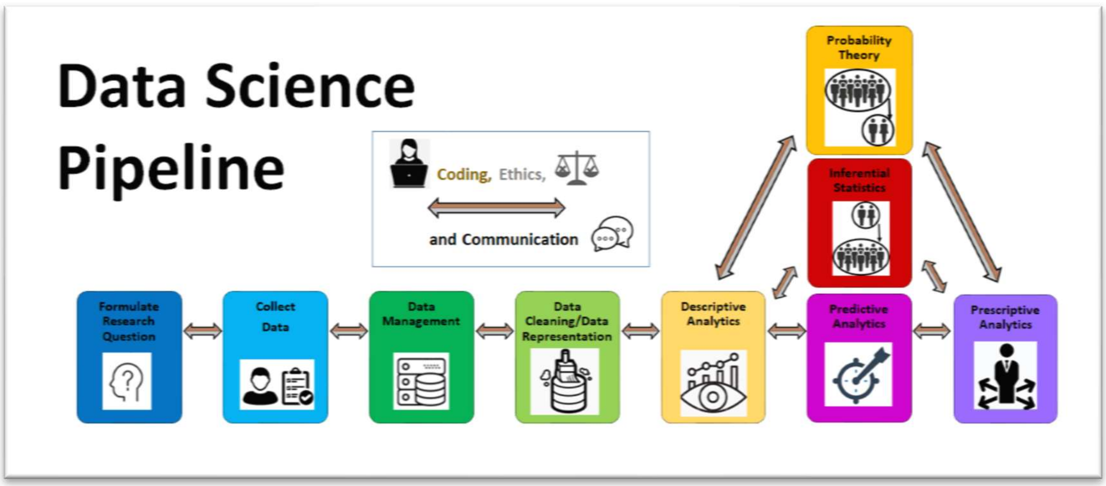
Unit 20: Variable Selection with Regularization

Case Studies:

- To introduce the **basics of machine learning, regularized logistic regression,** and **cross-validation** 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,
    - Age,
    - Whether a landline or cell phone was used for the survey



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 $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ 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

Sections 5.1, 6.1, and 6.2 *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. <u>PROBLEM</u>: WHAT ARE SOME ALGORITHMS THAT CAN HELP US <u>EFFICIENTLY</u> IDENTIFY MODELS THAT HAVE "HIGH PARSIMONY"?

1. **<u>Review</u>:** What methods have learned so far that help us determine which set of explanatory variables should be included in the "most parsimonious" logistic regression model?

2. **<u>Review</u>:** In general, why would we want a model that is parsimonious?

3. **<u>Total Number of Possible Models</u>:** If we have *P* available explanatory variables to model in a logistic regression model, how many possible logistic regression models can we make and evaluate whether they are the "most parsimonious"?

4. **<u>Unit 20 Problem</u>:** What we will explore in this unit is some algorithms that can help us ***efficiently*** identify which explanatory variables to include in a regression model such that the resulting model will have **high parsimony?**

2. ONE IDEA: USE BACKWARDS ELIMINATION OR FORWARD SELECTION ALGORITHMS

We can use what we call a **backwards elimination algorithm** to help us select a good combination of explanatory variables to include in *a model* that achieve some goal we are trying to reach. For instance, what we fill in the **black** blanks below is how we can use a backwards elimination algorithm that finds a model with ____________________, however you can use this same structure select a model that optimizes some other metric as well.

Backwards Elimination Algorithm

Goal: Find the model with the ______________ ________________.

Steps:

1. **Fit a "current model" and find the _________________ of this model.**

   In the beginning, your "current model" should include _____________ possible explanatory variables you are considering.

2. **For each explanatory variable in your "current model" do the following:**

   a. **Fit a "test model".**

      Your "test model" should be every explanatory variable in the "current model" ________________ the explanatory variable you are considering.

   b. **Find the ________________ of this "test model".**

3. **If NONE of the "test models" from step (2) had a _________________ that was _________________ than the _______________________ for the "current model", then STOP THE ALGORITHM, and return the "current model" as your "final model".**

4. **Otherwise, choose the "test model" from step (2) that had the _________________ ___________________, and set your new "current model" to be this "test model". Then go back to step (2).**

We can use what we call a **forward selection algorithm** to help us select a good combination of explanatory variables to include in *a model* that achieve some goal we are trying to reach. For instance, what we fill in the **black** blanks below is how we can use a backwards elimination algorithm that finds a model with _____________________, however you can use this same structure select a model that optimizes some other metric as well.

Forward Selection Algorithm

Goal: Find the model with the ______________ ________________.

Steps:

1. **Fit a "current model" and find the _________________ of this model.**

   In the beginning, your "current model" should include _____________ possible explanatory variables you are considering.

2. **For each explanatory variable in your "current model" do the following:**

   a. **Fit a "test model".**

      Your "test model" should be every explanatory variable in the "current model" ________________ the explanatory variable you are considering.

   b. **Find the ________________ of this "test model".**

3. **If NONE of the "test models" from step (2) had a _________________ that was _________________ than the _______________________ for the "current model", then STOP THE ALGORITHM, and return the "current model" as your "final model".**

4. **Otherwise, choose the "test model" from step (2) that had the _________________ ___________________, and set your new "current model" to be this "test model". Then go back to step (2).**

Caveat: **Forward selection algorithms** and **backwards elimination algorithms** that try to find a model with a _______________ _____________________, are **not guaranteed to find** the model with the ________________ ________________ out of all possible models!

Pros of Using Backwards Elimination (or Forward Selection) to Help find a Parsimonious Model

Cons of Using Backwards Elimination (or Forward Selection) to Help find a Parsimonious Model

See the exercise in the Jupyter notebook section 2.

3. <u>Another Idea</u>: Use a Regularization term in your Regression Model

We can also use a *single* model with a **regularization term** to help us select a good combination of explanatory variables to include in *a model* that achieve some goal we are trying to reach.

3.1 <u>Recap</u> of Non-Regularized Logistic Regression Models and AIC/BIC

<u>Non-Regularized Regression Objective Function</u>

1. **<u>Main Problem</u>**: Recall from previous units, when we fit a logistic regression model (ie. a **non-regularized logistic regression model)** we seek to find the optimal values of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ that __________________ the log likelihood function **LLF().**

$$\underset{\hat{\beta}_0,\hat{\beta}_1,\dots,\hat{\beta}_p}{\text{Max}} LLF(\hat{\beta}_0,\hat{\beta}_1,\dots,\hat{\beta}_p) = \underset{\hat{\beta}_0,\hat{\beta}_1,\dots,\hat{\beta}_p}{\max} \{y_1\log(\mathrm{p}_1) + (1-y_1)\log(1-p_1) + \cdots + y_n\log(\mathrm{p}_\mathrm{n}) + (1-y_n)\log(1-p_n)\},$$

$$\text{Where, } \mathrm{p_i} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\cdots+\hat{\beta}_p x_{ip}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\cdots+\hat{\beta}_p x_{ip}}}$$

We call the LLF the **objective function** as it is the function we are trying to optimize (ie. either maximize or minimize) in a given problem.

2. **<u>Equivalent Problem</u>**: This problem is equivalent to trying to find the optimal values of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ that ______________ "-2" times the log likelihood function, ie. **-2*LLF().**

$$\underset{\hat{\beta}_0,\hat{\beta}_1,\dots,\hat{\beta}_p}{\text{Min}} -2\cdot LLF(\hat{\beta}_0,\hat{\beta}_1,\dots,\hat{\beta}_p) = \underset{\hat{\beta}_0,\hat{\beta}_1,\dots,\hat{\beta}_p}{\min} -2\cdot[y_1\log(\mathrm{p}_1) + (1-y_1)\log(1-p_1) + \cdots + y_n\log(\mathrm{p}_\mathrm{n}) + (1-y_n)\log(1-p_n)],$$

$$\text{Where, } \mathrm{p_i} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\cdots+\hat{\beta}_p x_{ip}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\cdots+\hat{\beta}_p x_{ip}}}$$

<u>Model Selection</u>

Also, recall from previous units that the AIC for a given model (that has already been _____________) is calculated as:

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

<u>Parsimonious Models</u>

Ideally, we want the AIC of an __________________ to be ___________, which we can get when the:

- LLF is ___________________, and the
- p (ie. number of slopes) is __________________.

Penalty Terms

We call ________________ the **penalty term** in the AIC equation, as the ____________ slopes we have in the model, the more the AIC function for the model is penalized.

Main Idea Behind Regularization Terms

Use an **objective function** that tries to find the optimal values of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ that:

1. **_____________** "-2" times the log likelihood function, (ie. **-2*LLF())** and

2. also uses a **penalty term** (ie. a **regularization term)** that penalizes models with a large number of ___________ slopes for the explanatory variables.

$$\underset{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p}{\text{Min}} -2 \cdot LLF(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p) + \boldsymbol{penalty\ term}$$

Main Types of Regularization Terms

There are three common types of regularization terms that we will discuss in this unit.

1. Ridge Regression regularization term (L2 Penalty)
2. LASSO regularization term (L1 Penalty)
3. Elastic net regularization term (Combination of L2 and L1 Penalty)

3.2 Regularization term 1: Ridge Regression (L2 Penalty)

In a **Logistic Ridge Regression model**, we seek to find the optimal values of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ that minimize the following objective function (for a given $\lambda > 0$).

$$\underset{\hat{\beta}_0,\hat{\beta}_1,\dots,\hat{\beta}_p}{\text{MIN}} \; -2 \cdot LLF\left(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p\right) + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2 + \cdots + \hat{\beta}_p^2)$$

The best solutions select values of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, in which

- $LLF\left(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p\right)$ is ________________, which means the model will be encouraged to have ______________________________.
- $\hat{\beta}_1^2 + \hat{\beta}_2^2 + \cdots + \hat{\beta}_p^2$ is _________________, which means the model will be encouraged to have ______________________________.

A higher value of λ means

- The model will be __________ likely to select values of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ in which $\hat{\beta}_1^2 + \hat{\beta}_2^2 + \cdots + \hat{\beta}_p^2$ is low, at the expense of creating models with high value for $LLF\left(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p\right)$.

3.3 REGULARIZATION TERM 2: LASSO (L1 PENALTY)

In a **LASSO logistic regression model**, we seek to find the optimal values of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize the following objective function (for a given $\lambda > 0$).

$$\underset{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p}{\text{MIN}} -2 \cdot LLF\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right) + \lambda(|\hat{\beta}_1| + |\hat{\beta}_2| + \cdots + |\hat{\beta}_p|)$$

The best solutions select values of $\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_p$, in which

- $LLF\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right)$ is ___________________, which means the model will be encouraged to have

  ______________________.

- $|\hat{\beta}_1| + |\hat{\beta}_2| + \cdots + |\hat{\beta}_p|$ is ____________________, which means the model will be encouraged to have __________________________________.

A higher value of λ means

- The model will be ____________ likely to select values of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ in which $|\hat{\beta}_1| + |\hat{\beta}_2| + \cdots + |\hat{\beta}_p|$ is low, at the expense of creating models with high value for $LLF\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right)$.

3.4 COMPARING LASSO (L1) PENALTY TO THE RIDGE REGRESSION (L2) PENALTY

Why would you want to use ridge regression penalty over a LASSO penalty, and vice versa?

Ridge Regression (L2) Penalty $\hat{\boldsymbol{\beta}}_1^2 + \hat{\boldsymbol{\beta}}_2^2 + \cdots + \hat{\boldsymbol{\beta}}_p^2$

LASSO (L1) Penalty $|\hat{\boldsymbol{\beta}}_1| + |\hat{\boldsymbol{\beta}}_2| + \cdots + |\hat{\boldsymbol{\beta}}_p|$

3.5 Regularization term 3: Elastic Net (L2 and L1 Penalty Combination)

In an **Elastic Net Logistic regression model**, we seek to find the optimal values of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ that minimize the following objective function (for a given $\lambda > 0 \; and \; 0 \leq \alpha \leq 1$)

$$\underset{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p}{\text{MIN}} -2 \cdot LLF\left(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p\right) + \lambda \left[\alpha\left(|\hat{\beta}_1| + |\hat{\beta}_2| + \cdots + |\hat{\beta}_p|\right) + \tfrac{1-\alpha}{2}\left(\hat{\beta}_1^2 + \hat{\beta}_2^2 + \cdots + \hat{\beta}_p^2\right)\right]$$

<u>The best solutions select values of $\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1, \dots, \hat{\boldsymbol{\beta}}_p$, in which</u>

- $LLF\left(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p\right)$ is ___________________, which means the model will be encouraged to have

  ______________________.

- $|\hat{\beta}_1| + |\hat{\beta}_2| + \cdots + |\hat{\beta}_p|$ AND $(\hat{\beta}_1^2 + \hat{\beta}_2^2 + \cdots + \hat{\beta}_p^2)$ are ____________________, which means the model will be encouraged to have __________________________________.

<u>A α closer to 1 means</u>

- The model will be ____________ likely to select values of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ that resemble solutions to the

  _______________________.

<u>A α closer to 0 means</u>

- The model will be ____________ likely to select values of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ that resemble solutions to the

  ______________________.

<u>A higher value of λ means</u>

- The model will be ____________ likely to select values of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ in which $|\hat{\beta}_1| + |\hat{\beta}_2| + \cdots + |\hat{\beta}_p|$ AND $(\hat{\beta}_1^2 + \hat{\beta}_2^2 + \cdots + \hat{\beta}_p^2)$ are low, at the expense of creating models with high value for $LLF\left(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p\right)$.

3.5 How to Use Regularized Regression to select a Parsimonious Model

1. Fit a *single* regression model with a **regularization term.**

2. [OPTIONAL STEP]: Examine the resulting slopes of this this fitted model with **regularization term**, and use this to determine generate some ideas of possible “reduced models” to test.

3. [OPTIONAL STEP]: Fit each of these “reduced models” as well as the “full model” (ie. the model with all possible explanatory variables you are considering) and use the AIC, BIC, and/or log-likelihood ratio test to determine which model is best.

Go to the Unit 20 notebook, section 3 for an exercise in using regularized logistic regression models.