Biostats VHM 802 Course
Winter 2015, Atlantic Veterinary College, PEI
Henrik Stryhn

Brief solution to Additional exercise 4.4

The data consist of recordings related to 189 women giving birth; our interest here is in predicting low birth weight (< 2800 *g*) from the other variables. The birth weight itself is not contained in the file but it is available (Hosmer & Lemeshow, *Applied Logistic Regression*).

y_i = low birthweight (0,1 $\sim$ no,yes),
x_{1i} = age of mother (years),
x_{2i} = weight of mother (pounds),
x_{3i} = smoking during pregnancy (0,1 $\sim$ no,yes),
x_{4i} = no. of premature labours,
x_{5i} = hypertension (0,1 $\sim$ no,yes),
x_{6i} = uterine irritability (0,1 $\sim$ no,yes),
x_{7i} = no. of visits (to physician),
$\texttt{race}_i$ = mother's race (1,2,3 $\sim$ white, black, other).

for the ith birth/mother in the dataset, $i = 1, \ldots, 189$.

Initially we consider the causal relations among the variables. The purpose of the analysis, although not stated explicitly, is probably to obtain a good prediction model for low birthweight from variables recorded prior to birth. Three variables (age, weight, race) are determined prior to gestation, and any relations between them are probably of minor interest. The five other predictors refer to events occurring during gestation, and while there may very well be relations between them it is not clear if some are definitely intermediate for others. The three variables determined prior to gestation may act as confounders for all these five variables.

The basic statistical model is a logistic regression for the probability of low birthweight, $p_i = \mathrm{P}(y_i = 1)$, for the ith birth. The predictors age, weight, no. of premature labours and visits are quantitative, smoking, hypertension and uterine irritability are dichotomous, and race is categorical. If we as a starting point model all quantitative and dichotomous predictors by a linear relation and include only main effects (both of these assumptions will be evaluated below), the initial model can be written,

$$\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \beta_7 x_{7i} + \alpha_{\mathrm{race}(i)}.$$

The first step in the analysis is to assess the strength of association between each predictor and the binary outcome. We do this twice, both using simple (univariate) logistic regressions and using standard basic statistical procedures (two-sample t-tests and Pearson X^2-statistics for two-way tables, respectively, for continuous and binary/categorical predictors).

Predictor	Simple logistic regression Estimate	SE	P^a	Simple statistics Statistic	P
age	-0.051	.032	.097	$t = 1.77$	.078
weight	-0.014	.006	.015	$t = 2.52$	.013
race	0.84/0.64	.46/.35	.082	$X^2 = 5.00$	.082
smoking	0.704	.320	.027	$X^2 = 4.92$	.026
prem. labour	0.802	.317	.009	$z^b = 3.58$	.0007
hypertens.	1.214	.608	.045	$X^2 = 4.39$	$.052^c$
uter. irr.	0.947	.417	.023	$X^2 = 5.40$	.020
visits	-0.135	.157	.379	$z^b = 1.24$	.239

[a] P-value for likelihood-ratio test
[b] Mann-Whitney-Wilcoxon two-sample rank test (with exact P-value)
[c] P-value (two-sided) from Fisher's exact test

It is seen that all predictors, except visits, on their own have a significant or close to significant association with low birthweight. The strongest univariate association is with premature labour. The correlations between the quantitative predictors are small ($-0.14 \leq r \leq 0.22$) and do not signal any potential problems with collinearity. The next step is to assess the linearity of the quantitative predictors. Our approach involves examining the linear trend generated by categorising the predictor into groups with roughly equal percentages of the data (using the `lintrend` command in Stata) or with suitably chosen categories, examining the significance of a quadratic term and possibly examining the fit of alternative models for the predictor.

- age: Linear trends with 4 and 8 groups show a noisy pattern that could perhaps be considered linear, in absence of another obvious interpretation of the pattern. The quadratic term is clearly non-significant. In conclusion, we take the effect of age as linear.

- weight: The plot of linear trends based on 4 and 8 groups show somewhat different pictures; the former clearly indicates a drop in risk from the first quartile (around 110 pounds) relative to the other three quartiles, but this is less clear with 8 groups. A smoothed (lowess) curve shows a roughly linear decline in logit(p) up to about 130 pounds, then a constant risk until maybe 190 pounds and a further drop (but there are only 7 observations > 190, so this part should not be interpreted too strongly). The quadratic term is nonsignificant. Replacing weight by an indicator of low weight (less than 110 pounds) gives a better fit (and stronger significance) than a linear relation for weight. It is not clear which is preferable, so we'll examine both further with multivariable models.

- prem. labour: Only 6 women had more than 1 premature labour, so it is obvious to turn this variable into a dichotomous variable (0, ≥ 1).

- visits: Only 12 women had more than 2 visits, but as the proportion of low birthweight is considerably larger among these than those with 1-2 visits, we construct a trichotomous version of this predictor (0, 1-2, ≥ 3). In a univariate analysis, it gets somewhat close to significance ($P = 0.14$).

Our first multivariable model contains all predictors, with prem. labour in its dichotomous form (say x^*_{4i}), visits in its trichotomous form, and weight either as measured or in its dichotomous form (say x^*_{2i}), as discussed above. In the multivariable model, the effect of visits is clearly nonsignificant (P around 0.5 by the Wald test), and as this predictor was also identified as the least interesting apriori we decide to drop it at this point.

We now turn to the modelling of weight. A model including both candidate terms, x_2 and x^*_2, shows that the linear term has a slightly stronger fit ($P = 0.20$ compared to $P = 0.31$); note that the

situation is now reversed compared to the univariate analysis. A comparison between the estimates of the two analyses with x_2 or x_2^* separately shows substantial differences in the estimates for age and hypertension; both effects are stronger with x_2^* than x_2. A notable correlation between the estimates for weight and hypertension disappears when weight is modelled as a dichotomous predictor. At this point we have had arguments in favour of both ways of modelling weight, and the choice is not obvious so probably one should do the entire analysis in two versions. We will however concentrate on the results when weight is modelled as a dichotomous predictor (following Hosmer & Lemeshow, *Applied Logistic Regression*).

The model has 128 covariate patterns which gives too little replication to make residuals and Pearson goodness-of-fit statistics useful. However, the Hosmer-Lemeshow statistic with 10 groups is clearly nonsignificant ($X^2 = 4.81, \mathrm{df} = 8,\ P = 0.78$). The large number of covariate patterns is mainly caused by having age in the model. For model-checking purposes it is desirable to categorize age; alternatively, with an estimate of $-0.046\,(.038)$ and $P = 0.21$ one might drop age from the model. As age is a potential confounder for the five gestation period predictors, one should check for confounding first. The largest effect change is for x_2^*, from 1.17 to 1.28, and thus not an indication of strong confounding. We therefore drop age from the model.

The number of covariate patterns drop to 44, and the Pearson goodness-fit-statistic is nonsignificant ($X^2 = 43.7,\ \mathrm{df} = 36,\ P = 0.18$). The strongest negative residual is for covariate pattern # 13 with 4 negatives among white, light, smoking mothers with uterine irritability; this pattern also has the strongest delta-beta statistic. The strongest positive residual is for a single observation, and therefore of less interest (and of little influence). Note that leverage is less useful here with all predictors categorical. We rerun the model without covariate pattern 13. Major changes are seen in most estimates, and the coefficient for uterine irritability goes from nonsignificant ($P = 0.12$) to significant ($P = 0.025$). This difference is clearly alarming, and noteworthy in any reporting of the analysis. We will however continue with the full dataset. By the substantial impact of covariate pattern 13 on the effect of uterine irritability, we keep this predictor in the model despite its quite high P-value. No other predictors can be removed at the 0.05 significance level (based on Wald tests, but even if the likelihood-ratio test would shift a P-value slightly above 0.05 we would not take this as "evidence" to drop that variable).

We continue the analysis by examining interactions. First, we reintroduce age into the model to assess any interactions (because these would be biologically interesting). However, no interactions with age show up as significant. Only a single interaction shows up as significant among the other variables: weight (dichotomous) × uterine irritability ($P = 0.016$). This effect improves the modelling of covariate pattern 13 considerably, by lowering the predicted probability to 0.30 (from 0.52) and thereby reducing the residuals and influence diagnostics. Note that the interaction becomes non-significant ($P = 0.14$) without covariate pattern 13. The effect of race is now slightly above significant ($P = 0.068$) but we keep it in the model anyway. The Stata listings below (next page) gives the set of parameter estimates and odds-ratios for the final model.

The coefficient of the interaction parameter for weight and uterine irritability has opposite sign of the two main effect coefficients. This means that the effect of both factors changes with the presence of the other factor. Specifically, for women with no uterine irritability high weight is protective with an odds-ratio of about 1/4. Also, for women with low weight uterine irritability is protective with an odds-ratio of about 1/2 (although not significantly different from 1). In the presence of uterine irritability, high weight becomes a risk (the odds-ratio is about 2.5 ($\exp(2.3203 - 1.4039)$), but not significantly different from 1), and for high weight women uterine irritability is associated with a clearly significant odds-ratio of almost 5 ($\exp(2.3203 - 0.7697)$). The interpretation of the main effects of the other predictors is left as an exercise for the reader.

```
. logit bw prelab01 hyp smoker i.race i.wt01##i.uter

Logistic regression                               Number of obs   =        189
                                                  LR chi2(8)      =      41.50
                                                  Prob > chi2     =     0.0000
Log likelihood = -96.587694                       Pseudo R2       =     0.1768
------------------------------------------------------------------------------
          bw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    prelab01 |   1.108304   .4555631     2.43   0.015      .215417    2.001191
         hyp |   1.523986   .6624491     2.30   0.021     .2256101    2.822363
      smoker |   .8925826   .4111835     2.17   0.030     .0866777    1.698487
             |
        race |
          2  |   1.097761   .5204055     2.11   0.035     .0777851    2.117737
          3  |   .7697201   .4501838     1.71   0.087    -.1126239    1.652064
             |
     1.wt01  |  -1.403891    .460945    -3.05   0.002    -2.307327   -.5004552
     1.uter  |  -.7308647   .7637157    -0.96   0.339     -2.22772    .7659905
             |
   wt01#uter |
        1 1  |   2.320326   .9633119     2.41   0.016     .4322696    4.208383
             |
       _cons |  -1.027687   .5488153    -1.87   0.061    -2.103345    .0479713
------------------------------------------------------------------------------

. logit bw prelab01 hyp smoker i.race i.wt01##i.uter, or

Logistic regression                               Number of obs   =        189
                                                  LR chi2(8)      =      41.50
                                                  Prob > chi2     =     0.0000
Log likelihood = -96.587694                       Pseudo R2       =     0.1768
------------------------------------------------------------------------------
          bw | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    prelab01 |   3.029217       1.38     2.43   0.015     1.240379    7.397865
         hyp |   4.590489   3.040965     2.30   0.021     1.253087    16.81654
      smoker |   2.441427   1.003874     2.17   0.030     1.090545    5.465674
             |
        race |
          2  |   2.997448   1.559888     2.11   0.035      1.08089    8.312308
          3  |   2.159162   .9720196     1.71   0.087     .8934867    5.217739
             |
     1.wt01  |   .2456393   .1132262    -3.05   0.002      .099527    .6062546
     1.uter  |   .4814925   .3677233    -0.96   0.339     .1077739    2.151124
             |
   wt01#uter |
        1 1  |   10.17899   9.805547     2.41   0.016     1.540751    67.24771
             |
       _cons |   .3578337   .1963846    -1.87   0.061     .1220475    1.049141
------------------------------------------------------------------------------
```