(1, 1pt) Data were collected from a total of 196=195+1 anglers.

(2, 2pts) Since a logit link is used, increasing the cost of a fishing trip by $100 would multiply the odds of a "Yes" response by a factor of $\exp(-100 \times 0.001752) = 0.839$, i.e., a 16% decrease in the odds. Equivalently, the log-odds of a "Yes" response would decrease by $100 \times 0.001752 = 0.1752$.

(3, 2pts) This tests the null hypothesis that none of the predictors have an effect on the response, by comparing the null model (which includes only the intercept and no predictors) to the full model (which includes all predictors). Since the deviance difference of 41.62 is large relative to a chi-squared distribution with 9 degrees of freedom, we reject the null hypothesis and conclude that at least one predictor is significantly associated with the response.

(4, 2pts) An infinite increase in cost would be required. Due to the nature of the logistic transformation, the predicted probability never reaches exactly 0 or 1 for any finite value of a predictor, i.e. it only gets closer and closer as the predictor becomes extremely large or small.

(5, 2pts) Holding all other predictors constant and for a given increase in cost, the model suggests that men are more likely than women to say they would *not* still go fishing. This indicates that men are more price-sensitive than women. Specifically, the odds ratio is $\exp(1.003908) = 2.728926$, meaning that men are about 2.7 times more likely than women to choose not to fish in response to a cost increase.

It would be incorrect to conclude that "women are more likely to fish than men" based on this result. In fact, more men than women fish overall. What the analysis shows is that among those who fish, men are more likely than women to stop fishing if the price increases.

(6, 2pts) No, overdispersion is not a major concern in this case because the data is sparse. With only 196 anglers and many predictors — some of which have multiple quantitative levels — there is likely only one observation, or at most a few, for each combination of predictor levels. Overdispersion typically arises in grouped data with multiple observations per predictor combination. In sparse data settings like this, overdispersion is less likely to occur, and the standard binomial variance assumption is generally appropriate.

(7, 2pts) A better way to assess the statistical significance of the North vs. South Carolina effect is to use a deviance-based approach. This approach is generally more reliable than using the z-value from the analysis output, especially in small-sample or sparse-data settings. In such cases, the Hauck–Donner effect can lead to an overestimation of the standard error in the denominator of the z-statistic, resulting in less accurate inference.

(8, 1pt) A Poisson distribution was used because the response variable Fatalities is a count variable with no obvious upper bound.

(9, 2pts) The degrees of freedom associated with the residual deviance are calculated as follows: $(1997 - 1967 + 1) - 1 - 1 = 29$.The residual deviance of 336.72 is very large relative to a chi-squared distribution with 29 degrees of freedom, so we reject the hypothesis that Model 1 provides an adequate fit to the data.

(10, 3pts) (i) Include additional informative predictors or explanatory terms, such as a quadratic effect of Year, to better capture systematic variation in the mean structure of the response. (ii) Modify the variance structure to account for overdispersion, for example by introducing a dispersion parameter σ^2 such that $\mathrm{Var}(Y_x)=\sigma^2\times\mathrm{E}(Y_x)$. (iii) Consider alternative distributions for the response variable, such as the negative binomial distribution, which is more suitable for handling overdispersed count data.

(11, 1pt) The Pearson chi-squared statistic is the sum of the squared Pearson residuals and, under a good model fit, is expected to be roughly equal to the number of degrees of freedom (i.e., 29). This implies that most Pearson residuals should be approximately 1 in absolute value. While some variation is expected, a residual with an absolute value of 13.233 is exceptionally large in this context and suggests a poor fit for that observation.

(12, 3pts) To test the coefficient of Year using a deviance-based approach, the null model includes only the intercept, while the alternative model is the full model (Accidents $\sim$ Year). Therefore, the deviances corresponding to the null and alternative models are 38.740 (null deviance in the output) and 27.565 (residual deviance in the output), respectively. The difference in deviance is 11.175, which is large relative to a chi-squared distribution with 1 degree of freedom. Thus, the deviance-based test rejects the null hypothesis that Year has no effect. This conclusion is consistent with the result from the Wald test.

Since the residual deviance of 27.565 is not large compared to its associated degrees of freedom (29), there is no strong evidence of overdispersion. As a result, the use of the chi-squared test (rather than an F-test) is appropriate in this context.

(13, 1pt) Since this is a log-linear model, the estimated coefficient of Year, -0.04415, can be interpreted as follows: for each additional year, the expected number of train accidents decreases multiplicatively by a factor of $\exp(-0.04415) = 0.9568$. This corresponds to an approximate 4.32% decrease in the expected number of train accidents per year.

(14, 1pt) From Figure 1, the accident rate in 1967 appears to be approximately 15 accidents per billion kilometers. Given that there were 7 accidents in that year, we estimate the total distance traveled to be $7/15 = 0.4666667$ billion kilometers.

(15, 2pts) An offset of the form log(Train.km) was used in Model 3 because this is a rate model, where the response variable is the number of accidents and Train.km serves as the size variable. The rate of accidents is modeled as log(E(Accidents)/Train.km) $= \beta_0 + \beta_1\times$Year. Rearranging this gives log(E(Accidents)) $= \beta_0 + \beta_1\times$Year + log(Train.km), which defines the structure of a Poisson GLM with a log link and an offset term log(Train.km) to adjust for the size variable.

(16, 1pt) The interpretation is similar to that in Problem 13, except that here we are modeling the rate, i.e., the expected number of train accidents per billion kilometers traveled, rather than the expected count. For each additional year, the accident rate decreases multiplicatively by a factor of $\exp(-0.04145) = 0.9594$. This corresponds to an approximate 4.06% decrease in the accident rate per year.

(17, 1 pt) A curve — because the fitted rate, i.e., fitted value of E(Accidents)/Train.km, is given by $\exp(\hat{\beta}_0 + \hat{\beta}_1 \times \text{Year})$, which is an exponential function of Year.

(18, 1 pt) A possible reason the estimated coefficients for Year in Models 2 (a log-linear model) and 3 (a rate model) are similar is that the size variable, Train.km, is nearly constant over time. When switching from a log-linear model to a rate model, the coefficient for Year remains similar if the size variable does not vary much. This is plausible in this case, as the amount of railway traffic in a country is likely to remain relatively stable from year to year.