Logistic Regression

1. Gestatational age was dicotomized to form a new variable x, which equals 1 if gestational age is greater than or equal to 30, and 0 otherwise.

   The variable y equals 1 if the mother is toxemic, and otherwise y is 0.

	non toxemic(y=0)	toxemic(y=1)	
$x = 0$	55	6	61
$x = 1$	24	15	39
	79	21	100

   The estimated odds ratio is $(15/24)/(6/55) = 5.73$. We could construct a CI for population odds ratio as before. However, logistic regression allows us to do this in a more general fashion.

   - We suppose that $y = 1$ with probability p, and $y = 0$ with probability (1-p).
   - Logistic regression associates p with co-variates according to the following statistical model.

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

   - When $x = 0$, $log\left(\frac{p_0}{1-p_0}\right) = \alpha$
   - When $x = 1$, $log\left(\frac{p_1}{1-p_1}\right) = \alpha + \beta$
   - The subscript on p has been used to identify the particular value of x under consideration.
   - It follows that

$$log\left(\frac{p_1}{1-p_1}\frac{1-p_0}{p_0}\right) = \beta$$

   - **β is the log odds ratio**

- To test the null hypothesis that x does not influence the probability of toxemia, we are formally testing the hypothesis $H_0 : \beta = 0$. Alternatively, we might construct a confidence interval for β.

Here is a partial output from a logistic regression.

```
Logistic Regression Table

                                                 Odds     95% CI
Predictor       Coef    SE Coef       Z  P      Ratio  Lower  Upper
Intercep    -2.21557   0.429919   -5.15 0.000
X            1.74557   0.541446    3.22 0.001  5.73   1.98  16.56
```

- the estimated log odds ratio is $\hat{\beta} = 1.746$, with estimated standard error equal to $\widehat{s.e.}(\hat{\beta}) = .541$.
- exponentiating the estimated log odds ration we get the estimated odds ratio $exp(1.74557) = 5.73$
- A $100(1-\alpha)\%$ confidence interval for the log odds ratio is given by

$$\hat{\beta} \pm Z_{\alpha/2}\widehat{s.e.}(\hat{\beta})$$

  For example, a 95% CI for log(OR) is
  $1.74557 \pm 1.96(.541)$, or $(.684, 2.81)$
- Exponentiating the endpoints of this interval, we get a 95% CI for the odds ratio OR (exp(.684),exp(2.81)), which is (1.98, 16.56)
- The hypothesis of unit odds ratio $H_0 : OR = 1$ vs $H_A : OR \neq 1$ is equivalent to the test $H_0 : \beta = 0$ vs $H_A : \beta \neq 0$
  - the observed test statistic is

$$Z_{obs} = \frac{\hat{\beta}}{\widehat{s.e.}(\hat{\beta})} = 1.74557/.541446 = 3.22$$

– the p-value is $2P(Z > |Z_{obs}|) = 2P(Z > 3.22) = .001$

2. Dichotomizing gestational age might result in a loss of information.

   We can relate probability of disease to the continuous predictor variable x (gestational age), as

$$log\left(\frac{p_x}{1-p_x}\right) = \alpha + \beta x$$

   whish assumes that the log odds of disease is linear in x

   - βx is a log odds ratio, associated with an increase of x in the independent variable.

$$\beta x = log\left(\frac{p_x}{1-p_x}\frac{1-p_0}{p_0}\right)$$

   - This means that β **is the log odds ratio associated with an increase of 1 unit in** x
   - and therefore, that $exp(\beta)$ **is the odds ratio associated with a unit increase in** x

A logistic regression give the following partial output.

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Intercept	-16.2117	4.07500	-3.98	0.000			
gestage	0.501729	0.134441	3.73	0.000	1.65	1.27	2.15

- As gestational age increases by 1, the estimated log odds ratio is .502
- As gestational age increases by 1, the estimated odds ratio is $exp(.5017) = 1.65$
- The $100(1-\alpha)\%$ confidence interval for the log odds ratio is the same as $100(1-\alpha)\%$ confidence interval for β, which is given by

  $$\hat{\beta} \pm Z_{\alpha/2}\widehat{s.e.}(\hat{\beta})$$

  eg. a 95% confidence interval for β is $.501729 \pm 1.96(.134441)$, or (.238,.765)
- The confidence interval for the log odds ratio is the same as the confidence interval for β. This means that the confidence interval for the odds ratio is formally equivalent to the confidence interval for $exp(\beta)$. A 95% confidence interval for the OR, $exp(\beta)$ associated with a unit increase in x is (exp(.238),exp(.765)), or $(1.27, 2.15)$.

  An odds ratio equal to 1 means that there is no change in odds (hence no change in probability of event of interest) in the groups being compared. Because 1 is not contained in the 95%CI for the odds ratio, we formally reject the null hypothesis that the odds ratio is 1.

- We can carry out a formal test of $H_0 : \beta = 0$ vs $H_A : \beta \neq 0$, because when $\beta = 0$, the odds ratio is 1.
  - the test statistic is

$$Z_{obs} = \frac{\hat{\beta}}{\widehat{s.e.}(\hat{\beta})} = .501729/.134441 = 3.73$$

  - the p-value is $2P(Z > |Z_{obs}|) = 2P(Z > 3.73) \approx .000$

3. **(multiple) logistic regression of toxemia on mother's age and gestational age**

Let x_1 denote gestational age, x_2 denote mother's age, and $y = 1$ for toxemic pregnancy, otherwise 0.

The statistical model is

$$log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

```
Logistic Regression Table
                                                            Odds
Predictor               Coef      SE Coef        Z       P    Ratio
Constant            -16.1914      4.08049    -3.97   0.000
gestage (X1)        0.512590     0.143495     3.57   0.000    1.67
momage  (X2)      -0.0122518    0.0539993    -0.23   0.821    0.99
```

- A test of $H_0 : \beta_2 = 0$ vs $H_A : \beta_2 \neq 0$ has p-value .821.
- As in multiple regression, this is a test of whether mom's age has a significiant effect, given that gestational age is already in the model.
- The p-value is large, so the null hypothesis is not rejected, and we can reduce the model to include only gestational age.