STAT 659 — Solution to Exam 1 Spring 2015

1. (a) The marginal OR is 2.1488 while the partial ORs are 2.7021 and 1.8763. Since the partial ORs indicate a similar association to the marginal ORs, Simpson's paradox is not present.

(b)

	(i) the test of equal odds ratios	(ii) the test of partial association
Value of test statistic	0.8084	30.6184
$P-$value	0.3684	< 0.0001
Conclusion	There is insufficient evidence to indicate that the ORs relating `smoke` to `depresss` differ for males and females.	There is strong evidence of partial association between `smoke` and `depress`, controlling for `gender`.

(c) A 95% confidence interval for the common OR between `smoke` and `depress` is $(1.7791, 3.4639)$ for the Mantel-Haenszel interval or $(1.7870, 3.4717)$ for the logit interval. Since we cannot conclude that the ORs differ for males and females based on the BD Test, it is appropriate to report a single interval for OR.

2. (a) The estimated proportion of breast-cancer patients having their first child at age 30 or later is $\hat{\pi} = (463 + 220)/3220 = 0.212$. A 95% confidence interval for π is given by

$$0.212 \pm 1.96\sqrt{\frac{(0.212)(1-0.212)}{3220}} = 0.212 \pm 0.014 \quad \text{or} \quad (0.198, 0.226).$$

(b) The estimated odds ratio for having breast cancer for women in the oldest age group ($\geq$35) relative to those in the youngest age group ($<$20) is given by

$$\hat{\theta} = \frac{220 \times 1422}{626 \times 320} = 1.562.$$

The 95% confidence interval for the log odds ratio is given by

$$\log(1.562) \pm 1.96\sqrt{1/220 + 1/320 + 1/626 + 1/1422} = 0.445 \pm .196.$$

The 95% confidence interval for the OR is $(e^{0.250}, e^{0.641}) = (1.284, 1.899)$. Thus, a woman in the oldest age group has odds of having breast cancer somewhere from 1.284 to 1.899 times the odds of a woman in the youngest age group.

(c) We can reject $H_0 : \pi_{ij} = \pi_i \times \pi_j$, all i, j since $X^2 = 24.9562$ and $G^2 = 24.6488$, both with a $p-$value$< .0001$. Thus, there is strong indication of an association between age group and the presence of breast cancer.

(d) The standardized residuals for the < 20 and ≥ 35 cells are larger than 3, with that for $30-34$ great than 2. These cells indicate that fewer young women than expected had breast cancer whereas more than expected women in the two oldest groups had breast cancer.

3. To test $H_0 : \pi_W = 3/4,\ \pi_Y = 3/16,\ \pi_G = 1/16$, we compute the expected counts under H_0 and compute Pearson's chi-squared statistic:

Color	White	Yellow	Green
Number of Progeny	143	46	19
Probability under H_0	0.75	0.1875	0.0625
Expected frequency	156	39	13

Since

$$X^2 = \frac{(143-156)^2}{156} + \frac{(46-39)^2}{39} + \frac{(19-13)^2}{13} = 5.109 < 5.99 = \chi^2_{2,.05},$$

we do not reject H_0 and conclude that there is insufficient evidence to indicate that the proportions differ from those predicted by the genetic model.

4. (a) We reject $H_0 : \beta_{\texttt{risk}} = \beta_{\texttt{ses0}} = \beta_{\texttt{ses1}} = 0$ since $G^2 = 389.35 - 378.66 = 10.69 > 7.81 = \chi^2_{3,.05}$. Thus, we cannot simultaneously eliminate `risk` and `ses` from the complete model.

   (b) Keeping the other variables constant, the difference in the linear predictors for a person in a crowded setting and exposed to passive smoke relative to another person not in a crowded setting and not exposed to smoke is

$$\log(\hat{\mu}_1) - \log(\hat{\mu}_2) = \hat{\beta}_{\texttt{passive}} + \hat{\beta}_{\texttt{crowding}} = 0.3181 + 0.5062 = 0.8243.$$

   Thus,

$$\mu_1/\mu_2 = e^{0.8243} = 2.28.$$

   The estimated mean number of lower respiratory infections for the child in a crowded setting and exposed to passive smoke is 2.28 times that of another child not in a crowded setting and not exposed to smoke.

   (c) Model 3 has the lowest AIC_C. To compare it to model 2, we compute $G^2 = 380.9 - 380 = 0.9 < \chi^2_{1,.05}$. There is no need to include `race` in Model 3. To compare it to model 2, we compute $G^2 = 386.3 - 380.9 = 5.4 > \chi^2_{1,.05}$. Thus, we need to include `passive` in the model. Thus, Model 3 should be used.

   (d) The Poisson model A has a deviance/df ratio of 1.3769 indicating a possible lack of fit. The negative binomial model C has a deviance/df ratio of 0.9189, indicating a reasonable fit. Also, the 95% confidence interval for the dispersion parameter is (0.4432,1.3848) which does not include zero. Thus, there is overdispersion relative to the Poisson model.