Lecture 7: Logistic Regression

1 Birthweight Data

This example will use data studying risk factors for low birth weight in $n = 189$ women. We are interested in modeling the probability of low birth weight as a function of mother's age, mother's weight, mother's race and whether the mother smokes or not. The variables are coded as:

1. low: $1 =$ low birth weight ($<$2500g), $0 =$ not low birth weight
2. age: age of mother in years
3. lwt: weight of mother in pounds
4. race: white, black, other
5. smoke: $1 =$ Yes, $0 =$ No

```
rm(list=ls())
data = read.table("birthweight.txt", header = T)

names(data)
```

```
## [1] "low"   "age"   "lwt"   "race"  "smoke"
```

```
dim(data)
```

```
## [1] 189   5
```

```
attach(data)
summary(data)
```

```
##       low              age             lwt           race
##  Min.   :0.0000   Min.   :14.00   Min.   : 80.0   black:26
##  1st Qu.:0.0000   1st Qu.:19.00   1st Qu.:110.0   other:67
##  Median :0.0000   Median :23.00   Median :121.0   white:96
##  Mean   :0.3122   Mean   :23.24   Mean   :129.7
##  3rd Qu.:1.0000   3rd Qu.:26.00   3rd Qu.:140.0
##  Max.   :1.0000   Max.   :45.00   Max.   :250.0
##      smoke
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3915
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

1.1 Model Fit

```
logit = glm(low ~ age + lwt + smoke, family = binomial(link=logit))
summary(logit)
```

```
##
## Call:
## glm(formula = low ~ age + lwt + smoke, family = binomial(link = logit))
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.2825  -0.8648  -0.6930   1.2620   2.0080
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.36546    1.01528   1.345   0.1787
## age         -0.03949    0.03268  -1.208   0.2269
## lwt         -0.01205    0.00611  -1.972   0.0487 *
## smoke        0.67331    0.32581   2.067   0.0388 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 222.92  on 185  degrees of freedom
## AIC: 230.92
##
## Number of Fisher Scoring iterations: 4
```

```r
logit$coefficients
```

```
## (Intercept)         age         lwt       smoke
##  1.36546519 -0.03948808 -0.01204677  0.67331120
```

```r
exp(logit$coefficients)
```

```
## (Intercept)         age         lwt       smoke
##   3.9175450   0.9612814   0.9880255   1.9607189
```

Looking at the coefficient estimates, we can interpret the effect of each of the predictors on the probability of low birth weight. For example, $exp(\hat{\beta}_{smoke}) = 1.96$ means that a smoker ($smoke = 1$) is almost twice as likely to have a baby with low birth weight compared to a non-smoker, holding other covariates constant. In terms of log odds, being a smoker increases the log odds by $\hat{\beta}_{smoke} = 0.67$.

On the other hand, a year increase in mother's age means that a mother is $1/exp(\hat{\beta}_{age}) = 1.04$ times less likely to have a baby with low birth weight, holding other covariates constant. In terms of the log odds, for every one year increase in age, the log odds of low birth weight decreases by $\hat{\beta}_{age} = -0.04$.

1.2 Inference: Goodness-of-Fit (GOF)

```r
logit$null.deviance - logit$deviance
```

```
## [1] 11.75578
```

```r
qchisq(.05,3,lower.tail=FALSE)
```

```
## [1] 7.814728
```

The difference between the residual deviance and null deviance is used to test H_0 : intercept only model vs. H_a : proposed model. We reject the null of the intercept only model in favor of the model that includes age, lwt, and $smoke$ because $11.76 > \chi^2_3 = 7.81$.

1.3 Inference: Predictor Effects

Referring to the model output, we find that the Wald hypothesis tests indicate that *lwt* and *smoke* are statistically significant at the $\alpha = .05$ level. We can also calculate 95% Wald confidence intervals for the log odds and for the odds for each predictor individually. For example,

```
summary(logit)$coefficients
```

```
##                Estimate  Std. Error   z value   Pr(>|z|)
## (Intercept)  1.36546519 1.015282993  1.344911 0.17865400
## age         -0.03948808 0.032681317 -1.208277 0.22694076
## lwt         -0.01204677 0.006110314 -1.971546 0.04866141
## smoke        0.67331120 0.325810503  2.066573 0.03877440
```

```
lb = logit$coefficients[4] - qnorm(.05/2,lower.tail=FALSE) * summary(logit)$coefficients[4,2]
ub = logit$coefficients[4] + qnorm(.05/2,lower.tail=FALSE) * summary(logit)$coefficients[4,2]
c(lb,ub)
```

```
##      smoke      smoke
## 0.03473435 1.31188805
```

```
exp(c(lb,ub))
```

```
##    smoke    smoke
## 1.035345 3.713178
```

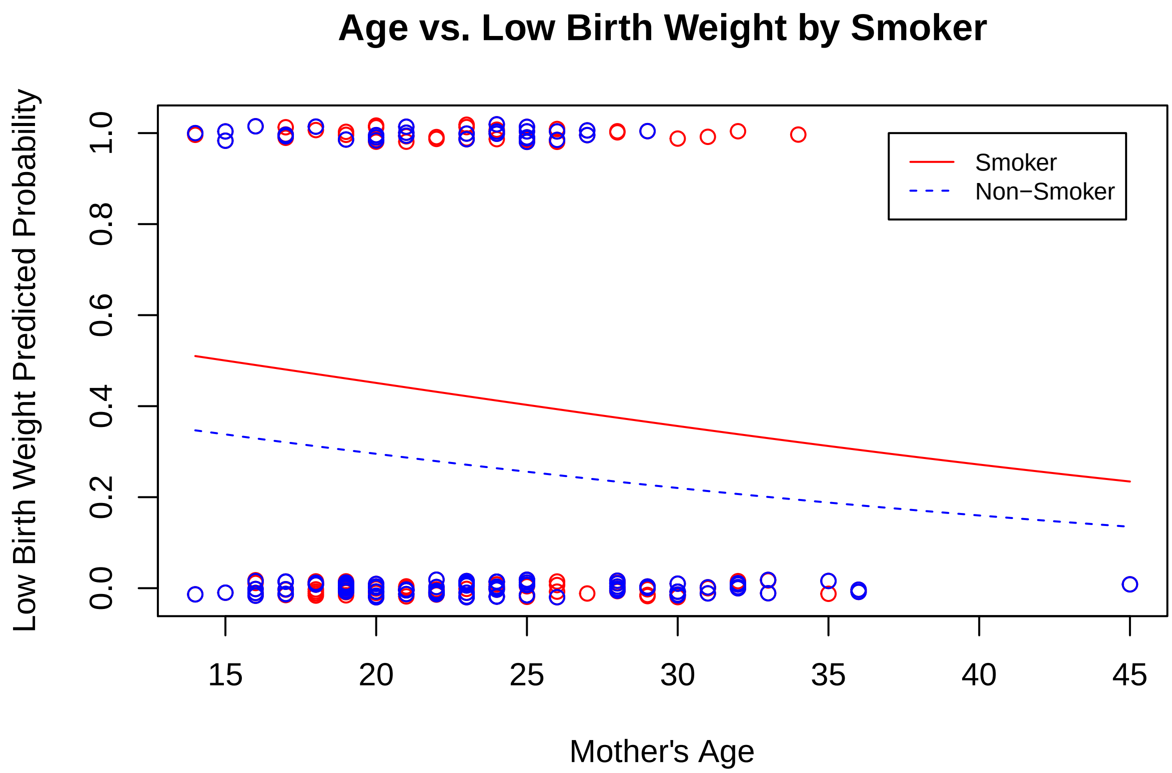
The 95% CI for β_{smoke} is $(0.03, 1.31)$ and for $exp(\beta_{smoke})$ is $(1.04, 3.71)$. Thus we are 95% confident that the interval $(1.04, 3.71)$ contains the true $exp(\beta_{smoke})$. The CI for β_{smoke} does not contain 0 and the CI for $exp(\beta_{smoke})$ does not contain 1, consistent with the finding that the predictor *smoke* is statistically significant.

1.4 Assessing Predictor Effects

We can look at the predicted probabilities to visualize the effects of the covariates. For example, consider the effect of *age* on the probability of a low birth weight baby for a 120 pound smoker. To do this, we evaluate the linear predictor at fixed values of *lwt* and *smoke* over the range of *age* and apply the logit tranformation to obtain the predicted probabilities.

```
# 120 pound smoker #
newX = cbind(1, seq(from=min(age), to=max(age), by=1), 120, 1)
eta = newX %*% logit$coefficients
pred.prob = exp(eta) / (1 + exp(eta))
jitter.low = jitter(low,.1)
plot(age, jitter.low, main = "Age vs. Low Birth Weight by Smoker", xlab = "Mother's Age", ylab = "Low B
points(age[smoke==0], jitter.low[smoke==0], col='blue')
lines(seq(from=min(age), to=max(age), by=1), pred.prob, col='red')

# 120 pound non-smoker #
newX = cbind(1, seq(from=min(age), to=max(age), by=1), 120, 0)
eta = newX %*% logit$coefficients
pred.prob = exp(eta) / (1 + exp(eta))
lines(seq(from=min(age), to=max(age), by=1), pred.prob, col='blue',lty=2)
legend(37, 1, legend=c("Smoker", "Non-Smoker"), col=c("red", "blue"), lty=1:2, cex=.75)
```



1.5 Model Comparison

We can use deviances to compare nested models. For example, should we include race as a covariate?

```r
Black = race == "black"
Other = race == "other"

logit2 = glm(low ~ age + lwt + smoke + Black + Other, family = binomial(link=logit))
summary(logit2)
```

```
##
## Call:
## glm(formula = low ~ age + lwt + smoke + Black + Other, family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5140  -0.9058  -0.5882   1.3039   2.0423
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.325551   1.109218   0.293   0.7691
## age         -0.023248   0.034094  -0.682   0.4953
## lwt         -0.012341   0.006349  -1.944   0.0519 .
## smoke        1.056942   0.379887   2.782   0.0054 **
## BlackTRUE    1.225850   0.516441   2.374   0.0176 *
## OtherTRUE    0.941416   0.416340   2.261   0.0237 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 214.66  on 183  degrees of freedom
## AIC: 226.66
##
## Number of Fisher Scoring iterations: 4
```

```r
diff.dev = deviance(logit) - deviance(logit2)
diff.dev
```

```
## [1] 8.252206
```

```r
qchisq(.95,2)
```

```
## [1] 5.991465
```

The difference between the residual deviances is used to test H_0 : reduced model ($\beta_{black} = \beta_{other} = 0$) vs. H_a : full model ($\beta_{black} \neq 0$ or $\beta_{other} \neq 0$). We reject the null of the reduced model in favor of the model that includes $race$ because $8.25 > \chi^2_2 = 5.99$.