Discrete Response Model
Lecture 2

datascience@berkeley

Probability of Success and the Corresponding Confidence Intervals

Probability of Success

As shown earlier, the estimate for π is

$$\hat{\pi} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \quad + \hat{\beta}_p x_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \quad + \hat{\beta}_p x_p}}$$

To find a confidence interval for π, consider again the logistic regression model with only one explanatory variable x:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x \quad \text{or} \quad \pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Wald Confidence Interval

To find a Wald confidence interval for π, we need to first find an interval for $\beta_0 + \beta_1 x$ (or equivalently for logit(π)):

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm Z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x)}$$

where

$$\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x) = \mathrm{Var}(\hat{\beta}_0) + x^2\mathrm{Var}(\hat{\beta}_1) + 2x\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)$$

and $\mathrm{Var}(\hat{\beta}_0)$, $\mathrm{Var}(\hat{\beta}_1)$, and $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)$ are obtained from the estimated covariance matrix for the parameter estimates.

To find the $(1 - \alpha)100\%$ Wald confidence interval for π, we use the $\exp(\cdot)/[1+\exp(\cdot)]$ transformation:

$$\frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x \pm Z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x)}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x \pm Z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x)}}}$$

Wald Confidence Interval (cont.)

For a model with p explanatory variables, the interval is

$$\frac{e^{\hat\beta_0+\hat\beta_1 x_1+\cdots+\hat\beta_p x_p \pm Z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat\beta_0+\hat\beta_1 x_1+\cdots+\hat\beta_p x_p)}}}{1+e^{\hat\beta_0+\hat\beta_1 x_1+\cdots+\hat\beta_p x_p \pm Z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat\beta_0+\hat\beta_1 x_1+\cdots+\hat\beta_p x_p)}}}$$

where

$$\mathrm{Var}(\hat\beta_0+\hat\beta_1 x_1+\cdots+\hat\beta_p x_p)=\sum_{i=0}^{p} x_i^2\,\mathrm{Var}(\hat\beta_i)+2\sum_{i=0}^{p-1}\sum_{j=i+1}^{p} x_i x_j\,\mathrm{Cov}(\hat\beta_i,\hat\beta_j)$$

and $x_0 = 1$. Verify on your own that the interval given for p explanatory variables is the same as the original interval given for p = 1 explanatory variable.

Profile Likelihood Ratio Interval

Profile LR confidence intervals for π can be found as well, but they can be much more difficult computationally to find than for OR. This is because a larger number of parameters are involved.

For example, the one explanatory variable model $\text{logit}(\pi) = \beta_0 + \beta_1 x$ is a linear combination of β_0 and β_1. The numerator of $-2\log(\Lambda)$ involves maximizing the likelihood function with a constraint for this linear combination.

- The **mcprofile package** provides a general way to compute profile likelihood ratio intervals.
  - Earlier versions sometimes produced questionable results; current versions generally do not have problems.

Recommend using the following approach with this package:

1. Calculate a Wald interval.
2. Calculate a profile likelihood ratio interval with the **mcprofile** package.
3. Use the profile LR interval as long as it is not outlandishly different than the Wald and there are no warning messages given by R when calculating the interval. Otherwise, use the Wald interval.

Example

Consider the model with only distance as the explanatory variable:

$$\text{logit}(\hat{\pi}) = 5.8121 - 0.1150\text{distance}$$

where the results from glm() are saved in the object mod.fit.

Wald Interval

Estimate the probability of success for a distance of 20 yards:

```
> linear.pred<-mod.fit$coefficients[1] +
+     mod.fit$coefficients[2]*20
> linear.pred
(Intercept)
   3.511547
> exp(linear.pred)/(1+exp(linear.pred))
(Intercept)
  0.9710145
```

Use the predict() function

```
> predict.data<-data.frame(distance = 20)
> predict(object = mod.fit, newdata = predict.data, type = "link")
       1
3.511547
> predict(object = mod.fit, newdata = predict.data, type = "response")
        1
0.9710145
```

Example (cont.)

Note that the predict() function is a generic function, so predict.glm() is actually used to perform the calculations. Also, notice the argument value of type = "link" calculates $\hat{\beta}_0 + \hat{\beta}_1 x$ (equivalently, $\text{logit}(\hat{\pi})$).

To find the Wald confidence interval, we can calculate components of the interval for $\beta_0 + \beta_1 x$ through adding arguments to the predict() function:

```
> linear.pred<-predict(object = mod.fit, newdata =
+     predict.data, type = "link", se = TRUE)
> linear.pred
$fit
       1
3.511547

$se.fit
[1] 0.1732707

$residual.scale
[1] 1

> pi.hat<-exp(linear.pred$fit) / (1 + exp(linear.pred$fit))
> CI.lin.pred<-linear.pred$fit + qnorm(p = c(alpha/2, 1- alpha/2))*linear.pred$se
> CI.pi<-exp(CI.lin.pred)/(1+exp(CI.lin.pred))
> CI.pi
[1] 0.9597647 0.9791871
> data.frame(predict.data, pi.hat, lower = CI.pi[1], upper = CI.pi[2])
  distance    pi.hat     lower     upper
1       20 0.9710145 0.9597647 0.9791871
```

The 95% Wald confidence interval for π is **$0.9598 < \pi < 0.9792$**; thus, the probability of success for the placekick is quite high at a distance of 20 yards.

Example: A More General Case

Using the original Wald confidence interval equation again, we can also calculate more than one interval at a time and include more than one explanatory variable. Below is an example using the estimated model.

$$\text{logit}(\hat{\pi}) = 5.8932 - 0.4478\text{change} - 0.1129\text{distance}$$

```
> alpha<-0.05
> linear.pred<-predict(object = mod.fit2, newdata =
+     predict.data, type = "link", se = TRUE)
> CI.lin.pred.x20<-linear.pred$fit[1] + qnorm(p =
+     c(alpha/2, 1-alpha/2)) * linear.pred$se[1]
> CI.lin.pred.x30<-linear.pred$fit[2] + qnorm(p =
+     c(alpha/2, 1-alpha/2)) * linear.pred$se[2]
> round(exp(CI.lin.pred.x20)/(1+exp(CI.lin.pred.x20)), 4)  #CI for distance = 20
[1] 0.9404 0.9738
> round(exp(CI.lin.pred.x30)/(1+exp(CI.lin.pred.x30)), 4) #CI for distance = 30
[1] 0.8493 0.9159
```

Example: Profile Likelihood Ratio Interval

```
> library(mcprofile)
Loading required package: ggplot2
Need help? Try the ggplot2 mailing list: http://groups.google.com/group/ggplot2.
Warning message:
package 'ggplot2' was built under R version 3.2.4
> K<-matrix(data = c(1, 20), nrow = 1, ncol = 2)
> linear.combo<-mcprofile(object = mod.fit, CM = K)  #Calculate -2log(Lambda)
> ci.logit.profile<-confint(object = linear.combo, level = 0.95) #CI for beta_0 + beta_1 * x
> ci.logit.profile

   mcprofile - Confidence Intervals

level:          0.95
adjustment:     single-step

   Estimate lower upper
C1     3.51  3.19  3.87

> names(ci.logit.profile)
[1] "estimate"    "confint"     "CM"           "quant"        "alternative" "level"       "adjust"
> exp(ci.logit.profile$confint)/(1 + exp(ci.logit.profile$confint))
      lower     upper
1 0.9603165 0.979504
```

The 95% interval for π is $0.9603 < \pi < 0.9795$, which is similar to the Wald interval due to the large sample size.

Berkeley SCHOOL OF INFORMATION