Nondetects And Data Analysis:
Logistic Regression with NDs

Dennis R. Helsel, Ph.D
Practical Stats

Binary Logistic Regression

Predicts: the probability of getting a 1
= (1 – prob of getting a 0)

using a familiar equation:

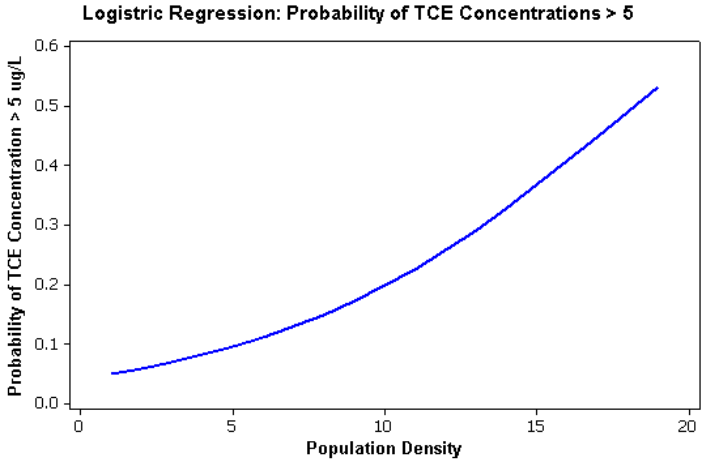
$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots$$

Application to censored data:

Predict the probability of a concentration above a reporting limit (1, a "hit") as a function of one or more X variables

2

Logistic Regression Equation



3

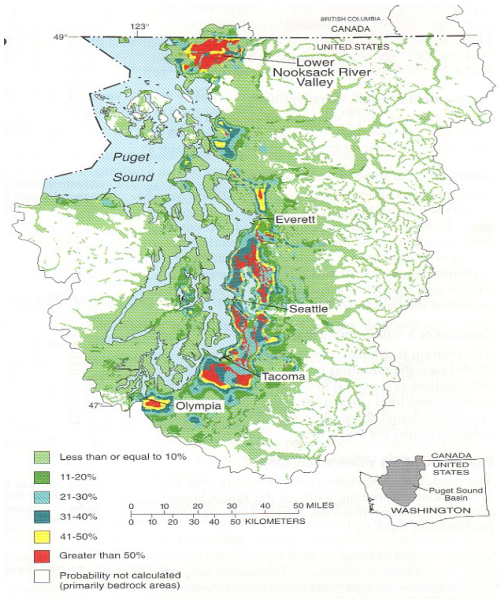
One Use of Logistic Regression

Compute from the equation over a grid of points, contour and map the probability of detecting TCE, nitrate or other contaminants.

Modeled as a function of geology, land use, and well depth.

From Erwin and Tesoriero (1997)

USGS Fact Sheet 061-97



4

Binary Logistic Regression

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots$$

Where Y is a "logit", the log of the odds:

$$Y = \log\left[\frac{prob(Event)}{1 - prob(Event)}\right]$$

5

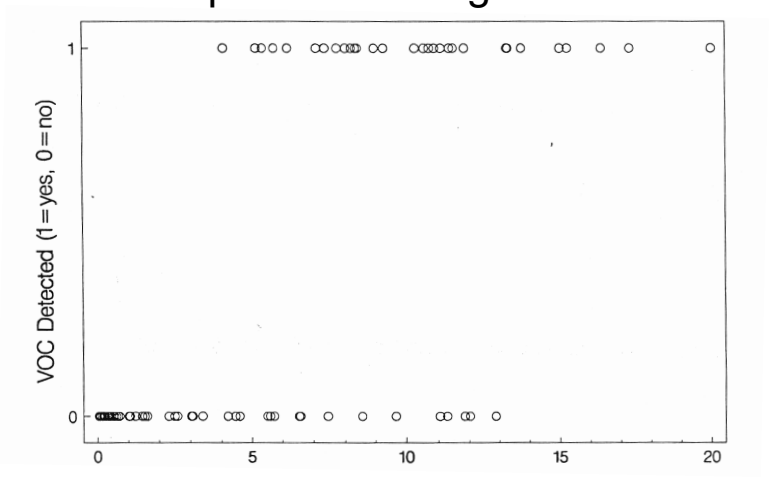
Binary Logistic Regression

Applied to data with nondetects, and Y=1 is a concentration at or above the DL:

$$Y = \log\left[\frac{\text{prob}(\text{Detect})}{\text{prob}(\text{Nondetect})}\right]$$

6

Example: VOCs in ground water



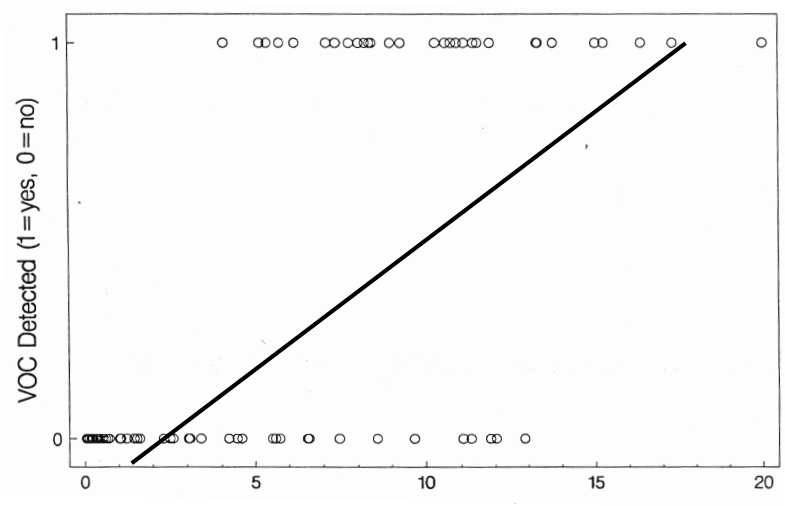
The Y variable has only 0s and 1s. How do we get a probability?

7

OLS regression would not fit these data well

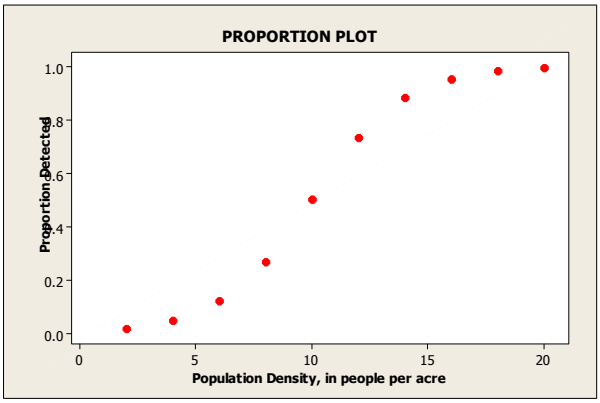
Residuals would look nothing like a normal distribution, as required by OLS

Predicted values could go below 0 or above 1



8

1950s: Split into categories and compute proportion of 1s (detects)



Logistic regression solves this problem in a better way

9

The S-shaped curve has the formula

$$\log\left[\frac{\text{Prob}(y=1)}{1-\text{Prob}(y=1)}\right] = \beta_0 + \beta_1 X$$

or

$$\text{Prob [y=1]} = \frac{\exp(\beta_0 + \beta_1 X)}{[1+\exp(\beta_0 + \beta_1 X)]}$$

where *X* may be a vector of explanatory variables

10

Binary Logistic Regression

$$\log\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 X$$

- Solve iteratively for the best estimates of the coefficients β_0 and β_1 by Maximum Likelihood Estimation.
- No assumption of normality of residuals or constant variance required. It is still a parametric model, with an S shape. It assumes the X's have a linear relationship to the logit Y scale.

11

Log likelihood function

$$\text{lnL} = \sum_{i=1}^{n} \left(y_i \ln[\hat{p}_i] + (1 - y_i) \ln[1 - \hat{p}_i] \right)$$

when y = 1 when y = 0

where $\hat{p}$ are the estimated probabilities, and y_i are the observed values (1 or 0)

lnL is a negative number, which is maximized (brought close to 0).

12

Example TCELogReg.rda

```
> load(TCELogReg)
> attach(TCELogReg)
```

Again, TCE concentrations in ground water. There were 3 reporting limits, so the highest at 5 ug/L was used.

A column "GT5" with 1 if TCE ≥ 5 ug/L, 0 otherwise has been added for use in logistic regression.

Explanatory variables: population density, %industrial land use, and depth to water.

13

Logistic regression in R: glm

```
> GLM.1 <- glm(GT5 ~ DEPTH + PctIND + POPDEN, family=binomial(logit) )
> summary(GLM.1)
Coefficients:
             Estimate Std. Error z value    Pr(>|z|)
(Intercept) -2.851419   0.544281  -5.239 0.000000162 ***
DEPTH       -0.001336   0.001701  -0.785     0.43223
PctIND       0.013361   0.040892   0.327     0.74387
POPDEN       0.153922   0.051981   2.961     0.00307 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 182.69  on 246  degrees of freedom
Residual deviance: 169.82  on 243  degrees of freedom
AIC: 177.82
```

General Linear Model (glm)

14

Measures of Error

R reports: `Residual deviance: 169.82`
for the DEPTH + PctIND + POPDEN model

Residual deviance = D_{model} = -2lnL . It's a measure of error.

From this we can compute lnL, the Log-Likelihood was = -84.91

- lnL or Deviance D in themselves provide little information to the user.
- Computed from the sum of errors for each observation, so there's no "good" or "bad" Deviance. The magnitude depends on the number of obs.
- The difference in Deviance (error) between models determines whether one model is better than another

15

Overall Likelihood ratio test G

The overall likelihood ratio test compares errors for a model to errors for a null model (with no x variables). This is not given by default in the output of glm, so we'll first compute the null model, and then use the anova command to perform the chi-square test of difference in models.

$$G = (\text{Deviance}_{\text{null}} - \text{Deviance}_{\text{model}})$$

```
> GLM.0 <- glm(GT5 ~ 1, family=binomial(logit))
> anova(GLM.0, GLM.1, test="Chisq")
Analysis of Deviance Table
Model 1: GT5 ~ 1
Model 2: GT5 ~ DEPTH + PctIND + POPDEN
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       246     182.69
2       243     169.82  3   12.872 0.004922 **
```

- The test statistic has a small p-value. Rejecting the null hypothesis means that the GLM.1 model is better than no model at all.
- With no x variables in GLM.0, the best guess for the probability of a 1 is the observed proportion of events: (# 1's / n). By rejecting this it states that a better prediction of %1s can be computed with the X variable values.

16

The null model

Log likelihood for the null (no x variable) model (L_0) L_{null} = -91.35

D_{null} = -2*(-91.35)= 182.69

Rejecting this null hypothesis says that using these x variables, a better prediction of the probability of a 1 is possible. But is this the best model? Should all 3 variables be in the model?

For that we need either AIC, or for nested models the partial likelihood or partial Wald's tests.

17

Model Selection
Step 1. Check for multicolinearity

Check for multicolinearity using VIFs

```
> vif(GLM.1)

Coefficients:
   DEPTH    PctIND    POPDEN
1.104578 1.020438 1.125963
```

No multcollinearity found. The reported p-values can therefore be trusted.

For the definition of the VIF, see the multiple regression lecture.

18

Step 2a. Test whether to transform an X variable

A quick test to determine whether the relationship between the log(odds) and the X variables are linear is using the residualPlots command in the car package. You'll get plots (unless you specify plot = FALSE) and a test, where the null hypothesis is that the X variable is sufficiently linear and no transform is needed. Small p-values indicate you should transform that X variable.

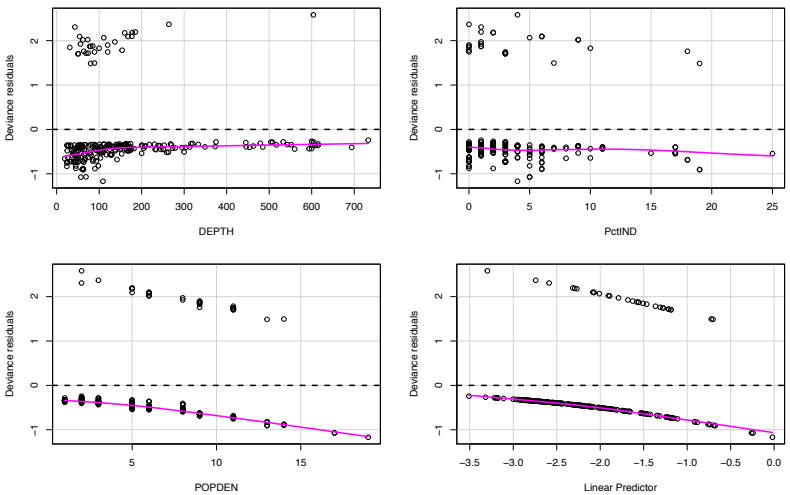
```
> residualPlots(GLM.1, type = "deviance")
       Test stat   Pr(>|Test stat|)
DEPTH     0.8353           0.360732
PctIND    0.2380           0.625636
POPDEN    8.6565           0.003259 **
```

The small p-value for POPDEN indicates that it should be transformed

19

Residuals Plots

It would be nice if these residuals plots would function like crPlots. However, they are more difficult to interpret. The two groups are because (observed - predicted) will be negative for the observed Y=0 data, and positive for the Y=1 data. The smooth is staying within the Y=0 data, indicating that the regression may not predict the Y=1 data well.



20

Step 2b. Test whether to transform an X variable

Box and Tidwell (1962) provided a simple method to determine whether to transform an X variable in a regression model. This pre-dates component+residuals plots (crPlots) now used in OLS. For logistic regression, some attempts at crPlots have been made but aren't yet satisfactory. The Box-Tidwell (BT) procedure is still useful.

1. Construct a variable $X_c = X*\log(X)$ to detect curvature in X, adding it to the regression
2. If the slope b_c on X_c is significant, use X^t instead of X in a subsequent regression model, where t is the power transform coefficient. Note that t=0 for a log transform.
3. A guide to the appropriate transformation to use is $t = 1 + (b_c / b_X)$ where b_c is the slope on X_c and b_X is the slope on X in the regression model. Pick a simple power transform near the value of t, and use the model with lowest AIC.

21

Step 2b. Test whether to transform an X variable

Construct the Box-Tidwell variables to test curvature:

```
> TCELogReg$BT.depth <- DEPTH*log(DEPTH)
> TCELogReg$BT.popden <- POPDEN*log(POPDEN)
```

PctIND is a percentage, so don't try to transform it. Leave its units alone.

```
> GLM.2 <- glm(GT5 ~ DEPTH + PctIND + POPDEN + BT.depth + BT.popden,
   family=binomial(logit), data=TCELogReg)
```

22

Test whether to transform an X variable

```
>  summary(GLM.2)
Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.659383   2.004852  -3.820 0.000133 ***
DEPTH        0.040528   0.030652   1.322 0.186113
PctIND       0.009287   0.041161   0.226 0.821495
POPDEN       1.834279   0.696051   2.635 0.008407 **
BT.depth    -0.006564   0.004800  -1.367 0.171469
BT.popden   -0.549212   0.227635  -2.413 0.015835 *

    Null deviance: 182.69  on 246  degrees of freedom
Residual deviance: 159.62  on 241  degrees of freedom
AIC: 171.62
```

Not significant. Don't transform DEPTH

Is significant, so use POPDEN^t instead of POPDEN as the X variable

$t = 1 + (b_c / b_X) = 1 + (-0.549/ 1.83) \cong 0.7$ Try square root or log of POPDEN

23

Which transformation to use?

the one with lower AIC

```
> sqrt.popden <- sqrt(POPDEN)
>  GLM.3 <- glm(GT5 ~ DEPTH + PctIND + sqrt.popden,
     family=binomial(logit))
>  summary(GLM.3)
            Estimate Std. Error z value   Pr(>|z|)
(Intercept) -4.055547   0.860918  -4.711 0.00000247 ***
DEPTH       -0.001265   0.001686  -0.750    0.45309
PctIND       0.009692   0.041107   0.236    0.81361
sqrt.popden  0.900216   0.280947   3.204    0.00135 **

    Null deviance: 182.69  on 246  degrees of freedom
Residual deviance: 167.19  on 243  degrees of freedom
AIC: 175.19
```

```
> TCELogReg$lnPOPDEN <- log(POPDEN)
>  GLM.4 <- glm(GT5 ~ DEPTH + PctIND + lnPOPDEN,
     family=binomial(logit))
>  summary(GLM.4)
Estimate Std. Error z value   Pr(>|z|)
(Intercept) -3.787534   0.800436  -4.732 0.00000222 ***
DEPTH       -0.001360   0.001684  -0.808    0.41916
PctIND       0.007113   0.041091   0.173    0.86257
lnPOPDEN     1.150996   0.357714   3.218    0.00129 **

    Null deviance: 182.69  on 246  degrees of freedom
Residual deviance: 165.10  on 243  degrees of freedom
AIC: 173.1
```

better 2-variable model

24

AIC
A cost-benefit analysis

$$AIC = 2p + G$$

2 * # parameters (including intercept).

Improves fit, but decreases degrees of freedom.

Cost

Unexplained noise, as expressed by the overall likelihood ratio test. Reducing it is the

Benefit

25

Step 3. Which X variables to keep in the model?

To compare nested models and pick the best one, use partial likelihood ratio tests:

$$G_{partial} = 2\,(\ln L_c - \ln L_s) = D_s - D_c$$

where $\ln L_c$ and D_c are for the more complex model (more X variables), and $\ln L_S$ and D_s are for the simpler model.

Compare $G_{partial}$ to a chi-square distribution with degrees of freedom equal to the number of additional variables in the more complex model.

Partial tests for models that differ by only 1 variable are reported in the regression output section, to determine each variable's effect.

26

Partial Tests

These tests compare two nested models to select the better one.

```
>  GLM.4 <- glm(GT5 ~ DEPTH + PctIND + lnPOPDEN, family=binomial(logit))
>  summary(GLM.4)

             Estimate Std. Error z value   Pr(>|z|)
(Intercept) -3.787534   0.800436  -4.732 0.00000222 ***
DEPTH       -0.001360   0.001684  -0.808    0.41916
PctIND       0.007113   0.041091   0.173    0.86257
lnPOPDEN     1.150996   0.357714   3.218    0.00129 **

    Null deviance: 182.69  on 246  degrees of freedom
Residual deviance: 165.10  on 243  degrees of freedom
AIC: 173.1
```

The p-value for PctIND compares this 3-variable model to a 2-variable model with only DEPTH and lnPOPDEN as the variables. Since the p-value is large, do not reject the null hypothesis that the slope coeff. for PctIND equals 0, and provides no explanatory power. Drop this variable from the model and check whether the AIC has improved. It should.

27

Partial Tests

```
>  GLM.5 <- glm(GT5 ~ DEPTH + lnPOPDEN, family=binomial(logit))
>  summary(GLM.5)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -3.774908   0.797359  -4.734 0.0000022 ***
DEPTH       -0.001350   0.001685  -0.801    0.4230
lnPOPDEN     1.158548   0.355085   3.263    0.0011 **

Null deviance: 182.69  on 246  degrees of freedom
Residual deviance: 165.13  on 244  degrees of freedom
AIC: 171.13
```

Three-variable model (GLM.4) with PctIND had an AIC of: `AIC: 173.1`
The AIC is smaller for the 2-variable model. It is better than the 3-variable model.

DEPTH does not have a significant p-value. How does this compare to a 1-variable model?

28

Partial Tests -- Compare 2 nested models

```
> anova (GLM.4, GLM.5, test = "Chisq")
Analysis of Deviance Table

Model 1: GT5 ~ lnPOPDEN + PctIND + DEPTH
Model 2: GT5 ~ DEPTH + lnPOPDEN
  Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
1       243     165.10
2       244     165.13 -1 -0.029573   0.8635
```

The null hypothesis is that the simpler model is better -- the additional variable(s) not in the simpler model do not affect the deviance very much, and should be dropped. The p-value of 0.8635 says to not reject this null hypothesis. Therefore the simpler 2-variable model is better - nothing much is lost by dropping out PctIND. This agrees with the AIC values.

29

Partial Tests

```
>  GLM.6 <- glm(GT5 ~ lnPOPDEN, family=binomial(logit))
>  summary(GLM.6)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)  -4.1362     0.7030  -5.884 0.00000000401 ***
log(POPDEN)   1.2515     0.3497   3.579      0.000345 ***

    Null deviance: 182.69  on 246  degrees of freedom
Residual deviance: 165.85  on 245  degrees of freedom
AIC: 169.85
```

Two-variable model (GLM.5) had an AIC of: `AIC: 171.13`

This AIC for this 1-variable model is smaller. If the p-value on the variable that was dropped (DEPTH) had been 0.10 or less, I'd choose the 2-variable model. It was not, so this 1-variable model is better.

30

Compare 2 nested models

```
> anova(GLM.6, GLM.4, test="Chisq")
Analysis of Deviance Table

Model 1: GT5 ~ lnPOPDEN
Model 2: GT5 ~ DEPTH + PctIND + lnPOPDEN
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       245     165.85
2       243     165.10  2  0.74794    0.688
```

The null hypothesis is that both slope coefficients for the variables not in the simpler model equal 0, and the variables should be dropped. The p-value of 0.688 says to not reject this null hypothesis. Therefore the simpler 1-variable model is better - nothing much is lost by dropping the other two variables out. This agrees with the AIC values.

31

How to compare non-nested models?

Choose the model with the smallest AIC

Model	2p	-2lnL	AIC
lnPOPDEN	4	165.85	169.85 ←
Depth	4	179.53	183.53
%indlu	4	181.97	185.97
lnPOPDEN, depth	6	165.93	171.93
lnPOPDEN, %indlu	6	165.83	171.83
Depth, %indlu	6	178.91	184.91
lnPOPDEN, depth, %indlu	8	165.10	173.10
POPDEN, depth, %indlu	8	169.82	177.82

32

Automated Model Selection

After first deciding whether or not to transform any of the X variables, you may use the bestglm function in the bestglm package. It minimizes the BIC (default), AIC, or other criterion to determine which set of explanatory variables has the most ability to correctly predict the log odds.

NOTE: bestglm requires a dataset whose columns are first all of the X variables, with the last column being the Y 0/1 variable). Create that with the data.frame command.

```
> TCEbest <- data.frame(PctIND, DEPTH, lnPOPDEN, GT5)
> install.packages("bestglm")
> library("bestglm")
> bestglm(TCEbest, family = binomial(logit), IC = "AIC")
Morgan-Tatar search since family is non-gaussian.
AIC Best Model:
             Estimate Std. Error   z value          Pr(>|z|)
(Intercept) -4.136235  0.7029967 -5.883719 0.000000004011498
lnPOPDEN     1.251458  0.3496746  3.578921 0.000345015951851
```

33

Step 4. Interpreting Coefficients

What does a slope of 1.2515 for log(POPDEN) mean?

```
             Estimate Std. Error z value      Pr(>|z|)
(Intercept)  -4.1362     0.7030  -5.884 0.00000000401 ***
log(POPDEN)   1.2515     0.3497   3.579      0.000345 ***
```

- The slope has a positive sign, so the probability of TCE ≥ 5 increases as log(POPDEN) increases, and so as POPDEN itself increases.
- For a unit increase in log(POPDEN) the log-odds increases by 1.25. This corresponds to a ($e^{1.25}$) = 3.49 multiplier (called the 'odds ratio') to the odds [p /(1-p)] . This can be printed out by the computation:

```
>  exp(coef(GLM.6))  # Exponentiated coefficients ("odds ratios")
(Intercept)    lnPOPDEN
 0.01598292  3.49543425
```

34

Interpreting Negative Coefficients

Suppose the 2-variable model GLM.5 had been chosen. What would a negative coefficient for DEPTH mean?

```
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -3.774908   0.797359  -4.734 0.0000022 ***
DEPTH       -0.001350   0.001685  -0.801    0.4230
lnPOPDEN     1.158548   0.355085   3.263    0.0011 **
```

```
>  exp(coef(GLM.5))  # odds ratios for a negative slope
(Intercept)       DEPTH    lnPOPDEN
  0.0229392   0.9986511   3.1853043
```

The slope of -0.0013 says the probability of TCE ≥ 5 decreases as depth increases. For a unit increase in depth, there is an ($e^{-0.0013}$) = 0.9987 multiplier, or a 0.13% decrease, in the odds [p /(1-p)] for TCE ≥ 5.

35

Step 5. Evaluating the chosen model

a. Everyone wants an r^2. Likelihood r^2 are available for logistic regression. As with OLS, they are not helpful for deciding between models with different numbers of X variables. The model with more X variables will always have a higher r^2. They are helpful for deciding which power transformation of an X variable to use. For GT5 ~ POPDEN, r2 = 0.091 while for GT5 ~ lnPOPDEN, r2 = 0.126. Therefore the log units are better.

b. Graphs

c. Measure Predictive Ability
Pair up the data. Will be n(n-1)/2 = M combinations of data pairs

36

5a. Nagelkerke r^2

There are several "pseudo-r²" methods. See the correlation video, and https://statisticalhorizons.com/r2logistic

For logistic regression, these tend to be much lower than for ordinary least-squares regression. One of the most recommended versions is the Nagelkerke (or rescaled likelihood) r-squared:

$$R_N^2 = \frac{1-\exp(\frac{-Gmode_l}{n})}{1-\exp(D_0/n)}$$

where n is the number of observations, G_{model} is the model likelihood ratio test statistic, and D_0 is the Deviance of the null model. So R_N^2 compares the model to the null model in a "proportion of likelihood explained" type of statistic, though not a proportion of variance explained scale as in least-squares regression. See

Nagelkerke, N. J. D. 1991. A note on the general definition of the coefficient of determination. Biometrika, 78:3, 691-692.

For the one-variable log(POPDEN) model:

$$R_N^2 = \frac{1-exp\left(-\frac{16.84}{247}\right)}{1-exp\left(-\frac{182.69}{247}\right)} = 0.126$$

A small amount of variation has been explained by the model. There are probably other X variables not currently in this dataset that should be added to this model before its use for prediction is reliable.

37

5a. Nagelkerke r^2

Several model evaluation statistics are computed using the logistic regression command in the rms package (yes, another package to install and load)

```
> install.packages("rms")
```

click the box next to rms to load it. Then,

```
> lrm6 <- lrm(GT5~lnPOPDEN)    # logistic regression model -- lrm
> lrm6
Logistic Regression Model

                      Model Likelihood     Discrimination    Rank Discrim.
                         Ratio Test           Indexes           Indexes

 Obs           247    LR chi2      16.84    R2        0.126    C       0.726
  0            217    d.f.             1    g         1.097    Dxy     0.452
  1             30    Pr(> chi2) <0.0001    gr        2.996    gamma   0.500
 max |deriv| 6e-11                          gp        0.094    tau-a   0.097
                                            Brier     0.101

          Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept -4.1362 0.7030 -5.88  <0.0001
 lnPOPDEN   1.2515 0.3497  3.58  0.0003
```

Nagelkerke (rescaled likelihood) r^2

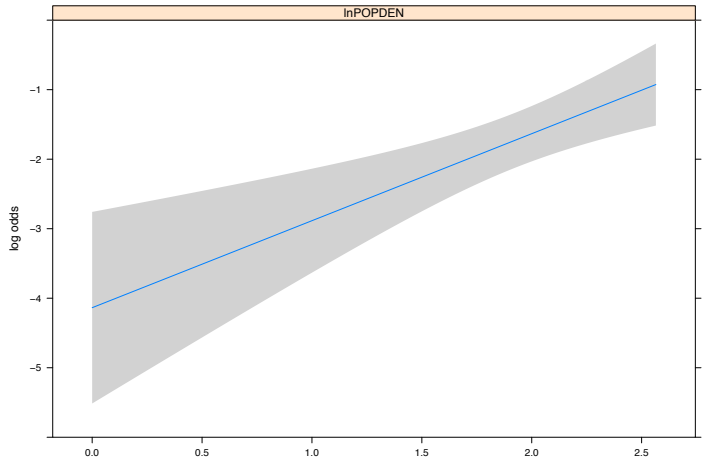
38

Step 5b. Plots of model effects

Plot the predicted probability against one or more X variables

```
> d6 = datadist(lnPOPDEN, GT5)  # 2 lines to tell R what original variables were used
> options(datadist = "d6")
> plot(Predict(lrm6))
```

Gray area shows the 95% confidence bands for the model.

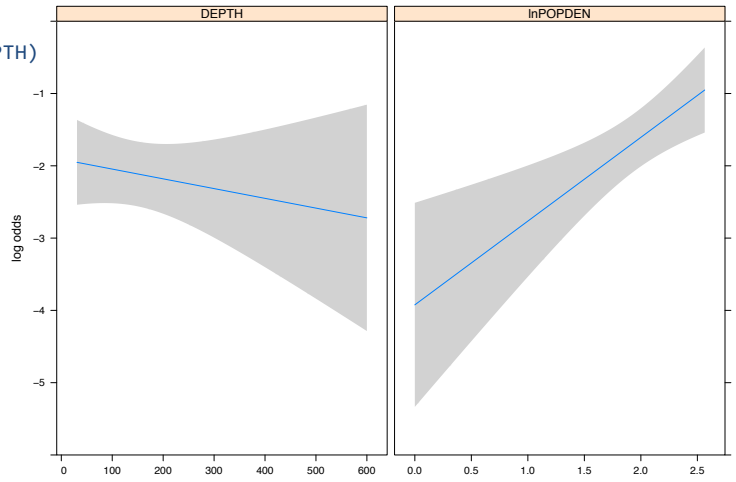


39

Step 5b. Plots of model effects

If we recreate the 2-variable GLM.5 model using the lrm command:

```
> d5 = datadist(lnPOPDEN, DEPTH, GT5)
> options(datadist = "d5")
> lrm5 <- lrm(GT5 ~ lnPOPDEN + DEPTH)
> plot(Predict(lrm5))
```

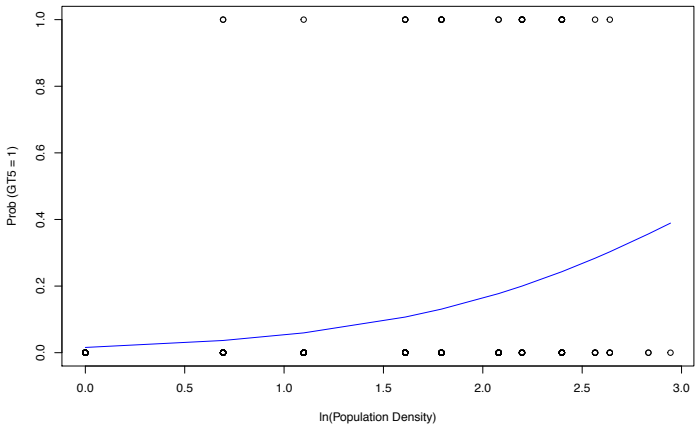


40

Step 5b. Plot the Predicted Probabilities

```
> pof1 <- exp(GLM.6$linear.predictors) / (1 + exp(GLM.6$linear.predictors))
> pred.6df <- data.frame(lnPOPDEN, GT5, pof1)
> psort <- order(lnPOPDEN)
> plot(lnPOPDEN, GT5,
    xlab = "ln(Population Density)",
    ylab = "Prob (GT5 = 1)")
> lines(lnPOPDEN[psort],
    pof1[psort], col = "blue")
```

Note that the predicted probability never goes above 0.4. Assuming a 0.5 cutoff, the model isn't predicting a detection above 5 ug/L for any observation. That is its biggest problem.



41

Step 5c. Measure predictive ability

- The goal is to determine how well the predicted probabilities match the observed data (0 or 1).
- There are n(n-1)/2 = M combinations of data pairs. Many of the measures look at the change in prediction from an observed 0 to an observed 1, or from an observed 1 to a 0.
- All but one tau-a drops out data pairs where both values are observed 0s, or observed 1s. M' is the number of pairs with different observed values. M' < M
- Do the predicted probabilities (of a 1) increase when going from an observed 0 to a 1? This is a concordant result. C = number of concordant results.
- If the predicted probability (of a 1) decreases when going from an observed 0 to a 1, this is a discordant result. D = number of discordant results.
- If the probability (of a 1) does not change when going from an observed 0 to a 1, this is a tie. T = number of tied results.
- M' = C + D + T

42

Step 5c. Measure predictive ability

Measures of Association		For GLM.6	Comment
Kendall's tau-a	$= \frac{C-D}{n(n-1)/2}$	= 0.097	Is always small due to many 1-1 and 0-0 ties. Not helpful
Kendall's tau-b (3 most useful)	$= \frac{C-D}{\sqrt{(C+D+T_x)(C+D+T_y)}}$	= 0.223	tau adjusted for ties. Compute w/ cor.test. Higher is better.
Somer's Dxy (3 most useful)	$= 2AUC - 1$ $= \frac{C-D}{C+D+T}$	= 0.452	Higher is better. Ignores 0-0 and 1-1 pairs.
AUC	$= \frac{C+0.5T}{C+D+T}$	= 0.726	Area under ROC curve. Higher is better. If 0.8 <= C < 0.9, excellent discrimination. If 0.7 <= C < 0.8, acceptable discrimination
Brier score	$= \frac{\sum_{i=1}^{n}(\hat{p}_i - y_i)^2}{n}$	= 0.103	Mean squared error of prediction. Lower is better.

None of these tell you that the predictions for GLM.6 are wrong for all data above 5 ug/L

43

Step 5c. Measure predictive ability

```
> lrm6 <- lrm(GT5~lnPOPDEN)    # logistic regression model -- lrm
> lrm6
Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes          Indexes
 Obs           247    LR chi2      16.84    R2       0.126    C       0.726
  0            217    d.f.             1    g        1.097    Dxy     0.452
  1             30    Pr(> chi2) <0.0001    gr       2.996    gamma   0.500
 max |deriv| 6e-11                          gp       0.094    tau-a   0.097
                                            Brier    0.101

           Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept -4.1362 0.7030 -5.88  <0.0001
 lnPOPDEN   1.2515 0.3497  3.58  0.0003

> cor.test(GLM.6$linear.predictors, GT5, method = "kendall")
            Kendall's rank correlation tau
z = 4.0468, p-value = 5.193e-05
alternative hypothesis: true tau is not equal to 0
 tau   0.2227567
```

Brier:	Brier score
C:	AUC
Dxy:	Somer's Dxy
tau-a	Kendall's tau-a
cor.test tau:	Kendall's tau-b

44

Exercise

Atrazine was measured in streams throughout the midwestern United States, in an area where corn is heavily grown. The dataset is `ReconLogistic.RData`

Are atrazine detections at a reporting limit =1 (GT_1 variable) a predictable function of the following land-use and climate variables?

Build the best logistic regression model you can for the recon data.

Use the following 6 of the 8 possible X variables:

APPLIC	amount of pesticide applied
corn%	% of basin planted in corn
soilgp	soil permeability - from the Census of Agriculture
precip	amount of recent precipitation
dyplant	days since planting (~ since atrazine last applied)
fpctl	percentile of streamflow

45