

STAT 601

--Assignment 6

NAIQING CAI
Ncai5@wisc.edu

1. Analyze the data and write a brief report.

We can find that y (number of children applied for LEP), is from poisson distribution. Thus, we fit it in a log linear poisson regression model.

(1) Fit a log linear poisson regression model with one factor as follows:

a. $Y \sim SES$

$$\log(E(Y_{ijk})) = \log(n_{ijk}) + \beta_0 + \sum_{i=1}^5 \beta_i * SES_i$$

b. $Y \sim mental\ health$

$$\log(E(Y_{ijk})) = \log(n_{ijk}) + \beta_0 + \sum_{j=1}^3 \alpha_j * mental_j$$

(2) Fit a log linear poisson regression model with two factors as follows:

a. $Y \sim SES + mental\ health$

$$\log(E(Y_{ijk})) = \log(n_{ijk}) + \beta_0 + \sum_{i=1}^5 \beta_i * SES_i + \sum_{j=1}^3 \alpha_j * mental_j$$

b. $Y \sim SES + mental\ health + SES:mental\ health$

$$\log(E(Y_{ijk})) = \log(n_{ijk}) + \beta_0 + \sum_{i=1}^5 \beta_i * SES_i + \sum_{j=1}^3 \alpha_j * mental_j + \sum_i \sum_j \gamma * SES_i * mental_j$$

Model	Deviance	df
Null	217.4	23
One Factor Models		
SES	160.943	18
Mental	103.87	20
Two Factor Models		
SES+Mental	47.418	15
SES+Mental+SES:Mental	0	0

Table 1.1 Deviances for the poisson log-linear model

General Observations

Null model has a deviance of 217.4 on 23 degree of freedom, which does not pass the goodness of test, so we should reject the null hypothesis. The null model is not good.

Then, introducing the SES and Mental both lead to substantial deviance reduction on a small degree of freedom, so the p-value is small enough to reject the null hypothesis. In this way, we can say that SES and Mental are both significant in the model. The number of applications will be effected both by mental health status and SES. So, we add these two items into the model.

The additive model $Y \sim SES + mental\ health + SES:Mental\ health$ has a deviance of 0 on 0 degree of freedom and the associated p-value is really small. In this way, the additive model provides a good description of the data. So, we need to add the interactions into the model.

Detailed Observations

Based on the table attached in the appendix, since the coefficient for “mild mental” is 0.71465 larger than the reference term “good mental”, we find that LEP receive a higher volume of applications from children with mild mental status, compared to children with good mental status.

Also, there is some interaction between particular levels of SES and mental health status, the most significant ones at 5% level are SESE:mentalMild, SESF:mentalMild, SESE:mentalModerate, SESD:mentalWell, SESE:mentalWell and SESF:mentalWell since the p-value for their coefficients is really small so they are significant in the model. (The result is attached in the appendix) This also means that the interactions are essential part of the model.

2. In the Bradley-Terry model for ranking k competitors, parameters $\theta_1, \dots, \theta_k$ representing ‘abilities’ are introduced in such a way that the probability π_{ij} that competitor i beats j is a function of the difference in their abilities.

(a) Write out the 21×7 model matrix X for the Bradley-Terry model.

In the logit model, we have

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \theta_i - \theta_j$$

Home team won/lost record in 1987

Home Team	Away Team						
	Milwaukee	Detroit	Toronto	New York	Boston	Cleveland	Baltimore
Milwaukee	—	4-3	4-2	4-3	6-1	4-2	6-0
Detroit	3-3	—	4-2	4-3	6-0	6-1	4-3
Toronto	2-5	4-3	—	2-4	4-3	4-2	6-0
New York	3-3	5-1	2-5	—	4-3	4-2	6-1
Boston	5-1	2-5	3-3	4-2	—	5-2	6-0
Cleveland	2-5	3-3	3-4	4-3	4-2	—	2-4
Baltimore	2-5	1-5	1-6	2-4	1-6	3-4	—

Table 2.1 Home-team win/lose record

team Milwaukee: 1	team Detroit: 2	team Toronto: 3	team New York: 4
team Boston: 5	team Cleveland: 6	team Baltimore: 7	

Table 2.2 Index for each team

Pairs\competitors	$\Theta 1$	$\Theta 2$	$\Theta 3$	$\Theta 4$	$\Theta 5$	$\Theta 6$	$\Theta 7$
Milwaukee Detroit	1	-1	0	0	0	0	0
Milwaukee Toronto	1	0	-1	0	0	0	0
Milwaukee New York	1	0	0	-1	0	0	0
Milwaukee Boston	1	0	0	0	-1	0	0
Milwaukee Cleveland	1	0	0	0	0	-1	0
Milwaukee Baltimore	1	0	0	0	0	0	-1
Detroit Toronto	0	1	-1	0	0	0	0
Detroit New York	0	1	0	-1	0	0	0
Detroit Boston	0	1	0	0	-1	0	0
Detroit Cleveland	0	1	0	0	0	-1	0
Detroit Baltimore	0	1	0	0	0	0	-1
Toronto New York	0	0	1	-1	0	0	0
Toronto Boston	0	0	1	0	-1	0	0
Toronto Cleveland	0	0	1	0	0	-1	0
Toronto Baltimore	0	0	1	0	0	0	-1
New York Boston	0	0	0	1	-1	0	0
New York Cleveland	0	0	0	1	0	-1	0
New York Baltimore	0	0	0	1	0	0	-1
Boston Cleveland	0	0	0	0	1	-1	0
Boston Baltimore	0	0	0	0	1	0	-1
Cleveland Baltimore	0	0	0	0	0	1	-1

Table 2.3 21×7 model matrix X for the Bradley-Terry model

(b) Fit the Bradley Terry model to these data to obtain a ranking of the teams. Extend this model by including a home-team advantage effect (equal for all teams). Obtain the likelihood-ratio statistic. Comment briefly on the magnitude of home-field advantage.

Fit the Bradley Terry model

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \theta_i - \theta_j$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \theta_1 + \sum_{i=2}^7 \theta_i team_i$$

```
Bradley Terry model fit by glm.fit

Call: BTm(outcome = cbind(home.wins, away.wins), player1 = home.team,
         player2 = away.team, id = "team", data = baseball)

Coefficients:
              teamBoston   teamCleveland   teamDetroit   teamMilwaukee   teamNew York   teamToronto
                1.1077        0.6839        1.4364        1.5814        1.2476        1.2945

Degrees of Freedom: 42 Total (i.e. Null); 36 Residual
Null Deviance:    78.02
Residual Deviance: 44.05      AIC: 140.5
```

Table 2.4 Coefficients for Bradley Terry model

Then, we get the formula for the model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 1.1077 * Boston + 0.6839 * Cleveland + 1.4364 * Detroit \\ + 1.5814 * Milwaukee + 1.2476 * New York + 1.2945 * Toronto$$

The coefficients here are maximum likelihood estimates of Θ_i ($i=2,3,4,5,6,7$) with Θ_1 (the log-ability for Baltimore) set to zero as an identifying convention.

The coefficients represent each team's ability that the probability π_{ij} that competitor i beats j is a function of the difference in their abilities.

In the model, the reference team is Baltimore, estimated to be the weakest of these seven, with Milwaukee and Detroit the strongest.

Ranking of the teams

Based on the results, we can rank each teams with by their coefficient estimates.

1	Milwaukee
2	Detroit
3	Toronto
4	New York
5	Boston
6	Cleveland
7	Baltimore

Table 2.5 Ranking of teams (from high to low)

Model Extension

In order to extend the model, we need to add the home advantage as another explanatory variable.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \theta_1 + \sum_{i=2}^7 \theta_i team_i + \gamma * hom e$$

Coefficients:						
	Estimate	Std. Error	z value	Pr(> z)		
teamBoston	1.1438	0.3378	3.386	0.000710 ***		
teamCleveland	0.7047	0.3350	2.104	0.035417 *		
teamDetroit	1.4754	0.3446	4.282	1.85e-05 ***		
teamMilwaukee	1.6196	0.3474	4.662	3.13e-06 ***		
teamNew York	1.2813	0.3404	3.764	0.000167 ***		
teamToronto	1.3271	0.3403	3.900	9.64e-05 ***		
at.home	0.3023	0.1309	2.308	0.020981 *		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1						

Table 2.6 Extended Bradley Terry model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 1.1438 * Boston + 0.7047 * Cleveland + 1.4754 * Detroit + 1.6196 * Milwaukee \\ + 1.2813 * New York + 1.3271 * Toronto + 0.3023 * hom e$$

Obtain the likelihood-ratio statistic

```
Analysis of Deviance Table

Response: cbind(home.wins, away.wins)

Model 1: ~team
Model 2: ~team + at.home
      Resid. Df Resid. Dev Df Deviance
1          36    44.053
2          35    38.643  1    5.4106
```

Table 2.7 Anova table of likelihood ratio test

Based on the likelihood ratio test, we can find the likelihood ratio statistic and its p value.

$$G^2 = -2 \log \frac{L(R)}{L(F)} = 5.4106 \sim \chi^2_{df_R - df_F}$$

p-value= 0.02<0.05, so the home term should not be removed from the model since it is significant.

Comment on the Magnitude of home-field advantage

Based on the table above, we can find that the p-value of at.home = 0.02<0.025. So we conclude that the coefficient of at.home is not equal to 0, and home-field has an effect on the success probability of each team.

In this sense, the home team has an estimated odds-multiplier of $\exp(0.3023) = 1.35$ in its favour.

(c) Estimate the probability that Detroit beats Boston

(i) at Boston

We take Boston=1, home=1

We take Detroit =1, home=0

$$\log\left(\frac{\pi_{DB}}{1-\pi_{DB}}\right) = \theta_D - (\theta_B + \gamma) = 1.4754 - (1.1438 + 0.3023) = 0.0293$$

$$\pi_{DB} = 0.5073$$

So the probability that Detroit beats Boston is 0.5073

(ii) at Detroit

We take Boston=1, home=0

We take Detroit =1, home=1

$$\log\left(\frac{\pi_{DB}}{1-\pi_{DB}}\right) = \theta_D + \gamma - \theta_B = 1.4754 + 0.3023 - 1.1438 = 0.6339$$

$$\pi_{DB} = 0.6534$$

So the probability that Detroit beats Boston is 0.6534

(iii) on neutral territory.

We take Boston=1, home=0

We take Detroit =1, home=0

$$\log\left(\frac{\pi_{DB}}{1-\pi_{DB}}\right) = \theta_D - \theta_B = 1.4754 - 1.1438 = 0.6339$$

$$\pi_{DB} = 0.58215$$

So the probability that Detroit beats Boston is 0.58215

(d) Does the extended model fit the data? Comment briefly on any patterns in the residuals.

	ability	s.e.
Baltimore	0.000000	0.000000
Boston	1.1438027	0.3378422
Cleveland	0.7046945	0.3350014
Detroit	1.4753572	0.3445518
Milwaukee	1.6195550	0.3473653
New York	1.2813404	0.3404034
Toronto	1.3271104	0.3403222

Table 2.8 Ability of extended model

Based on the above model, we can find that the ability of the extended model is good. So, it fits the data well.

Comments on residuals pattern

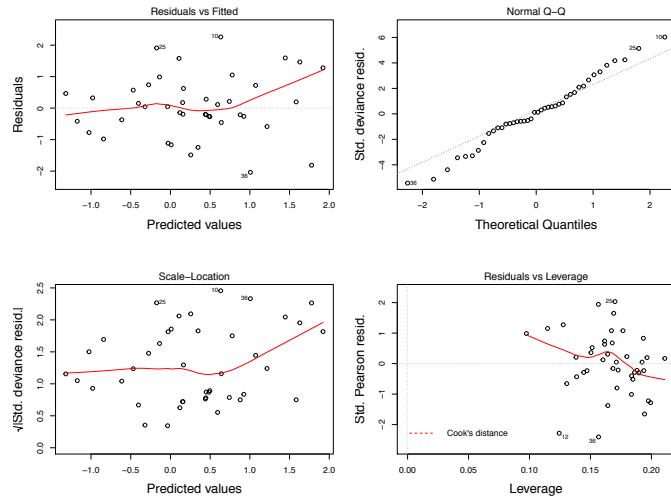


Figure 2.9 Residual Patterns

- (1) Normal distribution: Based on the qq-plot, we can find that the assumption that residuals are from normal distribution is satisfied.
- (2) Equal variance: Based on the residuals v.s fitted values plot, we can find that the assumption that equal variance is not also satisfied. In this way, we'd better to do some transformation to the data.

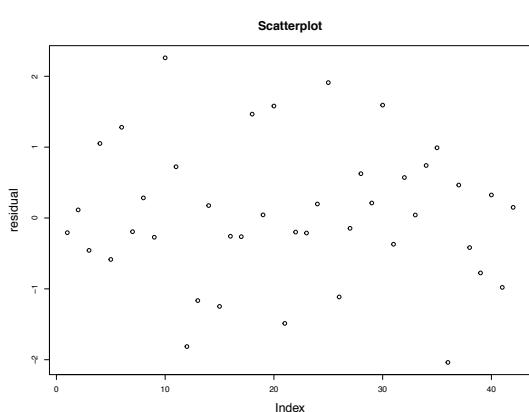


Figure 2.10 Scatterplot of Residuals

- (3) Independence: Based on the scatterplot above, we can conclude that the residuals are independent with each other so that the assumption of independence are satisfied.

3. Table 1 gives the mean number of children born per woman, the women being classified by place, education, and years since first marriage. Any systematic variation in the number of children is of interest. See fiji.txt for the dataset.

(a) Fit an appropriate model describing how the number of children varies with marital age, mother's abode and education. Give a brief synopsis of the arguments justifying your formulation and choice of model, including checks for model adequacy.

Fit an appropriate model

$$Y_{ijkl} \stackrel{iid}{\sim} Poisson(\mu_{ijkl})$$

$$Y_{ijk} \stackrel{iid}{\sim} Poisson(n_{ijk}\mu_{ijkl})$$

μ_{ijkl} : mean number of children born by l-th woman in (i,j,k)-th group

Thus, we can fit the model as follows:

$$\log(E(Y_{ijk})) = \log(n_{ijk}) + \beta_0 + \sum_{i=1}^5 \beta_i * marriage_i + \sum_{j=1}^3 \alpha_j * education_j + \gamma * place$$

Give a brief synopsis of the arguments

i : years since first marriage

j : education level

k : born place

$$marriage_i = \begin{cases} 1, & \text{when } i\text{th marriage group} (< 5 \text{ is a base line}) \\ 0, & \text{otherwise} \end{cases}$$

$$education_j = \begin{cases} 1, & \text{when } j\text{th education group} (\text{none is a base line}) \\ 0, & \text{otherwise} \end{cases}$$

$$place = \begin{cases} 0, & \text{urban} \\ 1, & \text{rural} \end{cases}$$

In this way, then we can analyze the data by fitting the **poisson model** to group totals.

```
Call: glm(formula = round(y) ~ marriage1 + marriage2 + marriage3 +
    marriage4 + marriage5 + edu1 + edu2 + edu3 + place1, family = poisson(link = "log"),
    data = XX, offset = log.n)

Coefficients:
(Intercept)   marriage1   marriage2   marriage3   marriage4   marriage5       edu1
              -0.08410     0.99715     1.41220     1.66473     1.84435     2.03017     0.03750
                  edu2       edu3      place1
                 -0.04710    -0.21093     0.06107

Degrees of Freedom: 45 Total (i.e. Null); 36 Residual
Null Deviance: 3035
Residual Deviance: 29.83      AIC: 346.3
```

Table 3.1 Log-linear poisson model

Based on the r summary table, we can get the formula as follows:

$$\log(E(Y)) = \log(n) - 0.0841 + 0.99715marriage_1 + 1.41220marriage_2 + 1.66473marriage_3 + 1.84435marriage_4 + 2.0302marriage_5 + 0.0375education_1 - 0.0471education_2 - 0.2109education_3 - 0.06107place$$

Checks for model adequacy (assumption)

We need to check the following to terms:

- (1) Independent

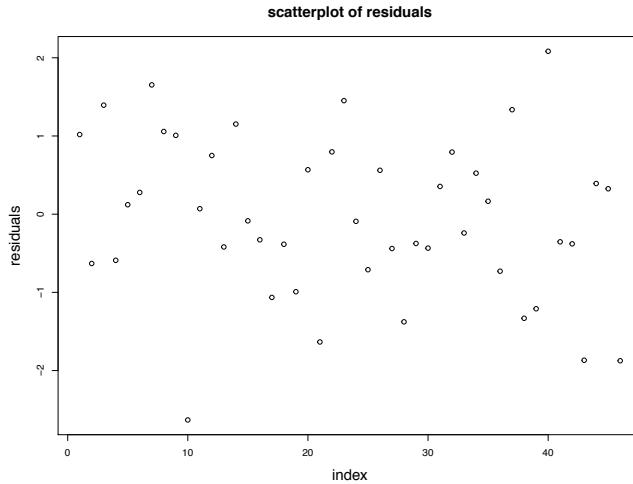


Figure 3.2 Scatterplot of Residuals

Based on the scatterplot above, we can conclude that the residuals are independent with each other so that the assumption of independence are satisfied.

- (2) Equal Variance & Normal distribution

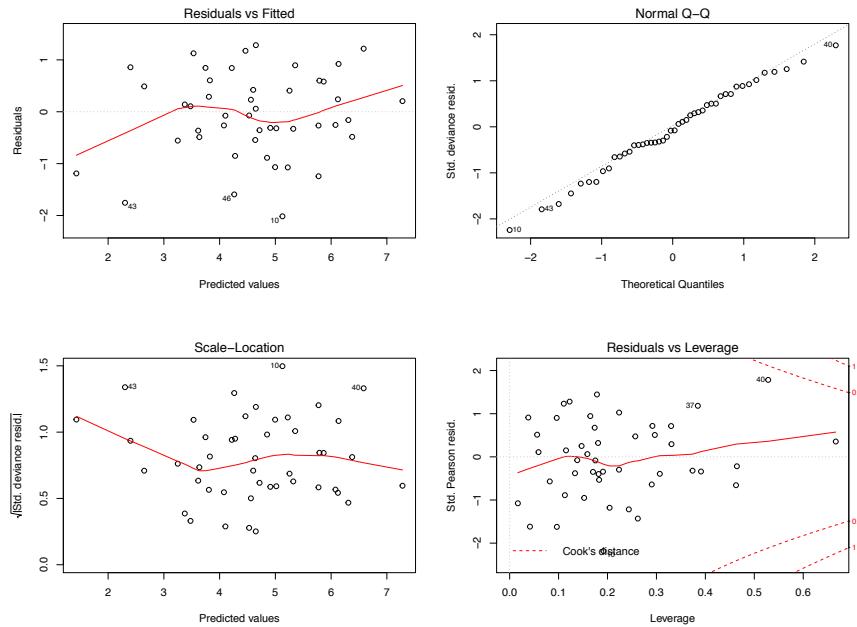


Figure 3.3 Checking Adequacy

Based on the qq-plot, we can find that the assumption that residuals are from normal distribution is satisfied.

Based on the residuals v.s fitted values plot, we can find that the assumption that equal variance is not satisfied since the variance of residuals are changed with the change of fitted values. In this way, we'd better to do some transformation to the data.

(3) Model Extension

Analysis of Deviance Table

Model: poisson, link: log

Response: round(y)

Terms added sequentially (first to last)

	Df	Deviance	Resid.	Df	Resid.	Dev
NULL				45	3035.43	
marriage	5	2973.82	40	61.61		
edu	3	25.68	37	35.93		
abode	1	6.09	36	29.83		
marriage:edu	14	11.29	22	18.54		
marriage:abode	5	3.82	17	14.72		
edu:abode	3	3.67	14	11.05		
marriage:edu:abode	14	11.05	0	0.00		

Table 3.4 Anova table for Extended Model

Based on the the anova table above, we can find that the interactions are not significant in the model, so we don't need to put them in the model and the original model is adequate.

(b) Explain the meaning of all parameters in your model. Comment on the major factors affecting fertility.

Meaning of all parameters

$$\log(E(Y_{ijkl})) = \beta_0 + \sum_{i=1}^5 \beta_i * marriage_i + \sum_{j=1}^3 \alpha_j * education_j + \gamma * place$$

β_0 : intercept(number of children born by l-th woman in (<5,none edu,rural) group)

β_i : representing how # of children changes with the change of years of marriage

$$marriage_i = \begin{cases} 1, & \text{when } i\text{-th marriage group} (<5 \text{ is a base line}) \\ 0, & \text{otherwise} \end{cases}$$

α_j : representing how # of children changes with the change of level of education

$$education_j = \begin{cases} 1, & \text{when } j\text{-th education group} (\text{none is a base line}) \\ 0, & \text{otherwise} \end{cases}$$

γ : representing how the change of place of born effect # of children

$$place = \begin{cases} 0, & \text{urban} \\ 1, & \text{rural} \end{cases}$$

Comment

```

Call:
glm(formula = round(y) ~ marriage1 + marriage2 + marriage3 +
    marriage4 + marriage5 + edu1 + edu2 + edu3 + place1, family = poisson(link = "log"),
    data = XX, offset = log.n)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-2.01562 -0.48720 -0.07375  0.55851  1.28367 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.08410   0.05876 -1.431   0.1524    
marriage1    0.99715   0.05869 16.990 <2e-16 ***  
marriage2    1.41220   0.05648 25.003 <2e-16 ***  
marriage3    1.66473   0.05628 29.580 <2e-16 ***  
marriage4    1.84435   0.05667 32.546 <2e-16 ***  
marriage5    2.03017   0.05546 36.606 <2e-16 ***  
edu1        0.03750   0.02450  1.531   0.1259    
edu2        -0.04710   0.03377 -1.395   0.1631    
edu3        -0.21093   0.06147 -3.431   0.0006 ***  
place1       0.06107   0.02484  2.458   0.0140 *    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 .’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 3035.429 on 45 degrees of freedom
Residual deviance: 29.831 on 36 degrees of freedom
AIC: 346.3

Number of Fisher Scoring iterations: 4

```

Table 3.5 Estimates for log-linear poisson model

Based on the table above, we can make a comment on the major factors affecting fertility.

Major factors are:

Marriage is the most effective factor among the three factors. Also, women with marriage years over 25years will have the highest fertility. That is said, the longer the marriage, the more the children to be born.

Education (secondary education) is the second effective factor. Women with highest level education will have least children than the other three kinds of education levels. That is said the higher the education level, the less the children to be born.

Place is the least effective factor. For this factor, women in rural will have more children than those in urban.

(c) Construct a 95% confidence interval for the mean number of children born to an urban woman with upper elementary education after ten years of marriage.

For 10-14 years:

"marriage1"=0,"marriage2"=1,"marriage3"=0,"marriage4"=0,"marriage5"=0,"edu1"=0,"edu2"=1,

"edu3"=0,"place1"=0,"log.n"=0

95% confidence interval

$$X_h'\hat{\beta} = 3.60022$$

$$se(X_h'\hat{\beta}) = 0.1404$$

$$CI \text{ for } X_h'\hat{\beta} \text{ is } (X_h'\hat{\beta} - z_{1-\alpha/2} * se(X_h'\hat{\beta}), X_h'\hat{\beta} + z_{1-\alpha/2} * se(X_h'\hat{\beta}))$$

$$CI \text{ for } X_h'\hat{\beta} \text{ is } (3.32496, 3.87548)$$

For 15-19 years:

"marriage1"=0,"marriage2"=0,"marriage3"=1,"marriage4"=0,"marriage5"=0,"edu1"=0,"edu2"=1,"edu3"=0,"place1"=0,"log.n"=0

95% confidence interval

$$X'_h \hat{\beta} = 4.634514$$

$$se(X'_h \hat{\beta}) = 0.1789317$$

CI for $X'_h \hat{\beta}$ is (4.2838, 4.98522)

For 20-24 years:

"marriage1"=0, "marriage2"=0, "marriage3"=0, "marriage4"=1, "marriage5"=0, "edu1"=0, "edu2"=1,

"edu3"=0, "place1"=0, "log.n"=0

95% confidence interval

$$X'_h \hat{\beta} = 5.5464$$

$$se(X'_h \hat{\beta}) = 0.2194622$$

CI for $X'_h \hat{\beta}$ is (5.116, 5.9765)

For 25+ years:

"marriage1"=0, "marriage2"=0, "marriage3"=0, "marriage4"=0, "marriage5"=1, "edu1"=0, "edu2"=1,

"edu3"=0, "place1"=0, "log.n"=0

95% confidence interval

$$X'_h \hat{\beta} = 6.678994$$

$$se(X'_h \hat{\beta}) = 0.2509071$$

CI for $X'_h \hat{\beta}$ is (6.187216, 7.170772)

(d) Estimate the lifetime average number of children born to rural women with secondary education. Give 90% confidence limits.

Based on the context, we take new dataset as

"marriage1"=0, "marriage2"=0, "marriage3"=0, "marriage4"=0, "marriage5"=1, "edu1"=0, "edu2"=0, "edu3"=1, "place1"=1, "log.n"=0

90% confidence interval

$$X'_h \hat{\beta} = 6.026728$$

$$se(X'_h \hat{\beta}) = 0.3873574$$

*CI for $X'_h \hat{\beta}$ is ($X'_h \hat{\beta} - z_{1-\alpha/2} * se(X'_h \hat{\beta}), X'_h \hat{\beta} + z_{1-\alpha/2} * se(X'_h \hat{\beta})$)*

CI for $X'_h \hat{\beta}$ is (5.389582, 6.663874)

4. The file byss.txt contains information, obtained from a survey conducted by a large textile company, on the prevalence of byssinosis, a lung disease to which cotton workers are subject. The file lists the observed prevalence of byssinosis (affected, not affected), by race (white = 1; non white = 2), sex (male = 1; female = 2), smoking habits (two levels), length of employment (three levels), and dustiness of the work environment (three levels). In the last three cases, higher-numbered categories denote larger values (more smoking, longer employment and increased dustiness). Parts (a) and (b) are based on the assumption that the main-effects linear logistic model is substantially correct.

(a) Fit the main-effects linear logistic model. Explain how the residual degrees of freedom is calculated for the deviance.

Fit the model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 * \text{race} + \beta_2 * \text{sex} + \beta_3 * \text{smokinghabit} + \sum_{i=1}^2 \alpha_i * \text{length of employment}_i + \sum_{j=1}^2 \gamma_j * \text{dustiness}_j$$

Explanatory Variables:

Race=1, if non-white, otherwise 0

Sex=1, if female, otherwise 0

Smoking habit=1, if =2, otherwise 0

Employment1=1, if =2, otherwise 0

Employment2=1, if =3, otherwise 0

Dustiness1=1, if=2, otherwise 0

Dustiness2=1, if=3, otherwise 0

Then we can have the result as follows:

```
Call: glm(formula = cbind(affected, not_affected) ~ race + sex + smok +
empl + dust, family = binomial(link = "logit"), data = byss)

Coefficients:
(Intercept)      race2       sex2       smok2      empl2      empl3      dust2      dust3 
-5.3172       0.1163     0.1239     0.6413     0.5641     0.7531     0.1507     2.7306 

Degrees of Freedom: 64 Total (i.e. Null);  57 Residual
Null Deviance:    322.5
Residual Deviance: 43.27      AIC: 165.9
```

Table 4.1 Linear Logistic Model

Based on the above table, we can find that the linear logistic model is the following form:

$$\begin{aligned} \log\left(\frac{p}{1-p}\right) = & -5.3172 + 0.1163 * \text{race} + 0.1239 * \text{sex} + 0.6413 * \text{smokinghabit} \\ & + 0.5641 * \text{employment}_1 + 0.7531 * \text{employment}_2 + 0.1507 * \text{dustiness}_1 + 2.7306 * \text{dustiness}_2 \end{aligned}$$

Explanation

Residual degrees of freedom is 57 is which is calculated for the residual deviance is 43.27.

Null degrees of freedom is 64 which is calculated for the null deviance is 322.5.

Thus, we can find that:

$$\text{residual degrees of freedom}(64) = \text{null degree of freedom}(57) - \# \text{ of parameters (8)} + 1$$

(b) Interpret the coefficient of sex(2). Construct an approximate 90% confidence interval for the odds ratio (males vs females) of contracting byssinosis.

Interpret the coefficient of sex(2)

sex(2) means female, coefficient of sex(2) is beta2 in the model.

$$\log\left(\frac{p_{female}}{1-p_{female}}\right) - \log\left(\frac{p_{male}}{1-p_{male}}\right) = \beta_2$$

Beta2 represents the difference between log(odds) for female and log(odds) for male in prevalence of byssinosis.

90% confidence interval

$$maleodds = \log\left(\frac{p_{male}}{1-p_{male}}\right)$$

$$femaleodds = \log\left(\frac{p_{female}}{1-p_{female}}\right)$$

$$\log(OR) = \log\left(\frac{p_{male}/(1-p_{male})}{p_{female}/(1-p_{female})}\right) = -\beta_2$$

$$OR = \exp(-\beta_2)$$

So the 90% confidence interval for the odds ratio (males vs females) of contracting byssinosis is the 90% confidence interval for the coefficient **exp(-beta2)**.

90% confidence interval for -beta2: (-0.4973715, 0.2565643)

90% confidence interval for exp(-beta2): (0.608127, 1.292482)

(c) Drop the least significant factor from the model, proceeding until all the remaining factors are significant at the 5% level. Interpret the reduced model thus obtained.

Drop the least significant factor

Coefficients:					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.3172	0.3028	-17.559	< 2e-16	***
race2	0.1163	0.2072	0.562	0.574426	
sex2	0.1239	0.2288	0.542	0.587983	
smok2	0.6413	0.1944	3.299	0.000971	***
empl2	0.5641	0.2617	2.156	0.031091	*
empl3	0.7531	0.2161	3.484	0.000493	***
dust2	0.1507	0.2860	0.527	0.598194	
dust3	2.7306	0.2153	12.681	< 2e-16	***

Table 4.2 Full model

Based on the above table, we find that the least significant factor is sex2. So we drop it from the model.

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-5.2383	0.2633	-19.892	< 2e-16 ***
race2	0.1169	0.2068	0.565	0.572115
smok2	0.6213	0.1908	3.257	0.001128 **
empl2	0.5464	0.2593	2.107	0.035109 *
empl3	0.7357	0.2133	3.450	0.000561 ***
dust2	0.1684	0.2841	0.593	0.553386
dust3	2.6860	0.1978	13.577	< 2e-16 ***

Table 4.3 Drop step one

Based on the above table, we find that the least significant factor is race2. So we drop it from the model.

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-5.1722	0.2349	-22.014	< 2e-16 ***
smok2	0.6210	0.1908	3.255	0.001133 **
empl2	0.5060	0.2490	2.032	0.042119 *
empl3	0.6728	0.1813	3.710	0.000207 ***
dust2	0.1682	0.2841	0.592	0.553915
dust3	2.7175	0.1898	14.314	< 2e-16 ***

Table 4.4 Drop step two

Interpret the reduced model

Now, all the remaining factors are significant at the 5% level.

The remaining model is:

$$\log\left(\frac{p}{1-p}\right) = -5.1722 + 0.621 * \text{smokinghabit} + 0.5060 * \text{employment}_2 \\ + 0.6728 * \text{employment}_3 + 0.1682 * \text{dustiness}_2 + 2.7175 * \text{dustiness}_3$$

Parameters:

Intercept: representing log(odds) for group (white, male, smoke habit level=1, employment level=1, dust level=1), which is -5.1722

Coefficient of smoking habit: representing difference between log(odds) for smoke habit level=2 and log(odds) for smoke habit level=1

Coefficient of employment2: representing difference between log(odds) for employment level=2 and log(odds) for employment level=1

Coefficient of employment3: representing difference between log(odds) for employment level=3 and log(odds) for employment level=1

Coefficient of dust2: representing difference between log(odds) for dust level=2 and log(odds) for dust level=1

Coefficient of dust3: representing difference between log(odds) for dust level=3 and log(odds) for dust level=1

(d) Beginning with the complete main-effects model, look for significant interactions by fitting each of the ten models main effects + one interaction. In judging the significance of interactions, you should bear in mind, at least informally, the effects of selection. After detecting the significant interactions, remove insignificant main effects as described in (c), except for those that are included in interactions. Interpret the model thus obtained.

Fitting each of the ten models main effects + one interaction

Complete main-effects model

$$\log\left(\frac{p}{1-p}\right) = -5.3172 + 0.1163 * race + 0.1239 * sex + 0.6413 * smokinghabit \\ + 0.5641 * employment_1 + 0.7531 * employment_2 + 0.1507 * dustiness_1 + 2.7306 * dustiness_2$$

1st model: main effects + race*sex

Coefficients:					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.31210	0.30396	-17.476	< 2e-16 ***	
race2	0.10172	0.22180	0.459	0.646530	
sex2	0.09406	0.28045	0.335	0.737342	
smok2	0.63995	0.19457	3.289	0.001005 **	
empl2	0.57060	0.26429	2.159	0.030850 *	
empl3	0.75788	0.21810	3.475	0.000511 ***	
dust2	0.15163	0.28604	0.530	0.596046	
dust3	2.72954	0.21502	12.694	< 2e-16 ***	
race2:sex2	0.07393	0.39816	0.186	0.852691	

The coefficient of race*sex is 0.0739, and p-value is 0.8527>0.05.
So it is insignificant interaction and should be removed.

2nd model: main effects + race*smoking habit

Coefficients:					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.4914	0.3445	-15.939	< 2e-16 ***	
race2	0.4568	0.3569	1.280	0.200569	
sex2	0.1401	0.2287	0.612	0.540319	
smok2	0.8447	0.2672	3.162	0.001567 **	
empl2	0.5794	0.2624	2.208	0.027227 *	
empl3	0.7709	0.2171	3.551	0.000383 ***	
dust2	0.1517	0.2860	0.530	0.595805	
dust3	2.7362	0.2151	12.718	< 2e-16 ***	
race2:smok2	-0.4480	0.3832	-1.169	0.242439	

The coefficient of race*smoking habit is -0.448, and p-value is 0.2424>0.25.
So it is insignificant interaction and should be removed.

3rd model: main effects + race*employment

Coefficients:					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.5939	0.3759	-14.882	< 2e-16 ***	
race2	0.4641	0.3343	1.388	0.165046	
sex2	0.1079	0.2287	0.472	0.637048	
smok2	0.6500	0.1946	3.340	0.000836 ***	
empl2	0.9499	0.3920	2.423	0.015387 *	
empl3	1.0572	0.3260	3.243	0.001183 **	
dust2	0.1462	0.2861	0.511	0.609327	
dust3	2.7412	0.2155	12.718	< 2e-16 ***	
race2:empl2	-0.6804	0.5566	-1.222	0.221530	
race2:empl3	-0.5571	0.4732	-1.177	0.239041	

The coefficient of race*employment are -0.6804 and -0.5571, and p-value are 0.2215>0.05 and 0.239>0.05. So it is insignificant interaction and should be removed.

4th model: main effects + race*dustiness

Coefficients:					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.3553	0.3175	-16.869	< 2e-16 ***	
race2	0.2186	0.3742	0.584	0.559152	
sex2	0.1246	0.2296	0.543	0.587272	
smok2	0.6410	0.1944	3.297	0.000977 ***	
empl2	0.5869	0.2639	2.224	0.026177 *	
empl3	0.7701	0.2184	3.526	0.000421 ***	
dust2	0.0712	0.3443	0.207	0.836157	
dust3	2.8057	0.2570	10.915	< 2e-16 ***	
race2:dust2	0.2560	0.6161	0.416	0.677769	
race2:dust3	-0.1880	0.4158	-0.452	0.651275	

The coefficient of race*dustiness are 0.2560 and -0.188, and p-value are 0.678>0.05 and 0.65 > 0.05.

So it is insignificant interaction and should be removed.

5th model: main effects + sex* smoking habit

Coefficients:					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.4078	0.3260	-16.590	< 2e-16 ***	
race2	0.1146	0.2076	0.552	0.580826	
sex2	0.3496	0.3508	0.997	0.318988	
smok2	0.7575	0.2432	3.115	0.001838 **	
empl2	0.5569	0.2621	2.125	0.033584 *	
empl3	0.7436	0.2166	3.432	0.000599 ***	
dust2	0.1511	0.2859	0.529	0.597111	
dust3	2.7363	0.2156	12.694	< 2e-16 ***	
sex2:smok2	-0.3492	0.4156	-0.840	0.400750	

The coefficient of sex* smoking habit is -0.3492, and p-value is 0.4001>0.05. So it is insignificant interaction and should be removed.

6th model: main effects + sex*employment

Coefficients:					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.3586	0.3060	-17.514	< 2e-16	***
race2	0.1047	0.2065	0.507	0.611934	
sex2	0.3266	0.2911	1.122	0.261852	
smok2	0.6237	0.1949	3.201	0.001370	**
empl2	0.6291	0.2908	2.164	0.030476	*
empl3	0.8710	0.2392	3.641	0.000271	***
dust2	0.1651	0.2862	0.577	0.563950	
dust3	2.7191	0.2138	12.717	< 2e-16	***
sex2:empl2	-0.2006	0.6233	-0.322	0.747558	
sex2:empl3	-0.5098	0.4354	-1.171	0.241664	

The coefficient of sex*employment are -0.2 and -0.51, and p-value are 0.748>0.05 and 0.242>0.05.

So it is insignificant interaction and should be removed.

7th model: main effects + sex* dustiness

Coefficients:					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.3538	0.3211	-16.674	< 2e-16	***
race2	0.1127	0.2065	0.546	0.585324	
sex2	0.2602	0.3161	0.823	0.410400	
smok2	0.6582	0.1946	3.383	0.000717	***
empl2	0.5030	0.2615	1.923	0.054470	.
empl3	0.6979	0.2158	3.234	0.001219	**
dust2	-0.3949	0.5464	-0.723	0.469908	
dust3	2.8488	0.2482	11.479	< 2e-16	***
sex2:dust2	0.7410	0.6499	1.140	0.254253	
sex2:dust3	-1.2609	0.6823	-1.848	0.064618	.

The coefficient of sex* dustiness are 0.741 and -1.261, and p-value are 0.254>0.05 and 0.065<0.1.

So sex2*dust2 is insignificant interaction and should be removed.

But sex2*dust3 is significant interaction and should be added to the model.

8th model: main effects + smoking habit*employment

Coefficients:					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.1401	0.3324	-15.464	<2e-16	***
race2	0.1091	0.2076	0.525	0.599	
sex2	0.1363	0.2285	0.596	0.551	
smok2	0.3878	0.2866	1.353	0.176	
empl2	0.4934	0.5319	0.928	0.354	
empl3	0.3545	0.3726	0.951	0.341	
dust2	0.1581	0.2860	0.553	0.580	
dust3	2.7379	0.2152	12.725	<2e-16	***
smok2:empl2	0.1160	0.5953	0.195	0.845	
smok2:empl3	0.5416	0.4106	1.319	0.187	

The coefficient of smoking habit*employment are 0.116 and 0.542, and p-value are 0.845>0.05 and 0.187>0.05.

So it is insignificant interaction and should be removed.

9th model: main effects + smoking habit*dustiness

Coefficients:					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.16967	0.36542	-14.147	< 2e-16	***
race2	0.10873	0.20776	0.523	0.600729	
sex2	0.06514	0.23453	0.278	0.781196	
smok2	0.49103	0.34304	1.431	0.152318	
empl2	0.54542	0.26233	2.079	0.037607	*
empl3	0.73095	0.21643	3.377	0.000732	***
dust2	0.55796	0.43803	1.274	0.202739	
dust3	2.35282	0.39702	5.926	3.1e-09	***
smok2:dust2	-0.69091	0.58194	-1.187	0.235128	
smok2:dust3	0.46890	0.43745	1.072	0.283767	

The coefficient of smoking habit*dustiness are -0.6901 and 0.4689, and p-value are 0.2351>0.05 and 0.2838>0.05.

So it is insignificant interaction and should be removed.

10th model: main effects + employment*dustiness

Coefficients:					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.05503	0.33336	-15.164	< 2e-16	***
race2	0.09204	0.20422	0.451	0.65222	
sex2	0.12785	0.23120	0.553	0.58025	
smok2	0.63711	0.19454	3.275	0.00106	**
empl2	-0.24846	0.56461	-0.440	0.65990	
empl3	0.44499	0.34566	1.287	0.19797	
dust2	0.05584	0.42843	0.130	0.89629	
dust3	2.34715	0.31196	7.524	5.32e-14	***
empl2:dust2	0.86588	0.88100	0.983	0.32569	
empl3:dust2	-0.07815	0.61641	-0.127	0.89911	
empl2:dust3	1.11876	0.63926	1.750	0.08010	.
empl3:dust3	0.51177	0.40497	1.264	0.20633	

The coefficient of employment*dustiness habit are 0.8659, -0.078, 1.1188 and 0.512, and p-value are 0.33>0.05, 0.899>0.05, 0.08<0.1 and 0.21>0.05.

So empl2*dust3 is significant interaction and should be remained. The others should be removed.

Final model with interaction

Based on the above analysis, and remove insignificant main effects described in (c), then we find that we should include sex, smoking habit, employment, dust and interaction dust*sex, as explanatory variables in the final model since the anova table can show that it is the best adequate model. And then we can get the results as follows:

$$y \sim \text{sex} + \text{smokinghabit} + \text{employment} + \text{dust} + \text{sex} * \text{dust}$$

```

Call:
glm(formula = y ~ sex + smokinghabit + employment2new + employment3new +
dust3new + sex * dust3new, family = binomial(link = "logit"))

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-0.7536 -0.1845 -0.1673 -0.1330  3.2768 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -5.3640    0.2844 -18.860 < 2e-16 ***
sex          0.4612    0.2679   1.722 0.085114 .  
smokinghabit 0.6585    0.1946   3.385 0.000713 *** 
employment2new 0.4617    0.2510   1.839 0.065891 .  
employment3new 0.6406    0.1836   3.489 0.000485 *** 
dust3new      2.9514    0.2269  13.007 < 2e-16 *** 
sex:dust3new -1.4592    0.6616  -2.205 0.027420 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1477.2 on 5418 degrees of freedom
Residual deviance: 1192.2 on 5412 degrees of freedom
AIC: 1206.2

```

Number of Fisher Scoring iterations: 7

Table 4.5 Coefficients for the final model

	Df	Deviance	Resid.	Df	Resid.	Dev
NULL			5418		1477.2	
sex	1	41.344	5417		1435.8	
smokinghabit	1	9.854	5416		1426.0	
employment2new	1	0.150	5415		1425.8	
employment3new	1	5.942	5414		1419.9	
dust3new	1	221.379	5413		1198.5	
sex:dust3new	1	6.295	5412		1192.2	

Table 4.6 Anova table for the final model

Interpretation

This model represents that sex, smoking habit, employment, dust and interaction dust*sex, these items are significant in the final model and they effect the response variable in a large degree.

(e) You are required to write a short report giving details of the excess risk associated with cotton dust. How fast does the risk increase with dust level? If necessary, give separate figures for males and females or for smokers and non-smokers.

a. General Analysis

(1) Fitting the model

```
Call: glm(formula = y ~ dust2new + dust3new, family = binomial(link = "logit"))
```

```
Coefficients:
(Intercept)      dust2new      dust3new
-4.3962        0.1304       2.7151
```

```
Degrees of Freedom: 5418 Total (i.e. Null); 5416 Residual
Null Deviance: 1477
Residual Deviance: 1225           AIC: 1231
```

Table 4.7 Model y~dust2+dust3

Then we get the model as follows:

$$\log\left(\frac{p}{1-p}\right) = -4.3962 + 0.1304 * dustiness_2 + 2.7151 * dustiness_3$$

(2) Analysis

Based on the model, we can calculate how fast does the risk increase with dust level.

The OR represents how fast the risk increase with dust level.

For dust level=1 and dust level=2

$$\log\left(\frac{p_2}{1-p_2}\right) - \log\left(\frac{p_1}{1-p_1}\right) = \log(OR_{21}) = 0.1304 \Rightarrow OR_{21} = 1.1393$$

So the dust level=2 has 1.1393 times risk than dust level=1

For dust level=2 and dust level=3

$$\log\left(\frac{p_3}{1-p_3}\right) - \log\left(\frac{p_2}{1-p_2}\right) = \log(OR_{32}) = 2.5847 \Rightarrow OR_{32} = 13.259$$

So the dust level=3 has 13.259 times risk than dust level=2

For dust level=1 and dust level=3

$$\log\left(\frac{p_3}{1-p_3}\right) - \log\left(\frac{p_1}{1-p_1}\right) = \log(OR_{31}) = 2.7151 \Rightarrow OR_{31} = 15.106$$

So the dust level=3 has 15.106 times risk than dust level=1

b. Categorical Analysis

(1) For males and females

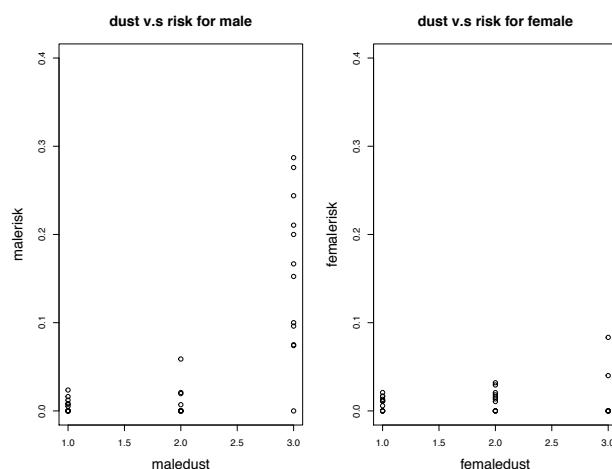


Figure 4.8 Dust v.s Risk in sex

Based on the above figures, we conclude that the increased dustiness will both affect (increase) the male and female's risk of prevalence of byssinosis. But we can also see that the increasing level of dust will affect more strongly on male instead of female. This is say that male will increase the risk with the increase of dust level at a rate which is higher than that of female.

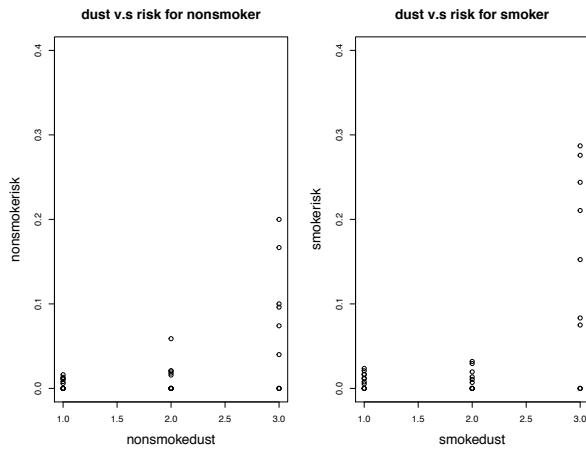
(2) For smokers and non-smokers

Figure 4.9 Dust v.s Risk in smoking habit

Based on the above figures, we conclude that the increased dustiness will both affect (increase) the smoker and nonsmoker's risk of prevalence of byssinosis. And the increasing level of dust will affect smokers slightly stronger than nonsmokers. This is say that smokers will increase the risk with the increase of dust level at a rate which is slightly higher than that of nonsmokers.

(f) Does this analysis suggest that the aetiology of byssinosis is related to sex or race? Explain.

Yes, some analysis suggests that the aetiology of byssinosis is related to sex. Since sex*dust is a significant interaction in the model, so the sex term should be remained in the model.

But the race itself and the interactions including it both have little significance in the model. So there is no analysis suggest that the aetiology of byssinosis is related to race.

Appendix

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.82864	0.14744	25.967	< 2e-16 ***
SESB	-0.13976	0.21619	-0.646	0.51797
SESC	0.26570	0.19597	1.356	0.17516
SESD	0.71465	0.17994	3.972	7.14e-05 ***
SESE	0.52807	0.18590	2.841	0.00450 **
SESF	0.43404	0.18927	2.293	0.02184 *
mentalMild	0.71465	0.17994	3.972	7.14e-05 ***
mentalModerate	0.23180	0.19743	1.174	0.24037
mentalWell	0.33024	0.19330	1.708	0.08755 .
SESB:mentalMild	0.13976	0.26080	0.536	0.59203
SESC:mentalMild	-0.15504	0.24201	-0.641	0.52176
SESD:mentalMild	-0.30919	0.22385	-1.381	0.16720
SESE:mentalMild	-0.49665	0.23560	-2.108	0.03503 *
SESF:mentalMild	-0.71465	0.24606	-2.904	0.00368 **
SESB:mentalModerate	0.06830	0.28723	0.238	0.81203
SESC:mentalModerate	-0.15176	0.26652	-0.569	0.56908
SESD:mentalModerate	-0.43129	0.25021	-1.724	0.08476 .
SESE:mentalModerate	-0.59953	0.26518	-2.261	0.02377 *
SESF:mentalModerate	-0.50550	0.26755	-1.889	0.05884 .
SESB:mentalWell	0.02393	0.28268	0.085	0.93254
SESC:mentalWell	-0.38153	0.26753	-1.426	0.15383
SESD:mentalWell	-0.59687	0.24878	-2.399	0.01643 *
SESE:mentalWell	-1.10343	0.27922	-3.952	7.75e-05 ***
SESF:mentalWell	-1.54840	0.31475	-4.919	8.68e-07 ***

Figure: Summary R table for problem 1

```
#1
SES=c("A","A","A","A","B","B","B","C","C","C","D","D","D","E","E",
      "E","F","F","F","F")
mental=c("Well","Mild","Moderate","Impaired","Well","Mild","Moderate","Impaired",
       ,"Well","Mild","Moderate","Impaired","Well","Mild","Moderate","Impaired","Well",
       "Mild","Moderate","Impaired","Well","Mild","Moderate","Impaired")
y=c(64,94,58,46,57,94,54,40,57,105,65,60,72,141,77,94,36,97,54,78,21,71,54,71)
#regression
glm1=glm(y~mental,family = poisson(link="log"))
summary(glm1)
glm2=glm(y~SES,family = poisson(link="log"))
summary(glm2)
glm11=glm(y~SES+mental,family = poisson(link="log"))
summary(glm11)
anova(glm11)
glm12=glm(y~SES+mental+SES*mental,family = poisson(link="log"))
anova(glm12)
summary(glm12)
#
install.packages("BradleyTerry2")
library(BradleyTerry2)
data("baseball",package = "BradleyTerry2")
head(baseball)
```

```
#original model
baseballModel1=BTm(cbind(home.wins,away.wins),home.team, away.team,data =
baseball,id="team")
baseballModel1
summary(baseballModel1)
#extended model
baseball$home.team=data.frame(team=baseball$home.team,at.home=1)
baseball$away.team=data.frame(team=baseball$away.team,at.home=0)
baseballModel2=update(baseballModel1,formula= ~ team+at.home)
summary(baseballModel2)
#Obtain the likelihood-ratio statistic
anova(baseballModel1,baseballModel2)
1-pchisq(5.4106,1)
#d
BTabilities(baseballModel1)
par(mfrow=c(2,2))
plot(baseballModel2,cex.lab=1.5,cex.main=1.5)
residual=residuals(baseballModel2)
par(mfrow=c(1,1))
plot(residual,main = "Scatterplot",cex.lab=1.5,cex.main=1.5)


---


#3
#a
#read data
fiji=read.table("fiji.txt",header = T)
fijinew=fiji[c(-24,-48),]
fijinew$marriage=factor(fijinew$marriage)
fijinew$edu=factor(fijinew$edu)
fijinew$abode=factor(fijinew$abode)
y=fijinew$tot*fijinew$average
log.n=log(fijinew$tot)
#model
glm1=glm(data=XX,round(y)~marriage1+marriage2+marriage3+marriage4+
          marriage5+edu1+edu2+edu3+place1,offset=log.n,family =
poisson(link="log"))
glmnew=glm(data=fijinew,round(y)~marriage+edu+abode,offset=log.n,family =
poisson(link="log"))
glmnewnew=glm(data=fijinew,round(y)~marriage+edu+abode+marriage*edu+marriag
e*abode+edu*abode+marriage*edu*abode,offset=log.n,family = poisson(link="log"))
#checking
install.packages("MASS")
library(MASS)
install.packages("lmtest")
install.packages("nortest")
library(nortest)
library(lmtest)
library(base)
```

```
par(mfrow=c(2,2))
plot(glm1)
#scatterplot
par(mfrow=c(1,1))
plot(studres(glm1),ylab="residuals",xlab = "index",main = "scatterplot of
residuals",cex.lab=1.5,cex.main=1.5)
## fit vs res ##
plot(glm1$fitted.values ,studres(glm1),xlab="Fitted values",
      ylab="Studentized residuals",
      main="Residual vs Fitted",cex.lab=1.5,cex.main=1.5)
abline(h=0);abline(h=3,lty=2);abline(h=-3,lty=2)
## res QQ ##
qqnorm(studres(glm1),ylab="Studentized residuals",
      ylim=c(-2,2),cex.lab=1.5,cex.main=1.8)
qqline(studres(glm1))
#b
summary(glm1)
#c&d
#predict
pnew1=predict.glm(glm1,newdata =
data.frame("marriage1"=0,"marriage2"=1,"marriage3"=0,"marriage4"=0,
"marriage5"=0,"edu1"=0,"edu2"=1,"edu3"=0,"place1"=0,"log.n"=0),
type="response",se.fit = T)
pnew2=predict.glm(glm1,newdata =
data.frame("marriage1"=0,"marriage2"=0,"marriage3"=1,"marriage4"=0,
"marriage5"=0,"edu1"=0,"edu2"=1,"edu3"=0,"place1"=0,"log.n"=0),
type="response",se.fit = T)
pnew3=predict.glm(glm1,newdata =
data.frame("marriage1"=0,"marriage2"=0,"marriage3"=0,"marriage4"=1,
"marriage5"=0,"edu1"=0,"edu2"=1,"edu3"=0,"place1"=0,"log.n"=0),
type="response",se.fit = T)
pnew4=predict.glm(glm1,newdata =
data.frame("marriage1"=0,"marriage2"=0,"marriage3"=0,"marriage4"=0,
"marriage5"=1,"edu1"=0,"edu2"=1,"edu3"=0,"place1"=0,"log.n"=0),
type="response",se.fit = T)
pnew25=predict.glm(glm1,newdata =
data.frame("marriage1"=0,"marriage2"=0,"marriage3"=0,"marriage4"=0,"marriage5"=
1,"edu1"=0,"edu2"=0,"edu3"=1,"place1"=1,"log.n"=0), type="response",se.fit = T)
#4
#a
byss=read.table("byss.txt",header = T)
byss
length(byss$affected)
byss$race=factor(byss$race)
byss$sex=factor(byss$sex)
byss$smok=factor(byss$smok)
```

```
byss$empl=factor(byss$empl)
byss$dust=factor(byss$dust)
m1=glm(cbind(affected,not_affected)~race+sex+smok+empl+dust,data=byss,family =
binomial(link="logit"))
summary(m1)
m2=glm(cbind(affected,not_affected)~race+smok+empl+dust,data=byss,family =
binomial(link="logit"))
summary(m2)
m3=glm(cbind(affected,not_affected)~smok+empl+dust,data=byss,family =
binomial(link="logit"))
summary(m3)
#d
glminter1=glm(cbind(affected,not_affected)~race+sex+smok+empl+dust+race*sex,dat
a=byss,family = binomial(link="logit"))
glminter1$coefficients
summary(glminter1)
glminter2=glm(cbind(affected,not_affected)~race+sex+smok+empl+dust+race*smok,d
ata=byss,family = binomial(link="logit"))
summary(glminter2)
glminter3=glm(cbind(affected,not_affected)~race+sex+smok+empl+dust+race*empl,d
ata=byss,family = binomial(link="logit"))
summary(glminter3)
glminter4=glm(cbind(affected,not_affected)~race+sex+smok+empl+dust+race*dust,da
ta=byss,family = binomial(link="logit"))
summary(glminter4)
glminter5=glm(cbind(affected,not_affected)~race+sex+smok+empl+dust+sex*smok,da
ta=byss,family = binomial(link="logit"))
summary(glminter5)
glminter6=glm(cbind(affected,not_affected)~race+sex+smok+empl+dust+sex*empl,da
ta=byss,family = binomial(link="logit"))
summary(glminter6)
glminter7=glm(cbind(affected,not_affected)~race+sex+smok+empl+dust+sex*dust,dat
a=byss,family = binomial(link="logit"))
summary(glminter7)
glminter8=glm(cbind(affected,not_affected)~race+sex+smok+empl+dust+smok*empl,
data=byss,family = binomial(link="logit"))
summary(glminter8)
glminter9=glm(cbind(affected,not_affected)~race+sex+smok+empl+dust+smok*dust,d
ata=byss,family = binomial(link="logit"))
summary(glminter9)
glminter10=glm(cbind(affected,not_affected)~race+sex+smok+empl+dust+empl*dust,
data=byss,family = binomial(link="logit"))
summary(glminter10)
#e
#dust and risk
plot(dust1,y)
```

```
plot(dust2,y)
glmdust=glm(y~dust2new+dust3new,family = binomial(link="logit"))
par(mfrow=c(1,1))
abline(glmdust)
curve(expr =
exp(glmdust$coefficients[1]+glmdust$coefficients[2]*x)/(1+exp(glmdust$coefficients[1]+glmdust$coefficients[2]*x)),add = T)
#for sex
byss=read.table("byss.txt",header = T)
malerisk=c()
maledust=c()
femalerisk=c()
femaledust=c()
byss=read.table("byss.txt",header = T)
for (a in 1:length(byss$sex)){
if (byss$sex[a]==1){
maleriska=byss$affected[a]/(byss$affected[a]+byss$not_affected[a])
malerisk=c(malerisk,maleriska)
maledust=c(maledust,byss$dust[a])
} else {
femaleriska=byss$affected[a]/(byss$affected[a]+byss$not_affected[a])
femalerisk=c(femalerisk,femaleriska)
femaledust=c(femaledust,byss$dust[a])
}
}
par(mfrow=c(1,2))
plot(maledust,malerisk,main = "dust v.s risk for
male",cex.lab=1.5,cex.main=1.5,ylim=c(0,0.4))
plot(femaledust,femalerisk,main = "dust v.s risk for
female",cex.lab=1.5,cex.main=1.5,ylim = c(0,0.4))
#for smoker
byss=read.table("byss.txt",header = T)
nonsmokerisk=c()
nonsmokedust=c()
smokerisk=c()
smokedust=c()
for (a in 1:length(byss$smok)){
if (byss$smok[a]==1){
nonsmokeriska=byss$affected[a]/(byss$affected[a]+byss$not_affected[a])
nonsmokerisk=c(nonsmokerisk,nonsmokeriska)
nonsmokedust=c(nonsmokedust,byss$dust[a])
} else {
smokeriska=byss$affected[a]/(byss$affected[a]+byss$not_affected[a])
smokerisk=c(smokerisk,smokeriska)
smokedust=c(smokedust,byss$dust[a])
}
}
```

```
}  
par(mfrow=c(1,2))  
plot(nonsmokedust,nonsmokerisk,main = "dust v.s risk for  
nonsmoker",cex.lab=1.5,cex.main=1.5,ylim=c(0,0.4))  
plot(smokedust,smokerisk,main = "dust v.s risk for  
smoker",cex.lab=1.5,cex.main=1.5,ylim=c(0,0.4))
```
