Logistic Regression and the ROC curve

Schwartz

September 30, 2017

Odd, even at best

In 2015 Leicester City was given 5000 to 1 odds to win the English Premier League. Actually, these are the longest odds ***ever seen*** for ***any*** top tier sporting league... ***ever***. To put this in perspective, the current odds out of Vegas for “the most unlikely team to win the 2016/2017 NFL season” – woefully disastrous Cleveland* Browns – are 200 to 1.

Since the clubs inception in 1890, Leicester City has only managed to appear in the Premier league 10 seasons. They had only been promoted the previous season and just barely escaped relegation in their final match that season. Only five teams – Arsenal, Chelsea, Liverpool, Man. City, and Man. U. – have held the trophy for the past 21 seasons.

Only a few stout souls put money down on Leicester City last year. And when Leicester City (*literally against all odds*) won the premiership last season in absolutely stunning, unbelievable, and unprecedented fashion, those stout souls got paid. Everyone, that is, except for John Micklethwait. John M has made the same bet – 20 pounds ($29) that Leicester will win their division – every August for the past 20 years. Every year, that is, except this one. Last year he moved from London to New York and missed placing his bet. That’s a pity for John M because if he had made his bet he would have won 100,000 pounds, or $145,355.

Overall, $3,000 was bet on Leicester City last season. The *unprecedented* $15,000,000 payout nearly bankrupted the bookmakers. John M got $0.

*Cleveland’s 52-year championship drought ended with the 2015/16 NBA season

Odds

$$\text{Odds} = \frac{p}{1-p} \implies \text{p} = \frac{Odds}{1+Odds} = \frac{1}{1+Odds^{-1}}$$

$$1-p = \frac{1}{1+Odds}$$

Objectives

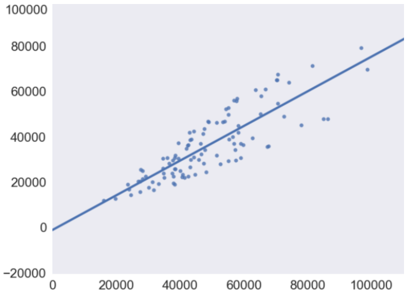
Morning

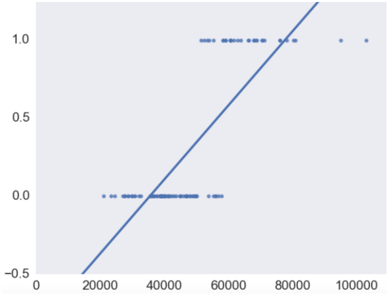
- Know why logistic regression is a thing:
  - Classification vs. Regression
  - Link functions
- Interpreting Logistic Regression
  - Fitted Values (probabilities)
  - Coefficients (log odds ratios)

Afternoon

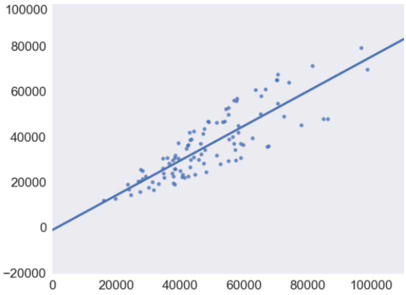
- T+, T-, F+, F- and other terminology
  - Confusion Matricies
- Thresholding Classification rules
  - ROC curves

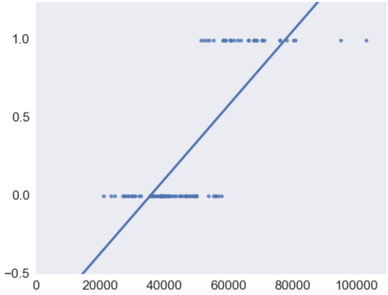
Linear Regression





Linear Regression

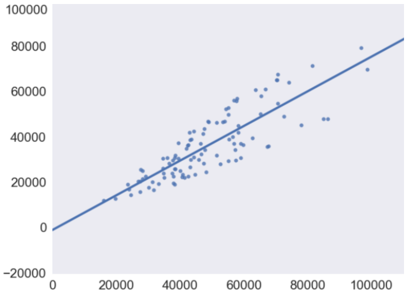


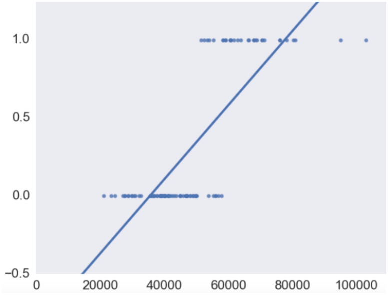


Is this satisfactory

for predicting probability?

Linear Regression

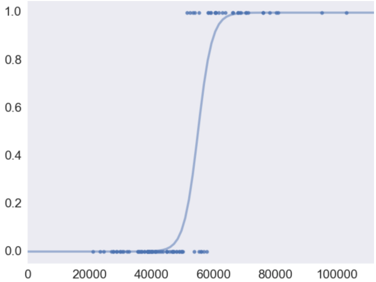




Is this satisfactory

for predicting probability?

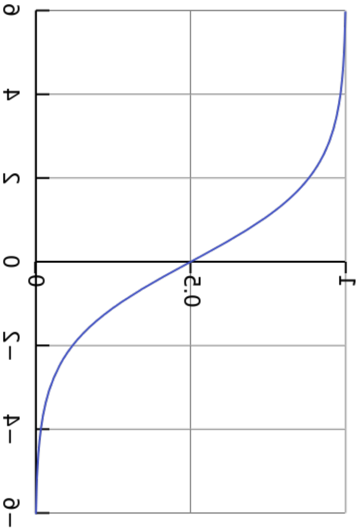
How about this instead $\Longrightarrow$



Link functions

- The "logit"

$$g(p) = \log\left(\frac{p}{1-p}\right)$$



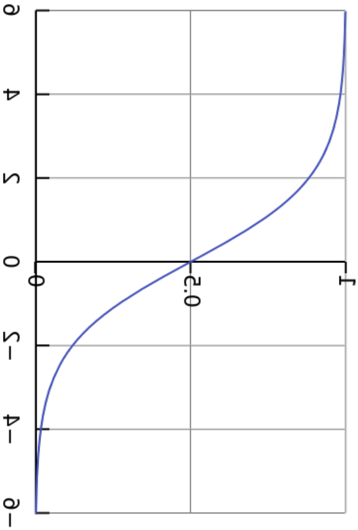
Link functions

- The "logit"

$$g(p) = \log\left(\frac{p}{1-p}\right)$$

- maps

$$p \in [0,1] \mapsto Z \in \mathbb{R}$$



Link functions

- The "logit"

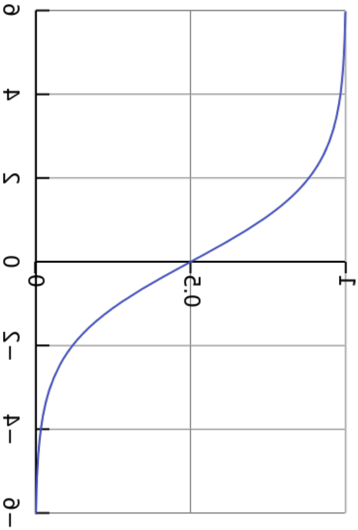
$$g(p) = \log\left(\frac{p}{1-p}\right)$$

- maps

$$p \in [0,1] \mapsto Z \in \mathbb{R}$$

Probabilities are from 0 to 1

But log odds go $-\infty$ to ∞!



Link functions

- The "logit"

$$g(p) = \log\left(\frac{p}{1-p}\right)$$

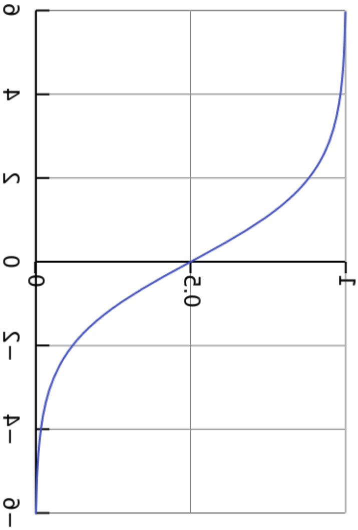
- maps

$$p \in [0,1] \mapsto Z \in \mathbb{R}$$

Probabilities are from 0 to 1

But log odds go $-\infty$ to ∞!

Don't be at odds with odds!



Logistic "regression"

- In Linear Model Regression, for *real valued outcomes*, we used

$$\hat{Y}_i$$

Logistic "regression"

- In Linear Model Regression, for *real valued outcomes*, we used

$$\hat{Y}_i = \mathrm{E}\left[Y\right] = Z = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

Logistic “regression”

- In Linear Model Regression, for *real valued outcomes*, we used

$$\hat{Y}_i = \mathrm{E}\left[Y\right] = Z = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

- For a *binary* outcome Y, we instead define

$$\hat{Y}_i = \Pr\left(Y = 1\right)$$

Logistic "regression"

- In Linear Model Regression, for *real valued outcomes*, we used

$$\hat{Y}_i = \mathrm{E}\left[Y\right] = Z = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

- For a *binary* outcome Y, we instead define

$$\hat{Y}_i = \Pr\left(Y = 1\right) = \mathrm{E}\left[Y\right]$$

Logistic "regression"

- In Linear Model Regression, for *real valued outcomes*, we used

$$\hat{Y}_i = \mathrm{E}[Y] = Z = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

- For a *binary* outcome Y, we instead define

$$\hat{Y}_i = \Pr(Y = 1) = \mathrm{E}[Y] = g^{-1}(Z)$$

because how else can Z stay between 0 and 1??

Logistic "regression"

- In Linear Model Regression, for *real valued outcomes*, we used

$$\hat{Y}_i = \mathrm{E}[Y] = Z = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

- For a *binary* outcome Y, we instead define

$$\hat{Y}_i = \Pr(Y = 1) = \mathrm{E}[Y] = g^{-1}(Z) = \frac{\exp(Z)}{1 + \exp(Z)}$$
$$= \frac{1}{1+\exp(-Z)}$$

Logistic "regression"

- In Linear Model Regression, for *real valued outcomes*, we used

$$\hat{Y}_i = \mathrm{E}\left[Y\right] = Z = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

- For a *binary* outcome Y, we instead define

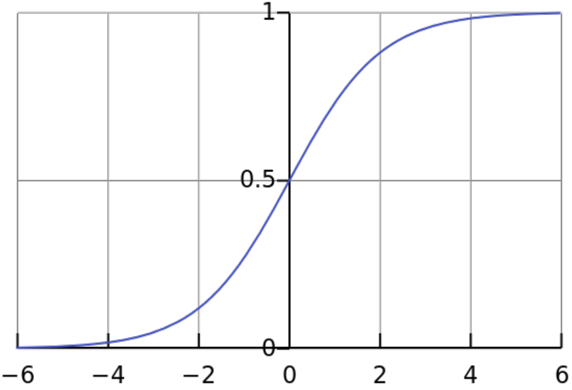
$$\hat{Y}_i = \Pr\left(Y = 1\right) = \mathrm{E}\left[Y\right] = g^{-1}(Z) = \frac{\exp(Z)}{1 + \exp(Z)}$$
$$= \frac{1}{1+\exp(-Z)}$$

So $g(p) = Z = \log\left(\frac{p}{1-p}\right) \in \mathbb{R}$ (which is called the logit function)

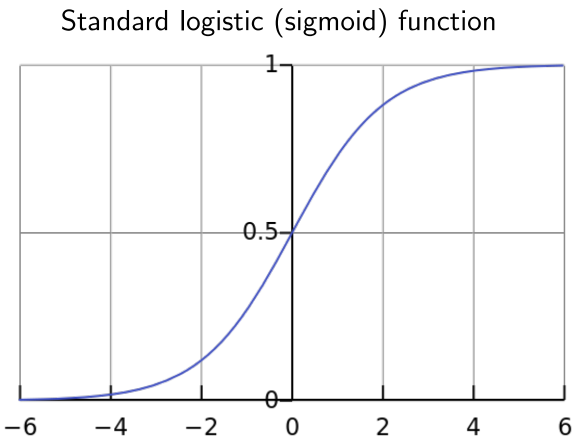
and $\underline{Z = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m \in \mathbb{R} \text{ models the log odds}}$

Linear model on log odds $\Longrightarrow$ transformed to probabilities

Standard logistic (sigmoid) function



Linear model on log odds $\Longrightarrow$ transformed to probabilities



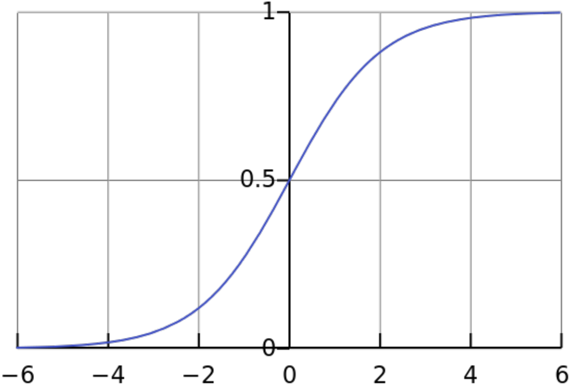
$$Z = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

(linear model on log odds)

Linear model on log odds $\Longrightarrow$ transformed to probabilities

Standard logistic (sigmoid) function

$\hat{Y}$



$$Z = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

(linear model on log odds)

Logarithmic Scale

- So
$$\Pr(Y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_m)}}$$

Logarithmic Scale

- So

$$\Pr(Y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_m)}}$$

- And we can quickly see then that

$$\frac{Pr(Y = 1|x)}{Pr(Y = 0|x)} =$$

Logarithmic Scale

- So
$$\Pr(Y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_m)}}$$
- And we can quickly see then that
$$\frac{Pr(Y = 1|x)}{Pr(Y = 0|x)} = \exp(\beta_0)\exp(\beta_1 x_1)\cdots\exp(\beta_m x_m)$$

Logarithmic Scale

- So
$$\Pr(Y = 1|x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_m)}}$$
- And we can quickly see then that
$$\frac{Pr(Y = 1|x)}{Pr(Y = 0|x)} = \exp(\beta_0)\exp(\beta_1 x_1)\cdots\exp(\beta_m x_m)$$
- So for a 1-unit increase in x_j

Logarithmic Scale

- So
$$\Pr(Y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_m)}}$$
- And we can quickly see then that
$$\frac{Pr(Y = 1|x)}{Pr(Y = 0|x)} = \exp(\beta_0)\exp(\beta_1 x_1)\cdots\exp(\beta_m x_m)$$
- So for a 1-unit increase in x_j
there is a $\exp(\beta_j)$ multiplicative increase in *the odds*

Logarithmic Scale

- So
$$\Pr(Y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_m)}}$$
- And we can quickly see then that
$$\frac{Pr(Y = 1|x)}{Pr(Y = 0|x)} \quad = \quad \exp(\beta_0) \exp(\beta_1 x_1) \cdots \exp(\beta_m x_m)$$
- So for a 1-unit increase in x_j
there is a $\exp(\beta_j)$ multiplicative increase in *the odds*
- I.e., *the odds* are linear in x on a multiplicative, i.e., odds increase with x on a *logorithmic* scale with base $\exp(\beta_j)$

Logarithmic Scale

- So
$$\Pr(Y = 1|x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_m)}}$$
- And we can quickly see then that
$$\frac{Pr(Y = 1|x)}{Pr(Y = 0|x)} \quad = \quad \exp(\beta_0)\exp(\beta_1 x_1)\cdots\exp(\beta_m x_m)$$
- So for a 1-unit increase in x_j
  there is a $\exp(\beta_j)$ multiplicative increase in *the odds*
- I.e., *the odds* are linear in x on a multiplicative, i.e., odds increase with x on a *logorithmic* scale with base $\exp(\beta_j)$
- The *log* odds $log\left(\frac{Pr(Y=1|x)}{Pr(Y=0|x)}\right)$ are on a linear scale
  $(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_m)$

The Odds Ratio (OR)

- Equivalently, $\exp(\beta_j)$ is the *odds ratio (OR)* between 1-unit differences in x_j (e.g., 0 versus 1) when other x's are constant

$$\exp(\beta_j) = \frac{Pr(Y=1|x_j+1, x_{-j})/Pr(Y=0|x_j+1, x_{-j})}{Pr(Y=1|x)/Pr(Y=0|x)}$$

since

$$\begin{aligned} & \frac{Pr(Y=1|x_j+1, x_{-j})}{Pr(Y=0|x_j+1, x_{-j})} \\ = \ & \exp(\beta_0)\exp(\beta_1 x_1)\cdots\exp(\beta_j(x_j+1))\cdots\exp(\beta_m x_m) \\ = \ & \exp(\beta_0)\exp(\beta_1 x_1)\cdots\exp(\beta_j x_j) exp(\beta_j)\cdots\exp(\beta_m x_m) \end{aligned}$$

and

$$\begin{aligned} & \frac{Pr(Y=1|x)}{Pr(Y=0|x)} \\ = \ & \exp(\beta_0)\exp(\beta_1 x_1)\cdots\exp(\beta_j x_j)\cdots\exp(\beta_m x_m) \end{aligned}$$

The Odds Ratio (OR)

- Equivalently, $\exp(\beta_j)$ is the *odds ratio (OR)* between 1-unit differences in x_j (e.g., 0 versus 1) when other x's are constant

$$\exp(\beta_j) = \frac{Pr(Y = 1|x_j + 1, x_{-j})/Pr(Y = 0|x_j + 1, x_{-j})}{Pr(Y = 1|x)/Pr(Y = 0|x)}$$

since

$$\begin{aligned}
&\frac{Pr(Y = 1|x_j + 1, x_{-j})}{Pr(Y = 0|x_j + 1, x_{-j})} \\
&= \exp(\beta_0)\exp(\beta_1 x_1)\cdots\exp(\beta_j(x_j + 1))\cdots\exp(\beta_m x_m) \\
&= \exp(\beta_0)\exp(\beta_1 x_1)\cdots\exp(\beta_j x_j)exp(\beta_j)\cdots\exp(\beta_m x_m)
\end{aligned}$$

and

$$\begin{aligned}
&\frac{Pr(Y = 1|x)}{Pr(Y = 0|x)} \\
&= \exp(\beta_0)\exp(\beta_1 x_1)\cdots\exp(\beta_j x_j)\cdots\exp(\beta_m x_m)
\end{aligned}$$

- So β_j is the change in $log(OR)$ for one unit changes in x_j...

Logistic Regression *Likelihood* and *Deviance*

- Likelihood

$$f(\mathbf{Y}|\beta, \mathbf{x}) = \prod \left(\frac{1}{1 + e^{-\mathbf{x}_i^T \boldsymbol{\beta}}} \right)^{Y_i} \left(\frac{1}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right)^{1-Y_i}$$

Logistic Regression *Likelihood* and *Deviance*

- Likelihood

$$f(\mathbf{Y}|\beta, \mathbf{x}) = \prod \left(\frac{1}{1 + e^{-\mathbf{x}_i^T \boldsymbol{\beta}}} \right)^{Y_i} \left(\frac{1}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right)^{1-Y_i}$$

- Deviance

$$D_M = \; -2 \left(\log f(\mathbf{Y}|\hat{\beta}, \mathbf{x}) - \log f(\mathbf{Y}|\mathbf{Y}) \right)$$
$$\overset{\text{approx.}}{\sim} \; \chi^2_{n-p-1}$$

$$n = \text{sample size}$$
$$p = \text{number of coefficients in model } M$$
$$f(\mathbf{Y}|\mathbf{Y}) = \text{saturated model } (\mathbf{Y} \text{ perfectly predicted})$$

More Deviance

- In logistic regression
  $D_M = -2\left(\log f(\mathbf{Y}|\hat{\beta}, \mathbf{x}) - \log f(\mathbf{Y}|\mathbf{Y})\right) =$

More Deviance

- In logistic regression
  $D_M = -2\left(\log f(\mathbf{Y}|\hat{\beta}, \mathbf{x}) - \log f(\mathbf{Y}|\mathbf{Y})\right) =$

$$-2\sum Y_i log(p_i) + (1 - Y_i) log(1 - p_i) - Y_i log(Y_i) - (1 - Y_i) log(1 - Y_i)$$

More Deviance

- In logistic regression
  $D_M = -2\left(\log f(\mathbf{Y}|\hat{\beta}, \mathbf{x}) - \log f(\mathbf{Y}|\mathbf{Y})\right) =$

$$-2\sum Y_i log(p_i) + (1 - Y_i) log(1 - p_i) - Y_i log(Y_i) - (1 - Y_i) log(1 - Y_i)$$

[show this]

More Deviance

- In logistic regression
  $D_M = -2\left(\log f(\mathbf{Y}|\hat{\beta},\mathbf{x}) - \log f(\mathbf{Y}|\mathbf{Y})\right) =$

$$-2\sum Y_i log(p_i) + (1 - Y_i) log(1 - p_i) - Y_i log(Y_i) - (1 - Y_i) log(1 - Y_i)$$

[show this] $\overset{\text{approx.}}{\sim} \chi^2_{n-p-1}$

More Deviance

- In logistic regression
  $D_M = -2\left(\log f(\mathbf{Y}|\hat{\beta},\mathbf{x}) - \log f(\mathbf{Y}|\mathbf{Y})\right) =$

$$-2\sum Y_i log(p_i) + (1-Y_i) log(1-p_i) - Y_i log(Y_i) - (1-Y_i) log(1-Y_i)$$

[show this] $\overset{\text{approx.}}{\sim} \chi^2_{n-p-1}$

- In linear regression

$$D_M = \frac{RSS}{\sigma^2} \qquad \text{[show this]}$$
$$= \frac{\sum(Y_i - \hat{Y})^2}{\sigma^2} = (n-p-1)\frac{s^2}{\sigma^2} \overset{\text{approx.}}{\sim} \chi^2_{n-p-1}$$

More Deviance

- In logistic regression
  $D_M = -2\left(\log f(\mathbf{Y}|\hat{\beta}, \mathbf{x}) - \log f(\mathbf{Y}|\mathbf{Y})\right) =$

$$-2\sum Y_i log(p_i) + (1 - Y_i) log(1 - p_i) - Y_i log(Y_i) - (1 - Y_i) log(1 - Y_i)$$

[show this] $\overset{\text{approx.}}{\sim} \chi^2_{n-p-1}$

- In linear regression

$$D_M = \frac{RSS}{\sigma^2} \qquad \text{[show this]}$$

$$= \frac{\sum(Y_i - \hat{Y})^2}{\sigma^2} = (n - p - 1)\frac{s^2}{\sigma^2} \overset{\text{approx.}}{\sim} \chi^2_{n-p-1}$$

[what are *residuals*?]

More Deviance

- In logistic regression
  $$D_M = -2\left(\log f(\mathbf{Y}|\hat{\beta}, \mathbf{x}) - \log f(\mathbf{Y}|\mathbf{Y})\right) =$$

$$-2\sum Y_i log(p_i) + (1 - Y_i) log(1 - p_i) - Y_i log(Y_i) - (1 - Y_i) log(1 - Y_i)$$

[show this] $\overset{\text{approx.}}{\sim} \chi^2_{n-p-1}$

- In linear regression

$$D_M = \frac{RSS}{\sigma^2} \qquad \text{[show this]}$$
$$= \frac{\sum(Y_i - \hat{Y})^2}{\sigma^2} = (n - p - 1)\frac{s^2}{\sigma^2} \overset{\text{approx.}}{\sim} \chi^2_{n-p-1}$$

[what are *residuals*?] [what are *"residuals"* in logistic regression?]

Fitting Logistic Regression

- MLE

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmax}} \prod \left(\frac{1}{1 + e^{-\boldsymbol{x}^T \boldsymbol{\beta}}} \right)^{Y_i} \left(\frac{1}{1 + e^{\boldsymbol{x}^T \boldsymbol{\beta}}} \right)^{1 - Y_i}$$

$$\Longleftrightarrow$$

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\ D_{\beta} = \underset{\beta}{\text{argmax}} \left(\log f(\mathbf{Y}|\hat{\beta}, \mathbf{x}) - \log f(\mathbf{Y}|\mathbf{Y}) \right)$$

Fitting Logistic Regression

- MLE

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \prod \left(\frac{1}{1 + e^{-\mathbf{x}^T \boldsymbol{\beta}}} \right)^{Y_i} \left(\frac{1}{1 + e^{\mathbf{x}^T \boldsymbol{\beta}}} \right)^{1 - Y_i}$$

$$\Longleftrightarrow$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ D_{\beta} = \underset{\beta}{\operatorname{argmax}} \left(\log f(\mathbf{Y}|\hat{\beta}, \mathbf{x}) - \log f(\mathbf{Y}|\mathbf{Y}) \right)$$

- Closed form solution not available (like with linear regression)
  - Optimization done via Newton-Rhapson or Gradient Decent
  - Coefficient standard errors can also be numerically estimated!

Fitting Logistic Regression

- MLE

$$\hat{\boldsymbol\beta} = \underset{\boldsymbol\beta}{\operatorname{argmax}} \prod \left(\frac{1}{1+e^{-\mathbf{x}^T\boldsymbol\beta}}\right)^{Y_i} \left(\frac{1}{1+e^{\mathbf{x}^T\boldsymbol\beta}}\right)^{1-Y_i}$$

$$\Longleftrightarrow$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ D_{\beta} = \underset{\beta}{\operatorname{argmax}} \left(\log f(\mathbf{Y}|\hat{\beta},\mathbf{x}) - \log f(\mathbf{Y}|\mathbf{Y})\right)$$

- Closed form solution not available (like with linear regression)
  - Optimization done via Newton-Rhapson or Gradient Decent
  - Coefficient standard errors can also be numerically estimated!
- Convergence difficulties will be encountered if
  - too many features ($n/p < 10$) or data is sparse/imbalanced

Fitting Logistic Regression

- MLE

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmax}} \prod \left(\frac{1}{1+e^{-\mathbf{x}^T\boldsymbol{\beta}}}\right)^{Y_i} \left(\frac{1}{1+e^{\mathbf{x}^T\boldsymbol{\beta}}}\right)^{1-Y_i}$$

$$\Longleftrightarrow$$

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\ D_{\beta} = \underset{\beta}{\text{argmax}} \left(\log f(\mathbf{Y}|\hat{\beta}, \mathbf{x}) - \log f(\mathbf{Y}|\mathbf{Y})\right)$$

- Closed form solution not available (like with linear regression)
  - Optimization done via Newton-Rhapson or Gradient Decent
  - Coefficient standard errors can also be numerically estimated!
- Convergence difficulties will be encountered if
  - too many features ($n/p < 10$) or data is sparse/imbalanced
- Coefficient standard errors will be compromised when
  - predicted probabilities are only $\sim$1 or $\sim$0 (separated classes)
  - There is covariate multicollinearity (as with linear regression)

Fitting Logistic Regression

- MLE

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmax}} \prod \left(\frac{1}{1+e^{-\mathbf{x}^T\boldsymbol{\beta}}}\right)^{Y_i} \left(\frac{1}{1+e^{\mathbf{x}^T\boldsymbol{\beta}}}\right)^{1-Y_i}$$

$$\Longleftrightarrow$$

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}}\ D_{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmax}} \left(\log f(\mathbf{Y}|\hat{\boldsymbol{\beta}},\mathbf{x}) - \log f(\mathbf{Y}|\mathbf{Y})\right)$$

- Closed form solution not available (like with linear regression)
  - Optimization done via Newton-Rhapson or Gradient Decent
  - Coefficient standard errors can also be numerically estimated!
- Convergence difficulties will be encountered if
  - too many features ($n/p < 10$) or data is sparse/imbalanced
- Coefficient standard errors will be compromised when
  - predicted probabilities are only $\sim$1 or $\sim$0 (separated classes)
  - There is covariate multicollinearity (as with linear regression)
- What if, for some λ, we choose $\boldsymbol{\beta}$ to minimize

$$-\prod \left(\frac{1}{1+e^{-\mathbf{x}_i^T\boldsymbol{\beta}}}\right)^{Y_i} \left(\frac{1}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}\right)^{1-Y_i} + \lambda||\boldsymbol{\beta}||^2?$$

Fitting Logistic Regression

- MLE

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmax}} \prod \left(\frac{1}{1+e^{-\boldsymbol{x}^T\boldsymbol{\beta}}}\right)^{Y_i} \left(\frac{1}{1+e^{\boldsymbol{x}^T\boldsymbol{\beta}}}\right)^{1-Y_i}$$

$$\Longleftrightarrow$$

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}}\ D_{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmax}} \left(\log f(\mathbf{Y}|\hat{\boldsymbol{\beta}}, \mathbf{x}) - \log f(\mathbf{Y}|\mathbf{Y})\right)$$

- Closed form solution not available (like with linear regression)
  - Optimization done via Newton-Rhapson or Gradient Decent
  - Coefficient standard errors can also be numerically estimated!
- Convergence difficulties will be encountered if
  - too many features ($n/p < 10$) or data is sparse/imbalanced
- Coefficient standard errors will be compromised when
  - predicted probabilities are only $\sim$1 or $\sim$0 (separated classes)
  - There is covariate multicollinearity (as with linear regression)
- What if, for some λ, we choose $\boldsymbol{\beta}$ to minimize

$$-\prod \left(\frac{1}{1+e^{-\mathbf{x}_i^T\boldsymbol{\beta}}}\right)^{Y_i} \left(\frac{1}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}\right)^{1-Y_i} + \lambda|\boldsymbol{\beta}|_1?$$

Pseudo R^2

- McFadden's *pseudo* $R^2 = 1 - D_M/D_0$

Pseudo R^2

- McFadden's *pseudo* $R^2 = 1 - D_M/D_0$
  Proportion of deviance explained

Pseudo R^2

- McFadden's *pseudo* $R^2 = 1 - D_M/D_0$
  Proportion of deviance explained

- Compare to linear model $R^2 = 1 - RSS/TSS$
  Proportion of variance explained

Pseudo R^2

- McFadden's *pseudo* $R^2 = 1 - D_M/D_0$
  Proportion of deviance explained
- Compare to linear model $R^2 = 1 - RSS/TSS$
  Proportion of variance explained
- `http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm`

Model Comparison

Remember, $D_M = -2(logf(\mathbf{Y}|\hat{\theta}_M) - logf(\mathbf{Y}|\mathbf{Y}))$, so

Model Comparison

Remember, $D_M = -2(logf(\mathbf{Y}|\hat{\theta}_M) - logf(\mathbf{Y}|\mathbf{Y}))$, so

- comparison of (*reduced* and *full*) models can be done using

$$D_R - D_F = -2\left(\log f(Y|\hat{\theta}^R) - \log f(Y|\hat{\theta}^F)\right) \overset{\text{approx.}}{\sim} \chi^2_k$$

where model R is nested in model F with k fewer parameters

Model Comparison

Remember, $D_M = -2(logf(\mathbf{Y}|\hat{\theta}_M) - logf(\mathbf{Y}|\mathbf{Y}))$, so

- comparison of (*reduced* and *full*) models can be done using

$$D_R - D_F = -2\left(\log f(Y|\hat{\theta}^R) - \log f(Y|\hat{\theta}^F)\right) \overset{\text{approx.}}{\sim} \chi^2_k$$

where model R is nested in model F with k fewer parameters

- For *non-nested* models, compare

$$AIC: \; -2\log f(Y|\hat{\theta}) + 2k$$
$$BIC: \; -2\log f(Y|\hat{\theta}) + k\log(n)$$

Model Comparison

Remember, $D_M = -2(logf(\mathbf{Y}|\hat{\theta}_M) - logf(\mathbf{Y}|\mathbf{Y}))$, so

- comparison of (*reduced* and *full*) models can be done using

$$D_R - D_F = -2\left(\log f(Y|\hat{\theta}^R) - \log f(Y|\hat{\theta}^F)\right) \overset{\text{approx.}}{\sim} \chi^2_k$$

where model R is nested in model F with k fewer parameters

- For *non-nested* models, compare

$$AIC: \; -2\log f(Y|\hat{\theta}) + 2k$$
$$BIC: \; -2\log f(Y|\hat{\theta}) + k\log(n)$$

How else could you compare nested or non-nested models?

Uses for logistic regression?

Uses for logistic regression?

- Predict probabilities

Uses for logistic regression?

- Predict probabilities
- Classify outcomes (based on probabilities)

Uses for logistic regression?

- Predict probabilities
- Classify outcomes (based on probabilities)
- Identify feature associations with class labels

Uses for logistic regression?

- Predict probabilities
- Classify outcomes (based on probabilities)
- Identify feature associations with class labels

- Balancing observational comparison groups on *propensity scores* $\Pr(T|x)$ which controls bias from group covariate composition differences

Confusion Matrix (and questions)

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

Confusion Matrix (and questions)

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

Which cells are Type I and Type II error?

Confusion Matrix (and questions)

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

Which cells are Type I and Type II error?

What is the power of a test?

Confusion Matrix (and questions)

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

Which cells are Type I and Type II error?

What is the power of a test? $\Pr(\text{Reject } H_0 \mid H_A \text{ True})$

Confusion Matrix (and questions)

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

Which cells are Type I and Type II error?

What is the power of a test? Pr(Reject H_0 | H_A True)

And the α-significance level?

Confusion Matrix (and questions)

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

Which cells are Type I and Type II error?

What is the power of a test? Pr(Reject H_0 | H_A True)

And the α-significance level? Pr(Reject H_0 | H_0 True)

Sensitivity & Specificity

- Sensitivity: % of "true H_A" tests <u>correctly called</u> $\left(\frac{\#TP}{\#TP+\#FN}\right)$

  "How **sensitive** are we to variations from H_0?"

Sensitivity & Specificity

- Sensitivity: % of "true H_A" tests correctly called $\left(\frac{\#TP}{\#TP+\#FN}\right)$

  "How **sensitive** are we to variations from H_0?"

  - Also called *True Positive Rate*
  - Also called *Recall*

Sensitivity & Specificity

- Sensitivity: % of "true H_A" tests correctly called $\left(\frac{\#TP}{\#TP+\#FN}\right)$

  "How **sensitive** are we to variations from H_0?"

  - Also called *True Positive Rate*
  - Also called *Recall*

- Specificity: % of "true H_0" tests correctly called $\left(\frac{\#TN}{\#TN+\#FP}\right)$

  "How **specific** must evidence against H_0 be?"

Sensitivity & Specificity

- Sensitivity: % of "true H_A" tests correctly called $\left(\frac{\#TP}{\#TP+\#FN}\right)$

  "How **sensitive** are we to variations from H_0?"

  - Also called *True Positive Rate*
  - Also called *Recall*

- Specificity: % of "true H_0" tests correctly called $\left(\frac{\#TN}{\#TN+\#FP}\right)$

  "How **specific** must evidence against H_0 be?"

  - Also called *True Negative Rate*
  - The "1 minus" related *False Positive Rate* is $\left(\frac{\#FP}{\#TN+\#FP}\right)$

Sensitivity & Specificity

- Sensitivity: % of "true H_A" tests correctly called $\left(\frac{\#TP}{\#TP+\#FN}\right)$

  "How **sensitive** are we to variations from H_0?"

  - Also called *True Positive Rate*
  - Also called *Recall*

- Specificity: % of "true H_0" tests correctly called $\left(\frac{\#TN}{\#TN+\#FP}\right)$

  "How **specific** must evidence against H_0 be?"

  - Also called *True Negative Rate*
  - The "1 minus" related *False Positive Rate* is $\left(\frac{\#FP}{\#TN+\#FP}\right)$

- Test *power*, $1 - \beta = \Pr(\text{Reject } H_0 \mid H_A \text{ True})$, ***IS* Sensitivity**
- α-significance level, $\Pr(\text{Reject } H_0 \mid H_0 \text{ True})$, ***IS* 1-Specificity**

Sensitivity & Specificity

- Sensitivity: % of "true H_A" tests correctly called $\left(\frac{\#TP}{\#TP+\#FN}\right)$

  "How **sensitive** are we to variations from H_0?"

  - Also called *True Positive Rate*
  - Also called *Recall*

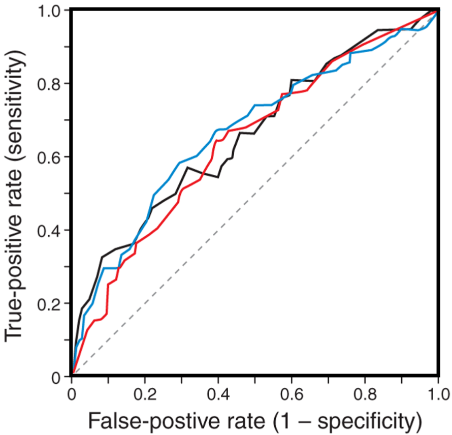
- Specificity: % of "true H_0" tests correctly called $\left(\frac{\#TN}{\#TN+\#FP}\right)$

  "How **specific** must evidence against H_0 be?"

  - Also called *True Negative Rate*
  - The "1 minus" related *False Positive Rate* is $\left(\frac{\#FP}{\#TN+\#FP}\right)$

- Test *power*, $1-\beta = \Pr(\text{Reject } H_0 \mid H_A \text{ True})$, ***IS* Sensitivity**
- α-significance level, $\Pr(\text{Reject } H_0 \mid H_0 \text{ True})$, ***IS* 1-Specificity**
- I.e, Type I & II error rates **are 1-Specificity & 1-Sensitivity**

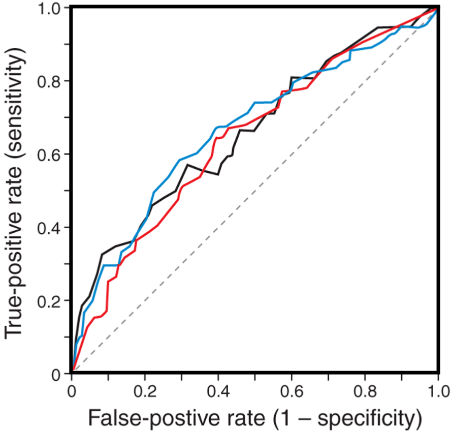
ROC/AUC



1.0
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1.0
True-positive rate (sensitivity)
False-postive rate (1 – specificity)

ROC/AUC

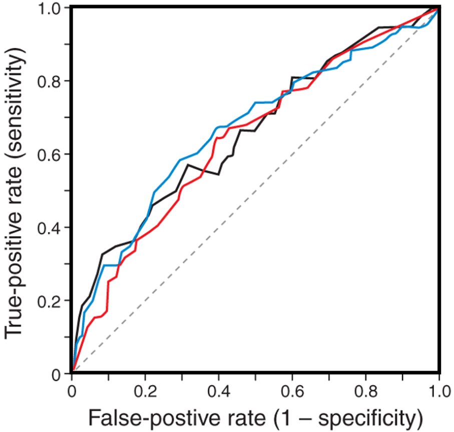
$Power = 1 - \beta$



α

ROC/AUC

Power $= 1 - \beta$



α

`https://www.youtube.com/watch?v=JAQC59ArFJw`
`https://www.youtube.com/watch?v=bhvvxNUbIpo`

Other BIGGIES: *precision, FDR, and accuracy*

- Precision: % **positives** <u>called **correctly**</u> $\left(\frac{\#TP}{\#TP+\#FP}\right)$

Other BIGGIES: *precision, FDR, and accuracy*

- Precision: % **positives** called **correctly** $\left(\frac{\#TP}{\#TP+\#FP}\right)$

  "How ***P****recise* are we in our '***P****ositives*' (when we reject H_0)?"

Other BIGGIES: *precision, FDR, and accuracy*

- Precision: % **positives** called **correctly** $\left(\frac{\#TP}{\#TP+\#FP}\right)$

  "How ***P****recise* are we in our '***P****ositives*' (when we reject H_0)?"

  - Also called *Positive Predicted Value*

Other BIGGIES: *precision, FDR, and accuracy*

- Precision: % **positives** <u>called **correctly**</u> $\left(\frac{\#TP}{\#TP+\#FP}\right)$

  "How ***P****recise* are we in our '***P****ositives*' (when we reject H_0)?"

  - Also called *Positive Predicted Value*

- False Discovery Rate (FDR): % **positives** <u>called **incorrectly**</u> $\left(\frac{\#FP}{\#TP+\#FP}\right)$

Other BIGGIES: *precision, FDR, and accuracy*

- Precision: % **positives** called **correctly** $\left(\frac{\#TP}{\#TP+\#FP}\right)$

  "How ***P****recise* are we in our '***P****ositives*' (when we reject H_0)?"

  - Also called *Positive Predicted Value*

- False Discovery Rate (FDR): % **positives** called **incorrectly** $\left(\frac{\#FP}{\#TP+\#FP}\right)$

  "What's the error rate in our significant tests?"

Other BIGGIES: *precision, FDR, and accuracy*

- Precision: % **positives** called **correctly** $\left(\frac{\#TP}{\#TP+\#FP}\right)$

  "How ***P****recise* are we in our '***P****ositives*' (when we reject H_0)?"

  - Also called *Positive Predicted Value*

- False Discovery Rate (FDR): % **positives** called **incorrectly** $\left(\frac{\#FP}{\#TP+\#FP}\right)$

  "What's the error rate in our significant tests?"

  - *VERY* useful concept of multiple testing contexts
    a.k.a. for a *multiplicity adjustment* (where it's called a *q-value*)

Other BIGGIES: *precision, FDR, and accuracy*

- Precision: % **positives** called **correctly** $\left(\frac{\#TP}{\#TP+\#FP}\right)$

  "How ***P****recise* are we in our '***P****ositives*' (when we reject H_0)?"

  - Also called *Positive Predicted Value*

- False Discovery Rate (FDR): % **positives** called **incorrectly** $\left(\frac{\#FP}{\#TP+\#FP}\right)$

  "What's the error rate in our significant tests?"

  - *VERY* useful concept of multiple testing contexts
    a.k.a. for a *multiplicity adjustment* (where it's called a *q-value*)

- Accuracy: % of tests we we correctly call $\left(\frac{\#TP+\#TN}{\#Total}\right)$

Other BIGGIES: *precision, FDR, and accuracy*

- Precision: % **positives** <u>called **correctly**</u> $\left(\frac{\#TP}{\#TP + \#FP}\right)$

  "How ***P**recise* are we in our '***P**ositives*' (when we reject H_0)?"

  - Also called *Positive Predicted Value*

- False Discovery Rate (FDR): % **positives** <u>called **incorrectly**</u> $\left(\frac{\#FP}{\#TP + \#FP}\right)$

  "What's the error rate in our significant tests?"

  - *VERY* useful concept of multiple testing contexts
    a.k.a. for a *multiplicity adjustment* (where it's called a *q-value*)

- Accuracy: % of tests we <u>we correctly call</u> $\left(\frac{\#TP + \#TN}{\#Total}\right)$

  "Do we call hypotheses **accurately**?"

And just a one or two more...

		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 $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	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 $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	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–) $= \frac{\text{FNR}}{\text{TNR}}$	

Thanks, Wiki!

(You're the besht!)