Logistic Regression

Today . . .

- Generalized linear models
- Binary logistic regression

The general linear model (ordinary least squares, OLS) that underlies common analyses such as multiple regression assumes continuous and normally distributed outcomes.

Some common outcomes do not fit those assumptions:

- Binary outcomes (e.g., alive vs. dead, pregnant or not pregnant, etc.)
- Multinomial outcomes (e.g., single, married, divorced; rating scale data)
- Counts (e.g., number of arguments; number of serious life events).

Why not just use OLS?

- In some cases OLS might produce a reasonable approximation, but it might not be the best approximation available.
- In some cases, OLS will produce absurd results. Using a binary outcome in a multiple regression can produce predicted values that exceed the bounds of 0 and 1.
- In some cases it makes no sense (a multiple categorical variable with no order to the categories).

Sometimes a non-normal variable (e.g., counts) can be transformed to be approximately normal and then OLS can be used.

But the outcome then is in an unfamiliar metric.

An alternative—generalized linear models—assumes a different (non-normal) distribution for the outcome and then attempts to model that directly.

Generalized linear models have three basic features:

- Random component—the response variable—that has a known (or strongly expected) underlying distribution.
- A systematic component that reflects the predictor or explanatory variables for the response.
- A link function that transforms the expected value for the response to a linear relation with the predictors. The link function connects the random and systematic components.

The simplest generalized linear model addresses the simplest kind of outcome—a binary variable.

- The problem with using a simple OLS regression is that the outcome is bounded by 0 and 1. It makes no sense to speak of values beyond those limits, but values beyond those limits might be predicted.
- When the limits are respected, the predictors will be nonlinearly related to the outcome.
- It would be better if the outcome were unbounded and the model a linear one.

A single trial binary outcome has a Bernoulli distribution with parameter, π, equal to the probability of the outcome Y=1 occurring.

The bounded probability can be transformed to an unbounded value using the logit link function:

$$Logit = ln\left[\frac{\pi}{1-\pi}\right]$$

A linear model for the logit can be related back to the original probability metric through the inverse of the link function.

$$P = \frac{e^{Logit}}{1 + e^{Logit}}$$

Logistic regression builds a linear model for logits. This links the linear model to the probability (π) metric via the link function, so interpretation is relatively easy. It has some distinct advantages over other approaches (e.g., discriminant analysis):

- Continuous and categorical predictors can be used
- Interactions can be included
- Nonlinearity (some forms) can be tested
- It generalizes well to more than two categorical outcomes.

Logistic regression shares many of the same limitations with OLS regression:

- Multicollinearity can limit interpretation
- A linear model is assumed, with provision for certain kinds of nonlinearity
- Outliers can distort the analysis

In logistic regression, the predictors are linearly related to the logit—the natural log of the odds:

$$Logit_i = ln\left[\frac{P_i}{1 - P_i}\right]$$

These range from -∞ (P = 0) to ∞ (P = 1).

For ease of interpretation, the logits can be transformed back to probabilities:

$$P_i = \frac{e^{Logit_i}}{1 + e^{Logit_i}}$$

The link function provides a way of moving between the response variable that is analyzed and the response variable that might be most easily interpreted.

In logistic regression . . .

$$P_i = \frac{e^{L_i}}{1 + e^{L_i}}$$

Probability that event occurs for Person i

$$L_i = B_0 + \sum_{k=1}^{K-1} B_k X_{ik}$$

Linear combination of predictors that estimates the log of the odds.

$$LF = \prod_{i=1}^{N} \left[P_i^{Y_i}(1 - P_i)^{1 - Y_i}\right]$$

The likelihood function, or joint probability of the sample data given L. When this is maximized, the "best" linear combination of X has been found (the Bs are optimal).

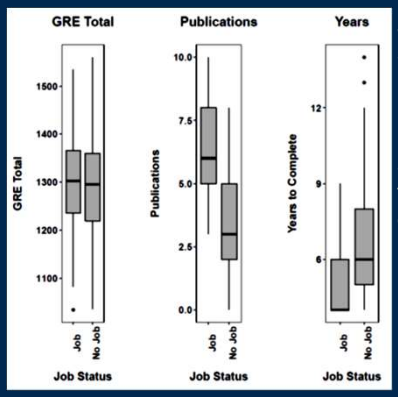
$$ln(LF) = \sum_{i=1}^{N} \left[Y_i ln(P_i) + (1 - Y_i) ln(1 - P_i)\right]$$

Mathematically easier to use the log of the likelihood function.

Y_i = 1 or 0.

Hypothetical data (N = 500) relating job search outcome (no job, job) to aggregate GRE score, number of publications, sex of applicant (men =1, women = 2) and years to finish degree.

Products of these predictors can be formed to test interactions and powers can be used to test polynomial nonlinearity.



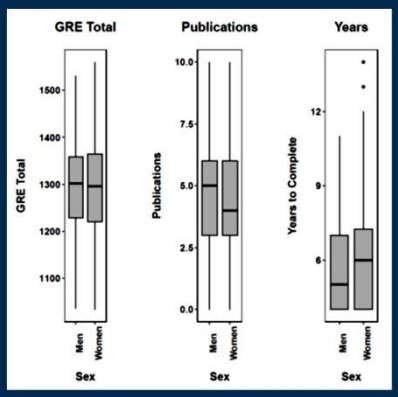
Job status is significantly related to number of publications and years to finish the degree.

	sex_F	
job_result	Men	Women
Job	0.3298	0.2372
No Job	0.6702	0.7628

```
Pearson's Chi-squared test with Yates' continuity
correction

data:  table_1
X-squared = 4.6, df = 1, p-value = 0.03
```

Women were less likely to get jobs than men.



Men and women differed significantly in the number of years to complete the degree.

```
Job_BLR_1 <- glm(job ~ sex_D, family = binomial("logit"), data = Job)
summary(Job_BLR_1)

## Coefficients:
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept)   -0.709      0.155   -4.57 0.0000048
## sex_D         -0.459      0.204   -2.25     0.025
##
```

The regression equation is interpreted in much the same manner as in standard multiple regression except that the predicted score is the logit—the natural log of the odds:

$$L = ln\left[\frac{P}{1-P}\right]$$

There is a significant sex difference in the log odds of getting a job.

```
## Coefficients:
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept)   -0.709      0.155   -4.57 0.0000048
## sex_D         -0.459      0.204   -2.25     0.025
##
```

For men:

$$L = -.709 - .459(0) = -.709$$

$$L = ln\left[\frac{P}{1-P}\right]$$

$$P = \left[\frac{e^L}{1+e^L}\right]$$

$$P = .33$$

Men have a probability of getting a job of .33

```
## Coefficients:
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept)   -0.709      0.155   -4.57 0.0000048
## sex_D         -0.459      0.204   -2.25     0.025
##
```

For women:

$$L = -.709 - .459(1) = -1.168$$

$$L = ln\left[\frac{P}{1-P}\right]$$

$$P = \left[\frac{e^L}{1+e^L}\right]$$

$$P = .24$$

Women have a probability of getting a job of .24

```
confint(Job_BLR_1)

## Waiting for profiling to be done...

##                2.5 %  97.5 %
## (Intercept) -1.0191 -0.4098
## sex_D       -0.8597 -0.0574

confint.default(Job_BLR_1)

##                2.5 %   97.5 %
## (Intercept) -1.0132 -0.40510
## sex_D       -0.8597 -0.05844

exp(cbind(OR = coef(Job_BLR_1), confint(Job_BLR_1)))

## Waiting for profiling to be done...
##                 OR  2.5 % 97.5 %
## (Intercept) 0.4921 0.3609 0.6638
## sex_D       0.6319 0.4233 0.9442
```

Confidence intervals can be placed around the coefficients (that predict logits) as well as the odds ratios (obtained by exponentiating the coefficients).

The odds of getting a job for men: .33/(1-.33) = .49

The odds of getting a job for women: .24/(1-.24) = .31

The odds ratio = .31/.49 = 0.63 = $e^B = e^{-.459}$

The odds of a woman getting a job are .63 times the odds of a man getting a job, not taking other predictors into account.

```
Coefficients:
            Estimate Std. Error z value  Pr(>|z|)
(Intercept)   -0.709      0.155   -4.57 0.0000048
sex_D         -0.459      0.204   -2.25     0.025
```

```
                 OR  2.5 % 97.5 %
(Intercept) 0.4921 0.3609 0.6638
sex_D       0.6319 0.4233 0.9442
```

```
Coefficients:
            Estimate Std. Error z value  Pr(>|z|)
(Intercept)   -0.709      0.155   -4.57 0.0000048
sex_D         -0.459      0.204   -2.25     0.025
```

$$Wald\ test = \left[\frac{B}{SE_B}\right]^2$$

The Wald test is chi-square distributed. With 1 df, the square root of the Wald test can be interpreted as a Z test.

```
Job_BLR_2 <- glm(job ~ gre_c, family = binomial("logit"), data = Job)
summary(Job_BLR_2)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.985718   0.100609   -9.80   <2e-16
gre_c        0.000706   0.000970    0.73     0.47
```

```
Job_BLR_3 <- glm(job ~ pubs_c, family = binomial("logit"), data = Job)
summary(Job_BLR_3)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7585     0.1745   -10.1   <2e-16
pubs_c        0.9491     0.0911    10.4   <2e-16
```

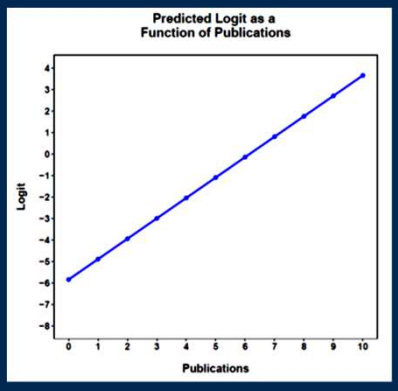
```
Job_BLR_4 <- glm(job ~ years_c, family = binomial("logit"), data = Job)
summary(Job_BLR_4)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.2775     0.1307   -9.77  < 2e-16
years_c      -0.5711     0.0786   -7.27  3.6e-13
```

```
Job_BLR_3 <- glm(job ~ pubs_c, family = binomial("logit"), data = Job)
summary(Job_BLR_3)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7585     0.1745   -10.1   <2e-16
pubs_c        0.9491     0.0911    10.4   <2e-16
```

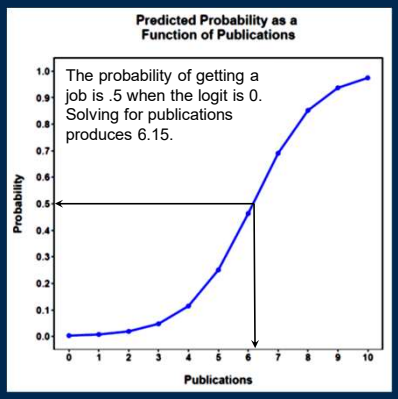
Publications are significantly related to the log odds of getting a job. Each additional publication increases the log of the odds by .949.

Each additional publication multiplies the odds of getting a job by 2.583.

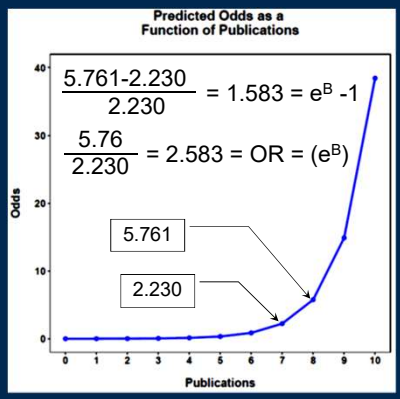
The odds of getting a job change by $100(e^B-1)\%$ with each additional publication. In this case, each publication increases the odds of getting a job by 158.3%.

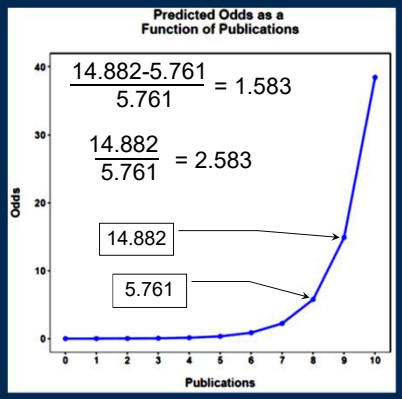


The regression equation can be used to predict the logit for different values of publications. This relationship must be linear.



The logits can be converted to probabilities for easier interpretation.





$$\frac{14.882-5.761}{5.761} = 1.583$$

$$\frac{14.882}{5.761} = 2.583$$

The best way to view the data is to enter all predictors simultaneously.

```
Job_BLR_5 <- glm(job ~ gre_c + pubs_c + years_c + sex_D, family = binomial("logit"),
    data = Job)
summary(Job_BLR_5)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.71205    0.46554   -7.97  1.5e-15
gre_c       -0.01470    0.00231   -6.37  1.8e-10
pubs_c       1.99614    0.22058    9.05  < 2e-16
years_c     -1.43390    0.18667   -7.68  1.6e-14
sex_D       -0.40619    0.35023   -1.16     0.25
```

Now the relationships are partialled relationships, controlling for other predictors in the model.

Controlling for GRE, publications, and years to degree, there is no longer a sex difference in the odds of getting a job.

```
                OR    2.5 %   97.5 %
(Intercept) 0.02443 0.009095  0.05691
gre_c       0.98540 0.980689  0.98963
pubs_c      7.36059 4.947572 11.78854
years_c     0.23838 0.160716  0.33501
sex_D       0.66619 0.333227  1.32310
```

The interpretation of e^B is now with the qualification, "all else equal" or "holding other predictors constant."

All else equal, each additional publication multiplies the odds of getting a job by 7.361.

Interactions can be easily incorporated into the model.

```
Job_BLR_6 <- glm(job ~ gre_c + pubs_c + years_c + sex_D + pubs_c:years_c +
    pubs_c:sex_D + years_c:sex_D, family = binomial("logit"), data = Job)
summary(Job_BLR_6)
```

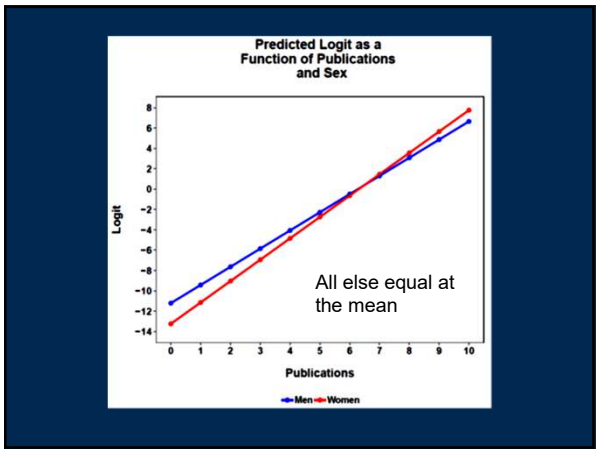
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.52816    0.60351   -5.85  5.0e-09
gre_c        -0.01476    0.00233   -6.32  2.6e-10
pubs_c        1.78635    0.28812    6.20  5.6e-10
years_c      -1.43511    0.32062   -4.48  7.6e-06
sex_D        -0.68367    0.68462   -1.00     0.32
pubs_c:years_c -0.02938    0.11205   -0.26     0.79
pubs_c:sex_D  0.31366    0.31912    0.98     0.33
years_c:sex_D  0.06856    0.32151    0.21     0.83
```

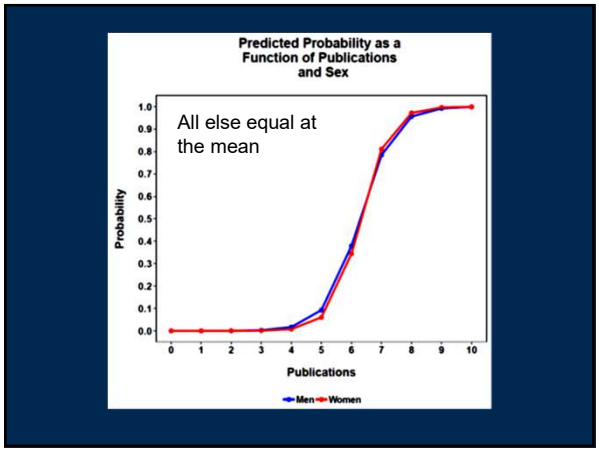

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.52816    0.60351   -5.85  5.0e-09
gre_c        -0.01476    0.00233   -6.32  2.6e-10
pubs_c        1.78635    0.28812    6.20  5.6e-10
years_c      -1.43511    0.32062   -4.48  7.6e-06
sex_D        -0.68367    0.68462   -1.00     0.32
pubs_c:years_c -0.02938    0.11205   -0.26     0.79
pubs_c:sex_D  0.31366    0.31912    0.98     0.33
years_c:sex_D  0.06856    0.32151    0.21     0.83
```

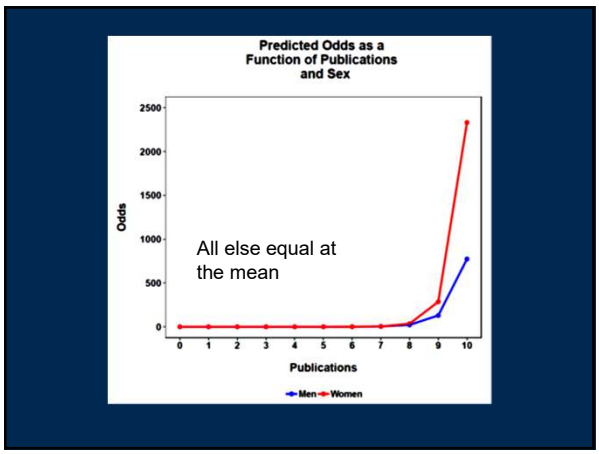

Controlling other predictors at their means, the simple relation between publications and logits for men and women:
Men: L = -3.53 + 1.786(pubs_c)
Women: L = -4.21 + 2.100(pubs_c)

```
                    OR    2.5 %   97.5 %
(Intercept)    0.02936 0.007827  0.08482
gre_c          0.98535 0.980569  0.98963
pubs_c         5.96764 3.559724 11.05927
years_c        0.23809 0.119984  0.42486
sex_D          0.50476 0.130603  1.98520
pubs_c:years_c 0.97104 0.776663  1.20548
pubs_c:sex_D   1.36843 0.723789  2.56885
years_c:sex_D  1.07097 0.574714  2.05900
```

Each additional publication, multiplies the odds of getting a job by e^B, or 5.968 ($e^{1.786}$) for men and 8.166 ($e^{2.100}$) for women. The odds ratio is 8.166/5.968 = 1.368, the value of e^B for the interaction. Although not significant, this indicates that publications benefit women more than men.


Predicted Logit as a Function of Publications and Sex
Logit
Publications
Men
Women
All else equal at the mean

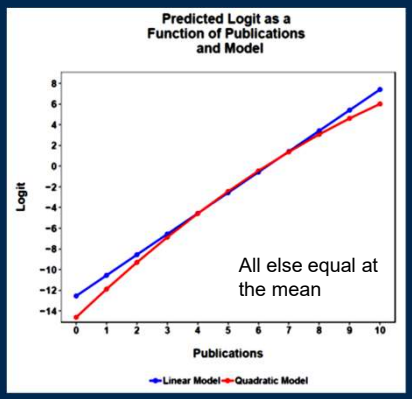

Predicted Probability as a Function of Publications and Sex
Probability
Publications
Men
Women
All else equal at the mean

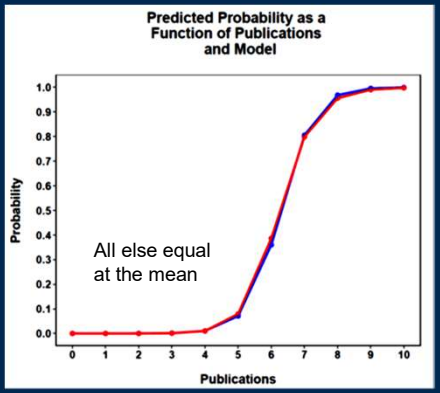

Predicted Odds as a Function of Publications and Sex
Odds
Publications
Men
Women
All else equal at the mean

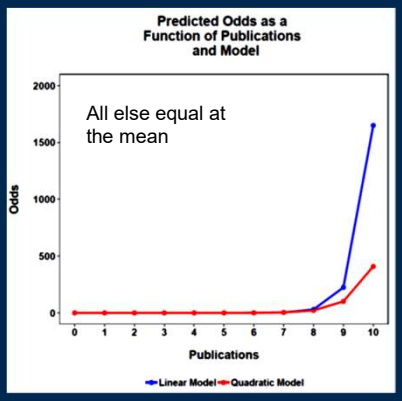
```
Job_BLR_7 <- glm(Job$job ~ gre_c + pubs_c + years_c + sex_D + I(pubs_c^2),
    family = binomial("logit"), data = Job)
summary(Job_BLR_7)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.67453    0.46818   -7.85  4.2e-15
gre_c       -0.01465    0.00229   -6.39  1.7e-10
pubs_c       2.16589    0.30909    7.01  2.4e-12
years_c     -1.41059    0.18569   -7.60  3.0e-14
sex_D       -0.40424    0.35093   -1.15     0.25
I(pubs_c^2) -0.07396    0.08421   -0.88     0.38
```

Nonlinearity can be incorporated into the model as well by adding powers of variables.







Classification: For each case, the predicted logit can be transformed to a probability, which can then be used to classify participants into their expected job categories. These predicted classifications can be compared to the known job search outcome. The quality of these classifications can be assessed in the same way as was done with discriminant analysis (e.g., t, τ).

For these data, 88.8% of the cases are classified correctly, with τ = .72, and $t(496) = 12.99$, $p < .001$.

Next time . . .

- Model diagnostics
- Ordinal and multinomial outcomes