# MSCR 509 Homework 3

*Anish Shah*

*February 10, 2020*

## Question 1

### Description

A logistic regression model was used to determine the association of low birth weight infants (Y) and maternal age (X1) and the smoking status (X2). The birth weight is coded as 1 (low) and 0 (not low). The variable X1 is continuous and the variable X2 is coded as 1 for smokers and 0 otherwise.

### Answers

**Write down the mathematical form of the logistic regression model. Be sure to define all the variables.**

$$log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$p$ = probability of low birth weight
$\beta_0$ = intercept
$\beta_1$ = coefficient/weight of hte maternal age feature/parameter
$x_1$ = maternal age (continuous)
$\beta_2$ = coefficient/weight of smoking smoking status feature/parameter
$x_2$ = smoking status (binary)

**When Pr(Y=1) is modeled using logistic regression, the regression coefficients associated with maternal age and smoking status are 0.042 (SE=0.029, p-value=0.003) and 0.033 (SE=0.012, p-value=0.006). The intercept is 0.027. Interpret the effects of maternal age and smoking on the probability of a low birth weight infant. Be sure to interpret both the direction and size of the effect.**

With the assumption that hte $\alpha = 0.05$, the $beta_1$ and $beta_2$ coefficients have enough evidence that we can reject the null hypothesis. There is an increased $OR = 1.0429$ for LBW infant for every 1 year increase in maternal age. There is an increase $OR = 1.0336$ for smokers versus non-smokers.

**Assuming this model is appropriate, write down the predictive equation for predicting the probability of a low birth infant in the future.**

$$\hat{p} = \frac{e^{0.027+0.042x_1+0.033x_2}}{1+e^{0.027+0.042x_1+0.033x_2}}$$

**What is the predicted probability of having a low weight baby for a 35 years old woman who is a smoker?**

$P(Y=1) = 0.822$

**What is the predicted probability of not having a low weight baby for a 35 years old woman who is a smoker?**

$P(Y=0) = 1 - P(Y=1) = 1 - 0.822 = 0.178$

# Question 2

## Description

The treatment regimen for patients who have been diagnosed as having cancer of the prostate is crucially dependent upon whether or not the cancer has spread to the surrounding lymph nodes. Indeed, a laparotomy (a surgical incision into the abdominal cavity) may be performed to ascertain the extent of this nodal involvement. However, there are a number of variables that are indicative of nodal involvement that can be measured noninvasively, and the aim of a study reported by Brown (1980) was to determine whether a combination of variables could be used to forecast whether or not the cancer has spread to the lymph nodes. The response variable (Y) is the presence or absence of nodal involvement (Y: 1 = present, 0 = absent). The prognostic variables considered are:

- AGE - age of patient at diagnosis (in years)
- ACID - level of serum acid phosphatase (in King-Armstrong units)
- XRAY - the result of an X-ray examination (0 = negative, 1 = positive)
- SIZE - the size of the tumor as determined by a rectal examination (0 = small, 1 = large)
- GRADE - a summary of the pathological grade of the tumor determined from a biopsy (0 = less serious, 1 = more serious)
- SES - income (high, middle, low)

## Responses

**What is the proportion of subjects with presence of nodal involvement?**

Of the 92 patients, there are 36 patients with nodal involvement. This leads to a proportion of 0.3913.

**Write down the three estimated logistic regression equations for modeling the presence of nodal involvement (Y) in terms of independent variables: AGE (model 1), SES (model 2) and {AGE, SES} (model 3). [Use low income as reference for SES]**

These are the theoretical models for the regression.

$$(1) : log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 AGE$$
$$(2) : log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_2 SES_{middle} + \beta_3 SES_{high}$$
$$(3) : log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 AGE + \beta_2 SES_{middle} + \beta_3 SES_{high}$$

Hypotheses: $H_0 : \beta_i = 0$ and $H_1 : \beta_i \neq 0$, with an $\alpha = 0.05$. After estimating them, we can generate $\beta$ for the models. For SES, the *low* level is considered the reference level.

$$(1) : log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = 2.3791 + -0.0477 AGE$$
$$(2) : log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = -0.1911 + 0.5965 SES_{middle} + -1.4184 SES_{high}$$
$$(3) : log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = 2.2527 + -0.0412 AGE + 0.5612 SES_{middle} + -1.4079 SES_{high}$$

**What is the -2 log L for each of these three models? Please specify the number of parameters associated with each model. SAS output provides a likelihood ratio test for each of these models - write down the null hypothesis corresponding to this likelihood ratio test.**

$$(1) : -2logL = 121.3324\ (df = 2), H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$$
$$(2) : -2logL = 111.797\ (df = 3), H_0 : \beta_2, \beta_3 = 0, H_1 : \beta_2, \beta_3 \neq 0$$
$$(3) : -2logL = 110.5812\ (df = 4), H_0 : \beta_1, \beta_2, \beta_3 = 0, H_1 : \beta_1, \beta_2, \beta_3 \neq 0$$

**We plan to compare model 3 vs. model 2. In other words, we would like to test the hypothesis that the coefficient corresponding to AGE is zero. Construct a likelihood ratio test by hand to test this hypothesis - include all steps, and be sure to specify your significance level. What is your conclusion? Provide Wald test statistic and p value for age in the presence of SES.**

$$(3) : log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 AGE + \beta_2 SES_{middle} + \beta_3 SES_{high}$$
$$(2) : log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_2 SES_{middle} + \beta_3 SES_{high}$$

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$
$\alpha = 0.05$

In the nested models above (model 2 and model 3), we can assess hypothesis if age is a significant parameter in predicting lymph node spread.

1. Model 3 has a $-2logL = 110.5812$, and model 2 hasa $-2logL = 111.797$.
2. The difference in log likelihood is **1.2157**. With 1 degree of freedom, and an $\alpha = 0.05$, the corresponding $\chi^2 = qchisq(0.95, df = 1)$.
3. The likelihood ratio test results, compared to our specificied $\chi^2$ cut-off, is $1.2157 < 3.8415$.
4. Based on the Wald test statistic, we have a value of 1.1992, df=1, and $p = 0.2735$, which is also less then our $\chi^2$ value.
5. We do not have sufficient evidence to reject $H_0$, and we can thus conclude that age is not needed in this model.

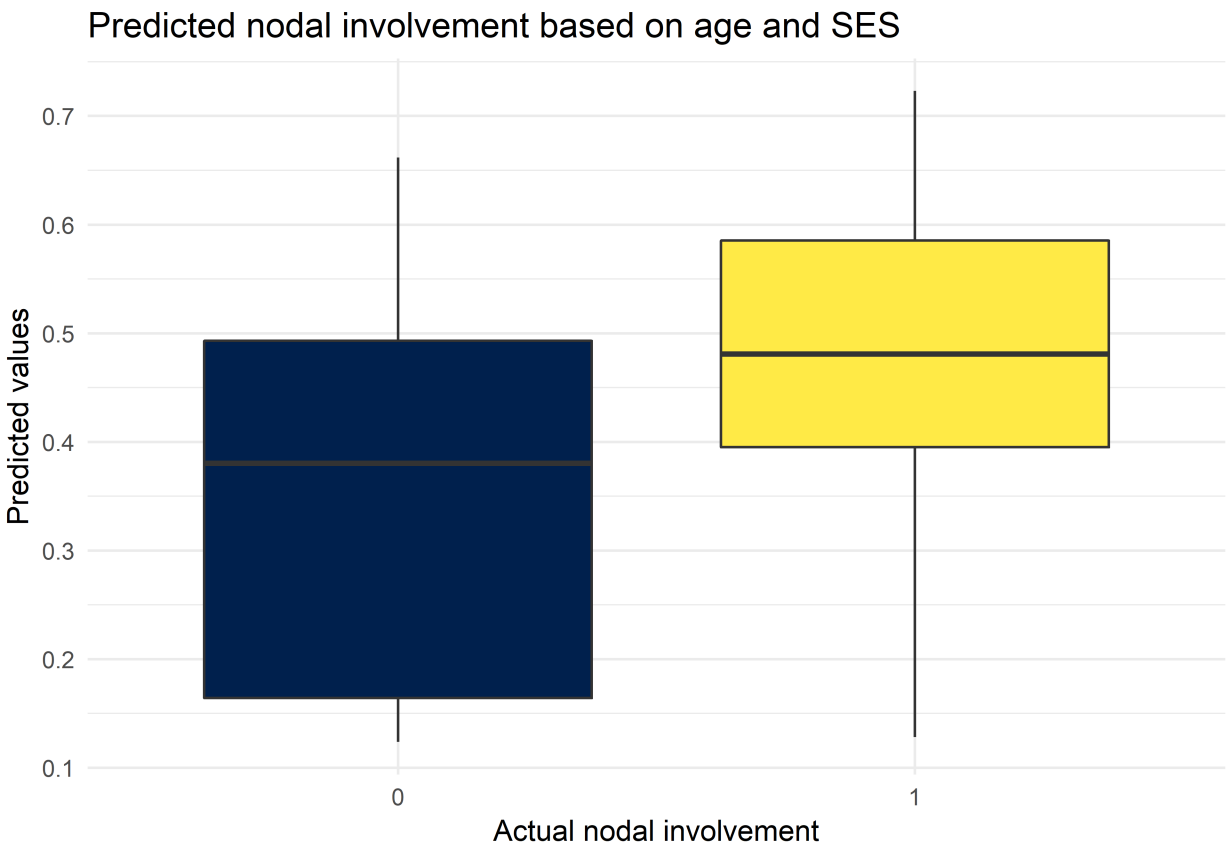
**We would like to compare model 1 and model 2. Can we perform a likelihood ratio test? Please be specific in your explanation. What else can be used to compare these two models? Compare the models and state your conclusion.**

$$(2) : log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_2 SES_{middle} + \beta_3 SES_{high}$$
$$(1) : log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 AGE$$

The features/parameters in the above models are not nested. Thus, we cannot use the likelihood ratio tests. If the null hypothesis were true, these models would not be the same. Non-nested testing is needed. We can instead compare the models using the *Akaike information criteria* (AIC).

For model 2, the $AIC = 117.797$. For model 1, the $AIC = 125.3324$. As model 2 has a lower AIC, this is suggestive that model 2 is a better model, however this is not a statistical test. It just suggests that the fit for model 2 is better than model 1.

**Use SAS to create a boxplot for predicted P vs observed from model 3. Attach your boxplot with homework.**




**Use SAS and model 3 to report predicted p for two patients: 55 year old with high income and 55 year old with middle income.**

$$(3) : log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 AGE + \beta_2 SES_{middle} + \beta_3 SES_{high}$$

With the appropriate $\beta$ values...

$$(3) : log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = 2.2527 + -0.0412 AGE + 0.5612 SES_{middle} + -1.4079 SES_{high}$$

For a patient that is 55 years old and has a high-income class, the predicted value is $p = -1.421$. For a patient that is 55 years old and has a middle-income class, the predicted value is $p = 0.5479$.