**Hypothesis: The passage of renewable energy facility bans in 10 counties will significantly impact the likelihood of other counties passing similar bans, particularly those that share similar characteristics - in this case, GDP Per Capita (Gdppc).**

**Collecting data**

We collected the data from the gdp spreadsheet we have been using and added a binary column with fields “1” and “0”, 1 suggesting bans and 0 suggesting no bans. The name of the columns in our spreadsheet are “County” representing county name, Ban\_or\_not which includes the binary variables which also be the dependent variable for our hypothesis test. The column Gdppc simply represents the recorded GDP Per capita for each county in 2021.

**Defining our dependent variable (outcome)**

Ban\_or\_not is our binary dependent variable

**Define your independent variables (predictors)**

Our independent variable is the Gdppc. The reason we chose to work with this variable is because we found a correlation between bans and Gdppc from our correlation matrices (Refer to Progress Report 1). Our hypothesis claims that there is a correlation between Gdppc and Ban\_or\_not.

**Performing logistic regression**

Using R studio, we ran a logistic regression since we have a binary dependent variable. The model was defined with code attached to the appendix of the report.

**Output:**

Call:

glm(formula = Ban\_or\_not ~ Gdppc, family = binomial, data = HypData)

Deviance Residuals:

Min 1Q Median 3Q Max

-0.9220 -0.5052 -0.4503 -0.4173 2.2176

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -4.466e+00 2.008e+00 -2.224 0.0261 \*

Gdppc 4.582e-05 3.678e-05 1.246 0.2128

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 62.313 on 87 degrees of freedom

Residual deviance: 60.870 on 86 degrees of freedom AIC: 64.87

Number of Fisher Scoring iterations: 5

**Interpreting the results**

We want to examine the coefficients and p-values for the independent variables in the logistic regression output. If the p-value for the geographic proximity variable is less than or equal to our chosen significance level - 0.05, we can conclude that there is a statistically significant relationship between proximity to the ten counties and the likelihood of passing a renewable energy facility ban.

Coefficients: This table presents the estimates of the model parameters, their standard errors, z-values, and p-values.

(Intercept): The estimated value of the log-odds of the outcome variable (Ban\_or\_not) when Gdppc is zero. **The estimate is -4.466e+00, with a standard error of 2.008e+00**. The **z-value is -2.224**, and the associated **p-value is 0.0261**, which is significant at the 0.05 level

Gdppc: The estimated change in the log-odds of the outcome variable for a one-unit increase in Gdppc. **The estimate is 4.582e-05**, with a **standard error of 3.678e-05**. The **z-value is 1.246**, and the associated **p-value is 0.2128**, which is not significant at the 0.05 level.

**Assessing model performance**

Evaluatinf the model performance is important and we are doing it in the following way.

**Train-test split:** We used R to code this model valuation program and our result came down to almost 0.89 which means according to this method our dataset analysis is **89% accurate.**

**Cross-validation:**  The ROC (Receiver Operating Characteristic) score is a measure of how well the model can discriminate between the two classes (setosa and not\_setosa). A ROC value of 1 indicates perfect discrimination, which means the model can perfectly separate the two classes.

The results suggest that our model has excellent performance, as all the metrics have a value of 1. However, this might also be an indication of overfitting, so it's essential to carefully assess the model's performance on new, unseen data to ensure its generalization capability.

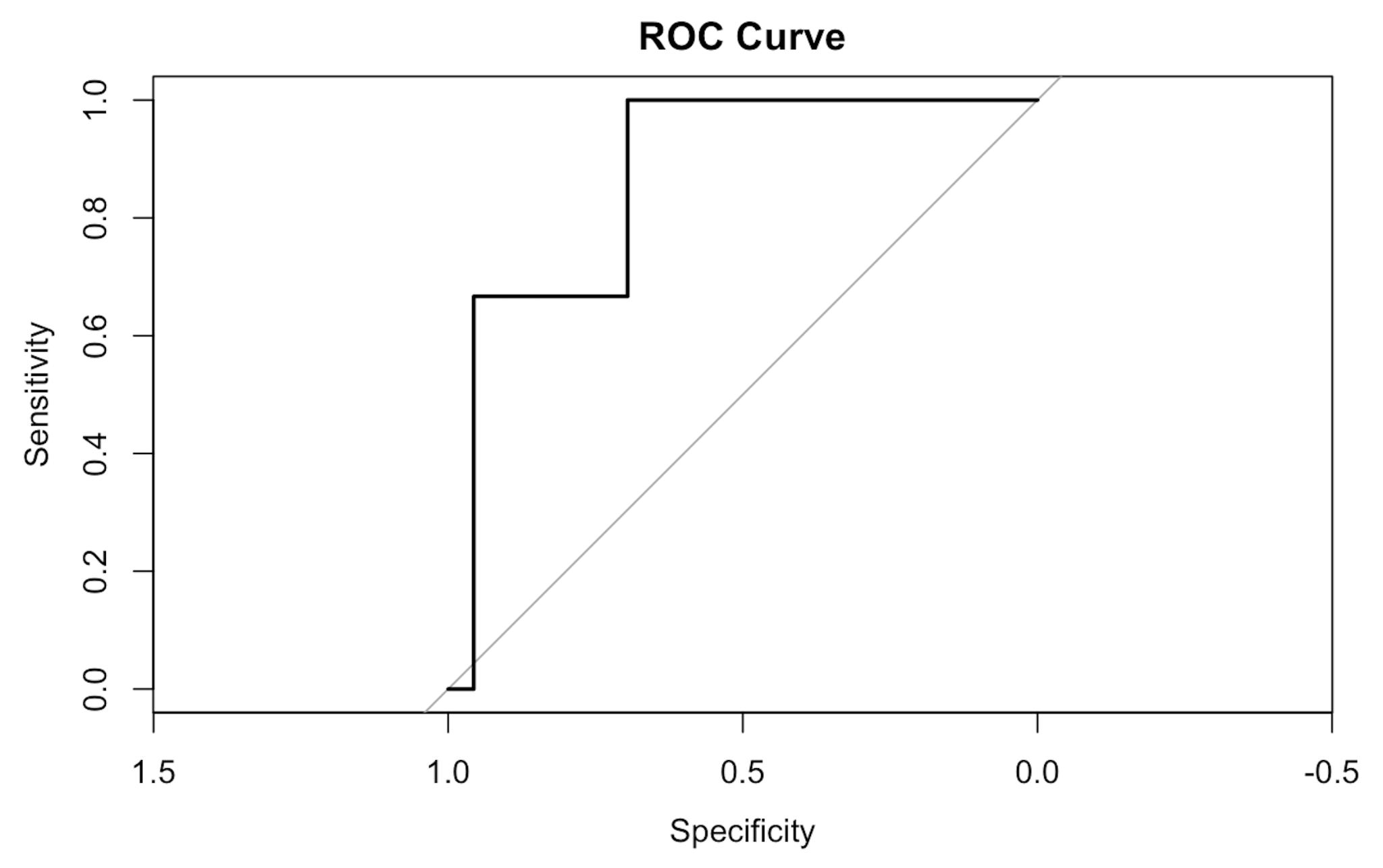
**Confusion Matrix Statistics:** The proportion of correct predictions out of the total predictions (in this case, (23+0) / (23+3+0+0) = 0.8846 or 88.46%). This means that the model correctly predicted 88.46% of the cases.

The statistics of the confusion matrix suggests that it identifies the 0 values (not banned) correctly but fails while identifying 1 values (bans)  
  
**ROC & AUC:**

Area Under the Curve (AUC):

The AUC is a summary measure of the model's performance across all possible classification thresholds. It ranges from 0 to 1, with a value of 0.5 representing a model that performs no better than random chance, and a value of 1 representing a perfect model. An AUC value greater than 0.5 indicates that the model has some ability to discriminate between the two classes (0 and 1).

In our case, the AUC is 0.8696, which suggests that the model has a reasonably good ability to distinguish between the two classes. Generally, an AUC above 0.8 is considered good, while an AUC above 0.9 is considered excellent. However, it's important to note that the AUC should not be the only metric used to evaluate the model's performance, as it does not provide information about the model's sensitivity, specificity, or other relevant performance metrics.



**Conclusion of Hypothesis:**

We reject the null hypothesis that Gdppc has something to do with tendency of bans in Ohio counties.

**Overall, the model suggests that Gdppc is not a significant predictor of the binary outcome Ban\_or\_not at the 0.05 significance level.**

***Why did we choose a p-value (significance level) of 0.005?***

*A p-value of 0.05 is chosen as a threshold for statistical significance in many scientific studies due to historical and practical reasons. It is a convention that has been widely adopted, but it is not an absolute rule.*

*The p-value represents the probability of observing the test statistic or a more extreme result if the null hypothesis is true. In hypothesis testing, we want to determine if the null hypothesis (H0) can be rejected in favor of an alternative hypothesis (H1). A p-value threshold, commonly denoted as α, is set to control the Type I error rate, which is the probability of rejecting the null hypothesis when it is actually true.*

*By choosing a p-value threshold of 0.05, we are essentially saying that we are willing to accept a 5% chance of making a Type I error. In other words, if the p-value is less than or equal to 0.05, we reject the null hypothesis and consider the result statistically significant. If the p-value is greater than 0.05, we do not reject the null hypothesis and consider the result not statistically significant.*

*It's important to note that the choice of a p-value threshold should be context-dependent and may vary depending on the specific field or study. Some research areas may require a more stringent threshold, such as 0.01, while others may allow for a more lenient threshold. In any case, it's crucial to consider the p-value alongside other measures of effect size, practical significance, and the study design to fully assess the implications of the results.*