Unit 10 Biostatistics Recitation

Ryan Wei

2023-07-07

Problem 4

You decide to conduct a case-control study of your own to test if smoking status during pregnancy (1 = never, 2 = past, 3 = current) is associated with infant mortality (0 = did not die within one year of birth, 1 = died within one year of birth).

Problem 4a

- Should you use linear or logistic regression to answer this question? Explain your reasoning.

Problem 4a

- Should you use linear or logistic regression to answer this question? Explain your reasoning.
- Binary outcome $\rightarrow$ Logistic regression

Problem 4b

```
. logistic mortality i.smoking, nolog

Logistic regression                             Number of obs     =        120
                                                LR chi2(2)        =       8.96
                                                Prob > chi2       =     0.0114
Log likelihood = -78.699822                     Pseudo R2         =     0.0538

------------------------------------------------------------------------------
   mortality | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     smoking |
          2 |   2.284615   1.045482     1.81   0.071     .9317246    5.601942
          3 |   3.994083   1.919169     2.88   0.004     1.557443    10.24288
            |
       _cons |   .4814815   .1625384    -2.17   0.030     .2484455    .9330997
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.

. logit mortality i.smoking, nolog

Logistic regression                             Number of obs     =        120
                                                LR chi2(2)        =       8.96
                                                Prob > chi2       =     0.0114
Log likelihood = -78.699822                     Pseudo R2         =     0.0538

------------------------------------------------------------------------------
   mortality |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     smoking |
          2 |   .8261977   .4576185     1.81   0.071     -.070718    1.723113
          3 |   1.384814   .4805031     2.88   0.004     .4430453    2.326583
            |
       _cons |  -.7308875   .3375798    -2.17   0.030    -1.392532   -.0692433
------------------------------------------------------------------------------
```

- What is the difference between the two outputs above and how are they related?

Problem 4b: Answer

- The first output provides the odds ratios comparing the odds of infant mortality in each group with the odds of infant mortality among babies born to mothers who were never smokers at the time of pregnancy.

Problem 4b: Answer

- The first output provides the odds ratios comparing the odds of infant mortality in each group with the odds of infant mortality among babies born to mothers who were never smokers at the time of pregnancy.
- The second output provides the coefficients used in the regression equation.

Problem 4b: Answer

- The first output provides the odds ratios comparing the odds of infant mortality in each group with the odds of infant mortality among babies born to mothers who were never smokers at the time of pregnancy.
- The second output provides the coefficients used in the regression equation.
- The odds ratios can be calculated by exponentiating the coefficients: $e^{\hat{\beta}_2} = \widehat{\text{OR}}_2$.
- The coefficients can be calculated by taking the natural log of the odds ratios $\hat{\beta}_2 = \ln(\widehat{\text{OR}}_2)$.

```
exp(1.384814);log(2.284615)
```

```
## [1] 3.994083
```

```
## [1] 0.8261975
```

Problem 4c

Write the fitted regression model. Be sure to define your variables.

```
. logistic mortality i.smoking, nolog

Logistic regression                             Number of obs     =        120
                                                LR chi2(2)        =       8.96
                                                Prob > chi2       =     0.0114
Log likelihood = -78.699822                     Pseudo R2         =     0.0538

------------------------------------------------------------------------------
   mortality | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     smoking |
          2 |   2.284615   1.045482     1.81   0.071     .9317246    5.601942
          3 |   3.994083   1.919169     2.88   0.004     1.557443    10.24288
            |
       _cons |   .4814815   .1625384    -2.17   0.030     .2484455    .9330997
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.

. logit mortality i.smoking, nolog

Logistic regression                             Number of obs     =        120
                                                LR chi2(2)        =       8.96
                                                Prob > chi2       =     0.0114
Log likelihood = -78.699822                     Pseudo R2         =     0.0538

------------------------------------------------------------------------------
   mortality |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     smoking |
          2 |   .8261977   .4576185     1.81   0.071     -.070718    1.723113
          3 |   1.384814   .4805031     2.88   0.004     .4430453    2.326583
            |
       _cons |  -.7308875   .3375798    -2.17   0.030    -1.392532   -.0692433
------------------------------------------------------------------------------
```

Problem 4c: Answer

- Model: $\text{logit}(\hat{p}) = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.73 + 0.83x_2 + 1.38x_3$
  - $x_2 = \begin{cases} 1 \text{ if past smoker} \\ 0 \text{ otherwise} \end{cases}$
  - $x_3 = \begin{cases} 1 \text{ if current smoker} \\ 0 \text{ otherwise} \end{cases}$
  - $p =$ the estimated probability of death within the first year of birth

Problem 4d

Based on what you know about the study, state which values of the first column in the Stata output (i.e. Odds Ratio/Coef.) are interpretable and interpret those values.

```
. logistic mortality i.smoking, nolog

Logistic regression                             Number of obs     =        120
                                                LR chi2(2)        =       8.96
                                                Prob > chi2       =     0.0114
Log likelihood = -78.699822                     Pseudo R2         =     0.0538

------------------------------------------------------------------------------
   mortality | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     smoking |
           2 |   2.284615   1.045482     1.81   0.071     .9317246    5.601942
           3 |   3.994083   1.919169     2.88   0.004     1.557443    10.24288
             |
       _cons |   .4814815   .1625384    -2.17   0.030     .2484455    .9330997
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.

. logit mortality i.smoking, nolog

Logistic regression                             Number of obs     =        120
                                                LR chi2(2)        =       8.96
                                                Prob > chi2       =     0.0114
Log likelihood = -78.699822                     Pseudo R2         =     0.0538

------------------------------------------------------------------------------
   mortality |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     smoking |
           2 |   .8261977   .4576185     1.81   0.071     -.070718    1.723113
           3 |   1.384814   .4805031     2.88   0.004     .4430453    2.326583
             |
       _cons |  -.7308875   .3375798    -2.17   0.030    -1.392532   -.0692433
------------------------------------------------------------------------------
```

Problem 4d: Answer

- Since this is a case-control study, we cannot estimate the probability or odds of the outcome, infant mortality. However, we can estimate the odds ratio for this outcome. Therefore, we cannot interpret the intercept (i.e. the value in the row for `_cons` in the Stata output), but we can interpret the slopes (i.e. the values in the rows for `smoking` in the Stata output).

Problem 4d: Answer

- Since this is a case-control study, we cannot estimate the probability or odds of the outcome, infant mortality. However, we can estimate the odds ratio for this outcome. Therefore, we cannot interpret the intercept (i.e. the value in the row for `_cons` in the Stata output), but we can interpret the slopes (i.e. the values in the rows for `smoking` in the Stata output).

- Interpretation
  - $\widehat{OR}_2$: The estimated odds ratio comparing the odds of infant mortality among babies born to mothers who were past smokers during pregnancy is 2.28 times (or 128% higher than) the odds of infant mortality among babies born to mothers who were never smokers during pregnancy.
  - $\widehat{OR}_3$: The estimated odds ratio comparing the odds of infant mortality among babies born to mothers who were current smokers during pregnancy is 3.99 times (or 299% higher than) the odds of infant mortality among babies born to mothers who were never smokers during pregnancy.