181.9308    -6.64   0.000    -1564.793   -851.5923
    race_oth |  -1583.594   195.7612    -8.09   0.000    -1967.303   -1199.885
    eth_hisp |  -1704.833   135.9056   -12.54   0.000    -1971.219   -1438.446
       _cons |   14140.71   950.2784    14.88   0.000     12278.08    16003.34
------------------------------------------------------------------------------

. predict yhat
(option xb assumed; fitted values)

. sum yhat

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        yhat |     19,386     3685.25    3499.218  -3592.533   17268.99

. qui reg exp_tot age female pcs

. predict res, rstandard
```

Linear/OLS model

- Not good at all. Predictions are negative, residuals not even close to normal, some large residuals. Unlikely that different specifications or covariates can account for shape of residuals

```
qnorm res, saving(qno.gph, replace)
kdensity res, saving(hisres.gph, replace)
graph combine qno.gph hisres.gph
graph export res.png, replace
```



Transformations

- You probably learned in intro classes that transformations of the outcome variable can improve model fit when there are violations of linear/OLS assumptions
- The most common for cost data is to take the log (the natural log; often we don't distinguish between log and ln) of the cost since taking the log of skewed data tend to produce distributions that look normal
- We will focus on the natural log (ln), but the ln transformation is part of the Box-cox type of transformation, given by:

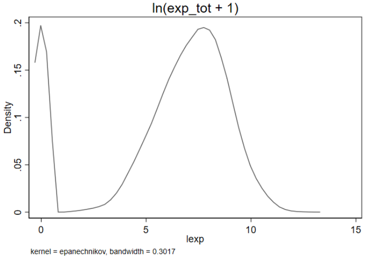
$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & if\, \lambda \neq 0 \\ ln(y) & if\, \lambda = 0 \end{cases}$$

- Box-cox models use MLE to find the parameter to transform the model (or outcome). See Stata help for command -boxcox-

Log transformation

- The most common transformation –the knee-jerk transformation– with skewed data is to use $ln(y)$ (called log-level model since we leave the covariates as they are)
- $ln(0)$ is undefined so we need to add 1 to the cost data without losing much, but it's a bit odd
- The outcome looks closer to normal but we have that peak for costs equal to 1 (the previous zeroes)

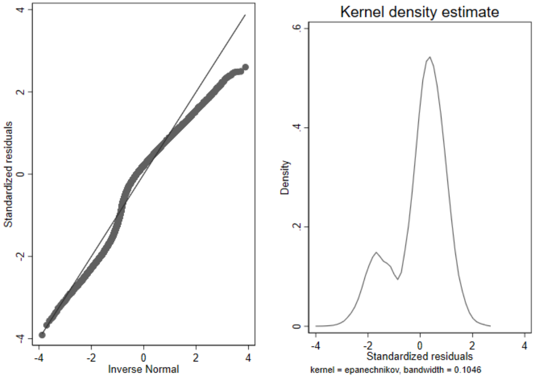
```
gen lexp = log(1+exp_tot)
kdensity lexp, title("ln(exp_tot + 1)")
graph export lexp.png, replace
```



Ln transformation

- Residuals look better, not great, but much better. Would be excellent without the zeroes

```
qui reg lexp age female pcs race*
predict resl, rstandard

qnorm resl, saving(qnol.gph, replace)
kdensity resl, saving(hisresl.gph, replace)
graph combine qnol.gph hisresl.gph
graph export resl.png, replace
```



Log transformation

- Note that the issue are the zeroes, transformed into *ln*(1). If we restricted the analysis to expenditures greater than zero, the ln transformation would be very reasonable. Box-Cox suggests so as well. In the output below θ would be the Box-Cox λ. We reject the null that is zero but it's close to zero
- See do file for today (the Box-Cox model doesn't change conclusions in terms of SEs and p-values in this example)

```
boxcox exp_tot1 age female pcs race* eth_hisp if exp_tot > 0, model(lhsonly) lrtest nolog nologlr
Fitting comparison model
Fitting full model
Fitting comparison models for LR tests
                                                  Number of obs   =     15,946
                                                  LR chi2(6)      =    4916.94
Log likelihood = -143350.1                        Prob > chi2     =      0.000

------------------------------------------------------------------------------
    exp_tot1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      /theta |   .0640056   .0039185    16.33   0.000     .0563255    .0716857
------------------------------------------------------------------------------

<...>
---------------------------------------------------------
   Test         Restricted     LR statistic      P-value
    H0:       log likelihood       chi2       Prob > chi2
---------------------------------------------------------
theta = -1      -179058.31     71416.43           0.000
theta =  0      -143483.87       267.54           0.000
theta =  1      -169451.21     52202.23           0.000
---------------------------------------------------------
```

Log transformation - interpretation

- Below is the fitted model (including observations with zero total expenditure). Now we need to face another problem: **how do we interpret the coefficients in the $ scale**?
- The estimated model is $E[ln(Y)|\mathbf{X}] = \beta_0 + \beta_1 X_1 + \cdots + \beta_j X_j$

```
reg lexp age female pcs race* eth_hisp

      Source |       SS           df       MS      Number of obs   =    19,386
-------------+----------------------------------   F(6, 19379)     =   1254.91
       Model |  53406.9521         6  8901.15868   Prob > F        =    0.0000
    Residual |  137456.862    19,379  7.09308336   R-squared       =    0.2798
-------------+----------------------------------   Adj R-squared   =    0.2796
       Total |  190863.814    19,385   9.8459538   Root MSE        =    2.6633

------------------------------------------------------------------------------
        lexp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0435419   .0012345    35.27   0.000     .0411221    .0459616
      female |   1.093938   .0386264    28.32   0.000     1.018227    1.169649
       pcs12 |  -.0654314   .0019186   -34.10   0.000     -.069192   -.0616708
     race_bl |  -1.020951   .0577763   -17.67   0.000    -1.134197    -.907704
    race_oth |   -.774305   .0792569    -9.77   0.000    -.9296554   -.6189545
    eth_hisp |  -1.793879   .0484721   -37.01   0.000    -1.888888   -1.698869
       _cons |   7.181796   .1349736    53.21   0.000     6.917236    7.446356
------------------------------------------------------------------------------
. di 100*(exp(_b[eth_hisp]) -1 )
-83.368614
```

Ln transformation - interpretation

- There is a shortcut (approximation) to interpret log-level model coefficients
- For continuous variables, we can interpret them as percent changes. For example, an additional point in the PCS12 score decreases expenditure by about 6.5%, holding other factors constant. An additional year of age increases expenditures by about 4.35%
- For dummy variables, we use $\Delta\% Y \approx 100(e^{\hat{\beta}_j} - 1)$
- So average expenditure for Hispanics is 83% lower than for whites, adjusting for other factors
- It's a convenient way to interpret models, **but we may still want to interpret models in the orginal scale**, \$
- (There is a modification for dummy variables called the "Kennedy transformation"; see DNM)

Ln transformation

- The log transformation is not an innocent transformation. The problem is easier to see using using the population model $ln(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_j X_j + \epsilon$
- Taking the exponent on both sides: $Y = e^{(\beta_0 + \beta_1 X_1 + \cdots + \beta_j X_j + \epsilon)} = e^{(\beta_0 + \beta_1 X_1 + \cdots + \beta_j X_j)} e^{\epsilon}$
- If we now take the conditional expectation we get: $E[Y|\mathbf{X}] = e^{(\beta_0 + \beta_1 X_1 + \cdots + \beta_j X_j)} E[e^{\epsilon}|\mathbf{X}] = e^{\beta_0} \times e^{\beta_1 X1} \times \cdots \times e^{\beta_j X_j} \times E[e^{\epsilon}|\mathbf{X}]$
- So taking the exponent of the estimated model is not going to give us what we want, although we could find a solution by trying to come up with $E[e^{\epsilon}|\mathbf{X}]$
- The bottom line of the story is that with the linear/OLS model we estimated $E[ln(Y)|\mathbf{X}]$, but $E[ln(Y)|\mathbf{X}] \neq ln(E[Y|\mathbf{X}])$
- If we could instead estimate $ln(E[Y|\mathbf{X}])$, exponentiation would give us what we want: $e^{ln(E[Y|\mathbf{X}])} = E[Y|\mathbf{X}]$

Duan's smearing factor

- From the previous slide, we could retransform the model back into the \$ scale if we find $E[e^{\epsilon}|\mathbf{X}]$
- The answer is just there in the formula: we can use the residuals of the model, $\hat{\epsilon}$ to estimate $E[e^{\epsilon}|\mathbf{X}]$
- If we assume that the error distributes normal the correction factor is $D_{norm} = e^{\frac{1}{2}\bar{\hat{\epsilon}}^2{}_i}$
- To relax the normality assumption, we can use Duan's smearing factor instead: $D_{smear} = \sum_{i=1}^{n} \frac{e^{\hat{\epsilon}_i}}{n}$
- Note that in these formulas the residual is the residual of the log-level model
- After we find the smearing factor, $E[Y|\mathbf{X}] = e^{\mathbf{X}'\beta} \times D_{smear}$
- Since we already know that marginal effects are based on predictions and we just found a way of calculating predictions in the dollar scale, we can then get marginal effects

Duan's smearing factor

- The steps are straightforward:
  1. Estimate the log-level model
  2. Estimate the model residuals $\hat{\epsilon}_i$
  3. Take the exponent of the residuals: $e^{\hat{\epsilon}_i}$
  4. The mean of step 3) is the smearing factor D_{smear}
- With the smearing factor in hand we can obtain predictions in the $ scale
- Again: this means that we can also find **marginal effects** in the $ scale
- Marginal and incremental effects are predictions

Example

- Below is example for positive expenditure where the Duan's smearing factor works best

```
qui reg lexp age female pcs race* eth_hisp if exp_tot > 0
predict epsilonhat, residual

* Predictions in ln scale
predict lyhat
* Exponent of predictions
gen explyhat = exp(lyhat)

* Duan's smearing factor
egen dduan = mean(exp(epsilonhat))
* Transform exponent of predictions
gen yhatduan = explyhat * dduan

sum yhatduan exp_tot if exp_tot > 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    yhatduan |     15,946    5090.052    5602.295   448.9315   65638.84
     exp_tot |     15,946    4480.262    10604.14          2     440524
```

Generalized Linear Models

- Rather than using retransformations that have many issues we can use Generalized Linear Models (GLM) that do not require retransformations (although with a catch)
- We will only scratch the surface of GLMs, but they are simple to implement with the tools we have learned. In fact, all the models we used so far are special cases of GLM models
- GLMs offer a unified theory for a class of regression models that have a distribution in the **exponential family** of distributions
- And it happens that the normal, binomial/bernoulli , probit, Poisson, and Gamma distributions are part of the exponential family

Generalized Linear Models - elements

- I'll follow Hardin and Hilbe (2018) in describing the key elements of GLMs
  1. A **random component** for the response Y that follows a distribution belonging to the exponential family (think of the error term ϵ in linear models)
  2. A **linear systematic component** relating the predictors $\mathbf{X}$ and coefficients, $\eta = \mathbf{X}'\beta = \beta_0 + \beta_1 X_1 + \cdots + \beta_j X_j$
  3. A **link function** relating the linear predictors to the fitted predictors. Function is monotonic, one-to-one, and differentiable. We can link the $E[Y]$ to the linear predictors: $E[Y] = g^{-1}(\eta) = g^{-1}(\mathbf{X}'\beta) = \mu$. In the linear/OLS model the function is the identify function: $E[Y] = \mathbf{X}'\beta$
  4. The variance may change with the covariates only as a function of the mean
  5. There is one Iterative Reweighted Least Squares algorithm (IRLS) (to compute estimates) that fits all members of the class
- We will focus on 1 to 3; 4 and 5 are more technical
- Although IRLS unifies GLM, Stata's default is MLE estimation. You can requests models to be estimated using IRLS with the irls option

Exponential family

- The exponential family density function can be written as
  $f(y;\theta,\phi) = e^{\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\}}$
- (**Go back to basics**: a probability density function gives you the values that a random variable can take –domain, support– and their probabilities)
- The θ parameter is the location parameter that relates to the mean (location), while the parameter ϕ relates to the scale (variance)
- If we observe $y_1, \ldots y_n$ independent observations we can write the log-likehood function as well:
  $ll(\theta, \phi, y_1 \ldots y_n) = \sum_{i=1}^{n}\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\}$
- Now, this is still a bit too abstract but the key is that by changing how we define θ and ϕ and how parameters relate to θ, we can estimate different pdf's that generate different models
- Essentially defining θ and ϕ defines different distributions, like the normal (Gaussian), binomial, Gamma, etc

GLM - normal/Gaussian family

- A GLM model with a **Gaussian/normal family** and an **identity link** is our standard linear/OLS model
- The Gaussian/normal density function in the exponential-family form means that $\theta = \mu$ and $b(\theta) = \frac{\mu^2}{2}$:

  $f(y; \mu, \sigma^2) = e^{\{-\frac{(y-\mu)^2}{2\sigma^2} - \frac{1}{2} ln(2\pi\sigma^2)\}}$
- That's the normal density that we saw in the MLE class written in a different way. In the MLE class it was $f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$
- We only need to show that $e^{-\frac{1}{2} ln(2\pi\sigma^2)} = \frac{1}{\sqrt{2\pi\sigma^2}}$, but that's straightforward once you remember two of the rules of exponents: $a^{\frac{x}{y}} = \sqrt[y]{a^x}$ and $a^{-x} = \frac{1}{a^x}$
- So if we assume a GLM with Gaussian family the likelihood function will be the same as before

GLM - normal/Gaussian family

- With covariates, we make μ a function of parameters
- The identity link implies $\mu = E[Y] = \mathbf{X}'\beta$
- Contrary to logistic regression, we don't need to worry about other links to constraint the values of Y. With logistic regression, we use the logit transofrmation but here it's just the identify function
- The log-likelihood becomes:

  $ll(\mu, \sigma^2; y) = \sum_{i=1}^{n} \{ \frac{y_i \mathbf{X}'\beta - (\mathbf{X}'\beta)^2/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2} ln(2\pi\sigma^2) \}$
- Again, this is in fact the same log-likelihood function we saw in the MLE class for the vanilla linear/OLS model
- Now the maximization problem is finding the vector β that maximizes the log-likelihood function. As before, Stata will do it numerically using the -glm- command, but the algorithm will be different than in the MLE class (you don't need to worry about that part)

Example

- At the start of the class we estimated the linear/OLS model below using the -reg- command

```
reg exp_tot age i.female pcs race* eth_hisp, robust

Linear regression                               Number of obs     =     19,386
                                                F(6, 19379)       =     198.97
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1283
                                                Root MSE          =     9121.6

------------------------------------------------------------------------------
             |               Robust
     exp_tot |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   53.67021   5.448849     9.85   0.000     42.98999    64.35042
             |
      female |
     Female  |   545.4941   138.9665     3.93   0.000     273.1078    817.8804
       pcs12 |   -255.709   13.96654   -18.31   0.000    -283.0846   -228.3334
     race_bl |  -1208.192   181.9308    -6.64   0.000    -1564.793   -851.5923
    race_oth |  -1583.594   195.7612    -8.09   0.000    -1967.303   -1199.885
    eth_hisp |  -1704.833   135.9056   -12.54   0.000    -1971.219   -1438.446
       _cons |   14140.71   950.2784    14.88   0.000     12278.08    16003.34
------------------------------------------------------------------------------
```

Example

- The same model is a GLM with Gaussian/normal family and identify link
- The "pseudo-likelihood" refers to the way GLM estimates the variance: it's a function of the mean (Nelder and Lee, 1992)

```
. glm exp_tot age i.female pcs race* eth_hisp, family(gaussian) link(identity) vce(robust)

Iteration 0:   log pseudolikelihood = -204273.44

Generalized linear models                         Number of obs   =     19,386
Optimization     : ML                             Residual df     =     19,379
                                                  Scale parameter =   8.32e+07
Deviance         =  1.61242e+12                   (1/df) Deviance =   8.32e+07
Pearson          =  1.61242e+12                   (1/df) Pearson  =   8.32e+07

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   21.07505
Log pseudolikelihood = -204273.4396               BIC             =   1.61e+12

------------------------------------------------------------------------------
             |               Robust
     exp_tot |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   53.67021   5.448006     9.85   0.000     42.99231     64.3481
             |
      female |
    Female   |   545.4941    138.945     3.93   0.000     273.1669    817.8213
       pcs12 |   -255.709   13.96438   -18.31   0.000    -283.0787   -228.3393
     race_bl |  -1208.192   181.9027    -6.64   0.000    -1564.715   -851.6697
    race_oth |  -1583.594   195.7309    -8.09   0.000    -1967.219   -1199.968
    eth_hisp |  -1704.833   135.8846   -12.55   0.000    -1971.161   -1438.504
       _cons |   14140.71   950.1314    14.88   0.000     12278.49    16002.94
------------------------------------------------------------------------------
```

GLM Gaussian family with identity link

- I used the robust SEs in both models
- Identical models. Note that in GLM the Wald test is z not t-student (asymptotically equivalent – that is, consistent)
- The deviance/Pearson statistics is analogous to the residual sum of squares
- We get BIC and AIC, although the formulas are slightly different for the GLM model in Stata
- **So what do we gain from using a GLM with identity link and Gaussian family?**
- Not much really. BUT, we are about to gain something
- **What about changing the link function**? Let's use the log link instead

GLM Gaussian family with log link

- The log-likelihood with the identity link was:
- $ll(\mu, \sigma^2; y) = \sum_{i=1}^{n} \{ \frac{y_i \mathbf{X}'\beta - (\mathbf{X}'\beta)^2/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2} ln(2\pi\sigma^2) \}$
- The log-likelihood with the log link is:

  $ll(\mu, \sigma^2; y) = \sum_{i=1}^{n} \{ \frac{y_i exp(\mathbf{X}'\beta) - (exp(\mathbf{X}'\beta))^2/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2} ln(2\pi\sigma^2) \}$
- So we changed $\mu = \mathbf{X}'\beta$ to $ln(\mu) = \mathbf{X}'\beta$, or equivalent to $ln(E[Y]) = \mathbf{X}'\beta$ since $\mu = E[Y] = e^{(\mathbf{X}'\beta)}$
- This may seem trivial, but in doing so **we just got rid of the retransformation problem**
- With GLM, we estimate $ln(E[Y]) = \mathbf{X}'\beta$, which means that if we take the exponent we have $E[Y] = e^{\mathbf{X}'\beta}$
- Remember, the problem with linear/OLS log-evel models is that we model $E[log(Y)] = \mathbf{X}'\beta$ and $E[ln(Y)|\mathbf{X}] \neq ln(E[Y|\mathbf{X}])$

GLM with log link

- The coefficients are in the ln scale, taking the exponent they become relative rates. Ignoring covariates (or fixing them at some value): $ln(E[Y_{female}]) - ln(E[Y_{male}]) = \beta_{female}$, so $\frac{E[Y_{female}]}{E[Y_{male}]} = e^{\beta_{female}}$

```
glm exp_tot age i.female pcs race* eth_hisp, family(gaussian) link(log) robust nolog
Iteration 7:   log pseudolikelihood = -204234.56

Generalized linear models                          Number of obs   =     19,386
Optimization     : ML                              Residual df     =     19,379
                                                   Scale parameter =   8.29e+07
Deviance         =  1.60596e+12                    (1/df) Deviance =   8.29e+07
Pearson          =  1.60596e+12                    (1/df) Pearson  =   8.29e+07

Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =   21.07104
Log pseudolikelihood = -204234.5627                BIC             =   1.61e+12

------------------------------------------------------------------------------
             |               Robust
     exp_tot |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .012103   .0023217     5.21   0.000     .0075526    .0166534
             |
      female |
    Female   |   .0534459   .0605223     0.88   0.377    -.0651757    .1720675
       pcs12 |  -.0431595   .0026255   -16.44   0.000    -.0483054   -.0380136
     race_bl |  -.1941958   .0639167    -3.04   0.002    -.3194703   -.0689213
    race_oth |  -.3461089    .107184    -3.23   0.001    -.5561856   -.1360322
    eth_hisp |  -.4321407   .0772245    -5.60   0.000     -.583498   -.2807834
       _cons |   9.699521   .2532714    38.30   0.000     9.203118    10.19592
------------------------------------------------------------------------------
```

GLM with log link

- Check relative costs

```
qui glm exp_tot age i.female pcs i.race* i.eth_hisp, family(gaussian) link(log) vce(robust)

* Take the exponent of the coefficient for female
di exp(_b[1.female])
1.0548999

* Check with predictive margins
margins i.female, post

Predictive margins                              Number of obs     =     19,386
Model VCE    : Robust

Expression   : Predicted mean exp_tot, predict()
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
       Male  |    3738.88   168.7492    22.16   0.000     3408.138    4069.622
     Female  |   3944.144   105.9751    37.22   0.000     3736.437    4151.852
------------------------------------------------------------------------------

. di _b[1.female]/_b[0.female]
1.0548999
```

GLM with log link

- With the eform option you can get the coefficients as relative rates or **relative costs** in this case

```
glm exp_tot age i.female pcs i.race* i.eth_hisp, ///
    family(gaussian) link(log) vce(robust) nolog eform

Generalized linear models                          Number of obs   =     19,386
Optimization     : ML                              Residual df     =     19,379
                                                   Scale parameter =   8.29e+07
Deviance         =  1.60596e+12                    (1/df) Deviance =   8.29e+07
Pearson          =  1.60596e+12                    (1/df) Pearson  =   8.29e+07

Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =   21.07104
Log pseudolikelihood = -204234.5627                BIC             =   1.61e+12
------------------------------------------------------------------------------
             |               Robust
     exp_tot |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.012177     .00235     5.21   0.000     1.007581    1.016793
             |
      female |
     Female  |     1.0549    .063845     0.88   0.377     .9369028    1.187758
       pcs12 |   .9577586   .0025146   -16.44   0.000     .9528427    .9626998
             |
     race_bl |
 Black race  |   .8234966   .0526352    -3.04   0.002     .7265338    .9334001
             |
    race_oth |
 Other race  |   .7074355   .0758257    -3.23   0.001     .5733921    .8728145
             |
    eth_hisp |
   Hispanic  |    .649118   .0501278    -5.60   0.000     .5579433    .7551919
       _cons |   16309.79   4130.803    38.30   0.000     9928.033    26793.74
------------------------------------------------------------------------------
```

GLM with log link - Marginal effects

- With the log link, marginal effects are $\frac{\partial E[Y]}{\partial x_k} = \beta_k e^{\mathbf{X}'\beta}$
- But we now have a bag of tricks and can use margins (note that I use factor variable syntax for all dummy variables)

```
. qui glm exp_tot age i.female pcs i.race* i.eth_hisp, family(gaussian) link(identity) vce(ro
> bust)

. margins, dydx(*)

Average marginal effects                          Number of obs     =     19,386
Model VCE    : Robust

Expression   : Predicted mean exp_tot, predict()
dy/dx w.r.t. : age 1.female pcs12 1.race_bl 1.race_oth 1.eth_hisp

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   53.67021   5.448006     9.85   0.000     42.99231     64.3481
             |
      female |
     Female  |   545.4941    138.945     3.93   0.000     273.1669    817.8213
       pcs12 |   -255.709   13.96438   -18.31   0.000    -283.0787   -228.3393
             |
     race_bl |
 Black race  |  -1208.192   181.9027    -6.64   0.000    -1564.715   -851.6697
             |
    race_oth |
 Other race  |  -1583.594   195.7309    -8.09   0.000    -1967.219   -1199.968
             |
    eth_hisp |
   Hispanic  |  -1704.833   135.8846   -12.55   0.000    -1971.161   -1438.504
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

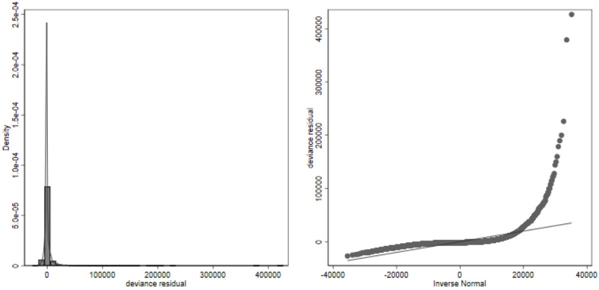
Big picture

- With a GLM model with Gaussian family and log link we don't have a retransformation problem anymore
- **But it doesn't mean that we did something that makes sense**. Remember, we took the ln so costs are closer to a normal distribution, which fits the assumptions of linear/OLS model
- We don't quite achieve this by retransforming $ln(E[Y])$. We will analyze the model residuals below
- Now, in this particular dataset, we still need to deal with the zeroes
- Note that we didn't have to add a 1 to the costs since we model $ln(E[Y])$
- We will deal with the non-normal costs issue soon

GLM with log link - Residuals

- Residuals in GLM models are of several types: Pearson, deviance and ascombe
- We will use the deviance residuals for the linear model. As the results below show, not good at all; in fact, it looks the same as with the linear/OLS model

```
qui glm exp_tot age i.female pcs i.race* i.eth_hisp, ///
    family(gaussian) link(log) vce(robust) nolog
predict double resloglink if e(sample), pearson

hist resloglink, kdensity saving(glmres.gph, replace)
qnorm resloglink, saving(glmresqnorm.gph, replace)
graph combine glmres.gph glmresqnorm.gph, xsize(10) ysize(5)
graph export glmg.png, replace
```

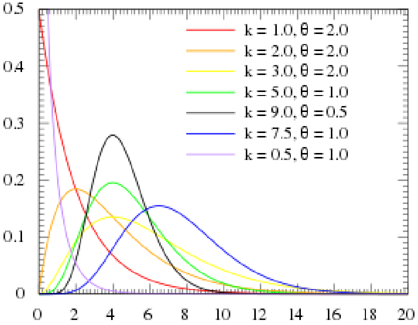


GLM - more options to solve main problem

- We know that the main source of problems is that costs do not distribute normal
- **So why not try other exponential family distributions instead of the normal distribution?**
- There is one option that is particularly appealing: the Gamma distribution
- The Gamma distribution has two parameters, the scale parameter and the shape parameter
- The domain or support is restricted to only positive continuous numbers, like cost data: $x \in (0, \infty)$

Gamma distribution

- In the exponential family distribution, the density is given by:
  $f(y;\mu,\phi) = exp\{\frac{y/\mu-(-ln\mu)}{-\phi} + \frac{1-\phi}{\phi}lny - \frac{ln\phi}{\phi} - ln\Gamma(\frac{1}{\phi})\}$
- $\Gamma()$ is the Gamma function: $\Gamma(n) = (n-1)!$
- Below are some examples of Gamma distributions from Wikipedia



GLM with log link and Gamma family

```
glm exp_tot age i.female pcs i.race* i.eth_hisp, ///
    family(gamma) link(log) nolog

Generalized linear models                         Number of obs   =     19,386
Optimization     : ML                             Residual df     =     19,379
                                                  Scale parameter =    5.03282
Deviance         =  29899.68159                   (1/df) Deviance =   1.542891
Pearson          =   97531.0098                   (1/df) Pearson  =    5.03282

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   17.62673
Log likelihood   = -170848.859                    BIC             =  -161415.7

------------------------------------------------------------------------------
             |                 OIM
     exp_tot |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .019701   .0010318    19.09   0.000     .0176786    .0217234
             |
      female |
     Female  |   .4257809   .0328566    12.96   0.000     .3613833    .4901786
       pcs12 |  -.0518467   .0015811   -32.79   0.000    -.0549456   -.0487477
             |
     race_bl |
 Black race  |  -.3074785   .0488381    -6.30   0.000    -.4031994   -.2117577
             |
    race_oth |
 Other race  |  -.5623746   .0667827    -8.42   0.000    -.6932663   -.4314829
             |
    eth_hisp |
   Hispanic  |  -.7306073   .0409377   -17.85   0.000    -.8108436    -.650371
       _cons |   9.471185   .1109449    85.37   0.000     9.253737    9.688633
------------------------------------------------------------------------------
. di exp(_b[1.female])
1.5307854
```

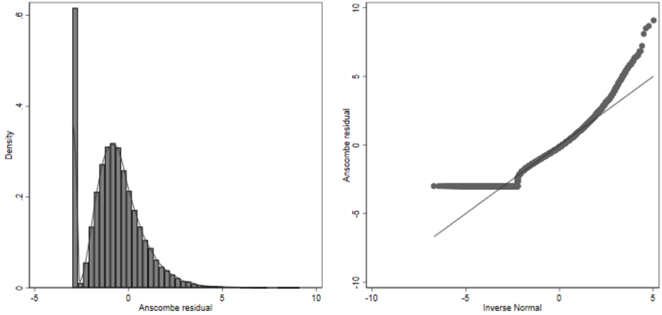
GLM with log link and Gamma family - margins

```
margins, dydx(*)

Average marginal effects                        Number of obs     =     19,386
Model VCE    : OIM

Expression   : Predicted mean exp_tot, predict()
dy/dx w.r.t. : age 1.female pcs12 1.race_bl 1.race_oth 1.eth_hisp

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   78.83696   5.019498    15.71   0.000     68.99892    88.67499
             |
      female |
     Female  |   1618.632    132.952    12.17   0.000     1358.051    1879.214
       pcs12 |  -207.4731   10.17648   -20.39   0.000    -227.4186   -187.5276
             |
     race_bl |
 Black race  |  -1107.139   162.2786    -6.82   0.000    -1425.199   -789.0789
             |
    race_oth |
 Other race  |  -1771.944   171.8934   -10.31   0.000    -2108.849   -1435.039
             |
    eth_hisp |
   Hispanic  |  -2306.634   123.1464   -18.73   0.000    -2547.996   -2065.271
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

GLM with log link and Gamma family - residuals

- We use ascombe residuals since these residuals follow an almost normal distribution. Not great –those zeroes!– but much better otherwise

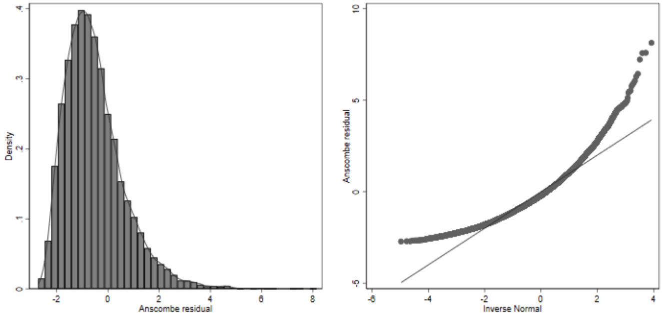
```
qui glm exp_tot age i.female pcs i.race* i.eth_hisp, ///
    family(gamma) link(log) nolog
predict double resgammalog if e(sample), anscombe
* Plot
hist resgammalog, kdensity saving(glmg.gph, replace)
qnorm resgammalog, saving(glmg1.gph, replace)
graph combine glmg.gph glmg1.gph, xsize(10) ysize(5)
graph export resgammalog.png, replace
```



No zeroes

- Let's estimate the model for only those with non-zero expenditure; tails a bit off
- Not bad, not great

```
glm exp_tot age i.female pcs i.race* i.eth_hisp if exp_tot > 0 , ///
    family(gamma) link(log) nolog
predict double resgammalog1 if e(sample), anscombe

hist resgammalog1, kdensity saving(glmg1.gph, replace)
qnorm resgammalog1, saving(glmg11.gph, replace)
graph combine glmg1.gph glmg11.gph, xsize(10) ysize(5)
graph export resgammalog1.png, replace
```



Big picture

- We started the search for a better model because cost data violated the assumptions of the standard linear/OLS model
- We ended up with a GLM with Gamma family and log link as a possible solution
- Our estimates of effects are quite different. This is due to effects being nonlinear with GLM models
- We get better SEs with log-level models and GLMs
- We also get better SEs with log-level models, but then have the retransformation problem, although we could interpret coefficients as percent changes
- In this example, with large sample size, nothing we did changed conclusions (all p-values are very low)
- **We could present parameters that are easier to interpret (dollar scale) but based conclusions on models with better SEs**

Changing family changes parameters estimates - nonlinearity

```
*** Same with only one dummy variable, we are essentially stratifying the sample
tabstat exp_tot if exp_tot >0, by(female)
female |      mean
-------+----------
  Male |  4144.966
Female |  4712.789
-------+----------
qui reg exp_tot i.female if exp_tot > 0
margins female
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
       Male  |   4144.966   131.1843    31.60   0.000      3887.83    4402.102
     Female  |   4712.789   109.2459    43.14   0.000     4498.655    4926.923
------------------------------------------------------------------------------
qui glm exp_tot i.female if exp_tot > 0, family(gamma) /*link(log)*/ nolog
margins female
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
       Male  |   4144.966   123.5248    33.56   0.000     3902.862    4387.071
     Female  |   4712.796   116.9595    40.29   0.000      4483.56    4942.033
------------------------------------------------------------------------------
qui glm exp_tot i.female if exp_tot > 0, family(gamma) link(log) nolog
margins female
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
       Male  |   4144.966   123.6304    33.53   0.000     3902.655    4387.277
     Female  |   4712.789   117.0592    40.26   0.000     4483.357     4942.22
------------------------------------------------------------------------------
```

Changing family changes parameters estimates - nonlinearity

- Only one continuous variable

```
qui reg exp_tot age if exp_tot > 0
margins, dydx(age)
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   124.8428   4.670387    26.73   0.000     115.6883    133.9973
------------------------------------------------------------------------------
qui glm exp_tot age if exp_tot > 0, family(gamma) /*link(log)*/ nolog
margins, dydx(age)
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   141.9435   12.99029    10.93   0.000      116.483     167.404
------------------------------------------------------------------------------
qui glm exp_tot age if exp_tot > 0, family(gamma) link(log) nolog
margins, dydx(age)
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   122.8156   6.838548    17.96   0.000     109.4123    136.2189
------------------------------------------------------------------------------
```

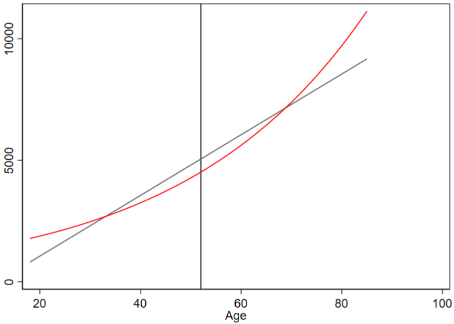
Changing family changes parameters estimates - nonlinearity

- One assumes a linear relationship, GLM with gamma and log link assumes a nonlinear relationship in the dollar scale
- Note that the marginal effects at age = 52 are nearly identical, as shown by the slope of the curve

```
qui reg exp_tot age if exp_tot > 0
predict double yhatols if e(sampple)
margins, dydx(age) at(age=53)
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   124.8428   4.670387    26.73   0.000     115.6883    133.9973
------------------------------------------------------------------------------

qui glm exp_tot age if exp_tot > 0, family(gamma) link(log) nolog
predict double yhatglmgamma if e(sample)
margins, dydx(age) at(age=52)
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   123.5738   6.333483    19.51   0.000     111.1604    135.9872
------------------------------------------------------------------------------

line yhatols age, sort || line yhatglmgamma age, color(red) sort xline(52) ///
     legend(off)
graph export g1.png, replace
```

Nonlinearity

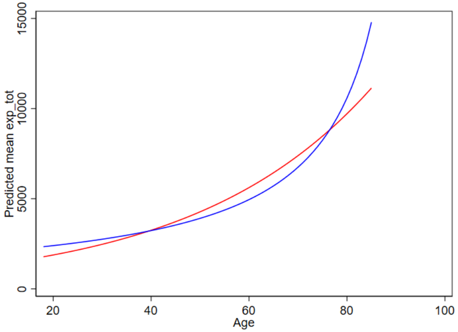
- Same as in logistic models. In the probability scale, nonlinear even if we enter age as linear in the model. Here, nonlinear in the dollar scale



Nonlinearity

- Without the log link, even more nonlinear, that's why marginal effects were so different (remember, marginal effects are averages). Blue line is the GLM with Gamma but identity link

```
qui glm exp_tot age if exp_tot > 0, family(gamma) /*link(log)*/ nolog
predict double yhatnolog if e(sample)
line yhatnolog age, sort

line yhatglmgamma age, color(red) sort ///
     legend(off) || line yhatnolog age, sort color(blue)
graph export g2.png, replace
```



Changing family changes parameters estimates - nonlinearity

- Two variables interacting in a nonlinear ways (no additive, separate effects as in linear/OLS models)

```
qui reg exp_tot i.female age if exp_tot > 0
margins, dydx(*)
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
     Female  |   624.5558   167.0181     3.74   0.000     297.1815    951.9301
         age |   125.0642   4.668862    26.79   0.000     115.9127    134.2157
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
qui glm exp_tot i.female age if exp_tot > 0, family(gamma) /*link(log)*/ nolog
margins, dydx(*)
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
     Female  |    503.465   201.5864     2.50   0.013      108.363     898.567
         age |   142.1625   13.46262    10.56   0.000     115.7763    168.5488
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
qui glm exp_tot i.female age if exp_tot > 0, family(gamma) link(log) nolog
margins, dydx(*)
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
     Female  |   919.9011   189.7723     4.85   0.000     547.9542    1291.848
         age |   126.0328    7.37149    17.10   0.000     111.5849    140.4806
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Choosing family and links

- How do we choose the best family and link in this problem?
- Previous work shows that GLM with Gamma link is a good fitting option but all datasets are different so "previous research" (often using simulations and particular datasets) is not a good guide
- We could compare models using BIC and AIC
- There are some formal tests (using Box-Cox) and the modified Park test (we won't cover them)
- A model with Gamma family but with different links will most likely be the winner

Compare models

- Gamma with log link seems the best fitting one

```
/// Compare models
* Null
qui glm exp_tot, family(gamma) link(log)
est sto nullm
* Log links
qui glm exp_tot age i.female pcs i.race* i.eth_hisp, family(gamma) link(log)
est sto glmgammalog
qui glm exp_tot age i.female pcs i.race* i.eth_hisp, family(gaussian) link(log)
est sto glmgaulog
* Gamma with other link
qui glm exp_tot age i.female pcs i.race* i.eth_hisp, family(gamma) link(power 0.5)
est sto glmgammapower5
* Poisson
qui glm exp_tot age i.female pcs i.race* i.eth_hisp, family(poisson) scale(x2) link(log)
est sto glmpoisoonlog

estimates stats nullm glmgammalog glmgaulog glmgammapower5 glmpoisoonlog
```

```
. estimates stats nullm glmgammalog glmgaulog glmgammapower5 glmpoisoonlog

Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
       nullm |     19,386          .  -178585.6       1    357173.3   357181.2
 glmgammalog |     19,386          .  -170848.9       7    341711.7   341766.8
   glmgaulog |     19,386          .  -204234.6       7    408483.1   408538.2
glmgammapo~5 |     19,386          .  -170864.8       7    341743.6   341798.7
glmpoisoon~g |     19,386          .  -6.06e+07       7    1.21e+08   1.21e+08
-----------------------------------------------------------------------------
```

GLM further topics

- Different combinations of families and links produce different models. Be careful interpreting coefficients since the choice of link affects estimation scale
- Count data? Either Poisson or negative binomial models (log link). Logit model? Binomial family with logit link. Probit? Binomial family with probit link (see next slide)
- One interesting thing about some models is that you can estimate an "offset." The offset has a coefficient constrained to be equal to 1. So we estimate $\eta = \mathbf{X}'\beta + 1$
- Poisson models are used to model count data (length of stay, number of deaths). With number of deaths, for example, we may want to use a denominator, say, population size in each county or hospital to model death rate rather than death counts
- The offset is the population and then the Poisson model is modeling a rate rather than a count:

  $log(deaths\_count) = \beta_0 + 1 \times log(popsize) + \beta_1 mask\_use$

  $log(\frac{deaths\_count}{pop\_size}) = \beta_0 + \beta_1 mask\_use$

GLM combinations and other Stata commands

- From Stata manual
- Note that GLM with Gamma links are used to model survival data (-streg- command). The Weibull model, used in survival analysis, has been proposed to model cost data as well

Some family() and link() combinations result in models already fit by Stata. These are

family()	link()	Options	Equivalent Stata command
gaussian	identity	*nothing* | irls | irls vce(oim)	regress
gaussian	identity	t(*var*) vce(hac nwest #) vfactor($\#_v$)	newey, t(*var*) lag(#) (see note 1)
binomial	cloglog	*nothing* | irls vce(oim)	cloglog (see note 2)
binomial	probit	*nothing* | irls vce(oim)	probit (see note 2)
binomial	logit	*nothing* | irls | irls vce(oim)	logit or logistic (see note 3)
poisson	log	*nothing* | irls | irls vce(oim)	poisson (see note 3)
nbinomial	log	*nothing* | irls vce(oim)	nbreg (see note 4)
gamma	log	scale(1)	streg, dist(exp) nohr (see note 5)

Two-part models

- Although we found the best fitting model with reasonable assumptions about the data generating process (Gamma is more realistic with cost data), we saw that the excess zeroes are a problem in this dataset
- A large proportion of zeroes are most command with inpatient data and with younger people – few people are hospitalized in a given year, so they will have zeroes
- It would be more unusual to find a large proportion of zeroes in outpatient costs in the Medicare population over 65 for example
- There is a class of models that seem odd but can deal with excess zeroes: two-part models
- Simple idea: estimate two models, one for the zeroes and one for the non-zeroes. Predictions are a combination of both models
- (There are other options: Tobit models, mixture models; for count data: Zero Inflated Poisson or ZIP models)

Two-part models

1) Estimate the probability that the cost is greater than zero conditional on covariates: $P(Y_i > 0|\mathbf{X}_i)$. This part could estimated using logit or probit models

2) For those observations with non-zero costs, estimate the expected costs conditional on covariates: $E(Y_i|y_i > 0, \mathbf{X}_i)$. This part can be model with linear models/OLS, GLM, or log-level models

- Predictions are obtained combining both parts (multiplication): $P(Y_i > 0|\mathbf{X}_i) \times E(Y_i|y_i > 0, \mathbf{X}_i)$
- And since we know that marginal effects are predictions, we can interpret models using marginal effects

Two-part model example

- Easy to estimate two-part models. See notes on lecture about marginal effects on how to compute them "by hand"
- Stata has a user-written command, twopm (install it typing: findit twopm)

```
gen     nonzero = 0
replace nonzero = 1 if exp_tot > 0 & exp_tot ~= .

* First part
qui logit nonzero age i.female pcs i.race* i.eth_hisp, nolog
predict double pnonzero

* Second part
qui glm exp_tot age i.female pcs i.race* i.eth_hisp, family(gamma) link(log)
predict double exphat

* Predictions
gen tpmhat = pnonzero * exphat

* Compare with observed
sum exp_tot tpmhat

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     exp_tot |     19,386     3685.25    9768.475          0     440524
      tpmhat |     19,386     3709.48    5250.387   70.83586   70825.74
```

Two-part model example -twopm- command

```
twopm exp_tot age i.female pcs i.race* i.eth_hisp, ///
    firstpart(logit, nolog) secondpart(glm, family(gamma) link(log) nolog)
margins, dydx(*)

Average marginal effects                        Number of obs     =     19,386

Expression   : twopm combined expected values, predict()
dy/dx w.r.t. : age 1.female pcs12 1.race_bl 1.race_oth 1.eth_hisp

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   64.59912   3.903919    16.55   0.000     56.94758    72.25066
             |
      female |
     Female  |   1078.981   112.5954     9.58   0.000     858.2979    1299.664
       pcs12 |  -188.0581   7.781285   -24.17   0.000    -203.3091    -172.807
             |
     race_bl |
 Black race  |  -913.5817   144.9708    -6.30   0.000    -1197.719   -629.4443
             |
    race_oth |
 Other race  |  -1625.517   150.2068   -10.82   0.000    -1919.917   -1331.117
             |
    eth_hisp |
   Hispanic  |  -1863.469    112.566   -16.55   0.000    -2084.095   -1642.844
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```