

Logistic Regression for Binary Response Variable

Logistic Regression applies in situations where the response (i.e., dependent) variable is qualitative with only two possible outcomes (0 vs 1, "yes" vs "no", "absent" vs "present" etc.). Because of this, the response and error terms follow the Binomial Distribution. If, as is often the case in regression, there is only one response for each set of independent variables (i.e., no replication) then the Binomial distribution is \( B(1,p) \) with number of trials = 1 and \( p \) = probability of success. This special case is known as the Bernoulli distribution. Since the response variable is binary, modeling response involves modeling the probability \( \pi \) that the response variable \( Y \) takes one of the two values (e.t., \( Y=1 \) versus \( 0 \)). The alternative probability \( P(Y=0) \) is therefore \( 1-\pi \). Kutner et al. (KNNL) in *Applied Linear Statistical Models 5th ed. p. 557* list three "special problems" arising from this: 1) non-Normal Distribution of error terms, 2) non-constant variance associated with values of the independent variables \( X \), and 3) the requirement that \( \pi \) must be bound between (0,1). All three rule out use of standard linear models involving the Normal (Gaussian) distribution.

The requirement that \( \pi \) remain bound between (0,1) suggest use of a sigmoidal function as the natural description of probability. Currently three functions are in common use: Logistic, Probit & Clog-log. The first two are symmetric, intended for modeling more-or-less equal numbers of each response in the dataset, whereas the last is asymmetric useful with unequal numbers. Although all are easily available in R as options, by far the most commonly used is the Logistic function shown in more detail here. Logistic Regression is one example of the class of Generalized Linear Models (GLM) in which "best fit" linear coefficients for the independent variables \( X \) (also termed the "systematic component") are estimated for transformed values of the response variable \( \pi \), with the function describing the transformation termed the "link function". When the relationship between \( \pi \) and \( X \) is expressed directly in Logistic Regression, the equation describes a non-linear function with sigmoidal shape. Zuur et al. 2009 (Za) in *Mixed Effects Models and Extensions in Ecology with R* usefully describe all GLM's as having three formal components: 1 - a statement of the distribution of \( Y \), 2 - a statement of the expected value of \( Y \), \( \exp(Y) \) and variance of \( Y \), \( \text{var}(Y) \), and 3 - a description of the link function. I'll follow their format here.

Assumptions:

Regression depends on specifying in Response & Independent variables in advance:

\[ Y = \text{vector of binary response variable (0 or 1), each row of } Y \text{ indicated by index } i. \]
\[ X = \text{matrix independent variables (columns) with observations of } X_i \text{ (rows) matched to } Y_i \text{ (rows of } Y). \]
\[ \beta = \text{vector of linear coefficients and } X\beta \text{ is the linear predictor (systematic component) of the model.} \]

Cases \( Y_i \) are independent.

Model:

\[ Y_i \sim B(1,\pi_i) \quad < \text{Y's are Bernoulli distributed with probability } \pi_i \text{ for each case.} \]
\[ E(Y_i) = \pi_i, \text{ and } \text{var}(Y_i) = \pi_i(1-\pi_i) \quad < \text{mean and variance defined.} \]

\[ \logit(\pi_i) = X\beta \text{ where } \logit(\pi_i) = \pi_i/(1-\pi_i) \quad < \text{for logit link to } \pi_i \text{ for Logistic Regression} \]

\[ \text{alternatively: } \pi_i = e^{X\beta}/(1+e^{X\beta}) \]

Estimation of Regression Coefficients:

Estimation is based on determining the maximum likelihood function given the data. Since a closed-form solution doesn’t exit, this requires interactive computation, here using \( \text{glm()} \) in the \{nlme\} package in R.
# GLM 020 LOGISTIC REGRESSION

library(nlme)
setwd("c:/DATA/Models")
D=read.table("KNNL1401.txt",header=T)
D
FM4=glm(Y~X,data=D,family=binomial)
summary(FM4)

> summary(FM4)

Call: glm(formula = Y ~ X, family = binomial, data = D)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
 -1.8992  -0.7509  -0.4140   0.7992   1.9624

Coefficients:                          Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.05970  1.25935  -2.430   0.0151 *
X               0.16149  0.06498   2.485   0.0129 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 34.296  on 24 degrees of freedom
Residual deviance: 25.425  on 23 degrees of freedom
AIC: 29.425

Number of Fisher Scoring iterations: 4

Fitted Values:

Results=cbind(X=D$X,Y=D$Y,Fitted=fitted(FM4))
Results

Example for first point X_1 = 14:

X_1 := 14  Y_1 := 0  < observed values of X & Y for the first case
β_0 := -3.05970  β_1 := 0.16149  < regression coefficients from above
π_1 := \frac{e^{(β_0+β_1X_1)}}{1 + e^{(β_0+β_1X_1)}}  π_1 = 0.3103  < fitted value for the first case π_1

Odds & Odds Ratio:

Odds := \frac{π_1}{1 - π_1}  Odds_1 = 0.45  = Odds (ratio of probability π_1 and its complement)
X_2 := X_1 + 1  < unit increase in independent variable X
π_2 := \frac{e^{(β_0+β_1X_2)}}{1 + e^{(β_0+β_1X_2)}}  π_2 = 0.346
### Logistic Regression

Odds ratio: 
\[
\frac{\pi_2}{1 - \pi_2} = 1.175 \quad e^{\beta_1} = 1.175 \quad \text{= Odds Ratio}
\]

\[^*\text{A unit increase in the independent variable X results in a 17.5% increase in the odds of Y (completing the task in this example).}\]

\[c := 15 \quad < \text{an arbitrary number of units in X (for estimating changes over a larger interval of X)}\]

\[e^{\beta_1} = 11.272 \quad < \text{Odds increase 11-fold over 15 months}\]

```r
Residuals = cbind(X = D$X, Y = D$Y,
                  Pearson = resid(FM1, type = "pearson"),
                  Deviance = resid(FM1, type = "deviance"))
```

**Pearson Residuals:**
\[
RP := \frac{Y_1 - \pi_1}{\sqrt{\pi_1 \cdot (1 - \pi_1)}} \quad RP = -0.6707 \quad < \text{pearson residual for the first case}
\]

**Deviance Residuals:**
\[
r_{dev1} := \text{sign}(Y_1 - \pi_1) \cdot \sqrt{-2 \cdot \left[ Y_1 \cdot \ln(\pi_1) + (1 - Y_1) \cdot \ln(1 - \pi_1) \right]}
\]
\[
\text{sign}(Y_1 - \pi_1) = -1 \quad < \text{function for sign}
\]
\[
r_{dev1} = -0.8619 \quad < \text{deviance residual for the first case}
\]

```r
Residuals
> X Y Pearson Deviance
1 14 0 -0.6706912 -0.8619095
2 29 0 -2.2517280 -1.8991601
3 6 0 -0.3515546 -0.4827618
4 25 1 0.6134073 0.7992195
5 18 1 1.0794748 1.2430150
6 4 0 -0.2991303 -0.4140037
7 18 0 -0.9263764 -1.1131881
8 12 0 -0.5706767 -0.7508920
9 22 1 0.7815336 0.9764504
10 6 0 -0.3515546 -0.4827618
11 30 1 0.4096546 0.5570209
12 11 0 -0.5264098 -0.6994248
13 30 1 0.4096546 0.5570209
14 5 0 -0.3242848 -0.4471928
15 20 1 0.9185019 1.1061148
16 13 0 -0.6186662 -0.8050748
17 9 0 -0.4479107 -0.6047172
18 32 1 0.3485663 0.4788856
19 24 0 -1.5037818 -1.5376242
20 13 1 1.6163806 1.6027798
21 19 0 -1.0042774 -1.1810373
22 4 0 -0.2991303 -0.4140037
23 28 1 0.4814490 0.6457104
24 22 1 0.7815336 0.9764504
25 8 1 2.4203309 1.9623537
```

### Plotting Curve:

```r
M = data.frame(X = D$X)  # USING ORIGINAL INDEPENDENT VALUES
M = na.omit(M)
Pred = predict(FM1, newdata = M, type = "response")
plot(x = D$X, y = D$Y, xlab = "X", ylab = "Y", pch = 20)
points(M$X, Pred, pch = 20, col = "red")
```
Comparison of Logistic, Probit & Clog-log fits:

```r
B = read.table("Boar.txt", header=T)
FM2 = glm(Tb ~ LengthCT, family = binomial(link="logit"), data = B) # (link="logit") DEFAULT
FM3 = glm(Tb ~ LengthCT, family = binomial(link="probit"), data = B)
FM4 = glm(Tb ~ LengthCT, family = binomial(link="clog-log"), data = B)

# PLOTTING CURVES:
M = data.frame(LengthCT = B$LengthCT) # USING ORIGINAL INDEPENDENT VALUES
M = na.omit(M)
# M = data.frame(LengthCT = seq(from = 46.5, to = 165, by = 1)) # USING NEW SEQUENCE
Pred2 = predict(FM2, newdata = M, type = "response")
Pred3 = predict(FM3, newdata = M, type = "response")
Pred4 = predict(FM4, newdata = M, type = "response")
plot(x = B$LengthCT, y = B$Tb, xlab = "Length", ylab = "Tb", pch = 20)
lines(M$LengthCT, Pred2, col = "red")
lines(M$LengthCT, Pred3, col = "blue")
lines(M$LengthCT, Pred4, col = "green")
legend("topleft", legend = c("logit", "probit", "clog-log"),
  lty = c(1, 1, 1), col = c("red", "blue", "green"), cex = 0.8)
```

data from Za boar.txt data:
Tb - Response Variable
LengthCT - Independent Variable

![Graph showing comparison of logistic, probit, and clog-log fits](image)
Multiple Logistic Regression:

C=read.table("ParasiteCod.txt",header=T)
C=na.omit(C)
C$fArea=factor(C$Area)
C$fYear=factor(C$Year)
FM5=glm(Prevalence~fArea*fYear+Length,family=binomial,data=C)
summary(FM5)

> summary(FM5)

Call:
  glm(formula = Prevalence ~ fArea * fYear + Length, family =
  binomial, data = C)

Deviance Residuals:
  Min       1Q   Median       3Q      Max
-2.0922  -0.9089  -0.4545   0.9678   2.2394

Coefficients:
  Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.003226   0.291973   0.011   0.99118
fArea2           -1.185849   0.276897  -4.283  1.85e-05 ***
fArea3           -1.136105   0.231248  -4.913 8.97e-07 ***
fArea4            0.728736   0.261815   2.783  0.00538 **
fYear2000        -0.383756   0.343877  -1.116   0.26444
fYear2001        -2.655704   0.433542  -6.126  9.03e-10 ***
Length           0.008516   0.004585   1.858  0.06324
fArea2:fYear2000  0.209035   0.503494   0.415   0.67802
fArea3:fYear2000  0.561158   0.443733   1.265   0.20600
fArea4:fYear2001  0.451582   0.588318   0.768   0.44274
fArea2:fYear2001  2.403050   0.493512   4.869  1.12e-06 ***
fArea4:fYear2001  2.115534   0.513489   4.120  3.79e-05 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1727.8 on 1247 degrees of freedom
Residual deviance: 1495.2 on 1235 degrees of freedom
(6 observations deleted due to missingness)
AIC: 1521.2

Number of Fisher Scoring iterations: 4

Wald Test of single β:

Hypotheses:
  H₀: β = 0
  H₁: β ≠ 0

Test Statistic:
  z = Estimate/std.error

Example calculations:
  β₁ := -1.185849   sβ₁ := 0.276897
  z₁ := β₁ / sβ₁  z₁ = -4.283
  β₆ := 0.008516   sβ₆ := 0.004585
  z₆ := β₆ / sβ₆  z₆ = 1.857

Sampling Distribution:
  If Assumptions hold and H₀ is true,
  then z ~ N(0,1)

Probability:
  P₁ := min[(2 · pnorm(z₁,0,1)),2 · (1 − pnorm(z₁,0,1))]  P₁ = 1.8469 × 10⁻⁵
  P₆ := min[(2 · pnorm(z₆,0,1)),2 · (1 − pnorm(z₆,0,1))]  P₆ = 0.0633

Decision Rule:

IF P < α THEN REJECT H₀, OTHERWISE ACCEPT H₀

The Wald test is a marginal test of single regression coefficients having a function similar to t-test in standard Linear Regression.

Note: Regression coefficients and standard errors are obtained from the maximum likelihood fit as first and second partial derivatives of the likelihood function. Summary information is obtained from the summary() wrapper of an object made by glm(). No attempt has been made to replicate the calculations here.
Confidence Interval for $\beta$:

$$\alpha := 0.05$$

$$C := \left| \text{qnorm}\left(1 - \frac{\alpha}{2}, 0, 1\right) \right|$$

$$C = 1.96$$

\[\beta_1 = -1.186\]  \[\text{CI}_1 = (-1.729, -0.643)\]

\[\beta_6 = 0.00852\]  \[\text{CI}_6 = (-0.00047, 0.0175)\]

Likelihood Ratio Test:

The Likelihood Ratio Test in GLM models serves a function similar to that of the General F Ratio test of Full Model (FM) versus Reduced Model (RM). Since model likelihoods are not reported by summary.glm or anova.glm, the Likelihood Ratio test of is conducted by consulting the Analysis of Deviance Table in anova(RM,FM) (for serial testing) or drop1 (for marginal testing).

```
c C
anova(FM5,test="Chisq")
drop1(FM5,test="Chisq")
```

Hypotheses:

- $H_0$: a specified subset of $\beta$'s = 0
- $H_1$: same subset of $\beta$'s <> 0

Test Statistic:

$$G^2 = -2\ln[L(R)/L(F)] = -2\ln(L(R) - 2\ln(L(F)) = \text{dev}_F - \text{dev}_R$$

Sampling Distribution:

If Assumptions hold and $H_0$ is true, then

$$G^2 \sim \chi^2(df_F - df_R)$$

\[<\text{ where } df_F & df_R \text{ are degrees of freedom of FM & RM}\]

Probability:

$$P = 1 - \text{pchisq}(\Delta\text{dev},\Delta\text{df})$$

\[<\text{ where } \Delta\text{dev} & \Delta\text{df} \text{ are differences in values between FM & RM}\]

Decision Rule:

IF $P < \alpha$ THEN REJECT $H_0$, OTHERWISE ACCEPT $H_0$
**Example Calculations:**

**For Interaction term (blue in table):**

\[
\text{dev}_F := 1422.7 \quad \text{dev}_R := 1474.7
\]

\[
\Delta \text{dev} := \text{dev}_R - \text{dev}_F \quad \Delta \text{dev} = 52
\]

\[
\text{df}_F := 1178 \quad \text{df}_R := 1184
\]

\[
\Delta \text{df} := \text{df}_R - \text{df}_F \quad \Delta \text{df} = 6
\]

\[
P := 1 - \text{pchisq}(\Delta \text{dev}, \Delta \text{df})
\]

\[
P = 1.865 \times 10^{-9}
\]

**For factor Length (green):**

\[
\text{dev}_F := 1422.7 \quad \text{dev}_R := 1424.9
\]

\[
\Delta \text{dev} := \text{dev}_R - \text{dev}_F \quad \Delta \text{dev} = 2.2
\]

\[
\text{df}_F := 1178 \quad \text{df}_R := 1179
\]

\[
\Delta \text{df} := \text{df}_R - \text{df}_F \quad \Delta \text{df} = 1
\]

\[
P := 1 - \text{pchisq}(\Delta \text{dev}, \Delta \text{df})
\]

\[
P = 0.138
\]

Note: slight differences here are due to rounding differences in hand calculation.

---

**Serial report of Likelihood Ratio Tests:**

```r
> anova(FM5,test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: Prevalence
Terms added sequentially (first to last)

| Df | Deviance | Resid. Df | Resid. Dev | F(>|Chi|) |
|----|----------|-----------|------------|-----------|
| NULL | 1640.7 | 1190 | 1640.7 |  |
| fArea | 3 | 113.171 | 1187 | 1527.5 < 2.2e-16 *** |
| fYear | 2 | 49.406 | 1185 | 1478.1 1.869e-11 *** |
| Length | 1 | 3.446 | 1184 | 1474.7 0.0634 |
| fArea:fYear | 6 | 52.031 | 1178 | 1422.7 1.838e-09 *** |

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Marginal report of Likelihood Ratio Tests:**

```r
> drop1(FM5,test="Chisq")
Single term deletions

Model: Prevalence ~ fArea * fYear + Length

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
<th>LRT</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>1422.7</td>
<td>1448.7</td>
<td>1422.7</td>
<td>1448.7</td>
</tr>
<tr>
<td>fArea</td>
<td>1</td>
<td>1424.9</td>
<td>1448.9</td>
<td>2.231</td>
</tr>
<tr>
<td>fYear</td>
<td>1</td>
<td>1474.7</td>
<td>1488.7</td>
<td>52.031</td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

---

**Explicit Test for Interaction:**

```r
> anova(RM1,FM5,test="Chisq")
Analysis of Deviance Table

Model 1: Prevalence ~ fArea + fYear + Length
Model 2: Prevalence ~ fArea * fYear + Length

| Df | Deviance | Resid. Df | Resid. Dev | F(>|Chi|) |
|----|----------|-----------|------------|-----------|
| 1 | 1184 | 1179 | 1422.7 | 6 52.031 1.838e-09 *** |
```

---

**Explicit test for factor Length:**

```r
> anova(RM2,FM5,test="Chisq")
Analysis of Deviance Table

Model 1: Prevalence ~ fArea * fYear
Model 2: Prevalence ~ fArea * fYear + Length

| Df | Deviance | Resid. Df | Resid. Dev | F(>|Chi|) |
|----|----------|-----------|------------|-----------|
| 1 | 1179 | 1178 | 1424.9 | |
| 2 | 1178 | 1178 | 1422.7 | 1 2.2313 0.1352 |
```
> summary(FM5)

Call:
  glm(formula = Prevalence ~ fArea * fYear + Length, family = binomial,
      data = C)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.0858  -0.8626  -0.4865   0.9625   2.2218

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                0.085255   0.295024   0.289   0.7726
fArea2                    -1.321373   0.285258  -4.632 3.62e-06 ***
fArea3                    -1.449183   0.243884  -5.942 2.81e-09 ***
fArea4                     0.300728   0.271107   1.109   0.2673
fYear2000                 -0.395069   0.343814  -1.149   0.2505
fYear2001                 -2.652010   0.433369  -6.120 9.39e-10 ***
Length                    0.006933   0.004653   1.490  0.1362
fArea2:fYear2000         -0.080344   0.507968  -0.158   0.8743
fArea3:fYear2000         -0.870585   0.450273  -1.933   0.0532 .
fArea4:fYear2000         -0.864622   0.592386  -1.460   0.1444
fArea2:fYear2001         -2.737488   0.532881  -5.137 2.79e-07 ***
fArea3:fYear2001         -2.718986   0.499459  -5.444 5.21e-08 ***
fArea4:fYear2001         -2.541437   0.518224  -4.904 9.38e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1640.7  on 1190  degrees of freedom
Residual deviance: 1422.7  on 1178  degrees of freedom
AIC: 1448.7

Number of Fisher Scoring iterations: 4

Note: P-values reported by Wald Test and Likelihood Ratio test for a single $\beta$ will not necessarily be equal.

As in linear regression, summary() are marginal reports.