PBSH 32410 / STAT 22401 Winter 2025 J. Dignam

Exam 2

March 11 10:00-12:00

Details: 2 hrs, page of notes (two-sided)

Important Topics

- **Multiple Linear Regression - General**
  - interpreting βs
  - categorical predictors/indicator variables and continuous variables
  - models with main effects vs. interactions - how different?

- **Multiple Linear Regression - Special Situations**
  - familiarity with model violations - recognize from plots, etc

- transformations on Y-
  - log transforms - what does this mean for β?
  - Box-Cox - what is being evaluated here?
- familiarity with transformation issues - why we use, etc

- **Logistic Regression**
  - relationship of frequency data to the model and odds ratios
  - basic interpretation of model coefficients - log odds ratio and odds ratios
  - calculating a probability of event (1) for a given set of covariate values

- **Poisson Regression**
  - relationship of mean count and rate data to the model and incidence rate ratios
  - basic interpretation of model coefficients - log of mean counts or rate, etc

Data on FEV1 predicted by age at three geographic centers

```
. plot fev1 age
    4.06 +
         |                          *
         |
         |      *    *
         |        *             * *     *               *
         |                            *      * * *                 *
    f    |                     *          *  *                       *
    e    |                            *               *       * *
    v    |        *                     *        * *    *                *
    1    | **                  *    *          *   *  *     *
         |                        *     * **     * *      * *      *     *
         |                                            *              * *  *
         |                   *            *                      *
         |                                         * **         *
         |                                         *                 *
         |                          *
         |                                                      *
    1.79 +                                            *         *
          +----------------------------------------------------------------+
               41                       age                             76
```

Linear Regression Models

```
. by center: sum age fev1
--------------------------------------------------------------------------------
-> center = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        21    62.04762    8.558482         41         76
        fev1 |        21    2.697619     .468614       1.79       3.47
--------------------------------------------------------------------------------
-> center = 2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        21    61.57143    9.452891         44         75
        fev1 |        21    3.226667    .3018664       2.67       3.69
--------------------------------------------------------------------------------
-> center = 3

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        24    60.04167    8.306515         42         73
        fev1 |        24     2.93375    .4365359       2.19       4.06
```

Linear Regression Models

```
. reg fev1 center2 center3

      Source |       SS           df       MS      Number of obs   =        66
-------------+----------------------------------   F(2, 63)        =      8.77
       Model |  2.95117175         2  1.47558588   Prob > F        =    0.0004
    Residual |  10.5974099        63  .168212855   R-squared       =    0.2178
-------------+----------------------------------   Adj R-squared   =    0.1930
       Total |  13.5485816        65  .208439718   Root MSE        =    .41014

------------------------------------------------------------------------------
        fev1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     center2 |    .529    .1265712      4.18   0.000     .2761152      .78198
     center3 |    .236     .122552      1.93   0.059    -.0087698    .4810317
       _cons |   2.697    .0894994     30.14   0.000     2.518769    2.876469
------------------------------------------------------------------------------
```

What does this model say and why does it reproduce means by center?

Linear Regression Models

```
. reg fev1 age center2 center3

      Source |       SS           df       MS      Number of obs   =        66
-------------+----------------------------------   F(3, 62)        =     12.31
       Model |  5.05636517         3  1.68545506   Prob > F        =    0.0000
    Residual |  8.49221647        62  .136971233   R-squared       =    0.3732
-------------+----------------------------------   Adj R-squared   =    0.3429
       Total |  13.5485816        65  .208439718   Root MSE        =     .3701

------------------------------------------------------------------------------
        fev1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0208577   .0053203    -3.92   0.000    -.0314928   -.0102226
     center2 |   .5191154   .1142423     4.54   0.000     .2907483    .7474825
     center3 |   .1942915   .1111012     1.75   0.085    -.0277966    .4163796
       _cons |   3.991788   .3398463    11.75   0.000     3.312445    4.671131
------------------------------------------------------------------------------
```

What is meaning of intercept here?

What is predicted FEV for 45 year old from center 2?

Linear Regression Models - with interaction effects

```
. reg fev1 age center2 center3 agebycent2 agebycent3
. . .
. . .
------------------------------------------------------------------------------
        fev1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0117189   .0097259    -1.20   0.233    -.0311736    .0077358
     center2 |   1.282685   .8193459     1.57   0.123    -.3562513     2.92162
     center3 |   1.052032   .8314677     1.27   0.211    -.6111514    2.715214
  agebycent2 |  -.0123307   .0131199    -0.94   0.351    -.0385744    .0139131
  agebycent3 |  -.0139804   .0134875    -1.04   0.304    -.0409595    .0129987
       _cons |   3.424748   .6089115     5.62   0.000     2.206744    4.642753
------------------------------------------------------------------------------
```

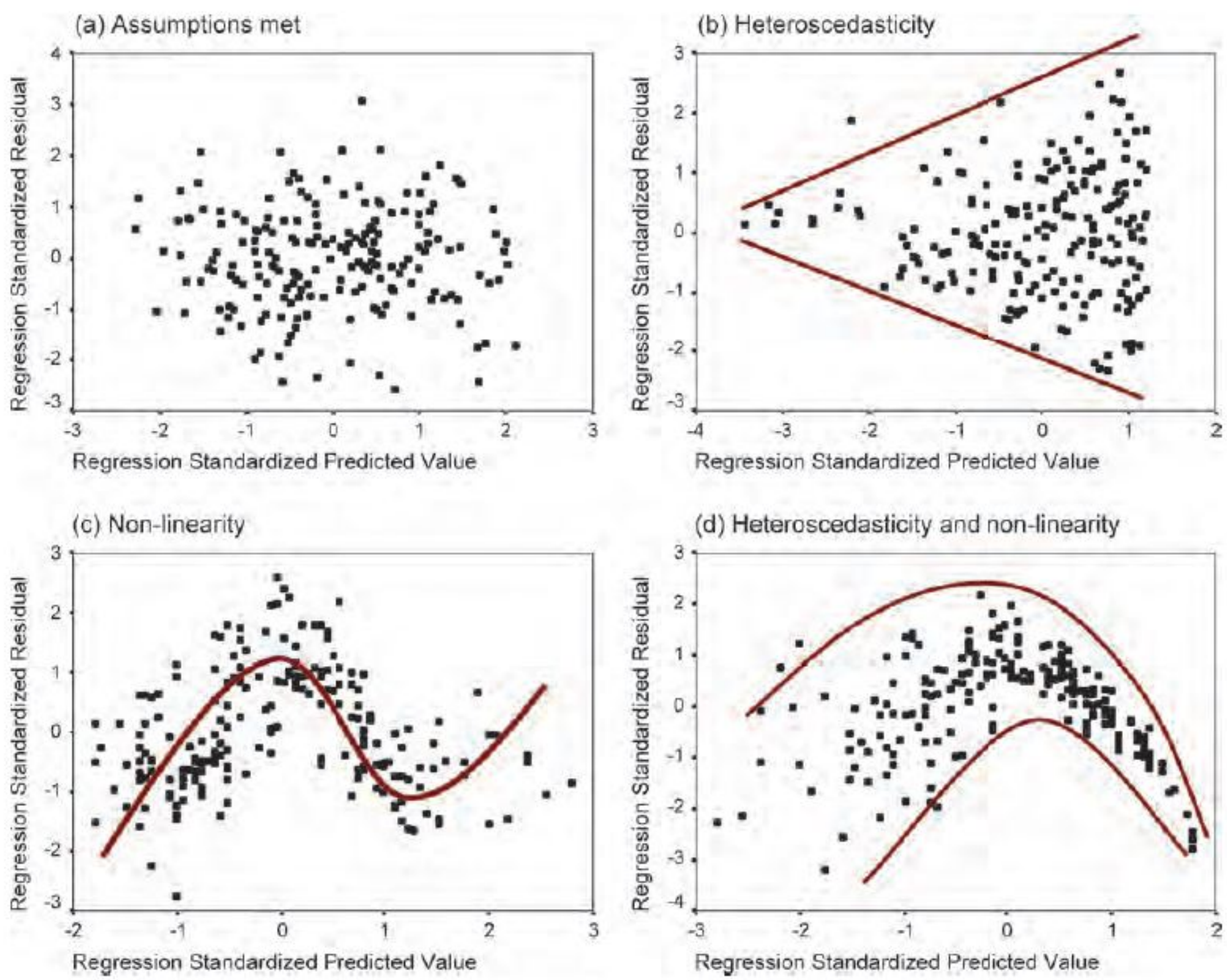
Slope in center 1: -.01172 per year of age

Slope in center 2: -.01172 - .01233 = -.02405 per year of age

Slope in center 2: -.01172 - .01398 = -.02570 per year of age

not statistically different here, based on tests on the interaction coefficients. What is another way to test?

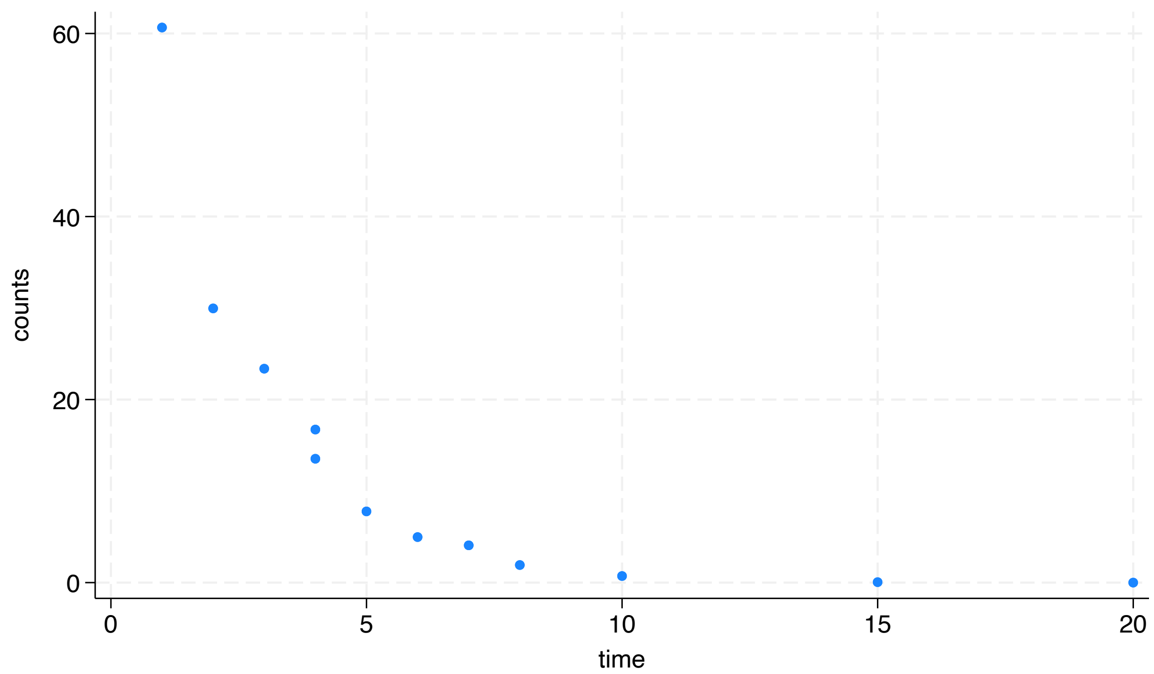
Some Diagnostic plots for OLS



(https://tobeneo.wordpress.com/2013/12/08/multiple-regression-iii/)

Transformations

We want to predict radioactivity counts at different times from contamination.



Is there a transformation that could help?

Transformations

```
. boxcox counts time
Fitting comparison model
. . .
Iteration 4:  Log likelihood = -39.643945

Fitting full model

Iteration 0:  Log likelihood =  -47.45472
. . .
Iteration 7:  Log likelihood = -4.4707423

                                                  Number of obs   =         12
                                                  LR chi2(1)      =      70.35
Log likelihood = -4.4707423                       Prob > chi2     =      0.000

------------------------------------------------------------------------------
      counts | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      /theta |   .0117264   .0096365     1.22   0.224    -.0071608    .0306136
------------------------------------------------------------------------------
Estimates of scale-variant parameters
----------------------------
```

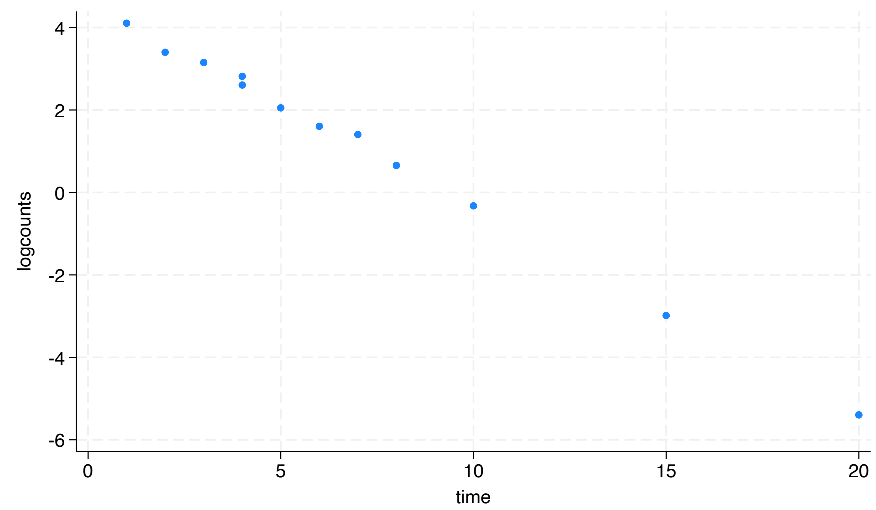
```
             | Coefficient
-------------+--------------
Notrans      |
        time |  -.4975041
       _cons |   4.664043
-------------+--------------
      /sigma |   .1194851
----------------------------
---------------------------------------------------------
   Test         Restricted     LR statistic
    H0:       log likelihood       chi2       Prob > chi2
---------------------------------------------------------
theta = -1      -86.978782       165.02           0.000
theta =  0      -5.1925572         1.44           0.230
theta =  1       -47.45472        85.97           0.000
---------------------------------------------------------
```

What is the suggested new response variable to predict w/time?

$$Y' = (Y^{\theta} - 1)/\theta$$

Since theta not different from zero, natural log transform is suggested. Do data look linear on log(counts) scale?

Transformations




```
. reg logcounts time
. . .
------------------------------------------------------------------------------
   logcounts | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        time |  -.5005141    .007411   -67.54   0.000    -.5170269   -.4840013
       _cons |   4.636288   .0657662    70.50   0.000     4.489752    4.782825
------------------------------------------------------------------------------
```

How to interpret?

Binary Outcome Data and Regression Models

Data on possible sources of food poisoning at a gathering

```
. list, clean

     crabsalad   psalad     n   isick
  1.         0        0    23       0
  2.         0        0     0       1
  3.         0        1    24       0
  4.         0        1    22       1
  5.         1        0     4       1
  6.         1        0    31       0
  7.         1        1    80       0
  8.         1        1   120       1
```

Of those who had neither crab or potato salad, 0 out of 23 sick,

Of those who had potato salad only, 24 were not sick, 24 were sick,

etc

```
. tab isick crabsalad [fweight=n]

           |       crabsalad
     isick |         0          1 |     Total
-----------+----------------------+----------
         0 |        47        111 |       158
         1 |        22        124 |       146
-----------+----------------------+----------
     Total |        69        235 |       304
```

Odds ratio for eating crab salad?

OR = (odds of illness in exposed) / (odds of illness in unexposed)

$$= \frac{124/111}{22/47} = 2.39$$

Binary Outcome Data and Regression Models

Running the logistic model:

```
. logit isick crabsalad [fweight=n]

Iteration 0:   log likelihood = -210.47984
. . .
Logistic regression                               Number of obs     =        304
                                                  LR chi2(1)        =       9.51
                                                  Prob > chi2       =     0.0020
Log likelihood = -205.72331                       Pseudo R2         =     0.0226

------------------------------------------------------------------------------
       isick |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   crabsalad |   .8698565   .2894904     3.00   0.003     .3024658    1.437247
       _cons |  -.7591051   .2583237    -2.94   0.003     -1.26541      -.2528
------------------------------------------------------------------------------
```

What is the odds ratio? $\exp(0.86985) = 2.39$

What is being tested? What is the inferential conclusion?

Binary Outcome Data and Regression Models

```
. logit isick crabsalad psalad [fweight=n]
. . .
Logistic regression                               Number of obs     =        304
                                                  LR chi2(2)        =      60.45
                                                  Prob > chi2       =     0.0000
Log likelihood = -180.25338                       Pseudo R2         =     0.1436
------------------------------------------------------------------------------
       isick |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   crabsalad |   .6097114   .3169859     1.92   0.054    -.0115696    1.230992
      psalad |   2.825864   .5361679     5.27   0.000     1.774994    3.876733
       _cons |  -3.007495   .5675813    -5.30   0.000    -4.119934   -1.895056
------------------------------------------------------------------------------
```

Which food item has greater risk? What is probability of being sick for those who had both?

$$\Pr(sick|both\ items) = \frac{\exp(-3.007 + 0.6097 + 2.8258)}{1 + \exp(-3.007 + 0.6097 + 2.8258)} = 0.6055$$

What is the log odds ratio comparing someone who had crab salad only to someone who had potato salad only?

Write out the difference:

$$(-3.007 + 2.826) - (-3.007 + .6097) = 2.216$$

and $\exp(2.216) = 9.17$

Note that OR for potato salad is $\exp(2.83) = 16.9$ and OR for crab salad is $\exp(.610) = 1.84$

ratio is 9.17 - OR for potato salad vs crab salad

Count Data and Regression Models

Ex: 3 different garden plantings are monitored for monarch butterfly visits. Counts are made over a fixed period:

	Gtype	count	g1	g3
1.	A	0	1	0
2.	A	3	1	0
3.	A	2	1	0
4.	A	2	1	0
5.	A	1	1	0
. . .				
9.	B	5	0	0
10.	B	9	0	0
11.	B	5	0	0
12.	B	5	0	0
13.	B	7	0	0
. . .				
17.	C	8	0	1
18.	C	14	0	1
19.	C	12	0	1
20.	C	12	0	1
21.	C	10	0	1

. . .

```
. by Gtype: sum count

--------------------------------------------------------------------------------------
-> Gtype = A

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       count |          8        1.75     1.28174          0          3

--------------------------------------------------------------------------------------
-> Gtype = B

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       count |          8         6.5    2.329929          5         11

--------------------------------------------------------------------------------------
-> Gtype = C

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       count |          8       11.75    3.284161          8         18
```

Count Data and Regression Models

```
.  poisson count g1 g3

Iteration 0:  Log likelihood = -50.669741
. . .
Poisson regression                                      Number of obs =     24
                                                        LR chi2(2)    =  66.46
                                                        Prob > chi2   = 0.0000
Log likelihood = -50.619718                             Pseudo R2     = 0.3963
------------------------------------------------------------------------------
       count | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
          g1 |  -1.312186   .3010969    -4.36   0.000    -1.902325   -.7220473
          g3 |   .5920511   .1728267     3.43   0.001      .253317    .9307852
       _cons |   1.871802    .138675    13.50   0.000     1.600004      2.1436
------------------------------------------------------------------------------
```

What are these numbers? Output is in $\log_e(counts)$

$\exp(1.871802) = 6.5$ - mean count in baseline (garden 2) group

$\exp(1.871802 - 1.312186) = 1.75$ - mean count in garden 1

$\exp(1.871802 + .5920511) = 11.75$ - mean count in garden 3

Rate Data and Regression Models

Here are skin cancer rates for two cites (Minneapolis (0) and Dallas (1)) and by age (in groups, midpoint of age used)

```
. list, clean

       cases   city    age     pyrs
  1.       1      0   19.5   172675
  2.      16      0   29.5   123065
  3.      30      0   29.5    96216
  4.      71      0   49.5    92051
  5.     102      0   59.5    72159
  6.     130      0   69.5    54722
  7.     133      0   79.5    32185
  8.      40      0   89.5     8328
  9.       4      1   19.5   181343
 10.      38      1   29.5   146207
 11.     119      1   39.5   121374
 12.     221      1   49.5   111353
 13.     259      1   59.5    83004
 14.     310      1   69.5    55932
 15.     226      1   79.5    29007
 16.      65      1   89.5     7503
```

Rate Data and Regression Models

Run the rate table by city:

```
. ir cases city pyrs

Incidence-rate comparison

                 |         city          |
                 |   Exposed   Unexposed |      Total
-----------------+-----------------------+-----------
           cases |      1242         523 |       1765
            pyrs |    735723      651401 |    1387124
-----------------+-----------------------+-----------
                 |                       |
  Incidence rate |  .0016881    .0008029 |   .0012724
                 |                       |
                 |      Point estimate   |    [95% conf. interval]
                 |-----------------------+-----------------------
 Inc. rate diff. |         .0008853      |    .0007688    .0010017
 Inc. rate ratio |         2.102587      |    1.896843    2.333224 (exact)
```

What is the interpretation?

Rate Data and Regression Models

Run the model with city as predictor

```
 poisson cases city, exposure(pyrs)
. . .
Poisson regression
------------------------------------------------------------------------------
       cases | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
        city |   .7431685   .0521268    14.26   0.000      .641002    .8453351
       _cons |  -7.127299   .0437269  -163.00   0.000    -7.213002   -7.041596
    ln(pyrs) |          1  (exposure)
------------------------------------------------------------------------------
```

Interpretation: Output in (natural) log(incidence rate) for baseline and log increase/decrease for covariate increment

$\exp(-7.127299) = 0.000803$ - incidence rate in Minn.

$\exp(-7.127299 + 0.7431685) = 0.00169$ - incidence rate in Dallas.

Incidence rate ratio is $0.00169/0.00080 = 2.10$ - same as that obtained from the incidence table