BWH - Biostatistics

Intermediate Biostatistics for Medical Researchers

Robert Goldman
Professor of Statistics
Simmons College

Introduction to Logistic Regression

Tuesday, April 4, 2017

Descriptive Aspects of Logistic Regression

Thus far we have looked at regression models in which the response variable is *quantitative* and the explanatory variables are a mixture of quantitative and qualitative.

Now we look at models in which the response variable is *qualitative* and the explanatory variables are, again, a mixture of quantitative and qualitative.

In this context, the response variable, Y might be (i) whether or not a patient survives a procedure, (ii) Whether an infant is low birth-weight or not, or (iii) whether or not a patient can return home or go on to long-term care following rehabilitation.

When the response variable is qualitative with just two categories a frequently used technique is called **logistic regression**.

Uses for Logistic Regression

Logistic regression can be used to create a prediction rule and to identify ‘risk’ factors that affect the likelihood of an outcome.

In recent years logistic regression has been used to create *propensity scores*. These scores are used in observational studies as estimates of the probabilities that each participant would choose/receive the experimental treatment.

The Burn data

SOURCE: Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression: Third Edition. These data are copyrighted by John Wiley & Sons Inc.

Hospital Discharge Status	0 = Alive 1 = Dead	Death
Age at admission	Years	Age
Gender	0 = Female 1 = Male	Gender
Race	0 = Non-White 1 = White	Race
Total burn surface area	0 - 100%	TBSA
Burn involved inhalation injury	1 = Yes 0 = No	INH
Flame involved in burn injury	1 = Yes 0 = No	Flame

```
> head(burn)
```

	Death	Age	Gender	Race	TBSA	INH_INJ	Flame
1	0	26.6	1	1	25.3	0	1
2	0	2.0	0	0	5.0	0	0
3	0	22.0	0	0	2.0	0	0
4	0	37.3	1	1	2.0	0	0
5	0	52.1	1	1	6.0	0	1
6	0	50.2	1	1	7.0	0	0

```
> tail(burn)
```

	Death	Age	Gender	Race	TBSA	INH_INJ	Flame
995	1	83.7	0	1	50.5	0	0
996	1	34.2	1	1	91.0	1	1
997	1	59.0	1	1	37.5	1	1
998	1	85.5	1	1	4.6	1	1
999	1	46.8	1	0	47.0	1	1
1000	1	40.8	1	1	1.2	1	1

In this case we shall construct models that relate whether or not a person will die to (i) Flame, (ii) TBSA, and (iii) Flame and TBSA, and finally, to all the available predictors.

In this case, the response variable (Y) can take two values (1 or 0)

Why does linear regression not work in this case?

```
model <- lm(death ~ TBSA, burn)
model
```

```
Coefficients:
(Intercept)         TBSA
  -0.009719     0.011792
```

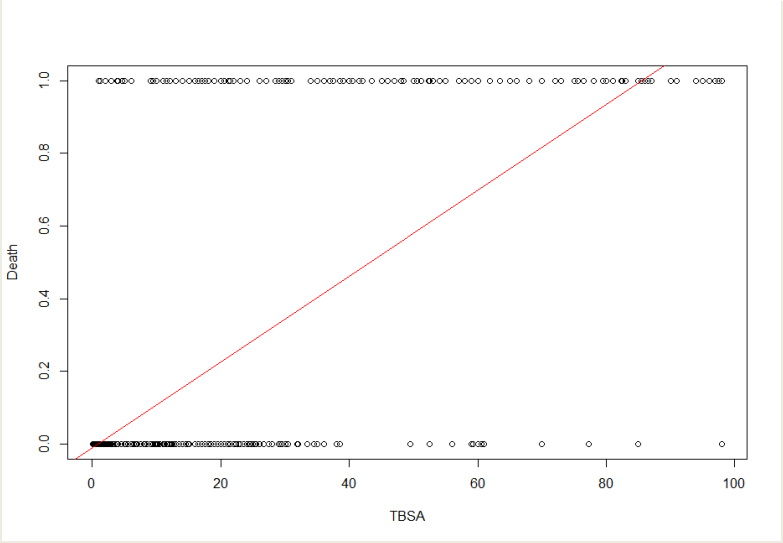
Death = -0.00972 + 0.01179TBSA

When TBSA = 50% Predicted Death = 0.5798

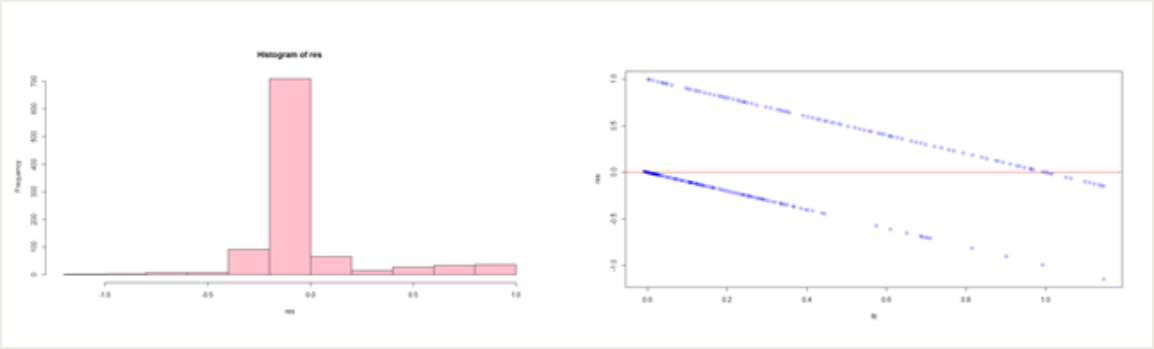
When TBSA = 0.1% Predicted Death = -0.0085

When TBSA = 99% Predicted Death = 1.157

```
model <- lm(Death ~ TBSA, burn)
plot(Death ~ TBSA, burn)
abline(model, col = "red")
```



If we used the model for inference, you will find that the conditions are not valid.



Some preliminary analyses

```
> table(burn$Death)
Death
  0   1
850 150
```

```
> t<- table(burn$Death, burn$Gender)
> t
     Gender
Death   0   1
    0 246 604
    1  49 101
```

```
> prop.table(t,2)
     Gender
Death         0         1
    0 0.8338983 0.8567376
    1 0.1661017 0.1432624
```

		Female	Male	All
Death	No	246	604	850
	Yes	49 (16.6%)	101 (14.3%)	150 (15%)
	All	295	705	1000

Race

Death	Non-White	White	All
No	356	494	850
Yes	55 (13.4%)	95 (16.1%)	150 (15%)
All	411	589	1000

INH_INJ

Death	No	Yes	All
No	800	50	850
Yes	78 (8.9%)	72 (59.0%)	150 (15%)
All	878	122	1000

Flame

Death	No	Yes	All
No	451	399	850
Yes	20 (4.2%)	130 (24.6%)	150 (15%)
All	471	529	1000

Death \ Flame	No	Yes	All
No	451	399	850
Yes	20 (4.2%)	130 (24.6%)	150 (15%)
All	471	529	1000

$$\hat{p}_N = \frac{20}{471} = 0.04246$$

$$\hat{p}_Y = \frac{130}{529} = 0.24575$$

$$\hat{O}_N = \frac{20}{451} = 0.04435 \qquad \hat{O}_Y = \frac{130}{399} = 0.32581$$

$$\widehat{OR} = 0.32581/0.04435 = 7.346.$$

Where a flame is involved, the burn victim's odds of death is 7.3 times the odds when a flame is not involved.

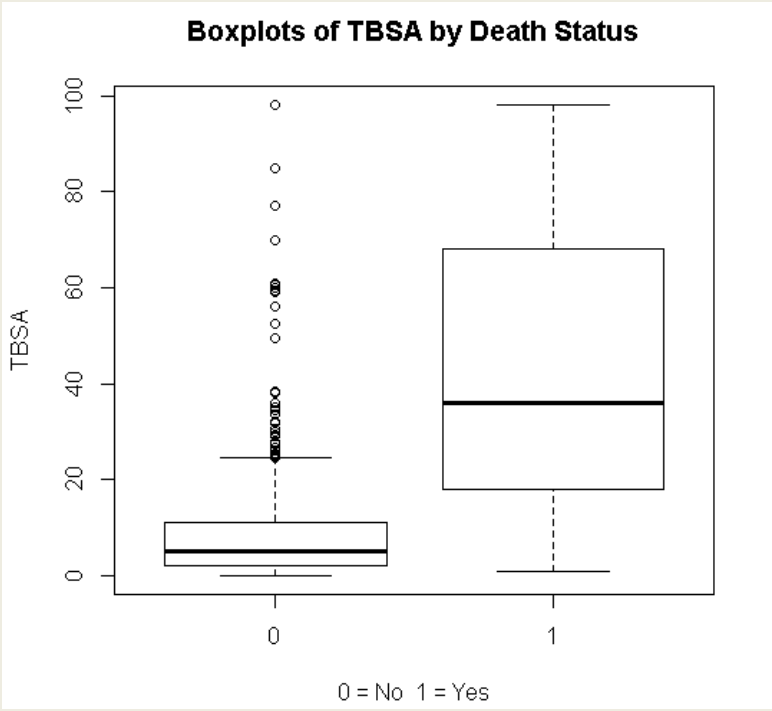
```
> tapply(burn$TBSA, burn$Death, summary)
No

$`0`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.100   2.000   5.000   8.505  11.000  98.000

Yes

$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   18.00   36.00   42.11   67.50   98.00
```

```
boxplot(burn$TBSA ~ burn$Death)
```



The Simple Logistic Regression Model

L**ogistic** regression models enable us to predict not Y but rather, the quantity p = P(Y = 1), the probability that a person will take the value Y = 1, as a function of the X variable(s). The simple logistic regression model is

$$P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Here, e = 2.718… is the base of natural logarithms.

The quantity

$$e^{\beta_0 + \beta_1 X}$$

must always be positive and can vary from 0 up to infinity. As a consequence

$$P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

must always lie between 0 and 1.

In simple linear regression (and multiple linear regression), statistical software uses the procedure called least squares to obtain, from the data,the ‘best’ values for the regression coefficients.

In the context of logistic regression, the software uses, not least squares, but a procedure called Maximum Likelihood Estimation to find the 'best' values for b_0 and b_1 from our data. The method seeks to find the values

$b_0 = \hat{\beta}_0$ and $b_1 = \hat{\beta}_1$

which are ‘most likely’ to have generated the sample of zeros or ones.

There are three ways to write the fitted model:

1. $\widehat{P(Y = 1)} = \hat{p} = \frac{e^{b0 + b1X}}{1 + e^{b0 + b1X}}$

This is an expression for the predicted probability that Y = 1.

2. $\frac{\hat{p}}{1 - \hat{p}} = \hat{O} = e^{b0 + b1X} = Exp(b_0 + b_1X)$

This is an expression for the predicted odds that Y = 1.

3. $\hat{L} = \ln(\frac{\hat{p}}{1 - \hat{p}}) = b_0 + b_1X$

This is an expression for the predicted log odds that Y = 1.

Logistic Regression when X is also 0/1

Here is the 'coefficients' output for a logistic regression when Flame is the explanatory variable.

```
model <- glm(Death ~ Flame, family =
binomial, burn)
 model
```

```
Coefficients:
(Intercept)         Flame
     -3.116         1.994
```

$$\widehat{P(Y=1)} = \hat{p} = \frac{e^{-3.116 + 1.994\text{Flame}}}{1+e^{-3.116 + 1.994\text{Flame}}}$$

$$\hat{O} = e^{-3.116 + 1.994\text{Flame}}$$

"No Flame" $\widehat{P(Y=1)} = \frac{e^{-3.116 + 1.994(0)}}{1+e^{-3.116 + 1.994(0)}} = 0.04245$

"Flame" $\widehat{P(Y=1)} = \frac{e^{-3.116 + 1.994(1)}}{1+e^{-3.116 + 1.994(1)}} = 0.24575$

These are the sample proportions we found earlier.

"No Flame" $\hat{O} = e^{-3.116 + 1.994(0)} = 0.04435$

"Flame" $\hat{O} = e^{-3.116 + 1.994(1)} = 0.32581$

These are the sample odds we found earlier.

When we have a 0/1 variable as the only explanatory variable, logistic regression returns predictions equal to the sample proportions and odds.

An important result!

X is a variable that takes values 0 or 1

The odds that $Y = 1 = e^{b0 + b1X}$

The odds ratio, $\widehat{OR} = \dfrac{\text{odds that } Y=1 \text{ when } X=1}{\text{odds that } Y=1 \text{ when } X=0}$

$$= \frac{e^{b0 + b1(1)}}{e^{b0 + b1(0)}}$$

$$= e^{b0 + b1 - b0} = e^{b1}$$

For our example $\widehat{OR} = e^{b1} = e^{1.994} = 7.346$

Logistic Regression When the Explanatory Variable is Quantitative (TBSA)

```
model <- glm(Death ~ TBSA, binomial, burn)
model
```

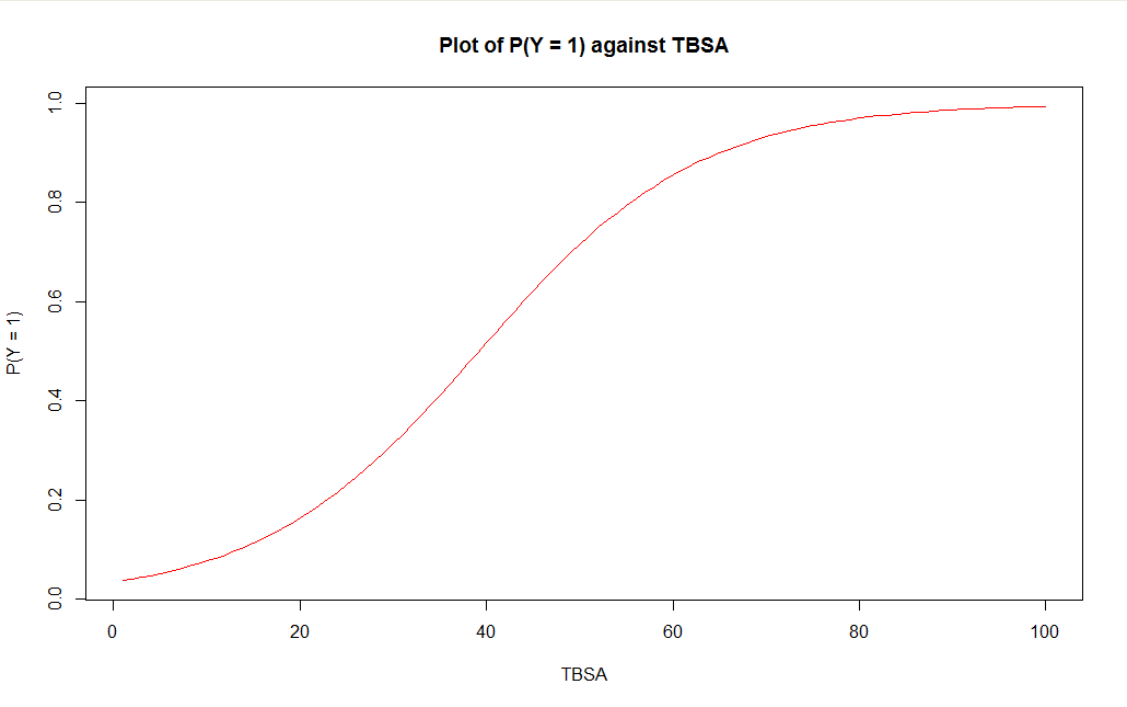
```
Coefficients:
(Intercept)         TBSA
   -3.34511      0.08537
```

$$\widehat{P(Y = 1)} = \hat{p} = \frac{e^{-3.34511 + 0.08537TBSA}}{1 + e^{-3.34511 + 0.08537TBSA}}$$

TBSA	$\widehat{P(Y = 1)}$
1%	0.036978
20%	0.162777
50%	0.715732
80%	0.970243
99%	0.993979

```
x <- seq(1,100)
z <- exp(-3.34511 + 0.08537*x)
y <- z/(1 + z)
plot(y ~ x, col = "red", type = "l",
     main = "Plot of P(Y = 1) against TBSA",
     xlab = "TBSA",
     ylab = "P(Y = 1)")
```

(The notation type = “l” connects the dots and omits the symbols.)



$\hat{O}$ = odds that Y= 1 = $e^{- 3.34511 - 0.08537TBSA}$

The predicted odds that a patient with TBSA of 20% will die is

$e^{- 3.34511 - 0.08537(20)} = 0.162777$

The predicted odds that a patient with TBSA of 80% will die is

$e^{- 3.34511 - 0.08537(80)} = 0.970243$

Earlier, we noted that when X is a 0/1 variable

$$\widehat{OR} = \frac{\text{odds that } Y=1 \text{ when } X=1}{\text{odds that } Y=1 \text{ when } X=0} = e^{b1}$$

Does e^{b1} have any similar interpretation when X is quantitative?

Yes!

$$e^{b1} = \frac{\text{odds that } Y=1 \text{ for } X}{\text{odds that } Y=1 \text{ for } X-1}$$

For our example, $b_1 = 0.08537$

So $e^{b1} = e^{-0.06647} = 1.08912$

For each additional 1% in TBSA, the predicted odds of dying change by a factor of 1.09.

In logistic regression where X is quantitative, e^{b1} is the factor by which the odds of Y = 1 change as X increases by one unit. In other words, e^{b1} is the odds (that Y = 1) ratio associated with being X as opposed to X - 1.

The odds (of dying) ratio associated with a TBSA of 21 rather than 20 is 1.08912

The odds (of dying) ratio associated with a TBSA of 81 rather than 80 is 1.08912

$$\frac{\text{Odds of dying with TBSA of 36}}{\text{Odds of dying with TBSA of 26}} =$$

$$\frac{\text{Odds of dying with TBSA of 26}}{\text{Odds of dying with TBSA of 36}} =$$

Classification Tables

The following code will assign a 1 if P(Y = 1) > 0.5 and a 0 if P(Y = 1) < 0.5 to preddeath.

```
model <- glm(Death ~ TBSA, binomial,
          burn)
fit <- fitted(model)
preddeath <- rep(0, 1000)
preddeath[fit >= 0.5] <- 1
n <- data.frame(burn$Death,
     preddeath)
table(burn$Death, preddeath)

head(n)
```

```
  burn.Death preddeath
1          0         0
2          0         0
3          0         0
4          0         0
5          0         0
6          0         0
```

```
tail(n)
```

```
     burn.Death preddeath
995           1         1
996           1         1
997           1         0
998           1         0
999           1         1
1000          1         0
```

```
preddeath
        0   1
  0 837  13
  1  82  68
```

	Predicted Death				
	No		Yes		All
Death?	No	837 (98.5%)	13		850
	Yes	82	68	(45.3%)	150
	All	919	81		1000

Death: p >0.5

		Predicted Death		
		No	Yes	All
Death?	No	837 (98.5%)	13	850
	Yes	82	68 (45.3%)	150
	All	919	81	1000

Death: p > 0.4

		Predicted Death		
		No	Yes	All
Death?	No	829 (97.5%)	21	850
	Yes	71	79 (52.7.3%)	150
	All	900	100	1000

TBSA + Flame

```
Model2 <- glm(Death ~ TBSA + Flame,
          binomial, burn)

summary(model2)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.105814   0.280726 -14.626  < 2e-16
TBSA         0.078119   0.006928  11.276  < 2e-16
Flame        1.267158   0.289756   4.373 1.22e-05
```

TBSA + Flame

$$\widehat{P(Y=1)} = \hat{p} = \frac{e^{-4.105814 + 0.078119TBSA + 1.267158Flame}}{1+e^{-4.105814 + 0.078119TBSA + 1.267158Flame}}$$

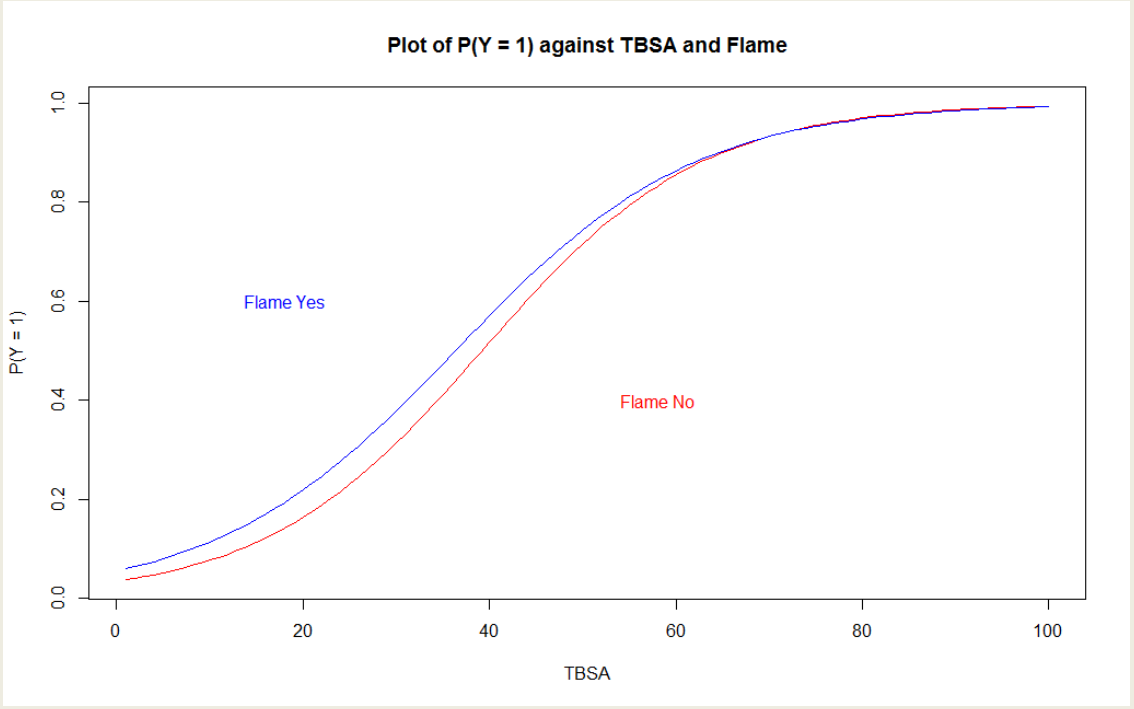
$b_1 = 0.07812 \qquad e^{b1} = e^{0.07812} = 1.0813$

Adj_OR for TBSA = 1.0813

$b_2 = 1.26716 \qquad e^{b2} = e^{1.26716} = 3.5508$

Adj_OR for Flame = 3.5508

```
x <- seq(1,100)
z1 <-  exp(-4.105814 + 0.078119*x)
y1 <- z1/(z1  + z1)
z2 <- exp(-2.838664 + 0.078119*x)
y2 <- z2/(1 + z2)
plot(y ~ x, col = "red", type = "l",
      main = "Plot of P(Y = 1) against TBSA and Flame",
      xlab = "TBSA",
      ylab = "P(Y = 1)")
lines(x, y2, col = "blue")
text(18, 0.6, "Flame Yes", col = "blue")
text(58, 0.4, "Flame No", col = "red")
```



Inferential Aspects of Logistic Regression

	Model	Odds ratio
Population	$P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X}}{1+e^{\beta_0 + \beta_1 X}}$	$OR = e^{\beta_1}$
Sample	$\widehat{P(Y = 1)} = \frac{e^{b_0 + b_1 X}}{1+e^{b_0 + b_1 X}}$	$\widehat{OR} = e^{b_1}$

For our example X is Flame or TBSA

- b_0 is an estimate for β_0
- b_1 is an estimate for β_1
- $\widehat{OR} = e^{b_1}$ is an estimate for $OR = e^{\beta_1}$

In logistic regression inferences can be based on either of two processes:

1. For large n, in repeated samples, the distribution of b_1 is approximately Normal with a mean of β_1.

2. Inferences can more reliably be based on the likelihood function—the probability of getting our sample.

Normal Inferences: Confidence Intervals

```
model <- glm(Death ~ TBSA, binomial, burn)
summary(model)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.345107   0.175648  -19.04   <2e-16
TBSA         0.085367   0.006956   12.27   <2e-16
```

A 95% confidence interval for β_1 is

$b_1 \pm 1.96\ SE(b_1)$

$0.0854 \pm 1.96\ (0.00696)$

$0.0854 \pm 0.01363 \rightarrow$ [0.0724 to 0.0997]

```
confint(model)
                   2.5 %      97.5 %
(Intercept) -3.70454223 -3.01470206
TBSA         0.07239926  0.09969456
```

A 95% confidence interval for OR = e^{β_1} is

$e^{0.0724}$ to $e^{0.0997}$ or 1.075 to 1.105

Normal Inferences: Hypothesis Testing

H_0: X and Y are independent

H_0: $\beta_1 = 0$ so $P(Y = 1) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$

H_0: $OR = e^{\beta_1} = 1$

H_A: X and Y are dependent

H_A: $\beta_1 \neq 0$ so $P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X}}{1+e^{\beta_0 + \beta_1 X}}$

H_A: $OR = e^{\beta_1} \neq 1$

1. Confidence Interval Approach to HT

A 95% confidence interval for β_1 is 0.0724 to 0.0997

A 95% confidence interval for OR 1.075 to 1.105

The first interval does not contain 0; the second does not contain 1.

We can reject the null hypothesis at the 5% level of significance.

The data suggest that the (population) odds ratio associated with a 1% increase in TBSA is significantly greater than 1.

2. The Wald Z Test

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.345107   0.175648  -19.04  <2e-16
TBSA         0.085367   0.006956   12.27  <2e-16
```

(a) Compute $Z_0 = \dfrac{b_1 - 0}{SE(b_1)} = \dfrac{0.085367}{0.006956} = 12.27$

(b) p-value = $P(Z < -12.27) + P(Z > 12.27)$

$= 2E\text{-}16$

Reject H_0 at the 1% level of significance.

The data strongly suggest that $\beta_1 > 0$.

3. The Wald Chi-Square Test

(a) Compute $Y = \left(\frac{b1}{SE(b1)}\right)^2 = \left[\frac{0.085367}{0.006956}\right]^2$

$= 12.27^2$

$= 150.55$

If H_0 is true, Y has a Chi-Square (1) distribution

p-value = P(Y > 150.55) = 2E-16

Reject H_0 at the 1% level of significance.

The data strongly suggest that $\beta_1 > 0$.

Inferences using the Likelihood Function

In logistic regression we estimate the coefficients β_0 and β_1 using a method called Maximum Likelihood Estimation (MLE). A likelihood function expresses the probability of obtaining the observed sample as a function of β_0 and β_1. The method of MLE asks: what values for β_0 and β_1 make our sample most likely?

The simplest situation to illustrate MLE is for the null case where p = P(Y = 1) is independent of X. That is

$$p = P(Y = 1) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

$$1 - p = P(Y = 0) = \frac{1}{1 + e^{\beta_0}}$$

Then, assuming independent observations

$$L(\beta_0) = \left(\frac{e^{\beta_0}}{1 + e^{\beta_0}}\right)^{150} \left(\frac{1}{1 + e^{\beta_0}}\right)^{850} \Leftarrow \text{Likelihood}$$

$$L_0(\beta_0) = \ln(L(\beta_0) = 850\,\beta_0 - 1000\ln(1 + e^{\beta_0})$$

Log Likelihood

To find the value for β_0 that maximizes $L_0(\beta_0)$:

$$\frac{dL_0}{d\beta_0} = 850 - 1000\frac{e^{\beta_0}}{1 + e^{\beta_0}} = 0$$

$$\hat{\beta}_0 = b_0 = -1.7346 \qquad e^{b_0} = \frac{150}{850} = \hat{O}$$

$$\widehat{P(Y = 1)} = \frac{150}{1000} = \hat{p}$$

In logistic regression the **deviance** plays the same role as the residual sum of squares in linear regression.

The **deviance** associated with a logistic regression model is

D = -2 ln(likelihood of the fitted model)

For our null model

$$\text{Likelihood} = L(b_0) = \left(\frac{e^{b_0}}{1+e^{b_0}}\right)^{150} \left(\frac{1}{1+e^{b_0}}\right)^{850}$$

$$= (0.15^{150})(0.85^{850})$$

$$D = -2\ln\left[(0.15^{150})(0.85^{850})\right]$$

$$= -2\left[150\ln(0.15) + 850\ln(0.85)\right]$$

$$= 845.42$$

```
model <- glm(Death ~ TBSA, binomial, burn)
summary(model)
```

```
    Null deviance: 845.42 on 999 degrees of freedom
Residual deviance: 538.65 on 998 degrees of freedom

AIC: 542.65
```

The Drop-in-Deviance Test

H_0: In the model $P(Y = 1) = \dfrac{e^{\beta_0 + \beta_1 TBSA}}{1 + e^{\beta_0 + \beta_1 TBSA}}$ $\beta_1 = 0$

H_A: $\beta_1 \neq 0$

```
anova(model, test = "Chisq")
     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                   999     845.42
TBSA  1   306.76        998     538.65 < 2.2e-16
```

Drop in Deviance

Deviance associated with TBSA

p-value for the Chi-Square test

$D_0 = 845.42$ $D_{TBSA} = 538.65$

$D_0 - D_{TBSA} = 306.76$

The Drop-in-deviance Chi-Square test can be used to compare two models so long as one is *nested* within the other. Model 1 is nested within model 2 if the predictor variables in model 1 are a subset of those in Model 2.

Here are several examples.

Example 1: Is it worth adding the variable Flame to a model predicting P(Y = 1) from TBSA?

1. Z test

```
model2 <- glm(Death ~ TBSA + Flame, binomial,
          burn)
summary(model2)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.105814   0.280726 -14.626  < 2e-16
TBSA         0.078119   0.006928  11.276  < 2e-16
Flame        1.267158   0.289756   4.373 1.22e-05
```

2. Drop-in-deviance Chi-Square test

```
model1 <- glm(Death ~ TBSA, binomial, burn)
anova(model1, model2, test = "Chisq")

Analysis of Deviance Table

Model 1: Death ~ TBSA
Model 2: Death ~ TBSA + Flame
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       998     538.65
2       997     516.68  1   21.978 2.758e-06
```

Example 2: Our current model (2) predicts P(Y = 1) from TBSA and Flame. Is it worth adding the remaining four potential predictors Age, Gender, Race, and INH_INJ?

```
model3 <- glm(Death ~ TBSA + Flame + Age +
      Gender + Race + INH_INJ, binomial, burn)
```

```
anova(model2, model3, test = "Chisq")
Analysis of Deviance Table

Model 1: Death ~ TBSA + Flame
Model 2: Death ~ TBSA + Flame + Age + Gender + Race
+ INH_INJ
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       997     516.68
2       993     336.46  4   180.21 < 2.2e-16 ***
```

Building a Logistic Regression Model

```
modelAge <- glm(Death ~Age, binomial, burn)
AIC(modelAge)
[1] 674.2585

modelGender <- glm(Death ~Gender, binomial,
burn)
AIC(modelGender)
[1] 848.5809

:    :    :    :    :    :    :    :    :    :

modelflame <- glm(Death ~ Flame, binomial, burn)
AIC(modelflame)
[1] 759.4591
```

	Variable	AIC
	Age	674.3
	Gender	848.6
	Race	848.0
√	TBSA	542.7
	INH_INJ	695.5
	Flame	759.5

The Complete Model

TBSA_Group = 1 if TBSA ≥ 50

= 0 otherwise

Age_Group = 1 if Age ≥ 32 [= median Age]

= 0 otherwise

```
m <- glm(Death ~ Gender + Race + INH_INJ + Flame
+ TBSA_Group + Age_Group, binomial, burn)
summary(m)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.5090     0.4246 -10.618  < 2e-16
Gender       -0.4682     0.2465  -1.900 0.057482
Race         -0.1794     0.2432  -0.738 0.460528
INH_INJ       1.7586     0.2810   6.258 3.90e-10
Flame         1.0432     0.2956   3.529 0.000417
TBSA_Group    3.1299     0.4079   7.673 1.68e-14
Age_Group     2.4375     0.3492   6.979 2.97e-12
```

```
Number of Fisher Scoring iterations: 6
```

Variable	Slope	Adj_OR	95% CI
Gender	- 0.468	0.626	0.387 - 1.019
Race	- 0.179	0.836	0.520 - 1.351
INH_INJ	1.759	5.807	3.356 -10.128
Flame	1.043	2.838	1.617 - 5.181
TBSA_Group	3.130	22.874	10.678 - 53.436
Age_Group	2.438	11.450	6.012 - 23.846

```
    Null deviance: 845.42  on 999  degrees of freedom

Residual deviance: 504.49  on 993  degrees of freedom
```

$D_0 - D_6 = 845.42 - 504.49 = 340.93$

This value can be compared to the Chi-Square distribution with 6 degrees of freedom.

Conditions for Inference in Logistic Regression

(a) Conditions we don't need

- No more condition that the Y values are approximately normal. Why not?

- No more condition that the standard deviation of the Ys not vary with the Xs.

- No more condition that the Ys are linearly related to the Xs.

(b) Conditions we do need

- We assume a linear relationship between the X variables and **logit** of Y

$$L = \log_e\left(\frac{\hat{p}}{1-\hat{p}}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots$$

It is hard to check the validity of this condition unless n is very large.

- We assume that the observations represent a random sample from some well-defined population.

- In logistic regression we estimate the population coefficients using a technique called maximum likelihood estimation (MLE). With this procedure, in large samples, the sample regression coefficients (b_1 and b_2, …etc) have approximately Normal distributions. What is a ‘large sample’ in the context of logistic regression?

Here are a couple of ‘rules of thumb’ that I am familiar with:

1. The smaller of the classes of the response variable (survived or died, in the case of the *burn* data) have at least 10 events per parameter in the model.

2. A minimum of 10 cases per explanatory variable in the model.

There are 1000 subjects in the *burn* data set. A total of 150 died and 850 survived.