**Section 2.2.5 – Interactions and transformations for explanatory variables (continued)**

Quadratic terms

Quadratic and higher order polynomials are needed when the relationship between an explanatory variable and logit(π) is not linear. To include this type of transformation in a formula argument, we can use the carat symbol ^ with the degree of the polynomial. However, as we saw earlier in this section, the carat symbol is used to denote the order of the interaction between explanatory variables. Thus,

formula = y ~ x1 + x1^2

would NOT be interpreted as . R interprets the x1^2 part as “all two-way interactions involving x1.” Because only one explanatory variable is given in x1^2, R interprets this as simply x1. Also, because x1 was already given in the formula argument, the variable is not duplicated, so  would be the model estimated.

To obtain a  term, we need to use the I() function with x1^2. The I() function instructs R to interpret arguments as it normally would. Thus,

formula = y ~ x1 + I(x1^2)

would be interpreted as .

Question: How would you code the model



into a formula argument?

Odds ratios involving polynomial terms are dependent on the explanatory variable of interest. For example, to find OR for x1 in , we form the odds ratio as:



Because the odds ratio is dependent on the explanatory variable value, the interpretation becomes

The odds of a success are  times as large for x1 = \_\_ + c than for x1 = \_\_

where you need to put in the appropriate value of x1. Also, this means multiple odds ratios may be needed to fully understand the effect of x1 on the response.

Wald confidence intervals are found in a similar manner as for interaction terms. For the model of , the interval is



where



Profile likelihood ratio intervals can be calculated as previously described in the notes.

Example: Placekicking (PlacekickDistanceSq.R, Placekick.csv)

This example is based on an exercise in the book. Its program is available on my course website.

Suppose x represents the distance of the placekick. Below is how we can estimate the model



using glm():

> mod.fit.distsq <- glm(formula = good ~ distance +

 I(distance^2), family = binomial(link = logit), data =

 placekick)

> summary(mod.fit.distsq)

Call:

glm(formula = good ~ distance + I(distance^2), family = binomial(link = logit),

 data = placekick)

Deviance Residuals:

 Min 1Q Median 3Q Max

-2.8625 0.2175 0.2175 0.4011 1.2865

Coefficients:

 Estimate Std. Error z value Pr(>|z|)

(Intercept) 7.8446831 1.0009079 7.838 4.59e-15 \*\*\*

distance -0.2407073 0.0579403 -4.154 3.26e-05 \*\*\*

I(distance^2) 0.0017536 0.0007927 2.212 0.027 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

 Null deviance: 1013.43 on 1424 degrees of freedom

Residual deviance: 770.95 on 1422 degrees of freedom

AIC: 776.95

Number of Fisher Scoring iterations: 6

The estimated models is



where x represents distance.

The p-value for the Wald test of H0:β2 = 0 vs. Ha: β2 ≠ 0 is 0.027 suggesting there is marginal evidence of a quadratic relationship between distance and the response. A LRT provides a similar p-value:

> mod.fit <- glm(formula = good ~ distance, family = binomial(link = logit), data = placekick)

> anova(mod.fit, mod.fit.distsq, test = "Chisq")

Analysis of Deviance Table

Model 1: good ~ distance

Model 2: good ~ distance + I(distance^2)

 Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 1423 775.75

2 1422 770.95 1 4.7904 0.02862 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Below is a plot of the estimated model, where I also include the estimate of :



The main difference between the two models appears to be for the larger distances. Given the small number of observations at those distances, it may be difficult to justify the need for the quadratic term. We will investigate this further in later chapters when performing model building to find the best overall model for this data set.

Please investigate on your own how to calculate the odds ratios. Some example code is given in the corresponding program to help you.

Other types of nonlinear representations for explanatory variables are possible. In particular, my book discusses the use of splines in a later chapter. These types of transformations have more flexibility in capturing the nonlinear relationship but at a cost of interpretability.