Biostat/Stat 571 Exercise #1

Answer Key/Comments

Question 2:

The King county data `KingCounty2001.data` contain a sample of size 2500 of data from the birth certificates for children born in King County WA in 2001. The data are restricted to singletons (i.e. no twins or triplets).

(a) We wish to summarize the evidence for the effectiveness of the *First Steps* program, designed to provide "maternity care necessary for to ensure healthy birth outcomes for low-income families".

The following is a general strategy that one could adopt:

- Examine closely the scientific question of interest as this will likely help considerably with some of the decisions you will be faced with below.
- Classify each of the covariates; outcome, predictor of interest, adjustment variables.
- Look at univariate and bivariate summaries
  - o this will help distinguish confounders, precision and nuisance
  - o variables (according to the data at least)
- Pick three models:
  (1) No adjustment
  (2) Adjusting for confounders identified by a priori consideration
  (3) Adjusting for additional (potential) confounders that seem reasonable.
- Model diagnostics for the specific model that we use to answer the question of interest

Since the main issue is whether or not a baby has low birth weight, I decided to dichotomise the continuous measure of birth weight (at 2500gm) and use this binary indicator as the outcome of interest. Given this and the prospect of probably needing to adjust for potential confounders of the outcome/First Steps association, this suggests that the model that we adopt will likely, be a logistic regression.

Before continuing it is important to consider the two groups that are being compared when we look at enrollement vs non-enrollement into the First Steps program. In particular, the First Steps program is aimed at "low income" mothers. Thus, those not enrolled in the program either (a) do not meet the eligibility criteria and presumiably in a "high" income bracket or (b) do meet the criteria but were not enrolled for some reason. Ideally, to assess the impact of the First Steps program we would perform the assessment amoung mothers of the same socio-economic status. Thus, in any analysis it will be necessary to include measures of income in the model. However, in the dataset we don't have such information and consequently we have to make do with two surrogate variables; welfare participation and education. This can be done by either including the variables in any model or restricting the analysis to an appropriate subset of the 2500 observations. In this analysis, adjustments were made by including education as a 3-level factor variable where the levels correspond to an lower than high school education, a high school education, and a post-high school (college) education. To address the impact of the First Steps program (in the "low income") I have included an interaction between participation in the program and the categorised education variable. In particular, the assessment will be based on the estimate of the impact of First Steps amoung the low (less than high school) education group.

Covariate		Weight > 2500gm	Weight ≤ 2500gm	Combined
		n = 2363	n = 127	n = 2500
First Steps Program	Yes	378 (15.9)	25 (19.7)	403 (16.1)
Gender	Female	1146 (48.3)	63 (49.6)	1209 (48.4)
Age (yrs)		30 (25,34)	29 (24,34)	30 (25,34)
Race	Asian	368 (15.5)	24 (18.9)	392 (15.7)
	Black	161 (6.8)	17 (13.4)	178 (7.1)
	Hispanic	211 (8.9)	9 (7.1)	220 (8.8)
	Other	30 (1.3)	1 (0.8)	31 (1.2)
	White	1603 (67.6)	76 (59.8)	1679 (67.2)
Married	Yes	1877 (79.1)	79 (62.2)	1956 (78.2)
Welfare Program	Yes	36 (1.5)	6 (4.7)	42 (1.7)
Smoking Status	Yes	154 (6.5)	21 (16.5)	175 (7)
Drinker ?	Yes	28 (1.2)	1 (0.8)	29 (1.2)
Prior Weight (lbs)		140 (125,161)	136 (115,153)	140 (125,160)
Weight Gain (lbs)		32 (25,40)	25 (20,30)	31 (25,40)
Education	Less than HS	259 (10.9)	18 (14.2)	277 (11.1)
	HS	447 (20.0)	33 (26.0)	510 (20.4)
	College	1637 (69.1)	76 (59.8)	1713 (68.5)
Gestation (wks)		39 (38,40)	35 (32,37)	39 (38,40)
Parity	0	1095 (46.1)	71 (55.9)	1166 (46.6)
	1	835 (35.2)	34 (26.8)	869 (34.8)
	2	294 (12.4)	13 (10.2)	307 (12.3)
	≥ 3	149 (6.3)	9 (7.1)	158 (6.3)

Table 1: Univariate and Bivariate (vs outcome) summaries. For factor variables, counts and percentages (within columns) are given. For continuous variables, medians and inter-quartile ranges are provided.

Initially, we can look at some descriptives for the covariates both for the cases (i.e. those baby's whose weight is less than 2500gm) and controls (i.e. those baby's whose weight is greater than 2500 gm), as well as the combined group. From Table 1. it seems that the set of variables which seem to be univariately associated with the outcome contain marital status, smoking status, prior weight, weight gain, gestation period and the parity. Thus, this set was used as an intermediary set of adjustment variables. Table 2 provides the results of the three logistic regression models outlined above. In each case, the point estimate for the regression coefficient and associated standard error are provided. In addition, the likelihood ratio test statistic and associated p-value, to test the hypothesis of no treatment (First Steps) effect, are provided.

Model 1, which provides a crude analysis, indicates that for a low education (i.e. target population) mother who is enrolled in the First Steps program, the odds of her baby having a low birth weight are

Education Category	Model 1		Model 2		Model 3.	
	Odds Ratio	Approx. 95% CI	Odds Ratio	Approx. 95% CI	Odds Ratio	Approx. 95% CI
LRT[a]	3.882 (0.275)		2.523 (0.471)		2.539 (0.468)	
Less than High School	1.16	(0.44, 3.09)	1.69	(0.41, 6.97)	1.93	(0.46, 8.14)
High School	1.44	(0.61, 3.39)	1.39	(0.50, 3.90)	1.26	(0.43, 3.67)
College	0.53	(0.27, 1.03)	0.57	(0.24, 1.33)	0.57	(0.24, 1.36)

[a] Likelihood Ratio Test statistic and p-value (based on the χ^2_3 distribution).

Table 2: Logistic Regression results.

estimated to be approximately 16% higher than the corresponding odds for a low education mother that is not enrolled in the First Steps program. Model 2 indicates, that holding marital status, smoking status, prior weight, weight gain, gestation period and the parity constant, the odds for a low education mother in the First Steps program are estimated to be 69% higher than for a low education mother not enrolled. Here, by '*holding constant*' this selection of variables we mean that we are comparing two women with the same characteristics for this list of variables. Evidently, the 95% confidence intervals are very wide so that it is not clear to what extent we can depend on the point estimates for interpretation.

Overall, it is clear that there is no evidence in this data to support a positive impact on birth weight outcomes for the First Steps program.

Comments :

- Q: Why don't we worry about overdispersion ? Since each of the outcomes is the results of a Bernoulli trial, there is no potential for overdispersion in this case. For a Bernoulli(μ) random variable the variance is given by $\mu(1-\mu)$.
- In the above model, I haven't really considered any interactions between covariates that might be of interest. Typically, you could either pick a few interactions which might be relevant (such as the impact of the First Steps program depending on race)or, preferably, ask an "expert" in the scientific field.
- The variable selection processes may not be the best or most appropriate. As in the above comment, typically you would want to consider a more scientifically driven variable selection process. In this case it doesn't seem like it would make alot of difference though.
- Although not presented here, there are many diagnostic tools for logistic regression. Hosmer and Lemeshow (2002) is a great reference for these. They include assessing the appropriateness of the link, functional form for covariates, residual diagnostics, and influence statistics. There are also a series of techniques which compare observed values from the data and expected values from the model.

(b) Here, we wish to build a rule which will be used to predict whether a baby will be born with a weight of less than 2500 grams. Regardless of the choice of model-building technique, the model will need to be validated (internally). There are three main ways of acheiving this: *data-splitting*, *cross-validation*, and *bootstrapping*. See Section 6 of Harrell *et al* (1996) for more details. Here, I am going to use a single instance of data-splitting to illustrate the need for validation. In particular, of the 2500 observations I used 1500 as a training (model-building) data set, and the remaining 1000 for the validation dataset. After selected a subset of 1500 observations on which to build the data set, we can use the S-Plus function `step.glm` to find the model. This model is based on variables which will be available early on in the pregnancy (i.e. variables such as weight gain and gestation are not included). Note, since there were so few observations in the 'Other' race category, these were combined with the 'Hispanic' race group.

```
> summary(model.step, cor = F)
Call: glm(formula = low ~ newrace + married + smoker + wpre, family = binomial, data = king[ - index,  ], na.action = na.omit)
Coefficients:
                   Value  Std. Error    t value
 (Intercept)  1.762398525 0.639547005  2.7556982
newraceasian -0.620674535 0.316362310 -1.9619105
newraceblack -1.052856810 0.374909363 -2.8082969
newraceother -0.166300247 0.414375227 -0.4013277
     married  0.491241711 0.284013728  1.7296407
      smoker -1.064650639 0.361176890 -2.9477264
        wpre  0.008406429 0.003981048  2.1116118

(Dispersion Parameter for Binomial family taken to be 1 )
    Null Deviance: 595.5457 on 1499 degrees of freedom
Residual Deviance: 564.9728 on 1493 degrees of freedom
```

From this model we may compute a predicted probability of a baby being of low birth weight. Typically, we would then compare this fitted probability to some pre-specified cutoff. If the probability is greater than the cutoff, then we predict that the baby will have low birth weight. We could conceivably choose several criteria by which we can assess the predictive ability of a model. Again, Harrell *et al* (1996) provides many details. Here, I am going to use sensitivity = P(rule indicates baby has weight less than 2500gm | baby's actual weight is less than 2500gm) and specificity = P(rule indicates baby has weight greater than 2500gm | baby's actual weight is greater than 2500gm) as my criteria. Figure 1 provides the ROC curves (sensitivity vs 1-specificity), when the prediction rule given by the above model is applied to the training and the validation data.

The main issue in deciding upon a cutoff is the balance between sensitivity and specificity. One possiblity is to look at a weighted sum of the two quantities. This allows you to attach different weights to the sensitivity and specificity depending on which you felt to be more important. Typically, there are many factors which would be involved in a decision regarding weights. In particular, it is likely that an "expert" in the field would want to be consulted about relative costs of different missclassifications. Then, translation of these relative costs into weights will need to be done. This is unlikely to be easy and would require quite careful thought. Here, it seems that sensitivity would be the most important of the two. We would not want to missclassify a baby that ultimately has a low birth weight, while missclassifying a baby that is ultimately an acceptable weight is not as bad. Table 2 provides "optimal" cutoffs based on three different (and fairly arbitrary) weighting schemes. The first scheme treats sensitivity and specificity equally. The third treats sensitivity as twice as 'important' as specificity. The second scheme is intermediary.

Table 3 provides the sensitivity and specificity measures for the validation set for three potential cutoff points based on the above table. We can see that when applied to the validation set, the decision rule has better sensitivity but worse specificity. However, we have to be aware that these results depend on the way that the original dataset was split in two. Another configuration of the training and validation datasets would have resulted in different final results. This variabilty in the results can be address using either cross-validation or bootstrapping mentioned above.

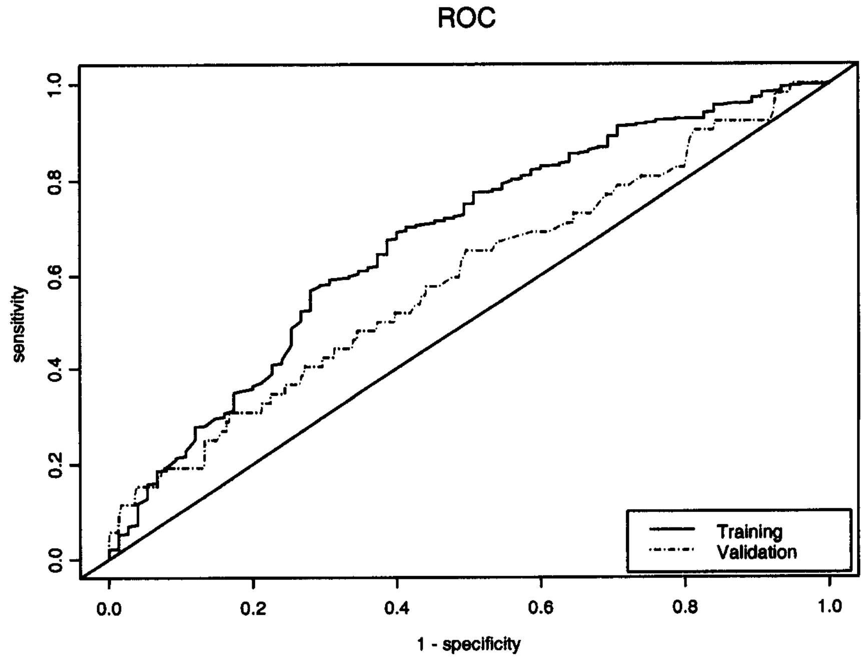

Figure 1: ROC Curves for the King County Birth 2001 data.

Weights = (1, 1)			Weights = (1.5, 1)			Weights = (2, 1)		
Cutoff	Sensitivity	Specificity	Cutoff	Sensitivity	Specificity	Cutoff	Sensitivity	Specificity
0.94755	69.3%	60.0%	0.90580	91.3%	29.3%	0.90580	91.3%	29.3%
0.94628	70.3%	58.7%	0.90663	91.2%	29.3%	0.90663	91.2%	29.3%
0.94960	67.6%	61.3%	0.90692	91.0%	29.3%	0.90692	91.0%	29.3%
0.94771	68.9%	60.0%	0.90736	90.8%	29.3%	0.90736	90.8%	29.3%
0.94636	70.2%	58.7%	0.93626	77.5%	49.3%	0.90477	91.4%	28.0%

Table 3: Cutoffs and associated sensitivity/specificity chosen on the basis of a weighted sum of the two probabilties.

(c) Suppose that the data that we have available to analyse resulted from a biased sampling scheme. In particular, where we sampled women who were not enrolled in the First Steps program (`firsteps = 0`) with probability 1/10, while women who did enroll in the program (`firsteps = 1`) were sampled with probability 1/3.

- One way to view this situation is as a missing data problem. In particular, there is missing data (to different degrees in the two groups defined by enrollement into the program) and the missingness is by *design* of the study. Here the missingness mechanism depends solely on enrollement into the First Steps program. In other words, the missingness is MAR (Missing At Random). In this case, as long as we account for the covariates that drive the missingness (i.e. we condition on them)

Cutoff	Sensitivity	Specificity
0.95	87.7%	19.2%
0.93	93.6%	15.4%
0.90	98.2%	11.5%

Table 4: Sensitivity and Specificity based on three cutoffs applied to the validation data set.

and as long as we model the mean correctly inference based on the observed data is valid.

- In part (a), since we are interested in the `firsteps` effect (i.e. it is in the model) we would not need to change our analysis (as long as we are happy with the mean model).
- In part (b), it would not necessarily be the case that we include enrollement into the First Steps program (especially given our analysis in part (a)). To adjust for the sampling scheme, we then need to incorporate weights where the weights are equal to the inverse of the probabilty of selection (i.e. 10 for the non-enrollees and 3 for the enrollees).
- One intuitive (perhaps?) was of thinking about the weighting is as follows. Consider the case where we sampled from each of the groups with probability equal to 1. This would effectively be equivalent to simple random sampling where the proportion of non-enrollees to enrollees in the sample is representative of the true proportion in the population. One can think of our sampling scheme as resulting in a dataset which contains 1/10 of the non-enrollees and 1/3 of the enrollees. Thus the proportion is no longer representative. By weighting each of the non-enrollees by 10 in the likelihood we are effectively creating 10 copies of each non-enrollee (similarly for the enrollees). The likelihood then contains contributions from the two groups in proportion to the contributions that would have been made via simple random sampling. One can think of this as creating some pseudo-population upon which our inference is based.
- We saw that for part (a), as long as the mean is modelled correctly, weighting adjustments are not strictly necessary. However, in general, it is probably best to use a weighting scheme.
    - More efficient estimation ?
    - More robust to mean model misspecification ?

S-Plus Code

```
options( contrasts = c("contr.treatment", "contr.poly") )
library( mass )
##
#### Source in any additional code. Please send me an e-mail if you would like copies of these functions.
##
source( "C:\\TA\\SPlus_Functions\\TableOne.SSC" )
source( "C:\\TA\\SPlus_Functions\\LogisticRegressionDiagnostics.SSC" )
source( "C:\\TA\\SPlus_Functions\\DiagnosticModels.SSC" )

##
#### Load in the data
##
king <- read.table( "C:\\TA\\571_W03\\data\\KingCounty2001_data.txt" )
names( king ) <- c( "gender", "plural", "age", "race", "parity", "married", "bwt", "smokeN",
"drinkN", "firstep", "welfare", "smoker", "drinker", "wpre", "wgain", "educ", "gest" )

king$firstep <- factor( king$firstep, labels = c("No", "Yes") )
king$welfare <- factor( king$welfare, labels = c("No", "Yes") )
king$married <- factor( king$married, labels = c("No", "Yes") )
king$newpar <- king$parity
king$newpar[king$newpar > 3] <- 3
king$newpar <- factor( king$newpar )
king$low <- factor( cut(king$bwt, breaks = c(0, 2500, 10000), labels = c("Yes", "No")) )
```

```
################
#### PART A ####
################

##
#### Univariate and Bivariate (according to treatment) summaries
##
## GetTableOne( king[,c(10,1,3,4,6,8,9,11,12,13,14,15,16,17,18)], king$low, c("No", "Yes") )

##
#### From these we can pick a subset of covariates which 'seem' associated with the outcome.
#### - marital status, smoking status, prior weight, weight gain, gest, parity.
#### This could have been done in any number of ways. For example, some stepwise procedure.
#### Without the assistance of a specialist our choices at this stage are fairly arbitrary.
##

##
#### Look at welfare vs education
##
boxplot( split(king$educ, king$welfare), xlab = "Welfare participation", ylab = "Education level" )
king$educcat <- rep( 0, 2500 ) # less than a high school education
king$educcat[king$educ == 12] <- 1 # high school education
king$educcat[king$educ > 12] <- 2 # more than a high school education
king$educcat <- factor( king$educcat, labels = c("Low", "HighSchool", "College") )

##
#### Model 1: No adjustment
##
model.1 <- glm( low ~ firstep*educcat, family = binomial, data = king )
model.1.null <- glm( low ~ educcat, family = binomial,data = king )
##
#### Model 2: Adjustments; married smoker wpre wgain newpar
##
model.2 <- glm( low ~ firstep*educcat + married + smoker + wpre + wgain + newpar + gest,
family = binomial, data = king )
model.2.null <- glm( low ~ educcat + married + smoker + wpre + wgain + newpar + gest,
family = binomial, data = king )
##
#### Model 3: Adjustments for all covariates
##
model.3 <- glm( low ~ firstep*educcat + gender + age + race + married + smoker + drinker + wpre + wgain + gest + newpar,
family = binomial, data = king )
model.3.null <- glm( low ~ educcat + gender + age + race + married + smoker + drinker + wpre + wgain + gest + newpar,
family = binomial, data = king )
##
#### Look at the results from the three models
##
lrt.stat.1 <- deviance(model.1.null) - deviance(model.1)
lrt.stat.2 <- deviance(model.2.null) - deviance(model.2)
lrt.stat.3 <- deviance(model.3.null) - deviance(model.3)
#
round( rbind( c(lrt.stat.1, 1 - pchisq(lrt.stat.1, 3)),
c(lrt.stat.2, 1 - pchisq(lrt.stat.2, 3)),
c(lrt.stat.3, 1 - pchisq(lrt.stat.3, 3)) ), 3 )
#
cbind( rbind( GetORs(c(0,1,0,0,0,0), model.1),
GetORs(c(0,1,0,0,1,0), model.1),
GetORs(c(0,1,0,0,0,1), model.1) ),
 rbind( GetORs(c(0,1,0,0,0,0,0,0,0,0,0,0,0,0), model.2),
GetORs(c(0,1,0,0,0,0,0,0,0,0,0,0,1,0), model.2),
GetORs(c(0,1,0,0,0,0,0,0,0,0,0,0,0,1), model.2) ),
 rbind( GetORs(c(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0), model.3),
  GetORs(c(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0), model.3),
GetORs(c(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1), model.3) ) )


################
#### PART B ####
################

##
#### We want to come up with a rule that we can use to help predict if a baby will
#### be classified as 'low birth weight', ie. bwt < 2500gms.
#### Need to pick 'reasonable' variables for the model. This probably won't include weight gain
#### gestation. Also, it is not clear if some of these variables are measured when the
#### baby is born. Ideally we would have a model that is based upon covariates
#### measured at the same time at which we would apply the prediction rule.
##
#### Note: race is the same as the old race except that hispanic and other have been combined.
```

```
####      newpar is also re-coded so that all mothers with parity >= 2 are together.
##
king$newrace <- as.numeric( king$race )
king$newrace[king$newrace == 3] <- 4
# re-code the whites to be the reference
king$newrace[king$newrace == 5] <- 0
king$newrace <- factor(king$newrace, labels = c("white", "asian", "black", "other"))
#
king$newpar2 <- as.numeric( king$newpar ) - 1
king$newpar2[king$newpar2 > 2] <- 2
king$newpar2 <- factor( king$newpar2 )

##
#### Generate a validation set
##
index <- sample( 1:2500, 1000 )
# To distinguish training and validation datasets
sub.weights <- rep(1, 2500)
sub.weights[index] <- 0


##
#### Get the prediction model
##
model.full <- glm( low ~ newrace + married + welfare + smoker + wpre + educ + newpar2,
family = binomial, data = king, weights = sub.weights, na.action = na.omit )
model.step <- stepAIC( model.full,
 scope = list(upper = ~ newrace + married + welfare + smoker + wpre + educ + newpar2, lower = ~ 1) )
summary( model.step, cor = F )

##
#### Compute the ROC curves for the training and validation datasets
##
GetROCv2( model.step$fitted[-index], model.step$y[-index], model.step$fitted[index], model.step$y[index] )

##
#### Evaluate Sensitivity and Specificity
##
cbind( EvalSensSpec(GetROC( model.step$fitted[-index], model.step$y[-index] ), c(1, 1)),
EvalSensSpec(GetROC( model.step$fitted[-index], model.step$y[-index] ), c(1.5, 1)),
EvalSensSpec(GetROC( model.step$fitted[-index], model.step$y[-index] ), c(2, 1)) )
GetSensSpec( 0.95, model.step$fitted[index], model.step$y[index] )
GetSensSpec( 0.93, model.step$fitted[index], model.step$y[index] )
GetSensSpec( 0.90, model.step$fitted[index], model.step$y[index] )
```