22401 HW6

Bin Yu

Mar 05, 2025

Question 1

(a)

The 2x2 table for recurrence is given by:

```
    . tabulate recur trt

           |          trt
     recur |         0          1 |     Total
-----------+----------------------+----------
         0 |       358        408 |       766
         1 |       164        112 |       276
-----------+----------------------+----------
     Total |       522        520 |     1,042
```

Odds Ratio (OR):
The odds of recurrence in the tamoxifen group are

$$\text{Odds}_1 = \frac{112}{408} \approx 0.2745,$$

and in the placebo group

$$\text{Odds}_0 = \frac{164}{358} \approx 0.4581.$$

Thus, the odds ratio is

$$\text{OR} = \frac{0.2745}{0.4581} \approx 0.5992.$$

Risk Difference (RD):
The risk of recurrence for the placebo group is

$$p_0 = \frac{164}{522} \approx 0.3141,$$

and for the tamoxifen group is

$$p_1 = \frac{112}{520} \approx 0.2154.$$

Hence, the risk difference is

$$\Delta = p_1 - p_0 \approx 0.2154 - 0.3141 \approx -0.0987.$$

This indicates a reduction of about 9.9 percentage points in recurrence with tamoxifen.

Relative Risk is given by:

$$RR = p_1/p_0 = \frac{0.2154}{0.3141} \approx 0.6858$$

The 2x2 table for death is presented below:

```
     tabulate dead trt

           |          trt
      dead |         0          1 |     Total
-----------+----------------------+----------
         0 |       337        368 |       705
         1 |       185        152 |       337
-----------+----------------------+----------
     Total |       522        520 |     1,042
```

Odds Ratio (OR):
The odds of death in the tamoxifen group are

$$\text{Odds}_1 = \frac{152}{368} \approx 0.4130,$$

and in the placebo group

$$\text{Odds}_0 = \frac{185}{337} \approx 0.5490.$$

Thus, the odds ratio is

$$\text{OR} = \frac{0.4130}{0.5490} \approx 0.7523.$$

Risk Difference (RD):
The risk of death in the placebo group is

$$p_0 = \frac{185}{522} \approx 0.3544,$$

and in the tamoxifen group is

$$p_1 = \frac{152}{520} \approx 0.2923.$$

The risk difference is therefore

$$\Delta = p_1 - p_0 \approx 0.2923 - 0.3544 \approx -0.0621.$$

This shows a reduction of about 6.2 percentage points in the risk of death with tamoxifen.

Relative Risk is given by:

$$RR = p_1/p_0 = \frac{0.2923}{0.3544} \approx 0.8248$$

(b)

Endpoint 1: Disease Recurrence Stata Output

```
. logit recur trt, or

Iteration 0:   log likelihood = -602.37472
```

```
Iteration 1:   log likelihood = -595.83931
Iteration 2:   log likelihood = -595.81568
Iteration 3:   log likelihood = -595.81568

Logistic regression                               Number of obs     =      1,042
                                                  LR chi2(1)        =      13.12
                                                  Prob > chi2       =     0.0003
Log likelihood = -595.81568                       Pseudo R2         =     0.0109

------------------------------------------------------------------------------
       recur | Odds Ratio   Std. Err.      z    P>|z|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
         trt |   .5992348   .0853156    -3.60   0.000     .4533232    .7921112
       _cons |   .4581006   .0431949    -8.28   0.000     .3808026     .551089
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Hypothesis Test

We test the null hypothesis that treatment has no effect on disease recurrence:

$$H_0 : \text{OR} = e^{\beta_1} = 1 \quad \text{(equivalently, } \beta_1 = 0\text{)}.$$

From the output, the estimated odds ratio is 0.5992348. Converting to the log scale:

$$\hat{\beta} = \ln(0.5992348) \approx -0.512.$$

The Wald test statistic is given by

$$Z = \frac{\hat{\beta}}{s.e.(\hat{\beta})} = -3.60,$$

with a corresponding p-value < 0.001.

Interpretation

Since the p-value is below 0.05, we reject the null hypothesis. This indicates that tamoxifen treatment is significantly associated with a reduction in disease recurrence. Therefore, the odds of recurrence in the tamoxifen group are approximately 60% compared to those in the placebo group.

Death

Stata Output

```
. logit dead trt, or

Iteration 0:   log likelihood = -655.85351
Iteration 1:   log likelihood = -653.55678
Iteration 2:   log likelihood = -653.55533
Iteration 3:   log likelihood = -653.55533

Logistic regression                               Number of obs     =      1,042
                                                  LR chi2(1)        =       4.60
                                                  Prob > chi2       =     0.0320
Log likelihood = -653.55533                       Pseudo R2         =     0.0035

------------------------------------------------------------------------------
        dead | Odds Ratio   Std. Err.      z    P>|z|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
         trt |   .7524089   .1000141    -2.14   0.032     .5798395    .9763377
       _cons |   .5489614   .0502315    -6.55   0.000     .4588328     .656794
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Hypothesis Test

We now test the null hypothesis that treatment has no effect on death:

$$H_0 : \text{OR} = e^{\beta_1} = 1 \quad (\text{i.e., } \beta_1 = 0).$$

The estimated odds ratio is 0.7524089, which on the log scale gives:

$$\hat{\beta} = \ln(0.7524089) \approx -0.285.$$

The test statistic is

$$Z = \frac{\hat{\beta}}{s.e.(\hat{\beta})} \approx -2.14,$$

with a p-value of 0.032.

Interpretation

Since the p-value is below 0.05, we reject H_0. This indicates that tamoxifen treatment is significantly associated with a reduction in death. Specifically, the odds of death in the tamoxifen group are about 75% compared to those in the placebo group.

(c)

We extend the logistic regression model for the recurrence endpoint by including additional continuous covariates: age at diagnosis (`age`), body mass index (`bmi`, in kg/m^2), and tumor size (`tumsiz`, in mm). The Stata output below provides the odds ratios and associated tests for each predictor.

Stata Output

```
logit recur trt age bmi tumsiz, or

Iteration 0:   log likelihood = -602.37472
Iteration 1:   log likelihood = -583.84502
Iteration 2:   log likelihood = -583.66653
Iteration 3:   log likelihood = -583.66651

Logistic regression                             Number of obs     =      1,042
                                                LR chi2(4)        =      37.42
                                                Prob > chi2       =     0.0000
Log likelihood = -583.66651                     Pseudo R2         =     0.0311

------------------------------------------------------------------------------
       recur | Odds Ratio   Std. Err.      z    P>|z|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
         trt |   .6077191   .0876392    -3.45   0.001     .4580902    .8062221
         age |   .9744114    .007044    -3.59   0.000     .9607027    .9883156
         bmi |   1.020568   .0134689     1.54   0.123      .994508    1.047311
      tumsiz |   1.018924   .0062295     3.07   0.002     1.006787    1.031207
       _cons |   .6972886   .3368405    -0.75   0.455     .2705335    1.797232
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Hypothesis Tests and Interpretations

For each predictor, we test the null hypothesis $H_0 : \beta_i = 0$ (equivalently, OR $= 1$) using the Wald test.

- **Treatment (`trt`):**
  OR $= 0.6077$ with a 95% CI of (0.4581, 0.8062), $z = -3.45$, and $p = 0.001$.
  Interpretation: Holding age, BMI, and tumor size constant, patients receiving tamoxifen have odds of recurrence that are multiplied by 0.6077 relative to those on placebo. In other words, tamoxifen is associated with a 39.2% reduction in the odds of recurrence (since $1 - 0.6077 \approx 0.3923$). This effect is significant.

- **Age (`age`):**
  OR $= 0.9744$ with a 95% CI of (0.9607, 0.9883), $z = -3.59$, and $p < 0.001$.
  Interpretation: Controlling for treatment, BMI, and tumor size, each additional year of age multiplies the odds of recurrence by 0.9744. That is, for every extra year, the odds of recurrence decrease by about 2.56% (since $1 - 0.9744 \approx 0.0256$). This effect is significant.

- **Body Mass Index (`bmi`):**
  OR $= 1.0206$ with a 95% CI of (0.9945, 1.0473), $z = 1.54$, and $p = 0.123$.
  Interpretation: Holding the other variables constant, a one unit increase in BMI multiplies the odds of recurrence by 1.0206. That is, for every extra BMI, the odds of recurrence increase by about 2.06% increase. However, this effect is not statistically significant.

- **Tumor Size (`tumsiz`):**
  OR $= 1.0189$ with a 95% CI of (1.0068, 1.0312), $z = 3.07$, and $p = 0.002$.
  Interpretation: Controlling for treatment, age, and BMI, each additional millimeter in tumor size multiplies the odds of recurrence by 1.0189. That is, for every extra tumor size, the odds of recurrence increase by about 1.89% increase.This effect is significant.

(d)

Change the variable to dead:

Stata Output

```
logit dead trt age bmi tumsiz, or

Iteration 0:   log likelihood = -655.85351
Iteration 1:   log likelihood = -629.67686
Iteration 2:   log likelihood = -629.48063
Iteration 3:   log likelihood = -629.48058

Logistic regression                             Number of obs     =      1,042
                                                LR chi2(4)        =      52.75
                                                Prob > chi2       =     0.0000
Log likelihood = -629.48058                     Pseudo R2         =     0.0402

------------------------------------------------------------------------------
        dead | Odds Ratio   Std. Err.      z    P>|z|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
         trt |   .7229534   .0986705    -2.38   0.017     .5532688    .9446793
         age |   1.031604   .0074625     4.30   0.000     1.017081    1.046334
         bmi |   1.023056   .0128355     1.82   0.069     .9982053    1.048525
      tumsiz |   1.027569   .0061317     4.56   0.000     1.015621    1.039658
       _cons |   .0290188   .014626    -7.02   0.000     .0108058    .0779292
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Hypothesis Tests and Interpretation

- **Age (`age`):**
  The odds ratio is 1.0316. This implies that, holding treatment, BMI, and tumor size constant, each additional year of age multiplies the odds of death by 1.0316, or a 3.16% increase in the odds of death per year. This effect is statistically significant ($z = 4.30$, $p < 0.001$).
  Comparison with Recurrence: In the recurrence model, age had an odds ratio of 0.9744 (i.e., each additional year was associated with a 2.56% decrease in the odds of recurrence). Hence, while older age appears protective against recurrence, it increases the risk of death.
- **Treatment (`trt`):**
  The odds ratio is 0.723, indicating that tamoxifen reduces the odds of death by about 27.7% compared to placebo (since $1-0.723 \approx 0.277$), controlling for age, BMI, and tumor size. This effect is statistically significant ($z = -2.38$, $p = 0.017$).

Non-Cancer Death (`ned` as Outcome) Stata Output

```
logit ned trt age bmi tumsiz, or

Iteration 0:   log likelihood = -274.67427
Iteration 1:   log likelihood = -245.29402
Iteration 2:   log likelihood = -239.21336
Iteration 3:   log likelihood = -239.1413
Iteration 4:   log likelihood = -239.14128

Logistic regression                             Number of obs     =      1,042
                                                LR chi2(4)        =      71.07
                                                Prob > chi2       =     0.0000
Log likelihood = -239.14128                     Pseudo R2         =     0.1294

------------------------------------------------------------------------------
         ned | Odds Ratio   Std. Err.      z    P>|z|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
         trt |   1.097409    .270827     0.38   0.706     .6765539     1.78006
         age |   1.141015    .022213     6.78   0.000     1.098298    1.185393
         bmi |   .9839115   .0230083    -0.69   0.488      .939834    1.030056
      tumsiz |   1.017106   .0103739     1.66   0.096     .9969753    1.037643
       _cons |   .0000321   .0000425    -7.81   0.000     2.39e-06    .0004309
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Hypothesis Tests and Interpretation

- **Age (`age`):**
  The odds ratio is 1.1410. This indicates that, after adjusting for treatment, BMI, and tumor size, each additional year of age multiplies the odds of non-cancer death by 1.1410, or each additional year of age is associated with a 14.10% increase of odds of non-cancer death. This effect is statistically significant ($z = 6.78$, $p < 0.001$).
- **Treatment (`trt`):**
  The odds ratio is 1.0974, suggesting that tamoxifen is associated with a 9.74% increase in the odds of non-cancer death relative to placebo. However, this effect is not statistically significant ($z = 0.38$, $p = 0.706$).

In summary, age is a significant predictor in both models. In the overall survival model (using `dead` as the outcome), each additional year of age increases the odds of death by about 3.16%, holding treatment, BMI, and

tumor size constant. In contrast, in the non-cancer death model (using `ned` as the outcome), the effect of age is more pronounced, with each additional year increasing the odds by approximately 14.10%.

This stronger association in the non-cancer death model suggests that older age may be more closely linked to deaths from causes other than cancer. Since mortality is inevitable and the risk of dying from other diseases generally increases with age, this result may indicate that, while cancer-related mortality is influenced by multiple factors, the overall vulnerability associated with aging plays a more substantial role in mortality. Regarding treatment, tamoxifen shows a protective effect in the overall survival model by reducing the odds of death by about 27.7%, but it is not significantly related to non-cancer death, which further supports the idea that its benefits are primarily linked to cancer-specific outcomes.

(e)

We fit the logistic regression model for endometrial cancer (`endo`) using treatment (`trt`) as the sole predictor. The Stata output is shown below.

Logistic Regression Output

```
logit endo trt, or
Iteration 0:   log likelihood = -74.243285
Iteration 1:   log likelihood = -71.973465
Iteration 2:   log likelihood = -71.766957
Iteration 3:   log likelihood = -71.766622
Iteration 4:   log likelihood = -71.766622

Logistic regression                             Number of obs     =      1,042
                                                LR chi2(1)        =       4.95
                                                Prob > chi2       =     0.0260
Log likelihood = -71.766622                     Pseudo R2         =     0.0334

------------------------------------------------------------------------------
        endo | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         trt |   3.738703   2.446306     2.02   0.044     1.036963    13.47965
       _cons |   .0057803   .0033469    -8.90   0.000     .0018582     .017981
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Interpretation of Logistic Regression

Controlling for baseline odds, the odds ratio for treatment is 3.74. This indicates that patients receiving tamoxifen (`trt = 1`) have approximately 3.74 times higher odds of developing endometrial cancer compared to those receiving placebo (`trt = 0`). The result is statistically significant ($p = 0.044$).

Predicted Probabilities

The predicted probabilities for endometrial cancer were computed using the logistic regression model and stored in the variable `prob_d`. The output from the `tabulate prob_d` command is as follows:

```
tabulate prob_d

   Pr(endo) |      Freq.     Percent        Cum.
------------+-----------------------------------
   .0057471 |        522       50.10       50.10
   .0211538 |        520       49.90      100.00
```

```
------------+-----------------------------------
      Total |      1,042      100.00
```

These predicted probabilities correspond to the two treatment groups:

- For the placebo group (`trt = 0`), the predicted probability of endometrial cancer is approximately 0.00575 (or 0.57%).
- For the tamoxifen group (`trt = 1`), the predicted probability is about 0.02115 (or 2.12%).

These predictions are calculated by applying the logistic regression model:

$$p = \frac{\exp(L)}{1 + \exp(L)},$$

where L is the linear predictor (i.e., $L = \beta_0 + \beta_1 \cdot \texttt{trt}$).

2×2 Table and Odds Ratio Calculation

The following 2×2 table was generated to summarize the distribution of endometrial cancer by treatment group:

```
 tabulate endo trt

           |          trt
      endo |         0          1 |     Total
-----------+----------------------+----------
         0 |       519        509 |     1,028
         1 |         3         11 |        14
-----------+----------------------+----------
     Total |       522        520 |     1,042
```

For the placebo group (`trt = 0`):

$$\text{Odds}_0 = \frac{\text{Number of endo} = 1}{\text{Number of endo} = 0} = \frac{3}{519}.$$

For the tamoxifen group (`trt = 1`):

$$\text{Odds}_1 = \frac{\text{Number of endo} = 1}{\text{Number of endo} = 0} = \frac{11}{509}.$$

Thus, the odds ratio is calculated as:

$$\text{OR} = \frac{\text{Odds}_1}{\text{Odds}_0} = \frac{\frac{11}{509}}{\frac{3}{519}} = \frac{11 \times 519}{3 \times 509} \approx 3.738.$$

This is consistent with the logistic regression output.

Overall Conclusion

Both the logistic regression analysis and the 2×2 table indicate that tamoxifen treatment is associated with a significantly increased risk of endometrial cancer. The odds ratio of approximately 3.74 suggests that tamoxifen increases the odds of developing endometrial cancer by 3.74 times compared to placebo. However, the absolute risk remains very low, as evidenced by the predicted probabilities of 0.57% for the placebo group and 2.12% for the tamoxifen group.

(f)

We use the logistic regression model (omitting BMI) given by:

$$\text{logit}(p) = \beta_0 + \beta_{\text{trt}} \cdot \text{trt} + \beta_{\text{age}} \cdot \text{age} + \beta_{\text{tumsiz}} \cdot \text{tumsiz},$$

```
. logit recur trt age tumsiz

Iteration 0:   log likelihood = -602.37472
Iteration 1:   log likelihood = -584.99952
Iteration 2:   log likelihood = -584.84224
Iteration 3:   log likelihood = -584.84222

Logistic regression                                 Number of obs     =      1,042
                                                    LR chi2(3)        =      35.07
                                                    Prob > chi2       =     0.0000
Log likelihood = -584.84222                         Pseudo R2         =     0.0291

------------------------------------------------------------------------------
       recur |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         trt |  -.4916501   .1439548    -3.42   0.001    -.7737963   -.2095039
         age |  -.0236281   .0070497    -3.35   0.001    -.0374453   -.0098109
      tumsiz |   .0197753   .0060813     3.25   0.001     .0078561    .0316945
       _cons |   .0222022   .4143408     0.05   0.957    -.7898909    .8342953
------------------------------------------------------------------------------
```

The estimated coefficients are:

$$\beta_0 = 0.0222022, \quad \beta_{\text{trt}} = -0.4916501, \quad \beta_{\text{age}} = -0.0236281, \quad \beta_{\text{tumsiz}} = 0.0197753.$$

The predicted probability is computed as:

$$p = \frac{\exp(L)}{1 + \exp(L)},$$

where L is the linear predictor.

Case 1: Age = 50, Tumor Size = 30 mm, Tamoxifen (trt = 1)

$$L_1 = 0.0222022 + (-0.4916501)(1) + (-0.0236281)(50) + (0.0197753)(30).$$

Thus,

$$L_1 = 0.0222022 - 0.4916501 - 1.181405 + 0.593259 \approx -1.057594.$$

The predicted probability is:

$$p_1 = \frac{\exp(-1.057594)}{1 + \exp(-1.057594)} \approx \frac{0.347}{1.347} \approx 0.2576 \quad (25.8\%).$$

Case 2: Age = 50, Tumor Size = 30 mm, Placebo (trt = 0)

$$L_2 = 0.0222022 + (-0.4916501)(0) + (-0.0236281)(50) + (0.0197753)(30).$$

Thus,

$$L_2 = 0.0222022 - 0 - 1.181405 + 0.593259 \approx -0.565944.$$

The predicted probability is:

$$p_2 = \frac{\exp(-0.565944)}{1+\exp(-0.565944)} \approx \frac{0.568}{1.568} \approx 0.3622 \quad (36.2\%).$$

Case 3: Age = 65, Tumor Size = 10 mm, Tamoxifen (trt = 1)

$$L_3 = 0.0222022 + (-0.4916501)(1) + (-0.0236281)(65) + (0.0197753)(10).$$

Thus,

$$L_3 = 0.0222022 - 0.4916501 - 1.5358265 + 0.197753 \approx -1.807521.$$

The predicted probability is:

$$p_3 = \frac{\exp(-1.807521)}{1+\exp(-1.807521)} \approx \frac{0.164}{1.164} \approx 0.1409 \quad (14.1\%).$$

Case 4: Age = 65, Tumor Size = 10 mm, Placebo (trt = 0)

$$L_4 = 0.0222022 + (-0.4916501)(0) + (-0.0236281)(65) + (0.0197753)(10).$$

Thus,

$$L_4 = 0.0222022 - 0 - 1.5358265 + 0.197753 \approx -1.315871.$$

The predicted probability is:

$$p_4 = \frac{\exp(-1.315871)}{1+\exp(-1.315871)} \approx \frac{0.268}{1.268} \approx 0.2114 \quad (21.1\%).$$

We can check whether it is consistent with stata output:

```
. * Case 1: age = 50, tumor size = 30 mm, tamoxifen (trt = 1)
. margins, at(trt=1 age=50 tumsiz=30)

Adjusted predictions                            Number of obs     =      1,042
Model VCE    : OIM

Expression   : Pr(recur), predict()
at           : trt             =           1
               age             =          50
               tumsiz          =          30

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .2577692   .0223182    11.55   0.000     .2140264     .301512
------------------------------------------------------------------------------

.
. * Case 2: age = 50, tumor size = 30 mm, placebo (trt = 0)
. margins, at(trt=0 age=50 tumsiz=30)
```

```
Adjusted predictions                            Number of obs     =      1,042
Model VCE    : OIM

Expression   : Pr(recur), predict()
at           : trt             =          0
               age             =         50
               tumsiz          =         30

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .3621729   .0242055    14.96   0.000      .314731    .4096148
------------------------------------------------------------------------------


.
. * Case 3: age = 65, tumor size = 10 mm, tamoxifen (trt = 1)
. margins, at(trt=1 age=65 tumsiz=10)

Adjusted predictions                            Number of obs     =      1,042
Model VCE    : OIM

Expression   : Pr(recur), predict()
at           : trt             =          1
               age             =         65
               tumsiz          =         10

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .1409377   .0192462     7.32   0.000     .1032158    .1786597
------------------------------------------------------------------------------


.
. * Case 4: age = 65, tumor size = 10 mm, placebo (trt = 0)
. margins, at(trt=0 age=65 tumsiz=10)

Adjusted predictions                            Number of obs     =      1,042
Model VCE    : OIM

Expression   : Pr(recur), predict()
at           : trt             =          0
               age             =         65
               tumsiz          =         10

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .2115057   .0254104     8.32   0.000     .1617023    .2613092
------------------------------------------------------------------------------
```

As a result, they are consistent.

Summary of Predicted Probabilities:

Case	Age	Tumor Size (mm)	Treatment	Probability
1	50	30	Tamoxifen (trt = 1)	25.8%
2	50	30	Placebo (trt = 0)	36.2%
3	65	10	Tamoxifen (trt = 1)	14.1%
4	65	10	Placebo (trt = 0)	21.1%

Table 1: Predicted probabilities of recurrence for various covariate patterns.

(g)

Now we compare the logistic regression results for three endpoints—breast cancer recurrence, breast cancer death, and endometrial cancer—and discusses the argument supporting the continued use of tamoxifen based on both relative and absolute risks.

1. Breast Cancer Recurrence

Logistic Regression Output:

```
logit recur trt age tumsiz

Iteration 0:   log likelihood = -602.37472
Iteration 1:   log likelihood = -584.99952
Iteration 2:   log likelihood = -584.84224
Iteration 3:   log likelihood = -584.84222

Logistic regression                             Number of obs     =      1,042
                                                LR chi2(3)        =      35.07
                                                Prob > chi2       =     0.0000
Log likelihood = -584.84222                     Pseudo R2         =     0.0291

------------------------------------------------------------------------------
       recur |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         trt |  -.4916501   .1439548    -3.42   0.001    -.7737963   -.2095039
         age |  -.0236281   .0070497    -3.35   0.001    -.0374453   -.0098109
      tumsiz |   .0197753   .0060813     3.25   0.001     .0078561    .0316945
       _cons |   .0222022   .4143408     0.05   0.957    -.7898909    .8342953
------------------------------------------------------------------------------
```

Predicted Probabilities by Treatment (using `margins`):

```
margins, at(trt=(0 1))

Predictive margins                              Number of obs     =      1,042
Model VCE    : OIM

Expression   : Pr(recur), predict()

1._at        : trt             =           0
2._at        : trt             =           1
```

```
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |    .310867   .0199772    15.56   0.000     .2717124    .3500215
          2  |   .2180468   .0180181    12.10   0.000      .182732    .2533617
------------------------------------------------------------------------------
```

Interpretation: Patients on placebo have a predicted recurrence probability of about 31.1%, whereas those on tamoxifen have about 21.8%. This indicates a substantial reduction in the absolute risk of recurrence with tamoxifen treatment.

2. Breast Cancer Death

Logistic Regression Output (including BMI):

```
logit dead trt age bmi tumsiz

Iteration 0:   log likelihood = -655.85351
Iteration 1:   log likelihood = -629.67686
Iteration 2:   log likelihood = -629.48063
Iteration 3:   log likelihood = -629.48058

Logistic regression                             Number of obs     =      1,042
                                                LR chi2(4)        =      52.75
                                                Prob > chi2       =     0.0000
Log likelihood = -629.48058                     Pseudo R2         =     0.0402

------------------------------------------------------------------------------
        dead |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         trt |  -.3244105   .1364825    -2.38   0.017    -.5919113   -.0569098
         age |   .0311148   .0072339     4.30   0.000     .0169367     .045293
         bmi |    .022794   .0125463     1.82   0.069    -.0017963    .0473842
      tumsiz |    .027196   .0059672     4.56   0.000     .0155005    .0388915
       _cons |  -3.539813   .5040185    -7.02   0.000    -4.527671   -2.551954
------------------------------------------------------------------------------
```

Predicted Probabilities by Treatment (using `margins`):

```
margins, at(trt=(0 1))

Predictive margins                              Number of obs     =      1,042
Model VCE    : OIM

Expression   : Pr(dead), predict()

1._at        : trt             =           0
2._at        : trt             =           1
```

```
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .3574313   .0205736    17.37   0.000     .3171079    .3977548
          2  |   .2898839   .0193716    14.96   0.000     .2519163    .3278515
------------------------------------------------------------------------------
```

Interpretation: The predicted probability of death is approximately 35.7% for the placebo group and 28.9% for the tamoxifen group. It is also a substantial reduction.

3. Endometrial Cancer

Logistic Regression Output (including age and tumor size):

```
logit endo trt age tumsiz

Iteration 0:   log likelihood = -74.243285
Iteration 1:   log likelihood = -71.575745
Iteration 2:   log likelihood = -71.288921
Iteration 3:   log likelihood = -71.288498
Iteration 4:   log likelihood = -71.288498

Logistic regression                             Number of obs     =      1,042
                                                LR chi2(3)        =       5.91
                                                Prob > chi2       =     0.1161
Log likelihood = -71.288498                     Pseudo R2         =     0.0398

------------------------------------------------------------------------------
        endo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         trt |   1.293811   .6550783     1.98   0.048     .0098815    2.577741
         age |   .0281477   .0296761     0.95   0.343    -.0300163    .0863118
      tumsiz |   .0004843   .0242703     0.02   0.984    -.0470846    .0480533
       _cons |  -6.728525   1.889838    -3.56   0.000    -10.43254    -3.02451
------------------------------------------------------------------------------
```

Predicted Probabilities by Treatment (using `margins`):

```
margins, at(trt=(0 1))

Predictive margins                              Number of obs     =      1,042
Model VCE    : OIM

Expression   : Pr(endo), predict()

1._at        : trt             =           0
2._at        : trt             =           1

------------------------------------------------------------------------------
             |            Delta-method
```

```
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+------------------------------------------------------------------
         _at |
          1  |    .005822   .0033518     1.74   0.082    -.0007474    .0123915
          2  |   .0208875   .0062347     3.35   0.001     .0086676    .0331073
------------------------------------------------------------------------------
```

Interpretation: For endometrial cancer, the predicted probability is very low in both groups: about 0.58% for the placebo group and 2.09% for the tamoxifen group. Although tamoxifen is associated with a relative increase in risk (with an estimated odds ratio of approximately 3.29 from the regression coefficients), the absolute risk remains very small.

Therefore, while tamoxifen increases the relative risk of endometrial cancer (with a predicted probability of approximately 2.1% versus 0.6% for placebo), the absolute risk for this adverse outcome is extremely small. In contrast, tamoxifen significantly reduces the risk of breast cancer recurrence and death, with absolute reductions of approximately 9.3 percentage points (31.1% to 21.8%) for recurrence and 7 percentage points (35.7% to 28.9%) for death. Given that breast cancer events and mortality occur at much higher rates than endometrial cancer, the substantial benefits in lowering recurrence and mortality far outweigh the modest absolute increase in the risk of endometrial cancer. This favorable benefit–risk profile supports the continued use of tamoxifen in appropriate patients.

Question 2

(a)

We first create a binary outcome variable, `fail`, where

$$\texttt{fail} = \begin{cases} 1, & \text{if } \texttt{damaged} > 0 \quad \text{(any damage)}, \\ 0, & \text{if } \texttt{damaged} = 0 \quad \text{(no damage)}. \end{cases}$$

This is done in Stata as follows:

```
gen fail = (damaged > 0)
```

Next, we fit a logistic regression model predicting the probability of an O-ring failure by temperature (`temp`):

Stata Output

```
Iteration 0:   log likelihood = -14.133576
Iteration 1:   log likelihood = -10.302864
Iteration 2:   log likelihood = -10.157825
Iteration 3:   log likelihood = -10.157596
Iteration 4:   log likelihood = -10.157596

Logistic regression                             Number of obs     =         23
                                                LR chi2(1)        =       7.95
                                                Prob > chi2       =     0.0048
Log likelihood = -10.157596                     Pseudo R2         =     0.2813

------------------------------------------------------------------------------
        fail | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        temp |   0.7928171  0.0858118    -2.14   0.032     0.6412715  0.9801761
```

```
       _cons |   3412315   2.52e+07      2.04   0.041     1.787897   6.51e+12
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Predicted Probabilities

We compute the predicted probabilities of failure as follows:

```
tabulate prob_f
```

Pr(fail)	Freq.	Percent	Cum.
0.0227033	1	4.35	4.35
0.0356414	1	4.35	8.70
0.0445405	1	4.35	13.04
0.0690441	2	8.70	21.74
0.0855436	2	8.70	30.43
0.129546	1	4.35	34.78
0.1580491	1	4.35	39.13
0.2299683	4	17.39	56.52
0.2736211	1	4.35	60.87
0.3220941	1	4.35	65.22
0.3747243	3	13.04	78.26
0.4304931	1	4.35	82.61
0.602681	1	4.35	86.96
0.8288448	1	4.35	91.30
0.8593166	1	4.35	95.65
0.9392478	1	4.35	100.00
Total	23	100.00	

Interpretation

The logistic regression model yields an estimated odds ratio for temperature of approximately 0.7928 (95% CI: 0.6413 to 0.9802) with $p = 0.032$. This indicates that for each one–degree Fahrenheit increase in launch temperature, the odds of an O–ring failure decrease by about 20.7% (since $1 - 0.7928 \approx 0.2072$). Or each one-degree Fahrenheit increase multiply the odds of an O-ring failure by 0.793. This effect is significant.

The predicted probabilities (`prob_f`) for O–ring failure across the 23 observations range from roughly 2.3% to 93.9%. These predictions are obtained by applying the logistic function

$$p = \frac{\exp(L)}{1 + \exp(L)},$$

where $L = \beta_0 + \beta_1 \cdot \texttt{temp}$.

(b)

After dropping observation Flight #18:

Stata Output

```
drop if flightno == 18
logit fail temp, or
Iteration 0:   log likelihood = -12.890958
```

```
Iteration 1:   log likelihood = -7.6133038
Iteration 2:   log likelihood = -7.1953417
Iteration 3:   log likelihood = -7.1885097
Iteration 4:   log likelihood = -7.1884884
Iteration 5:   log likelihood = -7.1884884

Logistic regression                               Number of obs     =         22
                                                  LR chi2(1)        =      11.40
                                                  Prob > chi2       =     0.0007
Log likelihood = -7.1884884                       Pseudo R2         =     0.4424

------------------------------------------------------------------------------
        fail | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        temp |   0.6969644   0.1223276   -2.06   0.040     0.4940968   0.9831259
       _cons |   1.46e+10   1.73e+11     1.98   0.048     1.238225    1.72e+20
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Interpretation

With flight #18 excluded, the estimated odds ratio for temperature is approximately 0.697 (95% CI: 0.494 to 0.983, $p = 0.040$). This indicates that for each one–degree Fahrenheit increase in temperature, the odds of an O–ring failure decrease by about 30.3% (calculated as $1 - 0.697 \approx 0.303$). Or each one-degree Fahrenheit increase multiply the odds of an O-ring failure by 0.697. The statistical significance (p = 0.040) suggests that this effect is significant.

(c)

Using the fitted logistic regression model

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot \text{temp},$$

```
     logit fail temp

Iteration 0:   log likelihood = -12.890958
Iteration 1:   log likelihood = -7.6133038
Iteration 2:   log likelihood = -7.1953417
Iteration 3:   log likelihood = -7.1885097
Iteration 4:   log likelihood = -7.1884884
Iteration 5:   log likelihood = -7.1884884

Logistic regression                               Number of obs     =         22
                                                  LR chi2(1)        =      11.40
                                                  Prob > chi2       =     0.0007
Log likelihood = -7.1884884                       Pseudo R2         =     0.4424

------------------------------------------------------------------------------
        fail |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        temp |   -.361021   .1755149    -2.06   0.040    -.7050238   -.0170181
       _cons |    23.4033   11.83166     1.98   0.048     .2136786    46.59293
------------------------------------------------------------------------------
```

we have the estimated coefficients

$$\beta_0 = 23.4033 \quad \text{and} \quad \beta_1 = -0.361021,$$

we calculate the linear predictor for a launch temperature of 31°F:

$$L = 23.4033 + (-0.361021 \times 31) \approx 23.4033 - 11.19165 = 12.21165.$$

The predicted probability of an O–ring failure is given by

$$P(Y = 1|X) = \frac{\exp(L)}{1 + \exp(L)}.$$

Substituting the computed value of L,

$$P(Y = 1|X) \approx \frac{\exp(12.21165)}{1 + \exp(12.21165)} \approx \frac{201118.59}{201119.59} \approx 0.999995.$$

Interpretation: At a launch temperature of 31°F, the model predicts an almost certain probability of an O–ring failure (99.99%). Based on this result, I would have advised against launching on that day.

(d)

Using the logistic regression model based on all launches except flight #18, we predicted the probability of damage (O–ring failure) and classified an observation as "damaged" if the predicted probability was ≥ 0.50:

```
    . logit fail temp

Iteration 0:   log likelihood = -12.890958
Iteration 1:   log likelihood = -7.6133038
Iteration 2:   log likelihood = -7.1953417
Iteration 3:   log likelihood = -7.1885097
Iteration 4:   log likelihood = -7.1884884
Iteration 5:   log likelihood = -7.1884884

Logistic regression                             Number of obs     =         22
                                                LR chi2(1)        =      11.40
                                                Prob > chi2       =     0.0007
Log likelihood = -7.1884884                     Pseudo R2         =     0.4424

------------------------------------------------------------------------------
        fail |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        temp |   -.361021   .1755149    -2.06   0.040    -.7050238   -.0170181
       _cons |    23.4033   11.83166     1.98   0.048     .2136786    46.59293
------------------------------------------------------------------------------
. predict phat
(option pr assumed; Pr(fail))
. list fail temp phat
     +-------------------------+
     | fail   temp       phat |
     |-------------------------|
  1. |    1     53      .9862 |
  2. |    1     57   .9440178 |
  3. |    1     58   .9215857 |
```

```
  4. |    1     63   .6590322 |
  5. |    0     66   .3955413 |
     |-----------------------|
  6. |    0     67   .3132219 |
  7. |    0     68   .2411985 |
  8. |    0     69   .1813629 |
  9. |    1     70   .1337545 |
 10. |    0     72   .0697715 |
     |-----------------------|
 11. |    0     73   .0496786 |
 12. |    0     75   .0247645 |
 13. |    0     76   .0173905 |
 14. |    0     78   .0085238 |
 15. |    0     79   .0059562 |
     |-----------------------|
 16. |    0     81   .0029022 |
 17. |    1     70   .1337545 |
 18. |    0     67   .3132219 |
 19. |    0     70   .1337545 |
 20. |    0     76   .0173905 |
     |-----------------------|
 21. |    0     67   .3132219 |
 22. |    0     70   .1337545 |
     +-----------------------+
```

The following 2×2 table summarizes the predicted classification versus the actual damage status:

```
     estat classification

Logistic model for fail

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         4             0  |          4
     -     |         2            16  |         18
-----------+--------------------------+-----------
   Total   |         6            16  |         22

Classified + if predicted Pr(D) >= .5
True D defined as fail != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   66.67%
Specificity                     Pr( -|~D)  100.00%
Positive predictive value       Pr( D| +)  100.00%
Negative predictive value       Pr(~D| -)   88.89%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    0.00%
False - rate for true D         Pr( -| D)   33.33%
False + rate for classified +   Pr(~D| +)    0.00%
False - rate for classified -   Pr( D| -)   11.11%
--------------------------------------------------
Correctly classified                        90.91%
--------------------------------------------------
```

From the table, the number of correctly classified observations is the sum of true positives and true negatives:

$$\text{Correctly Classified} = 4 + 16 = 20 \quad \text{(out of 22 observations)},$$

which corresponds to an overall accuracy of approximately 90.91%. Therefore, based on the model (excluding flight #18), 20 out of 22 launches were correctly classified as either damaged or not damaged when using a threshold of 0.50 for the predicted probability. This indicates that the model performs well in distinguishing between launches with and without O–ring damage.

(e)

When deciding on future launches using the predicted probability of O–ring damage (based on temperature), the most critical consideration is the risk of a false negative—i.e., predicting that no damage will occur when, in fact, damage is present. A false negative could lead to a launch that is actually unsafe, with potentially catastrophic consequences.

Below are three classification tables generated with different cut-points.

Classification Table with Cutoff = 0.5

```
Logistic model for fail

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         4             0  |          4
     -     |         2            16  |         18
-----------+--------------------------+-----------
   Total   |         6            16  |         22

Classified + if predicted Pr(D) >= .5
True D defined as fail != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   66.67%
Specificity                     Pr( -|~D)  100.00%
Positive predictive value       Pr( D| +)  100.00%
Negative predictive value       Pr(~D| -)   88.89%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    0.00%
False - rate for true D         Pr( -| D)   33.33%
False + rate for classified +   Pr(~D| +)    0.00%
False - rate for classified -   Pr( D| -)   11.11%
--------------------------------------------------
Correctly classified                        90.91%
--------------------------------------------------
```

Classification Table with Cutoff = 0.67

```
Logistic model for fail

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         3             0  |          3
     -     |         3            16  |         19
```

```
-----------+-------------------------+-----------
   Total   |         6            16 |         22

Classified + if predicted Pr(D) >= .67
True D defined as fail != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   50.00%
Specificity                     Pr( -|~D)  100.00%
Positive predictive value       Pr( D| +)  100.00%
Negative predictive value       Pr(~D| -)   84.21%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    0.00%
False - rate for true D         Pr( -| D)   50.00%
False + rate for classified +   Pr(~D| +)    0.00%
False - rate for classified -   Pr( D| -)   15.79%
--------------------------------------------------
Correctly classified                        86.36%
--------------------------------------------------
```

Classification Table with Cutoff = 0.33

```
Logistic model for fail

              -------- True --------
Classified |         D            ~D  |      Total
-----------+-------------------------+-----------
     +     |         4             1  |          5
     -     |         2            15  |         17
-----------+-------------------------+-----------
   Total   |         6            16  |         22

Classified + if predicted Pr(D) >= .33
True D defined as fail != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   66.67%
Specificity                     Pr( -|~D)   93.75%
Positive predictive value       Pr( D| +)   80.00%
Negative predictive value       Pr(~D| -)   88.24%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    6.25%
False - rate for true D         Pr( -| D)   33.33%
False + rate for classified +   Pr(~D| +)   20.00%
False - rate for classified -   Pr( D| -)   11.76%
--------------------------------------------------
Correctly classified                        86.36%
--------------------------------------------------
```

Comparison and Analysis

- **Cutoff = 0.5:**
  - **Sensitivity:** 66.67% (i.e., two-thirds of the true damage cases are detected)
  - **Specificity:** 100% (all non-damaged cases are correctly identified)
  - **PPV and NPV:** 100% and 88.89%, respectively

  - **Overall Accuracy:** 90.91%

- **Cutoff = 0.67:**
  - **Sensitivity:** Drops to 50% (a 16.67 percentage point decrease, meaning half of the true damage cases are missed)
  - **Specificity:** Remains at 100%
  - **NPV:** Decreases to 84.21%
  - **Overall Accuracy:** Declines to 86.36%
- **Cutoff = 0.33:**
  - **Sensitivity:** Remains at 66.67% (same as the 0.5 cutoff)
  - **Specificity:** Decreases to 93.75% (a slight increase in false positives, that some none-damage cases are classified as damaged)
  - **PPV:** Drops to 80.00%
  - **Overall Accuracy:** 86.36%

Discussion:

The key metric in this context is **sensitivity**, as failing to identify a damaged O–ring (a false negative) could have catastrophic implications. Comparing the three cut-points:

- Increasing the cutoff to 0.67 improves specificity to 100% but at the cost of reducing sensitivity to only 50%. This means that half of the actual damage cases would be missed—a risk that is unacceptable for launch safety.
- Lowering the cutoff to 0.33 does not improve sensitivity compared to 0.5 (both remain at 66.67%) but does reduce specificity (from 100% to 93.75%) and lowers the positive predictive value.
- The cutoff of 0.5 strikes a balance by achieving high sensitivity (66.67%) and perfect specificity, resulting in the highest overall accuracy (90.91%).

Conclusion: For deciding on future launches, the most relevant quantity is sensitivity because missing a potential O–ring damage (false negative) is far more dangerous than a false positive. Based on the comparisons, a cutoff of 0.5 is the most reasonable choice, as it minimizes the risk of false negatives while maintaining perfect specificity.