ORIGIN := 0

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 numbr 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 π that the response variable Y takes one of the two values (e.t., Y=1 versus 0). The alternative probabilility P(Y=0) is therefore (1-π). Kutner et al. (KNNL) in *Applied Linear Statistical Models 5th ed.* p. 557 list three "special problems" arrising 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 π must be bound between (0,1). All three rule out use of standard linear models involving the Normal (Gaussian) distribution.

The requirement that π remain bound between (0,1) suggest use of a sigmoidal function as the natural discription 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 π, with the function describing the transformation termed the "link function". When the relationship between π 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, 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 = vector of binary response variable (0 or 1), each row of Y indicated by index i.
X= matrix independent variables (columns) with observations of X_i (rows) matched to Y_i (rows of Y).
β = vector of linear coefficients and $X\beta$ is the linear predictor (systematic component) of the model.
Cases Y_i are independent.

Model:

$Y_i \sim B(1,\pi_i)$ **< Y's are Bernoulli distributed with probability π_i for each case.**
$E(Y_i) = \pi_i$, and $var(Y_i) = \pi_i(1-\pi_i)$ **< mean and variance defined.**

$logit(\pi_i) = X\beta$ where $logit(\pi_i) = \pi_i/(1-\pi_i)$
alternatively: $\pi_i = e^{(X\beta)}/(1+e^{(X\beta)})$ **< for logit link to π_i for Logistic Regression**

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 interative computation, here using 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(FM1)
```

data from KNNL Table 14.1
Y is the response variable (0 or 1)
X is the indepenent variable

```
> D
    X Y  fittedY
1  14 0 0.310262
2  29 0 0.835263
3   6 0 0.109996
4  25 1 0.726602
5  18 1 0.461837
6   4 0 0.082130
7  18 0 0.461837
8  12 0 0.245666
9  22 1 0.620812
10  6 0 0.109996
11 30 1 0.856299
12 11 0 0.216980
13 30 1 0.856299
14  5 0 0.095154
15 20 1 0.542404
16 13 0 0.276802
17  9 0 0.167100
18 32 1 0.891664
19 24 0 0.693379
20 13 1 0.276802
21 19 0 0.502134
22  4 0 0.082130
23 28 1 0.811825
24 22 1 0.620812
25  8 1 0.145815
```

```
> summary(FM1)

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(FM1))
Results
```

```
> Results
    X Y     Fitted
1  14 0 0.31026237
2  29 0 0.83526292
3   6 0 0.10999616
4  25 1 0.72660237
5  18 1 0.46183704
6   4 0 0.08213002
7  18 0 0.46183704
8  12 0 0.24566554
9  22 1 0.62081158
10  6 0 0.10999616
11 30 1 0.85629862
12 11 0 0.21698039
13 30 1 0.85629862
14  5 0 0.09515416
15 20 1 0.54240353
16 13 0 0.27680234
17  9 0 0.16709980
18 32 1 0.89166416
19 24 0 0.69337941
20 13 1 0.27680234
21 19 0 0.50213414
22  4 0 0.08213002
23 28 1 0.81182461
24 22 1 0.62081158
25  8 1 0.14581508
```

Example for first point $X_1 = 14$:

$X_1 := 14 \quad Y_1 := 0$ **< observed values of X & Y for the first case**

$\beta_0 := -3.05970 \quad \beta_1 := 0.16149$ **< regression coefficients from above**

$$\pi_1 := \frac{e^{(\beta_0+\beta_1 \cdot X_1)}}{1 + e^{(\beta_0+\beta_1 \cdot X_1)}} \qquad \pi_1 = 0.3103$$ **< fitted value for the first case π_1**

Odds & Odds Ratio:

$$Odds_1 := \frac{\pi_1}{1 - \pi_1} \qquad Odds_1 = 0.45$$ **= Odds (ratio of probability π_1 and its complement)**

$X_2 := X_1 + 1$ **< unit increase in independent variable X**

$$\pi_2 := \frac{e^{(\beta_0+\beta_1 \cdot X_2)}}{1 + e^{(\beta_0+\beta_1 \cdot X_2)}} \qquad \pi_2 = 0.346$$

$$Odds_2 := \frac{\pi_2}{1-\pi_2} \qquad Odds_2 = 0.529$$

$$\frac{Odds_2}{Odds_1} = 1.175 \qquad e^{\beta_1} = 1.175$$ **= Odds Ratio**

^ 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).

$c := 15$ **< an arbitrary number of units in X (for estimating changes over a larger interval of X)**

$e^{c\beta_1} = 11.272$ **< Odds increase 11-fold over 15 months**

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

```
> 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
```

Pearson Residuals:

$$r_P := \frac{Y_1 - \pi_1}{\sqrt{\pi_1 \cdot (1-\pi_1)}} \qquad r_P = -0.6707$$ **< pearson residual for the first case**

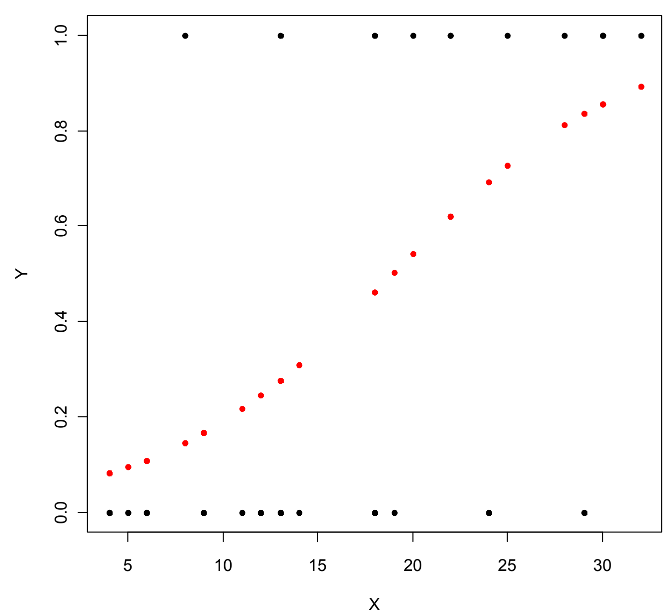
Deviance Residuals:

$$r_{dev1} := \text{sign}(Y_1 - \pi_1) \cdot \sqrt{-2 \cdot \left[Y_1 \cdot \ln(\pi_1) + (1 - Y_1) \cdot \ln(1-\pi_1) \right]}$$

$\text{sign}(Y_1 - \pi_1) = -1$ **< function for sign**

$r_{dev1} = -0.8619$ **< deviance residual for the first case**

```
#PLOTTING CURVE:
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")
```

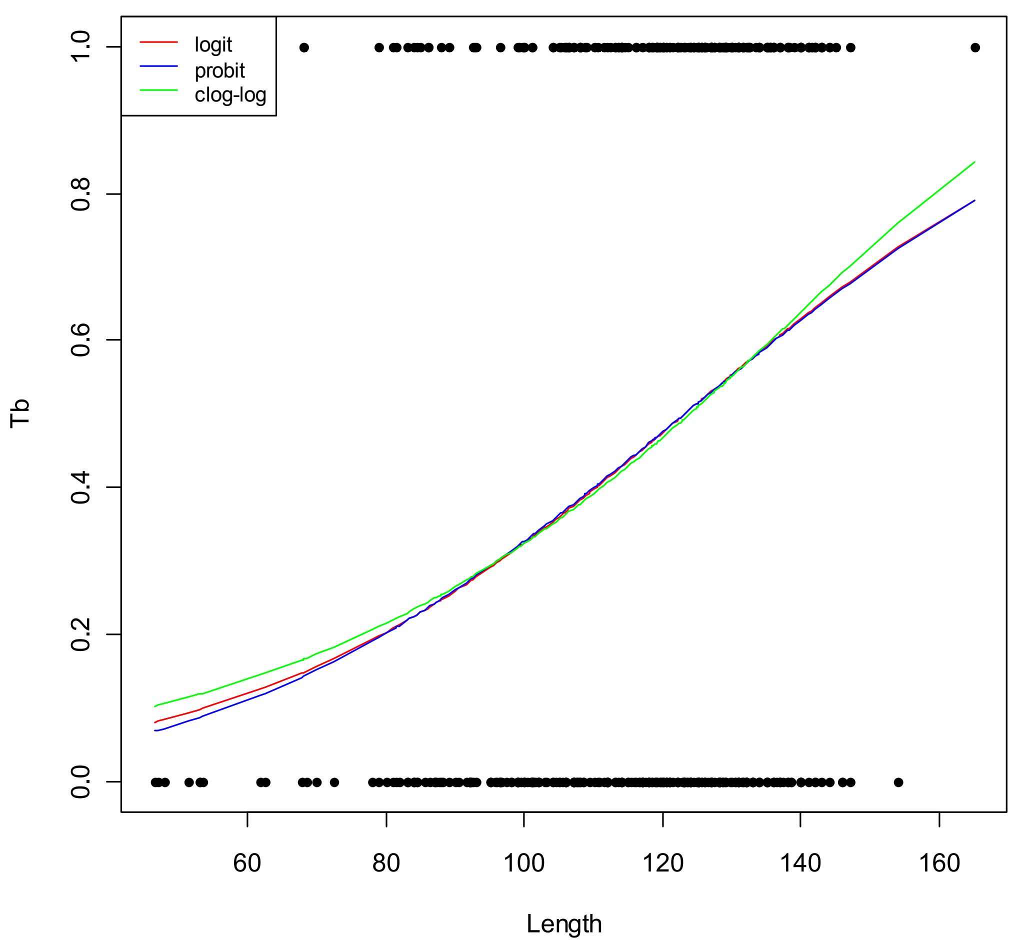


Comparison of Logistic, Probit & Clog-log fits:

```
B=read.table("Boar.txt",header=T)
B
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="cloglog"),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



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)
```

Za ParasiteCod data:
Prevalence of parasite = response.

Multiple independent variables X evaluated for a "best fit" model as in standard Linear Regression.

Wald Test of single β:

Hypotheses:

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

Test Statistic:

z = Estimate/std.error

Example calculations:

$\beta_1 := -1.185849 \qquad s_{\beta_1} := 0.276897$

$z_1 := \dfrac{\beta_1}{s_{\beta_1}} \qquad z_1 = -4.283$

$\beta_6 := 0.008516 \qquad s_{\beta_6} := 0.004585$

$z_6 := \dfrac{\beta_6}{s_{\beta_6}} \qquad z_6 = 1.857$

```
> 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:fYear2000  0.451582   0.588318   0.768  0.44274
fArea2:fYear2001  2.595866   0.528472   4.912 9.01e-07 ***
fArea3: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
```

Sampling Distribution:

If Assumptions hold and H_0 is true, then z ~ N(0,1)

Probability:

$P_1 := \min\left[\left(2 \cdot \text{pnorm}\left(z_1, 0, 1\right)\right), 2 \cdot \left(1 - \text{pnorm}\left(z_1, 0, 1\right)\right)\right] \qquad P_1 = 1.8469 \times 10^{-5}$

$P_6 := \min\left[\left(2 \cdot \text{pnorm}\left(z_6, 0, 1\right)\right), 2 \cdot \left(1 - \text{pnorm}\left(z_6, 0, 1\right)\right)\right] \qquad P_6 = 0.0633$

Decision Rule:

IF P < α THEN REJECT H_0, OTHERWISE ACCEPT H_0

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 β:

$\alpha := 0.05$

$C := \left|\text{qnorm}\left(1 - \frac{\alpha}{2}, 0, 1\right)\right| \qquad C = 1.96$

	estimate	confidence interval
$CI_1 := \left(\beta_1 - C \cdot s_{\beta_1} \quad \beta_1 + C \cdot s_{\beta_1}\right)$	$\beta_1 = -1.186$	$CI_1 = (-1.729 \quad -0.643)$
$CI_6 := \left(\beta_6 - C \cdot s_{\beta_6} \quad \beta_6 + C \cdot s_{\beta_6}\right)$	$\beta_6 = 0.00852$	$CI_6 = (-0.00047 \quad 0.0175)$

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

```
> C
     Sample Intensity Prevalence Year Depth Weight Length Sex Stage Age Area fArea fYear
1         1         0          0 1999   220    148     26   0     0   0    2     2  1999
2         2         0          0 1999   220    144     26   0     0   0    2     2  1999
3         3         0          0 1999   220    146     27   0     0   0    2     2  1999
4         4         0          0 1999   220    138     26   0     0   0    2     2  1999
5         5         0          0 1999   220     40     17   0     0   0    2     2  1999
...
1248   1248        84          1 2001   260    910     48   2     1   4    2     2  2001
1249   1249        89          1 2001   260   1414     56   1     1   6    2     2  2001
1250   1250        90          1 2001   228    224     31   1     1   2    4     4  2001
1251   1251       104          1 2001   140    690     43   2     1   3    4     4  2001
1252   1252       125          1 2001   140    754     44   2     1   3    4     4  2001
1253   1253       128          1 2001   140   1270     55   2     4   7    4     4  2001
1254   1254       257          1 2001   228    370     35   2     1   3    4     4  2001
```

Hypotheses:

H_0: a specified subset of β's = 0

H_1: same subset of β's <> 0

Test Statistic:

$G^2 = -2\ln[L(R)/L(F)] = -2\ln(L(R) - 2\ln(L(F)) = dev_F - dev_R$

< where dev_R & dev_R are residual deviances calculated from the Maximum Likelihood fits of FM & RM

Sampling Distribution:

If Assumptions hold and H_0 is true,

then $G^2 \sim \chi^2(df_F - df_R)$ **< where df_F & df_R are degrees of freedom of FM & RM**

Probability:

P = 1 - pchisq(Δdef,Δdf) **< where Δdev & Δdf 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):

$dev_F := 1422.7$ $dev_R := 1474.7$

$\Delta dev := dev_R - dev_F$ $\Delta dev = 52$

$df_F := 1178$ $df_R := 1184$

$\Delta df := df_R - df_F$ $\Delta df = 6$

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

$P = 1.865 \times 10^{-9}$

For factor Length (green):

$dev_F := 1422.7$ $dev_R := 1424.9$

$\Delta dev := dev_R - dev_F$ $\Delta dev = 2.2$

$df_F := 1178$ $df_R := 1179$

$\Delta df := df_R - df_F$ $\Delta df = 1$

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

$P = 0.138$

Serial report of Likelihood Ratio Tests:

> 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 P(>|Chi|)
NULL                         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:

> drop1(FM5,test="Chisq")

```
Single term deletions

Model:
Prevalence ~ fArea * fYear + Length
            Df Deviance    AIC    LRT   Pr(Chi)
<none>           1422.7 1448.7
Length       1   1424.9 1448.9  2.231    0.1352
fArea:fYear  6   1474.7 1488.7 52.031 1.838e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
RM1=glm(Prevalence~fArea+fYear+Length,family=binomial,data=C)
anova(RM1,FM5,test="Chisq")

RM2=glm(Prevalence~fArea*fYear,family=binomial,data=C)
anova(RM2,FM5,test="Chisq")
```

Explicit Test for Interaction:

> anova(RM1,FM5,test="Chisq")

```
Analysis of Deviance Table

Model 1: Prevalence ~ fArea + fYear + Length
Model 2: Prevalence ~ fArea * fYear + Length
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      1184     1474.7
2      1178     1422.7  6   52.031 1.838e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Explicit test for factor Length:

> anova(RM2,FM5,test="Chisq")

```
Analysis of Deviance Table

Model 1: Prevalence ~ fArea * fYear
Model 2: Prevalence ~ fArea * fYear + Length
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      1179     1424.9
2      1178     1422.7  1   2.2313    0.1352
```

summary(FM5)

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

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

> 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
```