Count data page 1

Count data

1. Estimating, testing proportions

100 seeds, 45 germinate. We estimate probability p that a plant will germinate to be 0.45 for this population. Is a 50% germination rate a reasonable possibility? Pr{ 45 or less germinate in 100 trials if p=0.5 } = ???

Binomial:

n independent trials

Each trial results in success or failure

p=probability of success same on every trial

X= observed number of successes in n trials

Pr{X=r} = n!/[r!(n-r)!] pr(1-p)n-r

Recall: 0!=1, 1!=1, 2!=2\*1=2, 3!=3\*2\*1=6, 4!= 4\*3\*2\*1=24 etc.

Example:

soft drinks - 2 cups, one has artificial sweetener, 1 sugar

Task: guess which has artificial sweetener

6 people, 5 get it right. Can they tell the difference?

H0: p=0.5 (indistinguishable)

Pr{ 5 or 6 right if p=0.5} = [6!/(5!1!)(0.5)5(0.5)] + [6!/(6!0!)(0.5)6] = 6/64 + 1/64 = 7/64 = 0.11 > 0.05

not significant.

Logistic Regression

Idea: p=probability of germinating = function of some variables (maybe temperature, moisture, or both).

Example:

Temperatures

Germinating 70 73 78 64 67 71 77 85 82

Not germ. 50 63 58 72 67 75

Temperature

Idea: Regress Germ (0 or 1) on Temperature:

Analysis of Variance

Sum of Mean

Source DF Squares Square F Value Pr > F

Model 1 1.09755 1.09755 5.70 0.0328

Error 13 2.50245 0.19250

Corrected Total 14 3.60000

Parameter Estimates

Parameter Standard

Variable DF Estimate Error t Value Pr > |t|

Intercept 1 -1.55013 0.90756 -1.71 0.1114

temperature 1 0.03066 0.01284 2.39 0.0328

Predicted response (= estimated probability of germination)

Y p temperature

1 0.59591 70

1 0.68789 73

1 0.84117 78

1 0.41197 64

1 0.50394 67

1 0.62657 71

1 0.81052 77

1 1.05578 85 <--

1 0.96380 82

0 -0.01724 50 <--

0 0.38131 63

0 0.22802 58

0 0.65723 72

0 0.50394 67

0 0.74920 75

\* Normal residuals ?

\* Reasonable predicted probabilities? 1.0558? -0.0172 ?

Better idea: Map 0<p<1 into L = logit =ln(p/(1-p)) then model expected value of L as

E{L} =  + (temperature)

or equivalently,

E{L } =  + X where X=(temperature-70) (approximate centering for nicer graphics)

"Likelihood" = probability of sample = p(1-p)ppp(1-p)p ...p

Use p for germinated, 1-p for not germinated. Values of p are all different, depending on temperature.

Substitute p= e + X/(1+e + X ) and thus 1-p = 1/(1+e + X )

"Maximum Likelihood Estimates"

Likelihood = [e + (70-70) /(1+e + (70-70) ) ][ e + (73-70)/(1+e + (73-70) )]… [ 1/(1+e + (75-70) )]

Graph Likelihood vs.  and  and find values  and  of that maximize.





For statistical reasons we often plot -2*ln*(likelihood). Here this quantity has been truncated at a value such that the elliptical region forms a 95% confidence region for the true (,)





Maximum likelihood theory also gives standard errors (large sample approximations) . Use "PROC LOGISTIC" to fit or "PROC CATMOD" . We get

Pr{ Germinate } = eL /(1+eL )=1/(1+e-L) where L =.4961 + 0.1821\*X and X=temperature-70.

Title " ";

**Data** seeds; Input Germ $ **1**-**3** n @; Y=(Germ="Yes");

If Germ=" " then Y=**.**;

do i=**1** to n; input temp @; X =temp-**70**; output; end;

cards;

Yes 9 64 67 70 71 73 77 78 82 85

No 6 50 58 63 67 72 75

23 46 48 50 52 54 56 58 60 62 64 66 68

70 72 74 76 78 80 82 84 86 88 90

**PROC** **LOGISTIC** data=seeds order=data;

model germ=X / itprint ctable pprob=**.6923**;

output out=out1 predicted=p xbeta=logit;

**proc sort** data=out1; by X;

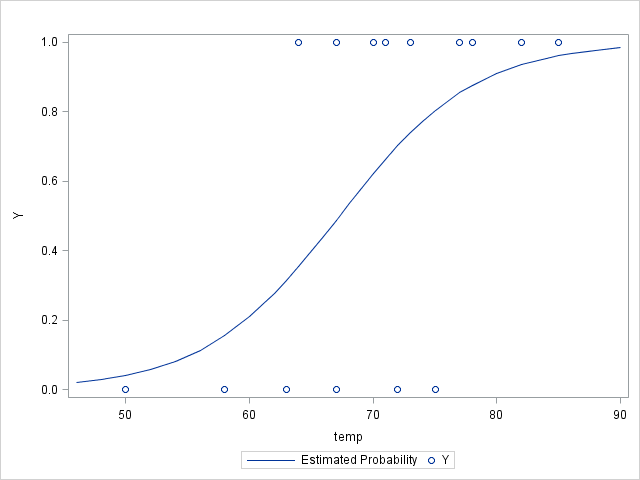
**proc** **sgplot**;

series X=temp Y=p;

scatter X=temp Y=Y;

**run**;

**quit**;



The LOGISTIC Procedure

Model Information

Data Set WORK.SEEDS

Response Variable Germ

Number of Response Levels 2

Model binary logit

Optimization Technique Fisher's scoring

Number of Observations Read 38

Number of Observations Used 15

Response Profile

Ordered Total

Value Germ Frequency

1 Yes 9

2 No 6

Probability modeled is Germ='Yes'.

NOTE: 23 observations were deleted due to missing values for the

response or explanatory variables.

Maximum Likelihood Iteration History

Iter Ridge -2 Log L Intercept X

0 0 20.190350 0.405465 0

1 0 15.205626 0.388433 0.127740

2 0 14.878609 0.478775 0.171150

3 0 14.866742 0.495409 0.181644

4 0 14.866718 0.496105 0.182141

Last Change in -2 Log L 0.0000235249

Last Evaluation of Gradient

Intercept X

1.5459414E-6 0.0000962751

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Intercept

Intercept and

Criterion Only Covariates

AIC 22.190 18.867

SC 22.898 20.283

-2 Log L 20.190 14.867

Testing Global Null Hypothesis: BETA=0

Test Chi-Square DF Pr > ChiSq

Likelihood Ratio 5.3236 1 0.0210

Score 4.5731 1 0.0325

Wald 3.1024 1 0.0782

Analysis of Maximum Likelihood Estimates

Standard Wald

Parameter DF Estimate Error Chi-Square Pr > ChiSq

Intercept 1 0.4961 0.6413 0.5984 0.4392

X 1 0.1821 0.1034 3.1024 0.0782

Odds Ratio Estimates

Point 95% Wald

Effect Estimate Confidence Limits

X 1.200 0.980 1.469

Association of Predicted Probabilities and Observed Responses

Percent Concordant 79.6 Somers' D 0.611

Percent Discordant 18.5 Gamma 0.623

Percent Tied 1.9 Tau-a 0.314

Pairs 54 c 0.806

Classification Table

Correct Incorrect Percentages

Prob Non- Non- Sensi- Speci- False False

Level Event Event Event Event Correct tivity ficity POS NEG

0.692 5 4 2 4 60.0 55.6 66.7 28.6 50.

ctable pprob=.6923;

Print out a "classification table" using "prior probability" 0.6923.In SAS, each observation is removed, the logistic regression refit, and if the new estimated p exceeds 0.6823, a 1 is predicted,otherwise 0.

I used 0.6823 because one observation is classified differently with this careful method than it would have been if just the full data model were used.

Odds: p/(1-p) L is ln(odds) = ln(p/(1-p)).

At X, let’s say L is L0 = + X and p becomes p0 =exp(L0)/(1+exp(L0)) odds: exp(L0)

At X+1, L is L1 = + X+1) and p becomes p1 =exp(L1)/(1+exp(L1)) odds: exp(L1)

Each L is a logarithm of odds so their difference () is the logarithm of the odds RATIO. (remember that ln(A/B)=ln(A)-ln(B) ). Thus we see that

(+ X+1))(+ X)) =

is the log of the odds RATIO and exp() is the odds ratio itself. For numbers Y near 0, like Y= 0.18 for example, exp(Y) is approximately 1+Y so if the log odds ratio is 0.18 then approximately the odds becomes 1.18.

Small example from notes organized in a table :

Actual

Non Event

Decision Event

Non Event 4 4 (8)

Event 2 5 (7)

(6) (9)

Your DECISION is either correct or not correct.

Number of correct events (say event and be correct) is 5 in lower right corner

Number of incorrect events

Does this mean that you *said* it was an event and were wrong OR

Does this mean that it *was* an event and you said it was not?

It is the first – you said event and were incorrect twice (2 of 7 declarations of event were wrong)

It is the number of event *decisions* that were wrong. Lower left corner.

Number of correct non events 4

Number of incorrect non events 4

(you said non event when it was really an event 4 times)

Out of 8 declarations of non event, 4 were incorrect.

Classification Table

Correct Incorrect Percentages

Prob Non- Non- Sensi- Speci- False False

Level Event Event Event Event Correct tivity ficity POS NEG

0.692 5 4 2 4 60.0 55.6 66.7 28.6 50.

Actual

Non Event

Decision Event

Non Event 4 4 (8)

Event 2 5 (7)

(6) (9)

Sensitivity Pr{say event given that it is an event} which we denote as

Pr{say event | event}=Pr{say event and it was an event}/Pr{true event} = 5/9 =0.556

Specificity Pr{say non event | non event} = 4/6 = 0.667

Proportion False positives = Proportion of positive decisions that are false = Pr{non event | say event} = 2/7=.286

Proportion False negatives = Proportion of negative decisions that are false = 4(upper right) divided by 8 = 4/8 = 0.50