PBHS 32410 / STAT 22401 **Problem Set 6** **Winter 2025**

Possible points 96

1. [48 pts total] Breast cancer and tamoxifen: The anti-estrogen drug tamoxifen is among the most widely used treatments in breast cancer. Data in tam_12.dta on the Canvas website are from a large clinical trial of tamoxifen vs placebo in women with estrogen receptor-positive tumors confined to the breast. Variables that we will use are mostly self-explanatory, but further information is given below:

Note: this dataset is more appropriately analyzed as time to event data, to take different amounts of follow-up into account. However, these analyses considering events occurring over the trial time period do lend useful information about treatment and other factors related to events after disease onset.

(a) [8 pts] Investigate whether tamoxifen (variable *trt*: 0 = placebo, 1 = tamoxifen) reduces disease recurrence (variable *recur*) and death (variable *death*). Construct simple 2x2 tables of treatment by outcome for these two response variables and compute the odds ratio for each endpoint . Also compute the risk difference.

(b) [8 pts] Execute the logistic regression models to obtain the odds ratio with associated test for each endpoint. Interpret the result.

(c) [8 pts] Additional (continuous scale) variables that may be related to prognosis and possibly the effect of tamoxifen include age at diagnosis (*age*), body mass index (*bmi*, kg/m^2), and tumor size in mm (*tumsiz*). Add these to the logistic regression model above for the recurrence endpoint to obtain odds ratios with associated tests. Interpret the result.

(d) [8 pts] Change the endpoint to survival (variable *death* and re-examine the effect of age. How is it different from the effect on recurrence? Then, use the variable *ned* as the outcome (this is death with no evidence of disease recurrence or other cancer, in other words, non-cancer death). How do age and treatment relate to this endpoint?

(e) [8 pts] Tamoxifen has the potential to increase risk for some undesirable outcomes, including endometrial (uterine) cancer. Use the logistic regression model predicting endometrial cancer (endpoint *endo*) with treatment. Also make the associated 2x2 table. Interpret the result (note: this dataset is slightly different so results will not be identical to what we looked at in class)

(f) [8 pts] For the endpoint *recur*, calculate the probability of recurrence for patients with the following covariate values (you can omit BMI from the model):

1. age 50, tumor size 30mm, tamoxifen treatment
2. age 50, tumor size 30mm, placebo treatment

3. age 65, tumor size 10mm, tamoxifen treatment

4. age 65, tumor size 10mm, placebo treatment

(g) [8 pts] Considering the risk associated with tamoxifen, what would be the argument supporting its continued use? (hint: consider the absolute risks of endometrial cancer and breast cancer events/death in addition to the relative benefits/risks of tamoxifen)?

2. [40 pts, Exercise 12.3 modified from C&H Space Shuttle launches:. The NASA Space Shuttle was the primary means of human-accompanied space travel throughout the 1980's until the mid-2000's. The program suffered two catastrophic events, marring the total safety record despite popularity and public support. The dataset for this exercise contains information on launches leading up to the first major incident, and involve measurements of damage to a key component of the shuttle rocket engine in relation to temperature at launch time.

On the course website in Orings_12.dta

(a) [8 pts] Fit a logistic regression predicting the probability of an O-ring failure by temperature. Note: you need to make a (0,1) response variable from the data field coding no damage (0) vs. any damage (value 1 or greater) for each launch as the outcome variable Interpret the temperature coefficient.

(b) [6 pts] The data for flight #18 that was launched when the launch temperature was 75 degrees Fahrenheit was thought to be problematic for other reasons, and possibly should be omitted. Fit a logistic regression model to the reduced data set leaving flight #18 out. Interpret the temperature coefficient.

(c) [6 pts] From the fitted model above, find the probability of an O-ring failure when the temperature at launch was 31 degrees Fahrenheit. This was the temperature forecast for the day of the launching of the fatal Challenger flight on January 20, 1986. Would you have advised the launching on that particular day?

(d) [8 pts] From the model based on all launches except #18, (from (b) above) predict the probability of damage. Classify all those with probability $\geq .50$ as predicted to be damaged. Make a 2x2 table of predicted to be damaged vs. actually damaged. How many were predicted correctly overall?

(e) [12 pts] The classification table above provides numerous summaries such as the positive predictive value and negative predictive value. For deciding on future launches based on probability of O-ring damage (where you would be putting in a temperature as the predictor in order to make the decision), which quantity derived from the 2x2 table might be most relevant in choosing your cut-point? Try moving the cut-point down to .33 and up to .67 and see how the results change.