Simple Logistic Regression – Part 2

SDS 291

April 8, 2020

Example: Female Gender as a Function of Height

```
Call:
glm(formula = Gender ~ Hgt, family = binomial, data = Pulse)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.77443  -0.34870  -0.05375   0.32973   2.37928

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  64.1416     8.3694   7.664 1.81e-14 ***
Hgt          -0.9424     0.1227  -7.680 1.60e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
logitmod=glm(Gender ~ Hgt, family = binomial, data=Pulse)
summary(logitmod)
```

Odds

Definition:

$$\frac{\pi}{1-\pi} = \frac{P(Yes)}{P(No)}$$ is the odds of Yes.

Try some odds for the Height/Gender data.

$$Odds = \frac{\pi}{1-\pi} \Leftrightarrow \pi = \frac{Odds}{1+Odds}$$

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  64.1416     8.3694   7.664 1.81e-14 ***
Hgt          -0.9424     0.1227  -7.680 1.60e-14 ***
---
```

Taller ← Shorter		
69”	**68”**	**67”**

Log(Odds)

Odds

π

Odds Ratio

A common way to compare two groups is to look at the *ratio* of their odds.

$$OR = \frac{Odds_{69}}{Odds_{68}} = \frac{0.4125}{1.0584} = 0.3896915$$

$$OR = \frac{Odds_{68}}{Odds_{67}} = \frac{1.0584}{2.7161} = 0.3896915$$

$$OR = e^{\beta_1} = e^{-0.9424} = 0.3896915$$

Every additional inch in height is associated with 0.39 times the odds of being female than a person 1 inch shorter, on average.

Conditions for Inference for Logistic Regression

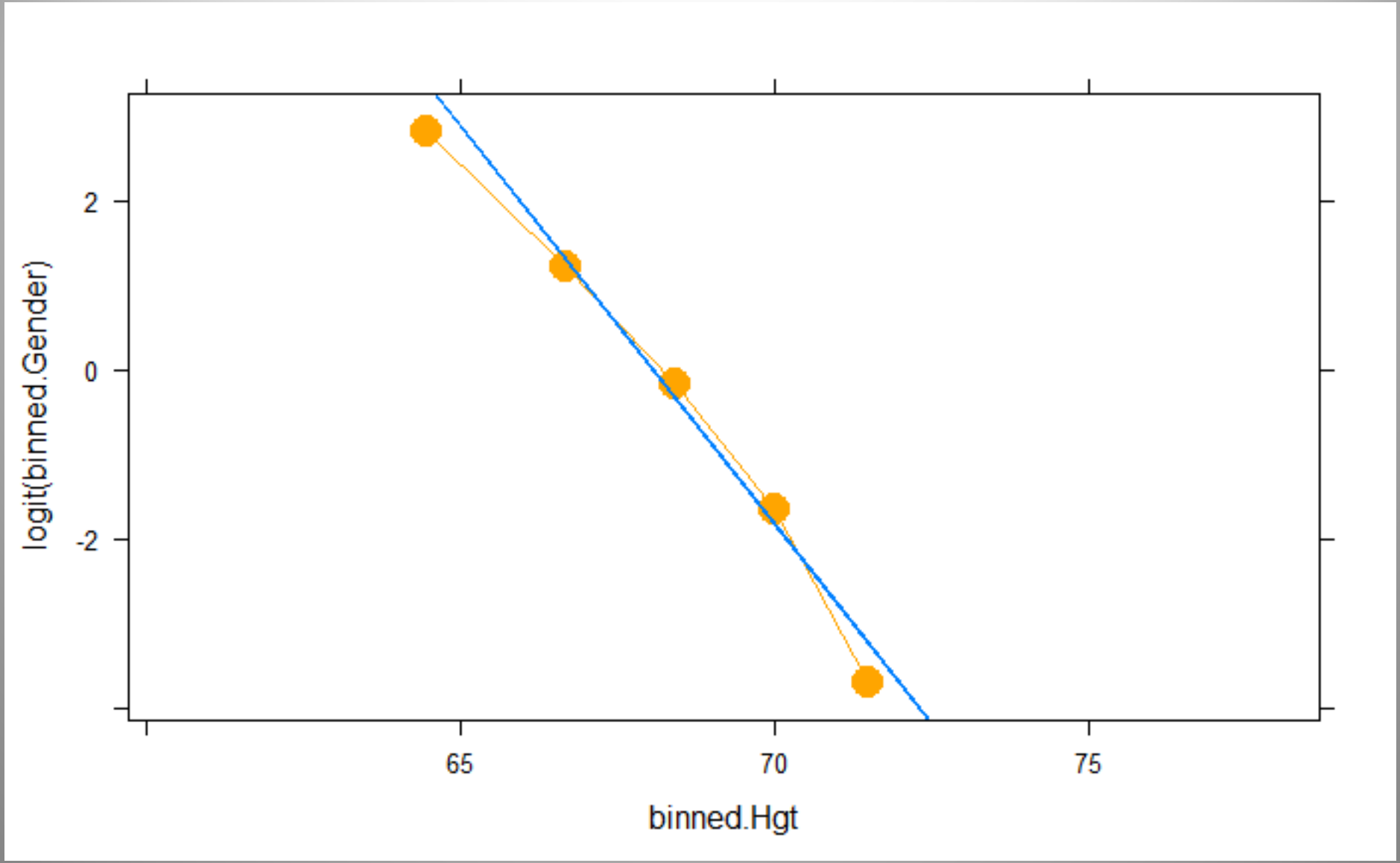
Linearity: The logits (log odds) should have a linear relationship with the predictor. [For binomial ("short form") data, we can check this with a plot.]

Independence: No pairing or clustering of data.

Random: Either a random sample from a population OR random assignment within an experiment.

~~**Normality**~~: This does not apply. The responses are 0/1.

~~**Constant variance**~~: Also does not apply. In fact, variability in Y is highest when π is near ½ and lowest when π is near 0 or 1.


logit(binned.Gender)
2
0
-2
65
70
75
binned.Hgt

Output for Gender Logistic

$$e^{-0.9424} = 0.3896915$$

Some sort of tests?

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  64.1416     8.3694   7.664 1.81e-14 ***
Hgt          -0.9424     0.1227  -7.680 1.60e-14 ***
---
```

Similar tests for logistic regression?

Recall: "Ordinary" Regression

```
Call:
lm(formula = Gender ~ Hgt, data = Pulse)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90150 -0.20348 -0.00216  0.20574  0.80311

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.343647   0.397563   18.47   <2e-16 ***
Hgt         -0.100658   0.005817  -17.30   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3305 on 230 degrees of freedom
Multiple R-squared:  0.5656,    Adjusted R-squared:  0.5637
F-statistic: 299.5 on 1 and 230 DF,  p-value: < 2.2e-16
```

Tests for individual coefficients

Compare models

Test for overall fit

Test for Individual Coefficients

$H_0: \beta_i = 0$
$H_1: \beta_i \neq 0$

$$t.s. = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$$

Supplied by *R*

Interpret as with individual t-tests in ordinary regression.

p-value = 2P(Z > |t.s.|)

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  64.1416     8.3694   7.664 1.81e-14 ***
Hgt          -0.9424     0.1227  -7.680 1.60e-14 ***
---
```

CI for Slope and Odds Ratio

From logistic model for Gender:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  64.1416     8.3694   7.664 1.81e-14 ***
Hgt          -0.9424     0.1227  -7.680 1.60e-14 ***
---
```

CI for slope: $\hat{\beta}_1 \pm z^* SE(\hat{\beta}_1)$

$$-0.9424 \pm 1.96(0.1227) = (-1.183, -0.702)$$

Exponentiate CI for slope

CI for odds ratio: $(e^{-1.183}, e^{-0.702}) = (0.30, 0.49)$

We are 95% confident that the odds of being female go down by a factor somewhere between 0.30 and 0.49 for every extra inch in height.

From logistic model for - height example:

	OR	2.5 %	97.5 %
(Intercept)	7.183333e+27	2.600201e+21	5.701922e+35
Hgt	3.896868e-01	2.984392e-01	4.843415e-01

```
logitmod<-glm(Gender ~ Hgt, family = binomial, data=Pulse)
exp(cbind(OR = coef(logitmod), confint(logitmod)))
```

We are 95% confident that the odds of being female go down by a factor somewhere between 0.29 and 0.48 for every extra inch in height.

Estimating Parameters in Logistic Regression

Parameters are chosen to *maximize* the *likelihood* of the observed sample (maximum likelihood estimation).

If the i^{th} point is YES ($y_i = 1$), calculate $\hat{\pi}_i$.

If the i^{th} point is NO ($y_i = 0$), calculate $1 - \hat{\pi}_i$.

Likelihood: $$L = \prod \hat{\pi}_i^{\,y_i} (1 - \hat{\pi}_i)^{1-y_i}$$

Here the estimated probabilities, $\hat{\pi}$, come from a model. What model is best?

(See Exercise 9.9 and likelihood-play.xls)

A Simple(?!) Max Likelihood Problem

If the i^{th} point is YES ($y_i = 1$), calculate $\hat{\pi}$.

If the i^{th} point is NO ($y_i = 0$), calculate $1 - \hat{\pi}$.

Likelihood: $$L = \prod \hat{\pi}^{y_i} (1 - \hat{\pi})^{1 - y_i}$$

Suppose $n = 20$ and there are 14 total successes.

$$L = \hat{\pi}^{14} (1 - \hat{\pi})^{6}$$

Calculus shows that the best estimate is $$\hat{\pi} = \frac{14}{20}$$

Evaluating Overall Effectiveness

1. Likelihood of the sample
   - Output below gives *Log-likelihood*
     - * Always negative
     - * Smaller -2*Log-likelihood* is better

Automated

```
'log Lik.' -67.81383 (df=2)
```

```
logLik(logitmod)
```

Manual

```
[1] -67.81383
```

```
pi<- fitted.values(logitmod)
likelihood <- ifelse(Pulse$Gender == 1, pi, 1 - pi)
log(prod(likelihood))
```

Example: Predict Gender via Weight

```
Call:
glm(formula = Gender ~ Wgt, family = binomial, data = Pulse)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.04279  -0.39608  -0.01075   0.53321   2.27350

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 16.67791    2.20912   7.550 4.37e-14 ***
Wgt         -0.10962    0.01458  -7.519 5.52e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 321.00  on 231  degrees of freedom
Residual deviance: 155.15  on 230  degrees of freedom
AIC: 159.15

Number of Fisher Scoring iterations: 6

'log Lik.' -77.57435 (df=2)
```

Hgt

```
'log Lik.' -67.81383 (df=2)
```

Evaluating Overall Fit

2. Test for overall fit
   (Similar to regression ANOVA)
   t.s. = G = improvement in $-2\ln(L)$ over a model with just a constant term
   Compare to χ^2 with k d.f.
   ↖ # predictors

```
Likelihood ratio test

Model 1: Gender ~ Hgt
Model 2: Gender ~ 1
  #Df   LogLik Df  Chisq Pr(>Chisq)
1   2  -67.814
2   1 -160.500 -1 185.37  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
require(lmtest)
lrtest(logitmod)
```

Deviance in R Output

```
Call:
glm(formula = Gender ~ Hgt, family = binomial, data = Pulse)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.77443  -0.34870  -0.05375   0.32973   2.37928

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  64.1416     8.3694   7.664 1.81e-14 ***
Hgt          -0.9424     0.1227  -7.680 1.60e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 321.00  on 231  degrees of freedom
Residual deviance: 135.63  on 230  degrees of freedom
AIC: 139.63

Number of Fisher Scoring iterations: 6
```

-2LL

Constant Model

This Model

How much better do we do with height than with nothing?

$G = 321.00 - 135.63 = 185.37$

Test for Overall Model—Logistic

Is there something effective in the model?

$H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 \quad \text{vs.} \quad \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

Same odds for all X

Odds are linear function of X

$$t.s. = G = -2\ln(L_0) - (-2\ln(L_1))$$ Compare to χ^2_1.

Improvement in $-2\ln(L)$ when using linear function of X.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  64.1416     8.3694   7.664 1.81e-14 ***
Hgt          -0.9424     0.1227  -7.680 1.60e-14 ***
---
```

	Taller ← Shorter		
	69”	**68”**	**67”**
$\text{Log(Odds)} = -0.9424Hgt$	-0.8856	0.0568	0.9992
$\text{Odds} = e^{64.14-0.9424Hgt}$	0.4125	1.0584	2.7161
$\pi = \frac{e^{64.14-0.9424Hgt}}{1+e^{64.14-0.9424Hgt}}$	0.2920	0.5142	0.7309

Example: Gender as a function of Smoking?

Gender	Smoker	Non-Smoker
Female	9	101
Male	17	105
Total	26	206

$$\widehat{\pi_{smoker}} = \frac{9}{26} = 0.35$$

$$Odds_{Smoker} = \frac{9}{17} = 0.53$$

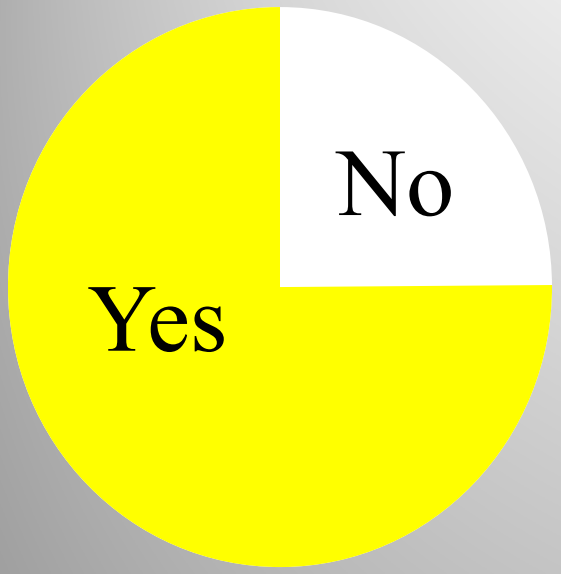
$$\widehat{\pi_{smoker}} = \frac{odds}{1 + odds} = \frac{0.53}{1.53} = 0.35$$

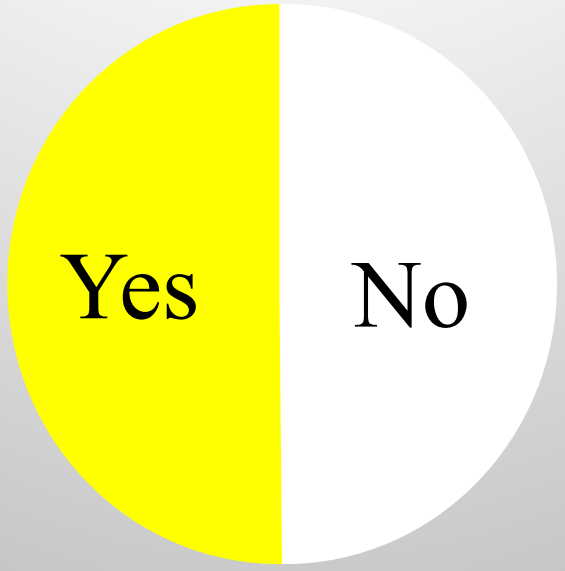
$$\widehat{\pi_{non-smoker}} = \frac{101}{206} = 0.49$$

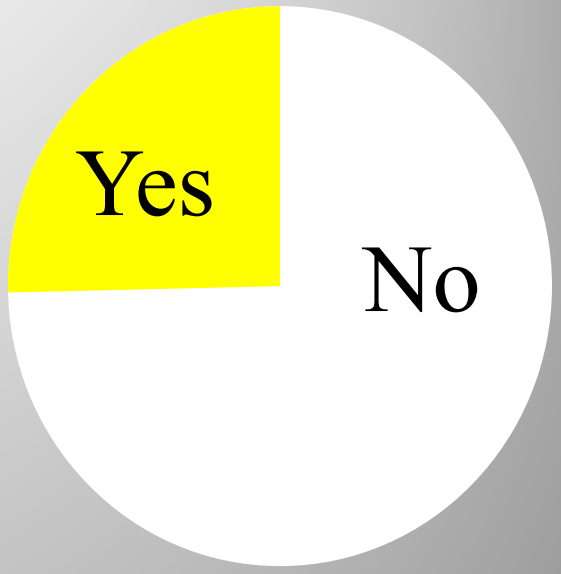
$$Odds_{Non-Smoker} = \frac{101}{105} = 0.96$$

$$OR = \frac{{}^{9}/_{17}}{{}^{101}/_{105}} = 0.55$$

Odds are 0.55 times lower of a smoker being Female than a non-smoker.


No
Yes


Yes
No


Yes
No

The Logistic Reg Max Likelihood Problem

Likelihood: $$L = \prod \hat{\pi}_i^{\,y_i} (1 - \hat{\pi}_i)^{1-y_i}$$

where $$\hat{\pi}_i = \frac{e^{\beta_0 + \beta_1 * x_i}}{1 + e^{\beta_0 + \beta_1 * x_i}}$$

For the putting data this is:

$$\mathrm{L} = \left[\frac{e^{\beta_0 + \beta_1 * 3}}{1 + e^{\beta_0 + \beta_1 * 3}}\right]^{84} * \left[\frac{1}{1 + e^{\beta_0 + \beta_1 * 3}}\right]^{17} * \left[\frac{e^{\beta_0 + \beta_1 * 4}}{1 + e^{\beta_0 + \beta_1 * 4}}\right]^{88} * \left[\frac{1}{1 + e^{\beta_0 + \beta_1 * 4}}\right]^{31} * etc$$

Maximize this with respect to β_0 and β_1 .

Example: Golf Putts

Length	3	4	5	6	7
Made	84	88	61	61	44
Missed	17	31	47	64	90
$\hat{\pi}$	0.826	0.730	0.605	0.465	0.330

When Length = 3, the model predicts

$$\frac{e^{3.256-0.5666*3}}{1+e^{3.256-0.5666*3}} = 0.826$$

The Likelihood component for Length = 3 is

$$0.826^{84}0.174^{17}$$

Length	3	4	5	6	7
Made	84	88	61	61	44
Missed	17	31	47	64	90
$\hat{\pi}$	0.826	0.730	0.605	0.465	0.330

Combining for all the data, the Likelihood is:

$$L = 0.826^{84}0.174^{17}0.730^{88}0.270^{31}\cdots 0.330^{44}0.670^{90}$$

$$\ln(L) = 84\ln(0.826) + 17\ln(0.174) + \cdots + 90\ln(0.670)$$

$$\ln(L) = -359.9$$

Coefficients are chosen to make $\ln(L)$ as large as possible.