MET CS 555 Term Project

❖ Research scenario and research question

The scenario revolves around analyzing the "Census Income" Data set, originally extracted by Barry Becker from the 1994 Census database. This data set contains various demographic and employment-related information of individuals, such as their income, age, workclass, education, marital status, among others. The primary goal of this analysis is to predict whether an individual's income exceeds $50K/year, leveraging the dataset's comprehensive attributes.

Given the initial scenario and data set information, 3 specific research questions emerges. The questions are follows:**1) Is there a significant difference in the proportion of individuals earning over $50K/year between men and women? 2) Does sex significantly predict the likelihood of having an income level above $50K? 3)How do sex and age together predict the likelihood of having an income level above $50K?**

❖ Describe the data set

Main data source: https://archive.ics.uci.edu/dataset/2/adult

After importing the original data set into RStudio, the focus was narrowed down to three variables: sex and income. Age(name "V1" in the data set): a continuous variable. Sex(name "V10" in the data set）: It is a binary variable, categorized as Male or Female. Income(name "V15" in the data set）: This variable represents whether an individual's annual income exceeds $50K. It is also categorical, with values indicating either ">50K" (more than $50,000) or "<=50K" (less than or equal to $50,000).

The sex variable was encoded as binary (1 for Male, 0 for Female), and the income variable was similarly encoded (1 for incomes over $50K, 0 for incomes at or below $50K).

Due to the large size of the original dataset (32,561 entries), a random sample of 1,000 entries was selected.

Summarize the data relating to Income by sex:

Population	Population Description	Sample Size	Count of high Income	Sample Proportion
1	male	n1=678	202	$\hat{p_1}$=202/678=0.298
2	female	n2=322	30	$\hat{p_2}$=30/322=0.093

❖ Describe the statistical methods

1) Is there a significant difference in the proportion of individuals earning over $50K/year between men and women?

A two-sample test for proportions is conducted, specifically a two-proportion z-test.

Formally test (at the **alpha=0.05** level) whether the proportion of people with higher Income (over $50K) is the same across men and women based on this effect measure.

Calculate the risk difference: 0.298-0.093=0.205

1) State your hypotheses and alpha level
H0: . $P_1 = P_2$ (The proportion of males with a high income is equal to the proportion of females with a high income.)
H1: $P_1 \neq P_2$
alpha level=0.05
2) State your test statistic
Based on the null hypothesis that P1=P2

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

3) State your decision rule
If p-value < 0.05, we reject the null hypothesis.
If p-value >= 0.05, we fail to reject the null hypothesis.
4) Perform your test and report your test statistic value, df, and p-value.
X-squared = 51.376, df = 1, p-value = 7.628e-13
5) State your conclusion
P-value=7.628e-13 here is below alpha level of 0.05, we reject the null hypothesis. The male and female group are not equal in their proportion of people with have high Income.

Result: The results indicated a significant difference between genders, with 29.8% of men and 9.3% of women earning above this threshold. The risk difference of 20.5 percentage points and the extremely low p-value (much less than the 0.05 alpha level) provide strong evidence against the null hypothesis of equal proportions. Therefore, we conclude that there is a statistically significant difference in the proportion of high earners between the male and female populations in this sample.

2) Does sex significantly predict the likelihood of having an income level above $50K?

A logistic regression with sex as the only explanatory variable is conducted.

1) State your hypotheses and alpha level
H0: $\beta = 0$
H1: $\beta \neq 0$
alpha level=0.05
2) State your test statistic (give us the formula)

$$z = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

3) State your decision rule
If p-value < 0.05, we reject the null hypothesis.
If p-value >= 0.05, we fail to reject the null hypothesis.
4) Perform your test.
Z-value = -6.153 P-value=7.63e-10
5) State your conclusion
P value = 7.63e-10 here is below alpha level of 0 point 0.05, we reject the null hypothesis.
Sex is associated with odds of having an income level above $50K. Females have 0.3062358 times the odds of having an income level above $50K compared to males.
We are 95% confidence that the true odds ratio for sex is between 0.2100491 and 0.4464505.
The c-statistic for this model is 0.6142.

Result: Pvalue = 7.63e-10 suggests that there is a statistically significant association between sex and the odds of having an income level above $50K. The odds ratio of 0.3062358 indicates that females have approximately 30.62% of the odds of having a high income compared to males, suggesting that males are more likely to have a high income than females. The 95% confidence interval for the odds ratio ranges from 0.2100491 to 0.4464505, which does not include 1. This reinforces the conclusion that the odds of females having a high income are significantly different from males.
The c-statistic for this model is 0.6142. A c-statistic of 0.5 suggests no discrimination (i.e., the model is no better than random chance at predicting the outcome), whereas a c-statistic of 1 indicates perfect discrimination. A value of 0.6142 indicates that the model has a modest ability to discriminate between those who do and do not have a high income based on sex.

3)How do sex and age together predict the likelihood of having an income level above $50K?

A multiple logistic regression is conducted to predict Income level from sex and age.

Z_{sex}=-5.513, P_{sex}=3.53e-08, Z_{age} =5.52, P_{age} =3.24e-08.

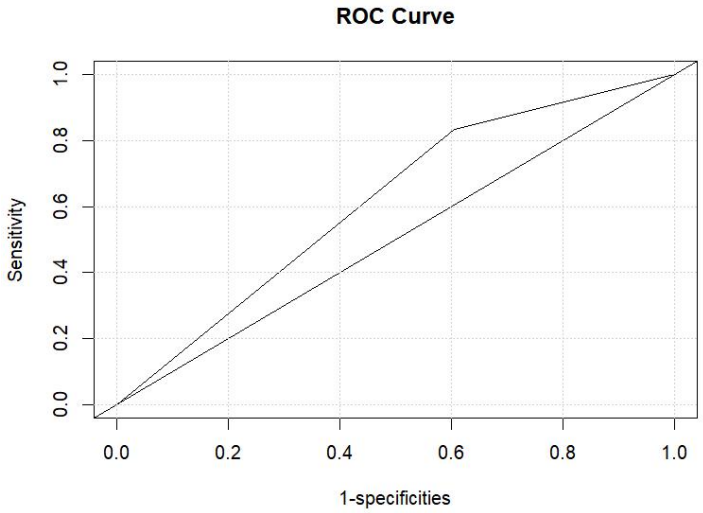
Both p-values are significantly less than the alpha level of 0.05. This indicates that we have statistically significant evidence to reject the null hypothesis for both

Sex and Age. In other words, both Sex and Age are statistically significantly associated with the likelihood of an income level above $50K.

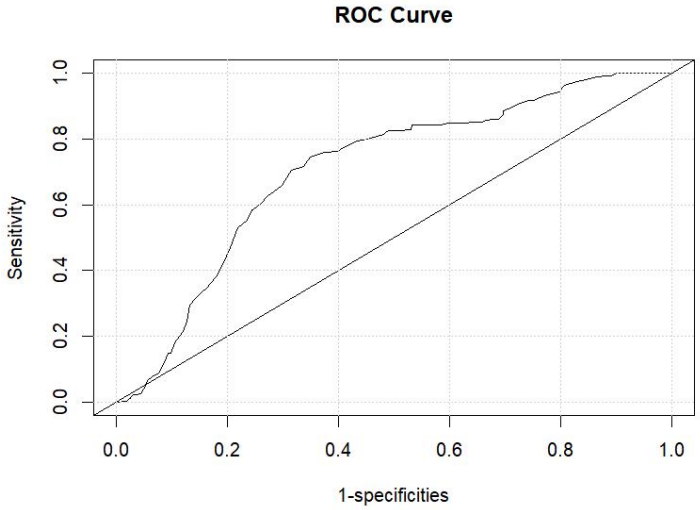
Females have 0.341049 times the odds of having an income level above $50K compared to males. For each one-unit increase in Age, the odds of having an income level above $50K increase by about 3.18%.

The c-statistic for this model is 0.7064.

ROC in research question 2:



ROC in research question 3:



Model with Sex as the Only Predictor:

The analysis shows that there's a significant difference between men and women when it comes to the odds of earning more than $50,000 a year. Specifically, women have about 30.62% of the odds of earning over $50K compared to men, indicating

that men are significantly more likely to have a high income than women.
The model's ability to differentiate between those who do and do not earn over $50K based on sex alone is modest, with a c-statistic of 0.6142. This suggests that while sex is a factor, it's not a strong predictor on its own.

Model with Both Sex and Age:

When we include both sex and age in the analysis, we find that both factors are significantly associated with income level. Women still have lower odds (about 34.10% of the odds compared to men) of having a high income, but age also plays a role. Specifically, for each additional year of age, the odds of earning over $50K increase by approximately 3.18%.
The model's ability to distinguish between individuals earning over $50K improves with the inclusion of age, as indicated by a higher c-statistic of 0.7064. This means the model is better at predicting who earns over $50K when considering both a person's sex and their age.

Which Model Does Better?

The model that includes both sex and age as predictors does a better job of identifying individuals with a high income. This is evident from the higher c-statistic (0.7064 compared to 0.6142), which indicates a better predictive ability. By considering age along with sex, we get a clearer picture of the factors that contribute to higher earnings.

- Conclusion and limitation

In our investigation into what affects people's chances of earning more than $50,000 a year, we found two important things. First, being a man significantly increases your chances compared to being a woman. Second, as people get older, their chances of earning more than $50K also increase, regardless of whether they are a man or a woman.

The models identify associations between sex, age, and high income but do not establish causality. There could be underlying factors not captured by these variables that influence the likelihood of earning over $50K/year. For example, education, work experience, industry, and geographic location can significantly impact income levels but are not included in the analysis.