UPPSALA UNIVERSITY
Department of Mathematics
Rolf Larsson

Exam in Mathematical Statistics
Analysis of Categorical Data 1MS370
2022–01–10
Solutions

1. A discrete random variable Y has probability mass function

$$p(y; \mu) = \left(1 - \frac{1}{\mu}\right)^{y-1} \frac{1}{\mu},$$

for $y = 0, 1, 2, \ldots$ and $\mu > 1$.

(a) Does this distribution belong to the exponential family, and in that case, why? (2p)

Solution: The distribution belongs to the exponential family if and only if the probability mass function/density can be written in the form

$$p(y; \mu) = a(\mu)b(y) \exp\{yQ(\mu)\},$$

where $Q(\mu)$ is the natural parameter. In our case, we have

$$p(y; \mu) = \frac{1}{\mu}\left(1 - \frac{1}{\mu}\right)^{-1} \exp\left\{y \log\left(1 - \frac{1}{\mu}\right)\right\},$$

which is of exponential family form with $a(\mu) = \frac{1}{\mu}\left(1 - \frac{1}{\mu}\right)^{-1} = \frac{1}{\mu-1}$, $b(y) = 1$ and natural parameter $Q(\mu) = \log\left(1 - \frac{1}{\mu}\right)$.

(b) Suggest an appropriate link function $g(\mu)$. (2p)

Solution: According to the general advice, we take $g(\mu) = Q(\mu)$, so we have

$$g(\mu) = \log\left(1 - \frac{1}{\mu}\right) = \log\left(\frac{\mu - 1}{\mu}\right).$$

(c) Let x be an explanatory variable that can take any real value. Discuss if the GLM $g(\mu) = \alpha + \beta x$ is a suitable model. (2p)

Solution: It is not a suitable model, because it can not take positive values (note: $\mu > 1$). This imposes a serious restriction on the function $\alpha + \beta x$.

2. The numbers of deaths in motor vecicle accidents in the counties of Stockholm and Norrbotten for the years 1976 and 1996, men and women, are given in the following table. (Data from Statistics Sweden.)

year	gender	Stockholm	Norrbotten
1976	men	99	34
	women	53	15
1996	men	46	13
	women	21	7

(a) For the 1976 data, estimate the odds ratio, calculate a 95% confidence interval for the odds ratio and interpret. (2p)

Solution: The odds ratio, θ, is estimated as

$$\hat{\theta} = \frac{99 \cdot 15}{34 \cdot 53} \approx 0.8241.$$

A confidence interval for $\log\theta$ with approximate confidence level 95% is

$$\begin{aligned} I_{\log\theta} &= \log 0.8241 \pm 1.96\sqrt{\frac{1}{99} + \frac{1}{34} + \frac{1}{53} + \frac{1}{15}} \\ &= -0.1935 \pm 0.6931 = (-0.8866,\ 0.4996). \end{aligned}$$

Because $-0.8866 < \log\theta < 0.4996$ is equivalent to

$$0.41 \approx \exp(-0.8866) < \theta < \exp(0.4996) \approx 1.65,$$

we have the confidence interval for θ with approximate confidence level 95% as $(0.41,\ 1.65)$.

Because this interval covers 1, there is no evidence of any association (dependence) between gender and region.

(b) For the 1976 data, test if gender is independent of county with respect to tendency of dying in motor vehicle accidents. Use the Pearson X^2 test or the likelihood ratio test. (2p)

Solution: We want to test H_0: gender is independent of county vs H_1: $\neg H_0$ (dependence). The row sums are 133, 68 and the column sums are 152 and 49. The total number of observations is 201. The expected counts under H_0 are

$$e_{11} = \frac{133 \cdot 152}{201} \approx 100.58, \quad e_{12} = \frac{133 \cdot 49}{201} \approx 32.42,$$
$$e_{21} = \frac{68 \cdot 152}{201} \approx 51.42, \quad e_{22} = \frac{68 \cdot 49}{201} \approx 16.58,$$

and since these are all greater than 5, χ^2 approximation of the distribution for the statistic is allowed.

The Pearson X^2 test statistic is

$$X^2 = \frac{(99 - 100.58)^2}{100.58} + \frac{(34 - 32.42)^2}{32.42} + \frac{(53 - 51.42)^2}{51.42} + \frac{(15 - 16.58)^2}{16.58}$$
$$\approx 0.30.$$

The number of degrees of freedom is $(2 - 1) \cdot (2 - 1) = 1$, and since $0.30 < 3.84 \approx \chi_1^2(0.05)$, we may not reject H_0 at the 5% level.

At this risk level, we have no evidence for dependence.

An alternative is the G^2 (log likelihood ratio) test:

$$G^2 = 2\left\{99 \log\left(\frac{99}{100.58}\right) + 34 \log\left(\frac{34}{32.42}\right) + 53 \log\left(\frac{53}{51.42}\right) + 15 \log\left(\frac{15}{16.58}\right)\right\}$$
$$\approx 0.30,$$

leading to the same conclusion as above.

(c) Use the data from both 1976 and 1996 to test if gender is independent of county with respect to tendency of dying in motor vehicle accidents, conditional on the year. (3p)

Solution: Test H_0: gender and county are independent conditional on the year vs
H_1: $\neg H_0$. We use the Cochran-Mantel-Haenzel test. To this end, we need the marginal sums $n_{1+1} = 152$, $n_{+11} = 133$, $n_{++1} = 201$, $n_{2+1} = 68$, $n_{+21} = 49$, $n_{1+2} = 59$, $n_{+12} = 67$, $n_{++2} = 87$, $n_{2+2} = 28$ and $n_{+22} = 20$. We then get

$$\begin{aligned}
\mu_{111} &= \frac{n_{1+1}n_{+11}}{n_{++1}} \approx 100.58, \;\; \mu_{112} = \frac{n_{1+2}n_{+12}}{n_{++2}} \approx 45.44 \\
\mathrm{var}(n_{111}) &= \frac{n_{1+1}n_{2+1}n_{+11}n_{+21}}{n_{++1}^2(n_{++1}-1)} \approx 8.34, \\
\mathrm{var}(n_{112}) &= \frac{n_{1+2}n_{2+2}n_{+12}n_{+22}}{n_{++2}^2(n_{++2}-1)} \approx 3.40,
\end{aligned}$$

which yields the CMH statistic

$$\frac{(99 - 100.58 + 46 - 45.44)^2}{8.34 + 3.40} \approx 0.089.$$

Because $0.089 < 3.84 = \chi_1^2(0.05)$, we may not reject H_0 at the 5% level. On this risk level (nor on any other sensible levels), there is no evidence of conditional dependence.

3. Let $P(Y = 1|x) = \pi(x) = F(\alpha+\beta x) = 1-P(Y = 0|x)$. Consider the following suggestions for the function F (in all cases, $-\infty < z < \infty$):

   (*i*) $F(z) = \frac{\exp(2z)}{1+\exp(z)+\exp(2z)}$,

   (*ii*) $F(z) = z^2$,

   (*iii*) $F(z) = (1 + z^{-1})^{-1}$.

   (a) Which (if any) of the suggestions gives a suitable model, and which do not? Why or why not? (3p)

   Solution: Because $\pi(x)$ is a probability, it is needed that $F(x)$ takes values only in the interval $[0, 1]$. This is fulfilled only for the function in (*i*). In fact, this function increases monotonely from 0 to 1 as z ranges from $-\infty$ to $+\infty$, which is a desirable property. The monotone increasing property is seen from the derivative w.r.t z:

$$F'(z) = \frac{2e^{2z}(1 + e^z + e^{2z}) - e^{2z}(e^z + 2e^{2z})}{(1 + e^z + e^{2z})^2} = \frac{2e^{2z} + e^{3z}}{(1 + e^z + e^{2z})^2} > 0.$$

   (b) Take your favourite choice of function F from above. For which x (as a function of α and β) is it true that the function $\pi(x) = 1/2$? (3p)

   Solution: With $z = \alpha + \beta x$, we need to solve $F(z) = 1/2$. This gives the equation

$$2e^{2z} = 1 + e^z + e^{2z},$$

   i.e., with $u = e^z$, $u^2 - u - 1 = 0$, which yields

$$u = \frac{1}{2} \pm \sqrt{\frac{1}{4} + 1} = \frac{1 \pm \sqrt{5}}{2}.$$

   Because $u > 0$, we choose the plus sign and get

$$\alpha + \beta x = z = \log\left(\frac{1+\sqrt{5}}{2}\right),$$

   i.e.

$$x = \frac{1}{\beta}\left\{\log\left(\frac{1+\sqrt{5}}{2}\right) - \alpha\right\}.$$

4. The numbers of employed people (converted to full-time employees) in different teaching categories at Uppsala university in 2011, by gender, are given in the table below.

Categories are in Swedish: 'Professor' means full professor, 'lektor' is a lecturer with a PhD exam, 'adjunkt' is a lecturer without a PhD exam and 'fo-ass' is short for 'forskarassistent', which was a time-limited position (similar to today's 'biträdande lektor') with a high percentage of research time for persons who recently got their PhD.

	men	women
Professor	438	126
Lektor	248	249
Adjunkt	64	104
Fo-ass	108	73

(a) From the table, what is the proportion of lecturers ('lektor') that are women divided by the proportion of professors that are women? Also, calculate the proportion of lecturers that are women divided by the proportion of people in the category 'adjunkt' that are women. (1p)

Solution: For the total number of women, w say, the required proportions are

$$\frac{249/w}{126/w} = \frac{249}{126} \approx 1.976, \quad \frac{249/w}{104/w} = \frac{249}{104} \approx 2.394.$$

(b) Consider the baseline-category model

$$\log\left\{\frac{\pi_j(x)}{\pi_1(x)}\right\} = \alpha_j + \beta_j x,$$

where $j = 2$ for 'lektor', $j = 3$ for 'adjunkt' and $j = 4$ for 'fo-ass'. The baseline category, $j = 1$, corresponds to professor.

Moreover, $\pi_j(x) = P(Y = j|x)$ with $x = 0$ for men and 1 for women.

Such a model was estimated based on the data, where $j = 1, 2, 3, 4$ correspond to 'professor', lektor', 'adjunkt' and 'fo-ass', respectively.

The parameter estimates were $\hat{\alpha}_2 = -0.5688$, $\hat{\alpha}_3 = -1.9234$, $\hat{\alpha}_4 = -1.4001$, $\hat{\beta}_2 = 1.2500$, $\hat{\beta}_3 = 1.7315$, $\hat{\beta}_4 = 0.8543$.

Estimate the proportions in (a) as probability ratios from the model and comment. (2p)

Solution: We have

$$\frac{\pi_j(x)}{\pi_1(x)} = \exp(\alpha_j + \beta_j x),$$

and inserting estimates for $j = 2$ and with $x = 1$, we get

$$\exp(-0.5688 + 1.2500) \approx 1.976.$$

The second probability ratio here is

$$\frac{\pi_2(1)}{\pi_3(1)} = \frac{\pi_2(1)/\pi_1(1)}{\pi_3(1)/\pi_1(1)} = \exp(\alpha_1 + \beta_2 - \alpha_3 - \beta_3),$$

which is estimated by

$$\exp(-0.5688 + 1.2500 + 1.9234 - 1.7315) \approx 2.394.$$

These numbers are very close to (in fact equal to, to the third decimal place) the corresponding numbers in (a). This is no surprise, since the model has so many parameters.

(c) Consider a shift of the numbering so that the baseline category $j = 1$ is 'adjunkt', and we have $j = 2$ for 'lektor', $j = 3$ for 'professor' and $j = 4$ for 'fo-ass'.
What do you think $\hat{\alpha}_3$ would be, and why? (2p)

Solution: Because

$$\log\left\{\frac{\pi_3(x)}{\pi_1(x)}\right\} = -\log\left\{\frac{\pi_1(x)}{\pi_3(x)}\right\}$$

for $x = 0, 1$, then, in case of 'exact' estimation, it is clear that $\hat{\alpha}_3$ from the first model will shift sign in the second model. Now, the estimation is not 'exact', but close to, as we have so good the model fit. Hence, we would expect the estimate 1.9234.

(d) If the teacher categories could be ordered in a natural way, suggest a suitable statistical model that takes such ordering into account. (But you don't need to estimate the model parameters.) Do not forget to impose parameter restrictions, if needed. (2p)

Solution: The natural choice is the ordered logit model:

$$\text{logit}\{P(Y \leq j|x)\} = \alpha_j + \beta x,$$

where $j = 1, 2, 3$, $\alpha_1 \leq \alpha_2 \leq \alpha_3$. The ordering of the α_j, as well as having β constant (not depending on j) is necessary in order to avoid logical inconsistencies with the distribution function.

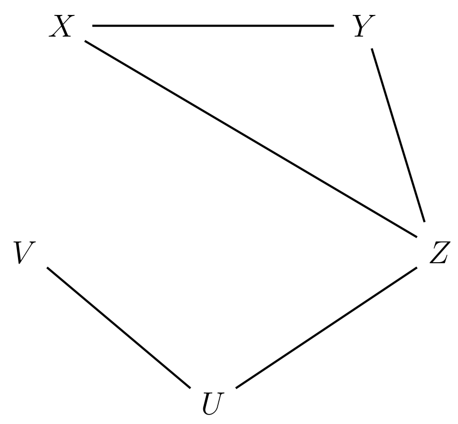

Figure 1: Model, problem 5.

5. Suppose we have variables X, Y, Z, U, V and a loglinear model described by the graph in figure 1.

   (a) Give the name (symbol) of a model that may be described by this graph expressed in a form like (X, Y, YZ) (which is not the model under question here). Also, write down the model equation (in a form like $\log \mu_{ijklm} = \lambda + \lambda_i^X + ...$). (2p)

   Solution: The edges represent interactions and the nodes represent variables. Hence, this is a model with main effects, and pairwise interactions between (X, Y), (Y, Z), (X, Z), (Z, U) and (U, V). There might also be a (X, Y, Z) interaction present. (The models with or without it are represented by the same graph.)
   Hence, the model symbol is either (XYZ, ZU, UV) or (XY, XZ, YZ, ZU, UV).
   The model equation for (XY, XZ, YZ, ZU, UV) is

$$\begin{aligned}\log \mu_{ijklm} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^U + \lambda_m^V \\ &\quad + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{kl}^{ZU} + \lambda_{lm}^{UV},\end{aligned}$$

   where μ_{ijklm} is the expected count. For the model (XYZ, ZU, UV), we add the term λ_{ijk}^{XYZ} on the right-hand side.

   (b) Is X independent of U? Why or why not? (1p)

   Solution: No, because there is a connection between X and U in the graph via (Y and) Z. This can also be seen from the model equation, by the presence of the terms λ_{ik}^{XZ} and λ_{kl}^{ZU} (and λ_{ij}^{XY}, λ_{jk}^{YZ}).

   (c) Is X conditionally independent of U given Z? Why or why not? (1p)

   Solution: Yes, because crossing out Z, we delete the connection between X and V in the graph.

In the model equation, λ_l^Z, λ_{jk}^{YZ} and λ_{kl}^{ZU} are crossed out. The terms λ_{ij}^{XY} and λ_{lm}^{UV} are still there, but there is no more term left including both X and Z, which in turn is the variable that connects to U.

(d) Is X conditionally independent of U given Y? Why or why not? (1p)

Solution: No, crossing out Y does not change the fact that there is a connection from X to U via Z. In the model equation, λ_{ij}^{XY} and λ_{jk}^{YZ} are crossed out, but λ_{ik}^{XZ} remains.

(e) If we sum over the possible values that the variable V can take, which model do we get, and why? (2p)

Solution: We sum over m in the equation for μ_{ijklm} and get

$$\begin{aligned}
\mu_{ijkl+} &= \sum_m \mu_{ijklm} \\
&= \sum_m \exp\left(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^U + \lambda_m^V\right. \\
&\left.+\lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{kl}^{ZU} + \lambda_{lm}^{UV}\right) \\
&= \exp\left(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^U + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{kl}^{ZU}\right) \\
&\cdot \sum_m \exp\left(\lambda_m^V + \lambda_{lm}^{UV}\right) \\
&= \exp\left(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^U + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{kl}^{ZU}\right) \\
&\cdot \exp(\tilde{\lambda}_l^U),
\end{aligned}$$

which means that

$$\log \mu_{ijkl+} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \bar{\lambda}_l^U + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{kl}^{ZU},$$

where $\bar{\lambda}_l^U = \lambda_l^U + \tilde{\lambda}_l^U$. Hence, we have a (XY, XZ, YZ, ZU) model.
If we start with the (XYZ, ZU, UV) model, in this way we end up with the (XYZ, ZU) model.

6. We have a data set from a survey of workers in the US cotton industry, which records whether they were suffering from the lung desease byssinosis, as well as the values of three categorical variables: dustiness of the workplace, smoking status of the worker and the length of employment.

   Let the probability for subject i to suffer of the desease, $P(Y=1|X=i, Z=j, W=k)=\pi_{ijk}$, fulfill the model

$$\text{logit}(\pi_{ijk})=\alpha+\beta_i^X+\beta_j^Z+\beta_k^W,$$

   where $i=1,2,3$ is the degree of dustiness (`dust`), $j=1,2$ correspond to smoker/non smoker (`smoke`), and $k=1,2,3$ correspond to length of employment (`emple`).

   The coefficients with index equal to one were set to zero. The R print from the estimation is given here.

```
> m=glm(yes/n~factor(dust)+factor(smoke)+factor(emple),family=binomial(link=logit),
weights=n);summary(m)

Call:
glm(formula = yes/n ~ factor(dust) + factor(smoke) + factor(emple),
    family = binomial(link = logit), weights = n)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6239  -0.8119  -0.2576   0.3307   1.6605

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.8336     0.1525 -12.026  < 2e-16 ***
factor(dust)2   -2.5493     0.2614  -9.753  < 2e-16 ***
factor(dust)3   -2.7175     0.1898 -14.314  < 2e-16 ***
factor(smoke)2  -0.6210     0.1908  -3.255 0.001133 **
factor(emple)2   0.5060     0.2490   2.032 0.042119 *
factor(emple)3   0.6728     0.1813   3.710 0.000207 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 322.527  on 64  degrees of freedom
Residual deviance:  43.882  on 59  degrees of freedom
  (7 observations deleted due to missingness)
AIC: 162.56

Number of Fisher Scoring iterations: 5
```

   A modified model, including an interaction effect between dustiness and smoking,

$$\text{logit}(\pi_{ijk})=\alpha+\beta_i^X+\beta_j^Z+\beta_k^W+\beta_{ij}^{XZ},$$

   was also estimated.

The coefficients with at least one index equal to one were set to zero. The R print from the estimation is given here.

```
> m=glm(yes/n~factor(dust)+factor(smoke)+factor(emple)+factor(dust):factor(smoke),
family=binomial(link=logit),weights=n);summary(m)

Call:
glm(formula = yes/n ~ factor(dust) + factor(smoke) + factor(emple) +
    factor(dust):factor(smoke), family = binomial(link = logit),
    weights = n)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4325  -0.7602  -0.2795   0.4335   1.3123

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.7573     0.1555 -11.301  < 2e-16 ***
factor(dust)2                 -2.9576     0.3565  -8.295  < 2e-16 ***
factor(dust)3                 -2.8325     0.2230 -12.701  < 2e-16 ***
factor(smoke)2                -0.9573     0.2751  -3.480 0.000502 ***
factor(emple)2                 0.4990     0.2499   1.997 0.045869 *
factor(emple)3                 0.6638     0.1819   3.649 0.000264 ***
factor(dust)2:factor(smoke)2   1.1807     0.5490   2.151 0.031497 *
factor(dust)3:factor(smoke)2   0.4864     0.4338   1.121 0.262201
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 322.527  on 64  degrees of freedom
Residual deviance:  39.031  on 57  degrees of freedom
  (7 observations deleted due to missingness)
AIC: 161.71

Number of Fisher Scoring iterations: 5
```

(a) Estimate the probability π_{221} for the two models, and compare to the empirical proportion of diseased for this category, which is $5/278 \approx 0.0180$. (3p)

Solution: Take the main effects model. We may express the estimated probabilities as

$$\pi_{ijk} = \frac{\exp\left(\alpha + \beta_i^X + \beta_j^Z + \beta_k^W\right)}{1 + \exp\left(\alpha + \beta_i^X + \beta_j^Z + \beta_k^W\right)} = \frac{1}{1 + \exp\left\{-\left(\alpha + \beta_i^X + \beta_j^Z + \beta_k^W\right)\right\}}.$$

Because $\beta_1^W = 0$, we may insert our estimates and get

$$\hat{\pi}_{221} = \frac{1}{1 + \exp\left\{-\left(-1.8336 - 2.5493 - 0.6210\right)\right\}} = \frac{1}{1 + \exp(5.0039)} \approx 0.0067.$$

For the model including the interaction, we similarly get

$$\begin{aligned}
\hat{\pi}_{221} &= \frac{1}{1+\exp\{-(-1.7573-2.9576-0.9573+1.1807)\}} \\
&= \frac{1}{1+\exp(4.4915)} \approx 0.0111.
\end{aligned}$$

We can note that the second (and bigger) model results in an estimate closer to the empirical one, 0.0180.

(b) Test the first model (without interaction) vs the second model (including the interaction) at the 5% level and draw a conclusion. (2p)

Solution: Test H_0: the first model holds vs H_1: the second model holds. We may use that under H_0, the difference of residual deviances is asymptotically χ^2 with the number of degrees of freedom (df) equal to the corresponding difference between the numbers of df for the models. In our case, the number of df is $59 - 57 = 2$ (this we also see because two more parameters are estimated in the second model). We get the observed test statistic

$$43.882 - 39.031 = 4.851 < 5.99 = \chi^2_2(0.05),$$

and so, we may not reject H_0 at the 5% level.

We have no evidence that the second model is better than the first one.

(c) Residual analysis for the model including the interaction is given in figures 2-6 below. Looking at these figures, do you think that the model assumptions are satisfied? (2p)

Solution: From the histogram and the qq plot, we want to check if the standardized residuals are normally distrubuted or not. Looking at the histogram, this seems a bit uncertain, but in the qq plot, the observations are roughly on a straight line. Hence, there does not seem to be any serious evidence against normality.

Figures 4-6 help us to see if the dispersion (variance) of the standardized residuals is different for different values of the explanatory variables. It does not seem so, the vertical spreads appear faily constant when moving horizontally in the graphs.

Conclusively, the residual analysis does not cause any serious doubts that the model is not correct.

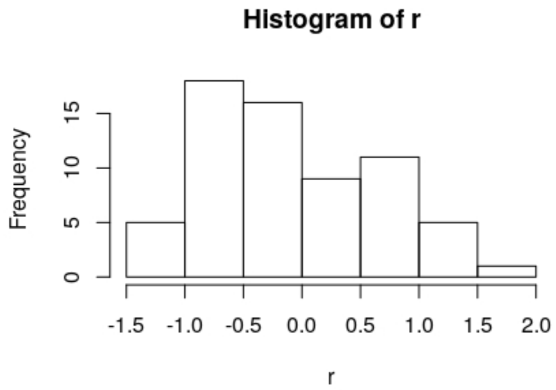

Figure 2: Histogram of standardized residuals.

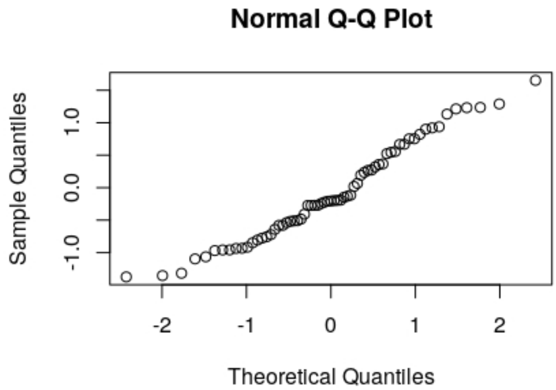

Figure 3: QQ plot of standardized residuals.

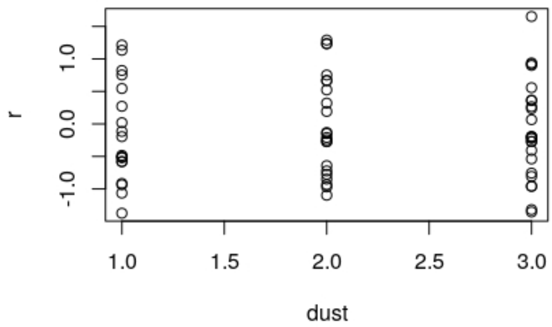

Figure 4: Plot of standardized residuals vs the dust variable.

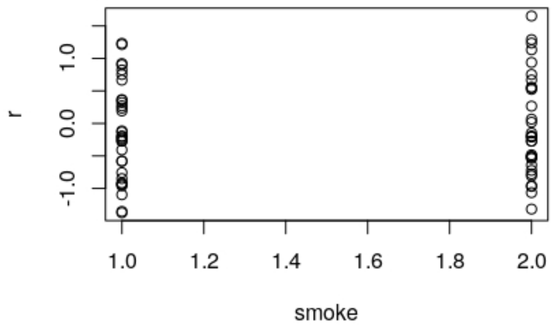

Figure 5: Plot of standardized residuals vs the smoke variable.

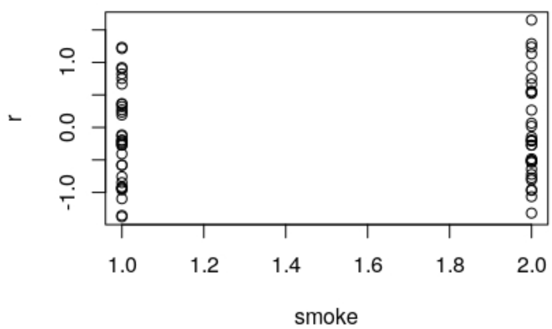

Figure 6: Plot of standardized residuals vs the length of employment variable.

APPENDIX B

Chi-Squared Distribution Values

	Right-Tailed Probability						
df	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	10.22	13.36	15.51	17.53	20.09	21.96	26.12
9	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	23.83	28.41	31.41	34.17	37.57	40.00	45.32
25	29.34	34.38	37.65	40.65	44.31	46.93	52.62
30	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	45.62	51.80	55.76	59.34	63.69	66.77	73.40
50	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	77.58	85.53	90.53	95.02	100.4	104.2	112.3
80	88.13	96.58	101.8	106.6	112.3	116.3	124.8
90	98.65	107.6	113.1	118.1	124.1	128.3	137.2
100	109.1	118.5	124.3	129.6	135.8	140.2	149.5

Categorical Data Analysis, Third Edition. Alan Agresti.