Binary data – logistic regression continued

Type of Outcome/Exposure Variables

OUTCOME VARIABLE

EXPOSURE VARIABLE

	Continuous	Binary
1 group	One-group t-test	Exact binomial test [or] normal approximation test
2 groups	Two-group t-test	χ^2 test [or] Fisher's exact test
>2 groups	ANOVA	χ^2 test [or] Fisher's exact test
Continuous	Linear regression	Logistic regression

Logistic Regression

- Logistic regression relates the predictor variable(s) to the **log odds of the outcome** (log odds of being in the category of interest)

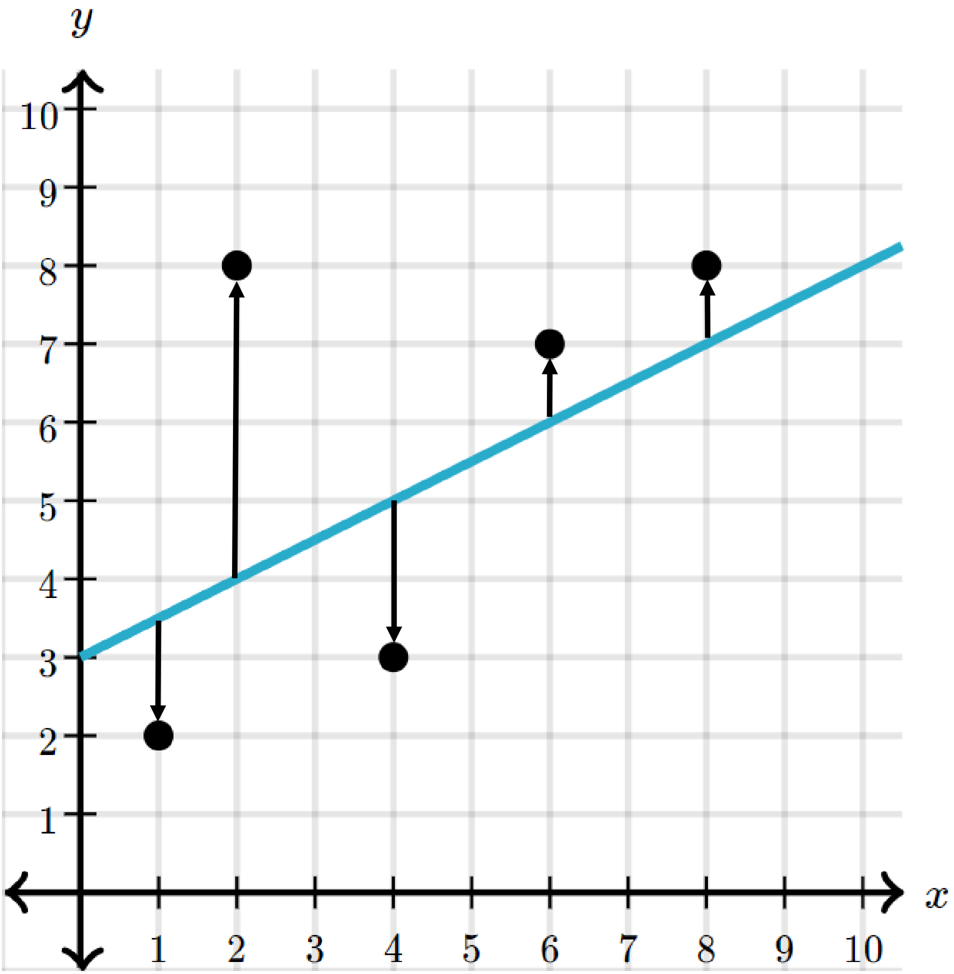
If y is the binary outcome variable, then let p be the probability that y equals the category of interest. The logistic regression model is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

REVIEW

Linear Regression Coefficients

- In linear regression, statistical software estimates coefficients by finding the equation that minimizes the residuals
  - Called the **ordinary least squares (OLS)** method



Logistic Regression Coefficients

- In logistic regression, statistical software estimates coefficients by finding the equation that maximizes the probability of getting our sample results
  - Called the **maximum likelihood estimation (MLE)** method

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Model Convergence

- Can’t directly solve for MLE estimates of logistic regression coefficients
- Statistical software finds the coefficient estimates using an iterative process
  - Basically “guess-and-check” until it finds the best answer
  - When it finds the best answer, we say that the model has **converged**
- Beware: sometimes the model doesn’t converge
  - Usually because the sample size is too small for the number of predictor variables in your model → either decrease the number of predictors or increase the sample size to get a model that converges

A Note on e

- Key: Regression models must capture the uncertainty in each subject's outcome

In linear regression, we do this by adding the residual on to the equation…

$$y = \beta_0 + \beta_1 x + e$$

…or by indicating that the equation is for the predicted values ($\hat{y}$), not the observed values (y).

$$\hat{y} = \beta_0 + \beta_1 x$$

In logistic regression, we model the log odds of the outcome directly (without a residual):

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

This is because we're only modeling the probability of the outcome, not the actual outcome itself. Our uncertainty in the outcome itself (0 or 1) is captured in this probability.

REVIEW

Assumptions of Linear Regression

- **Independence** of the observations
- **Linearity** of the relationship between the predictor(s) and outcome
- **Constant variance** of the residuals
- **Normality** of the residuals
- Absence of **multicollinearity**

Assumptions of Logistic Regression

- **Independence** of the observations
- **Linearity** of the relationship between the predictor(s) and ~~outcome~~ the log odds of the outcome
- ~~**Constant variance** of the residuals~~
- ~~**Normality** of the residuals~~
- Absence of **multicollinearity**
- Large sample size

Too small may lead to the model not converging

Confidence Intervals

- Recall: $\exp(\hat{\beta}_1)$ is the estimated odds ratio of the outcome for a one unit increase in the predictor variable
  - When this is the only predictor in the model, this is an **unadjusted odds ratio**
  - When there are other predictor variables in this model, this is an **adjusted odds ratio**

- We can also provide a range of possible values for the adjusted or unadjusted odds ratio in the population ($\exp(\beta_1)$)
  - Interpretation: "We are 95% confident that the adjusted/unadjusted odds ratio of [outcome variable] for a one unit increase in [predictor variable] is between _____ and _____."
  - What value would indicate that there is no significant association between the predictor and the outcome?

Low Birth Weight Data

- Information on 100 low birth weight infants born in two teaching hospitals in Boston, Massachusetts

Variable	Description
birthwtlbs	Birth weight of the baby (pounds)
hemorrhage	Germinal matrix hemorrhage (No, Yes). This is a type of brain bleed in a premature baby.
apgar	Apgar score (integers, min=0, max=10). This is a scoring system used for assessing the clinical status of a newborn. 7 or higher is generally considered normal, 4-6 is low, and 3 or below is critically lo
apgarnormal	Binary indicator variable for a normal Apgar score. Equals 1 when Apgar score is Normal (7-10) and equals 0 otherwise.

Find the dataset (lowbwt.xlsx) and the full data dictionary (lowbwt Data Dictionary.pdf) in the Data Module on the Canvas site

Example: Apgar Score

- Calculate and interpret a 95% confidence interval for the adjusted odds ratio of having a normal Apgar score for a one pound increase in birth weight, adjusting for germinal matrix hemorrhage status.

Let y be the normal Apgar score variable

Let p be the probability of having a normal apgar score (y=1)

Let x1 be birthweight (in pounds)

Lex x2 be germinal matrix hemorrhage status (indicator variable,

Example: Apgar Score

Logistic regression (outcome is log odds of *apgarnormal*):

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.5386     0.9137  -1.684  0.09219
birthwtlbs          0.9195     0.3755   2.449  0.01433
hemorrhage[T.Yes]  -1.7156     0.6494  -2.642  0.00825
```

Exponentiated coefficient estimates:

```
(Intercept)        birthwtlbs hemorrhage[T.Yes]
  0.2146772         2.5080151         0.1798562
```

Confidence intervals:

```
                    Estimate      2.5 %     97.5 % exp(Estimate)      2.5 %    97.5 %
(Intercept)       -1.5386197 -3.3808788  0.2326678     0.2146772 0.03401755 1.2619622
birthwtlbs         0.9194917  0.1998267  1.6825953     2.5080151 1.22119115 5.3794992
hemorrhage[T.Yes] -1.7155976 -3.1090470 -0.5080707     0.1798562 0.04464348 0.6016552
```

log odds confidence intervals

odds CI

We are 95% confident that the true adjusted odds ratio of having a normal Apgar score for a one pound increase in birth weight is between 1.22 and 5.38, controlling for germinal matrix hemorrhage status.

Hypothesis Testing

- We can also perform a hypothesis test for the association between a predictor variable and the binary outcome variable in the population
  - When there is only one predictor in the model, we're testing the **unadjusted** association
  - When there are other predictor variables in this model, we're testing the **adjusted** association, controlling for the other predictor variables in the model

$$H_0: \beta_1 = 0 \qquad H_A: \beta_1 \neq 0$$

Reject H_0 when p-value ≤ α
Fail to reject H_0 when p-value > α

Note: Can perform test on the log odds (β_1) or odds ($\exp(\beta_1)$) scale – both are measures of association between the two variables.

H0: exp(Beta1) = 1
HA: exp(Beta1) != 1

Example: Apgar Score

- Is the association between birth weight and having a normal Apgar score statistically significant after adjusting for germinal matrix hemorrhage status?

H0: beta1 = 0

HA: beta1 != 0

Let p be the probability of having a normal apgar score

Let x1 be birth weight (in pounds)

Let x2 be germinal matrix hemorrhage indicator variable

Example: Apgar Score

Logistic regression (outcome is log odds of *apgarnormal*):

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.5386	0.9137	-1.684	0.09219
birthwtlbs	0.9195	0.3755	2.449	0.01433
hemorrhage[T.Yes]	-1.7156	0.6494	-2.642	0.00825

p-value = 0.014

Since the p-value is less than 0.05, we reject the null hypothesis and conclude that there is evidence to suggest that there is an association between birth weight and having a normal Apgar score, after adjusting for germinal matrix hemorrhage status.

Comparing Logistic Regression Models

- In linear regression, we used adjusted R^2 to compare two models
  - Larger adjusted R^2 → better model fit

- In logistic regression, an equivalent R^2 statistic does not exist
  - A number of different measures can be used to compare models
    - AIC
    - BIC
    - Log likelihood
    - Pseudo R^2 (there are many)
    - Deviance
    - Etc…
  - We'll focus on **AIC** (Akaike Information Criterion)
  - Smaller AIC → better model fit

Example: Apgar Score

- Fit the following models to examine the association between birth weight and having a normal Apgar score, with and without adjusting for germinal matrix hemorrhage status:
  - Model 1: apgarnormal ~ birthwtlbs
  - Model 2: apgarnormal ~ birthwtlbs + hemorrhage

- Is it necessary to adjust for germinal matrix hemorrhage status in the model?

Example: Apgar Score

Model 1 (unadjusted model):

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7393     0.8854  -1.965   0.0495
birthwtlbs    0.8951     0.3620   2.472   0.0134
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 134.60  on 99  degrees of freedom
Residual deviance: 128.14  on 98  degrees of freedom
AIC: 132.14
```

Model 2 (adjusted model):

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.5386     0.9137  -1.684  0.09219
birthwtlbs         0.9195     0.3755   2.449  0.01433
hemorrhage[T.Yes] -1.7156     0.6494  -2.642  0.00825
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 134.60  on 99  degrees of freedom
Residual deviance: 120.19  on 97  degrees of freedom
AIC: 126.19
```

Since AIC in model 2 is smaller, model 2 is the better model.

Important Points

- Model convergence: what it means and reasons why a model may not converge
- Logistic regression assumptions
- Confidence interval for logistic regression coefficients (interpretation)
- Hypothesis test for logistic regression coefficients (set up and interpretation)
- Using AIC to compare models