Logistic Regression

via GLM
2008 US Election

- Some analysts say that Obama’s data science team basically won him the election
- For the first time, a team used data and statistical methods to model voter behavior
  - As opposed to ‘gut feelings’
- He was also the first president to advertise in a video game
- How did they do it?
Table A12.1: Logistic Regression Predicting Obama Vote Preference (Respondents in the Nonbattleground Only)

<table>
<thead>
<tr>
<th></th>
<th>B Coefficient</th>
<th>Standard Error</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.397</td>
<td>.375</td>
<td>4.04</td>
</tr>
<tr>
<td>Female (1=yes, 0=no)</td>
<td>-.082</td>
<td>.087</td>
<td>.922</td>
</tr>
<tr>
<td>Age (in years)</td>
<td>-.004</td>
<td>.003</td>
<td>.996</td>
</tr>
<tr>
<td>Black (1=yes, 0=no)</td>
<td>2.859</td>
<td>.255</td>
<td>17.445</td>
</tr>
<tr>
<td>Hispanic (1=yes, 0=no)</td>
<td>1.043</td>
<td>.154</td>
<td>2.838</td>
</tr>
<tr>
<td>Education (in years)</td>
<td>.102</td>
<td>.021</td>
<td>1.107</td>
</tr>
<tr>
<td>Household income (in thousands)</td>
<td>-.001</td>
<td>.001</td>
<td>.999</td>
</tr>
<tr>
<td>Republican (1=yes, 0=no)</td>
<td>-1.988</td>
<td>.117</td>
<td>.137</td>
</tr>
<tr>
<td>Democrat (1=yes, 0=no)</td>
<td>1.728</td>
<td>.101</td>
<td>5.632</td>
</tr>
<tr>
<td>Ideology (1=very liberal to 5=very conservative)</td>
<td>-.896</td>
<td>.045</td>
<td>.408</td>
</tr>
<tr>
<td>Number of days saw presidential campaign information on TV news in past week</td>
<td>-.009</td>
<td>.019</td>
<td>.991</td>
</tr>
<tr>
<td>Number of days heard about presidential campaign on talk radio in past week</td>
<td>-.045 **</td>
<td>.016</td>
<td>.956</td>
</tr>
<tr>
<td>Number of days saw presidential campaign information in newspapers in past week</td>
<td>.050 ***</td>
<td>.015</td>
<td>1.052</td>
</tr>
<tr>
<td>Number of days saw presidential campaign information on Internet in past week</td>
<td>.007</td>
<td>.015</td>
<td>1.007</td>
</tr>
<tr>
<td>Difference in national 30-second television advertising spending by campaigns (Obama – McCain and RNC) (Per $100,000 by week - 9/02/08 to Election Day)</td>
<td>.004 **</td>
<td>.001</td>
<td>1.004</td>
</tr>
</tbody>
</table>

N = 5,049

Cox & Snell R-square = .492
Nagelkerke R-square = .656
Percent Correct = 84.4

*p < .10  ** p < .05  *** p < .01  **** p < .001

Data: NAES08 telephone survey. Dates: 9/02/08 to 11/03/08.
## Background

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Independent Variable</th>
<th>Display</th>
<th>Parametric Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical</td>
<td>Categorical</td>
<td>Contingency Table</td>
<td>Chi-square Log-linear</td>
</tr>
<tr>
<td>Numerical</td>
<td>Categorical</td>
<td>Bar Charts</td>
<td>ANOVA</td>
</tr>
<tr>
<td>Numerical</td>
<td>Numerical</td>
<td>Scatterplot</td>
<td>Regression</td>
</tr>
<tr>
<td>Numerical</td>
<td>Numerical &amp; Categorical</td>
<td>Grouped Scatterplot</td>
<td>ANCOVA GLM</td>
</tr>
<tr>
<td>Binary</td>
<td>Numerical &amp; Categorical</td>
<td>Scatterplot</td>
<td>Logistic</td>
</tr>
</tbody>
</table>
Requirements

- Binary outcome \{0,1\}
  - Can later generalize this to multiple groups
- Numerical or categorical predictor
  - Can later add more predictors
- Independent observations
- Predictors not related
- Complete model
- No Normality assumption on error distribution
  - In fact, no error term on model at all!
LBW Dataset

- **Data description:**
  - **low:** indicator of birth weight less than 2.5kg
  - **age:** mother's age in years
  - **lwt:** mother's weight (lbs) at last menstrual period
  - **race:** mother's race ("white", "black", "other")
  - **smoke:** smoking status during pregnancy
  - **ht:** history of hypertension
  - **ui:** presence of uterine irritability
  - **ftv:** number of physician visits during first trimester
  - **ptl:** number of previous premature labours
  - **bwt:** birth weight in grams
What’s wrong with SLR?

- Outcome Y is either \{0, 1\}

\[
Y | X \sim Bernoulli(\pi)
\]

\[
E[Y | X] = \pi
\]

\[
\text{Var}[Y | X] = \pi(1 - \pi)
\]

- So two things:
  - Y is not Normal
  - \text{Var}(Y|X) is not constant
What’s wrong with SLR?

This:
What’s wrong with SLR?

And this:
What’s wrong with SLR?

- $E[Y|X] = \pi$, so model $P(Y = 1) = \pi(X) = X\beta$
- Looks fine, except a linear regression model is not bounded by $(0,1)$, and a probability IS
  - This is a major defect, and will result in predictions outside $(0,1)$ which are nonsense
  - Also, the variance will not be constant
- Interpretation of $\beta$:
  - Change in probability that $Y = 1$, for a unit change in $X$
Goal

- Construct a model where the outcome is bounded between 0 and 1
- Regression coefficients ($\beta$s) are still meaningful
- Inference is still available
What, then?

- Something like this would be nice:
What is the formula for that curve?

Actually, there are a few ways to get an “S-shaped” curve like that, but one of them is:

\[
f(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}
\]

This is the **logistic function**.

It can also “S” the other way.

It has a nice interpretation of the parameters.
So what are we modelling?

\[ \pi = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \]

\[ \Rightarrow \frac{1}{\pi} = 1 + e^{-(\beta_0 + \beta_1 x)} \]

\[ \Rightarrow \log \left( \frac{1}{\pi} - 1 \right) = -(\beta_0 + \beta_1 x) \]

\[ \Rightarrow \log \left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 x \]

- The LHS is called the “log-Odds”, and is linear in the parameters
A Linear Model?

\[
\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x
\]

\[
g(\pi) = \beta_0 + \beta_1 x
\]

\[
g(\mu) = X\beta
\]

● This is the same “Generalized Linear Model” we saw last week
Generalized Linear Model

\[ g(\mu) = X\beta \]

- \( g(\mu) \) is called a **Link Function**

\( g(\mu) = \mu \) is vanilla regression – Identity Link
\( g(\mu) = \log(\mu) \) is Poisson regression – Log Link
\( g(\mu) = \log \left( \frac{\mu}{1-\mu} \right) \) is Logistic regression – Logit Link

- Use MLE to get parameter estimates
Interpretation of $\beta_0$

- In SLR, in the absence of predictors, the best estimate for $Y$ is the overall mean
  
  $E[Y] = \beta_0 \Rightarrow \bar{Y} = b_0$

```r
> set.seed(2015)
> y <- rnorm(100, 10, 25)
> summary(lm(y ~ 1))

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) 
|----------|------------|---------|---------|
| (Intercept) | 8.727 | 2.524 | 3.458 | 0.000804

> mean(y)
[1] 8.726752
```
Interpretation of $\beta_0$

- MLE for $\pi = P(Y = 1)$ is the sample proportion $\hat{p}$
- In logistic regression in the absence of predictors, what is $\beta_0$?

\[
\log \left( \frac{\pi}{1 - \pi} \right) = \beta_0
\]

\[
\pi = \frac{1}{1 + e^{-\beta_0}}
\]

- So $\beta_0$ is (a function of) the marginal probability that $Y = 1$, ignoring predictors
Interpretation of $\beta_0$

```r
> y <- runif(100) < 0.7  # Generate 1s and 0s with 70% 1s
> fit <- glm(y ~ 1, family= binomial)
> summary(fit)
Call:
glm(formula = y ~ 1, family = binomial)

Coefficients:

Estimate  Std. Error z value Pr(>|z|)  
(Intercept) 1.1527     0.2341    4.923  8.53e-07

Null deviance: 110.22  on 99  degrees of freedom
Residual deviance: 110.22  on 99  degrees of freedom
AIC: 112.22

> (1 + exp(-summary(fit)$coefficients[1]))^-1
[1] 0.76
> mean(y)
[1] 0.76
```
Interpretation of $\beta_1$
(Categorical Predictor)

\[ \text{logit}(\pi) = \beta_0 + \beta_1 \text{smoke} \]

- In SLR, $\beta_1$ is the difference in means between the target group and the reference group.
- In logistic regression, $\beta_1$ is the increase in the log-Odds that $Y = 1$, when moving from the reference group to the target group.
- $e^{\beta_1}$ is the **Odds Ratio (OR)**.
- Same result as the log-linear model.
  - $\beta_0$ still contains information about the sample proportion for the reference group.
Interpretation of $\beta_1$
(Categorical Predictor)

```r
> fit <- glm(low ~ smoke, family= binomial, data= lbw)
> summary(fit)
Call:
  glm(formula = low ~ smoke, family = binomial, data = lbw)

Coefficients:
            Estimate  Std. Error   z value  Pr(>|z|)
(Intercept)   -1.0871      0.2147    -5.062   4.14e-07
smokeYes      0.7041      0.3196     2.203    0.0276

> exp(summary(fit)$coefficients[2,1]) # OR
[1] 2.021944
> with(lbw, fisher.test(table(smoke, low))) # OR
sample estimates:
  odds ratio
    2.014137
```
Interpretation of $\beta_1$ (Numerical Predictor)

$$\text{logit}(\pi) = \beta_0 + \beta_1 \text{lwt}$$

- In SLR, $\beta_1$ is the increase in $Y$ for a unit change in $X$.
- In logistic regression, $\beta_1$ is the increase in the log-Odds that $Y = 1$, for a unit change in $X$.
- $e^{\beta_1}$ is the **Odds Ratio (OR)**

\[
\log(\text{OR}) = \log\left(\frac{\exp(\beta_0 + \beta_1(X + 1))}{\exp(\beta_0 + \beta_1X)}\right) = \log(e^{\beta_1})
\]
Interpretation of $\beta_1$ (Numerical Predictor)

> fit <- glm(low ~ lwt, family = binomial, data = lbw)
> summary(fit)

Call:
  glm(formula = low ~ lwt, family = binomial, data = lbw)

Coefficients:

|                | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 0.970969 | 0.780964   | 1.243   | 0.214    |
| lwt            | -0.013854| 0.006138   | -2.257  | 0.024    |

> exp(summary(fit)$coefficients[2,1]) # OR

[1] 0.9862412

- There is a 1.4% decrease in the odds of having a low birth weight baby, for each additional pound of mother’s weight.
Interpretation of $\beta$
(Mixed Predictors)

\[
\text{logit}(\pi) = \beta_0 + \beta_1 \text{lwt} + \beta_2 \text{smoke}
\]

- In SLR, $\beta_1$ is the increase in Y for a unit change in X, for each level of smoke
- $\beta_2$ is the difference between group means, controlling for mother’s last weight
- In logistic regression, $\beta_1$ is the increase in the log-Odds that $Y = 1$, for a unit change in X, for each level of smoke
- $\beta_2$ is the increase in the log-Odds that $Y = 1$, when moving from the reference group to the target group, controlling for mother’s last weight
Inference

- Build a CI for the coefficients (in the log-Odds scale) first, since the $\hat{\beta}^s$ are Normal
- Exponentiate the interval to get an interval for Odds
- Do not go further to get an interval for $\pi$
Interpretation of $\beta$ (Mixed Predictors)

```
> fit <- glm(low ~ lwt + smoke, family= binomial, data= lbw)
> exp(coef(fit))

(Intercept)     lwt   smokeYes
   1.8244464 0.9869026  1.9738381
```

```
> exp(confint(fit))
Waiting for profiling to be done...

           2.5 %     97.5 %
(Intercept) 0.4079558  9.248857
lwt           0.9744541    0.998051
smokeYes     1.0454986    3.745274
```

- Controlling for mother’s weight, we are 95% confident that smoking during pregnancy is associated with a 4.5% to 274.5% increase in the odds of having a low birth weight baby.
Visualize
Likelihood Ratio Tests

- Can be used to compare any two nested models
- Most powerful test
- In R,
  - `anova(reduced, full, test="LRT")`
- In SAS, compute:
  \[-2\{\log L_{\text{reduced}} - \log L_{\text{full}}\}\]
  - This is nicely displayed as output for you
Model Assumptions

- Independent observations
- Correct form of model
  - Linearity between logits & predictor variables
  - All relevant predictors included
  - All irrelevant predictors excluded, or multicollinearity
    - Unstable coefficients
    - High SEs
    - $\beta^s$ hard to interpret
    - Large p-values for important predictors
  - Need large samples for tests & CIs to be valid
Confounding

- Appleton et al. (1996) studied a cohort of 1314 UK women for 20 years, and found that smoking appeared to reduce mortality!
  - OR = 0.68; 95% CI (0.53, 0.88)
- After adjusting for age
  - OR = 1.5; 95% CI (1.1, 2.2)
- Younger women were smoking more than older women … oops
- Simpson’s paradox
Logistic Formula

- There are a number of ways to use `glm()` to do a logistic regression.
- The model formula is $y \sim X\beta$ and $y$ can be:
  - A Boolean vector
  - A factor
  - A 2-column matrix with Yes/No counts
- Demonstrate the latter
Smoking Dataset

```r
> library(SMPracticals)
> data(smoking)
> smoking

<table>
<thead>
<tr>
<th>age</th>
<th>smoker</th>
<th>alive</th>
<th>dead</th>
</tr>
</thead>
<tbody>
<tr>
<td>18-24</td>
<td>1</td>
<td>53</td>
<td>2</td>
</tr>
<tr>
<td>18-24</td>
<td>0</td>
<td>61</td>
<td>1</td>
</tr>
<tr>
<td>25-34</td>
<td>1</td>
<td>121</td>
<td>3</td>
</tr>
<tr>
<td>25-34</td>
<td>0</td>
<td>152</td>
<td>5</td>
</tr>
<tr>
<td>35-44</td>
<td>1</td>
<td>95</td>
<td>14</td>
</tr>
<tr>
<td>35-44</td>
<td>0</td>
<td>114</td>
<td>7</td>
</tr>
<tr>
<td>45-54</td>
<td>1</td>
<td>103</td>
<td>27</td>
</tr>
<tr>
<td>45-54</td>
<td>0</td>
<td>66</td>
<td>12</td>
</tr>
<tr>
<td>55-64</td>
<td>1</td>
<td>64</td>
<td>51</td>
</tr>
<tr>
<td>55-64</td>
<td>0</td>
<td>81</td>
<td>40</td>
</tr>
<tr>
<td>65-74</td>
<td>1</td>
<td>7</td>
<td>29</td>
</tr>
<tr>
<td>65-74</td>
<td>0</td>
<td>28</td>
<td>101</td>
</tr>
<tr>
<td>75+</td>
<td>1</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>75+</td>
<td>0</td>
<td>0</td>
<td>64</td>
</tr>
</tbody>
</table>
```
Smoking Dataset

```r
> fit <- glm(cbind(dead, alive) ~ smoker, data= smoking, family= binomial)
> summary(fit)
Coefficients:
    Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.78052    0.07962  -9.803  < 2e-16
smoker       -0.37858    0.12566  -3.013  0.00259

Null deviance: 641.5  on 13  degrees of freedom
Residual deviance: 632.3  on 12  degrees of freedom
AIC: 683.29

> exp(coef(fit)[2])
smoker
0.6848366
> exp(confint(fit))

   2.5 %     97.5 %
(Intercept) 0.3913417  0.5347849
smoker      0.5345661  0.8750872
```
Smoking Dataset

```r
> fit1 <- update(fit, ~ age + .)
> summary(fit1)

Coefficients:

|                | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -3.8601  | 0.5939     | -6.500  | 8.05e-11 |
| (...)          |          |            |         |          |
| smoker         | 0.4274   | 0.1770     | 2.414   | 0.015762 |

Null deviance: 641.4963 on 13 degrees of freedom
Residual deviance: 2.3809 on 6 degrees of freedom
AIC: 65.377

> exp(coef(fit1)[8])

smoker
1.533275

> exp(confint(fit1))

2.5 %        97.5 %
smoker 1.086868e+00 2.17732199
```
2008 US Election

- Now that we’re equipped, say something about the type of people who might vote for Obama

<table>
<thead>
<tr>
<th></th>
<th>B Coefficient</th>
<th>Standard Error</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.397</td>
<td>.375</td>
<td>4.04</td>
</tr>
<tr>
<td>Female (1=yes, 0=no)</td>
<td>-.082</td>
<td>.087</td>
<td>.922</td>
</tr>
<tr>
<td>Age (in years)</td>
<td>-.004</td>
<td>.003</td>
<td>.996</td>
</tr>
<tr>
<td>Black (1=yes, 0=no)</td>
<td>2.859</td>
<td>.255</td>
<td>17.445</td>
</tr>
<tr>
<td>Hispanic (1=yes, 0=no)</td>
<td>1.043</td>
<td>.154</td>
<td>2.838</td>
</tr>
<tr>
<td>Education (in years)</td>
<td>.102</td>
<td>.021</td>
<td>1.107</td>
</tr>
<tr>
<td>Household income (in thousands)</td>
<td>-.001</td>
<td>.001</td>
<td>.999</td>
</tr>
<tr>
<td>Republican (1=yes, 0=no)</td>
<td>-1.988</td>
<td>.117</td>
<td>.137</td>
</tr>
<tr>
<td>Democrat (1=yes, 0=no)</td>
<td>1.728</td>
<td>.101</td>
<td>5.632</td>
</tr>
<tr>
<td>Ideology (1=very liberal to 5=very conservative)</td>
<td>-.896</td>
<td>.045</td>
<td>.408</td>
</tr>
<tr>
<td>Number of days saw presidential campaign information on TV news in past week</td>
<td>-.009</td>
<td>.019</td>
<td>.991</td>
</tr>
<tr>
<td>Number of days heard about presidential campaign on talk radio in past week</td>
<td>-.045</td>
<td>.016</td>
<td>.956</td>
</tr>
</tbody>
</table>
Problems in Linear Regression

- Extrapolation
  - Shouldn’t make inference outside range of observed predictor variables; model may not be correct

- Multicollinearity
  - Unstable fitted equation
  - Large S.E.s for coefficients
  - MLEs may not converge

- Influential points
  - One removal may substantially change estimates

- Model building
  - Overfit to data
Problems in Logistic only

- Extra-binomial variation
  - Variance of $Y_i$ greater than $\pi_i(1 - \pi_i)$
  - Also called “Over-dispersion”
  - Doesn’t bias estimates, but S.E. smaller than it should be
  - Fix: add a dispersion parameter

- Complete (and quasi-complete) separation
  - Predictors perfectly predict whether $Y=1$ or $Y=0$
    - Cannot compute MLEs
  - Fix: simplify model, use penalized ML, or use a Bayesian analysis
Separation

```r
> fit <- glm(low ~ bwt, family= binomial, data= lbw)

Warning messages:
  1: glm.fit: algorithm did not converge
  2: glm.fit: fitted probabilities numerically 0 or 1 occurred

> with(lbw, plot(bwt, low))
```
Separation

- Especially a problem once you start including many categorical predictors with many levels
  - And it gets worse with interactions
- Need many observations at every combination of levels or MLEs won’t converge
- Moral: Don’t go building logistic regression models with the kitchen sink
  - Not that you should do that in linear regression either
Model Building with LBW data

- **Data description:**
  - low: indicator of birth weight less than 2.5kg
  - age: mother's age in years
  - lwt: mother's weight (lbs) at last menstrual period
  - race: mother's race ("white", "black", "other")
  - smoke: smoking status during pregnancy
  - ht: history of hypertension
  - ui: presence of uterine irritability
  - ftv: number of physician visits during first trimester
  - ptl: number of previous premature labours
  - bwt: birth weight in grams
Model Building

```r
> fit <- glm(low ~ (age + lwt + race + smoke + ht + ui + ftv + ptl)^2, family= binomial, data= lbw)
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> with(lbw, table(race, smoke, ht, ui))

, , ht = No, ui = No

<table>
<thead>
<tr>
<th></th>
<th>smoke</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>race</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>black</td>
<td>11</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>other</td>
<td>43</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>white</td>
<td>39</td>
<td>39</td>
<td></td>
</tr>
</tbody>
</table>

, , ht = Yes, ui = No

<table>
<thead>
<tr>
<th></th>
<th>smoke</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>race</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>black</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>other</td>
<td>4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>white</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

, , ht = Yes, ui = Yes

<table>
<thead>
<tr>
<th></th>
<th>smoke</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>race</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>black</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>other</td>
<td>8</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>white</td>
<td>4</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>
```

42
Model Building

```r
> summary(lbw)

        low   age     lwt   race  smoke   ht   ui   ftv   ptl  bwt  race2
No :130 Min. :14.00 Min. :80.0 black:26 No :115 No :177 No :161 No :100 No :159 Min. :709 other:93
Yes: 59  1st Qu.:19.00 1st Qu.:110.0 other:67 Yes: 74  Yes: 12  Yes: 28  Yes: 89  Yes: 30 1st Qu.:2414 white:96
Median :23.00 Median :121.0 white:96
Mean   :23.24 Mean   :129.7
3rd Qu.:26.00 3rd Qu.:140.0
Max.   :45.00 Max.   :250.0

> # Drop 'ht' and 'ui' due to low counts
> fitF <- glm(low ~ (age + lwt + race + smoke + ftv + ptl)^2, family= binomial, data= lbw)
> fitR <- glm(low ~ age + lwt + race + smoke + ftv + ptl, family= binomial, data= lbw)
> anova(fitR, fitF, test="LRT")

Analysis of Deviance Table

<table>
<thead>
<tr>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Df</th>
<th>Deviance</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>181</td>
<td></td>
<td>205.31</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>161</td>
<td>20</td>
<td>168.97</td>
<td>0.6958</td>
</tr>
</tbody>
</table>
```

Model 1: low ~ age + lwt + race + smoke + ftv + ptl
Model 2: low ~ (age + lwt + race + smoke + ftv + ptl)^2

Resid. Df Resid. Dev Df Deviance Pr(>Chi)
Model Building

> summary(fitR)
Coefficients:

     Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.618120   1.210892   1.336  0.18145
   age        -0.039827   0.037429  -1.064  0.28730
   lwt        -0.010331   0.006532  -1.582  0.11375
raceother   -0.352861   0.542815  -0.650  0.51565
racewhite   -1.149563   0.525935  -2.186  0.02883
 smokeYes    0.812987   0.406118   2.002  0.04530
ftvYes     -0.218199   0.363184  -0.601  0.54798
ptlYes      1.348631   0.451606   2.986  0.00282

> fitRR <- glm(low ~ lwt + race2 + smoke + ptl, family= binomial, data= lbw)
> anova(fitRR, fitR, test="LRT")
Analysis of Deviance Table

Model 1: low ~ lwt + race2 + smoke + ptl
Model 2: low ~ age + lwt + race + smoke + ftv + ptl

                     Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1                      184     207.64
2                      181     205.31  3     2.331  0.5066
Model Building

> summary(fitRR)
Call:
glm(formula = low ~ lwt + race2 + smoke + ptl, family = binomial, data = lbw)

Coefficients:

|            | Estimate | Std. Error | z value | Pr(>|z|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 0.418928 | 0.805964   | 0.520   | 0.60321  |
| lwt        | -0.010527| 0.006097   | -1.727  | 0.08424  |
| race2white | -1.026199| 0.382381   | -2.684  | 0.00728  |
| smokeYes   | 0.932813 | 0.385106   | 2.422   | 0.01543  |
| ptlYes     | 1.211286 | 0.435598   | 2.781   | 0.00542  |

Null deviance: 234.67 on 188 degrees of freedom
Residual deviance: 207.64 on 184 degrees of freedom
AIC: 217.64
Logistic Regression as Classification

- Have a set of predictors \( \{x_1^*, ..., x_p^*\} \) to predict \( y^* \)
- Fit a logistic model and pick a cut point
  - Like \( \hat{\pi}^* = 0.5 \)
- If \( \hat{\pi}^* > 0.5 \), predict \( y^* = 1 \)
- If \( \hat{\pi}^* < 0.5 \), predict \( y^* = 0 \)
  - You can use a cut point other than 0.5 too
- There are many other classifiers in use
LBW example

```r
> p_cutoff <- 0.5
> lbw$p <- fitRR$fitted.values
> lbw$pred <- ifelse(fitRR$fitted.values > p_cutoff, "Yes", "No")
> lbw$I.low <- ifelse(lbw$low == "Yes", 1, 0)
> with(lbw, table(pred, low))

    low
pred  No Yes
    No  117 37
   Yes  13 22

> require(caret)
> with(lbw, confusionMatrix(pred, low))

Confusion Matrix and Statistics

             Reference
Prediction      No Yes
    No  117 37
   Yes  13 22
```
LBW example

Accuracy : 0.7354
95% CI : (0.6665, 0.7968)
No Information Rate : 0.6878
P-Value [Acc > NIR] : 0.089649

Kappa : 0.307
Mcnemar's Test P-Value : 0.001143

Sensitivity : 0.9000
Specificity : 0.3729
Pos Pred Value : 0.7597
Neg Pred Value : 0.6286
Prevalence : 0.6878
Detection Rate : 0.6190
Detection Prevalence : 0.8148
Balanced Accuracy : 0.6364

'Positive' Class : No
LBW example

- That’s only good for one cutoff probability (0.5)
  - We would have to generate the table for each possible cutoff, which would be a lot of tables!
- Use a method that considers all at once
- Focus on sensitivity and specificity metrics
Sensitivity

- True Positive Rate (TPR)

\[ Sensitivity = \frac{\text{# of true positives}}{\text{# who have the disease}} \]

\[ = \frac{TP}{TP + FN} \]

\[ = P(+|D) \]

- aka Recall in ML
Specificity

- True Negative Rate (TNR)

\[
Sensitivity = \frac{\text{# of true negatives}}{\text{# who do not have the disease}}
\]

\[
= \frac{TN}{TN + FP}
\]

\[
= P(-|D^c)
\]
# Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>+</th>
<th>–</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>TP</td>
<td>FN</td>
<td>Sensitivity = TP / (TP + FN)</td>
</tr>
<tr>
<td>Dc</td>
<td>FP</td>
<td>TN</td>
<td>Specificity = TN / (TN + FP)</td>
</tr>
<tr>
<td></td>
<td>PPV = TP / (TP + FP)</td>
<td>NPV = TN / (TN + FN)</td>
<td></td>
</tr>
</tbody>
</table>
A negative result on a very sensitive test “rules out” the disease
- High sensitivity makes a good screening test

A positive result on a very specific test “rules in” the disease
- High specificity makes a good confirmatory test

A screening test followed by a confirmatory test is a good (albeit expensive) diagnostic procedure
ROC Curves

- Receiver Operating Characteristic
- A plot of sensitivity vs. specificity (complement)
- Originally designed to grade radar detection methods for German planes
- Decades later, their usefulness in classification problems was realized
  - But the name stuck
ROC Curve
(LBW example)

```
require(pROC)
lbw.roc <- with(lbw, roc(I.low, p, percent=T, auc=T, plot=T, auc.polygon=T, max.auc.polygon=T, print.auc=T, main= "ROC curve"))
lbw.roc$thresholds[order(-lbw.roc$sensitivities*lbw.roc$specificities)[1]]
```

```
[1] 0.3243768
```
Over-fitting

- Just because $\hat{\pi}^* = 0.324$ gave optimal results for this dataset, it doesn’t mean it will be the best cut point on future measurements

- **Cross-validation** is a more reliable technique
  - Train model on your *training set*
  - Validate model (find optimal cut point) on your *validation set*
  - You didn’t collect two datasets?
  - Just divide it into some fraction ($\frac{1}{2}$ and $\frac{1}{5}$ are my favourites)
Cross-validation
(LBW example)

set.seed(2015)
# Randomly divide the set in two
ind.trn <- sample(nrow(lbw), floor(nrow(lbw)/2))
lbw.trn <- lbw[ind.trn,] # Training set
lbw.val <- lbw[setdiff(1:nrow(lbw), ind.trn),] # Validation set

# Fit to training data
fitRR <- glm(low ~ lwt + race2 + smoke + ptl, family= binomial, data= lbw.trn)

# Predict l-Odds on validation set
lbw.val$lOdds <- predict(fitRR, subset(lbw.val, select=c(lwt,race2,smoke,ptl)))

# Predicted outcome for validation set
lbw.val$p <- (1 + exp(lbw.val$lOdds)^-1)^-1
Cross-validation (LBW example)

> lbw.roc <- with(lbw.val, roc(I.low, p, percent=T, auc=T, plot=T, auc.polygon=T, max.auc.polygon=T, print.auc=T, main= "ROC curve"))

> lbw.roc$thresholds[order(-lbw.roc$sensitivities*lbw.roc$specificities)[1]]
[1] 0.2033337
Ordinal Response

- If the response is not binary but ordinal, can use an **ordered logit** model instead
- Makes a proportional odds assumption, which we will discuss next week as well
  - The odds ratio does not change when moving up the ordinal scale, regardless of the step
- Not here
Categorical Response

- If the response is categorical and not binary, can use a **multinomial logit** model instead
  - Good luck with that
- Makes an assumption of **independence of irrelevant alternatives** which you may not want
  - And in any case, is often not true
    - See work by A. Tversky & D. Kahneman
    - “Losses loom larger than gains”
  - Also known as **binary independence**
- Most analysts I know just fit separate log. reg. models in this situation
IIA axiom

- If you prefer alternative A to B when they are the only two choices, introducing a third choice C should not make B preferable to A
- “Do you want an in-class exam or a take-home exam?”
  - “I’ll do the take-home exam please”
- “You could also do a project instead”
  - “In that case I’ll do an in-class exam”
- “... ?”