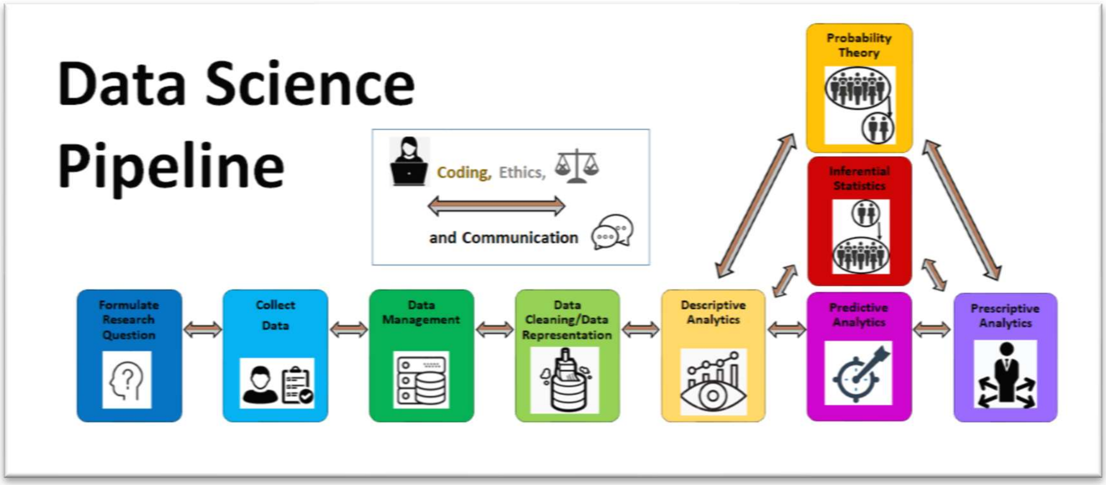
Unit 19: Logistic Regression Variable Selection

Case Studies:

- To introduce the concept of **using training data to *build* a model** and **using test data to *test* a model for it's predictive capabilities** we will, again, examine the relationship between a:
  - **Categorical response variable:** support for a certain opinion (favor/not in favor) and an
  - **Explanatory variables:**
    - Sex
    - Party, and
    - Age


Data Science Pipeline

Summary of Concepts:

1. Overfitting by using too many uninformative explanatory variables
2. Some pros and cons of overfitting vs. underfitting a model
3. Theory: Overfitting vs. Underfitting a Model
   3.1. A general goal of machine learning
   3.2. Properties of the **estimation function**
   3.3. Estimation function definitions
   3.4. Relationship between **bias**, **variance**, **overfitting**, **underfitting**, and **mean squared error** of a model
   3.5. Goal of selecting a model that will make good predictions on new data
4. Goal: Find a Parsimonious Model
5. More about Fitting a Logistic Regression Model
   5.1. How are the optimal values of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ determined in a logistic regression model?
   5.2. Where do we find the optimal log-likelihood function value for a given logistic regression model?
6. **Model Selection** with **Log Likelihood Ratio Test**
7. **Model Selection** with **AIC** and **BIC**

Additional Resources

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.
https://www.ime.unicamp.br/~dias/Intoduction%20to%20Statistical%20Learning.pdf

1. Overfitting by Using too many Uninformative Explanatory Variables

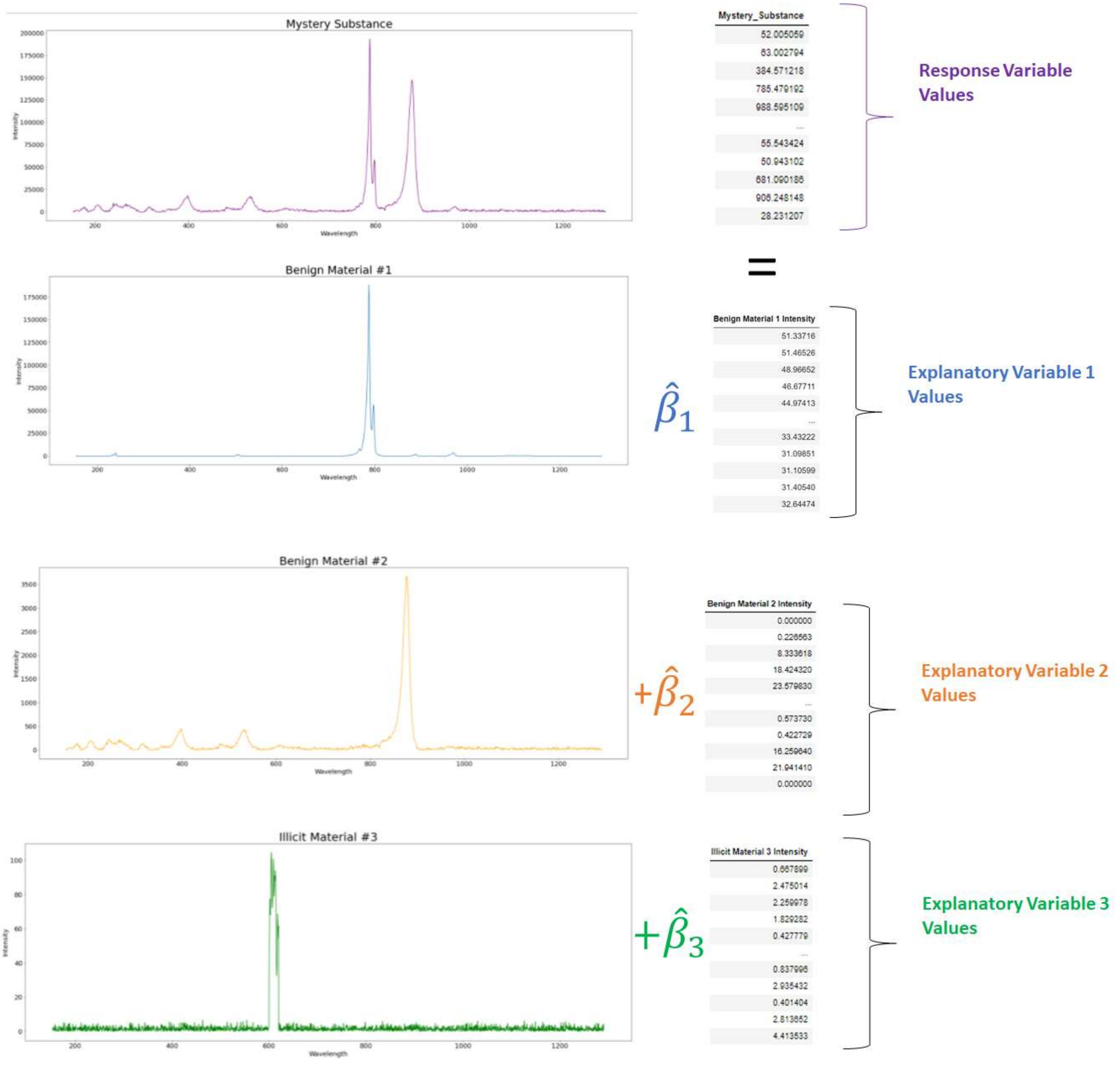
Ex: of the "spectral fingerprints" of *known* substances, the first two are benign substances, while the last is an "illicit" substance.

1. Which (if any) of the three substances below are we relatively sure the unknown substance is comprised of?

2. We would like to fit a linear regression model to help us identify which substances the mystery substance is comprised of.

   a. We will set the response variable values to be the spectral fingerprint of the mystery substance.
   b. We will set the three known spectral fingerprints to be the three explanatory variables.

   Ideally, which values of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ would we want to be non-zero in the linear regression model? Which would we want to be zero in the linear regression model?



Mystery_Substance
52.005059
63.002794
384.571218
785.479192
988.595109
...
55.543424
50.943102
681.090186
906.248148
28.231207

Benign Material 1 Intensity
51.33716
51.46526
48.96652
46.67711
44.97413
...
33.43222
31.09851
31.10599
31.40540
32.64474

Benign Material 2 Intensity
0.000000
0.226563
8.333618
18.424320
23.579830
...
0.573730
0.422729
16.259640
21.941410
0.000000

Illicit Material 3 Intensity
0.667899
2.475014
2.259978
1.829282
0.427779
...
0.837996
2.935432
0.401404
2.813652
4.413533

3. When we fit the linear regression model, we get the following values for the slopes:

$$\hat{\beta}_1 = \_\_\_\_\_\_\_\_\_\_, \hat{\beta}_2 = \_\_\_\_\_\_\_\_\_\_\_, \hat{\beta}_3 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

Why do you think we got a result like this? Do you think that this is evidence that the illicit material 3 is actually in the unknown substance?

IN GENERAL: Adding Explanatory Variables to a Model

For most fitted linear and logistic regression models, most explanatory variable value slopes $\hat{\beta}_i$ will have some

__________________ value, regardless of whether ____.

- In a **linear regression model**, adding an explanatory variable to the model will never

  ___________________ the ____________ of the model.

- In a **logistic regression model**, adding an explanatory variable to the model will never

  ___________________ the ____________________________________ of the model.

Ex: Choosing the Right Number of Variables in the Model

Knowing what we know about this example, we can say the following.

- A linear regression model that includes ***all* three explanatory variables** (ie. known substances) to predict the response variable (ie. unknown substance) will _____________________ the model.

- A linear regression model that includes ***just explanatory variable 1*** (ie. benign substance 1) to predict the response variable (ie. unknown substance) will _____________________ the model.

2. Some Pros and Cons of Overfitting vs. Underfitting a Model (via Too many or too little explanatory variables)

Overfitting a Model: Too many explanatory variables

Pros:

- The model, for the __________________ dataset, will have __________________ predictive power.

Cons:

- The model, for the __________________ dataset(s), my not have __________________ predictive power.
- The model may ________________ explanatory variables that have no

  _______________________________ with the response variable.

Underfitting a Model: Too few explanatory variables

Pros:

- The model, for the __________________ dataset, may have __________________ predictive power.

Cons:

- The model, for the __________________ dataset(s), may not have __________________ predictive power.
- The model may ________________ explanatory variables that have

  _______________________________ with the response variable.

3. Theory: Overfitting vs. Underfitting a Model

We want to be able to define "overfitting" and "underfitting" of a model in more mathematically precise terms. Let's consider the case of a linear regression model.

3.1. A General Goal of Machine Learning

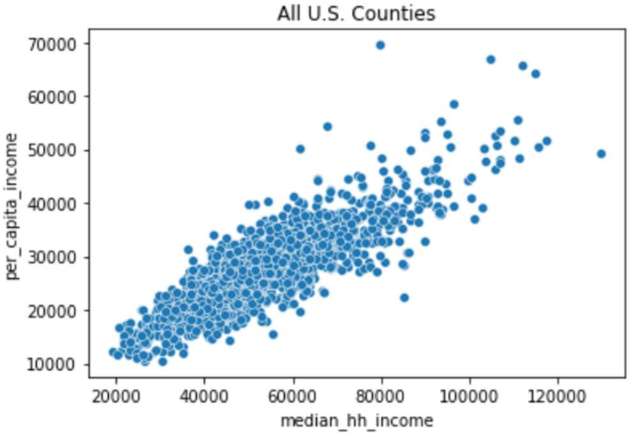
Actual Relationship Assumption: For some response variable Y and a set of p predictors $X = (X_1, X_2, \dots, X_p)$, we assume there is some underlying relationship between $Y$and X modeled with:

$$Y = f(X) + \epsilon$$

Properties:

- $f(X)$ is ____________________
- ϵ is a ____________________ called the **error (or noise) term.**

Ex: $Y = \beta_0 + \beta_1 X + \epsilon = 55609.4685 + 0.4116X + \epsilon$



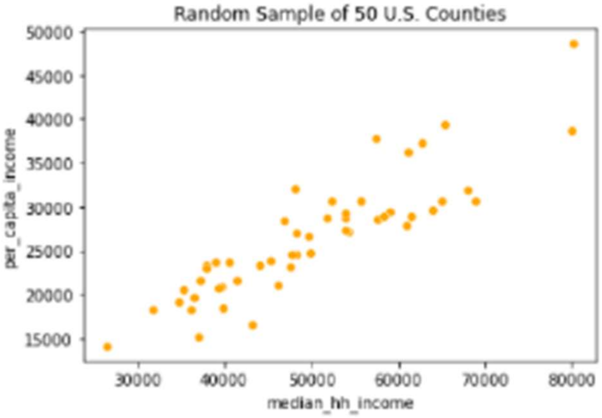
Goal of Machine Learning: Come up with an estimation function of $f(X)$, called:

$$\hat{f}(X) = \hat{Y}$$

Ex: $\hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X = 1589.8568 + 0.4985X.$

- In this example, $\hat{f}(X)$ is ________________________ for a given value of X.
- In this example, $\hat{f}(\,)$ was determined by using ______________________________________ on a

  ____________________________ which was a _____________________________ from the population.



3.2. PROPERTIES OF THE ESTIMATION FUNCTION

More about $\hat{f}(X)$

- Many Estimation Functions:
  We can create many different estimation functions $\hat{f}(X)$ for $f(X)$ using a variety of different models, datasets, and algorithms.

- Using Estimation Functions for Prediction:
  We can use our estimation function to make a prediction $\hat{f}(x_0) = \hat{y}_0$ of the response variable value for a given set of explanatory variable value inputs x_0.

- What we want to know about $\hat{f}(x_0)$.:
  - How will the predictions $\hat{f}(x_0)$ **vary** (based on different ways to create $\hat{f}()$)?
  - **How far away from $f(x_0)$** do we expect $\hat{f}(x_0)$ to be (based on different ways to create $\hat{f}()$)?
  - **How far away from y_0** do we expect $\hat{f}(x_0)$ to be (based on different ways to create $\hat{f}()$)?

3.3. Estimation Function Definitions

In order to quantify how $\hat{f}(x_0)$ will behave based on different datasets and models used to create this prediction function, we need the following definitions

Random Variable $\hat{f}(x_0)$ Definition:

- We can also define $\hat{f}(x_0)$ to be a ____________________________, defined by the following.
  - Random Experiment:
    - Randomly sample n observations from the population.
    - Fit a new linear regression model $\hat{f}(x)$ with this random sample.
  - Numerical Value Assigned to Outcome in Sample Space:
    - Make a prediction $\hat{y}_0 = \hat{f}(x_0)$ of the response variable for x_0.

$E[\hat{f}(x_0)]$ Definition:

If we were to collect many, many random samples of observations from the population and fit a linear regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_P x_P$ with each random sample and then predict the response variable value $\hat{y} = \hat{f}(x_0)$, then we would call the expected **average** of these predictions $\boldsymbol{E[\hat{f}(x_0)]}$**.**

We can think of this as: ____.

$Var[\hat{f}(x_0)]$ Definition:

If we were to collect many, many random samples of observations from the population and fit a linear regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_P x_P$ with each random sample and then predict the response variable value $\hat{y} = \hat{f}(x_0)$, then we would call the expected **variance** of these predictions $\boldsymbol{Var[\hat{f}(x_0)]}$**.**

We can think of this as: ____.

$Bias[\hat{f}(x_0)]$ Definition:

We define the **bias** of the random variable $\hat{\boldsymbol{f}}(\boldsymbol{x_0})$ as:

$$\boldsymbol{Bias}\big(\hat{f}(x_0)\big) = \boldsymbol{E}[\hat{f}(x_0)] - \boldsymbol{f}(\boldsymbol{x_0})$$

We can think of this as: ____.

Expected Mean Squared Error of x_0 Definition:

We define the **expected mean squared error of $\boldsymbol{x_0}$** with the random variable $\hat{\boldsymbol{f}}(\boldsymbol{x_0})$ as

$$\boldsymbol{E}\left(\left(\boldsymbol{y_0} - \hat{f}(x_0)\right)^2\right),$$

Where $y_0$is the actual response variable value that corresponds to x_0.

We can think of this as: ____.

3.4. Relationship between Bias, Variance, overfitting, Underfitting, and mean squared error of a model.

Here is a special property that links together everything we have talked about in this unit so far:

Bias-Variance Trade-Off Property:

$$\boldsymbol{E}\left(\left(\boldsymbol{y_0} - \hat{f}(x_0)\right)^2\right) = \left[\boldsymbol{Bias}\big(\hat{f}(x_0)\big)\right]^2 + \boldsymbol{Var}[\hat{f}(x_0)] + \boldsymbol{Var}[\epsilon]$$

An Overfit Model:

$$E\left(\left(y_0 - \hat{f}(x_0)\right)^2\right) = \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var[\hat{f}(x_0)] + Var[\epsilon]$$

- Will have ____________________ variables in the model.

- Will have _______________________ variance in the values of $\hat{f}(x_0)$.

- Will have $E[\hat{f}(x_0)]$ be _________________ from the true estimate $f(x_0)$, so thus the bias will be ____________________.

An Underfit Model:

$$E\left(\left(y_0 - \hat{f}(x_0)\right)^2\right) = \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var[\hat{f}(x_0)] + Var[\epsilon]$$

- Will have ____________________ variables in the model.

- Will have _______________________ variance in the values of $\hat{f}(x_0)$.

- Will have $E[\hat{f}(x_0)]$ be _________________ from the true estimate $f(x_0)$, so thus the bias will be ____________________.

3.5. GOAL OF SELECTING A MODEL THAT WILL MAKE GOOD PREDICTIONS ON *NEW DATA*

Here are some examples of things *you* can **“select” in a given model** of a dataset *X* with a response variable *Y*.

1. Which ________________________ you have included the model.

2. What type of regression model it is: ______________________________________.

3. Some other type of model (not a regression), for example

   ____.

Goal: In order to select a model that has a low error for new data (measured by

$E\left(\left(y_0 - \hat{f}(x_0)\right)^2\right)$), you should select one that will both have ________________bias

and _________________ variance for $\hat{f}(x_0)$

4. GOAL: FIND A "PARSIMONIOUS" MODEL

Definition

Thus, because we do not want to underfit or overfit a model, our goal is to find the **parsimonious model** which is a balance of the two. Specifically, a parsimonious model will find the ideal balance of:

- a __________________ number of explanatory variables to avoid _________________________ and
- a __________________ predictive power to avoid _________________________.

Methods

Depending on the model that we are using, there are various metrics/tests that help us find a parsimonious model. For logistic regression models, the following metrics/tests can help us find this parsimonious model:

1. **Log Likelihood Ratio Test**

2. **Pick the model with the Lowest AIC Score**

3. **Pick the model with the Lowest BIC Score**

5. More About Fitting a Logistic Regression Model

5.1. How are the optimal values of $\widehat{\boldsymbol{\beta}}_{\mathbf{0}}, \widehat{\boldsymbol{\beta}}_{\mathbf{1}}, \dots, \widehat{\boldsymbol{\beta}}_{\boldsymbol{p}}$ determined in a logistic regression model?

1. Logistic Regression Model Assumptions

If we are to fit a logistic regression model to a training dataset with n observations with response variable $y = (y_1, y_2, \dots, y_n)$, then we make the following model assumptions.

a. **Independence of Observations:** Each response variable observation $y_1, y_2, \dots, y_n$ in the training sample is independent.

b. **Bernoulli Random Variables:** The response variable values follow a Bernoulli distribution

$$y_i \sim Bern(p_i),$$

where

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip},$$

which is equivalent to

$$\mathrm{p_i} = \frac{e^{\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip}}}{1+e^{\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip}}},$$

2. Probability Mass Function of a *Single Random Variable* y_i

$$p(y_i | \beta_0, \beta_1, \dots, \beta_p X_i) = p_i^{y_i}(1-p_i)^{1-y_i} = \left\{ \begin{matrix} \underline{\hspace{2cm}} & if\ y_i = 1 \\ \underline{\hspace{2cm}} & if\ y_i = 0 \end{matrix} \right\}$$

$$\text{Where, } \mathrm{p_i} = \frac{e^{\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip}}}{1+e^{\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip}}}$$

3. **Probability Mass Function of a *All Random Variable* $\boldsymbol{y_1, \dots, y_n}$**

In binary response models such as logistic regression the **likelihood function (LF)** is the joint probability mass function of the responses viewed as a function of the parameters. For a logit model with independent Bernoulli responses, the likelihood function has the form

$$\begin{aligned} LF(\beta_0, \beta_1, \dots, \beta_p) &= p(y_1, \dots, y_n | \beta_0, \beta_1, \dots, \beta_p, X_1, \dots, X_n) \\ &= (p_1^{y_1}(1-p_1)^{1-y_1}) \cdot (p_2^{y_2}(1-p_2)^{1-y_2}) \cdot \dots \cdot (p_n^{y_n}(1-p_n)^{1-y_n}) \end{aligned}$$

$$\text{Where, } \mathrm{p_i} = \frac{e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}{1+e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}$$

4. **Goal:** Find the optimal values of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, that maximize the likelihood function $LF()$.

We use the method of **maximum likelihood** to find the optimal values of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ that maximize the likelihood of the data we observed (ie. $y_1, \dots, y_n$).

$$\begin{aligned} LF(\beta_0, \beta_1, \dots, \beta_p) &= p(y_1, \dots, y_n | \beta_0, \beta_1, \dots, \beta_p, X_1, \dots, X_n) \\ &= (p_1^{y_1}(1-p_1)^{1-y_1}) \cdot (p_2^{y_2}(1-p_2)^{1-y_2}) \cdot \dots \cdot (p_n^{y_n}(1-p_n)^{1-y_n}) \end{aligned}$$

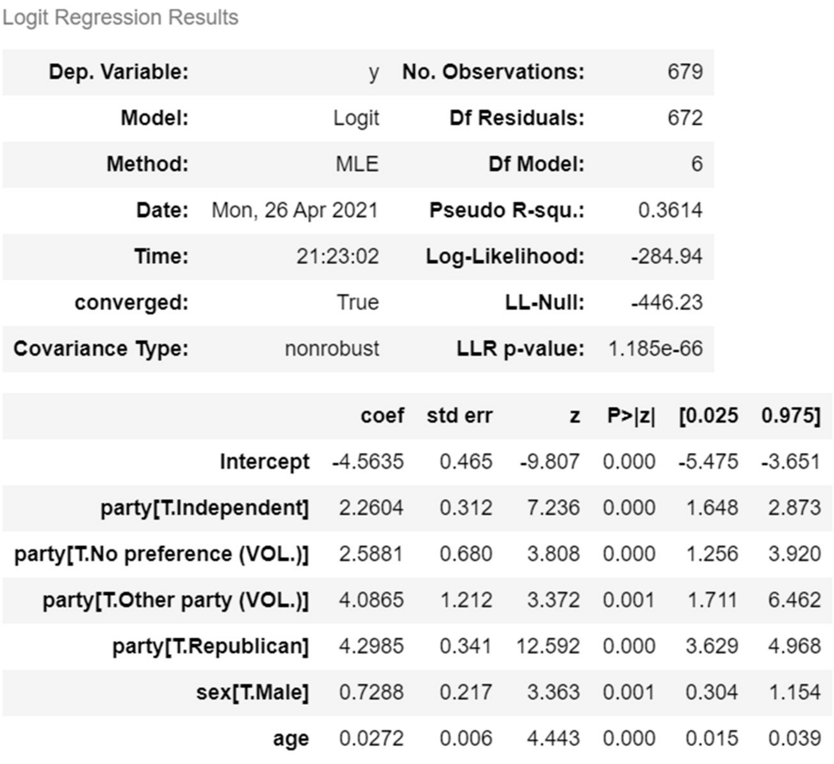
$$\text{Where, } \mathrm{p_i} = \frac{e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}{1+e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}$$

5. **Easier Goal:** Find the optimal values of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, that maximize the equivalent **log-likelihood function**.

$$LLF(\beta_0, \beta_1, \dots, \beta_p) = y_1 \log(\mathrm{p}_1) + (1 - y_1) \log(1 - p_1) + y_2 \log(\mathrm{p}_2) + (1 - y_2) \log(1 - p_2) + \cdots + y_n \log(\mathrm{p}_n) + (1 - y_n) \log(1 - p_n)$$

$$\text{Where, } \mathrm{p_i} = \frac{e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}$$

5.2. WHERE DO WE FIND THE OPTIMAL LOG-LIKELIHOOD FUNCTION VALUE FOR A GIVEN LOGISTIC REGRESSION MODEL?



Logit Regression Results

Dep. Variable:	y	**No. Observations:**	679
Model:	Logit	**Df Residuals:**	672
Method:	MLE	**Df Model:**	6
Date:	Mon, 26 Apr 2021	**Pseudo R-squ.:**	0.3614
Time:	21:23:02	**Log-Likelihood:**	-284.94
converged:	True	**LL-Null:**	-446.23
Covariance Type:	nonrobust	**LLR p-value:**	1.185e-66

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.5635	0.465	-9.807	0.000	-5.475	-3.651
party[T.Independent]	2.2604	0.312	7.236	0.000	1.648	2.873
party[T.No preference (VOL.)]	2.5881	0.680	3.808	0.000	1.256	3.920
party[T.Other party (VOL.)]	4.0865	1.212	3.372	0.001	1.711	6.462
party[T.Republican]	4.2985	0.341	12.592	0.000	3.629	4.968
sex[T.Male]	0.7288	0.217	3.363	0.001	0.304	1.154
age	0.0272	0.006	4.443	0.000	0.015	0.039

6. Model Selection with the Log Likelihood Ratio Test

Recall that in linear regression modeling it can be useful to test between two models using an analysis of variance F test, which compares the residual sums of squares for two, nested models. It allows us to test multiple parameters within one hypothesis test.

In logistic regression modeling, the F test is no longer applicable. However, the same general testing idea is possible by comparing log-likelihoods between two nested models. The change in log-likelihood is used as a large sample chi-square test of the null hypothesis that the simpler model is adequate.

1. First, we need to define two "nested" sample models.

Full Model (contains all the slopes)

$$\boldsymbol{log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \cdots \widehat{+\beta_p}x_p + \widehat{\beta_{p+1}}x_{p+1} + \cdots \widehat{\beta_{p+q}}x_{p+q}}$$

Where llf_{full} is the optimal log likelihood function value of the full model.

Reduced Model (contains just the slopes that you aren't testing)

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

Where llf_{red} is the optimal log likelihood function value of the reduced model.

2. Set up the hypotheses

$H_0: The\ reduced\ model\ is\ correct.$

$H_A: The\ reduced\ model\ is\ incorrect\ because\ at\ least\ one\ missing\ coefficient\ is\ non - zero.$

3. Calculate the test statistic:

We call the test statistic for this test the **log likelihood ratio**:

$$llr = -2(llf_{red} - llf_{full})$$

4. Calculate the p-value

$$\boldsymbol{p-value} = \boldsymbol{P}(\mathbf{X}^{\mathbf{2}}_{\boldsymbol{df=q}} \geq test\ stat)$$

5. Make a Decision

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."

6.1. CHI-SQUARED DISTRIBUTION

First, let's discuss this new distribution and some of it's properties.

***Random Variable that Follows the X^2-Distribution*:**

Definition: A continuous random variable is said to follow the **X^2-distribution with k degrees of freedom** if it has the following probability density function (pdf).

Short-Hand: _______________________

Probability Density Function:

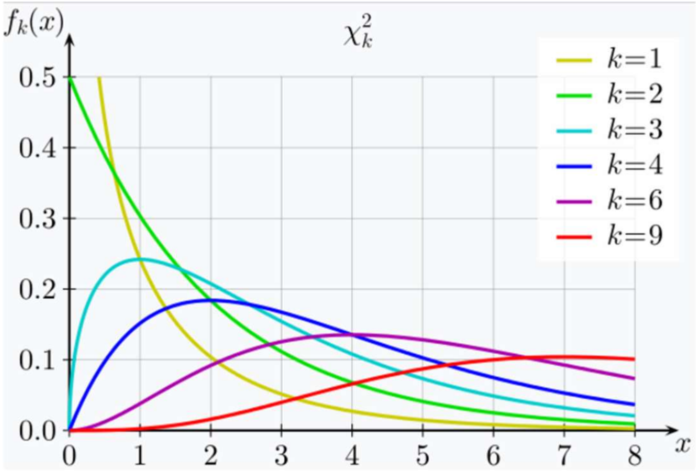
$$f(x) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, for\ x > 0$$

Parameters that Dictates Shape:____________________

Properties:

- Always ___________________________

Shapes: Can take on many different shapes, based on the parameter values.



Go to section 6 in the Unit 19 notebook for application of Log Likelihood Ratio test.

7. Model Selection with the AIC and BIC

AIC and BIC are criteria for evaluating a model that combine the likelihood assessment of fit with a penalty for complex models. Historically they were derived from different perspectives.

Akaike Information Criterion (AIC)

The AIC of a regression model with p slopes is calculated as:

$$AIC = -2 \cdot LLF + 2p$$

<u>Using AIC</u>:

- The model with the _______________ AIC is considered more of a

  ________________________________.

<u>How does AIC Help Pick out a Parsimonious Model</u>:

- As the number of slopes _____________ in a model…

$$AIC = -2 \cdot LLF + 2p$$

- As the number of slopes _____________ in a model…

$$AIC = -2 \cdot LLF + 2p$$

AIC helps us find a model with ideally a _________________ number of slopes and a

_________________ LLF (ie. predictive power).

Bayes Information Criterion (BIC)

The BIC of a regression model with p slopes is calculated as:

$$AIC = -2 \cdot LLF + \ln(n) \cdot p$$

Using BIC:

- The model with the _______________ BIC is considered more of a ________________________________.

How does BIC Help Pick out a Parsimonious Model:

- As the number of slopes _____________ in a model…

$$AIC = -2 \cdot LLF + \ln(n) \cdot p$$

- As the number of slopes _____________ in a model…

$$AIC = -2 \cdot LLF + \ln(n) \cdot p$$

AIC helps us find a model with ideally a _________________ number of slopes and a _________________ LLF (ie. predictive power).

AIC vs. BIC

BIC tends to favor _________________ more heavily than does AIC due to its ______________ penalty for large p.

What AIC and BIC *can* be used for…

What AIC and BIC *cannot* be used for…