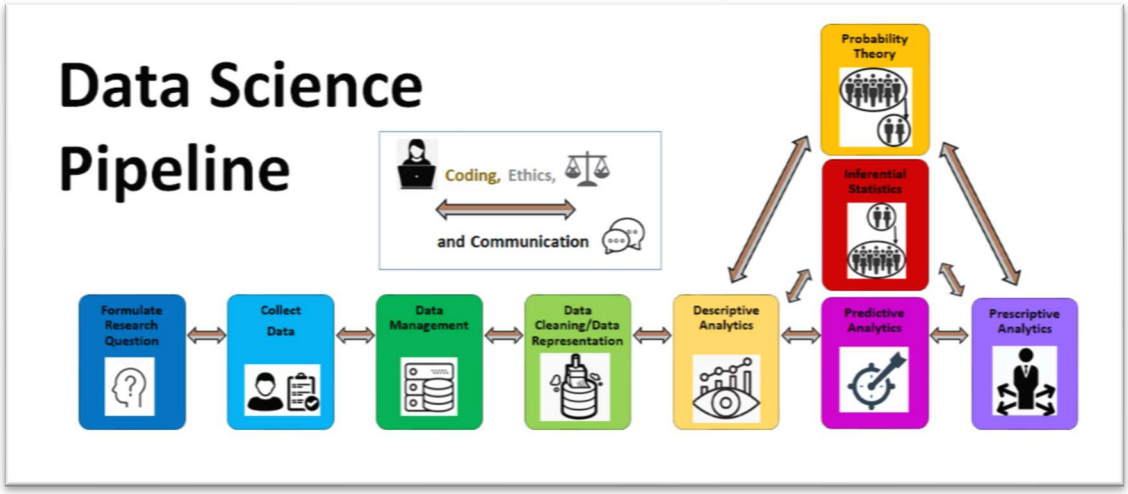
disapprove

approve

Unit 16: Logistic Regression – Part 2

Case Studies:

- To introduce the concept **multiple logistic regression *inference*,** we will examine the association between sex and age and approval for the president's foreign policy (in 2017)?
  - Response: Approve/disapprove of president's foreign policy
  - Explanatory(s):
    - Sex
    - Age



Summary of Concepts:

1. **Conducting inference on $\beta_0, \beta_1, \ldots, \beta_p$ in a Logistic Regression Model**
   - 1.1. Checking Conditions for Inference (and Model Fit)
   - 1.2. Creating a $(1-\alpha)\cdot 100\%$ confidence interval for $\beta_0, \beta_1, \ldots, \beta_p$
   - 1.3. Conducting hypothesis testing on $\beta_0, \beta_1, \ldots, \beta_p$
2. **Modeling Interaction Effects with Logistic Regression Models**

Additional Resources:

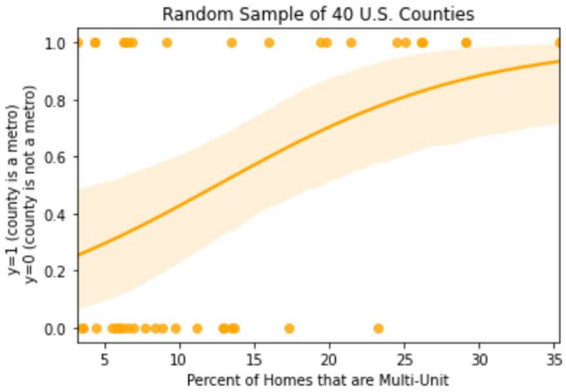
Section 8.4 in Diez, Barr, and Cetinkaya-Rundel, (2015), *OpenIntro Statistics* https://www.openintro.org/download.php?file=os3&redirect=/stat/textbook/os3.php

1. Conducting Inference on $\beta_0, \beta_1, \ldots, \beta_p$ on logistic Regression model

Like with our linear regression models, we can formulate a logistic regression model for either the sample data or the population data.

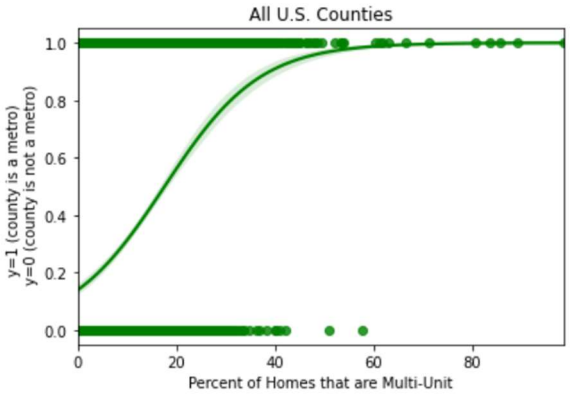
Sample Logistic Regression Model

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



Population Logistic Regression Model

$$log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$$



We can also think of $\hat{\beta}_i$ as a sample statistic that can be used as a point estimate for estimating β_i.

Like with multiple linear regression slopes and intercepts, we can do the following to conduct inference on $\beta_0, \beta_1, \ldots, \beta_p$.

1. **Create a confidence interval for β_i.**

2. **Conduct hypothesis testing for β_i.**
   a. Using a p-value
   b. Using a confidence interval.

1.1. Checking Conditions for Inference (and Model Fit) on $\beta_0, \beta_1, \ldots, \beta_p$ in a Logistic Regression Model

First check logistic regression conditions for model fit and inference.

1. Independence of Observations Condition:
    a. The sample is random.
    b. $n < 10\%$ of the population size.

2. Linearity Condition: the explanatory variables should have a linear relationship with the **log-odds** of the response variable success level.

    How to check in Python

    For each numerical explanatory variable, plot this explanatory variable on the x-axis and the (0/1 response variable on the y-axis in a scatterplot. Then fit and plot a simple logistic regression curve for each of these scatterplots. If each of your curves are "S-shaped" for each of these plots, then we can say that this condition is met.

3. No Multicollinearity Condition: the explanatory variables should not have a strong linear relationship with each other.

1.2. Creating a $(1-\alpha)\cdot 100\%$ Confidence Interval for $\beta_0, \beta_1, \ldots, or, \beta_p$ (individually) in a Logistic Regression Model

1. **First check logistic regression conditions for inference.**
   a. Independence of Observations Condition
      i. The sample is random.
      ii. $n < 10\%$ of the population size.
   b. Linearity Condition: the explanatory variables should have a linear relationship with the **log-odds** of the response variable success level.
   c. No Multicollinearity Condition: the explanatory variables should not have a strong linear relationship with each other.

2. **The $(1-\alpha)\cdot 100\%$ confidence interval for β_i is calculated by**:

$$(point\ estimate) \pm (critical\ value)(standard\ error)$$

$$\widehat{\beta_i} \pm z^* SE_{\beta_i}$$

Notation:

3. **The $(1-\alpha)\cdot 100\%$ confidence for e^{β_i} (the odds multiplier for x_i) is calculated by**:

$$(e^{\widehat{\beta_i}-z^* SE_{\beta_i}}, e^{\widehat{\beta_i}+z^* SE_{\beta_i}})$$

Remember from Unit 15, that e^{β_i} is easier to interpret than β_i. We call e^{β_i}, the odds multiplier for the explanatory variable x_i because if we were to increase this variable value by 1, we would expect the odds to increase by a multiple/factor of e^{β_i}.

After we have calculated a $(1-\alpha)\cdot 100\%$ confidence interval for β_i, we can just exponentiate the lower and upper bounds to get a confidence interval for e^{β_i}.

Ex: In the **Jupyter notebook**, create a 90% confidence interval for the slope that corresponds to the age variable in the population logistic regression model.

1.3. Conducting Hypothesis Testing on $\beta_0, \beta_1, \ldots, or, \beta_p$ (individually) in a logistic Regression model

1. **Set up the hypotheses**

   $H_0: \beta_i = 0$
   $H_A: \beta_i \neq 0$

   Equivalent to…

   $H_0: e^{\beta_i} = 1$
   $H_A: e^{\beta_i} \neq 1$

2. **Check the conditions for conducting inference on a population slope/intercept.**
   a. Independence of Observations Condition
      i. The sample is random.
      ii. $n < 10\%$ of the population size.
   b. Linearity Condition: the explanatory variables should have a linear relationship with the **log-odds** of the response variable success level.
   c. No Multicollinearity Condition: the explanatory variables should not have a strong linear relationship with each other.

3. **Calculate the point estimate (observed sample statistic)**

   $$\widehat{\beta_i}$$

4. **Calculate the p-value**

$$\boldsymbol{p - value = 2P(Z \geq |\frac{\widehat{\beta_i} - 0}{SE_{\widehat{\beta_i}}}|)}$$

5. Make a Decision

With a p-value

a. If $\mathbf{p-value} < \boldsymbol{\alpha}$**,** then we "reject the null hypothesis." And we say that "there IS sufficient evidence to suggest the alternative hypothesis."
b. If $\mathbf{p-value} \geq \boldsymbol{\alpha}$, then we "fail to reject the null hypothesis." And we say that "there IS NOT sufficient evidence to suggest the alternative hypothesis."

With a confidence interval for $\boldsymbol{\beta_i}$

a. If the **null value $\boldsymbol{\beta_i = 0}$ is NOT IN the confidence interval,** then we "reject the null hypothesis." And we say that "there IS sufficient evidence to suggest the alternative hypothesis."
b. If **null value $\boldsymbol{\beta_i = 0}$ is IN the confidence interval**, then we "fail to reject the null hypothesis." And we say that "there IS NOT sufficient evidence to suggest the alternative hypothesis."

With a confidence interval for $\boldsymbol{e^{\beta_i}}$

a. If the **null value $\boldsymbol{e^{\beta_i} = 1}$ is NOT IN the confidence interval,** then we "reject the null hypothesis." And we say that "there IS sufficient evidence to suggest the alternative hypothesis."
b. If **null value $\boldsymbol{e^{\beta_i} = 1}$ is IN the confidence interval**, then we "fail to reject the null hypothesis." And we say that "there IS NOT sufficient evidence to suggest the alternative hypothesis."

Ex: In the **Jupyter notebook**, create test whether there is sufficient evidence to suggest that the population slope for age is non-zero in the logistic regression model (which predicts the probability for approval for the president's foreign policy given age and sex.)

2. MODELING INTERACTION EFFECTS IN A LOGISTIC REGRESSION MODEL

Like in linear regression models, we can model the interaction effect between two explanatory variables x_i and x_j by the product of these two variables $x_i \cdot x_j$ in our logistic regression model and fitting a slope for this product as well as our other explanatory variables.

Ex: In the **Jupyter notebook**, create a logistic regression model that predicts the probability that a person in the sample approves of the president's foreign policy given:

- sex,
- age, and
- the interaction between sex and age.

Then, write out the resulting logistic regression model for the sample.

Finally, use this model to predict the probability that 19 year old male supported the president's foreign policy (in 2017).