Predictive Analytics Lecture 3

Adam Kapelner

Stat 422/722
at The Wharton School of the University of Pennsylvania

January 31 & February 1, 2017

The Coin Example from Last Class I

I want to explain the coin example from last class in the context of likelihood. Imagine you flip a coin three times and get heads, heads, tails; thus, $y_1 = 1, y_2 = 1, y_3 = 0$. There is a true probability of heads called θ. We don't know it.

What is the probability of the data? We employ the mass / density function:

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3; \theta) = \prod_{i=1}^{3} \mathbb{P}(Y_i = y_i; \theta) =$$

The Coin Example from Last Class I

I want to explain the coin example from last class in the context of likelihood. Imagine you flip a coin three times and get heads, heads, tails; thus, $y_1 = 1, y_2 = 1, y_3 = 0$. There is a true probability of heads called θ. We don't know it.

What is the probability of the data? We employ the mass / density function:

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3; \theta) = \prod_{i=1}^{3} \mathbb{P}(Y_i = y_i; \theta) = \prod_{i=1}^{3} \theta^{y_i} (1-\theta)^{1-y_i}$$

$$=$$

The Coin Example from Last Class I

I want to explain the coin example from last class in the context of likelihood. Imagine you flip a coin three times and get heads, heads, tails; thus, $y_1 = 1, y_2 = 1, y_3 = 0$. There is a true probability of heads called θ. We don't know it.

What is the probability of the data? We employ the mass / density function:

$$\mathbb{P}\left(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3; \theta\right) = \prod_{i=1}^{3} \mathbb{P}\left(Y_i = y_i; \theta\right) = \prod_{i=1}^{3} \theta^{y_i} \left(1-\theta\right)^{1-y_i}$$

$$= \left(\theta^{(1)} \left(1-\theta\right)^{1-(1)}\right) \left(\theta^{(1)} \left(1-\theta\right)^{1-(1)}\right) \left(\theta^{(0)} \left(1-\theta\right)^{1-(0)}\right)$$

The Coin Example from Last Class I

I want to explain the coin example from last class in the context of likelihood. Imagine you flip a coin three times and get heads, heads, tails; thus, $y_1 = 1, y_2 = 1, y_3 = 0$. There is a true probability of heads called θ. We don't know it.

What is the probability of the data? We employ the mass / density function:

$$
\begin{aligned}
\mathbb{P}(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3; \theta) &= \prod_{i=1}^{3} \mathbb{P}(Y_i = y_i; \theta) = \prod_{i=1}^{3} \theta^{y_i} (1-\theta)^{1-y_i} \\
&= \left(\theta^{(1)} (1-\theta)^{1-(1)}\right) \left(\theta^{(1)} (1-\theta)^{1-(1)}\right) \left(\theta^{(0)} (1-\theta)^{1-(0)}\right) \\
&= \theta^2 (1-\theta)
\end{aligned}
$$

And now we can calculate the probability of seeing the data assuming θ. Assume $\theta = 0.5$ then,

$$\mathbb{P}(Y_1 = 1, Y_2 = 1, Y_3 = 0; \theta = 0.5) = 0.5^2 (1 - 0.5) =$$

The Coin Example from Last Class I

I want to explain the coin example from last class in the context of likelihood. Imagine you flip a coin three times and get heads, heads, tails; thus, $y_1 = 1, y_2 = 1, y_3 = 0$. There is a true probability of heads called θ. We don't know it.

What is the probability of the data? We employ the mass / density function:

$$\begin{aligned}
\mathbb{P}(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3; \theta) &= \prod_{i=1}^{3} \mathbb{P}(Y_i = y_i; \theta) = \prod_{i=1}^{3} \theta^{y_i} (1-\theta)^{1-y_i} \\
&= \left(\theta^{(1)} (1-\theta)^{1-(1)}\right) \left(\theta^{(1)} (1-\theta)^{1-(1)}\right) \left(\theta^{(0)} (1-\theta)^{1-(0)}\right) \\
&= \theta^2 (1-\theta)
\end{aligned}$$

And now we can calculate the probability of seeing the data assuming θ. Assume $\theta = 0.5$ then,

$$\mathbb{P}(Y_1 = 1, Y_2 = 1, Y_3 = 0; \theta = 0.5) = 0.5^2(1-0.5) = 0.125$$

The Coin Example from Last Class II

Now we ask the inverse question.

The Coin Example from Last Class II

Now we ask the inverse question. If we saw this data $y_1 = 1$, $y_2 = 1$, $y_3 = 0$, what is the most likely model? Equivalently, what is the most likely value of θ?

The Coin Example from Last Class II

Now we ask the inverse question. If we saw this data $y_1 = 1$, $y_2 = 1$, $y_3 = 0$, what is the most likely model? Equivalently, what is the most likely value of θ? We first write down the likelihood function which it's easy because it's the same as the mass / density function

$$\mathcal{L}\left(\theta; Y_1 = 1, Y_2 = 1, Y_3 = 0\right) =$$

The Coin Example from Last Class II

Now we ask the inverse question. If we saw this data $y_1 = 1$, $y_2 = 1$, $y_3 = 0$, what is the most likely model? Equivalently, what is the most likely value of θ? We first write down the likelihood function which it's easy because it's the same as the mass / density function

$$\mathcal{L}(\theta; Y_1 = 1, Y_2 = 1, Y_3 = 0) = \mathbb{P}(Y_1 = 1, Y_2 = 1, Y_3 = 0; \theta) =$$

The Coin Example from Last Class II

Now we ask the inverse question. If we saw this data $y_1 = 1$, $y_2 = 1$, $y_3 = 0$, what is the most likely model? Equivalently, what is the most likely value of θ? We first write down the likelihood function which it's easy because it's the same as the mass / density function

$$\mathcal{L}(\theta; Y_1 = 1, Y_2 = 1, Y_3 = 0) = \mathbb{P}(Y_1 = 1, Y_2 = 1, Y_3 = 0; \theta) = \theta^2(1-\theta)$$

And now we pick the value of θ which maximizes the likelihood,

The Coin Example from Last Class II

Now we ask the inverse question. If we saw this data $y_1 = 1$, $y_2 = 1$, $y_3 = 0$, what is the most likely model? Equivalently, what is the most likely value of θ? We first write down the likelihood function which it's easy because it's the same as the mass / density function

$$\mathcal{L}(\theta; Y_1 = 1, Y_2 = 1, Y_3 = 0) = \mathbb{P}(Y_1 = 1, Y_2 = 1, Y_3 = 0; \theta) = \theta^2(1 - \theta)$$

And now we pick the value of θ which maximizes the likelihood,

$$\hat{\theta} := \underset{\theta \in \Theta}{\arg\max} \{\mathcal{L}(\theta; x)\}$$

How?

The Coin Example from Last Class II

Now we ask the inverse question. If we saw this data $y_1 = 1$, $y_2 = 1$, $y_3 = 0$, what is the most likely model? Equivalently, what is the most likely value of θ? We first write down the likelihood function which it's easy because it's the same as the mass / density function

$$\mathcal{L}(\theta; Y_1 = 1, Y_2 = 1, Y_3 = 0) = \mathbb{P}(Y_1 = 1, Y_2 = 1, Y_3 = 0; \theta) = \theta^2(1 - \theta)$$

And now we pick the value of θ which maximizes the likelihood,

$$\hat{\theta} := \underset{\theta \in \Theta}{\arg\max} \, \{\mathcal{L}(\theta; x)\}$$

How? Calculus 101: we take the derivative, set it equal to zero and solve for θ:

$$\frac{d}{d\theta}\left[\theta^2(1 - \theta)\right] =$$

The Coin Example from Last Class II

Now we ask the inverse question. If we saw this data $y_1 = 1$, $y_2 = 1$, $y_3 = 0$, what is the most likely model? Equivalently, what is the most likely value of θ? We first write down the likelihood function which it's easy because it's the same as the mass / density function

$$\mathcal{L}\left(\theta; Y_1 = 1, Y_2 = 1, Y_3 = 0\right) = \mathbb{P}\left(Y_1 = 1, Y_2 = 1, Y_3 = 0; \theta\right) = \theta^2(1-\theta)$$

And now we pick the value of θ which maximizes the likelihood,

$$\hat{\theta} := \arg\max_{\theta \in \Theta} \left\{\mathcal{L}\left(\theta; x\right)\right\}$$

How? Calculus 101: we take the derivative, set it equal to zero and solve for θ:

$$\frac{d}{d\theta}\left[\theta^2(1-\theta)\right] = \frac{d}{d\theta}\left[\theta^2 - \theta^3\right] =$$

The Coin Example from Last Class II

Now we ask the inverse question. If we saw this data $y_1 = 1$, $y_2 = 1$, $y_3 = 0$, what is the most likely model? Equivalently, what is the most likely value of θ? We first write down the likelihood function which it's easy because it's the same as the mass / density function

$$\mathcal{L}(\theta; Y_1 = 1, Y_2 = 1, Y_3 = 0) = \mathbb{P}(Y_1 = 1, Y_2 = 1, Y_3 = 0; \theta) = \theta^2(1-\theta)$$

And now we pick the value of θ which maximizes the likelihood,

$$\hat{\theta} := \underset{\theta \in \Theta}{\arg\max} \ \{\mathcal{L}(\theta; x)\}$$

How? Calculus 101: we take the derivative, set it equal to zero and solve for θ:

$$\frac{d}{d\theta}\left[\theta^2(1-\theta)\right] = \frac{d}{d\theta}\left[\theta^2 - \theta^3\right] = 2\theta - 3\theta^2$$

and set it equal to zero:

$$0 = 2\theta - 3\theta^2 = \theta(2 - 3\theta) \Rightarrow 0 = 2 - 3\theta \Rightarrow \hat{\theta}_{\mathrm{MLE}} = \frac{2}{3}$$

The Coin Example from Last Class II

Now we ask the inverse question. If we saw this data $y_1 = 1$, $y_2 = 1$, $y_3 = 0$, what is the most likely model? Equivalently, what is the most likely value of θ? We first write down the likelihood function which it's easy because it's the same as the mass / density function

$$\mathcal{L}(\theta; Y_1 = 1, Y_2 = 1, Y_3 = 0) = \mathbb{P}(Y_1 = 1, Y_2 = 1, Y_3 = 0; \theta) = \theta^2(1-\theta)$$

And now we pick the value of θ which maximizes the likelihood,

$$\hat{\theta} := \underset{\theta \in \Theta}{\arg\max} \; \{\mathcal{L}(\theta; x)\}$$

How? Calculus 101: we take the derivative, set it equal to zero and solve for θ:

$$\frac{d}{d\theta}\left[\theta^2(1-\theta)\right] = \frac{d}{d\theta}\left[\theta^2 - \theta^3\right] = 2\theta - 3\theta^2$$

and set it equal to zero:

$$0 = 2\theta - 3\theta^2 = \theta(2 - 3\theta) \Rightarrow 0 = 2 - 3\theta \Rightarrow \hat{\theta}_{\mathrm{MLE}} = \frac{2}{3}$$

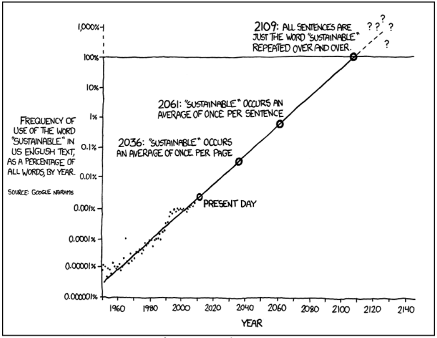
i.e. the most likely model for this data is a weighted coin with probability of heads of 2/3.

Extrapolation

Data driven approaches are all focused on accuracy during **interpolation** (estimation within a known range). **Extrapolation** (estimation outside of a known range) brings trouble.

Extrapolation

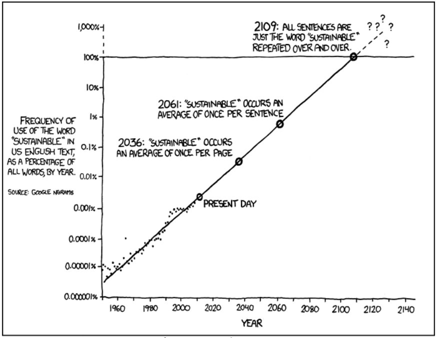
Data driven approaches are all focused on accuracy during **interpolation** (estimation within a known range). **Extrapolation** (estimation outside of a known range) brings trouble.



THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

Extrapolation

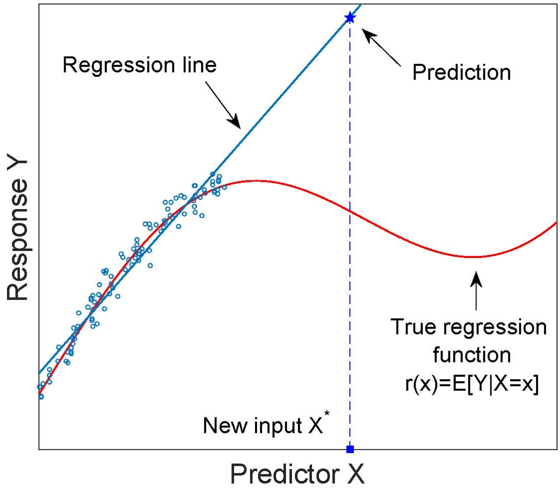
Data driven approaches are all focused on accuracy during **interpolation** (estimation within a known range). **Extrapolation** (estimation outside of a known range) brings trouble.



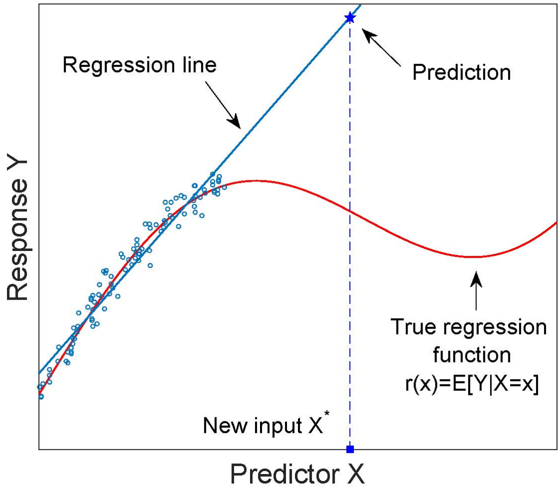
THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

It is important to ask the question for a new observation $\boldsymbol{x}^*$ if it is within the space of $\boldsymbol{x}$'s in the historical data. (Hardly anyone does this when $p > 2$... but you should)!

Reconciliation of These Silly Cartoons



Reconciliation of These Silly Cartoons



Be aware that extrapolation methods of different algorithms differ considerably! [R Demo]

Dataframe Design

We spoke a lot about featurization i.e. selecting the columns in the dataframe (these are the predictors to measure). Once we did this, we can then go out and sample observations and then measure each for their predictor values.

Dataframe Design

We spoke a lot about featurization i.e. selecting the columns in the dataframe (these are the predictors to measure). Once we did this, we can then go out and sample observations and then measure each for their predictor values.

But we didn't speak at all about selecting the observations themselves (assuming you have some modicum of control of selecting your data). Two things to consider:

Dataframe Design

We spoke a lot about featurization i.e. selecting the columns in the dataframe (these are the predictors to measure). Once we did this, we can then go out and sample observations and then measure each for their predictor values.

But we didn't speak at all about selecting the observations themselves (assuming you have some modicum of control of selecting your data). Two things to consider:

1. **Generalizability** refers to the ability of the model to generalize, or be **externally valid** when considering new observations. This comes down to sampling observations from the same population as your new data you wish to predict (pretty obvious). Sometimes difficult in practice!

Dataframe Design

We spoke a lot about featurization i.e. selecting the columns in the dataframe (these are the predictors to measure). Once we did this, we can then go out and sample observations and then measure each for their predictor values.

But we didn't speak at all about selecting the observations themselves (assuming you have some modicum of control of selecting your data). Two things to consider:

1. **Generalizability** refers to the ability of the model to generalize, or be **externally valid** when considering new observations. This comes down to sampling observations from the same population as your new data you wish to predict (pretty obvious). Sometimes difficult in practice! Extrapolation??

Dataframe Design

We spoke a lot about featurization i.e. selecting the columns in the dataframe (these are the predictors to measure). Once we did this, we can then go out and sample observations and then measure each for their predictor values.

But we didn't speak at all about selecting the observations themselves (assuming you have some modicum of control of selecting your data). Two things to consider:

1. **Generalizability** refers to the ability of the model to generalize, or be **externally valid** when considering new observations. This comes down to sampling observations from the same population as your new data you wish to predict (pretty obvious). Sometimes difficult in practice! Extrapolation??
2. Optimal Design

Dataframe Design

We spoke a lot about featurization i.e. selecting the columns in the dataframe (these are the predictors to measure). Once we did this, we can then go out and sample observations and then measure each for their predictor values.

But we didn't speak at all about selecting the observations themselves (assuming you have some modicum of control of selecting your data). Two things to consider:

1. **Generalizability** refers to the ability of the model to generalize, or be **externally valid** when considering new observations. This comes down to sampling observations from the same population as your new data you wish to predict (pretty obvious). Sometimes difficult in practice! Extrapolation??
2. Optimal Design

What to measure? ✓

Dataframe Design

We spoke a lot about featurization i.e. selecting the columns in the dataframe (these are the predictors to measure). Once we did this, we can then go out and sample observations and then measure each for their predictor values.

But we didn't speak at all about selecting the observations themselves (assuming you have some modicum of control of selecting your data). Two things to consider:

1. **Generalizability** refers to the ability of the model to generalize, or be **externally valid** when considering new observations. This comes down to sampling observations from the same population as your new data you wish to predict (pretty obvious). Sometimes difficult in practice! Extrapolation??
2. Optimal Design

What to measure? ✓ Who to measure it on???

Optimal Design for Inferring one Slope

Question: assume (a) OLS with $p = 1$ and

Optimal Design for Inferring one Slope

Question: assume (a) OLS with $p = 1$ and (b) we only care about inference for β_1 (intercept inessential) and

Optimal Design for Inferring one Slope

Question: assume (a) OLS with $p = 1$ and (b) we only care about inference for β_1 (intercept inessential) and (c) we are at liberty to sample any x values live in their set $\mathbb{X}$ e.g. $\in [x_m, x_M]$. What should these n values of x be?

Let $x_m = 0$, $x_M = 1$ and $n = 10$. The best inference for β_1 means ...

Optimal Design for Inferring one Slope

Question: assume (a) OLS with $p = 1$ and (b) we only care about inference for β_1 (intercept inessential) and (c) we are at liberty to sample any x values live in their set $\mathbb{X}$ e.g. $\in [x_m, x_M]$. What should these n values of x be?

Let $x_m = 0$, $x_M = 1$ and $n = 10$. The best inference for β_1 means ... $\mathbb{S}\mathrm{E}\left[\hat{\beta}_1\right]$ is minimum (since unbiasedness is guaranteed since we are using minimum SSE estimation). Design strategies for the x's:

Optimal Design for Inferring one Slope

Question: assume (a) OLS with $p = 1$ and (b) we only care about inference for β_1 (intercept inessential) and (c) we are at liberty to sample any x values live in their set $\mathbb{X}$ e.g. $\in [x_m, x_M]$. What should these n values of x be?

Let $x_m = 0$, $x_M = 1$ and $n = 10$. The best inference for β_1 means ... $\mathbb{S}\mathrm{E}\left[\hat{\beta}_1\right]$ is minimum (since unbiasedness is guaranteed since we are using minimum SSE estimation). Design strategies for the x's:

1. Random sampling

Optimal Design for Inferring one Slope

Question: assume (a) OLS with $p = 1$ and (b) we only care about inference for β_1 (intercept inessential) and (c) we are at liberty to sample any x values live in their set $\mathbb{X}$ e.g. $\in [x_m, x_M]$. What should these n values of x be?

Let $x_m = 0$, $x_M = 1$ and $n = 10$. The best inference for β_1 means ... $\mathbb{S}\mathrm{E}\left[\hat{\beta}_1\right]$ is minimum (since unbiasedness is guaranteed since we are using minimum SSE estimation). Design strategies for the x's:

1. Random sampling
2. Uniform spacing: $\{0, 0.111, 0.222, \ldots, 0.999\}$

Optimal Design for Inferring one Slope

Question: assume (a) OLS with $p = 1$ and (b) we only care about inference for β_1 (intercept inessential) and (c) we are at liberty to sample any x values live in their set $\mathbb{X}$ e.g. $\in [x_m, x_M]$. What should these n values of x be?

Let $x_m = 0$, $x_M = 1$ and $n = 10$. The best inference for β_1 means ... $\mathbb{S}\mathrm{E}\left[\hat{\beta}_1\right]$ is minimum (since unbiasedness is guaranteed since we are using minimum SSE estimation). Design strategies for the x's:

1. Random sampling
2. Uniform spacing: $\{0, 0.111, 0.222, \ldots, 0.999\}$
3. Something else?

[R demo]

Optimal Design: Split Between Extremes

Recall the formula from Stat 102 / 613:

$$\mathbb{SE}\left[\hat{\beta}_1\right] = \sqrt{\frac{MSE}{(n-1)s_x^2}} = \frac{RMSE}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

How can we make this small?

Optimal Design: Split Between Extremes

Recall the formula from Stat 102 / 613:

$$\mathbb{SE}\left[\hat{\beta}_1\right] = \sqrt{\frac{MSE}{(n-1)s_x^2}} = \frac{RMSE}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

How can we make this small?

1. Maximize n (duh) ... so we'll assume n is fixed by a budget constraint so now what?

Optimal Design: Split Between Extremes

Recall the formula from Stat 102 / 613:

$$\mathbb{SE}\left[\hat{\beta}_1\right] = \sqrt{\frac{MSE}{(n-1)s_x^2}} = \frac{RMSE}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

How can we make this small?

1. Maximize n (duh) ... so we'll assume n is fixed by a budget constraint so now what?
2. Minimize the numerator, $RMSE$ i.e. minimize the MSE i.e. minimize SSE. Can we do this?

Optimal Design: Split Between Extremes

Recall the formula from Stat 102 / 613:

$$\mathbb{SE}\left[\hat{\beta}_1\right] = \sqrt{\frac{MSE}{(n-1)s_x^2}} = \frac{RMSE}{\sqrt{\displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

How can we make this small?

1. Maximize n (duh) ... so we'll assume n is fixed by a budget constraint so now what?
2. Minimize the numerator, $RMSE$ i.e. minimize the MSE i.e. minimize SSE. Can we do this? Yes by picking the closest $\hat{\beta}_1$ to β_1 (which we already do).

Optimal Design: Split Between Extremes

Recall the formula from Stat 102 / 613:

$$\mathbb{SE}\left[\hat{\beta}_1\right] = \sqrt{\frac{MSE}{(n-1)s_x^2}} = \frac{RMSE}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

How can we make this small?

1. Maximize n (duh) ... so we'll assume n is fixed by a budget constraint so now what?
2. Minimize the numerator, $RMSE$ i.e. minimize the MSE i.e. minimize SSE. Can we do this? Yes by picking the closest $\hat{\beta}_1$ to β_1 (which we already do).
3. Maximize the denominator $(n-1)s_x^2$. Since n is already maximized, we can pick $x_1, \ldots, x_n$ to maximize s_x^2, the sample variance of the predictor. How?

Optimal Design: Split Between Extremes

Recall the formula from Stat 102 / 613:

$$\mathbb{SE}\left[\hat{\beta}_1\right] = \sqrt{\frac{MSE}{(n-1)s_x^2}} = \frac{RMSE}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

How can we make this small?

1. Maximize n (duh) ... so we'll assume n is fixed by a budget constraint so now what?
2. Minimize the numerator, $RMSE$ i.e. minimize the MSE i.e. minimize SSE. Can we do this? Yes by picking the closest $\hat{\beta}_1$ to β_1 (which we already do).
3. Maximize the denominator $(n-1)s_x^2$. Since n is already maximized, we can pick $x_1, \ldots, x_n$ to maximize s_x^2, the sample variance of the predictor. How? Put half of the x's at x_m and the other half at x_M thereby

Optimal Design: Split Between Extremes

Recall the formula from Stat 102 / 613:

$$\mathbb{SE}\left[\hat{\beta}_1\right] = \sqrt{\frac{MSE}{(n-1)s_x^2}} = \frac{RMSE}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

How can we make this small?

1. Maximize n (duh) ... so we'll assume n is fixed by a budget constraint so now what?
2. Minimize the numerator, $RMSE$ i.e. minimize the MSE i.e. minimize SSE. Can we do this? Yes by picking the closest $\hat{\beta}_1$ to β_1 (which we already do).
3. Maximize the denominator $(n-1)s_x^2$. Since n is already maximized, we can pick $x_1, \ldots, x_n$ to maximize s_x^2, the sample variance of the predictor. How? Put half of the x's at x_m and the other half at x_M thereby maximizing the distance from the x's to

Optimal Design: Split Between Extremes

Recall the formula from Stat 102 / 613:

$$\mathbb{SE}\left[\hat{\beta}_1\right] = \sqrt{\frac{MSE}{(n-1)s_x^2}} = \frac{RMSE}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

How can we make this small?

1. Maximize n (duh) ... so we'll assume n is fixed by a budget constraint so now what?
2. Minimize the numerator, $RMSE$ i.e. minimize the MSE i.e. minimize SSE. Can we do this? Yes by picking the closest $\hat{\beta}_1$ to β_1 (which we already do).
3. Maximize the denominator $(n-1)s_x^2$. Since n is already maximized, we can pick $x_1, \ldots, x_n$ to maximize s_x^2, the sample variance of the predictor. How? Put half of the x's at x_m and the other half at x_M thereby maximizing the distance from the x's to $\bar{x}$.

General Optimal Design of Linear Models

We seek the best linear approximation of $f(x)$ which is $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. We pick the $\boldsymbol{x}$'s to give us the best linear approximation. What criteria? JMP gives two ways:

1. Note: $\mathbb{V}\text{ar}\left[\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right] = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$

General Optimal Design of Linear Models

We seek the best linear approximation of $f(x)$ which is $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. We pick the $\boldsymbol{x}$'s to give us the best linear approximation. What criteria? JMP gives two ways:

1. Note: $\mathbb{V}\text{ar}\left[\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right] = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$

   D-optimality: maximize $\left|\boldsymbol{X}^T \boldsymbol{X}\right|$ — this minimizes the variance-covariance among the parameter estimates.

General Optimal Design of Linear Models

We seek the best linear approximation of $f(x)$ which is $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. We pick the $\boldsymbol{x}$'s to give us the best linear approximation. What criteria? JMP gives two ways:

1. Note: $\mathbb{V}\text{ar}\left[\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right] = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$

   D-optimality: maximize $\left|\boldsymbol{X}^T \boldsymbol{X}\right|$ — this minimizes the variance-covariance among the parameter estimates.

2. Note: $\mathbb{V}\text{ar}\left[\hat{Y}_1, \ldots, \hat{Y}_n\right] = \sigma^2 \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T$

General Optimal Design of Linear Models

We seek the best linear approximation of $f(x)$ which is $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. We pick the $\boldsymbol{x}$'s to give us the best linear approximation. What criteria? JMP gives two ways:

1. Note: $\mathbb{V}\text{ar}\left[\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right] = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$

   D-optimality: maximize $\left|\boldsymbol{X}^T \boldsymbol{X}\right|$ — this minimizes the variance-covariance among the parameter estimates.

2. Note: $\mathbb{V}\text{ar}\left[\hat{Y}_1, \ldots, \hat{Y}_n\right] = \sigma^2 \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T$

   I-optimality: minimize the average prediction variance over the design space. I'm unsure how JMP defined "design space".

[R Demo]

General Optimal Design of Linear Models

We seek the best linear approximation of $f(x)$ which is $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. We pick the $\boldsymbol{x}$'s to give us the best linear approximation. What criteria? JMP gives two ways:

1. Note: $\mathbb{V}\text{ar}\left[\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right] = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$

   D-optimality: maximize $\left|\boldsymbol{X}^T \boldsymbol{X}\right|$ — this minimizes the variance-covariance among the parameter estimates.

2. Note: $\mathbb{V}\text{ar}\left[\hat{Y}_1, \ldots, \hat{Y}_n\right] = \sigma^2 \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T$

   I-optimality: minimize the average prediction variance over the design space. I'm unsure how JMP defined "design space".

[R Demo] What did we learn?

General Optimal Design of Linear Models

We seek the best linear approximation of $f(x)$ which is $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. We pick the $\boldsymbol{x}$'s to give us the best linear approximation. What criteria? JMP gives two ways:

1. Note: $\mathbb{V}\text{ar}\left[\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right] = \sigma^2 \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$

   D-optimality: maximize $\left|\boldsymbol{X}^T\boldsymbol{X}\right|$ — this minimizes the variance-covariance among the parameter estimates.

2. Note: $\mathbb{V}\text{ar}\left[\hat{Y}_1, \ldots, \hat{Y}_n\right] = \sigma^2 \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$

   I-optimality: minimize the average prediction variance over the design space. I'm unsure how JMP defined "design space".

[R Demo] What did we learn? For linear models with no polynomials or interactions, keep the observations as close to the minimimums and maximums as possible. For linear models with polynomials and interactions (more non-parametric than parametric),

General Optimal Design of Linear Models

We seek the best linear approximation of $f(x)$ which is $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. We pick the $\boldsymbol{x}$'s to give us the best linear approximation. What criteria? JMP gives two ways:

1. Note: $\mathbb{V}\text{ar}\left[\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right] = \sigma^2 \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$

   D-optimality: maximize $\left|\boldsymbol{X}^T\boldsymbol{X}\right|$ — this minimizes the variance-covariance among the parameter estimates.

2. Note: $\mathbb{V}\text{ar}\left[\hat{Y}_1, \ldots, \hat{Y}_n\right] = \sigma^2 \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$

   I-optimality: minimize the average prediction variance over the design space. I'm unsure how JMP defined "design space".

[R Demo] What did we learn? For linear models with no polynomials or interactions, keep the observations as close to the minimimums and maximums as possible. For linear models with polynomials and interactions (more non-parametric than parametric), keep most towards the minimums and maximums and some in the center of the input space.

Modeling Categorical Responses

Previously the response y was continuous and via the OLS assumptions we obtained the statistical model,

$$Y \overset{ind}{\sim} \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,\, \sigma^2\right)$$

If the response y is categorical, can we still use this?

Modeling Categorical Responses

Previously the response y was continuous and via the OLS assumptions we obtained the statistical model,

$$Y \overset{ind}{\sim} \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,\, \sigma^2\right)$$

If the response y is categorical, can we still use this? No... the only elements in the support of the r.v. Y are the levels only. [JMP Churn Dataframe]

Modeling Categorical Responses

Previously the response y was continuous and via the OLS assumptions we obtained the statistical model,

$$Y \overset{ind}{\sim} \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,\, \sigma^2\right)$$

If the response y is categorical, can we still use this? No... the only elements in the support of the r.v. Y are the levels only. [JMP Churn Dataframe]

First, assume Y is binary i.e. zero or one. The model (AKA "classifier") we use is...

Modeling Categorical Responses

Previously the response y was continuous and via the OLS assumptions we obtained the statistical model,

$$Y \overset{ind}{\sim} \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,\, \sigma^2\right)$$

If the response y is categorical, can we still use this? No... the only elements in the support of the r.v. Y are the levels only. [JMP Churn Dataframe]

First, assume Y is binary i.e. zero or one. The model (AKA "classifier") we use is...

$$Y \sim \text{Bernoulli}\left(f(x_1, \ldots, x_p)\right)$$

since $\mathbb{E}\left[Y \mid x_1, \ldots, x_p\right] = f(x_1, \ldots, x_p)$, then f is still the conditional expectation function like before except now it varies only within

Modeling Categorical Responses

Previously the response y was continuous and via the OLS assumptions we obtained the statistical model,

$$Y \overset{ind}{\sim} \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,\, \sigma^2\right)$$

If the response y is categorical, can we still use this? No... the only elements in the support of the r.v. Y are the levels only. [JMP Churn Dataframe]

First, assume Y is binary i.e. zero or one. The model (AKA "classifier") we use is...

$$Y \sim \text{Bernoulli}\left(f(x_1, \ldots, x_p)\right)$$

since $\mathbb{E}\left[Y \mid x_1, \ldots, x_p\right] = f(x_1, \ldots, x_p)$, then f is still the conditional expectation function like before except now it varies only within $[0, 1]$ and it is the same as

Modeling Categorical Responses

Previously the response y was continuous and via the OLS assumptions we obtained the statistical model,

$$Y \overset{ind}{\sim} \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,\, \sigma^2\right)$$

If the response y is categorical, can we still use this? No... the only elements in the support of the r.v. Y are the levels only. [JMP Churn Dataframe]

First, assume Y is binary i.e. zero or one. The model (AKA "classifier") we use is...

$$Y \sim \text{Bernoulli}\left(f(x_1, \ldots, x_p)\right)$$

since $\mathbb{E}\left[Y \mid x_1, \ldots, x_p\right] = f(x_1, \ldots, x_p)$, then f is still the conditional expectation function like before except now it varies only within $[0, 1]$ and it is the same as $\mathbb{P}\left(Y = 1 \mid x_1, \ldots, x_p\right)$.

Linear $f(x)$?

We can model $f(x)$ as the simple linear function but this returns values smaller than 0 and larger than 1 and thus it cannot be the conditional expectation function! Why?

Linear $f(x)$?

We can model $f(x)$ as the simple linear function but this returns values smaller than 0 and larger than 1 and thus it cannot be the conditional expectation function! Why? Lines with nonzero slope vary between $(-\infty, +\infty)$.

Linear $f(x)$?

We can model $f(x)$ as the simple linear function but this returns values smaller than 0 and larger than 1 and thus it cannot be the conditional expectation function! Why? Lines with nonzero slope vary between $(-\infty, +\infty)$.

We need a “link function” to connect the linear function to the restricted support of the response:

$$\lambda(f_{\mathbb{R}}(x_1, \ldots, x_p)) = f(x_1, \ldots, x_p)$$

And the parametric assumption would be

$$\lambda(s_{\mathbb{R}}(x_1, \ldots, x_p; \theta_1, \ldots, \theta_\ell)) = s(x_1, \ldots, x_p; \theta_1, \ldots, \theta_\ell)$$

And assuming a linear form of $s_{\mathbb{R}}$,

Linear $f(x)$?

We can model $f(x)$ as the simple linear function but this returns values smaller than 0 and larger than 1 and thus it cannot be the conditional expectation function! Why? Lines with nonzero slope vary between $(-\infty, +\infty)$.

We need a "link function" to connect the linear function to the restricted support of the response:

$$\lambda(f_{\mathbb{R}}(x_1, \ldots, x_p)) = f(x_1, \ldots, x_p)$$

And the parametric assumption would be

$$\lambda(s_{\mathbb{R}}(x_1, \ldots, x_p; \theta_1, \ldots, \theta_\ell)) = s(x_1, \ldots, x_p; \theta_1, \ldots, \theta_\ell)$$

And assuming a linear form of $s_{\mathbb{R}}$,

$$\lambda(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p) = \ ?$$

Choice of λ?

We just need $\lambda : \mathbb{R} \to [0, 1]$. There are infinite λ's to choose from... but I've only seen three used:

Choice of λ?

We just need $\lambda : \mathbb{R} \to [0, 1]$. There are infinite λ's to choose from... but I've only seen three used:

1. Logistic link: $\lambda(w) = \frac{e^w}{1+e^w}$ (most common)

Choice of λ?

We just need $\lambda : \mathbb{R} \to [0, 1]$. There are infinite λ's to choose from... but I've only seen three used:

1. Logistic link: $\lambda(w) = \frac{e^w}{1+e^w}$ (most common)
2. Inverse normal (probit) link: $\lambda(w) = \Phi^{-1}(w)$ where Φ is the normal CDF function (somewhat common)

Choice of λ?

We just need $\lambda : \mathbb{R} \to [0, 1]$. There are infinite λ's to choose from... but I've only seen three used:

1. Logistic link: $\lambda(w) = \frac{e^w}{1+e^w}$ (most common)
2. Inverse normal (probit) link: $\lambda(w) = \Phi^{-1}(w)$ where Φ is the normal CDF function (somewhat common)
3. Complementary Log-log (cloglog) link: $\lambda(w) = \ln(-\ln(w))$ (rare!)

Choice of λ?

We just need $\lambda : \mathbb{R} \to [0,1]$. There are infinite λ's to choose from... but I've only seen three used:

1. Logistic link: $\lambda(w) = \frac{e^w}{1+e^w}$ (most common)
2. Inverse normal (probit) link: $\lambda(w) = \Phi^{-1}(w)$ where Φ is the normal CDF function (somewhat common)
3. Complementary Log-log (cloglog) link: $\lambda(w) = \ln(-\ln(w))$ (rare!)

Herein, we focus on the logistic link. First, short detour. Define $p := \mathbb{P}(Y = 1)$. We can think about probability in another way:

$$odds(Y = 1) := \frac{p}{1 - p}$$

So if odds of an event is 4 ("4:1"), what is p?

Choice of λ?

We just need $\lambda : \mathbb{R} \to [0, 1]$. There are infinite λ's to choose from... but I've only seen three used:

1. Logistic link: $\lambda(w) = \frac{e^w}{1+e^w}$ (most common)
2. Inverse normal (probit) link: $\lambda(w) = \Phi^{-1}(w)$ where Φ is the normal CDF function (somewhat common)
3. Complementary Log-log (cloglog) link: $\lambda(w) = \ln(-\ln(w))$ (rare!)

Herein, we focus on the logistic link. First, short detour. Define $p := \mathbb{P}(Y = 1)$. We can think about probability in another way:

$$odds(Y = 1) := \frac{p}{1 - p}$$

So if odds of an event is 4 ("4:1"), what is p? This means that the probability of the event happening is four times more likely than the complement happening. Or... of 4+1 trials, 4 will be a yes (on average). What is the range of the odds function?

Choice of λ?

We just need $\lambda : \mathbb{R} \to [0, 1]$. There are infinite λ's to choose from... but I've only seen three used:

1. Logistic link: $\lambda(w) = \frac{e^w}{1+e^w}$ (most common)
2. Inverse normal (probit) link: $\lambda(w) = \Phi^{-1}(w)$ where Φ is the normal CDF function (somewhat common)
3. Complementary Log-log (cloglog) link: $\lambda(w) = \ln(-\ln(w))$ (rare!)

Herein, we focus on the logistic link. First, short detour. Define $p := \mathbb{P}(Y = 1)$. We can think about probability in another way:

$$odds(Y = 1) := \frac{p}{1 - p}$$

So if odds of an event is 4 ("4:1"), what is p? This means that the probability of the event happening is four times more likely than the complement happening. Or... of 4+1 trials, 4 will be a yes (on average). What is the range of the odds function? $[0, \infty)$.

Why Logistic Link is Interpretable

Now let's take the log odds (called the logit function):

$$logit(Y = 1) := \ln\left(odds(Y = 1)\right) = \ln\left(\frac{p}{1-p}\right)$$

What is the range of the logit function?

Why Logistic Link is Interpretable

Now let's take the log odds (called the logit function):

$$logit(Y = 1) := \ln\left(odds(Y = 1)\right) = \ln\left(\frac{p}{1-p}\right)$$

What is the range of the logit function? All of $\mathbb{R}$. Hence, we can now set this equal to our

Why Logistic Link is Interpretable

Now let's take the log odds (called the logit function):

$$logit(Y = 1) := \ln\left(odds(Y = 1)\right) = \ln\left(\frac{p}{1-p}\right)$$

What is the range of the logit function? All of $\mathbb{R}$. Hence, we can now set this equal to our $s_{\mathbb{R}}$ function. In the linear modeling context,

$$s_{\mathbb{R}} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \quad = \quad logit(Y = 1) = \ln\left(\frac{p}{1-p}\right)$$

Why Logistic Link is Interpretable

Now let's take the log odds (called the logit function):

$$logit(Y = 1) := \ln\left(odds(Y = 1)\right) = \ln\left(\frac{p}{1-p}\right)$$

What is the range of the logit function? All of $\mathbb{R}$. Hence, we can now set this equal to our $s_{\mathbb{R}}$ function. In the linear modeling context,

$$s_{\mathbb{R}} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \quad = \quad logit(Y = 1) = \ln\left(\frac{p}{1-p}\right)$$

Thus, a change in the linear model becomes a linear change in log-odds.

$$e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p} \quad = \quad \frac{p}{1-p}$$

Why Logistic Link is Interpretable

Now let's take the log odds (called the logit function):

$$logit(Y = 1) := \ln\left(odds(Y = 1)\right) = \ln\left(\frac{p}{1-p}\right)$$

What is the range of the logit function? All of $\mathbb{R}$. Hence, we can now set this equal to our $s_{\mathbb{R}}$ function. In the linear modeling context,

$$s_{\mathbb{R}} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \quad = \quad logit(Y = 1) = \ln\left(\frac{p}{1-p}\right)$$

Thus, a change in the linear model becomes a linear change in log-odds.

$$\begin{aligned} e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p} &= \frac{p}{1-p} \\ (1-p)e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p} &= p \end{aligned}$$

Why Logistic Link is Interpretable

Now let's take the log odds (called the logit function):

$$logit(Y = 1) := \ln\left(odds(Y = 1)\right) = \ln\left(\frac{p}{1-p}\right)$$

What is the range of the logit function? All of $\mathbb{R}$. Hence, we can now set this equal to our $s_{\mathbb{R}}$ function. In the linear modeling context,

$$s_{\mathbb{R}} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \quad = \quad logit(Y = 1) = \ln\left(\frac{p}{1-p}\right)$$

Thus, a change in the linear model becomes a linear change in log-odds.

$$\begin{aligned} e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p} &= \frac{p}{1-p} \\ (1-p)e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p} &= p \\ e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p} &= p + pe^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p} \end{aligned}$$

Why Logistic Link is Interpretable

Now let's take the log odds (called the logit function):

$$logit(Y=1) := \ln\left(odds(Y=1)\right) = \ln\left(\frac{p}{1-p}\right)$$

What is the range of the logit function? All of $\mathbb{R}$. Hence, we can now set this equal to our $s_{\mathbb{R}}$ function. In the linear modeling context,

$$s_{\mathbb{R}} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \quad = \quad logit(Y=1) = \ln\left(\frac{p}{1-p}\right)$$

Thus, a change in the linear model becomes a linear change in log-odds.

$$\begin{aligned} e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p} &= \frac{p}{1-p} \\ (1-p)e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p} &= p \\ e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p} &= p + pe^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p} \\ p &= \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}} = \lambda(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p) \end{aligned}$$

Why Logistic Link is Interpretable

Now let's take the log odds (called the logit function):

$$logit(Y = 1) := \ln(odds(Y = 1)) = \ln\left(\frac{p}{1-p}\right)$$

What is the range of the logit function? All of $\mathbb{R}$. Hence, we can now set this equal to our $s_{\mathbb{R}}$ function. In the linear modeling context,

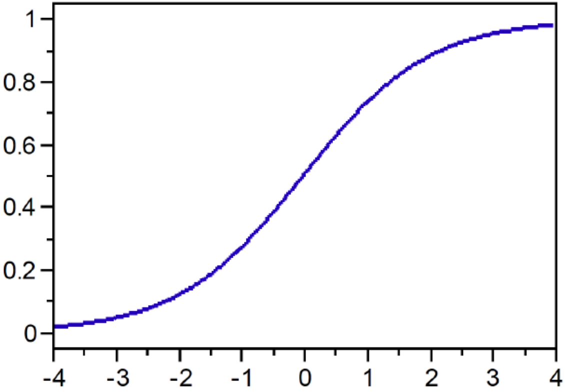
$$s_{\mathbb{R}} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \quad = \quad logit(Y = 1) = \ln\left(\frac{p}{1-p}\right)$$

Thus, a change in the linear model becomes a linear change in log-odds.

$$\begin{aligned} e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p} &= \frac{p}{1-p} \\ (1-p)e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p} &= p \\ e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p} &= p + pe^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p} \\ p &= \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}} = \lambda(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p) \end{aligned}$$

This is (I would say) the most interpretable link function situation we've got.

The Logistic Function (an "S" Shape)



How to Obtain a Model Fit

A model fit would mean we estimate $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$.

How to Obtain a Model Fit

A model fit would mean we estimate $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$. We initially did this estimation for regression (continuous y) by ...

How to Obtain a Model Fit

A model fit would mean we estimate $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$. We initially did this estimation for regression (continuous y) by ... defining a loss function, SSE, and finding the optimal solution via calculus. What do we do now??

Likelihood to the rescue. First the "logistic regression assumptions"

1. Independence among observations. Thus,

How to Obtain a Model Fit

A model fit would mean we estimate $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$. We initially did this estimation for regression (continuous y) by ... defining a loss function, SSE, and finding the optimal solution via calculus. What do we do now??

Likelihood to the rescue. First the "logistic regression assumptions"

1. Independence among observations. Thus,

$$\begin{aligned} & \mathbb{P}\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n\right) \\ = & \prod_{i=1}^{n} \mathbb{P}\left(Y_i = y_i \mid \boldsymbol{X}_1 = \boldsymbol{x}_i\right) \end{aligned}$$

2. Bernoulli Model. Thus,

How to Obtain a Model Fit

A model fit would mean we estimate $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$. We initially did this estimation for regression (continuous y) by ... defining a loss function, SSE, and finding the optimal solution via calculus. What do we do now??

Likelihood to the rescue. First the "logistic regression assumptions"

1. Independence among observations. Thus,

$$\begin{aligned} & \mathbb{P}\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n\right) \\ = & \prod_{i=1}^{n} \mathbb{P}\left(Y_i = y_i \mid \boldsymbol{X}_1 = \boldsymbol{x}_i\right) \end{aligned}$$

2. Bernoulli Model. Thus,

$$= \prod_{i=1}^{n} p^{y_i} (1-p)^{1-y_i}$$

How to Obtain a Model Fit

A model fit would mean we estimate $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$. We initially did this estimation for regression (continuous y) by ... defining a loss function, SSE, and finding the optimal solution via calculus. What do we do now??

Likelihood to the rescue. First the "logistic regression assumptions"

1. Independence among observations. Thus,

$$\begin{aligned} & \mathbb{P}\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n\right) \\ = & \prod_{i=1}^{n} \mathbb{P}\left(Y_i = y_i \mid \boldsymbol{X}_1 = \boldsymbol{x}_i\right) \end{aligned}$$

2. Bernoulli Model. Thus,

$$= \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

3. Linear Logistic conditional expectation. Thus,

Maximum Likelihood Estimates

$$
\begin{aligned}
&= \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}} \right)^{y_i} \left(1 - \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}} \right)^{1-y_i} \\
&= \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)
\end{aligned}
$$

This does not have a simple, closed form solution.

Maximum Likelihood Estimates

$$
\begin{aligned}
&= \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}} \right)^{y_i} \left(1 - \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}} \right)^{1-y_i} \\
&= \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)
\end{aligned}
$$

This does not have a simple, closed form solution. The computer iterates numerically using gradient methods. It usually uses the $\ln(\cdot)$ of above, since it's (1) numerically more stable and (2) the expression is easier to work with. When it "converges" on the values of the parameters that maximize the above, these are shipped to you as $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$. This is called "running a logistic regression". The above looks complicated but it is instant on a modern computer for most real-world datasets.

Prediction with Logistic Regression

Recall that predictions in linear regression were easy:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*$$

Prediction with Logistic Regression

Recall that predictions in linear regression were easy:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*$$

How do we use a logistic regression model to predict with new data $\boldsymbol{x}^*$?

Prediction with Logistic Regression

Recall that predictions in linear regression were easy:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*$$

How do we use a logistic regression model to predict with new data $\boldsymbol{x}^*$?

$$\hat{p} = \hat{p}(\boldsymbol{x}^*) =$$

Prediction with Logistic Regression

Recall that predictions in linear regression were easy:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*$$

How do we use a logistic regression model to predict with new data $\boldsymbol{x}^*$?

$$\hat{p} = \hat{p}(\boldsymbol{x}^*) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p}}$$

Note the predictions are for the conditional expectation function, the probability itself, the **estimated expected probability**. However, you may actually wish to predict the response, the 1 or the 0. What to do?

Prediction with Logistic Regression

Recall that predictions in linear regression were easy:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*$$

How do we use a logistic regression model to predict with new data $\boldsymbol{x}^*$?

$$\hat{p} = \hat{p}(\boldsymbol{x}^*) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p}}$$

Note the predictions are for the conditional expectation function, the probability itself, the **estimated expected probability**. However, you may actually wish to predict the response, the 1 or the 0. What to do?

You can create a **classification rule** which allows you to make a decision about the response based on the probability. What is the most intuitive classification rule?

Prediction with Logistic Regression

Recall that predictions in linear regression were easy:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*$$

How do we use a logistic regression model to predict with new data $\boldsymbol{x}^*$?

$$\hat{p} = \hat{p}(\boldsymbol{x}^*) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p}}$$

Note the predictions are for the conditional expectation function, the probability itself, the **estimated expected probability**. However, you may actually wish to predict the response, the 1 or the 0. What to do?

You can create a **classification rule** which allows you to make a decision about the response based on the probability. What is the most intuitive classification rule?

$$\hat{y} = \mathbb{1}_{\hat{p} \geq 0.5} :=$$

Prediction with Logistic Regression

Recall that predictions in linear regression were easy:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*$$

How do we use a logistic regression model to predict with new data $\boldsymbol{x}^*$?

$$\hat{p} = \hat{p}(\boldsymbol{x}^*) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p}}$$

Note the predictions are for the conditional expectation function, the probability itself, the **estimated expected probability**. However, you may actually wish to predict the response, the 1 or the 0. What to do?

You can create a **classification rule** which allows you to make a decision about the response based on the probability. What is the most intuitive classification rule?

$$\hat{y} = \mathbb{1}_{\hat{p} \geq 0.5} := \begin{cases} 1 & \text{if} \quad \hat{p} \geq 0.5 \\ 0 & \text{if} \quad \hat{p} < 0.5 \end{cases}$$

AKA the "most likely criterion".

Prediction with Logistic Regression

Recall that predictions in linear regression were easy:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*$$

How do we use a logistic regression model to predict with new data $\boldsymbol{x}^*$?

$$\hat{p} = \hat{p}(\boldsymbol{x}^*) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p}}$$

Note the predictions are for the conditional expectation function, the probability itself, the **estimated expected probability**. However, you may actually wish to predict the response, the 1 or the 0. What to do?

You can create a **classification rule** which allows you to make a decision about the response based on the probability. What is the most intuitive classification rule?

$$\hat{y} = \mathbb{1}_{\hat{p} \geq 0.5} := \begin{cases} 1 & \text{if} \quad \hat{p} \geq 0.5 \\ 0 & \text{if} \quad \hat{p} < 0.5 \end{cases}$$

AKA the "most likely criterion". We will return to prediction and evaluation of predictive performance later but first... inference.

Global Test in Logistic Regression

Recall in OLS regression, gaussian (normal) theory directly gave us t-tests and F-tests. Under the logistic regression assumptions, **we have no such analogous theory**! However, we can make use of the ...

Global Test in Logistic Regression

Recall in OLS regression, gaussian (normal) theory directly gave us t-tests and F-tests. Under the logistic regression assumptions, **we have no such analogous theory**! However, we can make use of the ... likelihood ratio test.

Global Test in Logistic Regression

Recall in OLS regression, gaussian (normal) theory directly gave us t-tests and F-tests. Under the logistic regression assumptions, **we have no such analogous theory**! However, we can make use of the ... likelihood ratio test. Recall:

$$LR := \max_{\theta \in \Theta} \mathcal{L}(\theta; x) \,/\, \max_{\theta \in \Theta_R} \mathcal{L}(\theta; x)$$

Let's now do a "whole model" / "global" / "omnibus" test:

Global Test in Logistic Regression

Recall in OLS regression, gaussian (normal) theory directly gave us t-tests and F-tests. Under the logistic regression assumptions, **we have no such analogous theory**! However, we can make use of the ... likelihood ratio test. Recall:

$$LR := \max_{\theta \in \Theta} \mathcal{L}(\theta; x) \,/\, \max_{\theta \in \Theta_R} \mathcal{L}(\theta; x)$$

Let's now do a "whole model" / "global" / "omnibus" test:

$$H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_p = 0, \quad H_a : \text{at least one is non-zero}$$

So Θ would be the space of all $\beta_0, \beta_1, \ldots, \beta_p$ and Θ_R will restrict the space to only β_0 with zeroes for all other "slope" parameters.

$$LR = \frac{\max_{\beta_0, \beta_1, \ldots, \beta_p} \mathcal{L}(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)}{\max_{\beta_0} \mathcal{L}(\beta_0, \beta_1 = 0, \ldots, \beta_p = 0; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)}$$

Global Test in Logistic Regression

Recall in OLS regression, gaussian (normal) theory directly gave us t-tests and F-tests. Under the logistic regression assumptions, **we have no such analogous theory**! However, we can make use of the ... likelihood ratio test. Recall:

$$LR := \max_{\theta \in \Theta} \mathcal{L}(\theta; x) \,/\, \max_{\theta \in \Theta_R} \mathcal{L}(\theta; x)$$

Let's now do a "whole model" / "global" / "omnibus" test:

$$H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_p = 0, \quad H_a : \text{at least one is non-zero}$$

So Θ would be the space of all $\beta_0, \beta_1, \ldots, \beta_p$ and Θ_R will restrict the space to only β_0 with zeroes for all other "slope" parameters.

$$LR = \frac{\max\limits_{\beta_0, \beta_1, \ldots, \beta_p} \mathcal{L}(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)}{\max\limits_{\beta_0} \mathcal{L}(\beta_0, \beta_1 = 0, \ldots, \beta_p = 0; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)}$$

So on top the computer iterates to find $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$, plugs it in and computes the likelihood and on the bottom the computer independently iterates to find $\left\{\hat{\beta}_0\right\}$, plugs it in and computes the likelihood, then together, the LR.

Partial Tests in Logistic Regression

We then look at $Q = 2\ln(LR)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped p parameters / degrees of freedom, we look at the critical $\chi^2_{p,\alpha}$ value.

Partial Tests in Logistic Regression

We then look at $Q = 2\ln(LR)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped p parameters / degrees of freedom, we look at the critical $\chi^2_{p,\alpha}$ value.

Let's say we want to test something like:

$$H_0 : \beta_1 = 0 \ \& \ \beta_2 = 0, \quad H_a : \text{at least one is non-zero}$$

We can again use the likelihood ratio test:

$$LR = \frac{\max\limits_{\beta_0,\beta_1,\ldots,\beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}{\max\limits_{\beta_0,\beta_3,\ldots,\beta_p} \mathcal{L}\left(\beta_0, \beta_1 = 0, \beta_2 = 0, \beta_3, \ldots, \beta_p = 0; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}$$

Partial Tests in Logistic Regression

We then look at $Q = 2\ln(LR)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped p parameters / degrees of freedom, we look at the critical $\chi^2_{p,\alpha}$ value.

Let's say we want to test something like:

$$H_0 : \beta_1 = 0 \;\&\; \beta_2 = 0, \quad H_a : \text{at least one is non-zero}$$

We can again use the likelihood ratio test:

$$LR = \frac{\max\limits_{\beta_0,\beta_1,\ldots,\beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}{\max\limits_{\beta_0,\beta_3,\ldots,\beta_p} \mathcal{L}\left(\beta_0, \beta_1 = 0, \beta_2 = 0, \beta_3, \ldots, \beta_p = 0; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}$$

We then look at $Q = 2\ln(LR)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped 2 parameters / degrees of freedom, we look at the critical $\chi^2_{2,\alpha}$ value.

Individual Tests in Logistic Regression

Let's say we want to test an individual slope coefficient:

$$H_0 : \beta_j = 0, \quad H_a : \beta_j \neq 0$$

(a la the "partial-F test"). We can again use the likelihood ratio test:

$$LR = \frac{\max_{\beta_0, \beta_1, \ldots, \beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}{\max_{\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots \beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_j = 0, \beta_{j+1}, \ldots \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}$$

Individual Tests in Logistic Regression

Let's say we want to test an individual slope coefficient:

$$H_0 : \beta_j = 0, \quad H_a : \beta_j \neq 0$$

(a la the "partial-F test"). We can again use the likelihood ratio test:

$$LR = \frac{\max\limits_{\beta_0,\beta_1,\ldots,\beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}{\max\limits_{\beta_0,\beta_1,\ldots,\beta_{j-1},\beta_{j+1},\ldots\beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_j = 0, \beta_{j+1}, \ldots \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}$$

We then look at $Q = 2\ln(LR)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped 1 parameter / degrees of freedom; thus we look at the critical $\chi^2_{1,\alpha}$ value.

Individual Tests in Logistic Regression

Let's say we want to test an individual slope coefficient:

$$H_0 : \beta_j = 0, \quad H_a : \beta_j \neq 0$$

(a la the "partial-F test"). We can again use the likelihood ratio test:

$$LR = \frac{\max\limits_{\beta_0, \beta_1, \ldots, \beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}{\max\limits_{\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots \beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_j = 0, \beta_{j+1}, \ldots \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}$$

We then look at $Q = 2\ln\left(LR\right)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped 1 parameter / degrees of freedom; thus we look at the critical $\chi^2_{1,\alpha}$ value.

There is something special about a χ^2 r.v. with one degree of freedom.

Individual Tests in Logistic Regression

Let's say we want to test an individual slope coefficient:

$$H_0 : \beta_j = 0, \quad H_a : \beta_j \neq 0$$

(a la the "partial-F test"). We can again use the likelihood ratio test:

$$LR = \frac{\max_{\beta_0,\beta_1,\ldots,\beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}{\max_{\beta_0,\beta_1,\ldots,\beta_{j-1},\beta_{j+1},\ldots\beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_j = 0, \beta_{j+1}, \ldots \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}$$

We then look at $Q = 2\ln\left(LR\right)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped 1 parameter / degrees of freedom; thus we look at the critical $\chi^2_{1,\alpha}$ value.

There is something special about a χ^2 r.v. with one degree of freedom. Note this cool fact from probability theory: $Q \sim \chi^2_1 \Rightarrow$

Individual Tests in Logistic Regression

Let's say we want to test an individual slope coefficient:

$$H_0 : \beta_j = 0, \quad H_a : \beta_j \neq 0$$

(a la the "partial-F test"). We can again use the likelihood ratio test:

$$LR = \frac{\max\limits_{\beta_0,\beta_1,\ldots,\beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}{\max\limits_{\beta_0,\beta_1,\ldots,\beta_{j-1},\beta_{j+1},\ldots\beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_j = 0, \beta_{j+1}, \ldots \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}$$

We then look at $Q = 2\ln\left(LR\right)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped 1 parameter / degrees of freedom; thus we look at the critical $\chi^2_{1,\alpha}$ value.

There is something special about a χ^2 r.v. with one degree of freedom. Note this cool fact from probability theory: $Q \sim \chi^2_1 \Rightarrow \sqrt{Q} \sim \mathcal{N}\left(0, 1\right)$ i.e. a "z-score". This is how JMP produces standard errors for logistic regression coefficients.

Telecom Company Churn Example

In marketing lingo, “churn” refers to a customer canceling their service. Studies suggest that it costs 5-10x more to acquire a new customer than to retain an old customer. Thus, predicting churn is of major interest!

Telecom Company Churn Example

In marketing lingo, "churn" refers to a customer canceling their service. Studies suggest that it costs 5-10x more to acquire a new customer than to retain an old customer. Thus, predicting churn is of major interest!

Here's a dataset from a telecom company (likely it's churn on Verizon / AT&T / T-Mobile /Sprint's cell-phone plan). We have 7,043 customers with 20 features. This is likely a nearly-mindless dump!! Churn is defined to be a complete cancellation of services in the next month period. Since we are predicting churn, define $y = 1$ to be churn, so the $\hat{p}$'s are estimates of probability of churning (this just makes everything easier to understand).

Telecom Company Churn Example

In marketing lingo, "churn" refers to a customer canceling their service. Studies suggest that it costs 5-10x more to acquire a new customer than to retain an old customer. Thus, predicting churn is of major interest!

Here's a dataset from a telecom company (likely it's churn on Verizon / AT&T / T-Mobile /Sprint's cell-phone plan). We have 7,043 customers with 20 features. This is likely a nearly-mindless dump!! Churn is defined to be a complete cancellation of services in the next month period. Since we are predicting churn, define $y = 1$ to be churn, so the $\hat{p}$'s are estimates of probability of churning (this just makes everything easier to understand).

We begin just trying to model y: churn vs. x: tenure (the number of months customer is currently subscribed for). What do you think the relationship will be i.e. what is the sign of $\partial/\partial x[f(x)]$ generally speaking?

Results of Simple Logistic Regression

[JMP demo]

Results of Simple Logistic Regression

[JMP demo] Which likelihood numbers are "best"? Highest likelihoods are good which means highest log likelihoods which means *smallest* negative log likelihoods.... yup it's confusing.

Results of Simple Logistic Regression

[JMP demo] Which likelihood numbers are “best”? Highest likelihoods are good which means highest log likelihoods which means *smallest* negative log likelihoods.... yup it’s confusing. Here’s what JMP did:

$$\begin{aligned} Q &= 2\ln(LR) = 2\left(\ell\left(\hat{\theta};x\right) - \ell\left(\hat{\theta}_R;x\right)\right) \\ &= 2\left(-(3595.9341) - -(4075.0729)\right) \\ &= 2\left(479.1389\right) = 958.2778 \end{aligned}$$

and $\chi^2_{1,5\%} = \texttt{qchisq(.95, 1)} = 3.84$. So this passes the test (comfortably). We reject H_0 and conclude that the model is linearly useful.

Results of Simple Logistic Regression

[JMP demo] Which likelihood numbers are "best"? Highest likelihoods are good which means highest log likelihoods which means *smallest* negative log likelihoods.... yup it's confusing. Here's what JMP did:

$$
\begin{aligned}
Q &= 2\ln(LR) = 2\left(\ell\left(\hat{\theta}; x\right) - \ell\left(\hat{\theta}_R; x\right)\right) \\
&= 2\left(-(3595.9341) - -(4075.0729)\right) \\
&= 2\left(479.1389\right) = 958.2778
\end{aligned}
$$

and $\chi^2_{1,5\%}$ = `qchisq(.95, 1)` = 3.84. So this passes the test (comfortably). We reject H_0 and conclude that the model is linearly useful. ALSO: equivalent to a test of one variable. We reject H_0 and conclude tenure has a linear effect on the log-odds of churn.

Results of Simple Logistic Regression

[JMP demo] Which likelihood numbers are "best"? Highest likelihoods are good which means highest log likelihoods which means *smallest* negative log likelihoods.... yup it's confusing. Here's what JMP did:

$$\begin{aligned} Q &= 2\ln(LR) = 2\left(\ell\left(\hat{\theta};x\right) - \ell\left(\hat{\theta}_R;x\right)\right) \\ &= 2\left(-(3595.9341) - -(4075.0729)\right) \\ &= 2\left(479.1389\right) = 958.2778 \end{aligned}$$

and $\chi^2_{1,5\%}$ = `qchisq(.95, 1)` = 3.84. So this passes the test (comfortably). We reject H_0 and conclude that the model is linearly useful. ALSO: equivalent to a test of one variable. We reject H_0 and conclude tenure has a linear effect on the log-odds of churn.

Also note that $LR = 1.22 \times 10^{208}$ but $e^{-3595.9341} = 0$ i.e. it's less than the smallest number a computer can represent (without special handling). Numbers are cool things... and logs are pretty powerful.

Basic Predictions I

Predict estimated expected probability of churn for someone who has 1 month of tenure

Basic Predictions I

Predict estimated expected probability of churn for someone who has 1 month of tenure

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1}} = \frac{e^{0.0273+(-0.0288)(1)}}{1 + e^{0.0273+(-0.0288)(1)}} = \frac{0.9985}{1.9985} = 0.500$$

How about 2 months of tenure?

Basic Predictions I

Predict estimated expected probability of churn for someone who has 1 month of tenure

$$\hat{p} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1}} = \frac{e^{0.0273+(-0.0288)(1)}}{1+e^{0.0273+(-0.0288)(1)}} = \frac{0.9985}{1.9985} = 0.500$$

How about 2 months of tenure?

$$\hat{p} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1}} = \frac{e^{0.0273+(-0.0288)(2)}}{1+e^{0.0273+(-0.0288)(2)}} = \frac{0.9702}{1.9702} = 0.492$$

i.e. a difference in about 0.8% as measured on a probability scale.

Basic Predictions II

Predict estimated expected probability of churn for someone who has 100 month of tenure

Basic Predictions II

Predict estimated expected probability of churn for someone who has 100 month of tenure

$$\hat{p} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1}} = \frac{e^{0.0273+(-0.0288)(100)}}{1+e^{0.0273+(-0.0288)(100)}} = 0.055$$

How about 101 months of tenure?

Basic Predictions II

Predict estimated expected probability of churn for someone who has 100 month of tenure

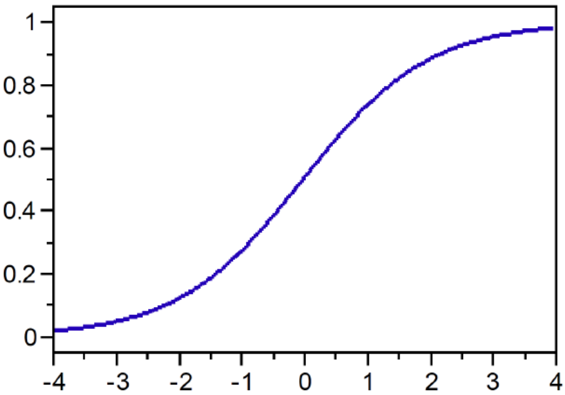
$$\hat{p} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1}} = \frac{e^{0.0273+(-0.0288)(100)}}{1+e^{0.0273+(-0.0288)(100)}} = 0.055$$

How about 101 months of tenure?

$$\hat{p} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1}} = \frac{e^{0.0273+(-0.0288)(101)}}{1+e^{0.0273+(-0.0288)(101)}} = 0.053$$

i.e. a difference in about 0.2% as measured on a probability scale (i.e. a 4x difference from before). But isn't the model supposedly to be linear??

The Logistic Function



A move of one unit in x when $x \approx 0$ is a much bigger move than one unit in x when $x \approx 3$

Parameter Standard Error

To add to the confusion... JMP prefers to calculate parameter estimates and standard error via the Wald test, which is similar to the likelihood ratio test. Thus, $761.00 \neq 958.28$ but, remarkably, they are about the same conceptually – both large and significant.

Parameter Standard Error

To add to the confusion... JMP prefers to calculate parameter estimates and standard error via the Wald test, which is similar to the likelihood ratio test. Thus, $761.00 \neq 958.28$ but, remarkably, they are about the same conceptually – both large and significant. The standard errors,

$$\underbrace{Q = z^2}_{\substack{\text{fact from}\\\text{probability}\\\text{theory}}} = \left(\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}\right)^2 \Rightarrow s_{\hat{\beta}_1} = \frac{\left|\hat{\beta}_1\right|}{\sqrt{Q}}$$

Parameter Standard Error

To add to the confusion... JMP prefers to calculate parameter estimates and standard error via the Wald test, which is similar to the likelihood ratio test. Thus, $761.00 \neq 958.28$ but, remarkably, they are about the same conceptually – both large and significant. The standard errors,

$$\underbrace{Q = z^2}_{\text{fact from probability theory}} = \left(\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}\right)^2 \Rightarrow s_{\hat{\beta}_1} = \frac{\left|\hat{\beta}_1\right|}{\sqrt{Q}}$$

are expectedly about the same

$$s_{\hat{\beta}_1} = \frac{|-0.0387682|}{\sqrt{761.00}} = 0.0014 \quad \text{(via the Wald test)}$$

$$s_{\hat{\beta}_1} = \frac{|-0.0387682|}{\sqrt{958.28}} = 0.0013 \quad \text{(via the LR test)}$$

Multivariate Logistic Regr. Interp. I

Now let's use all variables. Questions:

- Which variable(s) should we leave out?

Multivariate Logistic Regr. Interp. I

Now let's use all variables. Questions:

- Which variable(s) should we leave out? `customerID` — no causal mechanism.

Multivariate Logistic Regr. Interp. I

Now let's use all variables. Questions:

- Which variable(s) should we leave out? `customerID` — no causal mechanism.
- Why are we getting biased estimates and zeroes?

Multivariate Logistic Regr. Interp. I

Now let's use all variables. Questions:

- Which variable(s) should we leave out? `customerID` — no causal mechanism.
- Why are we getting biased estimates and zeroes? Perfect multicollinearity.

Multivariate Logistic Regr. Interp. I

Now let's use all variables. Questions:

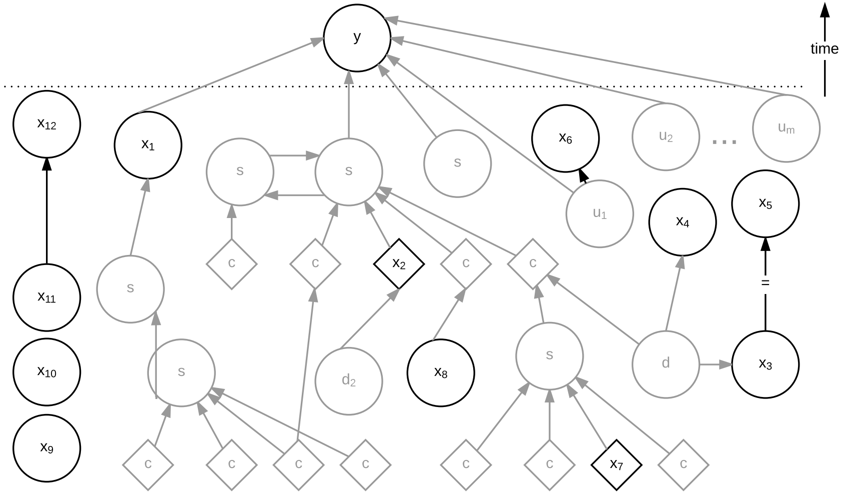
- Which variable(s) should we leave out? `customerID` — no causal mechanism.
- Why are we getting biased estimates and zeroes? Perfect multicollinearity. Solution?

Multivariate Logistic Regr. Interp. I

Now let's use all variables. Questions:

- Which variable(s) should we leave out? `customerID` — no causal mechanism.
- Why are we getting biased estimates and zeroes? Perfect multicollinearity. Solution? Kill columns until you don't get the error anymore. Turns out 6 have to be removed.

Realistic Predictors Illustration (updated)



Multivariate Logistic Regr. Interp. II

Now let's use all variables. Questions:

- Which variable(s) should we leave out? `customerID` — no causal mechanism.
- Why are we getting biased estimates and zeroes? Perfect multicollinearity.
- Are all these variables "significant"?

Multivariate Logistic Regr. Interp. II

Now let's use all variables. Questions:

- Which variable(s) should we leave out? `customerID` — no causal mechanism.
- Why are we getting biased estimates and zeroes? Perfect multicollinearity.
- Are all these variables "significant"? No. And further, after a Bonferroni correction, we "lose" two. Are they for sure insignificant?

Multivariate Logistic Regr. Interp. II

Now let's use all variables. Questions:

- Which variable(s) should we leave out? `customerID` — no causal mechanism.
- Why are we getting biased estimates and zeroes? Perfect multicollinearity.
- Are all these variables "significant"? No. And further, after a Bonferroni correction, we "lose" two. Are they for sure insignificant? No, maybe we didn't have enough **power** to detect them... [review power concept]
- Do the coefficients sign / size make sense intuitively?

Multivariate Logistic Regr. Interp. II

Now let's use all variables. Questions:

- Which variable(s) should we leave out? `customerID` — no causal mechanism.
- Why are we getting biased estimates and zeroes? Perfect multicollinearity.
- Are all these variables "significant"? No. And further, after a Bonferroni correction, we "lose" two. Are they for sure insignificant? No, maybe we didn't have enough **power** to detect them... [review power concept]
- Do the coefficients sign / size make sense intuitively? Yes?
- Which of the "driving" variables are *not* able to be manipulated?

Multivariate Logistic Regr. Interp. II

Now let's use all variables. Questions:

- Which variable(s) should we leave out? `customerID` — no causal mechanism.
- Why are we getting biased estimates and zeroes? Perfect multicollinearity.
- Are all these variables "significant"? No. And further, after a Bonferroni correction, we "lose" two. Are they for sure insignificant? No, maybe we didn't have enough **power** to detect them... [review power concept]
- Do the coefficients sign / size make sense intuitively? Yes?
- Which of the "driving" variables are *not* able to be manipulated? Senior citizen, total charges (?)

Multivariate Logistic Regr. Interp. II

Now let's use all variables. Questions:

- Which variable(s) should we leave out? `customerID` — no causal mechanism.
- Why are we getting biased estimates and zeroes? Perfect multicollinearity.
- Are all these variables "significant"? No. And further, after a Bonferroni correction, we "lose" two. Are they for sure insignificant? No, maybe we didn't have enough **power** to detect them... [review power concept]
- Do the coefficients sign / size make sense intuitively? Yes?
- Which of the "driving" variables are *not* able to be manipulated? Senior citizen, total charges (?)
- Should you (the manager) try to incentivize electronic checks?

Multivariate Logistic Regr. Interp. II

Now let's use all variables. Questions:

- Which variable(s) should we leave out? `customerID` — no causal mechanism.
- Why are we getting biased estimates and zeroes? Perfect multicollinearity.
- Are all these variables "significant"? No. And further, after a Bonferroni correction, we "lose" two. Are they for sure insignificant? No, maybe we didn't have enough **power** to detect them... [review power concept]
- Do the coefficients sign / size make sense intuitively? Yes?
- Which of the "driving" variables are *not* able to be manipulated? Senior citizen, total charges (?)
- Should you (the manager) try to incentivize electronic checks? Paperless billing?

Multivariate Logistic Regr. Interp. II

Now let's use all variables. Questions:

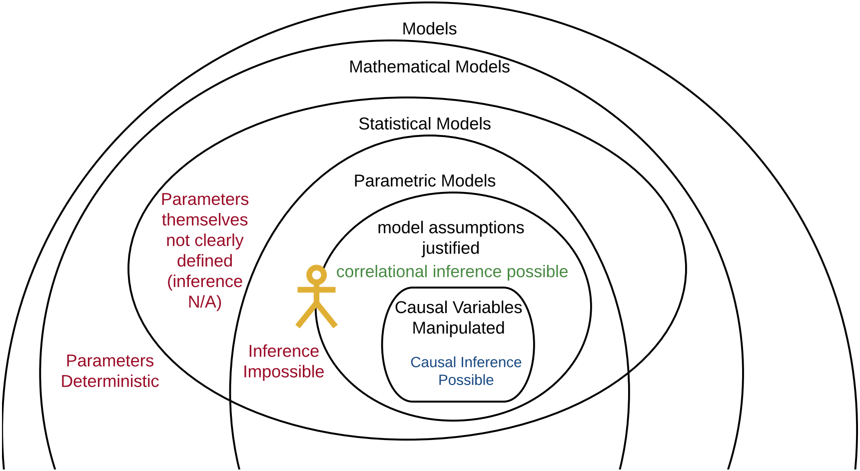
- Which variable(s) should we leave out? `customerID` — no causal mechanism.
- Why are we getting biased estimates and zeroes? Perfect multicollinearity.
- Are all these variables "significant"? No. And further, after a Bonferroni correction, we "lose" two. Are they for sure insignificant? No, maybe we didn't have enough **power** to detect them... [review power concept]
- Do the coefficients sign / size make sense intuitively? Yes?
- Which of the "driving" variables are *not* able to be manipulated? Senior citizen, total charges (?)
- Should you (the manager) try to incentivize electronic checks? Paperless billing? Dropping multiple phone lines?

Multivariate Logistic Regr. Interp. II

Now let's use all variables. Questions:

- Which variable(s) should we leave out? `customerID` — no causal mechanism.
- Why are we getting biased estimates and zeroes? Perfect multicollinearity.
- Are all these variables "significant"? No. And further, after a Bonferroni correction, we "lose" two. Are they for sure insignificant? No, maybe we didn't have enough **power** to detect them... [review power concept]
- Do the coefficients sign / size make sense intuitively? Yes?
- Which of the "driving" variables are *not* able to be manipulated? Senior citizen, total charges (?)
- Should you (the manager) try to incentivize electronic checks? Paperless billing? Dropping multiple phone lines? You can try these things if you have nothing to lose, but remember, they are not guaranteed to be causal! And they may backfire!! (Can you think of

Remember Where you At!!



Evaluating Logistic Regression Models

Many, many measures reported by JMP that we generally don't use:

Evaluating Logistic Regression Models

Many, many measures reported by JMP that we generally don't use:

- $R^2(U)$
- Generalized R^2
- Mean Negative Log probability
- "RMSE"
- Mean Absolute Deviation

Evaluating Logistic Regression Models

Many, many measures reported by JMP that we generally don't use:

- $R^2(U)$
- Generalized R^2
- Mean Negative Log probability
- "RMSE"
- Mean Absolute Deviation

And ones that we do use:

- AICc / BIC (for something a little bit different... we will come back to this in a couple of lectures)

Evaluating Logistic Regression Models

Many, many measures reported by JMP that we generally don't use:

- $R^2(U)$
- Generalized R^2
- Mean Negative Log probability
- "RMSE"
- Mean Absolute Deviation

And ones that we do use:

- AICc / BIC (for something a little bit different... we will come back to this in a couple of lectures)
- Misclassification Rate

Evaluating Logistic Regression Models

Many, many measures reported by JMP that we generally don't use:

- $R^2(U)$
- Generalized R^2
- Mean Negative Log probability
- "RMSE"
- Mean Absolute Deviation

And ones that we do use:

- AICc / BIC (for something a little bit different... we will come back to this in a couple of lectures)
- Misclassification Rate

We now cover evaluating classification models in general (not only in the context of logistic regression models specifically).

Probability Predictions $\Rightarrow$ Level Predictions

Recall that ... you can create a classification rule which allows you to make a decision about the response based on the probability. The most intuitive classification rule is:

Probability Predictions $\Rightarrow$ Level Predictions

Recall that ... you can create a classification rule which allows you to make a decision about the response based on the probability. The most intuitive classification rule is:

$$\hat{y} = \mathbb{1}_{\hat{p} \geq 0.5} :=$$

Probability Predictions $\Rightarrow$ Level Predictions

Recall that ... you can create a classification rule which allows you to make a decision about the response based on the probability. The most intuitive classification rule is:

$$\hat{y} = \mathbb{1}_{\hat{p} \geq 0.5} := \begin{cases} 1 & \text{if} \quad \hat{p} \geq 0.5 \\ 0 & \text{if} \quad \hat{p} < 0.5 \end{cases}$$

In regression, you examined functions of the residuals $e_i := y_i - \hat{y}_i$ to assess model fit. What is an analagous residual here?

Probability Predictions $\Rightarrow$ Level Predictions

Recall that ... you can create a classification rule which allows you to make a decision about the response based on the probability. The most intuitive classification rule is:

$$\hat{y} = \mathbb{1}_{\hat{p} \geq 0.5} := \begin{cases} 1 & \text{if} \quad \hat{p} \geq 0.5 \\ 0 & \text{if} \quad \hat{p} < 0.5 \end{cases}$$

In regression, you examined functions of the residuals $e_i := y_i - \hat{y}_i$ to assess model fit. What is an analagous residual here? There are four residuals, two representing errors. The best way to see them is to create the **confusion matrix**:

		$\hat{y}$ (decision) 1	$\hat{y}$ (decision) 0
y (truth)	1	true positive (TP)	false negative (FN)
	0	false positive (FP)	true negative (TN)

Why do "correlations rock" here??

Probability Predictions $\Rightarrow$ Level Predictions

Recall that ... you can create a classification rule which allows you to make a decision about the response based on the probability. The most intuitive classification rule is:

$$\hat{y} = \mathbb{1}_{\hat{p} \geq 0.5} := \begin{cases} 1 & \text{if} \quad \hat{p} \geq 0.5 \\ 0 & \text{if} \quad \hat{p} < 0.5 \end{cases}$$

In regression, you examined functions of the residuals $e_i := y_i - \hat{y}_i$ to assess model fit. What is an analagous residual here? There are four residuals, two representing errors. The best way to see them is to create the **confusion matrix**:

		$\hat{y}$ (decision) 1	$\hat{y}$ (decision) 0
y (truth)	1	true positive (TP)	false negative (FN)
	0	false positive (FP)	true negative (TN)

Why do "correlations rock" here?? We are purely evaluating predictive performance... no inferential claims!

Confusion Matrix for Churn Model

JMP gives us the matrix [JMP],

Confusion Matrix for Churn Model

JMP gives us the matrix [JMP], but they don't annotate it well. Here are some numbers I like to see:

		$\hat{y}$ 1	$\hat{y}$ 0	Totals	Model Errors
y	1	$TP = 1012$	$FN = 857$	$P = 1869$	$FNR = 45.9\%$
	0	$FP = 531$	$TN = 4632$	$N = 5163$	$FPR = 10.2\%$
	Totals	$\hat{P} = 1543$	$\hat{N} = 5489$	$n = 7032$	
	Use errors	$FDR = 34.3\%$	$FOR = 15.6\%$		$\boxed{ME = 19.7\%}$

There are other metrics commonly reported

- Sensitivity = Recall = $\frac{TP}{TP+FN} = \frac{TP}{P}$, the proportion of positives successfully recovered (large value = good model), 54.9% above
- Specificity = $\frac{TN}{TN+FP} = \frac{TN}{N}$, the proportion of negatives successfully recovered (large value = good model), 89.8% above

Misclassification Error

Already... what is one broad, general metric to evaluate the model?

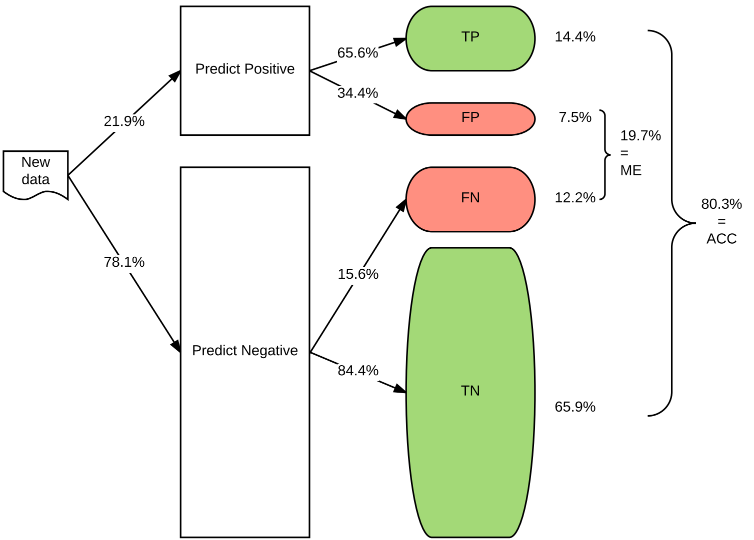
Misclassification Error

Already... what is one broad, general metric to evaluate the model? Misclassification error cost function (or Accuracy):

$$
\begin{aligned}
ME &:= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i \neq \hat{y}_i} \\
ACC &:= 1 - ME = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i = \hat{y}_i}
\end{aligned}
$$

This essentially treats both types of errors (the FN's and the FP's) equally (more on this later).

Production Classifier Flowchart



There's a Ton of Metrics...

From wikipedia...

		Predicted condition			
	Total population	Predicted Condition positive	Predicted Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
True condition	condition positive	**True positive**	**False Negative** (Type II error)	True positive rate (TPR), Sensitivity, Recall, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$
	condition negative	**False Positive** (Type I error)	**True negative**	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$
	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	

Others (from above) commonly used:

- False Discovery Rate (FDR) $= \frac{FP}{TP+FP} = \frac{FP}{\hat{P}}$, the proportion of negatives of those predicted to be positive (small value = good model)

There's a Ton of Metrics...

From wikipedia...

		Predicted condition			
	Total population	Predicted Condition positive	Predicted Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
True condition	condition positive	**True positive**	**False Negative** (Type II error)	True positive rate (TPR), Sensitivity, Recall, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$
	condition negative	**False Positive** (Type I error)	**True negative**	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$
	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

Others (from above) commonly used:

- False Discovery Rate (FDR) $= \frac{FP}{TP+FP} = \frac{FP}{\hat{P}}$, the proportion of negatives of those predicted to be positive (small value = good model)
- False Omission Rate (FOR) $= \frac{FN}{TN+FN} = \frac{FN}{\hat{N}}$, the proportion of postives of those predicted to be negative (small value = good model)

There's a Ton of Metrics...

From wikipedia...

		Predicted condition			
	Total population	Predicted Condition positive	Predicted Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
True condition	condition positive	**True positive**	**False Negative** (Type II error)	True positive rate (TPR), Sensitivity, Recall, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$
	condition negative	**False Positive** (Type I error)	**True negative**	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$
	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	

Others (from above) commonly used:

- False Discovery Rate (FDR) $= \frac{FP}{TP+FP} = \frac{FP}{\hat{P}}$, the proportion of negatives of those predicted to be positive (small value = good model)
- False Omission Rate (FOR) $= \frac{FN}{TN+FN} = \frac{FN}{\hat{N}}$, the proportion of postives of those predicted to be negative (small value = good model)
- Precision = Positive Predictive Value (PPV) = 1 - FDR $= \frac{TP}{TP+FP}$, the proportion of positives of those predicted to be positive (large value = good model)

Generalizing the Classification Rule

Recall the classification rule $\hat{y} = \mathbb{1}_{\hat{p} \geq 0.5}$. Using 0.5 is a principled default but we can use any rule $p_0 \in (0, 1)$:

$$\hat{y} = \mathbb{1}_{\hat{p} \geq p_0} :=$$

Generalizing the Classification Rule

Recall the classification rule $\hat{y} = \mathbb{1}_{\hat{p} \geq 0.5}$. Using 0.5 is a principled default but we can use any rule $p_0 \in (0, 1)$:

$$\hat{y} = \mathbb{1}_{\hat{p} \geq p_0} := \begin{cases} 1 & \text{if} \quad \hat{p} \geq p_0 \\ 0 & \text{if} \quad \hat{p} < p_0 \end{cases}$$

What happens when we change the p_0 threshold? If $p_0 \uparrow \quad \Rightarrow \quad \hat{P}$

Generalizing the Classification Rule

Recall the classification rule $\hat{y} = \mathbb{1}_{\hat{p} \geq 0.5}$. Using 0.5 is a principled default but we can use any rule $p_0 \in (0, 1)$:

$$\hat{y} = \mathbb{1}_{\hat{p} \geq p_0} := \begin{cases} 1 & \text{if} \quad \hat{p} \geq p_0 \\ 0 & \text{if} \quad \hat{p} < p_0 \end{cases}$$

What happens when we change the p_0 threshold? If $p_0 \uparrow \quad \Rightarrow \quad \hat{P} \downarrow$ and $\hat{N}$

Generalizing the Classification Rule

Recall the classification rule $\hat{y} = \mathbb{1}_{\hat{p} \geq 0.5}$. Using 0.5 is a principled default but we can use any rule $p_0 \in (0, 1)$:

$$\hat{y} = \mathbb{1}_{\hat{p} \geq p_0} := \begin{cases} 1 & \text{if} \quad \hat{p} \geq p_0 \\ 0 & \text{if} \quad \hat{p} < p_0 \end{cases}$$

What happens when we change the p_0 threshold? If $p_0 \uparrow \quad \Rightarrow \quad \hat{P} \downarrow$ and $\hat{N} \uparrow$. If $p_0 \downarrow \quad \Rightarrow \quad \hat{P}$

Generalizing the Classification Rule

Recall the classification rule $\hat{y} = \mathbb{1}_{\hat{p} \geq 0.5}$. Using 0.5 is a principled default but we can use any rule $p_0 \in (0, 1)$:

$$\hat{y} = \mathbb{1}_{\hat{p} \geq p_0} := \begin{cases} 1 & \text{if} \quad \hat{p} \geq p_0 \\ 0 & \text{if} \quad \hat{p} < p_0 \end{cases}$$

What happens when we change the p_0 threshold? If $p_0 \uparrow \quad \Rightarrow \quad \hat{P} \downarrow$ and $\hat{N} \uparrow$. If $p_0 \downarrow \quad \Rightarrow \quad \hat{P} \uparrow$ and $\hat{N}$

Generalizing the Classification Rule

Recall the classification rule $\hat{y} = \mathbb{1}_{\hat{p} \geq 0.5}$. Using 0.5 is a principled default but we can use any rule $p_0 \in (0, 1)$:

$$\hat{y} = \mathbb{1}_{\hat{p} \geq p_0} := \begin{cases} 1 & \text{if} \quad \hat{p} \geq p_0 \\ 0 & \text{if} \quad \hat{p} < p_0 \end{cases}$$

What happens when we change the p_0 threshold? If $p_0 \uparrow \quad \Rightarrow \quad \hat{P} \downarrow$ and $\hat{N} \uparrow$. If $p_0 \downarrow \quad \Rightarrow \quad \hat{P} \uparrow$ and $\hat{N} \downarrow$. Changing p_0 changes the column totals and obviously creates a whole new confusion matrix.

So now it's simple, vary p_0 and pick the best model according to your cost / error / loss function (the ME at the moment). Let's just do every p_0!

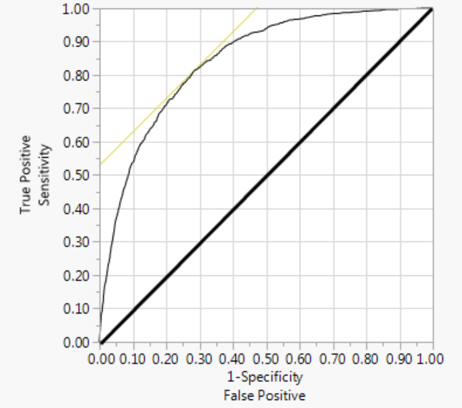
Receiver-Operator Characteristic Table

ROC Table

Prob	1-Specificity	Sensitivity	Sens-(1-Spec)	True Pos	True Neg	False Pos	False Neg
.	0.0000	0.0000	0.0000	0	5163	0	1869
0.8117	0.0000	0.0005	0.0005	1	5163	0	1868
0.8104	0.0000	0.0011	0.0011	2	5163	0	1867
0.8093	0.0000	0.0016	0.0016	3	5163	0	1866
0.8092	0.0000	0.0021	0.0021	4	5163	0	1865
0.8090	0.0000	0.0027	0.0027	5	5163	0	1864
0.8085	0.0000	0.0032	0.0032	6	5163	0	1863
0.8083	0.0000	0.0037	0.0037	7	5163	0	1862
0.8082	0.0000	0.0043	0.0043	8	5163	0	1861
0.8079	0.0000	0.0048	0.0048	9	5163	0	1860
0.8079	0.0000	0.0054	0.0054	10	5163	0	1859
0.8077	0.0000	0.0059	0.0059	11	5163	0	1858
0.8076	0.0002	0.0059	0.0057	11	5162	1	1858
0.8072	0.0002	0.0064	0.0062	12	5162	1	1857
0.8065	0.0002	0.0070	0.0068	13	5162	1	1856
0.8064	0.0002	0.0075	0.0073	14	5162	1	1855
0.8061	0.0002	0.0080	0.0078	15	5162	1	1854

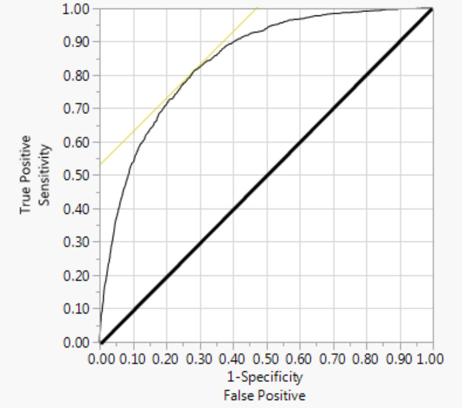
Here, `Prob` is what we denoted p_0. "Best model" is not defined here by highest ACC (lowest ME), it's determined by highest **specificity** + **sensitivity** or equivalently, the highest **sensitivity** - **(1 - specificity)**.JMP indicates that row with a $\star$. This is an arbitrary metric, but is a good default.

Receiver-Operator Characteristic Curve



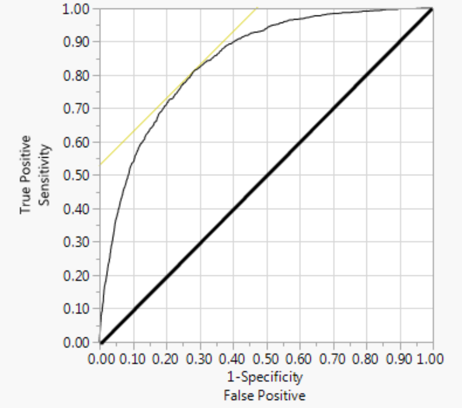
This is graphical illustration of the table. Each dot represents the sensitivity-specificity tradeoff for each p_0.

Receiver-Operator Characteristic Curve



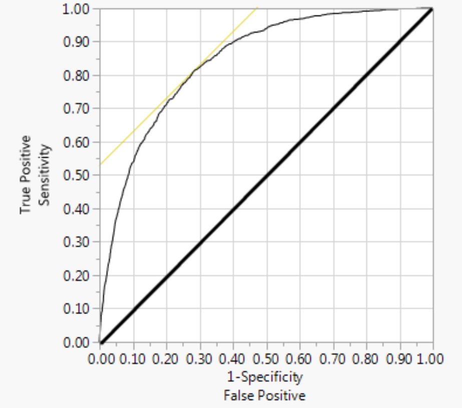
This is graphical illustration of the table. Each dot represents the sensitivity-specificity tradeoff for each p_0. The starred row of maximum sensitivity + specificity is indicated here by a yellow tangent line.

Receiver-Operator Characteristic Curve



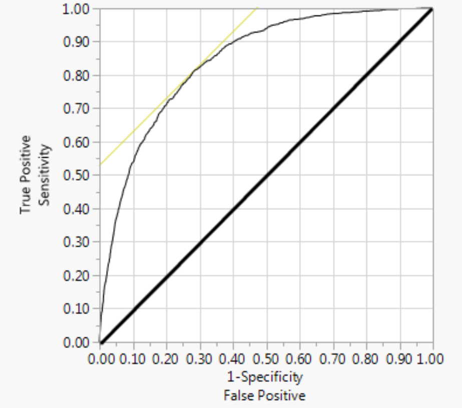
This is graphical illustration of the table. Each dot represents the sensitivity-specificity tradeoff for each p_0. The starred row of maximum sensitivity + specificity is indicated here by a yellow tangent line. I drew the diagonal line to indicate predictive performance that is expected "by chance". Why?

Receiver-Operator Characteristic Curve



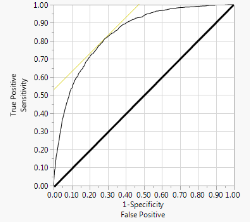
This is graphical illustration of the table. Each dot represents the sensitivity-specificity tradeoff for each p_0. The starred row of maximum sensitivity + specificity is indicated here by a yellow tangent line. I drew the diagonal line to indicate predictive performance that is expected "by chance". Why? Also, are there other ROC variants?

Receiver-Operator Characteristic Curve



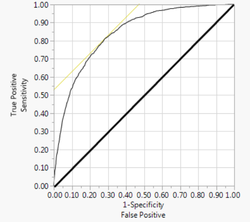
This is graphical illustration of the table. Each dot represents the sensitivity-specificity tradeoff for each p_0. The starred row of maximum sensitivity + specificity is indicated here by a yellow tangent line. I drew the diagonal line to indicate predictive performance that is expected "by chance". Why? Also, are there other ROC variants? Yes (HW?).

Area Under the Curve (AUC) Metric



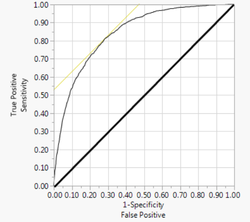
If you built a model by chance the “area under the curve” (or to the right of the curve) on the graph would be ...

Area Under the Curve (AUC) Metric



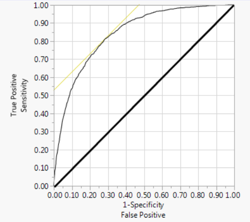
If you built a model by chance the "area under the curve" (or to the right of the curve) on the graph would be ... 0.5 since the graph is a unit square. Under the ROC curve itself (or to its right) is an area ...

Area Under the Curve (AUC) Metric



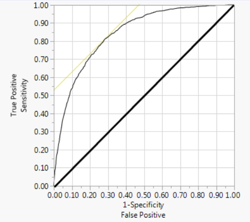
If you built a model by chance the "area under the curve" (or to the right of the curve) on the graph would be ... 0.5 since the graph is a unit square. Under the ROC curve itself (or to its right) is an area ... greater than 0.5. Here, it's 0.844.

Area Under the Curve (AUC) Metric



If you built a model by chance the "area under the curve" (or to the right of the curve) on the graph would be ... 0.5 since the graph is a unit square. Under the ROC curve itself (or to its right) is an area ... greater than 0.5. Here, it's 0.844. This metric is called AUC and is widely used as a metric to assess performance of all possible classifiers in this set of models together, it is a composite metric unlike *ME* or anything derived from an individual confusion table.

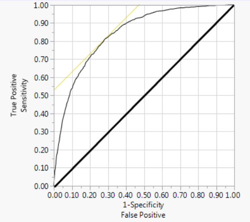
Area Under the Curve (AUC) Metric



If you built a model by chance the "area under the curve" (or to the right of the curve) on the graph would be ... 0.5 since the graph is a unit square. Under the ROC curve itself (or to its right) is an area ... greater than 0.5. Here, it's 0.844. This metric is called AUC and is widely used as a metric to assess performance of all possible classifiers in this set of models together, it is a composite metric unlike *ME* or anything derived from an individual confusion table.

AUC is nice to evaluate overall performance of all possible models... but at the end of the day...

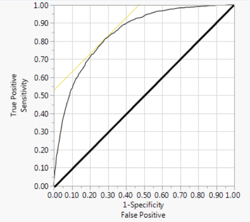
Area Under the Curve (AUC) Metric



If you built a model by chance the "area under the curve" (or to the right of the curve) on the graph would be ... 0.5 since the graph is a unit square. Under the ROC curve itself (or to its right) is an area ... greater than 0.5. Here, it's 0.844. This metric is called AUC and is widely used as a metric to assess performance of all possible classifiers in this set of models together, it is a composite metric unlike *ME* or anything derived from an individual confusion table.

AUC is nice to evaluate overall performance of all possible models... but at the end of the day... you ship **ONE** model!

Area Under the Curve (AUC) Metric



If you built a model by chance the "area under the curve" (or to the right of the curve) on the graph would be ... 0.5 since the graph is a unit square. Under the ROC curve itself (or to its right) is an area ... greater than 0.5. Here, it's 0.844. This metric is called AUC and is widely used as a metric to assess performance of all possible classifiers in this set of models together, it is a composite metric unlike *ME* or anything derived from an individual confusion table.

AUC is nice to evaluate overall performance of all possible models... but at the end of the day... you ship **ONE** model! So we still need a means of evaluating our one model from one confusion table.

Churn Example Where $p_0 = 0.10$

Churn Example Where $p_0 = 0.10$

$p_0 = 0.5$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1012$	$FN = 857$	$P = 1869$	$FNR = 45.9\%$
	0	$FP = 531$	$TN = 4632$	$N = 5163$	$FPR = 10.2\%$
	Totals	$\hat{P} = 1543$	$\hat{N} = 5489$	$n = 7032$	
	Use errors	$FDR = 34.3\%$	$FOR = 15.6\%$		$\boxed{ME = 19.7\%}$

Churn Example Where $p_0 = 0.10$

$p_0 = 0.5$		$\hat{y}$ 1	$\hat{y}$ 0	Totals	Model Errors
y	1	$TP = 1012$	$FN = 857$	$P = 1869$	$FNR = 45.9\%$
	0	$FP = 531$	$TN = 4632$	$N = 5163$	$FPR = 10.2\%$
	Totals	$\hat{P} = 1543$	$\hat{N} = 5489$	$n = 7032$	
	Use errors	$FDR = 34.3\%$	$FOR = 15.6\%$		$\boxed{ME = 19.7\%}$

$p_0 = 0.1$		$\hat{y}$ 1	$\hat{y}$ 0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	FP $= 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Churn Example Where $p_0 = 0.10$

$p_0 = 0.5$		$\hat{y}$ = 1	$\hat{y}$ = 0	Totals	Model Errors
y	1	$TP = 1012$	$FN = 857$	$P = 1869$	$FNR = 45.9\%$
	0	$FP = 531$	$TN = 4632$	$N = 5163$	$FPR = 10.2\%$
	Totals	$\hat{P} = 1543$	$\hat{N} = 5489$	$n = 7032$	
	Use errors	$FDR = 34.3\%$	$FOR = 15.6\%$		$\boxed{ME = 19.7\%}$

$p_0 = 0.1$		$\hat{y}$ = 1	$\hat{y}$ = 0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	FP $= 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Which numbers did not change?

Churn Example Where $p_0 = 0.10$

$p_0 = 0.5$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1012$	$FN = 857$	$P = 1869$	$FNR = 45.9\%$
	0	$FP = 531$	$TN = 4632$	$N = 5163$	$FPR = 10.2\%$
	Totals	$\hat{P} = 1543$	$\hat{N} = 5489$	$n = 7032$	
	Use errors	$FDR = 34.3\%$	$FOR = 15.6\%$		$\boxed{ME = 19.7\%}$

$p_0 = 0.1$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	FP $= 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Which numbers did not change? n, P and N. Why?

Churn Example Where $p_0 = 0.10$

$p_0 = 0.5$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1012$	$FN = 857$	$P = 1869$	$FNR = 45.9\%$
	0	$FP = 531$	$TN = 4632$	$N = 5163$	$FPR = 10.2\%$
	Totals	$\hat{P} = 1543$	$\hat{N} = 5489$	$n = 7032$	
	Use errors	$FDR = 34.3\%$	$FOR = 15.6\%$		$\boxed{ME = 19.7\%}$

$p_0 = 0.1$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	FP $= 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Which numbers did not change? n, P and N. Why? These are fixed according to the dataframe. All other numbers changed! What happend to our first means of evaluation, the Misclassification Error?

Churn Example Where $p_0 = 0.10$

$p_0 = 0.5$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1012$	$FN = 857$	$P = 1869$	$FNR = 45.9\%$
	0	$FP = 531$	$TN = 4632$	$N = 5163$	$FPR = 10.2\%$
	Totals	$\hat{P} = 1543$	$\hat{N} = 5489$	$n = 7032$	
	Use errors	$FDR = 34.3\%$	$FOR = 15.6\%$		$\boxed{ME = 19.7\%}$

$p_0 = 0.1$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	FP $= 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Which numbers did not change? n, P and N. Why? These are fixed according to the dataframe. All other numbers changed! What happend to our first means of evaluation, the Misclassification Error? It increased from 19.7% $\rightarrow$ 39.3%. So isn't this a worse model??

Churn Example Where $p_0 = 0.10$

$p_0 = 0.5$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1012$	$FN = 857$	$P = 1869$	$FNR = 45.9\%$
	0	$FP = 531$	$TN = 4632$	$N = 5163$	$FPR = 10.2\%$
	Totals	$\hat{P} = 1543$	$\hat{N} = 5489$	$n = 7032$	
	Use errors	$FDR = 34.3\%$	$FOR = 15.6\%$		$\boxed{ME = 19.7\%}$

$p_0 = 0.1$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	FP $= 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Which numbers did not change? n, P and N. Why? These are fixed according to the dataframe. All other numbers changed! What happend to our first means of evaluation, the Misclassification Error? It increased from 19.7% $\rightarrow$ 39.3%. So isn't this a worse model??

Not necessarily... It depends on what your goal is!

Asymmetric Costs in a Classifier

These are always two types of errors but the costs are not always the same.

$p_0 = 0.1$		$\hat{y}$: 1	$\hat{y}$: 0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	$FP = 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Imagine we really are the Telecom business manager. It costs 5-10x more to acquire a new customer than to engage a customer who is likely to churn. So you give an incentive package to those who are predicted to churn.

Asymmetric Costs in a Classifier

These are always two types of errors but the costs are not always the same.

	$p_0 = 0.1$	$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	$FP = 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Imagine we really are the Telecom business manager. It costs 5-10x more to acquire a new customer than to engage a customer who is likely to churn. So you give an incentive package to those who are predicted to churn. Which of the two types of errors specifically is *very* costly?

Asymmetric Costs in a Classifier

These are always two types of errors but the costs are not always the same.

$p_0 = 0.1$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	$FP = 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Imagine we really are the Telecom business manager. It costs 5-10x more to acquire a new customer than to engage a customer who is likely to churn. So you give an incentive package to those who are predicted to churn. Which of the two types of errors specifically is *very* costly? The *FN*. Who are they?

Asymmetric Costs in a Classifier

These are always two types of errors but the costs are not always the same.

$p_0 = 0.1$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	$FP = 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Imagine we really are the Telecom business manager. It costs 5-10x more to acquire a new customer than to engage a customer who is likely to churn. So you give an incentive package to those who are predicted to churn. Which of the two types of errors specifically is *very* costly? The *FN*. Who are they? These are those who you said were not going to churn *and they did*! Cost?

Asymmetric Costs in a Classifier

These are always two types of errors but the costs are not always the same.

	$p_0 = 0.1$	$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	$FP = 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Imagine we really are the Telecom business manager. It costs 5-10x more to acquire a new customer than to engage a customer who is likely to churn. So you give an incentive package to those who are predicted to churn. Which of the two types of errors specifically is *very* costly? The *FN*. Who are they? These are those who you said were not going to churn *and they did*! Cost? You need to acquire a new customer! The other type of error is less costly, the *FP*. Who are they?

Asymmetric Costs in a Classifier

These are always two types of errors but the costs are not always the same.

	$p_0 = 0.1$	$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	$FP = 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Imagine we really are the Telecom business manager. It costs 5-10x more to acquire a new customer than to engage a customer who is likely to churn. So you give an incentive package to those who are predicted to churn. Which of the two types of errors specifically is *very* costly? The *FN*. Who are they? These are those who you said were not going to churn *and they did*! Cost? You need to acquire a new customer! The other type of error is less costly, the *FP*. Who are they? These are the people you thought were going to churn and did not. Cost?

Asymmetric Costs in a Classifier

These are always two types of errors but the costs are not always the same.

$p_0 = 0.1$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	$FP = 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Imagine we really are the Telecom business manager. It costs 5-10x more to acquire a new customer than to engage a customer who is likely to churn. So you give an incentive package to those who are predicted to churn. Which of the two types of errors specifically is *very* costly? The *FN*. Who are they? These are those who you said were not going to churn *and they did*! Cost? You need to acquire a new customer! The other type of error is less costly, the *FP*. Who are they? These are the people you thought were going to churn and did not. Cost? Whatever the incentive package is.

Weighted Misclassification Error

We now define two costs: (1)

Weighted Misclassification Error

We now define two costs: (1) the cost of the FP denoted c_{FP} and

Weighted Misclassification Error

We now define two costs: (1) the cost of the FP denoted c_{FP} and (2) the cost of the FN denoted c_{FN}.

Weighted Misclassification Error

We now define two costs: (1) the cost of the *FP* denoted c_{FP} and (2) the cost of the *FN* denoted c_{FN}. We then define the weighted misclassification error evaluation metric:

$$ME_w := \frac{1}{n}\sum_{i=1}^{n} c_{FP}\mathbb{1}_{y_i=0\&\hat{y}=1} + c_{FN}\mathbb{1}_{y_i=1\&\hat{y}_i=0}$$

Weighted Misclassification Error

We now define two costs: (1) the cost of the *FP* denoted c_{FP} and (2) the cost of the *FN* denoted c_{FN}. We then define the weighted misclassification error evaluation metric:

$$ME_w := \frac{1}{n}\sum_{i=1}^{n} c_{FP}\mathbb{1}_{y_i=0\,\&\,\hat{y}=1} + c_{FN}\mathbb{1}_{y_i=1\,\&\,\hat{y}_i=0}$$

We now vary p_0 to locate the model that optimizes this error to be minimum.

Minimum Weighted Misclassification Error

Let's assume that $c_{FN} = \$1000$ and $c_{FP} = \$100$ just for the example's sake. Note: this is a **cost ratio** of 10:1.

	Prob	TP	TN	FP	FN	COST
1	0.8117	1	5163	0	1868	1868000
2	0.8104	2	5163	0	1867	1867000
3	0.8093	3	5163	0	1866	1866000
4	0.8092	4	5163	0	1865	1865000
5	0.8090	5	5163	0	1864	1864000
6	0.8085	6	5163	0	1863	1863000
7	0.8083	7	5163	0	1862	1862000
8	0.8082	8	5163	0	1861	1861000
9	0.8079	9	5163	0	1860	1860000

We now calculate the cost and find the minimum model (i.e. the p_0 to ship). [JMP]

Minimum Weighted Misclassification Error

Let's assume that $c_{FN} = \$1000$ and $c_{FP} = \$100$ just for the example's sake. Note: this is a cost ratio of 10:1.

	Prob	TP	TN	FP	FN	COST
1	0.8117	1	5163	0	1868	1868000
2	0.8104	2	5163	0	1867	1867000
3	0.8093	3	5163	0	1866	1866000
4	0.8092	4	5163	0	1865	1865000
5	0.8090	5	5163	0	1864	1864000
6	0.8085	6	5163	0	1863	1863000
7	0.8083	7	5163	0	1862	1862000
8	0.8082	8	5163	0	1861	1861000
9	0.8079	9	5163	0	1860	1860000

We now calculate the cost and find the minimum model (i.e. the p_0 to ship). [JMP] Or alternatively, we can select the model with the closest $FN/FP \approx 10:1$ to match the stakeholder preference of the desired cost ratio. Why would this be good?

Expected Value Calculation

You can also imagine assignment of both costs *and* benefits:

$p_0 = 0.1$		$\hat{y}$	
		1	0
y	1	b_{TP}	c_{FN}
	0	c_{FP}	b_{TN}

Expected Value Calculation

You can also imagine assignment of both costs *and* benefits:

	$p_0 = 0.1$	$\hat{y}$ 1	0
y	1	b_{TP}	c_{FN}
	0	c_{FP}	b_{TN}

and then use the confusion matrix to estimate probabilities:

	$p_0 = 0.1$	$\hat{y}$ 1	0
y	1	25.1%	1.3%
	0	40.0%	35.5%

Expected Value Calculation

You can also imagine assignment of both costs *and* benefits:

$p_0 = 0.1$		$\hat{y}$ 1	0
y	1	b_{TP}	c_{FN}
	0	c_{FP}	b_{TN}

and then use the confusion matrix to estimate probabilities:

$p_0 = 0.1$		$\hat{y}$ 1	0
y	1	25.1%	1.3%
	0	40.0%	35.5%

The expected value would be?

$$\begin{aligned} \mathbb{E}[T] &= p_{TP} \times b_{TP} + p_{TN} \times b_{TN} + p_{FP} \times c_{FP} + p_{FN} \times c_{FN} \\ &\approx \hat{p}_{TP} \times b_{TP} + \hat{p}_{TN} \times b_{TN} + \hat{p}_{FP} \times c_{FP} + \hat{p}_{FN} \times c_{FN} \end{aligned}$$

Highest expected value model is shipped (ex. from Provost & Fawcett, 2013).

$\hat{p}$'s as Ordinal Values

One final point... If we were on a mission to find the top m churners. What would we do?

$\hat{p}$'s as Ordinal Values

One final point... If we were on a mission to find the top m churners. What would we do? Sort the $\hat{p}$'s and return the top m.