NAME: ______________________________

PBHS 32410 /Stat 22401
Regression Analysis for Health and Social Research
FINAL EXAM –Winter 2024

INSTRUCTIONS: You have 2.5 hours to work on this exam (not counting download/printing/upload time). Please write your name on the top of the first page, work alone and use only permitted materials, and legibly show any relevant computations.

1. **Linear Regression Concepts: Multiple Choice.**

   (a) (3pt) (Circle all that apply.) In which of the following settings would you perform variable transformations to avoid violations of ordinary least squares (OLS) regression assumptions, and then proceed to perform OLS analysis (i.e., linear regression) on the transformed data?

   i. The variance of the error term increases as one of the predictors X increases.
   ii. One or more of the predictors is not normally distributed.
   iii. The response variable is a binary (0/1) variable.
   iv. There are potential confounders not included in the model.
   v. Response and predictor have some relationship, but it is not linear in the observed 'raw' data.

   (b) (3pt) (Circle all that apply.) Which of the following are true for variable selection in modeling?

   i. Variable selection methods are helpful if the number of predictors for the problem is large.
   ii. Variable selection algorithms/methods should not overrule scientific reasoning.
   iii. Forward selection and backward selection will always yield the same model.
   iv. For 5 predictors, there are as many as 32 main effects models to consider
   v. A stepwise procedure will yield the best model

   (c) (3pt) (Circle all that apply.) Suppose one has a continuous-scale Y variable that appears to be related to X via $Y = k\exp(\beta X)$. What approach(es) might be taken to for using X to model Y?

   i. Estimate directly via OLS of X predicting Y, since such a model is linear with respect to β
   ii. Apply a logarithmic transformation to Y
   iii. Fit a logistic regression model
   iv. Apply a logarithmic transformation to both sides of the equation
   v. This model is beyond the scope of any linear regression approach, so other methods (non-linear regression, etc) are needed

   (d) (3pt) (Circle all that apply.) Interaction effects in multiple regression

   i. Are needed to permit different intercepts depending on the value of a category variable
   ii. Generally require larger sample sizes (more observations) to detect reliably
   iii. Are needed to permit different slopes depending on combinations of predictor variable values
   iv. Are not included in degrees of freedom calculations since they are formed from variables already in the model

v. Expand the flexibility of linear models to deal with so-called effect modification

2. **Linear Regression Analysis: Model Fitting and Interpretation**

It is established that smoking impairs lung function. Much of the supportive data arises from studies of pulmonary function in adults who are long-time smokers. Whether such deleterious effects of smoking can be detected in younger individuals (adolescents) who smoke is unknown. To address this question, measures of lung function were made in 654 adolescents seen for a routine check up in a particular pediatric clinic. It was ascertained among those participating whether they were smokers.

A common measurement of lung function is the forced expiratory volume (FEV), which measures how much air (in liters) one can expel from the lungs in a short period of time. A higher FEV is usually associated with better respiratory function.

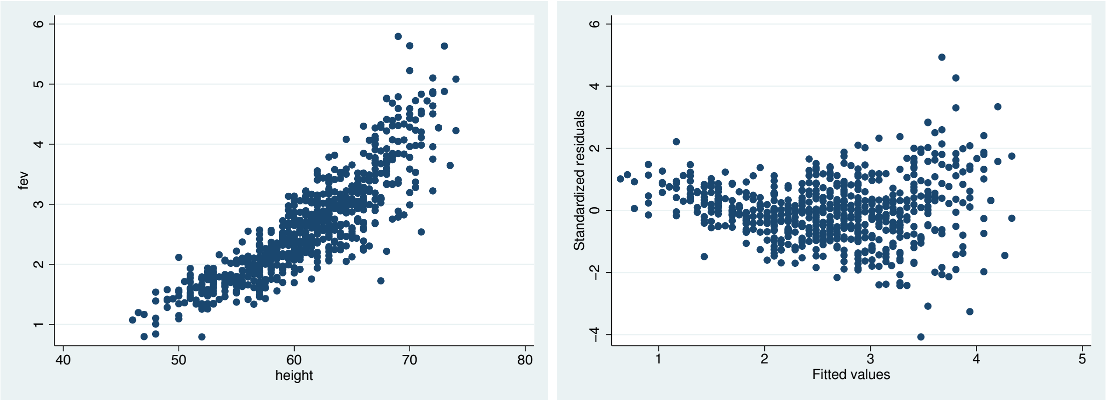
Since FEV is a volume measure, how much air one can expel may depend on height (directly and as a surrogate for other development measures). In the following problem, we employ a SLR with height (in inches) as predictor and FEV as response, we obtain the following model:

```
. reg fev height

      Source |       SS       df       MS              Number of obs =     654
-------------+------------------------------           F(  1,   652) = 1994.73
       Model |  369.985854     1  369.985854           Prob > F      =  0.0000
    Residual |  120.933979   652  .185481563           R-squared     =  0.7537
-------------+------------------------------           Adj R-squared =  0.7533
       Total |  490.919833   653  .751791475           Root MSE      =  .43068
------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |    0.131    .002955         *    *        .1261732     .137778
       _cons |   -5.432   .1814599   -29.94   0.000    -5.788995   -5.076363
------------------------------------------------------------------------------
```

(a) (2pt) Write out the fitted model (i.e., with the estimates) and interpret the model coefficients

(b) (2pt) Set up the test of whether the coefficient for height differs from zero at the $\alpha = .05$ level. Write out the null and alternative hypothesis and provide the t-statistic.



(a) FEV vs height

(b) Residual plot

(c) (2pt) What is the estimated FEV for height = 55 inches?

(d) (4pt) Based on the figures above, what OLS model assumption(s) may be violated?

(e) (4 pts) The following Box-Cox analysis was done to assess a possible transformation of the response variable to fix one of the above problems in the plots. The estimated theta (what we call λ in our notation) below is the relevant parameter for the Box-Cox transformation.

```
 boxcox fev height   . . .
 ------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      /theta |   .0856867   .0635592     1.35   0.178     -.038887    .2102604
------------------------------------------------------------------------------
```

From this result, what if any transformation is suggested? What is the conclusion based on?

(f) (4pt) Suppose one did re-fit the model, using a logarithmic transform of FEV, and thus would obtain a new set of $\hat{\beta}_0$ and $\hat{\beta}_1$. How would you interpret your new estimated slope coefficient.

3. (12 pts) It would certainly make sense in this analysis to adjust for other factors that possibly influence FEV. Below is a model taking sex (0=male, 1=female) into account and including a model term for interaction between height and sex as predictors

```
------------------------------------------------------------------------------
          fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
       height |   0.167  .0087714    19.08   0.000     .1501165    .1845641
          sex |   1.545   .3738427     4.13   0.000      .8115443    2.279714
height_by_sex |  -0.027   .0061193    -4.49   0.000    -.0394731   -.0154412
        _cons |  -7.409   .5415434   -13.68   0.000    -8.472863   -6.346092
------------------------------------------------------------------------------
```

(a) (3pt) From this model, whose slope for height is greater (males or females)?

(b) (3pt) Which group (males or females) has the larger intercept?

(c) (3pt) Based on the model results above, is including the interaction term important in modeling these data?

(d) (3pt) Predict FEV for males of height 51 inches and females of height 54 inches.

4. **Logistic Regression Analysis: Model Fitting and Interpretation**

In the Breast Cancer Screening Consortium (BCSC) study, factors related to positive mammography finding were evaluated in a large (> 180,000) cohort of women 35-85+. The following analyses investigate several key potential risk factors for breast cancer (outcome *inv* for invasive tumor), including age at screening (in 5-year age bins), breast density (ordinal score 1-4), hormone replacement therapy use (no/yes), and body mass index (4 ordinal categories) on a subset of the data.

Note: model output is on log odds ratio scale.

first run: age as a predictor

```
. logit inv age

Iteration 0:   log likelihood = -5995.3571
. . .
Iteration 3:   log likelihood = -5976.4159

Logistic regression                             Number of obs     =     58,558
                                                LR chi2(1)        =      37.88
                                                Prob > chi2       =     0.0000
Log likelihood = -5976.4159                     Pseudo R2         =     0.0032

------------------------------------------------------------------------------
         inv |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   0.084     .013594      6.18   0.000      .0574049    .1106922
       _cons |  -4.353    .0907415   -47.97   0.000     -4.531006   -4.175306
------------------------------------------------------------------------------
```

(a) (2pt) Interpret the odds ratio for age (age is coded numerically here (3,4,5, etc) in 5-year increments starting at age 45 for this subset).

(b) (4pt) compute the probability of a positive finding for age group 3 and for age group 10 (note: to compute, plug in age as coded, not actual ages)

model with more predictors

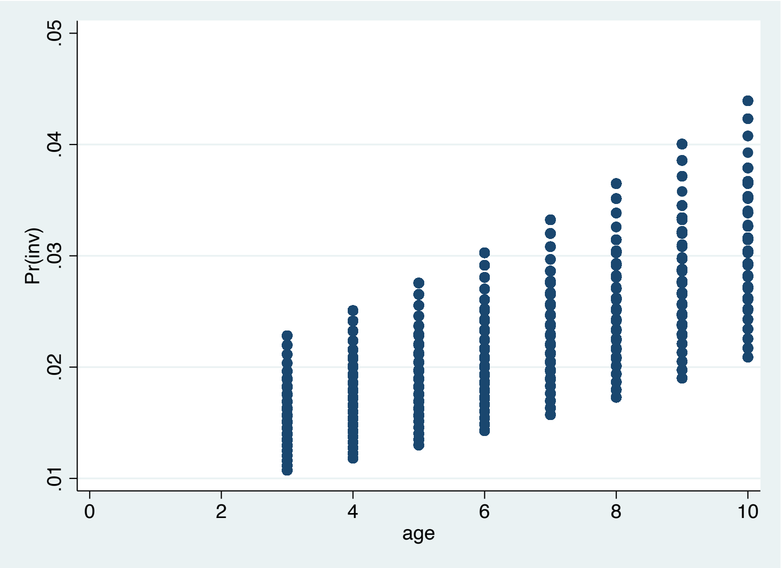
```
Logistic regression                             Number of obs     =     58,558
                                                LR chi2(4)        =      77.78
                                                Prob > chi2       =     0.0000
Log likelihood =  -5956.466                     Pseudo R2         =     0.0065
------------------------------------------------------------------------------
         inv |    Coef.   Std. Err.     z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .096   .0139421     6.93   0.000     .0693189    .1239709
         hrt |    .187   .0587141     3.19   0.001     .0724255    .3025805
     density |    .154   .0346053     4.45   0.000     .0862983    .2219488
         bmi |   -.039   .0298069    -1.31   0.191    -.0973909      .01945
       _cons |  -4.812   .1624104   -29.63   0.000    -5.130055   -4.493418
------------------------------------------------------------------------------
```

(c) (4pt) Interpret the odds ratio for hormone replacement therapy (variable hrt, coded 0,1). Write down the null and alternative hypothesis in terms of the odds ratio and the test result from the information given in the table above. How you would advise in common language the risk for a woman considering HRT?

(d) (4pt) Density is coded in 4 integer categories from 1 (least dense) to 4 (most dense). Based on the coefficient, what is the general relationship between density and breast cancer risk? What is the odds ratio contrasting individuals with density category 3 vs. density category 1?

(e) 4pt) Predict the probability of breast cancer for an individual in age category 5, hrt use, density level 2, and bmi category 2

(f) The following are the predicted probabilities from the model fit and data set, plotted by age group



(g) (4pt) There are seven age groups, with predicted probabilities within each. What is the maximum number of distinct predicted probabilities that can be produced within each age group (hint: this is driven by unique covariate combinations; also, not all combinations appear in the plot - only those represented in the dataset, so you can't just count the dots)?

(h) (5pt) What would be the covariate values (including the age code) for the smallest probability this model can possibly predict? For the largest probability?

5. **Ecology Modeling:** Data was collected on female horseshoe crabs and factors related to mate attraction. For a nesting female, the number of male crabs nearby (called satellites here, the outcome variable) were recorded, along with characteristics of the female (the predictors) including the following: color (ordinal categories), weight (continuous) and protective spine condition (3 ordered categories).

A model evaluating the spine variable as a predictor is fit. Here it is coded as categorical, with category 1 as the reference:

```
. poisson Sat sp2 sp3

Iteration 0:   log likelihood = -119.21797
Iteration 1:   log likelihood = -119.21776
Iteration 2:   log likelihood = -119.21776

Poisson regression                              Number of obs     =         45
                                                LR chi2(2)        =      25.16
                                                Prob > chi2       =     0.0000
Log likelihood = -119.21776                     Pseudo R2         =     0.0955
------------------------------------------------------------------------------
         Sat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          sp2 |  -1.218   .3961586    -3.07   0.002    -1.994614   -.4417008
          sp3 |  -0.803    .181528    -4.42   0.000    -1.159002   -.4474252
       _cons |   1.554  .1186782    13.10   0.000     1.322025    1.787235
------------------------------------------------------------------------------
```

(a) (4 pts) For each level of spine condition (reference, level 2, level 3), calculate the estimated mean number of male crabs hanging around.

A second model uses color (C) as an ordinal (i.e.,numeric) scale:

```
.. poisson Sat sp2 sp3  C
Iteration 0:   log likelihood = -113.53853
. . ;
Poisson regression                                Number of obs     =         45
                                                  LR chi2(3)        =      36.52
                                                  Prob > chi2       =     0.0000
Log likelihood = -113.53768                       Pseudo R2         =     0.1386
------------------------------------------------------------------------------
         Sat |    Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          sp2 |  -.971   .4006879    -2.42   0.015    -1.756578   -.1859109
          sp3 |  -.403   .2133987    -1.89   0.058    -.8220117    .0144957
            C |  -.410   .1248442    -3.29   0.001    -.6549518   -.1655717
        _cons |  2.619   .3370354     7.77   0.000     1.958942    3.280096
------------------------------------------------------------------------------
```

(b) (2 pts) What is the gain/loss in the mean number of males for each unit increase in the female crab's color score?

(c) (4 pts) What is the predicted mean number of males expected for a female crab of color 4 and spine condition category 3?

6. **Lifestyle factors associated with heart disease:** Two factors putatively associated with heart disease are cigarette smoking and alcohol consumption. Individuals were grouped into those engaging or not engaging in these activities, binned according to what might reasonably be considered sufficient use for possible risk elevation (i.e., exposed or not). Coronary heart disease diagnosis was recorded.

   (a) (4 pts) Here is a cross classification of CHD status with alcohol exposure.

	Alcohol Use		**Total**
Disease Status	yes	no	
CHD case	68	32	100
non-case	33	72	105
Total	101	104	205

   For CHD. what is the odds ratio associated with alcohol exposure vs not?

   (b) (4 pts) For the model executed below, check the OR from the table calculation. What is the inferential conclusion?

```
Logistic regression for grouped data            Number of obs     =        205
                                                LR chi2(1)        =      28.05
                                                Prob > chi2       =     0.0000
Log likelihood = -128.00904                     Pseudo R2         =     0.0987
------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         alc |    1.533   .3002462     5.11   ****     .9454585    2.122402
       _cons |   -0.810   .2124591    -3.82   0.000    -1.227342   -.3945179
------------------------------------------------------------------------------
```

   (c) (4 pts) Here are model parameters after adding smoking status (yes/no) .

```
. . .
------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         alc |   0.108   .4247293     0.26   0.798    -.7234958    .9414126
       smoke |   2.712   .4256251     6.37   0.000     1.878112    3.546532
       _cons |  -1.387     .25717    -5.39   0.000    -1.891143    -.883055
------------------------------------------------------------------------------
```

What are the odds ratios for smoking and for alcohol use now? What is the baseline odds of chd?

(d) (5 pts) Here is a cross classification of alcohol use and smoking. Does this offer an explanation for why the multivariable model produced the risks that it did?

```
|-----------------|
|    frequency    |
| row percentage  |
| cell percentage |
+-----------------+
           |        smoke
       alc |         0          1 |     Total
-----------+----------------------+----------
         0 |        85         19 |       104
           |     81.73      18.27 |    100.00
           |     41.46       9.27 |     50.73
-----------+----------------------+----------
         1 |        23         78 |       101
           |     22.77      77.23 |    100.00
           |     11.22      38.05 |     49.27
-----------+----------------------+----------
     Total |       108         97 |       205
           |     52.68      47.32 |    100.00
           |     52.68      47.32 |    100.00

          Pearson chi2(1) statistic =  71.4506
          Test of Ho: no association:  Pr <  0.0001
```

7. Skin cancer events and total exposure time were recorded for two cites (Minneapolis (0) and Dallas (1)) and by age (in groups, here treated as a numerical variable). The following model was estimated to relate location and age to incidence rates.

```
. poisson cases city age , exposure(pyrs)

Iteration 0:  Log likelihood = -144.07583
Iteration 1:  Log likelihood = -144.01259
Iteration 2:  Log likelihood = -144.01258
```

```
Poisson regression                                      Number of obs =      16
                                                        LR chi2(2)    = 2597.08
                                                        Prob > chi2   =  0.0000
Log likelihood = -144.01258                             Pseudo R2     =  0.9002
-------------------------------------------------------------------------------
       cases | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+-----------------------------------------------------------------
        city |   .7887091   .0522059    15.11   0.000     .6863875    .8910307
         age |   .0592264   .0012993    45.58   0.000     .0566798    .0617729
       _cons |  -10.23525   .0949384  -107.81   0.000    -10.42133   -10.04918
    ln(pyrs) |          1  (exposure)
-------------------------------------------------------------------------------
```

(a) (2 pts) From this model, what is the relative increase in the rate of skin cancer per unit of age?

(b) (2 pts) What is the relative increase in skin cancer incidence for Dallas over Minneapolis?