Example: Logistic Regression (Low Birth Weight)

A group of doctors at Baystate Medical Center (in Springfield, MA) sought to determine the factors associated with infant low birth weight (defined as birth weight <2.5 kg). The response variate is a 0/1 indicator for low birth weight (LOW), with the covariates given by the following characteristics of the newborn's mother: age in years (AGE); weight in pounds at last menstrual period (WT); race ("White", "Black", "Other"); smoking status during pregnancy (SMOKE); history of hypertension (HYP); presence of uterine irritability (UI); number of physician visits during the first trimester (FTV); history of premature labor (PTL).

(a) Compute descriptive statistics on all variables.

See the SAS code.

(b) Fit a main effects model based on all covariates and using PROC logistic.

$$
\begin{aligned}
logit(\pi) &= \beta_0 + \beta_1 AGE + \beta_2 WT + \beta_3 I(white) \\
&+ \beta_4 I(black) + \cdots + \beta_9 PTL
\end{aligned}
$$

(c) Carry out the Hosmer-Lemeshow goodness of fit test.

- Use lackfit option
- H_0: the model fits the data well
- HL test statistics: 4.0126, DF= 8
- HL p-value 0.856

- Fail to reject H_0. This implies that the logistic regression model fits the data well.

(d) What would be the impact on $\widehat{\boldsymbol{\beta}}$ of reversing the response variable?

- Model:

$logit(\pi_i) = X_i'\beta$, where $\pi_i = Pr(y_i = 1|X_i)$

- New Model:

$logit(\pi_i^*) = X_i'\beta^*$, where $\pi_i^* = Pr(y_i = 0|X_i)$

Since $logit(\pi_i^*) = logit(-\pi_i) = -logit(\pi_i)$,

$$\beta^* = -\beta$$

(e) Re-fit the main effects model, with $[LOW = 1]$ as the event. Which covariates are predictive of low birth weight?

- Based on Wald test p-values, WT, HYP and PTL are significant.

- You can also use stepwise selection (See the SAS code).

(f) Re-fit the model, with the AGE, RACE and FTV deleted. Interpret the parameter estimate for SMOKE and WT.

- $\widehat{\beta}_{smoke} = 0.5035; \quad \widehat{\beta}_{WT} = -0.0154$
- $\widehat{\beta}_{smoke}$: estimated log odds ratio of LBW between smoking status, adjusting for the other covariates.
- $\widehat{\beta}_{WT}$: estimated log odds ratio of LBW per pound increase in weight, adjusting for the other covariates.

(g) Interpret the intercept, $\widehat{\beta}_0$.

- $\widehat{\beta}_0 = 0.4723$
- $\widehat{\beta}_0$: estimated log odds of LBW when all the covariates = zero.

(h) How would you restructure the design matrix such that the intercept has a more appealing interpretation?

- WT cannot be zero, so use WT - $\overline{WT}$, instead of WT
- $\overline{WT}$ = average value of WT = 130
- Replace WT to WT - 130.

(i) Re-fit the model, carrying out your suggestion for improving the interpretation of the intercept. Compare $\widehat{\beta_0}$ based on the previous and current models, and reconcile any difference.

- Model (with WT)

$$logit(\pi) = \beta_0+\beta_1 WT+\beta_2 Smoke+\beta_3 Hyp+\beta_4 Ui+\beta_5 PTL$$

- Model (with WT -130)

$$\begin{aligned} logit(\pi) &= \beta_0^* + \beta_1^*(WT - 130) + \beta_2^* Smoke + \beta_3^* Hyp \\ &\quad +\beta_4^* Ui + \beta_5^* PTL \\ &= (\beta_0 + \beta_1 130) + \beta_1(WT - 130) + \beta_2 Smoke \end{aligned}$$

$$+\beta_3 Hyp + \beta_4 Ui + \beta_5 PTL$$

- $\widehat{\beta}_0^* = -1.5239$, which is $\widehat{\beta}_0 + 130\widehat{\beta}_{WT}$ of the previous model.

(j) Compare $\widehat{\beta}_{WT}$ based on the previous and current models. Comment.

$\widehat{\beta}_{WT}$s are the same.

(k) Carry out a test of whether the effect of SMOKE depends on either UI or PTL.

- Model: (S: smoke)

$$\begin{aligned} logit(\pi) &= \beta_0 + \beta_1 WT + \cdots + \beta_5 PTL \\ &+ \beta_6 S \times UI + \beta_7 S \times PTL \end{aligned}$$

- H0: $\beta_6 = \beta_7 = 0$

- Wald test statistic: $1.76 < 5.99 = \chi^2_{2,0.95}$

  Fail to reject H_0

(l) Based on the interaction model, interpret the parameter estimate for SMOKE.

- Log odds among smokers when UI = PTL=0

$$logit(\pi_{s=1}) = \beta_0 + \beta_1 WT + \beta_2 S + \beta_3 Hyp$$

- Log odds among non-smokers when UI = PTL=0

$$logit(\pi_{s=0}) = \beta_0 + \beta_1 WT + \beta_3 Hyp$$

- Now

$$\beta_{smoke} = \beta_2 = logit(\pi_{s=1}) - logit(\pi_{s=0})$$

- $\widehat{\beta}_{smoke} = 0.6094$
- $\widehat{\beta}_{smoke}$: estimated log odds ratio of LBW between smoking status when PTL=UI=0, adjusting for the other covariates.

(m) Based on the interaction model, interpret the parameter estimate for SMOKE×PTL parameter.

- Log odds ratio between smoker vs non-smoker when PTL=0 (S=Smoke)

$$\begin{aligned} & logit(\pi_{S=1}) - logit(\pi_{S=0}) \\ & = \beta_S + \beta_{S \times UI} UI \end{aligned} \quad (1)$$

- When PTL=1

$$\begin{aligned} & logit(\pi_{S=1}) - logit(\pi_{S=0}) \\ & = \beta_S + \beta_{S \times UI} UI + \beta_{S \times PTL} \end{aligned} \quad (2)$$

- Now

$$\beta_{S \times PTL} = (2) - (1)$$

or

$$exp(\beta_{S \times PTL}) = exp((2))/exp((1))$$

- $\widehat{\beta}_{S \times PTL} = 0.5245$
- $\widehat{\beta}_{S \times PTL}$: log odds ratio of LBW between smoking status is increased by 0.524 for women with history of PTL.

Use OR

- $\exp(\widehat{\beta}_{S\times PTL}) = 1.690$
- $\exp(\widehat{\beta}_{S\times PTL})$: odds ratio of LBW between smoking status is increased by 69% for women with history of PTL.

(n) Test whether the impact of SMOKE on low birth weight is affected by the weight of the mother.

- Model:

$$\begin{aligned} logit(\pi) &= \beta_0 + \beta_1 WT + \cdots + \beta_5 PTL \\ &+ \beta_6 S \times WT \end{aligned}$$

- Wald test

$$X_w = \frac{0.0174^2}{0.0134^2} = 1.7 < 3.84 = \chi^2_{1,0.95}$$

Fail to reject H_0

(o) Based on this latest interaction model, interpret the SMOKE×WT parameter estimate.

- $\widehat{\beta}_{S\times WT} = 0.0174$
- $\widehat{\beta}_{S\times WT}$: log odds ratio between smoking status is

increased by 0.0174 per pound increase in weight, adjusting for the other covariates.

- $\exp(\widehat{\beta}_{S\times WT}) = 1.018$
- $\exp(\widehat{\beta}_{S\times WT})$: odds ratio between smoking status is increased by 1.8% per pound increase in weight, adjusting for the other covariates.

(p) Based on this latest interaction model, interpret the SMOKE parameter estimate.

- $\widehat{\beta}_S = -1.674$
- $\widehat{\beta}_S$: estimated log odds ratio between smoking status when WT=0, adjusting for the other covariates.

(q) Re-fit the model, using $(WT - 130)$ in place of WT. Compare results with the preceding interaction model and comment on any differences.

- New model

$$\begin{aligned} logit(\pi) &= \beta_0^* + \beta_1^*(WT - 130) + \beta_2 S + \cdots \\ &+ \beta_6^* S \times (WT - 130) \end{aligned}$$

- From the original model

$$\begin{aligned} logit(\pi) &= \beta_0 + \beta_1 WT + \beta_2 S + \cdots + \beta_6 S \times WT \\ &= (\beta_0 + 130\beta_1) + \beta_1(WT - 130) \\ &+ (\beta_2 + 130\beta_6)S + \cdots + \beta_6 S \times (WT - 130) \end{aligned}$$

$\widehat{\beta}^*_{smoke}$ is changed to 0.5934.

$\widehat{\beta}_{smoke \times WT}$ and $\widehat{\beta}^*_{smoke \times WT}$ are the same.

(r) Re-fit the most recent model, this time using PROC GENMOD.

See the SAS code.