22401 HW5

Bin Yu

Feb 24, 2025

Question 1

(a)

A scatter plot with a fitted regression line and a simple regression analysis were performed using the following STATA commands:

```
rename FEV fev
twoway (scatter fev age) (lfit fev age)
regress fev age
```

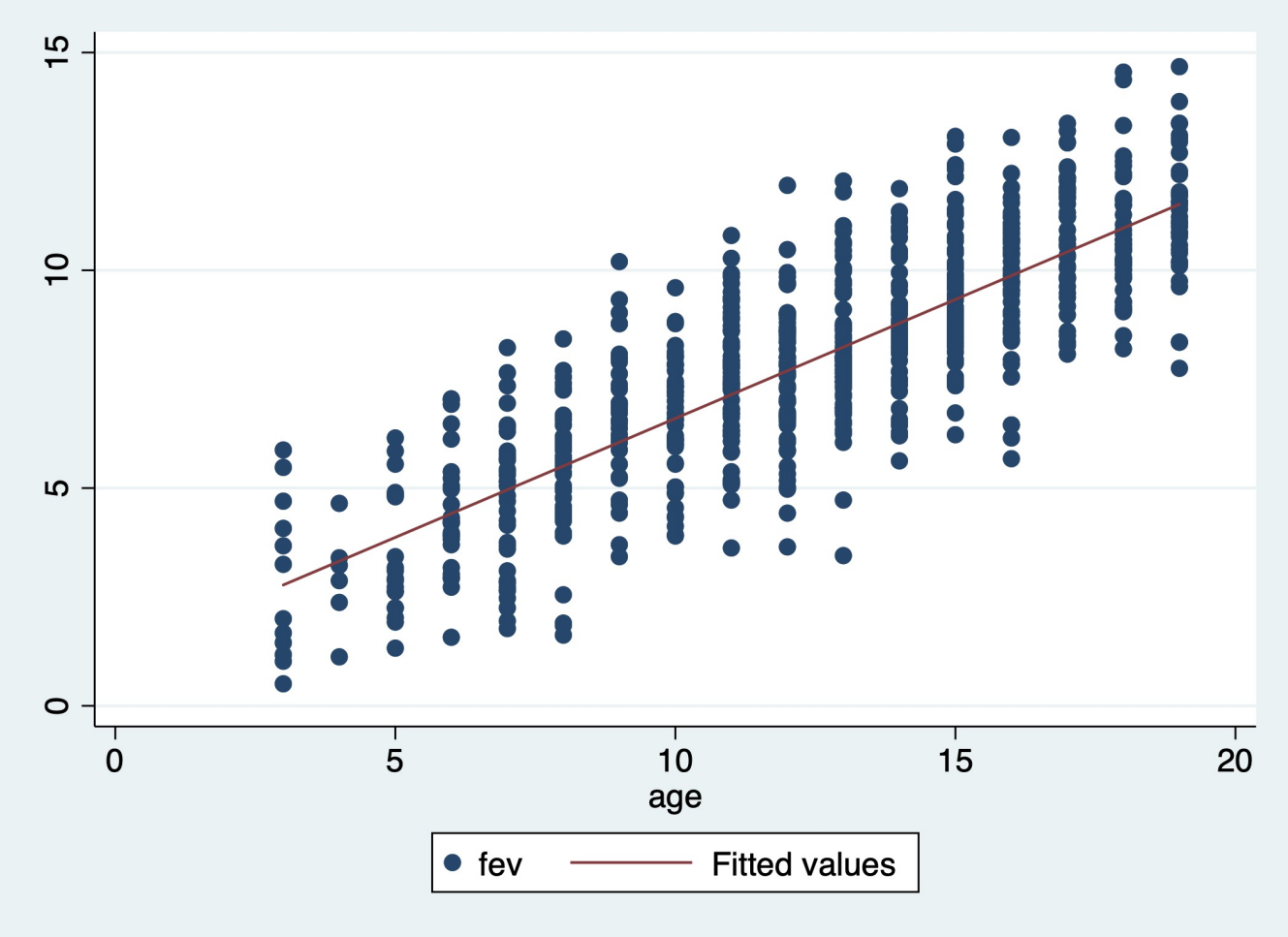

Figure 1: Scatter Plot for age and fev

Regression Output:

```
      Source |       SS           df       MS      Number of obs   = 708
-------------+----------------------------------   F(1, 706)       = 1446.11
       Model |  3366.09117         1  3366.09117   Prob > F        = 0.0000
    Residual |   1643.3436       706  2.32768216   R-squared       = 0.6720
-------------+----------------------------------   Adj R-squared   = 0.6715
       Total |  5009.43478       707  7.08548059   Root MSE        = 1.5257

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   0.5463371   0.0143668    38.03   0.000     0.5181304    0.5745439
       _cons |    1.13675   0.1861209     6.11   0.000     0.7713338    1.502167
------------------------------------------------------------------------------
```

Interpretation: The coefficient on `age` is approximately 0.5463, meaning that, on average, FEV increases by about 0.55 units for every additional year in age, and the effect is siginificant. With an R^2 of 0.6720, age explains about 67% of the variation in FEV among these children.

(b)

A box plot and a simple regression were used to examine the association between smoking status and FEV:

```
graph box fev, over(nsmoke)
regress fev nsmoke
```

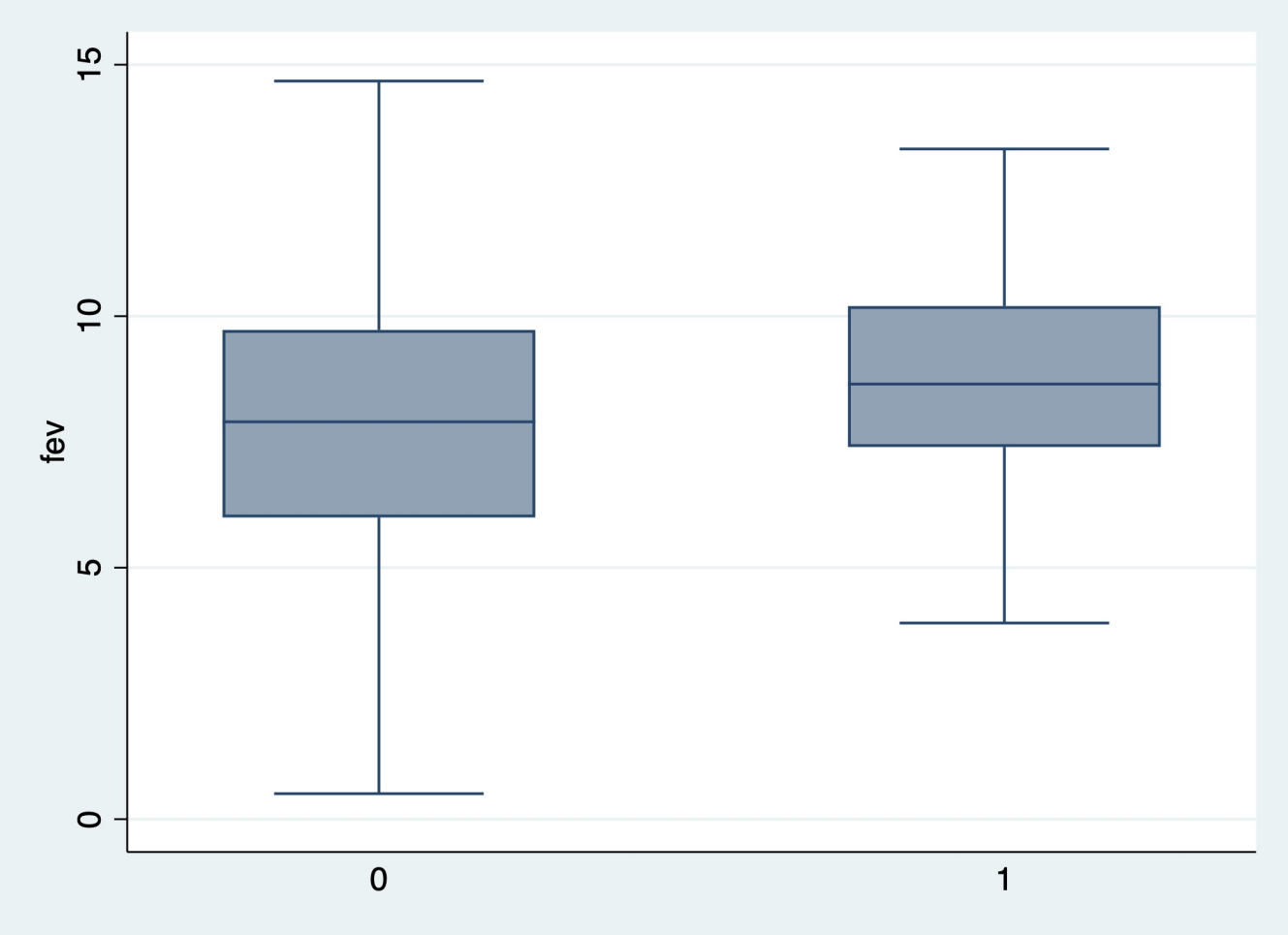

Figure 2: Box plot for smoking status and fev

Regression Output:

```
      Source |       SS           df       MS      Number of obs   = 708
-------------+----------------------------------   F(1, 706)       = 10.70
       Model |  74.7716509         1  74.7716509   Prob > F        = 0.0011
    Residual |  4934.66312       706  6.98960783   R-squared       = 0.0149
-------------+----------------------------------   Adj R-squared   = 0.0135
       Total |  5009.43478       707  7.08548059   Root MSE        = 2.6438

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      nsmoke |   1.010197   0.3088616     3.27   0.001     0.4037997    1.616594
       _cons |   7.751851   0.1057515    73.30   0.000     7.544226    7.959476
```

```
------------------------------------------------------------------------------
```

Interpretation: The simple regression shows a positive coefficient (approximately 1.0102) for `nsmoke`, indicating that children with `nsmoke` = 1 have, on average, a FEV that is about 1.01 units higher than those with `nsmoke` = 0, and the effect is significant. The R^2 is 0.0149, only explains about 1.5% of the variation in FEV among these children. This result is counterintuitive, as smoking is expected to reduce lung function. The likely explanation is confounding by age—if older children (who naturally have higher FEV) are more likely to smoke, the unadjusted analysis will misrepresent the true detrimental effect of smoking.

(c)

To control for confounding, both age and `nsmoke` were included in a multiple regression model:

```
regress fev age nsmoke
```

Regression Output:

```
      Source |       SS           df       MS      Number of obs   = 708
-------------+----------------------------------   F(2, 705)       = 737.47
       Model |  3389.36437         2  1694.68218   Prob > F        = 0.0000
    Residual |  1620.07041       705  2.29797221   R-squared       = 0.6766
-------------+----------------------------------   Adj R-squared   = 0.6757
       Total |  5009.43478       707  7.08548059   Root MSE        = 1.5159

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   0.5570716   0.0146679    37.98   0.000     0.5282736    0.5858696
      nsmoke |  -0.5791136   0.1819735    -3.18   0.002    -0.9363885   -0.2218387
       _cons |    1.07234   0.1860335     5.76   0.000     0.7070936    1.437586
------------------------------------------------------------------------------
```

Interpretation:

Observation: After controlling for age in the regression model, both age and the smoking indicator (`nsmoke`) remain statistically significant. Howeveer, the coefficient for `nsmoke` reverses its sign and becomes approximately -0.5791. This indicates that, when holding age constant, children who smoke have, on average, a forced expiratory volume (FEV) that is about 0.58 units lower than that of non-smokers on average. Meanwhile, the coefficient for age increases slightly (to approximately 0.5571), reinforcing the strong positive effect of age on FEV.

Explanation: A confounder is a variable that is related to both the independent variable (in this case, smoking status) and the dependent variable (FEV), which can lead to a misleading association if not properly controlled. In our analysis, age acts as a confounder because:

- **Age and FEV:** As children grow older, their lung capacity increases, resulting in a higher FEV.
- **Age and Smoking:** Older children are more likely to smoke compared to younger children.

Without adjusting for age, the simple regression of FEV on smoking (`nsmoke`) yielded a positive coefficient, falsely suggesting that smoking is associated with higher FEV. This is because the group of smokers includes more older children, whose naturally higher FEV levels drive the overall association.

By including age as a control variable in the multiple regression, we remove the confounding influence of age. This adjustment allows us to isolate the direct effect of smoking on FEV. The resulting negative coefficient for `nsmoke`

(approximately -0.5791) then reflects the true impact of smoking on lung function when comparing children of the same age.

(d)

To further explore the effect of age on FEV within each smoking group, separate regressions were performed.

For Non-Smokers ($nsmoke = 0$):

```
regress fev age if nsmoke==0
```

Output:

```
      Source |       SS           df       MS      Number of obs   = 625
-------------+----------------------------------   F(1, 623)       = 1351.13
       Model |  3186.05835         1  3186.05835   Prob > F        = 0.0000
    Residual |  1469.08182       623  2.35807676   R-squared       = 0.6844
-------------+----------------------------------   Adj R-squared   = 0.6839
       Total |  4655.14017       624  7.46016053   Root MSE        = 1.5356

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   0.5613188   0.0152708    36.76   0.000     0.5313303    0.5913072
       _cons |   1.021415   0.1931311     5.29   0.000     0.6421478    1.400681
------------------------------------------------------------------------------
```

For Smokers ($nsmoke = 1$):

```
regress fev age if nsmoke==1
```

Output:

```
      Source |       SS           df       MS      Number of obs   = 83
-------------+----------------------------------   F(1, 81)        = 72.43
       Model |  131.958525         1  131.958525   Prob > F        = 0.0000
    Residual |  147.564426        81  1.82178304   R-squared       = 0.4721
-------------+----------------------------------   Adj R-squared   = 0.4656
       Total |  279.522951        82  3.40881648   Root MSE        = 1.3497

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   0.4815888   0.0565856     8.51   0.000     0.3690012    0.5941764
       _cons |   1.613646   0.8528875     1.89   0.062    -0.0833329    3.310624
------------------------------------------------------------------------------
```

Interpretation:

- **Non-Smokers:** The estimated coefficient for `age` is approximately 0.5613. This means that, among non-smokers, for each additional year of age, the FEV increases by about 0.5613 units. It is a significant effect (p = 0.000). Furthermore, the R^2 value of 0.6844 indicates that age explains roughly 68.4% of the variation in FEV within the non-smoking group.

- **Smokers:** For smokers, the estimated coefficient for `age` is approximately 0.4816, indicating that each additional year of age is associated with an increase in FEV of about 0.4816 units on average. Compared to non-smokers, the effect of age is slightly lower among smokers. This effect is also significant. and R^2 value of 0.4721 suggests that age explains less of the variation in FEV for smokers, only 47.21%.

Interaction Between Age and Smoking To formally test whether the effect of age on FEV differs by smoking status, an interaction term was included:

```
regress fev c.age##i.nsmoke
```

Output:

```
      Source |       SS           df       MS      Number of obs   = 708
-------------+----------------------------------   F(3, 704)       = 492.49
       Model |  3392.78853         3  1130.92951   Prob > F        = 0.0000
    Residual |  1616.64625       704  2.29637251   R-squared       = 0.6773
-------------+----------------------------------   Adj R-squared   = 0.6759
       Total |  5009.43478       707  7.08548059   Root MSE        = 1.5154

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   0.5613188   0.0150697    37.25   0.000     0.5317319    0.5909057
    1.nsmoke |    0.592231   0.9763394     0.61   0.544    -1.324655    2.509117
             |
nsmoke#c.age|
           1  |   -0.07973   0.0652928    -1.22   0.222    -0.207922    0.048462
             |
       _cons |   1.021415   0.1905875     5.36   0.000     0.6472266    1.395602
------------------------------------------------------------------------------
```

F-test

```
. test 1.nsmoke#c.age

 ( 1)  1.nsmoke#c.age = 0

       F(  1,   704) =    1.49
            Prob > F =    0.2225
```

Interpretation:

Confounding vs. Effect Modification: Confounding: Confounding is a bias that we hope to prevent or control. It occurs when an extraneous variable is associated with both the exposure and the outcome, making it appear that there is an association between the two when, in fact, the relationship is misleading. In our study, age is a confounder because it is related to both smoking and FEV. Older children tend to have higher FEV and are also more likely to smoke. If we do not control for age, the simple regression of FEV on smoking may misleadingly suggest that smoking is associated with higher FEV, when the apparent effect is actually due to the influence of age.

Effect Modification: Effect modification, in contrast, is a real phenomenon where the effect of one variable (here, age) on the outcome (FEV) differs depending on the level of another variable (smoking status). In other words, effect modification examines whether the slope of the relationship between age and FEV changes across different

groups, such as smokers versus non-smokers. An interaction term in the regression model tests this hypothesis by determining if the effect of age on FEV is statistically significantly different between the groups.

Comparison in Our Analysis:

- In the confounding analysis (see part (c)), after controlling for age, the coefficient for the smoking indicator reversed from positive to negative. This change revealed that, once the confounding effect of age is removed, smoking is associated with a lower FEV.
- In the effect modification analysis, we included an interaction term between age and smoking:

```
regress fev c.age##i.nsmoke
test 1.nsmoke#c.age
 ( 1)  1.nsmoke#c.age = 0

       F(  1,   704) =    1.49
            Prob > F =    0.2225
```

The null hypothesis tested by the interaction term is that the slope of age (its effect on FEV) is the same for smokers and non-smokers.

Our test yielded an F-statistic with a p-value of 0.2225, indicating that there is no statistically significant difference in the effect of age on FEV between the two groups. Thus, while confounding was evident in the simple models, there is little evidence for effect modification, and the relationship between age and FEV (the slope) is similar regardless of smoking status.

Question 2

(a)

We first examine summary statistics for the infection counts overall and by net use (netting), where netting is coded as 0 (no net use) and 1 (net use).

The summary statistics for `infect` are as follows:

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      infect |         50        5.96    3.446145          0         15
```

To compute the overall variance, we square the standard deviation:

$$\text{Variance} = (3.446145)^2 \approx 11.88.$$

Under the Poisson model, the mean and the variance are assumed to be equal (i.e., equidispersion). In this case, the mean is 5.96, whereas the variance is approximately 11.88. Since the variance is roughly twice the mean, there is evidence of overdispersion relative to the Poisson assumption. Therefore, although the Poisson model can be fitted, the observed overdispersion suggests that the model may not be the best fit for these data. If the overdispersion is significant, alternative models such as the Negative Binomial regression should be considered.

Then calculate the mean and variance in different net use groups:

.

```
. * Summary statistics stratified by net use (netting: 0 = no, 1 = yes)
. by netting, sort: summarize infect

-------------------------------------------------------------------------------------------------------------
-> netting = 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      infect |         32      7.5625    3.110207          2         15

-------------------------------------------------------------------------------------------------------------
-> netting = 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      infect |         18    3.111111    1.778595          0          7

. tabstat infect, statistics(mean variance N) by(netting)

Summary for variables: infect
     by categories of: netting

 netting |      mean  variance         N
---------+------------------------------
       0 |    7.5625  9.673387        32
       1 |  3.111111  3.163399        18
---------+------------------------------
   Total |      5.96  11.87592        50
-----------------------------------------
```

I also used the good of fit test to see the results:

```
estat gof

    Deviance goodness-of-fit =  60.64895
    Prob > chi2(48)          =    0.1040

    Pearson goodness-of-fit  =  56.93861
    Prob > chi2(48)          =    0.1766
```

Since the p-value is 0.05, the overall model fits the data well.

Therefore, since under the Poisson model, the variance is assumed to be equal to the mean. In the netting group with net use (netting = 1), the mean and variance are very similar, suggesting that the Poisson model has a good fit for this subgroup. However, for the overall data (and for the netting = 0 group) the variance exceeds the mean, indicating some over dispersion. Despite this, the good of fit test suggests that Poisson model is still a reasonable starting point for the analysis.

(b)

A Poisson regression model, sometimes known as a log-linear model, takes the form on the log-count scale

$$\log E(Y \mid X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

The predicted mean of the Poisson model on the count scale is given by:

$$E(Y \mid X) = \exp\big(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\big).$$

Next, we fit a Poisson regression model using net use as the predictor variable. The Stata command used was:

```
poisson infect netting
```

Output:

```
Poisson regression                              Number of obs     =         50
                                                LR chi2(1)        =      42.45
                                                Prob > chi2       =     0.0000
Log likelihood = -116.13416                     Pseudo R2         =     0.1545

------------------------------------------------------------------------------
      infect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     netting |  -.8882219   .1482881    -5.99   0.000    -1.178861   -.5975825
       _cons |   2.023202   .0642824    31.47   0.000     1.897211    2.149193
------------------------------------------------------------------------------
```

In our analysis of the Zika infection data, the outcome Y is the infection count (`infect`) and the predictor is net use (`netting`). Thus, the model simplifies to:

$$\log E(\text{infect} \mid \text{netting}) = \beta_0 + \beta_1 \cdot \text{netting}.$$

(log-count scale)

$$E(\text{infect} \mid \text{netting}) = \exp\big(\beta_0 + \beta_1 \cdot \text{netting}\big).$$

(count scale)

we obtained the following estimates:

$$\hat{\beta}_0 = 2.0232 \quad \text{and} \quad \hat{\beta}_1 = -0.8882.$$

Thus, the estimated model on the log(counts) scale is:

$$\log E(\text{infect} \mid \text{netting}) = 2.0232 - 0.8882 \cdot \text{netting}.$$

Here, the intercept (2.0232) represents the log mean infection count for villages with no net use (netting = 0), and the coefficient for netting indicates the change in the log mean count when nets are used.

Exponentiating the linear predictor gives the predicted mean counts on the original scale:

$$E(\text{infect} \mid \text{netting} = 0) = \exp(2.0232) \approx 7.56,$$

$$E(\text{infect} \mid \text{netting} = 1) = \exp(2.0232 - 0.8882) = \exp(1.1350) \approx 3.11.$$

Thus, villages without nets are predicted to have an average of approximately 7.56 infections, whereas those with nets are predicted to have about 3.11 infections. This result indicates a substantial reduction in infection counts associated with the use of nets.

(c)

First, we verify that the Poisson model reproduces the mean infection counts by netting strata. When we fit the Poisson model without the IRR option:

```
 poisson infect netting

Iteration 0:   log likelihood = -116.13525
Iteration 1:   log likelihood = -116.13416
Iteration 2:   log likelihood = -116.13416

Poisson regression                                Number of obs     =         50
                                                  LR chi2(1)        =      42.45
                                                  Prob > chi2       =     0.0000
Log likelihood = -116.13416                       Pseudo R2         =     0.1545

------------------------------------------------------------------------------
      infect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     netting |  -.8882219   .1482881    -5.99   0.000    -1.178861   -.5975825
       _cons |   2.023202   .0642824    31.47   0.000     1.897211    2.149193
------------------------------------------------------------------------------
```

The estimated baseline incidence rate (i.e., for villages with no net use, netting $= 0$) is given by the constant:

$$\hat{E}(\text{infect} \mid \text{netting} = 0) = \exp(2.0232) \approx 7.56,$$

which exactly matches the observed mean for villages without nets (7.5625).

For villages with nets (netting $= 1$), the predicted mean is:

$$\hat{E}(\text{infect} \mid \text{netting} = 1) = \exp(2.0232 - 0.8882) = \exp(1.1350) \approx 3.11,$$

which is consistent with the observed mean for villages with nets (3.1111).

Next, we re-fit the Poisson model with the `irr` option:

```
 poisson infect netting, irr

Iteration 0:   log likelihood = -116.13525
Iteration 1:   log likelihood = -116.13416
Iteration 2:   log likelihood = -116.13416

Poisson regression                                Number of obs     =         50
                                                  LR chi2(1)        =      42.45
                                                  Prob > chi2       =     0.0000
Log likelihood = -116.13416                       Pseudo R2         =     0.1545

------------------------------------------------------------------------------
      infect |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     netting |   .4113866   .0610038    -5.99   0.000     .3076288      .55014
       _cons |     7.5625   .4861359    31.47   0.000     6.667271    8.577934
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.x
```

The output yields an incidence rate ratio (IRR) for `netting` of:

$$\text{IRR} = 0.4114 \quad (95\% \text{ CI: } [0.3076, 0.5501]).$$

The observed mean for villages without nets is 7.5625 and for villages with nets is 3.1111. The ratio of these means is:

$$\frac{3.1111}{7.5625} \approx 0.4114,$$

which exactly matches the estimated IRR of 0.4114.

This IRR indicates that, on average, villages that use nets have an infection rate that is only about 41.1% of the infection rate for villages that do not use nets.

Summary of the Result: The Poisson model reproduces the observed mean infection counts by netting strata: approximately 7.56 infections for villages without nets and 3.11 infections for villages with nets. The estimated IRR of 0.4114 implies a relative infection reduction of about 58.9% (i.e., $1 - 0.4114$) for villages with net use compared to those without, or we could say, on average, villages that use nets have an infection rate that is only about 41.1% of the infection rate for villages that do not use nets. In other words, the use of nets is associated with a substantial reduction in the infection rate.

(d)

```
. predict infect_hat
(option n assumed; predicted number of events)

. list netting infect infect_hat, clean

      netting   infect   infect~t
  1.        0        6     7.5625
  2.        0        8     7.5625
  3.        0       11     7.5625
  4.        0        8     7.5625
  5.        0        6     7.5625
  6.        0       14     7.5625
  7.        0       11     7.5625
  8.        0        5     7.5625
  9.        0       10     7.5625
 10.        0       10     7.5625
 11.        0        5     7.5625
 12.        0       10     7.5625
 13.        0        4     7.5625
 14.        0        4     7.5625
 15.        0       13     7.5625
 16.        0        7     7.5625
 17.        0        6     7.5625
 18.        0        6     7.5625
 19.        0        6     7.5625
 20.        0        9     7.5625
 21.        0        2     7.5625
 22.        0       15     7.5625
 23.        0        2     7.5625
 24.        0        9     7.5625
 25.        0        8     7.5625
 26.        0        5     7.5625
 27.        0        5     7.5625
```

28.	0	7	7.5625
29.	0	8	7.5625
30.	0	8	7.5625
31.	0	7	7.5625
32.	0	7	7.5625
33.	1	2	3.111111
34.	1	2	3.111111
35.	1	4	3.111111
36.	1	7	3.111111
37.	1	1	3.111111
38.	1	1	3.111111
39.	1	4	3.111111
40.	1	4	3.111111
41.	1	4	3.111111
42.	1	3	3.111111
43.	1	0	3.111111
44.	1	5	3.111111
45.	1	4	3.111111
46.	1	3	3.111111
47.	1	5	3.111111
48.	1	4	3.111111
49.	1	2	3.111111
50.	1	1	3.111111

The output shows that for every observation in the non-netting group (netting = 0), the predicted count is 7.5625, while for every observation in the netting group (netting = 1), the predicted count is 3.111111. Here, the first 32 observations (netting=0) all have $\widehat{\text{infect}} = 7.5625$, and the remaining 18 observations (netting=1) all have $\widehat{\text{infect}} = 3.111111$.

Interpretation:

We examine the observed infection counts separately for the two netting groups:

Non-Netting Group (netting = 0):

- **Minimum Count:** 2
- **Maximum Count:** 15
- **Range:** $15 - 2 = 13$
- **Mean:** 7.5625

Netting Group (netting = 1):

- **Minimum Count:** 0
- **Maximum Count:** 7
- **Range:** $7 - 0 = 7$
- **Mean:** 3.1111

The Poisson model assumes that the expected value equals the variance. Hence, for the non-netting group with a higher mean (≈ 7.56), we also expect a larger variance compared to the netting group, which has a lower mean (≈ 3.11). Indeed, the observed data are consistent with this assumption, as the non-netting group shows a wider range (and, by implication, larger variability) than the netting group.

The fitted Poisson model produced the following predicted values:

- For the non-netting group: $\widehat{E}(\text{infect} \mid \text{netting} = 0) = 7.5625$.
- For the netting group: $\widehat{E}(\text{infect} \mid \text{netting} = 1) = 3.1111$.

Thus, while the actual counts within each group vary (with a range of 13 in the non-netting group and 7 in the netting group), the Poisson model correctly captures the group means, and the higher mean in the non-netting group is accompanied by greater variability, as expected under the Poisson distribution.

Question 3

(a)

The incidence rate (IR) is defined as the total number of events (kidney stones) divided by the total person-time (years of follow-up). Given the summary statistics:

```
. summarize stones yrfu

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      stones |      1,423    .6872804    2.461781          0         60
        yrfu |      1,423    8.059175    6.656896          1   32.17808
```

the total number of stones is $\bar{S} \times N$ and the total person-years is $\bar{T} \times N$. Thus, the overall incidence rate is:

$$\text{IR} = \frac{\bar{S} \times N}{\bar{T} \times N} = \frac{\bar{S}}{\bar{T}}.$$

Substituting the given values, we get:

$$\text{IR} = \frac{0.6872804}{8.059175} \approx 0.0853 \quad \text{stones per person-year.}$$

This IR value indicates that, on average, there are approximately 0.0853 kidney stone episodes per person-year in this cohort.

(b)

Using the command:

```
ir stones sex yrfu
```

the following summary is obtained for kidney stone incidence rates by sex (with the coding: 0 = male, 1 = female):

```
. ir stones sex yrfu

                 | Sex [0=male, 1=female] |
```

```
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
Post-treatment n |       234         744  |        978
Years of post-tr |  3552.088    7916.118  |   11468.21
-----------------+------------------------+------------
                 |                        |
  Incidence rate |  .0658768    .0939855  |   .0852793
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Inc. rate diff. |        -.0281087       |   -.0389185   -.0172989
 Inc. rate ratio |         .7009249       |     .602557    .8129174 (exact)
 Prev. frac. ex. |         .2990751       |    .1870826     .397443 (exact)
 Prev. frac. pop |         .0926336       |
                 +-----------------------------------------------
                     (midp)   Pr(k<=234) =                  0.0000 (exact)
                     (midp) 2*Pr(k<=234) =                  0.0000 (exact)
```

The incidence rate ratio (IRR) for females relative to males is calculated as:

$$\text{IRR} = \frac{0.06588}{0.09399} \approx 0.7009,$$

with a 95% confidence interval of approximately $[0.6026, 0.8129]$.

This result indicates that the kidney stone incidence rate for females is about 70.09% of that for males.

(c)

We fit a Poisson regression model using `sex` as the only predictor, while including the follow-up time (`yrfu`) as the exposure variable. The Stata command used was:

```
poisson stones sex, exposure(yrfu) irr
```

Output:

```
Poisson regression                              Number of obs     =      1,423
                                                LR chi2(1)        =      23.83
                                                Prob > chi2       =     0.0000
Log likelihood = -2370.2832                     Pseudo R2         =     0.0050

------------------------------------------------------------------------------
      stones |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   0.7009249   0.0525347    -4.74   0.000     0.6051645    0.8118383
       _cons |   0.0939855   0.0034457   -64.50   0.000     0.087469     0.1009874
    ln(yrfu) |          1  (exposure)
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Interpretation: The estimated incidence rate ratio (IRR) for `sex` is 0.7009. Since `sex` is coded as 0 for males and 1 for females, this IRR indicates that, holding follow-up time constant, the rate of kidney-stone formation for

females is 70.1% of that for males. In other words, about a 0.7 fold risk for the female group on the kidney-stone formation rate.

Comparison of the Model-Based IRR with the Direct Calculation

The Poisson regression model with sex as the only predictor (and follow-up time as the exposure) yielded an IRR of 0.7009. Since `sex` is coded as 0 for males (the reference group) and 1 for females, this indicates that the incidence rate for females is 70.09% that of males—that is, females have a lower risk of kidney-stone formation.

To compare with the direct calculation from the incidence rates in (b):

- The incidence rate for males (sex = 0) is 0.09399 (or about 9.4 per 100 person-years).
- The incidence rate for females (sex = 1) is 0.06588 (or about 6.6 per 100 person-years).

Taking the ratio,

$$\frac{0.06588}{0.09399} \approx 0.7009,$$

which exactly reproduces the model-based IRR of 0.7009 as the slope for variable `sex`.

From the Poisson regression output, we have:

- The intercept ($\hat{\beta}_0$) represents the baseline incidence rate for the reference group (males, where sex = 0):

$$IR_{\text{male}} = \exp(\hat{\beta}_0) = 0.0939855.$$

- The coefficient for `sex` gives an incidence rate ratio (IRR) of 0.7009249. This indicates that the incidence rate for females is 70.09249% of that for males. Thus, the predicted incidence rate for females is:

$$IR_{\text{female}} = IR_{\text{male}} \times 0.7009249 = 0.0939855 \times 0.7009249 \approx 0.065882.$$

Taking the ratio of the female incidence rate to the male incidence rate:

$$\frac{IR_{\text{female}}}{IR_{\text{male}}} = \frac{0.065882}{0.0939855} \approx 0.7009249$$

also matches the Incidence rate ratio in (b)

Testing

Given the 95% confidence interval of (0.6052, 0.8118) for the IRR doesn't contain 0 and a $p - value < 0.001$, we conclude that the difference in kidney-stone formation rates between men and women is statistically significant.

(d)

In studies of incidence rates, the outcome is a count of events occurring over a period of time. Since different individuals may have different follow-up times, it is crucial to adjust for this variable to obtain unbiased estimates of the event rate.

Model without Exposure: If follow-up time were not included, the Poisson model would be specified as:

$$\log E(Y \mid X) = \beta_0 + \beta_1 X,$$

implicitly assuming that each individual has the same exposure time (e.g., $T = 1$). This model is appropriate only when follow-up time is constant across all subjects, which is rarely the case in real-world studies.

Model with Exposure:

When follow-up time (T) is included as an offset (log-transformed) in the model, the Poisson regression specification becomes:

$$\log E(Y \mid X, T) = \log(T) + \beta_0 + \beta_1 X,$$

or equivalently,

$$E(Y \mid X, T) = T \times \exp(\beta_0 + \beta_1 X).$$

This formulation correctly models the expected number of events as proportional to the amount of time under observation. By including $\log(T)$ as an offset, the model adjusts for differences in follow-up time, ensuring that the estimated rates (events per person-year) are comparable across individuals.

To obtain the incidence rate (i.e., events per unit time), we divide the expected count by the person-time:

$$\text{Incidence Rate} = \frac{E(Y \mid X, T)}{T} = \exp(\beta_0 + \beta_1 X).$$

To illustrate using a 2x2 table framework:

- **For the unexposed (reference) group ($X = 0$):**

  $$E(Y \mid X = 0, T = T_0) = T_0 \times \exp(\beta_0).$$

  Thus, the incidence rate is

  $$IR_0 = \frac{E(Y \mid X = 0, T = T_0)}{T_0} = \exp(\beta_0).$$

- **For the exposed group ($X = 1$):**

  $$E(Y \mid X = 1, T = T_1) = T_1 \times \exp(\beta_0 + \beta_1).$$

  Thus, the incidence rate is

  $$IR_1 = \frac{E(Y \mid X = 1, T = T_1)}{T_1} = \exp(\beta_0 + \beta_1).$$

Thus, the incidence rate ratio (IRR) comparing the exposed to the unexposed group is:

$$IRR = \frac{IR_1}{IR_0} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1).$$

Conclusion: In a setting that disease incidence is analyzed relative to exposure time. Instead of using a fixed denominator (e.g., the total number of individuals, N), the relevant denominator is the total person-time at risk—i.e., the sum of exposure times across all individuals. This yields incidence rates expressed per unit of time. By modeling rates (events per unit time) instead of simple proportions, the analysis reflects the risk of an event over time when individuals are observed for different lengths of time. This approach allows for meaningful comparisons between groups with varying at-risk periods, providing a more accurate assessment of the effect of exposures on disease incidence.

Therefore, using Poisson regression with an exposure offset enables us to model incidence rates and make meaningful comparisons.

(e)

We fit a Poisson regression model with `nx1` (indicator for having only one functional kidney) as the only predictor, using follow-up time (`yrfu`) as the exposure variable:

```
poisson stones nx1, exposure(yrfu) irr
```

The model output is:

```
Poisson regression                              Number of obs     =      1,423
                                                LR chi2(1)        =      34.08
                                                Prob > chi2       =     0.0000
Log likelihood = -2365.1612                     Pseudo R2         =     0.0072

------------------------------------------------------------------------------
      stones |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         nx1 |   0.4135836   0.072195    -5.06   0.000     0.2937500    0.5823027
       _cons |   0.0894829   0.0029124   -74.16   0.000     0.0839529    0.0953771
    ln(yrfu) |          1  (exposure)
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Test

We can test the value of an individual covariate in a Poisson regression model much like in simple or multiple linear regression. Specifically, we test

$$H_0 : \text{IRR} = \exp(\beta) = 1,$$

which is equivalent to

$$H_0 : \log(\text{IRR}) = \beta = 0,$$

versus

$$H_A : \beta \neq 0.$$

Theory from Generalized Linear Models and the estimation method shows that β is approximately normally distributed. Therefore, the test statistic is given by:

$$Z = \frac{\hat{\beta} - 0}{\widehat{\text{se}}(\hat{\beta})}.$$

In our output, the estimated z-score for the covariate `nx1` is -5.06 and the corresponding p-value is 0.000. This indicates that the null hypothesis is rejected, and thus the incidence rate ratio (IRR) for `nx1` is significantly different from 1, which means that individuals with only one functional kidney have different kidney-stone rates.

Interpretation:

The baseline incidence rate (for individuals with two functional kidneys, i.e., `nx1` $= 0$) is estimated as:

$$IR_{2 \text{ kidneys}} = \exp(\hat{\beta}_0) = 0.08948.$$

For individuals with only one functional kidney (`nx1` $= 1$), the incidence rate is:

$$IR_{1 \text{ kidney}} = 0.08948 \times 0.41358 \approx 0.0370.$$

This IRR of 0.4136 means that individuals with one functional kidney have a kidney-stone formation rate that is about 41.4% of the rate for individuals with two kidneys.

Thus, we conclude that, ignoring other effects, having only one functional kidney is associated with a significantly lower rate of kidney-stone formation compared to having two functional kidneys.

(f)

We fit a Poisson regression model with `age` as the only predictor, using follow-up time (`yrfu`) as the exposure variable:

```
poisson stones age, exposure(yrfu) irr
```

The output is as follows:

```
Poisson regression                              Number of obs     =      1,423
                                                LR chi2(1)        =      50.61
                                                Prob > chi2       =     0.0000
Log likelihood = -2356.8958                     Pseudo R2         =     0.0106

------------------------------------------------------------------------------
      stones |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   0.9816975   0.0025643    -7.07   0.000     0.9766844    0.9867364
       _cons |   0.1843083   0.0202868   -15.36   0.000     0.1485434    0.2286844
    ln(yrfu) |          1  (exposure)
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

The 95% confidence interval for the IRR is (0.9767, 0.9867), which doesn't contain 0, and the effect is statistically significant $p < 0.001$). Therefore, we conclude that age has a significant effect on the kidney-stone rates.

Interpretation: The estimated incidence rate ratio (IRR) for `age` is approximately 0.9817. This means that for every additional year of age, the rate of kidney-stone formation is multiplied by 0.9817, which means that if age increase by 1 year, then the kidney-stone rate will be 98.17% than before on average, or equivalently, decreases by about

$$100 \times (1 - 0.9817) \approx 1.83\% \text{ per year.}$$

If age increases by n years, the incidence rate is multiplied by:

$$\text{IRR}_n = (0.9817)^n.$$

Thus, the new incidence rate for a person n years older is:

$$IR_{\text{new}} = IR_{\text{baseline}} \times (0.9817)^n.$$

Therefore, ignoring other factors, older patients tend to have a slightly lower rate of kidney-stone formation.

(g)

We fit the following Poisson regression model using all three predictors:

```
poisson stones i.sex i.nx1 c.age, exposure(yrfu) irr
```

with the output:

```
Poisson regression                              Number of obs     =      1,423
```

```
                                                LR chi2(3)        =     102.68
                                                Prob > chi2       =     0.0000
Log likelihood = -2330.8625                     Pseudo R2         =     0.0216

------------------------------------------------------------------------------
      stones |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.sex |   0.7094036   0.0536933    -4.54   0.000     0.6116002     0.822847
       1.nx1 |   0.4764933   0.0838503    -4.21   0.000     0.3374953    0.6727379
         age |   0.9812308   0.002617    -7.10   0.000     0.9761149    0.9863736
       _cons |   0.2149314   0.0246577   -13.40   0.000     0.1716512    0.2691243
    ln(yrfu) |          1  (exposure)
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Model Specification: The mathematical expression for this Poisson regression model with an exposure offset is:

$$\log E(\text{stones} \mid \text{sex}, \text{nx1}, \text{age}, \text{yrfu}) = \log(\text{yrfu}) + \beta_0 + \beta_1 \,\text{sex} + \beta_2 \,\text{nx1} + \beta_3 \,\text{age},$$

or equivalently,

$$E(\text{stones} \mid \text{sex}, \text{nx1}, \text{age}, \text{yrfu}) = \text{yrfu} \times \exp\Big(\beta_0 + \beta_1 \,\text{sex} + \beta_2 \,\text{nx1} + \beta_3 \,\text{age}\Big).$$

Since the output reports exponentiated coefficients (IRRs), we have:

$$\exp(\beta_0) = 0.2149314, \quad \exp(\beta_1) = 0.7094036, \quad \exp(\beta_2) = 0.4764933, \quad \exp(\beta_3) = 0.9812308.$$

Taking logarithms, we obtain approximate estimates of the coefficients:

$$\beta_0 \approx \ln(0.2149314) \approx -1.537, \quad \beta_1 \approx \ln(0.7094036) \approx -0.343,$$

$$\beta_2 \approx \ln(0.4764933) \approx -0.742, \quad \beta_3 \approx \ln(0.9812308) \approx -0.019.$$

Prediction: We wish to predict the rate of kidney-stone formation for a 45-year-old female patient with only one functional kidney. For this patient, we set:

$$\text{sex} = 1, \quad \text{nx1} = 1, \quad \text{age} = 45,$$

and, to obtain the rate per person-year, we set yrfu $= 1$.

The predicted log rate is:

$$\log \hat{IR} = \beta_0 + \beta_1(1) + \beta_2(1) + \beta_3(45).$$

Substituting the estimated coefficients:

$$\log \hat{IR} \approx -1.537 - 0.343 - 0.742 - 0.019 \times 45.$$

$$\log \hat{IR} \approx -1.537 - 0.343 - 0.742 - 0.855 \approx -3.477.$$

Thus, the predicted incidence rate is:

$$\hat{IR} = \exp(-3.477) \approx 0.0309 \quad \text{stones per person-year.}$$

Or we can directly compute the predicted incidence rate using the multiplicative property of the IRRs. In our model, the estimated coefficients (expressed as IRRs) are:

$$\exp(\beta_0) = 0.2149314, \quad \exp(\beta_{\text{sex}}) = 0.7094036, \quad \exp(\beta_{\text{nx1}}) = 0.4764933, \quad \exp(\beta_{\text{age}}) = 0.9812308.$$

For a 45-year-old female patient with only one functional kidney (i.e., sex = 1, nx1 = 1, age = 45), the predicted incidence rate per person-year is given by:

$$\hat{IR} = \exp(\beta_0) \times \exp(\beta_{\text{sex}}) \times \exp(\beta_{\text{nx1}}) \times [\exp(\beta_{\text{age}})]^{45}.$$

Substituting the numerical values:

$$\hat{IR} = 0.2149314 \times 0.7094036 \times 0.4764933 \times (0.9812308)^{45}.$$

First, we compute the effect of age:

$$(0.9812308)^{45} \approx \exp\big(45 \times \ln(0.9812308)\big) \approx \exp(-0.855) \approx 0.426.$$

Then, the predicted incidence rate is:

$$\hat{IR} \approx 0.2149314 \times 0.7094036 \times 0.4764933 \times 0.426 \approx 0.0309 \quad \text{stones per person-year.}$$

Thus, a 45-year-old female patient with one functional kidney is predicted to have approximately 0.031 kidney stones per person-year.

When we run:

```
. display exp(_b[_cons] + _b[sex] + _b[nx1] + _b[age]*45)
.03097085
```

The predicted result is almost the same.