Logistic Model

For

Determining Income Range Based on Certain Demographics

My Goal

For this analysis, I selected a dataset from the UC Irvine Machine Learning Repository, which contains demographic information such as gender, race, income, age, hours worked per week, years of education, marital status, and work class. My goal is to predict whether an adult makes more than $50,000 in income based on a set of predictors, including hours worked per week, years of education, gender, and age, using multiple logistic regression.

I chose these particular predictors because they are straightforward and have clear relationships with income. For instance, it is reasonable to expect that age and education (as measured by years) might influence income, with older individuals or those with more education potentially earning higher wages. Similarly, the number of hours worked per week could directly affect income, as more hours worked often correspond to higher earnings.

While marital status and the highest level of education attained could also potentially influence income, I chose not to include them in this model. The marital status variable, for example, contains many classes (e.g., married, divorced, single), which would require complex encoding techniques. I could have reduced it to a binary feature (e.g., married vs. unmarried), but I felt this might oversimplify the data. Similarly, while the highest level of education attained might be an important factor, I already included years of education as a predictor, which captures much of the relevant information. Including both could introduce redundancy and make the model unnecessarily complex.

By focusing on the variables I selected, I aim to create a more streamlined and interpretable model while still testing if these key predictors are sufficient to predict income. I will first check the data to ensure a logistic regression model is appropriate. Then, I will assess the model's performance, including testing its predictive accuracy and calculating confidence intervals for the model's coefficients. If the model performs well, I'll use it to predict whether certain individuals, including myself, are likely to earn more than $50,000.

Checking Conditions

To assess if a logistic model is a good fit for my data, I will have to check three things: Randomness, Independence, and Linearity.

Randomness

To assess randomness in the data, we need to examine whether the outcomes of the variable we're predicting—whether an adult makes more than $50,000—appear random. In this dataset, it is assumed that we have a random sample of adults, as the data is derived from the US Census. However, we must note that the predictors we're using—age, years of education, gender, and hours worked per week—are not randomly assigned to each individual. For example, factors like gender and years of education are inherent characteristics, while hours worked per week could vary depending on an individual's job but are not strictly random (although there might be some randomness in the distribution of work hours depending on the employer). The core question is: Does an adult's income of $50K or more behave as if it were determined by a random process, like a spinner? In other words, does randomness play a significant role in determining income outcomes, or are there more predictable factors at work? I believe that factors such as gender and years of education strongly influence the likelihood of earning more than $50K, so I don't think the outcome is purely random. However, since hours worked per week might involve an element of randomness (depending on how work hours are assigned), some degree of randomness could be involved. Taking a more strict view, I would say randomness does apply to some extent, but with reservations. In a less strict sense, it's difficult to confirm whether randomness is a major factor. Ultimately, I will be able to draw more concrete conclusions after applying the multiple logistic regression model. Regardless of this uncertainty,

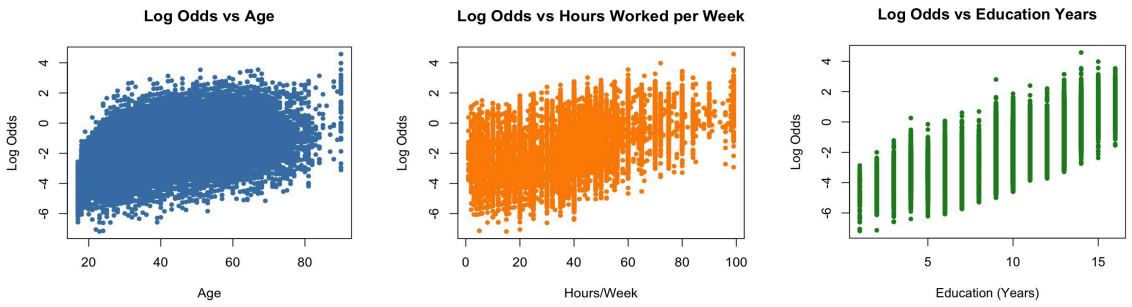
my primary goal remains to determine whether age, gender, years of education, and hours worked per week are strong predictors of whether an adult earns more than $50K annually.

Independence

One way independence can fail is if there are lurking variables that we haven't accounted for. In my dataset, marital status and relationship (e.g., whether the person is an only child, husband, wife, etc.) are included. For example, let's say one of my observations comes from a family with children, which could influence the number of hours worked per week. If two of my observations are from married individuals, one might think that marital status could affect the likelihood of making more than $50K in income. However, I believe marital status doesn't significantly impact the income outcome in this case, as I think income is primarily based on job earnings rather than other factors like family dynamics. Therefore, I would argue that marital status or relationship should not have a major effect on predicting income in this context.

Linearity

To check for linearity, we need to check the plots for log odds of income vs continuous variables. Below are the plots:



Off the bat we can see that log odds vs age isn't that linear so we might have to do a transformation. This might be a stretch but the trend seems to be logarithmic or square root, so I might transform age to log age and hope for linearity. Log odds versus hours worked per week

follows a somewhat linear trend the only thing that concerns me is that there are fewer points in the upper right corner so there might not be constant variance, but overall there doesn't seem to be that many curves. Log odds vs years of education is interesting. The only time I've seen straight lines like this is for categorical values. Although education years are discrete and not continuous, that could explain the vertical lines. But honestly, the model could be treated as education years as categorical, so it is worth seeing how the model would look like without education years as a predictor but as for now, I will keep education years since it is quantitative still. In conclusion, all three scatter plots follow a steady linear trend.

Multicollinearity

When working with multiple predictors, we want to make sure they aren't highly correlated with each other, because high correlation can cause problems in the model. It can make it difficult to determine each variable's unique contribution to the prediction. To check for this, we can use a correlation matrix. A correlation matrix shows the correlation coefficients between each pair of predictors. If we see a value close to 1 (or -1, if there's a negative relationship), it means those two predictors are highly correlated. This could cause issues in the model, so it's important to identify and address any strong correlations.

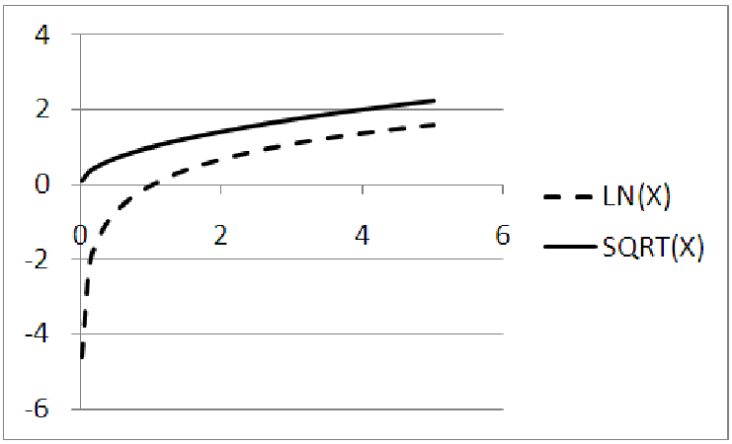
---

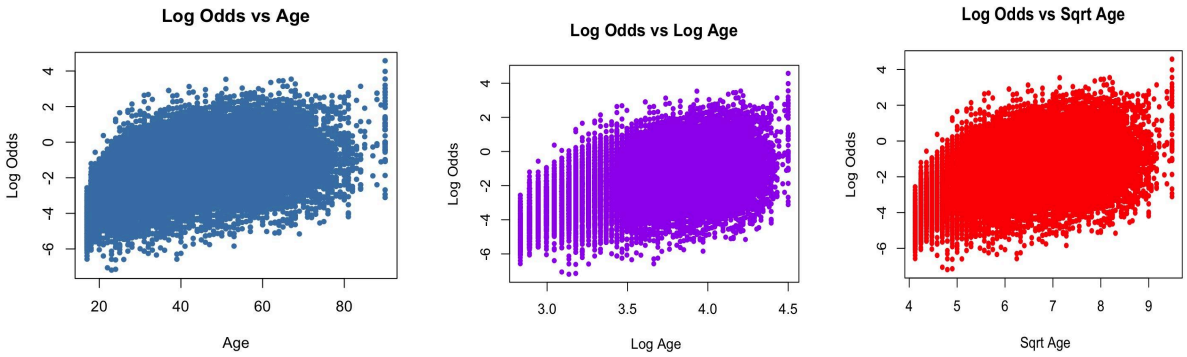
	age	eduyrs	gender	hours
age	1.00000000	0.030940376	0.088120022	0.07155834
eduyrs	0.03094038	1.000000000	0.009328018	0.14368891
gender	0.08812002	0.009328018	1.000000000	0.22855980
hours	0.07155834	0.143688909	0.228559799	1.00000000

Here we have no numbers correlated to each other, except when paired with themselves which is expected. So we don't have to worry about multicollinearity! Hooray! None of the variables have to go.

Transformation of Age

As stated above, we will transform the age predictor with a log transformation hoping it gives us a more linear trend. We will also do a square root transformation since both of those equations usually have a steady slow increase as shown below:

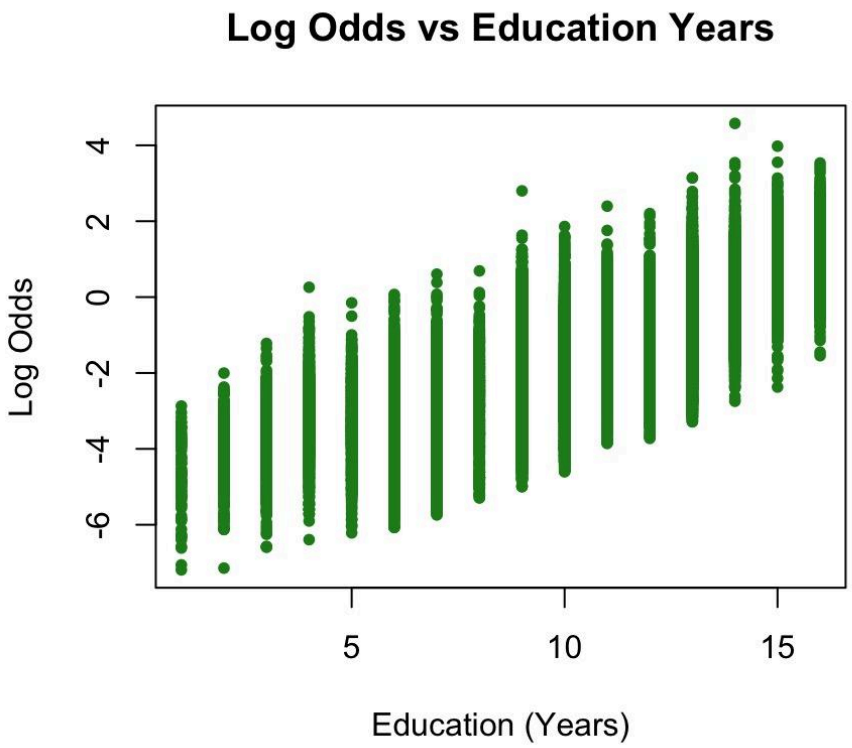




After doing the transformations, there doesn't seem to be much of a difference, so why fix something that's not broken, or why fix something that doesn't get better? I assess there is no need for a transformation of the age variable since there doesn't seem to be much of a change in scatterplots.

Comparing Models with AIC

One predictor I was skeptical about was education years. It gave a scatterplot like this:



It behaves like a categorical variable so let's see if there may be a reason to omit this variable. Why omit it? Well as stated in the beginning it would be difficult to hot encode if it is categorical since it would have so many classes.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.3198756  0.0638174  -83.36   <2e-16 ***
age          0.0412770  0.0008663   47.65   <2e-16 ***
gender       1.0481887  0.0291639   35.94   <2e-16 ***
hours        0.0404278  0.0009961   40.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 53751  on 48841  degrees of freedom
Residual deviance: 47155  on 48838  degrees of freedom
AIC: 47163
```

```
Call:
glm(formula = income ~ age + eduyrs + gender + hours, family = "binomial")

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.1097328  0.0945176  -96.38   <2e-16 ***
age          0.0449637  0.0009642   46.63   <2e-16 ***
eduyrs       0.3554416  0.0054058   65.75   <2e-16 ***
gender       1.1619445  0.0307122   37.83   <2e-16 ***
hours        0.0353741  0.0010481   33.75   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 53751  on 48841  degrees of freedom
Residual deviance: 41855  on 48837  degrees of freedom
AIC: 41865

Number of Fisher Scoring iterations: 5
```

As a rule of thumb, AIC, or Akaike Information Criterion, helps determine which model best fits the data by balancing goodness of fit with model complexity, with a lower value indicating a better model. And would you look at that, removing the years of education increases our AIC from 41865 to 47163! So it was not a mistake to keep years of education.

Assessing Fit of the Model

We know that our model is appropriate for our data but how well does it work? Luckily we can use R Studio to quickly assess.

```
Call:
glm(formula = income ~ age + eduyrs + gender + hours, family = "binomial")

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.1097328  0.0945176  -96.38   <2e-16 ***
age          0.0449637  0.0009642   46.63   <2e-16 ***
eduyrs       0.3554416  0.0054058   65.75   <2e-16 ***
gender       1.1619445  0.0307122   37.83   <2e-16 ***
hours        0.0353741  0.0010481   33.75   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 53751  on 48841  degrees of freedom
Residual deviance: 41855  on 48837  degrees of freedom
AIC: 41865

Number of Fisher Scoring iterations: 5
```

Based on the p-values for each coefficient, according to the model each coefficient is significant to the model for predicting if an adult earns more than $50K in income. The AIC is large but we mainly use the Akaike Information Criterion to compare to two models as we did above. The null deviance of 53,751 represents the error when predicting income using no predictors, while the residual deviance of 41,855 shows the error after including age, education years, gender, and hours worked per week. The substantial drop in deviance indicates that these predictors significantly improve the model’s ability to explain income levels.

Making Predictions

Since our model seems to be working well let's have it make some predictions. Let's use our equation for predicting log odds and probability of earning more than $50K in income:

$$log\left(\frac{P(income > \$50K)}{1 - P(income > \$50K)}\right) = -9.11 + 0.045 \cdot age + 0.355 \cdot eduyrs + 1.162 \cdot gender\_Male + 0.035 \cdot hours$$

$$\pi = \frac{e^{-9.11+0.045 \cdot age+0.355 \cdot eduyrs+1.162 \cdot gender\_Male+0.035 \cdot hours}}{1 + e^{-9.11+0.045 \cdot age+0.355 \cdot eduyrs+1.162 \cdot gender\_Male+0.035 \cdot hours}}$$

* Here pi represents the probability of the adult earning more than $50K in income.

I'm going to have some fun and use my information to see if the model can predict whether I make more than $50K in income (spoiler alert: I don't, so we'll see how well the model reflects real-life data). I'm 22 years old, have 18 years of education, identify as male (so my gender is coded as 1), and work 12 hours per week. Using Desmos, I calculated a probability of 0.4631, meaning the model predicts there's a 46% chance that I earn more than $50K. Since this is less than 50%, the model would classify me as earning less than or equal to $50K—so close! This result seems reasonable, especially considering my years of education. Many adults in the dataset likely don't have as much education, so perhaps the model gives more weight to high school education and beyond, rather than counting everything from preschool onward.
So not counting elementary and middle school, that leaves me with 8 years of education, so then my probability drops to 0.0241741, so the model predicts there's a 2.4% chance of my income being $50K or greater! Ouch!

Confidence Interval

Suppose we don't trust the small p-values of the coefficients, we don't believe the coefficient for education years is NOT zero. After all, the log odds vs years of education gave us

a weird scatterplot so we are skeptical. So let's do a confidence interval that will give us an estimated range of what the coefficient could be using the formula from the textbook below.

> We can also compute a confidence interval for the slope using
>
> $$\widehat{\beta}_1 \pm z^* \cdot SE_{\widehat{\beta}_1}$$
>
> where z^* is found using the normal distribution and the desired level of confidence.

Our estimated coefficient for years of education in predicting log odds of income for an adult being more than $50K is 0.3554416. That means for every increase in year of education, there is a 0.355 increase in log odds that the adult earns more than $50K. But what are log odds? Log odds are just a way of expressing how likely something is to happen — like a score the model gives based on a person's age, education, gender, and hours worked — where higher scores mean a greater chance of earning more than $50K, and lower scores mean a smaller chance. Let our confidence interval be 95% so our z^* will be 1.96. Checking our summary, the standard error is 0.0054058. So our confidence interval will be **(0.344846232, 0.366036968)**. We are 95% confident that each additional year of education increases the log odds of earning more than $50K by between 0.3448 and 0.3660. Since the entire interval is above 0, it supports that education is a significant positive predictor in your model.

Summary

In this analysis, we applied a multiple logistic regression model to predict whether an adult earns more than $50K in income based on four key predictors: age, gender, hours worked per week, and years of education. After checking the necessary assumptions, including the null deviance, residual deviance, and p-values of the coefficients, the model performed well, indicating that these predictors are strong enough to provide reliable income predictions. The model was able to successfully classify individuals based on their characteristics, as demonstrated by its accurate prediction for a 22-year-old male college senior working 12 hours per week, who the model correctly predicted would earn less than $50K. Overall, the results

show that age, gender, hours worked per week, and years of education are useful determinants of income in this dataset.

Appendix/Sources

Where the dataset came from: https://archive.ics.uci.edu/dataset/2/adult

Code

Data Cleaning:

```
In [14]:import pandas as pd
         from sklearn.preprocessing import OneHotEncoder
In [6]:df = pd.read_csv ("adult.csv")
In [8]:df.head(5)
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	ho
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0	
4	18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	0	

```
In [18]:df_encoded = pd.get_dummies(df, columns=['income', 'gender'], drop_fi
In [22]:df_encoded.head()
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	capital-gain	capital-loss	hours-per-week	n co
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	0	0	40	U
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	0	0	50	U
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	0	0	40	U
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	7688	0	40	U
4	18	?	103497	Some-college	10	Never-married	?	Own-child	White	0	0	30	U

```
In [28]:df_encoded['gender_Male'] = df_encoded['gender_Male'].astype(int)
In [30]:df_encoded['income_>50K'] = df_encoded['income_>50K'].astype(int)
In [32]:df_encoded.head(5)
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	capital-gain	capital-loss	hours-per-week	n co
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	0	0	40	U
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	0	0	50	U
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	0	0	40	U
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	7688	0	40	U
4	18	?	103497	Some-college	10	Never-married	?	Own-child	White	0	0	30	U

```
In [36]:df_encoded.to_csv('fil_adult.csv', index=False)
In [38]:df_encoded[['age', 'educational-num', 'hours-per-week', 'income_>50K',
```

	age	educational-num	hours-per-week	income_>50K	gender_Male
0	25	7	40	0	1
1	38	9	50	0	1
2	28	12	40	1	1
3	44	10	40	1	1
4	18	10	30	0	0
...	...	...	...	...	...
48837	27	12	38	0	0
48838	40	9	40	1	1
48839	58	9	40	0	0
48840	22	9	20	0	1
48841	52	9	40	1	0

48842 rows × 5 columns

R Studio Code:

```
adults = read.csv('fil_adult.csv')
View(adults)
age = adults$age
eduyrs = adults$educational.num
gender = adults$gender_Male
hours = adults$hours.per.week
income = adults$income_.50K

#fitting model
LogModel = glm(income~age+eduyrs+gender+hours, family = "binomial")
summary(LogModel)

#log odds
adults$prob <- predict(LogModel, type = "response")
adults$logit <- log(adults$prob / (1 - adults$prob))

# Age vs Logit
plot(age, adults$logit,
     main = "Log Odds vs Age",
     xlab = "Age",
     ylab = "Log Odds",
     pch = 20, col = "steelblue")

# Education Years vs Logit
plot(eduyrs, adults$logit,
     main = "Log Odds vs Education Years",
     xlab = "Education (Years)",
     ylab = "Log Odds",
     pch = 20, col = "forestgreen")

# Hours per Week vs Logit
plot(hours, adults$logit,
     main = "Log Odds vs Hours Worked per Week",
     xlab = "Hours/Week",
     ylab = "Log Odds",
     pch = 20, col = "darkorange")

#matrix
predictors <- data.frame(
  age = adults$age,
  eduyrs = adults$educational.num,
  gender = adults$gender_Male,
```

```
  hours = adults$hours.per.week
)

# Compute the correlation matrix
cor(predictors)

#fitting with transf
LogModel = glm(income ~ log(age) + eduyrs + gender + hours, family = "binomial")
plot(sqrt(age), adults$logit,
     main = "Log Odds vs Sqrt Age",
     xlab = " Sqrt Age",
     ylab = "Log Odds",
     pch = 20, col = "red")
summary(LogModel)

#model w/o yrs edu
LogModel2 = glm(income~age+gender+hours, family = "binomial")
summary(LogModel2)
```