Logistic Regression

Categorical outcomes

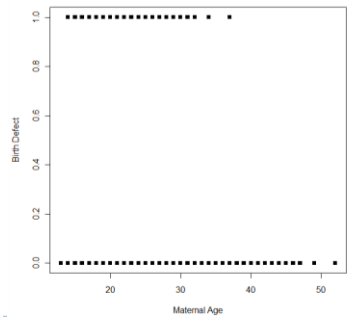
- Often in studies, we encounter outcomes that are not continuous, but instead fall into 1 of 2 categories
- For example:
  - Disease status (disease vs. no disease)
  - Alive or dead
  - Fire or no fire
  - Plant occurrence (present vs. absent)

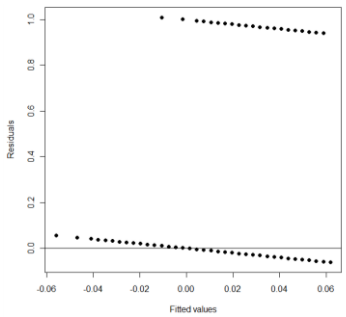
Modeling binary data

- So far, we have discussed cases where the dependent variable is continuous
- Logistic regression is a technique used when the dependent variable is categorical
- As with the other techniques, independent variables may be either continuous or categorical
- The technique can be extended for use with a multi-level categorical variable
  - Poisson regression or multinomial logistic regression

Why not just use linear regression?

- In the case of a binary response variable, the assumptions of linear regression are not valid:
  - The relationship between X and Y is nonlinear
  - Error terms are heteroscedastic
  - Error terms are not normally distributed





Why not just use linear regression?

- If you proceeded in light of these violations, the result would be:
  - Predicted values that are not possible (greater than a value of 1, smaller than a value of 0)
  - Magnitude of the effects of independent variables may be greatly underestimated
- So…what should we do?

Modeling a binary variable

- Recall the General Linear Model (GLM):

$$\hat{y} = a + bx$$

- Logistic regression is part of a family of models called the General*ized Linear Model*
- The main feature of these models is that instead of using *y* directly, it is modeled through what is called a "link" function (here G(•)):

$$G(y) = a + bx + e$$

Modeling a binary variable

- The *ized* ending to General comes from:
  - The model being linear after being transformed (after G(•))
  - The General Linear Model (linear regression) being a subset of the Generalized Linear Model:

$$G(y) = y$$

- For OLS regression, with a continuous response variable, the appropriate link function (above) is called the "identity link"
- The appropriate link function depends on the distribution of the response variable

Link functions

- Suggested models for use with dichotomous dependent variables include:
  - The logistic model
  - The probit model
    - These models define different link functions (G(•)) of the DV
- Logistic regression is considered to be the most versatile
  - After transforming the DV, logistic regression parallels least-squares regression

A step back

- If you have one dichotomous IV and one dichotomous DV, we can display them in a 2 x 2 table

	y		
x	Yes (1)	No (0)	Total
Yes (1)	a	b	a+b
No (0)	c	d	c+d
Total	a+c	b+d	

- We are interested in predicting the probability of a "yes" or "1" event
  - The overall probability of a "yes" is: $P = \frac{a+c}{a+b+c+d}$
  - The odds of "yes" for y would be: $\frac{P}{1-P}$ $\frac{\text{probability of a "yes"}}{\text{probability of a "no"}}$

Logits and logistic regression

- In logistic regression, a logistic transformation of the odds (referred to as logit) serves as the dependent variable:

$$\log(\text{odds}) = \text{logit}(P) = \log\left(\frac{P}{1-P}\right)$$

This is our "link" function
It represents the log of the odds of a "yes" answer, or the log-odds

- If we take the above dependent variable and add a regression equation for the independent variable, we get a logistic regression:

$$\text{logit}(P) = a + bx$$

No error term!

- As in least-squares regression, the relationship between the logit(*P*) and x is assumed to be linear
  - Log-odds changes linearly as a function of explanatory variables

Parameter estimation

- Within the framework of GLMs, ordinary least squares (OLS) parameter estimation replaced by maximum likelihood estimation (MLE)
- Likelihood measures how well a set of data support a particular value of a parameter
  - The probability of having obtained the observed data if the true parameter(s) equaled that value
- Calculate the probability of obtaining the sample data you observed for each possible value of the parameter
  - Compare this probability among the different values generated
- Value with the highest support (i.e., highest probability) is the maximum likelihood estimate
  - The best estimate of the parameter

Parameter interpretation

- Interpretation of parameters in analogous to simple linear regression, the slope is the expected change in logit(P) with a unit change in x
- Because the logit is the log of the odds, we exponentiate the logit to get the odds:

$$e^{y} = e^{(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n)}$$

  - The odds that a response is positive (i.e., y=1) when all x=0 is $e^{(\beta_0)}$
  - As x_i increases by 1 unit, the odds that y=1 changes by a multiplicative factor of $e^{(\beta_1)}$
    - Often referred to as the "odds-ratio"

Example 1: Continuous x

- We want to know if the probability of a certain birth defect is higher among women of a certain age
  - Outcome (y) = presence/absence of birth defect
  - Explanatory (x) = maternal age a birth

```
> bdlog<-glm(bd$casegrp~bd$MAGE,family=binomial("logit"))
```

Since "logit" is the default, you can actually use:

```
> bdlog<-glm(bd$casegrp~bd$MAGE,binomial)
```

```
> summary(bdlog)
```

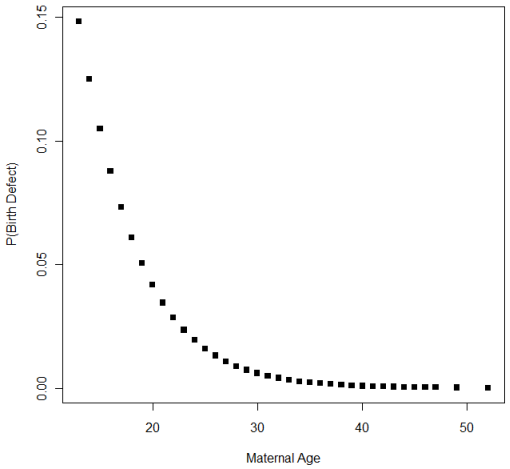
Example in R

```
Call:
glm(formula = bd$casegrp ~ bd$MAGE, family = binomial("logit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.56672  -0.24047  -0.14728  -0.08994   3.60539

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.82555    0.32491   2.541   0.0111 *
bd$MAGE     -0.19793    0.01489 -13.290   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    Null deviance: 2364.1  on 11892  degrees of freedom
Residual deviance: 2130.8  on 11891  degrees of freedom
AIC: 2134.8
```

$$\ln\left(\frac{P}{1-P}\right) = 0.8255 + -0.1979x$$

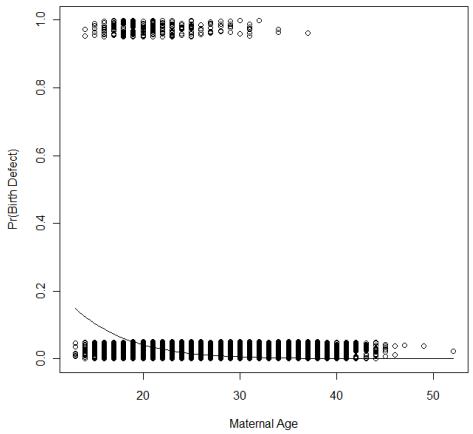
Plot

```
> plot(MAGE,
  fitted(glm(casegrp~
  MAGE,binomial)),
  xlab="Maternal Age",
  ylab="P(Birth Defect)",
  pch=15)
```



Another plot

```
> library(arm)
> jitter.binary<-
  function(a,jitt=.05)
  {ifelse (a==0,
  runif(length(a), 0,jitt),
  runif(length(a),1-
  jitt,1))}
> bd.jitter<-jitter.binary
  (bd$casegrp)
> plot(bd$MAGE,bd.jitter)
> curve(invlogit
  (coef(bd.lm1)[1]+coef(bd.
  lm1)[2]*x),add=TRUE)
```



Interpretation

$$\ln\left(\frac{P}{1-P}\right) = 0.8255 + -0.1979x$$

- This equation shows that for every 1 year increase in maternal age, the logit of the probability of a birth defect decreases by 0.1979
- We can predict the probability of a birth defect for a woman of a specific age using the regression equation

$$\ln\left(\frac{P}{1-P}\right) = 0.8255 + -0.1979(20) = -3.1325$$

$$\frac{e^{-3.1325}}{1+e^{-3.1325}} = 0.04179$$

A 20 year old woman has an average of a 0.04 chance of a birth defect

Interpretation

- We can use the odds ratio to measure how the fitted probability changes between different values of the explanatory variable

$$e^{\{B_i*(x_{i\max}-x_{i\min})\}}$$

$$e^{\{-0.1979*(25-20)\}} = e^{\{-0.9895\}} = 0.372$$

The odds of a 25 year old woman having a birth defect is about .37 (63% less) that of a 20 year old woman

$$e^{\{-0.1979*(20-25)\}} = e^{\{0.9895\}} = 2.690$$

The odds of a 20 year old woman having a birth defect is about 2.7 times that of a 25 year old woman

- *Note:*
  - *OR = 1 indicates a zero effect*
  - *OR >1 indicates an increase in odds*
  - *OR <1 indicates a decrease in odds*

Example 2: Categorical x

- Sometimes, it's easier to interpret logistic regression output if the x variables are categorical
- Suppose we categorize maternal age into 3 categories:

	Maternal Age		
Birth Defect	< 20 years	20-24 years	> 24 years
Yes	101	105	36
No	1385	3755	6511

```
> bd$magecat3 <- ifelse(bd$MAGE>25, c(1),c(0))
> bd$magecat2 <- ifelse(bd$MAGE>=20 & bd$MAGE<=25,
  c(1),c(0))
> bd$magecat1 <- ifelse(bd$MAGE<20, c(1),c(0))
```

Example in R

```
> bdlog2<-glm(casegrp~magecat1+magecat2,binomial)
> summary(bdlog2)
```

- Remember, with a set of dummy variables, you always put in one less variable than category

Example in R

```
Call:
glm(formula = casegrp ~ magecat1 + magecat2, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3752  -0.2349  -0.1050  -0.1050   3.2259

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.1977     0.1671 -31.101   <2e-16 ***
magecat1      2.5794     0.1964  13.137   <2e-16 ***
magecat2      1.6208     0.1942   8.345   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 2364.1  on 11892  degrees of freedom
Residual deviance: 2148.6  on 11890  degrees of freedom
AIC: 2154.6
```

Interpretation

$$\ln\left(\frac{P}{1-P}\right) = -5.1977 + 2.5794x + 1.6208x$$

- The predicted value (probability of birth defect) for women <20 years is given by:

$$\ln\left(\frac{P}{1-P}\right) = -5.1977 + 2.5794(1) + 1.6208(0)$$

- The predicted value for a women 20-24 years is:

$$\ln\left(\frac{P}{1-P}\right) = -5.1977 + 2.5794(0) + 1.6208(1)$$

- The predicted value for a woman >24 years is:

$$\ln\left(\frac{P}{1-P}\right) = -5.1977 + 2.5794(0) + 1.6208(0)$$

Interpretation: odds ratios

$$OR = \exp(\beta_n)$$

```
> exp(cbind(OR=coef(bdlog2),confint(bdlog2)))
Waiting for profiling to be done...
                      OR       2.5 %       97.5 %
(Intercept)  0.005529105 0.003910843  0.007544544
magecat1    13.189149619 9.066531887 19.622868917
magecat2     5.057367954 3.492376718  7.495720840
```

- This tells us that:
  - women <20 years have a 13 times greater odds of a birth defect than women >24 years
  - women 20-24 years have a 5 times greater odds of a birth defect than women >24 years

Significance testing in logistic regression

- Similar to linear regression, we have several hypotheses we want to test with logistic regression:
  - Logit or log of the odds is 0 vs. some other value
  - Whether estimates of IVs explain a significant portion of the variation in the DV
    - The contribution of individual regression coefficients
      - Wald's tests (similar to t-test)
    - The contribution of several coefficients simultaneously
      - Deviance tests (similar to F-tests)

The contribution of individual variables Example in R

```
> bd.log<-glm(casegrp~magecat1+magecat2+bthparity2+smoke,
  binomial, data=bd)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -4.9100     0.1870 -26.252  < 2e-16 ***
magecat1           2.2534     0.2073  10.872  < 2e-16 ***
magecat2           1.4732     0.1965   7.497 6.52e-14 ***
bthparity2parous  -0.5932     0.1497  -3.962 7.45e-05 ***
smokesmoker        0.6515     0.1546   4.213 2.52e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 2355.9  on 11880  degrees of freedom
Residual deviance: 2113.1  on 11876  degrees of freedom
  (12 observations deleted due to missingness)
AIC: 2123.1
```

This is Wald's test

Interpretation

$$\ln\left(\frac{P}{1-P}\right) = -4.91 + 2.25x_{agecat1} + 1.47x_{agecat2} + -.59x_{bthpar} + 0.65x_{smoke}$$

Test the effect of young maternal age on birth defect outcome, after accounting for birth parity and smoking

- Know `magecat1` has a significant effect on birth defect outcomes (Wald's)

```
                 Estimate Std. Error z value Pr(>|z|)
magecat1           2.2534     0.2073  10.872  < 2e-16 ***
```

$$e^{(2.25*(1-0))} = 9.52$$

A <20 year old woman's odds of a birth defect was 9.5 higher than a >24 year old woman, after adjusting for birth parity and smoking

Interpretation: odds ratios

```
> exp(cbind(OR=coef(bd.log),CI=confint(bd.log)))

Waiting for profiling to be done...
                           OR      2.5 %      97.5 %
(Intercept)       0.007372587 0.00502582  0.01047792
magecat1          9.520382522 6.40106854 14.45557517
magecat2          4.363334359 2.99877570  6.49390754
bthparity2parous 0.552540323 0.41049414  0.73886199
smokesmoker       1.918448991 1.40719166  2.58257928
```

Interpretation: prediction

$$\ln\left(\frac{P}{1-P}\right) = -4.91 + 2.25x_{agecat1} + 1.47x_{agecat2} + -.59x_{bthpar} + 0.65x_{smoke}$$

- What's the probability of a birth defect for a 21 year old smoker with no prior births?

$$\ln\left(\frac{P}{1-P}\right) = -4.91 + 2.25(0) + 1.47(1) + -.59(0) + 0.65(1) = -2.79$$

$$\frac{e^{-2.79}}{1+e^{-2.79}} = 0.0579$$

The contribution of groups of variables

- To compare full and reduced models in least squares regression we compared differences in the sizes of residuals between models
- Drop in deviance compares the change in deviance between a full and reduced set of variables
  - Large p-value means reduced model explains about the same amount of the variation as the full model
    - And, thus, you can probably leave out the variables

What variables can we consider dropping?

```
> anova(bd.log,test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: casegrp

Terms added sequentially (first to last)

             Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                         11880    2355.87
magecat1      1   130.58     11879    2225.28 < 2.2e-16 ***
magecat2      1    82.73     11878    2142.56 < 2.2e-16 ***
bthparity2    1    13.37     11877    2129.18 0.0002555 ***
smoke         1    16.10     11876    2113.09 6.022e-05 ***
```

Small p-values indicate that all variables are needed to explain the variation in y

Analysis of deviance: drop in deviance

- We want to compare the differences in the size of residuals between models

```
> bd.full<-glm(casegrp~magecat1+magecat2+bthparity2+smoke,
  binomial,data=bd)
> bd.red<-glm(casegrp~magecat1+magecat2+smoke,
  binomial,data=bd)

> 1-pchisq(deviance(bd.red)-deviance(bd.full),
  df.residual(bd.red)-df.residual(bd.full))
[1] 5.493537e-05
```

- Since the p-value is small, there is evidence that the addition of `birthparity2` explains a significant amount (more) of the deviance

Strategy for Analysis of Binary Data with Logistic Regression

- Identify the questions of interest
- Build models that include terms to answer questions via hypothesis tests and parameter estimates
- Examine the need for extra terms (interactions) as a way to assess model fit and adequacy
- Use Wald's tests to examine the effect of single variables
- Use analysis of deviance (drop in deviance) to compare full and reduced models to assess the contribution of several variables

Diagnostic tools for logistic regression

- Graphs of the data or the residuals are of less value with binary data because the response variable can only take on one of two values
  - No need to check for nonconstant variance or outliers, so diagnostics are somewhat simplified compared to linear regression
- Easiest way is to look at the model "goodness of fit" statistic, the AIC or *-2 log L* for logistic models
  - Ultimately the model with the **smallest** AIC is "best"

Goodness of fit statistics

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -4.9100     0.1870 -26.252  < 2e-16 ***
magecat1           2.2534     0.2073  10.872  < 2e-16 ***
magecat2           1.4732     0.1965   7.497 6.52e-14 ***
bthparity2parous  -0.5932     0.1497  -3.962 7.45e-05 ***
smokesmoker        0.6515     0.1546   4.213 2.52e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 2355.9  on 11880  degrees of freedom
Residual deviance: 2113.1  on 11876  degrees of freedom
AIC: 2123.1
```

-2 log L
AIC

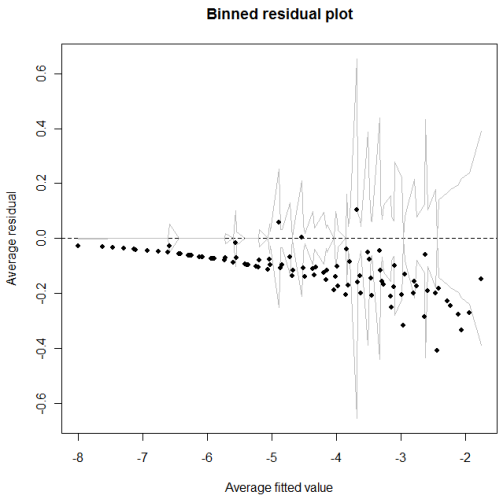
Binned residual plot

```
> x<-predict(bd.log)
> y<-resid(bd.log)
> binnedplot(x,y)
```

Plots the average residual and the average fitted (predicted) value for each bin, or category

Category is based on the fitted values

95% of all values should fall within the dotted line



Model building using stepwise regression

```
> load.package(MASS)
> library(MASS)
> stepAIC(glm(casegrp~magecat1+magecat2+bthparity2+smoke,binomial,data=bd))
Start:  AIC=2123.09
casegrp ~ magecat1 + magecat2 + bthparity2 + smoke

             Df Deviance    AIC
<none>            2113.1 2123.1
- smoke       1   2129.2 2137.2
- bthparity2  1   2129.4 2137.4
- magecat2    1   2178.8 2186.8
- magecat1    1   2251.1 2259.1

Call:  glm(formula = casegrp ~ magecat1 + magecat2 + bthparity2 + smoke,      family =
   binomial, data = bd)

Coefficients:
     (Intercept)          magecat1          magecat2  bthparity2parous   smokesmoker
         -4.9100            2.2534            1.4732           -0.5932          0.6515

Degrees of Freedom: 11880 Total (i.e. Null);  11876 Residual
Null Deviance:      2356
Residual Deviance: 2113         AIC: 2123
```

Next class…

Poisson - Used for count data of rare events where outcome can be 1,2,3,4….n

- Number of cases of cholera
- Number of forest fire events

- The link function for Poisson regression is the log function
  - G(y) = log(y)