432 Class 8 Slides

github.com/THOMASELOVE/432-2018

2018-02-08

Setup

```
library(skimr)
library(broom)
library(Hmisc)
library(rms)
library(tidyverse)
```

Today's Materials

- Logistic Regression and the Low Birth Weight data

```
lbw1 <- read.csv("data/lbw.csv") %>% tbl_df

lbw1 <- lbw1 %>%
    mutate(race_f = fct_recode(factor(race), white = "1",
                               black = "2", other = "3"),
         race_f = fct_relevel(race_f, "white", "black")) %>%
    mutate(preterm = fct_recode(factor(ptl > 0),
                                yes = "TRUE",
                                no = "FALSE")) %>%
    select(subject, low, lwt, age, ftv, ht, race_f,
           preterm, smoke, ui)
```

The `lbw1` data (n = 189 infants)

Variable	Description
`subject`	id code
`low`	indicator of low birth weight (< 2500 g)
`lwt`	mom's weight at last menstrual period (lbs.)
`age`	age of mother in years
`ftv`	count of physician visits in first trimester (0 to 6)
`ht`	history of hypertension: $1 =$ yes, $0 =$ no
`race_f`	race of mom: white, black, other
`preterm`	prior premature labor: $1 =$ yes, $0 =$ no
`smoke`	$1 =$ smoked during pregnancy, $0 =$ did not
`ui`	presence of uterine irritability: $1 =$ yes, $0 =$ no

Source: Hosmer, Lemeshow and Sturdivant, *Applied Logistic Regression* 3rd edition. Data from Baystate Medical Center, Springfield MA in 1986.

Goals for Today and Tuesday

1. Fit and evaluate the fit of a logistic regression model to predict the probability of a low birth weight (`low` = 1) using the mom's weight at her last menstrual period (`lwt`).
2. Fit and evaluate a larger logistic regression model to predict `low` on the basis of a larger group of predictors drawn from the available options, which include: `lwt`, `age`, `ftv`, `ht`, `race_f`, `preterm`, `smoke` and `ui`.
3. Learn about the use of both `glm` and `lrm` (from the `rms` package) to fit and evaluate logistic regression models.

EDA for Task 1

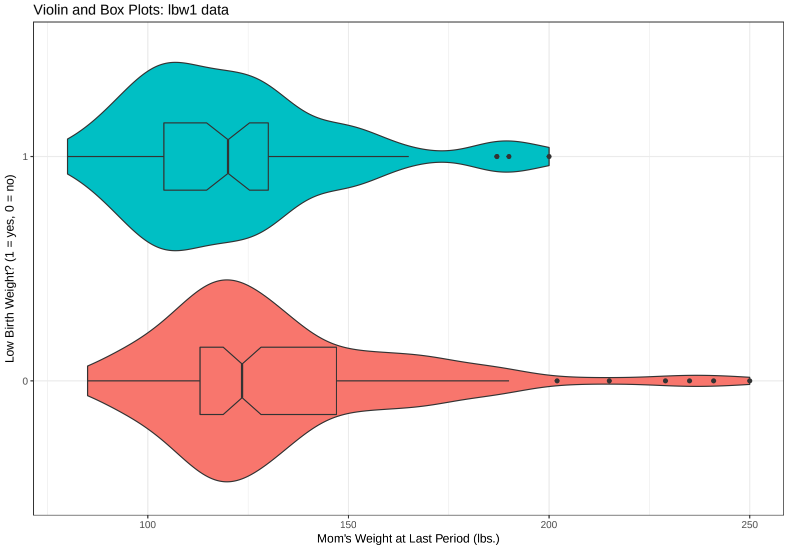
We want to look at the probability of a low birth weight (`low` = 1) on the basis of the mom's weight at her last menstrual period (`lwt`).

```
lbw1 %>% group_by(low) %>% skim(lwt)
```

```
> lbw1 %>% group_by(low) %>% skim(lwt)
Skim summary statistics
 n obs: 189
 n variables: 10
 group variables: low

Variable type: integer
 low variable missing complete   n   mean     sd p0 p25 median p75 p100     hist
   0      lwt       0      130 130 133.3  31.72 85 113  123.5 147  250
   1      lwt       0       59  59 122.14 26.56 80 104  120   130  200
```

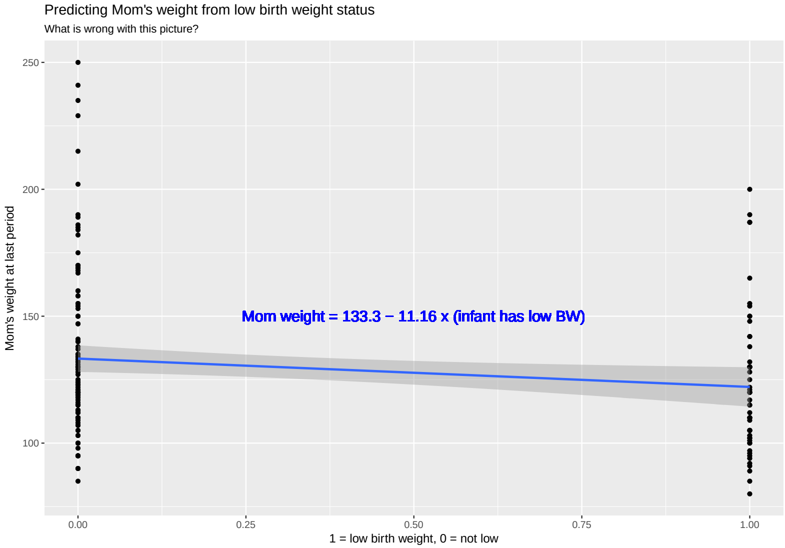
Can we predict Pr(low) effectively with `lwt`?



Code for Previous Slide

```
ggplot(lbw1, aes(x = factor(low), y = lwt,
                 fill = factor(low))) +
    geom_violin() +
    geom_boxplot(width = .3, notch = TRUE) +
    guides(fill = FALSE) +
    labs(x = "Low Birth Weight? (1 = yes, 0 = no)",
         y = "Mom's Weight at Last Period (lbs.)",
         title = "Violin and Box Plots: lbw1 data") +
    theme_bw() +
    coord_flip()
```

Working in Reverse: Can we predict `lwt` with `low`?



Working in Reverse: Predicting `lwt` with `low`

Easy to go in the other direction...

```
lm(lwt ~ low, data = lbw1)
```

```
Call:
lm(formula = lwt ~ low, data = lbw1)

Coefficients:
(Intercept)          low
     133.30       -11.16
```

Weight at Last Period = 133.3 - 11.16 * (baby is low bw)

- But that's reversing the outcome and predictor...

Can we fit a linear probability model? Sure, but . . .

```r
lm(low ~ lwt, data = lbw1)
```

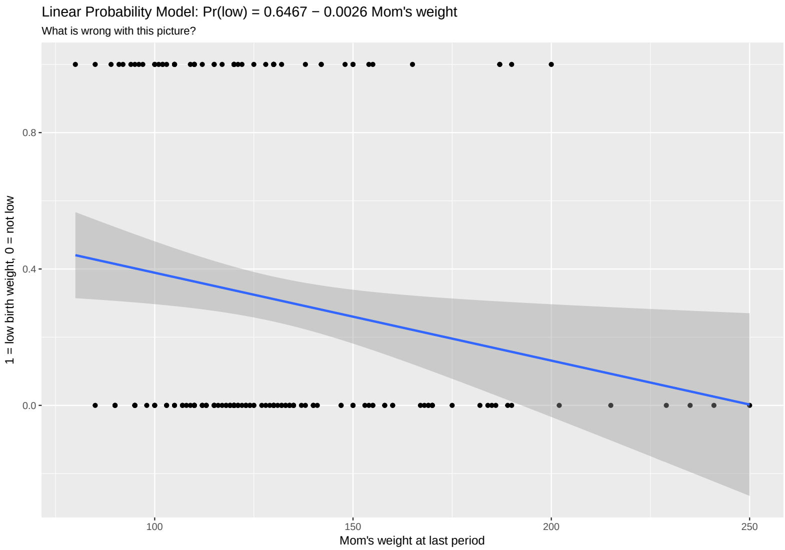
```

Call:
lm(formula = low ~ lwt, data = lbw1)

Coefficients:
(Intercept)          lwt
   0.646733    -0.002577
```

Pr(low birth weight) = 0.6467 - 0.0026 (Mom's weight at last period)

Plotting the Linear Probability Model



Fitting a Model to predict a Binary Outcome

Logistic regression is the most common model used when the outcome is binary. Our response variable is assumed to take on two values - zero or one, and we then describe the probability of a "one" response, given a linear function of explanatory predictors.

- Linear regression approaches to the problem of predicting probabilities are problematic for several reasons: not least of which being that they predict probabilities greater than one and less than zero.

Logistic regression is a non-linear regression approach, since the equation for the mean of the $0/1$ Y values conditioned on the values of our predictors $X_1, X_2, ..., X_k$ turns out to be non-linear in the β coefficients.

The Logit Link and Logistic Function

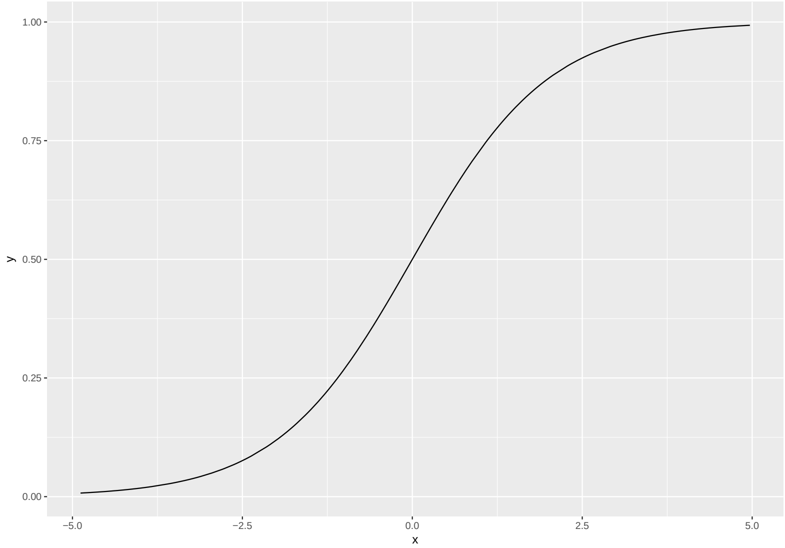
The particular link function we use in logistic regression is called the **logit link**.

$$logit(\pi) = log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

The inverse of the logit function is called the **logistic function**. If logit(π) $= \eta$, then $\pi = \frac{exp(\eta)}{1+exp(\eta)}$.

- The logistic function $\frac{e^x}{1+e^x}$ takes any value x in the real numbers and returns a value between 0 and 1.

The Logistic Function $y = \frac{e^x}{1+e^x}$



The logit or log odds

We usually focus on the **logit** in statistical work, which is the inverse of the logistic function.

- If we have a probability $\pi < 0.5$, then $logit(\pi) < 0$.
- If our probability $\pi > 0.5$, then $logit(\pi) > 0$.
- Finally, if $\pi = 0.5$, then $logit(\pi) = 0$.

Model 1

We'll use `glm` to get started.

```
model.1 <- glm(low ~ lwt, data = lbw1, family = binomial)
model.1
```

```
Call:  glm(formula = low ~ lwt, family = binomial, data = lbw1

Coefficients:
(Intercept)          lwt
    0.99831     -0.01406

Degrees of Freedom: 188 Total (i.e. Null);  187 Residual
Null Deviance:      234.7
Residual Deviance: 228.7    AIC: 232.7
```

Our logistic regression model

The logistic regression equation is:

$$logit(Pr(low = 1)) = log\left(\frac{Pr(low = 1)}{1 - Pr(low = 1)}\right) = 0.99831 - 0.01406 \times lwt$$

Suppose, for instance, that we are interested in making a prediction when Mom's weight at her last period, `lwt` = 130 lbs.

So we have:

$$logit(Pr(low = 1)) = 0.99831 - 0.01406x130 = -0.82949$$

Getting a Prediction from R for the Model

```
model.1 <- glm(low ~ lwt, data = lbw1, family = binomial)
```

To predict on the log odds scale, we use

```
predict(model.1, newdata = data.frame(lwt = 130))
```

```
         1
-0.8292596
```

The Model in terms of Odds

We can exponentiate to state the odds, rather than the log odds. For a Mom at 130 lbs, we have:

$$log\left(\frac{Pr(low = 1)}{1 - Pr(low = 1)}\right) = 0.99831 - 0.01406 \times 130 = -0.82949$$

and so we have

$$Odds(low = 1 | lwt = 130) = exp(-0.82949) = 0.4362717$$

Making a Prediction about Probability

$$Odds(low = 1|lwt = 130) = \frac{Pr(low = 1)}{1 - Pr(low = 1)} = 0.4362717$$

so

$$Pr(low = 1|lwt = 130) = \frac{Odds(low = 1|lwt = 130)}{1 + Odds(low = 1|lwt = 130)} = \frac{0.4362717}{1 + 0.4362717}$$

which is 0.304.

Obtaining a Prediction from R for Prob(low = 1)

```
model.1 <- glm(low ~ lwt, data = lbw1, family = binomial)
```

To predict on the probability scale, we can use

```
predict(model.1, newdata = data.frame(lwt = 130),
        type = "response")
```

```
        1
0.3038016
```

Plotting the Logistic Regression Model

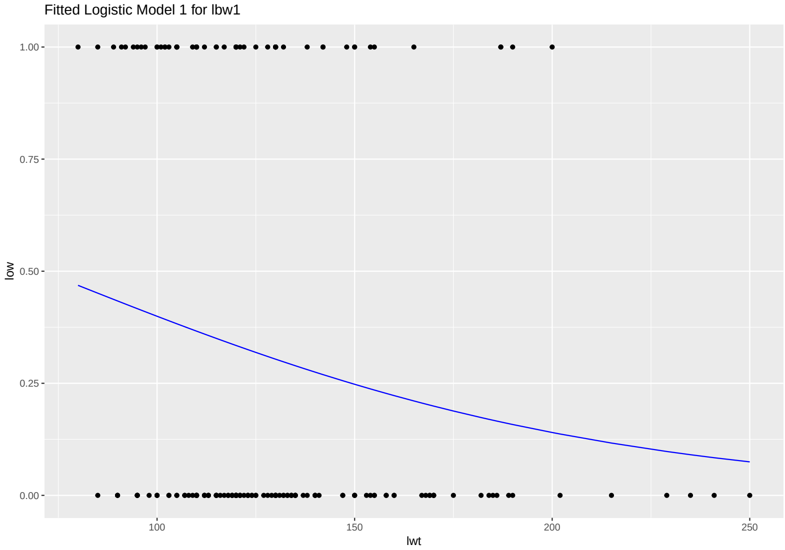
We can use the augment function from the broom package to get our fitted probabilities included in the data.

```
mod1.aug <- augment(model.1, lbw1,
                    type.predict = "response")

ggplot(mod1.aug, aes(x = lwt, y = low)) +
    geom_point() +
    geom_line(aes(x = lwt, y = .fitted), col = "blue") +
    labs(title = "Fitted Logistic Model 1 for lbw1")
```

- Results on next slide

Plotting the Logistic Regression Model



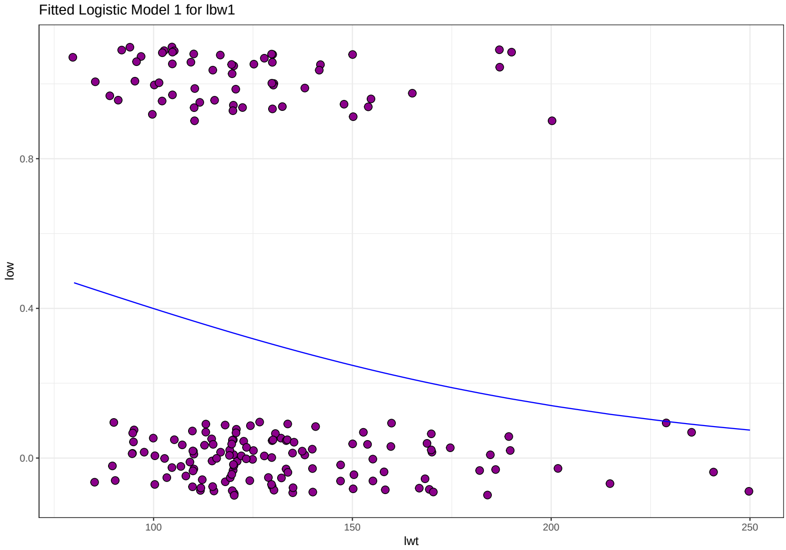
Cleaning up the plot

I'll add a little jitter on the vertical scale to the points, so we can avoid overlap, and also make the points a little bigger.

```
ggplot(mod1.aug, aes(x = lwt, y = low)) +
    geom_jitter(height = 0.1, size = 3, pch = 21,
                fill = "darkmagenta") +
    geom_line(aes(x = lwt, y = .fitted), col = "blue") +
    labs(title = "Fitted Logistic Model 1 for lbw1") +
    theme_bw()
```

- Results on next slide

Cleaned up Plot of Model 1



Studying the Model, Again

```
model.1
```

```

Call:  glm(formula = low ~ lwt, family = binomial, data = lbw1

Coefficients:
(Intercept)          lwt
    0.99831     -0.01406

Degrees of Freedom: 188 Total (i.e. Null);  187 Residual
Null Deviance:      234.7
Residual Deviance: 228.7    AIC: 232.7
```

- logit(Pr(low $= 1$)) $=$ 0.998 - 0.014 lwt
  - so ... as lwt increases, what happens to Pr(low $= 1$)?
  - if Harry's mom weighed 130 lbs and Sally's weighed 150 lbs, how can we compare the predicted Pr(low $= 1$) for Harry and Sally?

Comparing Harry (lwt = 130) to Sally (lwt = 150)

```
predict(model.1, newdata = data.frame(lwt = c(130, 150)),
        type = "response")
```

```
        1         2 
0.3038016 0.2477917 
```

- Harry's mom weighed 130 lbs, and his predicted probability of low birth weight is 0.304
- Sally's mom weighed 150 lbs, and her predicted $\Pr(\text{low} = 1) = 0.248$

Interpreting the Coefficients of the Model

```
coef(model.1)
```

```
(Intercept)         lwt
 0.99831432 -0.01405826
```

To understand the effect of `lwt` on `low`, try odds ratios.

```
exp(coef(model.1))
```

```
(Intercept)         lwt
  2.7137035   0.9860401
```

Suppose Charlie's Mom weighed one pound more than Harry's.

- The **odds** of low birth weight are 0.986 times as large for Charlie as Harry.
- In general, odds ratio comparing two subjects whose `lwt` differ by 1 pound is 0.986

Comparing Harry to Charlie

Charlie's mom weighed 1 pound more than Harry's. The estimated odds ratio for low birth weight from the model associated with a one pound increase in `lwt` is 0.986.

- If the odds ratio was 1, that would mean that Harry and Charlie had the same estimated odds of low birth weight, and thus the same estimated probability of low birth weight, despite having Moms with different weights.
- Since the odds ratio is less than 1, it means that Harry has a lower estimated odds of low birth weight than Charlie, and thus that Harry has a lower estimated probability of low birth weight than Charlie.
- If the odds ratio was greater than 1, it would mean that Harry had a higher estimated odds of low birth weight than Charlie, and thus that Harry had a higher estimated probability of low birth weight than Charlie.

The smallest possible odds ratio is ... ?

The rest of the model's output

```
Degrees of Freedom: 188 Total (i.e. Null);  187 Residual
Null Deviance:      234.7
Residual Deviance: 228.7    AIC: 232.7
```

Model	Null	Residual	Δ (`model.1`)
Deviance (lack of fit)	234.7	228.7	6.0
Degrees of Freedom	188	187	1

- Deviance accounted for by `model.1` is 6 points on 1 df
- Can compare to a χ^2 distribution for a p value via `anova`

AIC = 232.7, still useful for comparing models for the same outcome

anova on a glm model

```r
anova(model.1)
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: low

Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev
NULL                   188     234.67
lwt   1   5.9813       187     228.69
```

```r
pchisq(5.9813, 1, lower.tail = FALSE)
```

```
[1] 0.01445834
```

glance on model.1

```
glance(model.1)
```

```
  null.deviance df.null    logLik      AIC      BIC
1       234.672     188 -114.3453 232.6907 239.1742
  deviance df.residual
1 228.6907         187
```

- Deviance $= -2\times \log(\text{likelihood})$
- AIC and BIC are based on the deviance, but with differing penalties for complicating the model
- AIC and BIC remain useful for comparing multiple models for the same outcome

summary of model.1

```
> summary(model.1)

Call:
glm(formula = low ~ lwt, family = binomial, data = lbw1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0951  -0.9022  -0.8018   1.3609   1.9821

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.99831    0.78529   1.271   0.2036
lwt         -0.01406    0.00617  -2.279   0.0227 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 228.69  on 187  degrees of freedom
AIC: 232.69

Number of Fisher Scoring iterations: 4
```

Coefficients output

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.99831    0.78529   1.271   0.2036
lwt         -0.01406    0.00617  -2.279   0.0227 *
```

- We have a table of coefficients with standard errors, and hypothesis tests, although these are Wald z-tests, rather than the t tests we saw in linear modeling.
- `lwt` has a Wald Z of -2.279, yielding $p = 0.0227$
  - H_0: `lwt` does not have an effect on the log odds of `low`
  - H_A: `lwt` does have such an effect
- If the coefficient (on the logit scale) for `lwt` was truly 0, this would mean that:
  - the log odds of low birth weight did not change based on `lwt`,
  - the odds of low birth weight were unchanged based on `lwt` (OR = 1), and
  - the probability of low birth weight was unchanged based on the `lwt`.

Confidence Intervals for Coefficients

```r
coef(model.1)
```

```
(Intercept)         lwt 
 0.99831432 -0.01405826 
```

```r
confint(model.1, level = 0.95)
```

```
Waiting for profiling to be done...

                  2.5 %       97.5 %
(Intercept) -0.48116701  2.611748138
lwt         -0.02696198 -0.002650036
```

- The coefficient of `lwt` has a point estimate of -0.014 and a 95% confidence interval of (-0.027, -0.003).
- On the logit scale, this isn't that interpretable, but we will often exponentiate to describe odds ratios.

Odds Ratio Interpretation of exp(Coefficient)

```
exp(coef(model.1))
```

```
(Intercept)         lwt
  2.7137035   0.9860401
```

```
exp(confint(model.1, level = 0.95))
```

```
                2.5 %     97.5 %
(Intercept) 0.6180617 13.6228447
lwt         0.9733982  0.9973535
```

- Odds Ratio for `low` based on a one pound increase in `lwt` is 0.986 (95% CI: 0.973, 0,997).
  - Estimated odds of low birth weight will be smaller (odds < 1) for those with larger `lwt` values.
  - Smaller odds(low birth weight) $=$ smaller Prob(low birth weight).

Deviance Residuals

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0951  -0.9022  -0.8018   1.3609   1.9821
```

- The deviance residuals for each individual subject sum up to the deviance statistic for the model, and describe the contribution of each point to the model likelihood function. The formula is in the Course Notes.
- Logistic Regression is a non-linear model, and it doesn't come with either an assumption that the residuals will follow a Normal distribution, or an assumption that the residuals will have constant variance, so when we build diagnostics for the logistic regression model, we'll use different plots and strategies than we used in linear models.

Other New Things

```
(Dispersion parameter for binomial family taken to be 1)

Number of Fisher Scoring iterations: 4
```

- Dispersion parameters matter for some generalized linear models. For binomial family models like the logistic, it's always 1.
- The solution of a logistic regression model involves maximizing a likelihood function. Fisher's scoring algorithm needed just four iterations to perform this fit. The model converged, quickly.

How Well Does Our `model.1` Classify Subjects?

One possible rule: if predicted $\Pr(\text{low} = 1) \geq 0.5$, then we predict "low birth weight"

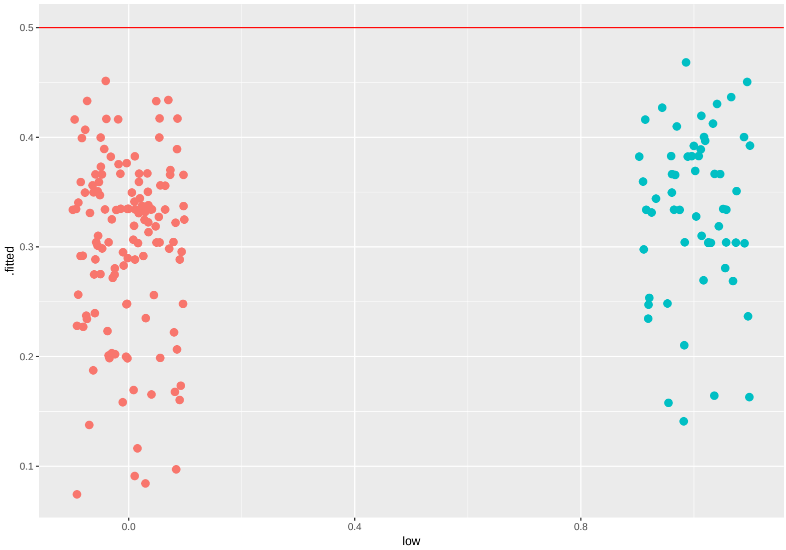
```
mod1.aug$rule.5 <- ifelse(mod1.aug$.fitted >= 0.5,
                        "Predict Low", "Predict Not Low")

table(mod1.aug$rule.5, mod1.aug$low)
```

```
                  0   1
  Predict Not Low 130  59
```

This rule might be a problem for us. What % are correct?

A plot of classifications with the 0.5 rule



How Well Does Our `model.1` Classify Subjects?

A new rule: if predicted $\Pr(\text{low} = 1) \geq 0.3$, then we predict "low birth weight"

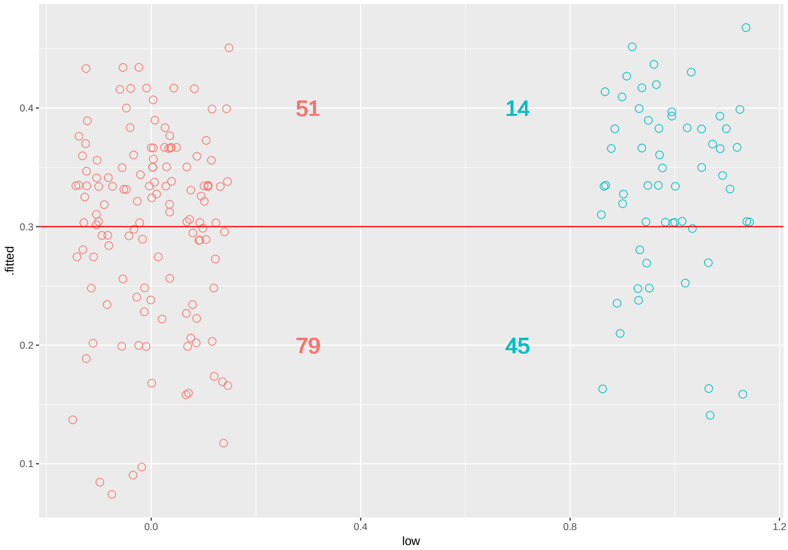
```
mod1.aug$rule.3 <- ifelse(mod1.aug$.fitted >= 0.3,
                          "Predict Low", "Predict Not Low")

table(mod1.aug$rule.3, mod1.aug$low)
```

```
                 0  1
  Predict Low     79 45
  Predict Not Low 51 14
```

What percentage of these classifications are correct?

A plot of classifications with the 0.3 rule



Coming Soon

- Receiver Operating Characteristic Curve Analysis
    - The C statistic (Area under the curve)
- Assessing Residual Plots for a Logistic Regression
- A "Kitchen Sink" Logistic Regression Model
    - Comparing Models
    - Interpreting Models with Multiple Predictors
- Logistic Regression using the `lrm` function
    - Nagelkerke R^2, Somers' d etc.
    - Validating Summary Statistics
    - Summaries of Effects
    - Plotting In-Sample Predictions
    - Influence
    - Calibration
    - Nomograms