4.2a **For Table 3.8 with scores (0, 0.5, 1.5, 4.0, 7.0) for alcohol consumption, ML fitting of the linear probability model for malformation has output (pictured). Interpret the model parameter estimates. Use the fit to estimate the relative risk of malformation that compares alcohol consumption levels 0 and 7.0.**

Since the linear probability model for malformation was performed here (presented below), we can do the parameter estimates interpretations like the following.

$$\hat{\pi}(alcohol) = \hat{\Pr}(mal = 1 \,|\, alcohol) = \hat{\beta}_0 + \hat{\beta}_1 * alcohol$$

(1) $\hat{\beta}_0$ (intercept): when the alcohol consumption score =0, that is, when the average number of drinks per day is 0, the probability of being malformation is about $\hat{\beta}_0 = 0.0025$.

(2) $\hat{\beta}_1$ (corresponding to alcohol score): when alcohol consumption score is increased by 1 unit (that is, by 1), the probability of being malformation would be increased by about $\hat{\beta}_1 = 0.001087$.

The estimated relative risk of malformation that compares alcohol consumption levels 0 to 7.0 is

$$r_{0vs7} = \frac{\hat{\pi}_0}{\hat{\pi}_{7.0}} = \frac{0.0025}{0.0025 + 0.001087 * 7.0} = \frac{0.0025}{0.010109} \approx 0.247$$

4.7

a. Log($\hat{\mu}$) = 1.6094 + 0.5878 x.
Since β = log(μ_B / μ_A), exp($\hat{\beta}$) = $\hat{\mu}_b$ / $\hat{\mu}_a$ = 1.80.
i.e., the mean is predicted to be 80 percent higher for treatment B (In fact, this estimate is simply the ratio of sample means.)

b. Wald test gives $z = .588/.176 = 3.33$, $z^2 = 11.1$, ($df = 1$), $P < 0.001$. Likelihood-ratio statistic equals $27.86 - 16.27 = 11.6$ with $df = 1$, $P < 0.001$. There is strong evidence against H_0 and a higher defect rate for treatment B.

c. Exponentiate 95% CI for β of $0.588 \pm 0.196(0.176)$ to get Wald CI of $\exp(0.242, 0.934) = (1.27, 2.54)$.

d. Normal approximation to binomial yields $z = (50 - 90)/\sqrt{140(.5)(.5)} = -6.76$ and very strong evidence against H_o

4.8

a. No evidence of overdispersion, since sample variance does not exceed sample mean for each treatment.

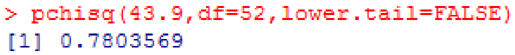
b. Fit is $log(\hat{\mu}) = 1.609 + .588x$.

c. Fit is $log(\hat{\mu}) = 1.946$. for both models, but $SE = 0.0845$ for Poisson and 0.099 for negative binomial(NB). (For NB, estimated dispersion parameter $= 0.055 > 0$, so fits are different.) Wald confidence intervals are $(5.93, 8.26)$ for Poisson and $(5.76, 8.51)$ for NB, the slightly wider one for the NB model reflecting the slight overdispersion for the overall sample.

4.12 Table 4.10 describes survival for 539 males diagnosed with lung cancer. The prognostic factors are histology (H) and state (S) of disease. The assumption of a constant rate over time is often not sensible, and this study divided the time scale (T) into two-month intervals and let the rate vary by the time interval. Let μ_{ijk} denote the expected number of deaths and t_{ijk} the total time at risk for histology *i* and state of disease *j*, in the follow-up time interval *k*.

a. The main effects model has deviance 43.9. Explain why df=52. Does the model seems to fit adequately?

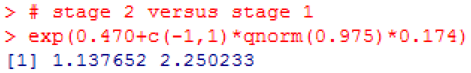
In this main effects model, N=7×9=63 and we have totally 11 parameters in the model (1 for intercept, 2 for histology H, 2 for stage S, and 6 for time scale T). Therefore, the degrees of freedom is df=N-p=63-11=52. The deviance is 43.9 with df=52, indicating that the model fits adequately (p-value=0.78, R code is shown below to calculate the p-value).



```
> pchisq(43.9,df=52,lower.tail=FALSE)
[1] 0.7803569
```

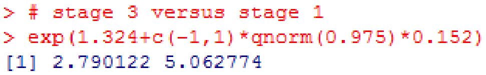
b. For this model, interpret the estimated effects of S, $\hat{\beta}_2^S - \hat{\beta}_1^S = 0.470\ (SE = 0.174)$, $\hat{\beta}_3^S - \hat{\beta}_1^S = 1.324\ (SE = 0.152)$.

Conditioning on the histology and time scale, the estimated death rate on Stage 2 of lung cancer is about $\exp(\hat{\beta}_2^S - \hat{\beta}_1^S) = \exp(0.470) \approx 1.60$ (95% CI: 1.14 – 2.25) times the estimated death rate on Stage 1 of lung cancer (R code to calculate 95% CI is shown below).



```
> # stage 2 versus stage 1
> exp(0.470+c(-1,1)*qnorm(0.975)*0.174)
[1] 1.137652 2.250233
```

Conditioning on the histology and time scale, the estimated death rate on Stage 3 of lung cancer is about $\exp(\hat{\beta}_3^S - \hat{\beta}_1^S) = \exp(1.324) \approx 3.76$ (95% CI: 2.79 – 5.06) times the estimated death rate on Stage 1 of lung cancer (R code to calculate 95% CI is shown below).



```
> # stage 3 versus stage 1
> exp(1.324+c(-1,1)*qnorm(0.975)*0.152)
[1] 2.790122 5.062774
```

c. The model that adds an S x H interaction term has deviance 41.5 with df=48. Test whether a significantly improved fit results by allowing this interaction.

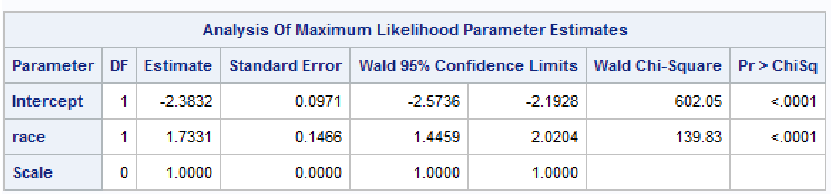
To test for the significance of the interaction term S× H, we can do the likelihood-ratio model comparison as follows.

$G^2(M_0 \mid M_1) = 43.9 - 41.5 = 2.4$ with df =4 (the difference in number of parameters between M_1 and M_0 is 4; p-value≈0.663), indicating not rejecting the null hypothesis. In other words, no significant improvement is made by the interaction term.

p.s. M_0 refers to main effects model and M_1 refers to the model with interaction term S× H.

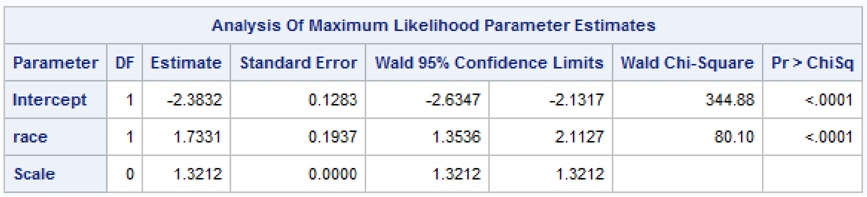
4.14 Refer to Table 14.6. Fit a loglinear model with an indicator variable for race, a) assuming a Poisson distribution, and b) allowed overdispersion with a quasi-likelihood approach. Compare results.

(a) Refer to Table 14.6 (textbook page 554). Fit a loglinear model with an indicator variable for race assuming a Poisson distribution. Part of SAS output table is shown below. Thus, the fitted model is $\log(\hat{\mu}) = -2.3832 + 1.7331 * race$ with $SE(\hat{\beta}_0) = 0.0971$ and $SE(\hat{\beta}_1) = 0.1466$



Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.3832	0.0971	-2.5736	-2.1928	602.05	<.0001
race	1	1.7331	0.1466	1.4459	2.0204	139.83	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

(b) Using the same dataset, fit the same loglinear model but allowing overdispersion with quasi-likelihood approach. Part of SAS output table is shown below. Thus, the fitted model result is $\log(\hat{\mu}) = -2.3832 + 1.7331 * race$ with $SE(\hat{\beta}_0) = 0.1283$ and $SE(\hat{\beta}_1) = 0.1937$



Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.3832	0.1283	-2.6347	-2.1317	344.88	<.0001
race	1	1.7331	0.1937	1.3536	2.1127	80.10	<.0001
Scale	0	1.3212	0.0000	1.3212	1.3212		

Compare the results: Coefficient estimates of these two models are the same, but their standard errors are obviously different. More specifically, the standard errors in quasi-likelihood approach are larger than those in the ordinary Poisson model because quasi-likelihood approach multiplies the ordinary standard error estimates by another factor which allows greater standard errors and then wider confidence intervals.

4.16 For binary data, define a GLM using the log link. Show that effects refer to the relative risk. Why do you think this link is not often used? [*Hint:* What happens if the linear predictor takes a positive value?]

For example, suppose we have the following GLM with log link for binary data.

$$\log(\pi) = \log[\Pr(Y=1)] = \beta_0 + \beta_1 X$$

When X=1, $\log(\pi_1) = \beta_0 + \beta_1$. When X=0, $\log(\pi_0) = \beta_0$

Thus, $\beta_1 = \log(\pi_1) - \log(\pi_0) = \log(\frac{\pi_1}{\pi_0}) = \log RR$. In other words, $RR = \exp(\beta_1)$.

Therefore, for binary data, the GLM with log link like the one shown above would provide an interpretation of relative risk (RR) using effects (e.g. the effect of covariate X here).

The reason why log link for binary data is not often used is because of the consideration of the range that probability π could take. Since π is between 0 and 1, it's hard to control exp(linear predictor) to be between 0 and 1.

4.18 Let Y_i be a bin(n_i, π_i) variate for group *i*, *i*=1, …, N, with {Y_i} independent. For the model that π_1= … =π_N, denote that common value by π. For observations {y$_i$}, show that $\hat{\pi} = (\sum_i y_i)/(\sum_i n_i)$. When all $n_i = 1$, for testing this model's fit in the N x 2 table, show that X^2=*N*. Thus goodness-of-fit statistics can be completely uninformative for ungrouped data.

Based on the given context, we can write the likelihood and do the calculations as follows.

$$L = \prod_{i=1}^{N} f(y_i \mid \pi) = \prod_{i=1}^{N} \binom{n_i}{y_i} \pi^{y_i} (1-\pi)^{n_i - y_i}$$

$$\Downarrow$$

$$l = \log L = \sum_{i=1}^{N} \log(f(y_i \mid \pi)) = \sum_{i=1}^{N} \log[\binom{n_i}{y_i} \; \pi^{y_i} (1-\pi)^{n_i - y_i}] = \sum_{i=1}^{N} [\log\binom{n_i}{y_i} + y_i \log(\pi) + (n_i - y_i)\log(1-\pi)]$$

$$\Downarrow$$

$$\frac{\partial l}{\partial \pi} = \frac{\sum_{i=1}^{N} y_i}{\pi} - \frac{\sum_{i=1}^{N} (n_i - y_i)}{1-\pi} \qquad let \; \frac{\partial l}{\partial \pi} = 0$$

$$\Downarrow$$

$$(1-\pi)\sum_{i=1}^{N} y_i - \pi \sum_{i=1}^{N} (n_i - y_i) = \sum_{i=1}^{N} y_i - \pi \sum_{i=1}^{N} n_i = 0 \quad \Rightarrow \quad \pi = \left(\sum_{i=1}^{N} y_i\right) \Big/ \left(\sum_{i=1}^{N} n_i\right)$$

When all n*i* = 1, we could have the following:

$$X^2 = \sum_i e_i^2 = \sum_i \left(\frac{y_i - \hat{\pi}}{\sqrt{\hat{\pi}(1-\hat{\pi})}} \right)^2 = \sum_i \frac{\left(y_i - \hat{\pi}\right)^2}{\hat{\pi}(1-\hat{\pi})} = \sum_i \frac{\left(y_i - \frac{\sum_i y_i}{N} \right)^2}{\frac{\sum_i y_i}{N}(1 - \frac{\sum_i y_i}{N})} = \frac{\frac{1}{N^2} \sum_i \left(N y_i - \sum_i y_i \right)^2}{\frac{1}{N^2} \left(\sum_i y_i \right) \left(N - \sum_i y_i \right)}$$

$$= \frac{\sum_i \left(N^2 y_i^2 + \left(\sum_i y_i \right)^2 - 2N y_i \left(\sum_i y_i \right) \right)}{\left(\sum_i y_i \right) \left(N - \sum_i y_i \right)} = \frac{N^2 \sum_i y_i^2 + N \left(\sum_i y_i \right)^2 - 2N \left(\sum_i y_i \right)^2}{\left(\sum_i y_i \right) \left(N - \sum_i y_i \right)}$$

$$= \frac{N^2 \sum_i y_i^2 - N \left(\sum_i y_i \right)^2}{\left(\sum_i y_i \right) \left(N - \sum_i y_i \right)} = \frac{N \left(N \sum_i y_i^2 - \left(\sum_i y_i \right)^2 \right)}{\left(\sum_i y_i \right) \left(N - \sum_i y_i \right)} = \frac{N \left(N \sum_i y_i - \left(\sum_i y_i \right)^2 \right)}{\left(\sum_i y_i \right) \left(N - \sum_i y_i \right)}$$

$$= \frac{N \left(\left(\sum_i y_i \right) \left(N - \left(\sum_i y_i \right) \right) \right)}{\left(\sum_i y_i \right) \left(N - \sum_i y_i \right)} = N$$

The key to the last but second step in calculation is $y_i = y_i^2$. Since all $n_i = 1$, y_i can be either 0 or 1. That is why $y_i = y_i^2$.

4.27 $f(y|k,\mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}(\frac{k}{\mu+k})^k(1-\frac{k}{\mu+k})^y \Rightarrow$ when k is known let $a(\mu) = (\frac{k}{\mu+k})^k, b(y) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}$, and $\exp[yQ(\mu)] = (1-\frac{k}{\mu+k})^y = \exp[y\log(1-\frac{k}{\mu+k})] = \exp[y\log(\frac{\mu}{\mu+k})]$ with the natural parameter $Q(\mu) = \log(\frac{\mu}{\mu+k})$. Thus, this distribution has exponential family form.

Problem #9

From what we knew, $Y|\lambda \sim Poi(\lambda)$, and $\lambda \sim Gamma(\mu, k)$.

$$\begin{aligned}
f(y) &= \int_0^\infty f(y,\lambda)d\lambda \\
&= \int_0^\infty P(Y=y|\lambda)\cdot f(\lambda)d\lambda \\
&= \int_0^\infty \frac{\lambda^y \cdot e^{-\lambda}}{y!} \cdot \frac{(\frac{k}{\mu})^k}{\Gamma(k)}\lambda^{k-1}\cdot e^{-\frac{\lambda k}{\mu}}d\lambda \\
&= \frac{(\frac{k}{\mu})^k}{\Gamma(k)\cdot\Gamma(y+1)}\cdot\int_0^\infty \lambda^{y+k-1}\cdot e^{-\lambda(1+\frac{k}{\mu})}d\lambda \\
&= \frac{\Gamma(y+k)\cdot(\frac{k}{\mu})^k}{\Gamma(k)\cdot\Gamma(y+1)\cdot(1+\frac{k}{\mu})^{k+y}} \\
&= \frac{\Gamma(y+k)}{\Gamma(k)\cdot\Gamma(y+1)}\cdot(\frac{k}{\mu+k})^k\cdot(1-\frac{k}{\mu+k})^y
\end{aligned}$$

$\therefore$ Y marginally follows Negative Binomial distribution, with $E(Y)=\mu=E(\lambda)$, $var(Y)=\mu+\frac{\mu^2}{k}$.
Poisson model assumes the mean λ is fixed, and $var(Y)=E(Y)=\lambda$.
While when λ is not fixed, e.g. in this case, λ follows a Gamma distribution. Y marginally has Negative Binomial distribution. If we still assumes Y marginally follows Poisson distribution, as in Poisson model, overdispersion is included.
Thus Negative Binomial model is a way to handle overdispersion for Poisson.

5.2 For a study using logistic regression to determine characteristics associated with remission in cancer patients, Table 5.11 shows the most important explanatory variable, a labeling index (LI) that measures proliferative activity of cells after a patient receives an injection of tritiated thymidine. It represents the percentage of cells that are "labeled." The response measured whether the patient achieved remission. Software reports Table 5.12 for a logistic regression model using L1 to estimate π = *P*(remission).

According to software output, the logistic regression model is: $\hat{\pi} = \dfrac{e^{-3.7771+0.1449li}}{1+e^{-3.7771+0.1449li}}$

a. Show how software obtained $\hat{\pi} = 0.068$ when LI=8

Using the equation in (1), when li = 8, $\hat{\pi} = \dfrac{e^{-3.7771+0.1449\times 8}}{1+e^{-3.7771+0.1449\times 8}} = \dfrac{0.0730}{1.0730} = 0.0680$

b. Show that $\hat{\pi} = 0.50$ when LI=26.0

if $\hat{\pi}$ = 0.50, LI can be solved from the equation in (1):

$$li = \frac{\ln\dfrac{\pi}{1-\pi} + 3.7771}{0.1449} = \frac{3.7771}{0.1449} = 26.07$$

c. Show that the rate of change in $\hat{\pi}$ is 0.009 when LI=8 and 0.036 when LI=26

From the equation in (1), $\hat{\pi}(8) = 0.0680$, then the rate of change is:

$$\frac{\partial\hat{\pi}}{\partial x} = 0.1449\hat{\pi}(1-\hat{\pi}) = 0.1449\times 0.068\times(1-0.068) = 0.009$$

From the equation in (1), $\hat{\pi}(26) = \dfrac{e^{-3.7771+0.1449\times 26}}{1+e^{-3.7771+0.1449\times 26}} = \dfrac{0.9903}{1.9903} = 0.4976$, then the rate of change is: $\dfrac{\partial\hat{\pi}}{\partial x} = 0.1449\hat{\pi}(1-\hat{\pi}) = 0.1449\times 0.4976\times(1-0.4976) = 0.036$

d. The lower quartile and upper quartile for LI are 14 and 28. Show that $\hat{\pi}$ increases by 0.42, from 0.15 to 0.57, between those values

Using the equation in (1), $\hat{\pi}(14) = \dfrac{0.1740}{1.1740} = 0.1482, \hat{\pi}(28) = \dfrac{1.3233}{2.3233} = 0.5696$, the increase of $\hat{\pi}$ is 0.5696-0.1482=0.42.

e. For a unit increase in LI, show that the estimated odds of remission multiply by 1.16

$$odds = e^{\beta} = e^{0.1449} = 1.16$$

f. Explain how to obtain the CI reported for the OR. Interpret.

The 95% CI for $\hat{\beta}$ is:

$$\left(0.1449-1.96\times 0.0593, 0.1449-1.96\times 0.0593\right)=\left(0.0287, 0.2611\right)$$

Then the 95% CI for odds ratio is: $\left(e^{0.0287}, e^{0.2611}\right)=\left(1.0291, 1.2984\right)$. Since the 95% CI of odds ratio does not contain 0, as *li* increases, the remission rate is significantly increases.

g. Construct a Wald test for the effect. Interpret

The test statistic of Wald test for *li* is: $z=\left(\dfrac{\hat{\beta}}{se(\hat{\beta})}\right)^2=\left(\dfrac{0.1449}{0.0593}\right)^2=5.9707$ and p-value is 0.0145<0.05, which indicates that the effect of *li* on the remission rate is significant.

h. Conduct a LRT for the effect showing how to construct the test statistic using the -2 log *L* values reported

The test statistic of likelihood-ratio = 34.372 – 26.073 = 8.289 and the p-value is 0.0039<0.05. Therefore the effect of *li* on the remission rate is significant.

i. Show how the software obtained the CI interval for π reported at LI=8.

Let $g\left(\hat{\beta}\right)=\hat{\pi}\left(li\right)=\dfrac{e^{-3.7771+\hat{\beta}\cdot li}}{1+e^{-3.7771+\hat{\beta}\cdot li}}$, then

$$Dg\left(\hat{\beta}\right)=\frac{\partial g\left(\hat{\beta}\right)}{\partial\hat{\beta}}=\frac{1}{\left(1+e^{-3.7771+\hat{\beta}\cdot li}\right)^2}\times e^{-3.7771+\hat{\beta}\cdot li}\times li$$

when li = 8, $Dg\left(\hat{\beta}\right)=\dfrac{1}{\left(1+e^{-3.7771+0.1449\times 8}\right)^2}\times e^{-3.7771+0.1449\times 8}\times 8=0.5070$

then

$$\begin{aligned} se\left\{g\left(\hat{\beta}\right)\right\} &= \sqrt{Dg\left(\hat{\beta}\right)\operatorname{cov}\left(\hat{\beta}\right)Dg\left(\hat{\beta}\right)} \\ &= \sqrt{0.5070\times 0.003521\times 0.5070} \\ &= 0.0301 \end{aligned}$$

Then the 95% CI for $\left(0.0680-1.96\times 0.0301, 0.0680+1.96\times 0.0301\right)=\left(0.0301, 0.1270\right)$.

5.14 Refer to the prediction equation $\mathbf{logit}(\hat{\pi}) = -\mathbf{10.071} - \mathbf{0.509}c + \mathbf{0.458}x$ for model 5.14 using quantitative color and width. The means and sds are $\bar{c} = \mathbf{2.44}$ and *s*=0.80 for color $\bar{x} = \mathbf{26.30}$ and *s*=2.11 for width. The standardized predictors [e.g., *x* = (width – 26.30)/2.11], explain why the estimated coefficients of *c* and *x* equal -0.41 and 0.97. Interpret these by comparing the partial effects of a sd increase in each predictor on the odds. Describe the color effect by estimating the change in $\hat{\pi}$ between the first and last color categories at the sample mean width.

Let β be the coefficient of the original covariate x and β' the coefficient of standardized covariate x_s, the partial effect before and after standardization should be the same:

$$\frac{\partial(\beta x)}{\partial x} = \beta \equiv \frac{\partial(\beta' x_s)}{\partial x} = \frac{\partial\left(\beta' \frac{x-\bar{x}}{s_x}\right)}{\partial x} = \frac{\beta'}{s_x}$$

$$\Rightarrow \beta' = \beta \cdot s_x$$

Therefore, after standardization, coefficient for c is $-0.509\times 0.8 = 0.41$ and coefficient for x is $0.458\times 2.11 = 0.97$.

$$\pi(c=1, x=26.30) = \frac{e^{-10.071-0.509\times 1+0.458\times 26.30}}{1+e^{-10.071-0.509\times 1+0.458\times 26.30}} = 0.8124$$

$$\pi(c=4, x=26.30) = \frac{e^{-10.071-0.509\times 4+0.458\times 26.30}}{1+e^{-10.071-0.509\times 4+0.458\times 26.30}} = 0.4846$$

As color changes from category 1 to category 4, the satellite rate decreases by 0.8124-0.4846=0.3278.

5.18 In a study designed to evaluate whether an educational program makes sexually active adolescents more likely to obtain condoms, adolescents were randomly assigned to two experimental groups. The educational group was provided to one group but not the other. Table 5.17 summarizes results of a logistic regression model for factors observed to influence teenagers to obtain condoms.

a. Find the parameter estimates for the fitted model using (1, 0) indicator variables for the first three predictors. Based on the corresponding CI for the log OR, determind the standard error for the group effect.

$$\because \textit{odds ratio} = e^{\beta} \textit{ and its } 95\% \; CI = \left(e^{\beta \mp 1.96 se\{\beta\}}\right)$$

$$\therefore \beta = \ln\left(\textit{odds ratio}\right),\; se\{\beta\} = \frac{\ln\left(95\% \; CI \textit{ upper limit}\right) - \beta}{1.96}$$

According to the output table, we can compute coefficient of each variable from odds ratio:

$$\beta_{grp} = \ln\left(4.04\right) = 1.40,\; se\{\beta_{grp}\} = (\log(13.9)\text{-}1.4)/1.96 = 0.63$$

$$\beta_{gen} = \ln\left(1.38\right) = 0.32,\; se\{\beta_{gen}\} = (\log(13.9)\text{-}1.4)/1.96 = 1.14$$

$$\beta_{ses} = \ln\left(5.82\right) = 1.76,\; se\{\beta_{ses}\} = (\log(18.28)\text{-}1.76)/1.96 = 0.58$$

$$\beta_{lnp} = \ln\left(3.22\right) = 1.17,\; se\{\beta_{lnp}\} = (\log(11.31)\text{-}1.17)/1.96 = 0.64$$

b. Explain why either the estimate of 1.38 for the OR for gender or the corresponding CI is incorrect. Show that if the reported interval is correct, 1.38 is actually the log OR and estimated OR equals 3.98.

According to the result of problem ***a***, $se\{\beta_{gen}\} = 1.14$, from this result, the lower limit of odds ratio should be $e^{0.32\text{-}1.96*1.14} = 0.1474 \neq \textit{lower limit read from table} = 1.23$. Therefore, the odds ratio is wrong. If the reported interval is right, then the odds ratio should be:

$$\because 95\% \; CI = \left(OR \cdot e^{-1.96 \cdot se\{\beta\}}, OR \cdot e^{1.96 \cdot se\{\beta\}}\right)$$

$$\therefore OR = \sqrt{\textit{upper limit} \times \textit{lower limit}} = \sqrt{1.23 \times 12.88} = 3.98$$

5.32 For an *I* x 2 contingency table, consider logistic model (5.4)

a. Given $\{\boldsymbol{\pi_i} > 0\}$, show how to find $\{\boldsymbol{\beta_i}\}$ satisfying $\boldsymbol{\beta_I} = 0$

Suppose the original model is $\log\frac{\pi_i}{1-\pi_i} = \alpha + \beta_i, i = 1, 2, ..., I$

Let $\tilde{\alpha} = \alpha - \beta_I, \tilde{\beta}_i = \beta_i - \beta_I$, now the parameter for category I is $\tilde{\beta}_I = \beta_I - \beta_I = 0$.

b. Prove that $\boldsymbol{\beta_1} = \boldsymbol{\beta_2} = \cdots = \boldsymbol{\beta_I}$ is the independence model. Find its likelihood equation show that $\widehat{\boldsymbol{\alpha}} = \mathbf{logit}[\frac{(\sum_i y_i)}{(\sum_i n_i)}]$

Based on the conclusion of problem *a*, we can find $\{\beta_i\}$ that satisfies $\beta_1 = \beta_2 = ... = \beta_I = 0$.

$$\left.\begin{array}{l} \Rightarrow logit(\pi_i) = \hat{\alpha} \Rightarrow \pi_1 = \pi_2 = ... = \pi_I \Rightarrow \frac{y_1}{n_1} = \frac{y_2}{n_2} = ... = \frac{y_I}{n_I} = \frac{\sum_i y_i}{\sum_i n_i} \\ \pi_i = \frac{y_i}{n_i} \end{array}\right\} \Rightarrow \hat{\alpha} = logit\left(\frac{\sum_i y_i}{\sum_i n_i}\right)$$

5.34 Construct the log-likelihood function for the model [$\mathbf{logit}[\boldsymbol{\pi}(\boldsymbol{x})] = \boldsymbol{\alpha} + \boldsymbol{\beta x}$ with independent binomial outcomes of y_0 successes in n_0 trials at *x*=0 and y_1 successes in n_1 trials at x=1. Derive the likelihood equations and show that $\widehat{\boldsymbol{\beta}}$ is the sample log OR.

$$\left.\begin{array}{l} \left.\begin{array}{l} logit[\pi(0)] = \alpha + \beta \cdot 0 = \alpha \\ \pi(0) = y_0/n_0 \end{array}\right\} \Rightarrow e^{\hat{\alpha}} = \frac{n_0 - y_0}{y_0} \\ \left.\begin{array}{l} logit[\pi(1)] = \alpha + \beta \cdot 1 = \alpha + \beta \\ \pi(1) = y_1/n_1 \end{array}\right\} \Rightarrow e^{\hat{\alpha}+\hat{\beta}} = \frac{n_1 - y_1}{y_1} \end{array}\right\} \Rightarrow \hat{\beta} = \ln\left(\frac{n_1 - y_1}{y_1} : \frac{n_0 - y_0}{y_0}\right)$$