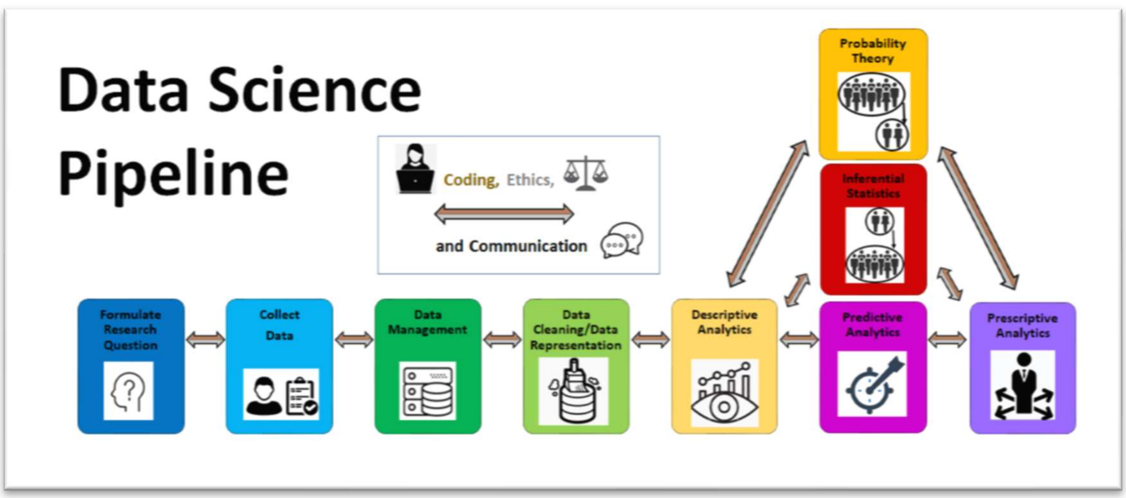
Unit 11: Logistic Regression – Part 1

Case Studies:

- To introduce the concept of **simple logistic regression**, we will examine the association between sex and approval for the direction the country is going in (in 2017)?
  - Response: Approve/disapprove of the direction
  - Explanatory: sex

- To introduce the concept of **multiple logistic regression**, we will examine the association between sex and age and approval for the direction the country is going in (in 2017)?
  - Response: Approve/disapprove of the direction
  - Explanatory(s):
    - Sex
    - Age



Summary of Concepts:

1. **Analyses for Associations**
2. **Association Analyses Summary:** (Numerical(s) and Categorical(s) Explanatory Variable(s)-> Categorical Response Variable (with 2 levels)
3. **Why should we not use linear regression to model a categorical response variable?**
4. **What curve can we fit instead when we have a categorical response variable with 2 levels?**
5. **Odds vs. Probability**
6. **Fitting a Simple Logistic Regression Model**
7. **Fitting a Multiple Logistic Regression Model**
8. **Making predictions with a logistic regression equation.**
9. **Interpreting $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots \hat{\beta}_p$ in a Logistic Regression Model**

Additional Resources:

Section 8.4 in Diez, Barr, and Cetinkaya-Rundel, (2015), *OpenIntro Statistics* https://www.openintro.org/download.php?file=os3&redirect=/stat/textbook/os3.php

1. Analyses for Associations

Questions to consider, when selecting an analysis to test an association.

1. Which variable is the **<u>response variable</u>** in this association?
    a. Is it a **categorical** or **numerical** variable?
    b. If it's a categorical variable, **how many levels** does it have?

2. Which variable(s) is the **<u>explanatory variable</u>** in this association?
    a. Is it a **categorical** or **numerical** variable?
    b. If it's a categorical variable, **how many levels** does it have?

3. How would you **<u>quantify this association</u>**?
    a. Difference between two summary statistics? What two summary statistics?
    b. With a model? What kind of model?

4. Are you interested in an association in a **<u>sample</u>** or a **<u>population</u>**?

5. When is it **<u>appropriate to use this test</u>** for association?

6. Can you use this model/test to **<u>make predictions</u>**?
    a. How would you quantify the performance of your predictions?

2. <u>ASSOCIATION ANALYSIS SUMMARY</u>:

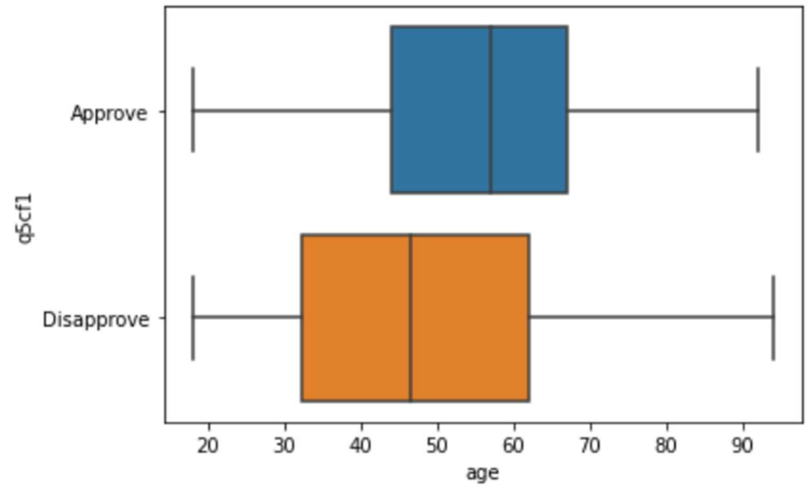
<u>RESPONSE</u>: CATEGORICAL (TWO LEVELS)

<u>EXPLANATORY</u>: NUMERICAL(S), CATEGORICAL(S) (ANY NUMBER OF LEVELS)

Research Questions about Associations	**Type of Variables Involved in the Association Test**	**<u>Explanatory Variables</u>:** **Numerical Variables and/or Categorical Variables** **<u>Response Variable</u>:** **Categorical Variable (two levels)**
	Example	Is there an association between **sex** and **age** and **opinion** on the direction that the country is going in (satisfied/dissatisfied)? 
	Type of Association *(Way to Quantify Association)*	**Logistic Regression Model** *(<u>linear relationship</u> between explanatory variables and response variable (y))*
Descriptive Analytics	**How to <u>Describe</u> an Association in a <u>Sample</u>?**	<u>Model</u>: • $\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \widehat{\beta_2}x_2 + \cdots \widehat{\beta_p}x_p$
	When is this analysis *(for the sample)* appropriate to use?	See upcoming unit.
Inferential Statistics	**How to <u>Infer</u> an Association for a <u>Population?</u>**	See upcoming unit.
	When is this analysis *(for the population)* appropriate to use?	See upcoming unit.
Predictive Analytics	**Making Predictions**	See upcoming unit.
	How to quantify the performance of your prediction(s)?	See upcoming unit.

3. WHY SHOULD WE NOT USE LINEAR REGRESSION TO MODEL A CATEGORICAL RESPONSE VARIABLE?

Suppose we wanted to fit a simple linear regression model, this time using **approval (or disapproval) of the president's foreign policy** as the response variable and **age** as the explanatory variable.

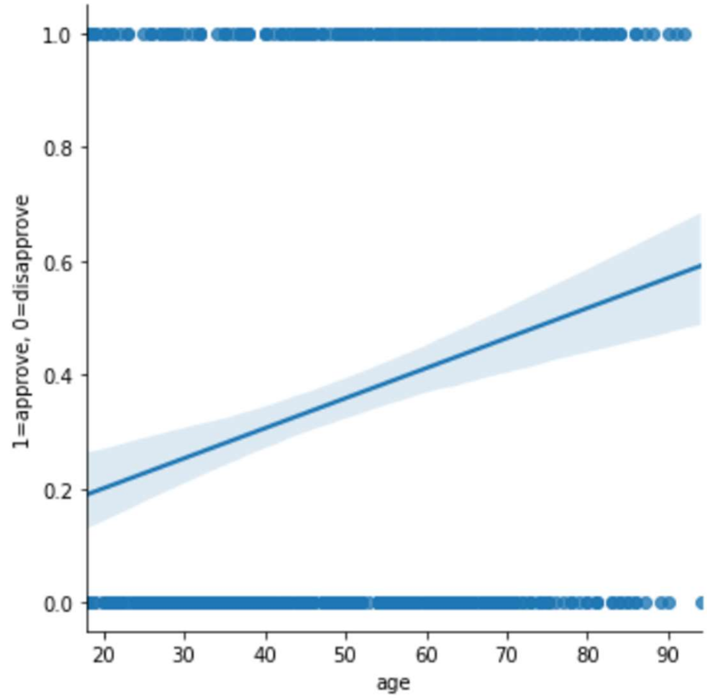


What we could *try* to do is create a new numerical variable **y** in which:

- y = ____________, when a survey respondent approves
- y = ____________, when a survey respondent disapproves.

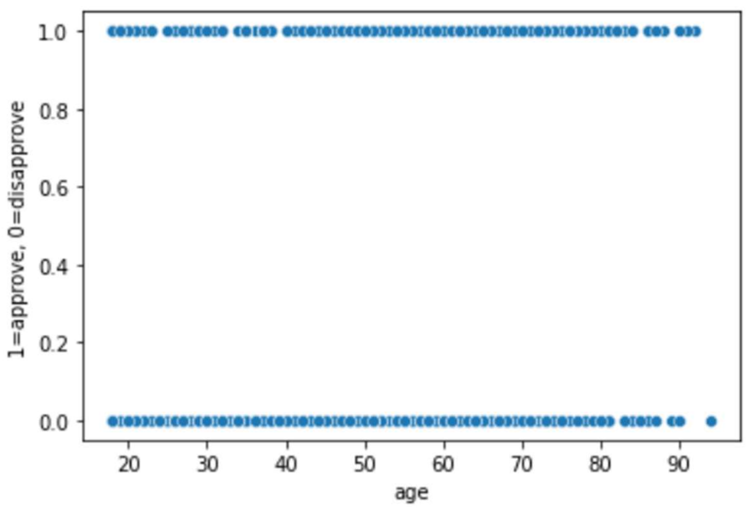
Would we expect a best fit line to provide a good fit of the data? What are some issues we might have when using this best fit line when making predictions?

	age	sex	q5cf1	y
1	70.0	Female	Disapprove	0
2	69.0	Female	Disapprove	0
4	70.0	Female	Disapprove	0
6	89.0	Female	Disapprove	0
7	92.0	Female	Approve	1



4. What curve can we fit instead when we have a categorical response variable (with 2 levels)?

What kind of curve would be a better fit of the data?



The curve that we just drew is called a ______________________________. This function has the following properties:

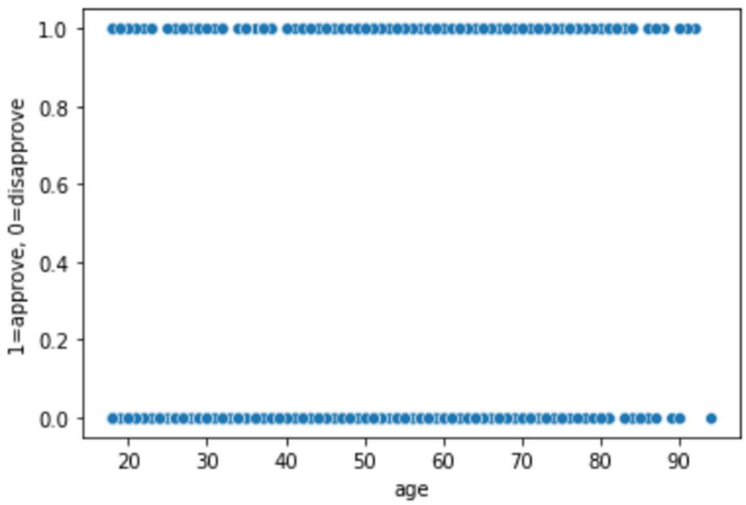
- **It's S-shaped and has two horizontal asymptotes at y = _______ and y = _______.**

- **Curves of *this type* are defined by the formula** $S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$

By multiplying a $\widehat{\beta_1}$ to x and then adding a $\widehat{\beta_0}$, we can horizontally stretch out the curve and horizontally shift the curve in an attempt to better fit our data.

$$\boldsymbol{S(\widehat{\beta}_0 + \widehat{\beta}_1 x) = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 x}}{e^{\widehat{\beta}_0+\widehat{\beta}_1 x}+1}}$$

Even with this better fitting logistic curve, in most datasets we will never have all data points (which have either a y=1 or a y=0) fit perfectly on this curve.



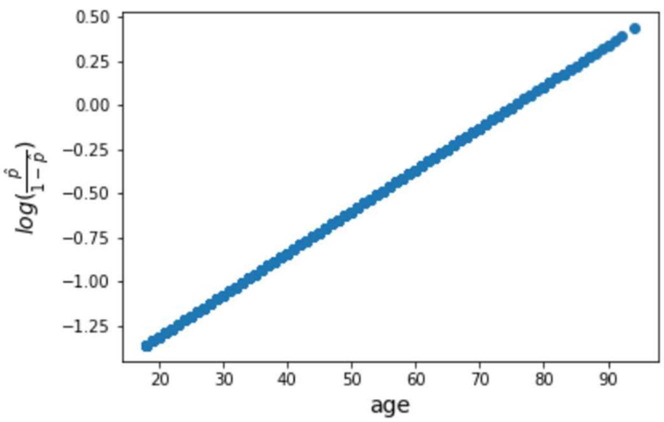
So how should we interpret predictions made with this curve then if our predictions are in between (0,1)?

$$\boxed{\boldsymbol{P(\widehat{Y=1}) = \widehat{p} = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 x}}{e^{\widehat{\beta}_0+\widehat{\beta}_1 x}+1}}}$$

In this case, we call $\widehat{\boldsymbol{p}}$ the **predictive probability,** which is the predicted probability that the corresponding response variable will be the ________________ (ie. y=1), for the given explanatory variable value x.

With some algebraic manipulation, we can convert this equation into what we call our **logistic regression model** for the sample data.

$$\boldsymbol{log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 x}$$



5. ODDS VS. PROBABILITY

In general, if $\boldsymbol{p}$ **is the probability of some "success" level happening**, then we call:

$\frac{p}{1-p}$ **= the odds <u>of</u> a success.**

In general, if $\boldsymbol{p}$ **is the probability of some "success" level happening**, then we call:

$log\left(\frac{p}{1-p}\right)$ **= the log-odds <u>of</u> a success** (odds <u>against</u> failure)

Ex: For instance, if the probability of a horse wining a race is $p = 1/10$, then

- the probability of a horse losing a race is $1 - p =$ ________
- thus, the odds of the horse winning the race is:

What does it mean to say that the odds **of** a given horse winning are 1 to 9? Think of it as "chances" in a box of tickets. For every 1 "chance" of winning there are 9 "chances" of losing.

In general, we define **odds for success are "number of successes" to "number of failures."**

We can equivalently say that the odds **against** a given horse winning are 9 to 1. We can still think of it as "chances" in a box of tickets. For every 1 "chance" of winning there are 9 "chances" of losing, but we flip the order in which we state these odds…

In general, we define **odds against success are "number of failures" to "number of successes."**

Ex: What are the odds **against** the horse winning?

In general, if $\boldsymbol{p}$ **is the probability of some "success" level happening**, then we call:

$\frac{\mathbf{1-p}}{\mathbf{p}}$ **= the odds <u>against</u> a success** (odds <u>of</u> failure)

Converting Odds (in "number of successes" to "number of failures" form) to Probability of Success

$$probability\ of\ success = p = \frac{Number\ of\ Chances\ of\ Success}{Total\ Number\ of\ Chances}$$

$$probability\ of\ failure = 1 - p = \frac{Number\ of\ Chances\ of\ Failure}{Total\ Number\ of\ Chances}$$

Converting Odds (in numerical form) to Probabilities

$$odds\ of\ success = \frac{p}{1-p} \quad and\ p = \frac{odds\ of\ success}{1 + odds\ of\ success}$$

$$odds\ of\ failure = \frac{1-p}{p} \quad and\ p = \frac{odds\ of\ success}{1 - odds\ of\ success}$$

<u>Ex:</u> If the odds of a horse winning a race are 1 to 24, what is the probability that the horse wins the race?

<u>Ex:</u> If there is a 40% chance of rain, what are the odds it won't rain?

6. Fitting a Simple Logistic Regression Curve

So how do we find these optimal values $\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1$ that best fit our data?

Efficient coefficient estimates are obtained for this model by the **method of maximum likelihood**.

We can use Python to estimate these optimal values $\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1$ that best fit our data.

Logit Regression Results

Dep. Variable:	y	**No. Observations:**	1503
Model:	Logit	**Df Residuals:**	1501
Method:	MLE	**Df Model:**	1
Date:	Sun, 25 Oct 2020	**Pseudo R-squ.:**	0.01091
Time:	20:52:45	**Log-Likelihood:**	-955.55
converged:	True	**LL-Null:**	-966.09
Covariance Type:	nonrobust	**LLR p-value:**	4.412e-06

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.9211	0.082	-11.208	0.000	-1.082	-0.760
sex[T.Male]	0.5027	0.110	4.563	0.000	0.287	0.719

Fit the logistic regression model.

Notation:

- Make sure to put a hat over the predictive probability p to indicate that this is a prediction.
- Put the variables into words.

7. Fitting a Multiple Logistic Regression Curve

Similarly, if we have multiple explanatory variables (or a categorical variable with >2 levels, which translates to more than one indicator variables), then we can fit multiple slopes $\hat{\beta}_1, \ldots, \hat{\beta}_p$ like we did with multiple linear regression.

$$log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \widehat{\beta_p} x_p$$

We also use the **method of maximum likelihood** to find these optimal values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that best fit our data. Python will also find these optimal values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ for us.

```
Optimization terminated successfully.
         Current function value: 0.612754
         Iterations 5
```

Logit Regression Results

Dep. Variable:	y	**No. Observations:**	691
Model:	Logit	**Df Residuals:**	688
Method:	MLE	**Df Model:**	2
Date:	Tue, 06 Apr 2021	**Pseudo R-squ.:**	0.06252
Time:	10:40:34	**Log-Likelihood:**	-423.41
converged:	True	**LL-Null:**	-451.65
Covariance Type:	nonrobust	**LLR p-value:**	5.457e-13

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.3609	0.288	-8.190	0.000	-2.926	-1.796
sex[T.Male]	0.8856	0.167	5.296	0.000	0.558	1.213
age	0.0260	0.005	5.466	0.000	0.017	0.035

Categorical explanatory variables are translated into indicator variables in the same way that they are for multiple linear regression models.

Ex: Use the Python summary output to formulate the logistic regression model.

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -2.3609 + .8858sex[T.Male] + 0.026age$$

8. Making Predictions with a Logistic Regression model

Use this logistic regression model to predict the following:

a. **The log-odds that a 20-year-old female supports the president's foreign policy *in the sample.***

b. **The odds that a 20-year-old female supports the president's foreign policy *in the sample.***

c. **The probability that a 20-year-old female supports the president's foreign policy *in the sample.***

9. INTERPRETING THE VALUES OF $\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2} \dots, \widehat{\beta_p}$

Easier Interpretation of the Logistic Regression Model

$$log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \widehat{\beta_p} x_p$$

We can exponentiate both sides of the logistic regression model above to help us interpret what the values of $\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2} \dots, \widehat{\beta_p}$ mean.

$$\left(\frac{\hat{p}}{1-\hat{p}}\right) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \widehat{\beta_p} x_p}$$

Remember that $\left(\frac{\hat{p}}{1-\hat{p}}\right)$ represents the predicted odds of the "success level" happening for a a given set of values for $x_1, x_2, \dots x_p$.

Interpreting $\widehat{\beta}_0$

By plugging in 0 for all x_i (for i=1,…,p), we get:

$\left(\frac{\hat{p}}{1-\hat{p}}\right) = e^{\hat{\beta}_0}$,

Thus, we call $e^{\hat{\beta}_0}$ the **baseline odds** as it is the ____________________ when $x_1 = 0, x_2 = 0, \dots x_p = 0$.

Ex: $\widehat{\beta_0} = -2.3609$

So, we say that the baseline odds of someone in the sample supporting the president's foreign policy are $e^{\widehat{\beta_0}} = 0.0943$ (or in other words, 943 to 10,000). This would represent a person who is (female and 0 years old, which is nonsensical).

Interpreting $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ $(\boldsymbol{i} = \mathbf{1}, \ldots, \boldsymbol{p})$, when $\boldsymbol{x}_{\boldsymbol{i}}$ is numerical:

We can use properties of exponents to represent the logistic regression equation as

$$\boldsymbol{odds} = \left(\frac{\widehat{\boldsymbol{p}}}{\mathbf{1} - \widehat{\boldsymbol{p}}}\right) = \boldsymbol{e}^{\widehat{\boldsymbol{\beta}}_0} \cdot \boldsymbol{e}^{\widehat{\boldsymbol{\beta}}_1 \boldsymbol{x}_1} \cdot \boldsymbol{e}^{\widehat{\boldsymbol{\beta}}_2 \boldsymbol{x}_2} \cdot \ldots \cdot \boldsymbol{e}^{\widehat{\boldsymbol{\beta}}_p \boldsymbol{x}_p}$$

If we were to increase $\boldsymbol{x}_{\boldsymbol{i}}$ by one and hold everything else constant, we would get:

$$\frac{\boldsymbol{odds}_{\boldsymbol{new}}}{\boldsymbol{odds}_{\boldsymbol{old}}} = \frac{\left(\frac{\widehat{\boldsymbol{p}}}{\mathbf{1} - \widehat{\boldsymbol{p}}}\right)_{\boldsymbol{new}}}{\left(\frac{\widehat{\boldsymbol{p}}}{\mathbf{1} - \widehat{\boldsymbol{p}}}\right)_{\boldsymbol{old}}} = \frac{\boldsymbol{e}^{\widehat{\boldsymbol{\beta}}_0} \cdot \boldsymbol{e}^{\widehat{\boldsymbol{\beta}}_1 \boldsymbol{x}_1} \cdot \ldots \cdot \boldsymbol{e}^{\widehat{\boldsymbol{\beta}}_i (\boldsymbol{x}_i + \mathbf{1})} \cdot \ldots \cdot \boldsymbol{e}^{\widehat{\boldsymbol{\beta}}_p \boldsymbol{x}_p}}{\boldsymbol{e}^{\widehat{\boldsymbol{\beta}}_0} \cdot \boldsymbol{e}^{\widehat{\boldsymbol{\beta}}_1 \boldsymbol{x}_1} \cdot \ldots \cdot \boldsymbol{e}^{\widehat{\boldsymbol{\beta}}_i (\boldsymbol{x}_i)} \cdot \ldots \cdot \boldsymbol{e}^{\widehat{\boldsymbol{\beta}}_p \boldsymbol{x}_p}} = \boldsymbol{e}^{\widehat{\boldsymbol{\beta}}_i}$$

Thus we call, $\boldsymbol{e}^{\widehat{\boldsymbol{\beta}}_i}$**,** the **odds multiplier** as this is the *multiple* we would expect the odds to increase (or decrease if $e^{\widehat{\beta}_i} < 0$) by if we increased x_i by 1 (holding everything else constant).

Ex: $\widehat{\boldsymbol{\beta}_1} = 0.0260$

So, we say that we would expect the odds of someone in the sample supporting the president's foreign policy to increase by a factor of $e^{\widehat{\boldsymbol{\beta}_1}} = 1.026$, all else held equal.

Interpreting $\widehat{\boldsymbol{\beta}}_i$ $(\boldsymbol{i = 1, \dots, p})$, when $\boldsymbol{x_i}$ is an indicator variable:

Ex: Using our fitted logistic regression model
$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -2.3609 + .08858 sex[T.Male] + 0.026 age$,

If we were to find the difference in the log-odds of a male supporting the president's foreign policy vs. a female supporting the president's foreign policy (everything else held equal), we would get:

$$\log\left(\frac{\hat{p}_{male}}{1-\hat{p}_{male}}\right) - \log\left(\frac{\hat{p}_{female}}{1-\hat{p}_{female}}\right)$$
$$= (-2.3609 + .8858(1) + 0.026age) - (-2.3609 + .8858(0) + 0.026age) = .8858$$

By using properties of logarithms, we also get

$$\log\left(\frac{\frac{\hat{p}_{male}}{1-\hat{p}_{male}}}{\frac{\hat{p}_{female}}{1-\hat{p}_{female}}}\right) = .8858$$

And by exponentiating both sides, we get:

$$\frac{odds_{male}}{odds_{female}} = \frac{\frac{\hat{p}_{male}}{1-\hat{p}_{male}}}{\frac{\hat{p}_{female}}{1-\hat{p}_{female}}} = e^{.8858} = 2.424$$

So this shows that $e^{.08858} = 2.424$ represents an **odds ratio**, which in this case represents that we expect that the odds that a male supports the president's foreign policy is 2.424 higher than the odds that a female supports the president's foreign policy.

Interpretation in General:

If $\boldsymbol{x_i}$ is an indicator variable, we expect that the odds that the level of the categorical explanatory variable in which $\boldsymbol{x_i = 1}$ is a success will be $e^{\widehat{\beta_i}}$ higher (or lower if $e^{\widehat{\beta_i}} < 0$) than the reference level for that categorical variable level.