Logistic Regression - Part I

Statistics 149

Spring 2006

Logistic Regression Model

Let Y_i be the response for the ith observation where

$$Y_i = \begin{cases} 1 & \text{Success} \\ 0 & \text{Failure} \end{cases}$$

Then

$$Y_i \sim Bin(1, \pi(x_i))$$

where x_i is the level of the predictor of observation i.

An equivalent way of thinking of this, is to model

$$\mu(Y_i|x_i) = \pi(x_i)$$

One possible model would be

$$\pi(x_i) = \beta_0 + \beta_1 x_i$$

As we saw last class, this model will give invalid values for π for extreme xs. In fact there is a problem if

$$x < \frac{-\beta_0}{\beta_1} \qquad \text{or} \qquad x > \frac{1-\beta_0}{\beta_1}$$

We would still like to keep a linear type predictor. One way to do this is to apply a **link function** $g(\cdot)$ to transform the mean to a linear function, i.e.

$$g(\pi) = \beta_0 + \beta_1 x$$

As mentioned last class, we want the function $g(\cdot)$ to transform the interval [0,1] to $(-\infty, \infty)$. While their are many possible choices for this, there is one that matches up with what we have seen before, the logit transformation

$$\text{logit}(\pi) = \log \frac{\pi}{1-\pi} = \log \omega = \eta$$

So instead of looking at things on the probability scale, lets look at things on the log odds (η) scale.

Transforming back gives

$$\frac{\pi}{1-\pi} = \omega = e^{\eta} = e^{\beta_0+\beta_1 x}$$

and

$$\pi = \frac{\omega}{1+\omega} = \frac{e^{\eta}}{1+e^{\eta}} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

Note the standard Bernoulli distribution results hold here

$$\mu(Y|X) = \pi = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \qquad \text{and} \qquad \mathrm{Var}(Y|X) = \pi(1-\pi) = \frac{e^{\beta_0+\beta_1 x}}{(1+e^{\beta_0+\beta_1 x})^2}$$

While the above just assumes a single predictor, it is trivial to extend this to multiple predictors, just set

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = X\beta$$

giving

$$\frac{\pi}{1-\pi} = \omega = e^{\eta} = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p} = e^{X\beta}$$

and

$$\pi = \frac{\omega}{1+\omega} = \frac{e^{\eta}}{1+e^{\eta}} = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}} = \frac{e^{X\beta}}{1+e^{X\beta}}$$

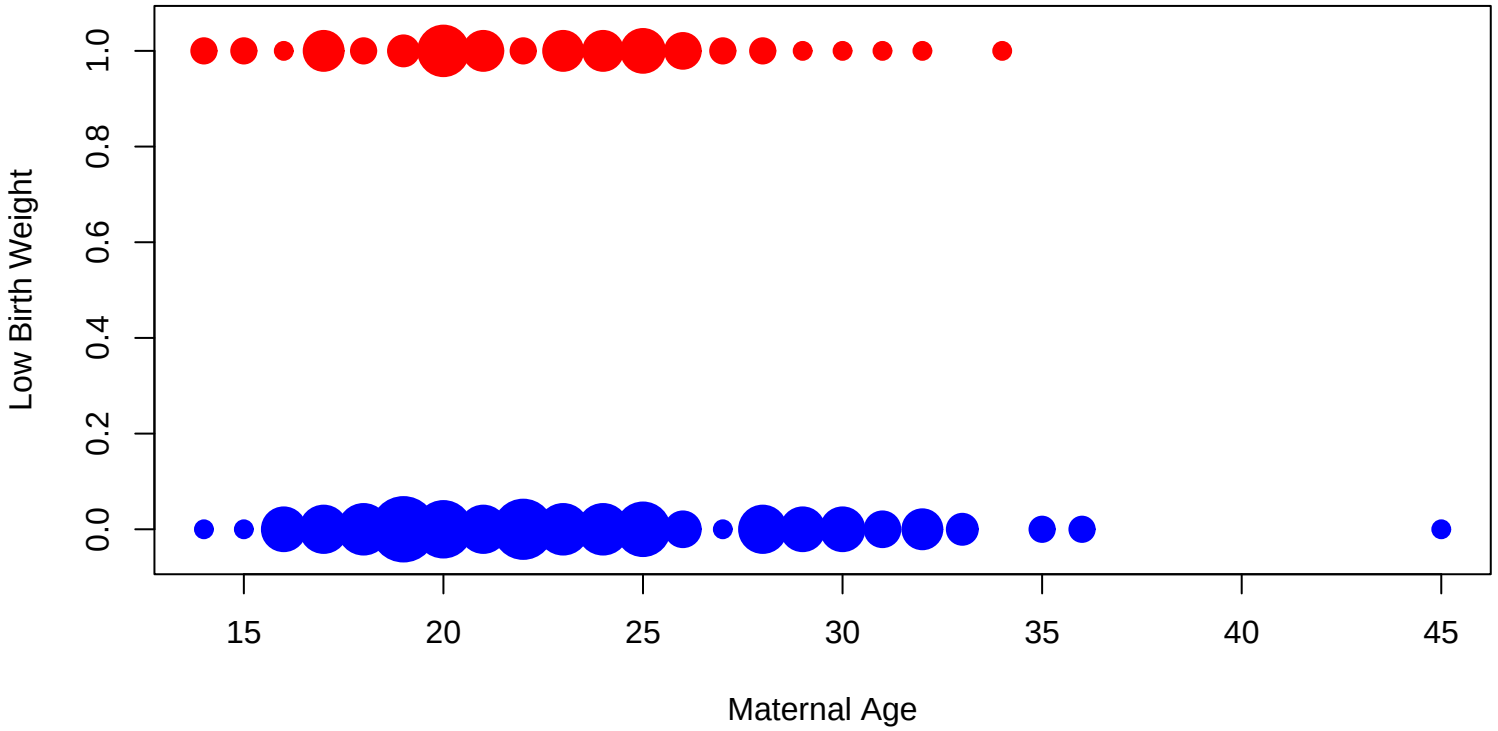
Example: Low Birth Weight in Infants

Hosmer and Lemeshow (1989) look at a data set on 189 births at Baystate Medical Center, Springfield, Mass during 1986, with the main interest being in low birth weight.

`low`: birth weight less than 2.5 kg (0/1)
`age`: age of mother in years
`lwt`: weight of mother (lbs) at last menstrual period
`race`: white/black/other
`smoke`: smoking status during pregnancy
`ptl`: number of previous premature labours
`ht`: history of hypertension (0/1)
`ui`: has uterine irritability (0/1)
`ftv`: number of physician visits in first trimester
`bwt`: actual birth weight (grams)

We will focus on maternal age for now.

This data set is available in **R** in the data frame `birthwt`, though you may need to give the command `library(MASS)` to access it.



Lets fit the model

$$\text{logit}(\pi(age)) = \beta_0 + \beta_1 age$$

in **R** with the function `glm`

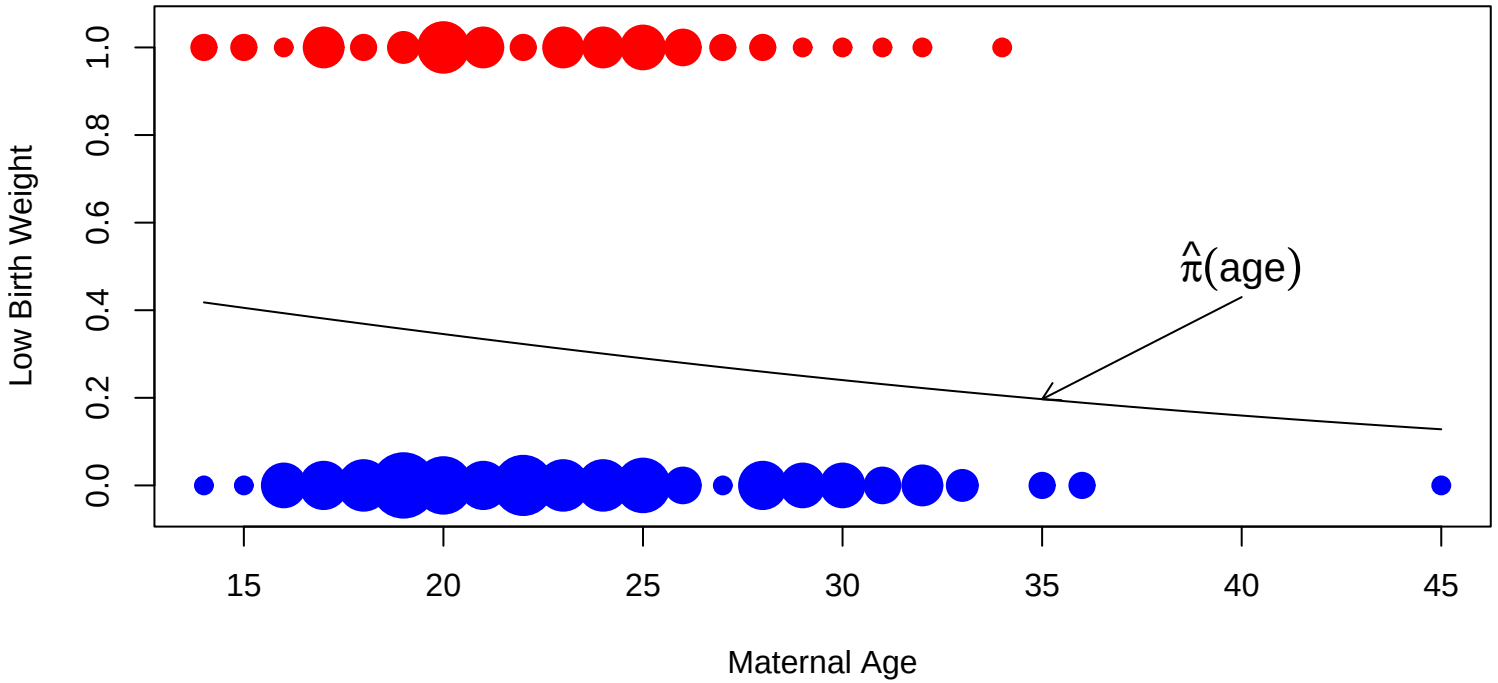
```
> birthwt.glm <- glm(low ~ age, data=birthwt, family=binomial)
> summary(birthwt.glm)

Call: glm(formula = low ~ age, family = binomial, data = birthwt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0402  -0.9018  -0.7754   1.4119   1.7800

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.38458    0.73212   0.525    0.599
age         -0.05115    0.03151  -1.623    0.105

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 231.91  on 187  degrees of freedom
AIC: 235.91
```

The fitted curves are

$$\hat{\eta}(age) = 0.385 - 0.051age$$

$$\hat{\omega}(age) = e^{0.385-0.051age}$$

$$\hat{\pi}(age) = \frac{e^{0.385-0.051age}}{1 + e^{0.385-0.051age}}$$



So in this case, it appears that age increases, the probability/odds of having a low birth weight baby decreases.

What is the effect of changing x on η, ω, and π? Lets see what happens as x goes to $x + \Delta_x$ in the single predictor case.

$$\eta(x + \Delta_x) = \beta_0 + \beta_1(x + \Delta_x) = \beta_0 + \beta_1 x + \beta_1 \Delta_x = \eta(x) + \beta_1 \Delta_x$$

So the log odds work the same way as linear regression. Changing x by one leads to a change in log odds of β_1.

$$\omega(x+\Delta_x) = e^{\beta_0+\beta_1(x+\Delta_x)} = e^{\beta_0+\beta_1 x+\beta_1\Delta_x} = \omega(x)\times e^{\beta_1\Delta_x} = \omega(x)\times(e^{\beta_1})^{\Delta_x}$$

So for this model, the changing x has a multiplicative effect on the odds. Increasing x by 1 leads to multiplying the odds by e^{β_1}. Increasing x by another 1 leads to another multiplication of e^{β_1}.

Another way of thinking of this is through the odds ratio

$$\frac{\omega(x+\Delta_x)}{\omega(x)} = e^{\beta_1 \Delta_x} = (e^{\beta_1})^{\Delta_x}$$

Note that the difference in odds depends on x (through the odds) as

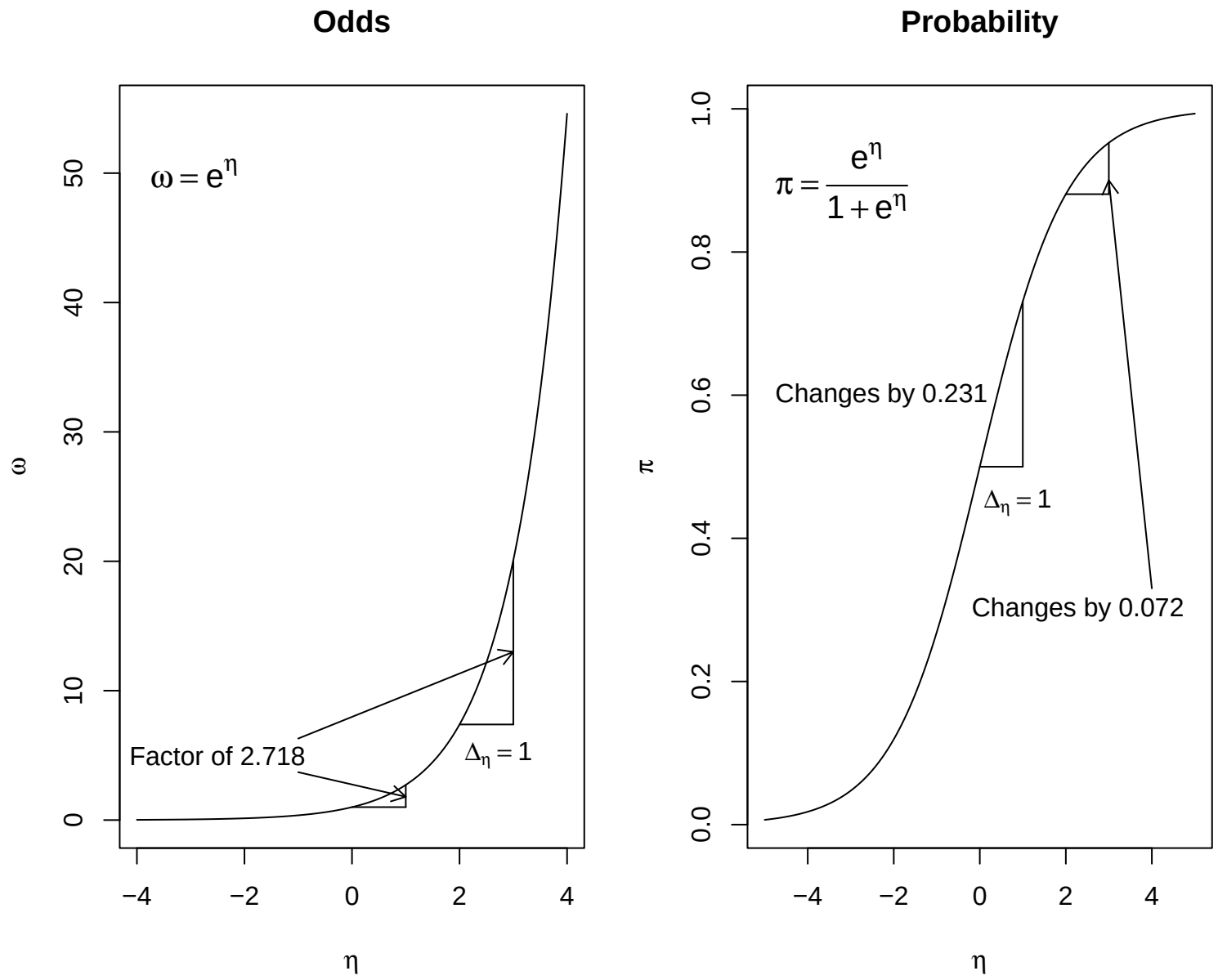
$$\omega(x+\Delta_x) - \omega(x) = \omega(x)(e^{\beta_1 \Delta_x} - 1)$$

So the bigger $\omega(x)$, the bigger the absolute difference.

For π there is not a nice relationship as $\pi(x)$ has an S shape.

$$\pi(x+\Delta_x) = \frac{e^{\beta_0+\beta_1(x+\Delta_x)}}{1+e^{\beta_0+\beta_1(x+\Delta_x)}} = \pi(x) + \text{ something ugly}$$

As can be seen in the following figure, the change $\pi(x)$ depends on $\pi(x)$ and not in a nice way. However the biggest changes occur when $\pi(x) \approx 0.5$ and the size of the change decreases as $\pi(x)$ approaches 0 and 1.


Odds
Probability
$\omega = e^{\eta}$
$\pi = \frac{e^{\eta}}{1+e^{\eta}}$
Factor of 2.718
$\Delta_\eta = 1$
Changes by 0.231
$\Delta_\eta = 1$
Changes by 0.072
ω
π
η
η
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4

These formulas also imply that the sign of β_1 indicates whether $\omega(x)$ and $\pi(x)$ increases ($\beta_1 > 0$) or decreases ($\beta_1 < 0$) as x increases.

So in the example, since $\hat{\beta}_1 =$ -0.051, older mothers should be less likely to have low birth weight babies (ignoring the effects of other predictors).

More precisely, each additional year of age lowers the odds of a low birth weight birth by a factor of $e^{-0.051} = 0.95$ per year.

Actually, there isn't enough evidence to declare age to be statistically significant, but the data does suggest the previous statements.

In the case of multiple predictors, you need to be a bit more careful. If there are no interaction terms in the model, e.g. nothing like

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

then the previous ideas go through if you fix all but one x_j and only allow one to vary.

For example for the model

$$\text{logit}\pi(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

The log odds satisfy

$$\begin{aligned}\eta(x_1 + \Delta_x, x_2) &= \beta_0 + \beta_1(x + \Delta_x) + \beta_2 x_2 \\ &= \beta_0 + \beta_1 x + \beta_1 \Delta_x + \beta_2 x_2 \\ &= \eta(x_1, x_2) + \beta_1 \Delta_x\end{aligned}$$

imply that the odds satisfy

$$\begin{aligned}\omega(x_1 + \Delta_x, x_2) &= e^{\beta_0 + \beta_1(x + \Delta_x) + \beta_2 x_2} \\ &= e^{\beta_0 + \beta_1 x + \beta_1 \Delta_x + \beta_2 x_2} \\ &= \omega(x_1, x_2) \times e^{\beta_1 \Delta_x} \\ &= \omega(x_1, x_2) \times (e^{\beta_1})^{\Delta_x}\end{aligned}$$

However for the interaction model

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

the log odds satisfy

$$\begin{aligned}
\eta(x_1 + \Delta_x, x_2) &= \beta_0 + \beta_1(x + \Delta_x) + \beta_2 x_2 + \beta_{12}(x_1 + \Delta_x)x_2 \\
&= \beta_0 + \beta_1 x + \beta_2 x_2 + \beta_{12} x_1 x_2 + \Delta_x(\beta_1 + \beta_{12} x_2) \\
&= \eta(x_1, x_2) + \Delta_x(\beta_1 + \beta_{12} x_2)
\end{aligned}$$

implying the effect of changing x_1 depends on the level of x_2 as

$$\begin{aligned}
\omega(x_1 + \Delta_x, x_2) &= e^{\beta_0 + \beta_1(x + \Delta_x) + \beta_2 x_2 + \beta_{12}(x_1 + \Delta_x)x_2} \\
&= e^{\beta_0 + \beta_1 x + \beta_2 x_2 + \beta_{12} x_1 x_2 + \Delta_x(\beta_1 + \beta_{12} x_2)} \\
&= \omega(x_1, x_2) \times e^{\Delta_x(\beta_1 + \beta_{12} x_2)}
\end{aligned}$$

Fitting the Logistic Regression Model

One approach you might think of fitting the model would be do to least squares on the data $(x_i, \text{logit}(Y_i))$. Unfortunately this approach has some problems.

- If $X \sim Bin(n, \pi)$ and $\hat{p} = \frac{X}{n}$, then

$$\text{Var}(\text{logit}(\hat{p})) \approx \frac{1}{n\pi(1-\pi)}$$

  so we don't have constant variance.

- If $Y \sim Bin(1, \pi)$, then $\text{logit}(Y)$ equals $-\infty$ or ∞.

While there are ways around this (weighted least squares & fudging the Y_is), a better approach is maximum likelihood.

Maximum Likelihood Estimation

Lets assume that that $Y_1, Y_2, \ldots, Y_n$ are an independent random sample from a population with a distribution described by the density (or pmf if discrete) $f(Y_i|\theta)$. The parameter θ might be a vector $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$.

Then the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(Y_i|\theta)$$

The maximum likelihood estimate (MLE) of θ is

$$\hat{\theta} = \arg\sup L(\theta)$$

i.e. the value of θ that maximizes the likelihood function. One way of thinking of the MLE is that its the value of the parameter that is most consistent with the data.

So for logistic regression, the likelihood function has the form

$$
\begin{aligned}
L(\beta) &= \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i} \\
&= \prod_{i=1}^{n} \left(\frac{\pi_i}{1-\pi_i} \right)^{y_i} (1-\pi_i) \\
&= \prod_{i=1}^{n} \omega_i^{y_i} (1-\pi_i) \\
&= \prod_{i=1}^{n} \left(e^{\beta_0+\beta_1 x_i} \right)^{y_i} \frac{1}{1+e^{\beta_0+\beta_1 x_i}}
\end{aligned}
$$

where $\text{logit}(\pi_i) = \log \omega_i = \beta_0 + \beta_1 x_i$.

One approach to maximizing the likelihood is via calculus by solving the equations

$$\frac{\partial L(\theta)}{\partial \theta_1} = 0, \quad \frac{\partial L(\theta)}{\partial \theta_2} = 0, \quad \ldots, \quad \frac{\partial L(\theta)}{\partial \theta_p} = 0$$

with respect to the parameter θ.

Note that when determining MLEs, it is usually easier to work with the log likelihood function

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

It has the same optimum since $\log$ is an increasing function and it is easier to work with since derivatives of sums are usually much nicer than derivatives of products.

Thus we can solve the score equations

$$\frac{\partial l(\theta)}{\partial \theta_1} = \sum_{i=1}^{n} \frac{\partial \log f(x_i|\theta)}{\partial \theta_1} = 0$$
$$\frac{\partial l(\theta)}{\partial \theta_2} = \sum_{i=1}^{n} \frac{\partial \log f(x_i|\theta)}{\partial \theta_2} = 0$$
$$\cdots$$
$$\frac{\partial l(\theta)}{\partial \theta_p} = \sum_{i=1}^{n} \frac{\partial \log f(x_i|\theta)}{\partial \theta_p} = 0$$

for θ instead.

For logistic regression, the log likelihood function is

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \left(y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i) \right) \\ &= \sum_{i=1}^{n} \left(y_i \log \omega_i + \log(1 - \pi_i) \right) \\ &= \sum_{i=1}^{n} \left(y_i(\beta_0 + \beta_1 x_i) - \log(1 + e^{\beta_0 + \beta_1 x_i}) \right) \end{aligned}$$

Normally there are not closed form solutions to these equations as can be seen from the score equations

$$\frac{\partial l(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} \left(y_i - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) = 0$$

$$\frac{\partial l(\beta)}{\partial \beta_1} = \sum_{i=1}^{n} \left(x_i y_i - x_i \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) = 0$$

These equations will need to be solved by numerical methods, such as Newton-Raphson or iteratively reweighted least squares.

However there are some special cases which we will discuss later where there are closed formed solutions ($\hat{\beta}$ is a nice function of (x_i, y_i))

Key Properties of MLEs

1. For large n, MLEs are nearly unbiased (they are consistent)

2. $\mathrm{Var}(\hat{\theta})$ can be estimated.

   The information matrix $I(\theta)$ is an $p \times p$ matrix with entries

   $$I_{ij} = -\frac{\partial^2 l(\theta)}{\partial\theta_i \partial\theta_j}$$

   Then the inverse of the observed information matrix satisfies

   $$\mathrm{Var}(\hat{\theta}) \approx I^{-1}(\hat{\theta})$$

3. Among approximately unbiased estimators, the MLE has a variance smaller than any other estimator.

4. For large n, the sampling distribution of an MLE is approximately normal.

   This implies we can get confidence intervals for θ_i easily.

5. If $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$, for any "nice" function $g(\cdot)$. (Transformations of MLEs are MLEs - Invariance property.)

   An example of where this is useful is the estimation of success probabilities in logistic regression. So

$$\hat{\pi}(x) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x}}$$

   is the MLE of

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

   So the fitted curve on the earlier plot is the MLE of the probabilities of a low birth weight for ages 14 to 45.

Least Squares Versus Maximum Likelihood in Normal Based Regression

One question some of you may be asking is, if maximum likelihood is so nice, why was least squares used earlier in the book for regression.

If $Y_i|X_i \overset{ind}{\sim} N(X_i\beta, \sigma^2), i = 1, \ldots, n$, then the least squares and the maximum likelihood estimates of β are exactly the same. The log likelihood function in this case can be written as

$$l(\beta, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2$$

So maximizing this with respect to β is the same as minimizing the least square criteria.

The one place where there is a slight difference is in estimating σ^2. The MLE is

$$\tilde{\sigma}^2 = \frac{SSE}{n-p} = \frac{n-p-1}{n-p}\hat{\sigma}^2$$

where $\hat{\sigma}^2$ is the usual unbiased method of moments estimator.

Inference on Individual βs

As mentioned before $\hat{\beta}_j$ is approximately normally distributed with mean β_j and a variance we can estimate. So we can base our inference on the result

$$z = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \overset{approx.}{\sim} N(0,1)$$

Thus an approximate confidence interval for β_j is

$$\hat{\beta}_j \pm z^*_{\alpha/2} SE(\hat{\beta}_j)$$

where $z^*_{\alpha/2}$ is the usual normal critical value.

While **R** doesn't give you these CIs directly, it does give you the information needed to calculate them.

```
> birthwt.glm <- glm(low ~ age, data=birthwt, family=binomial)
> summary(birthwt.glm)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0402  -0.9018  -0.7754   1.4119   1.7800

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.38458    0.73212   0.525    0.599
age         -0.05115    0.03151  -1.623    0.105

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 231.91  on 187  degrees of freedom
AIC: 235.91
```

In addition you can get the $\hat{\beta}$s and standard errors into vectors, and create confidence intervals by the following code

```
> betahat <- coef(birthwt.glm)
> betahat
(Intercept)         age
 0.38458192 -0.05115294
> se.betahat <- sqrt(diag(vcov(birthwt.glm)))
> se.betahat
(Intercept)         age
 0.73212479  0.03151376

> me.betahat <- qnorm(0.975) * se.betahat
> ci.betahat <- cbind(Lower=betahat - me.betahat,
                      Upper=betahat + me.betahat)
> ci.betahat   # 95% approximate CIs
                Lower      Upper
(Intercept) -1.0503563 1.81952014
age         -0.1129188 0.01061290
```

In addition, testing the hypothesis

$$H_0 : \beta_j = 0 \qquad \text{vs} \qquad H_A : \beta_j \neq 0$$

is usually done by Wald's test

$$z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

which is compared to the $N(0, 1)$ distribution.

This is given in the standard **R** output with the `summary` command

```
> birthwt.glm <- glm(low ~ age, data=birthwt, family=binomial)
> summary(birthwt.glm)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.38458    0.73212   0.525    0.599
age         -0.05115    0.03151  -1.623    0.105
```

So in this case age does not appear to statistically significant. However remember there are a number of potential confounders ignored in this analysis so you may not want to read too much into this.

As in regular regression, inference on the intercept usually is not interesting. In this case β_0 gives information about mothers with age $= 0$, a situation that can't happen.