Australian National University

Venue _____

STUDENT NUMBER | U | | | | | | | |

# Research School of Finance, Actuarial Studies and Statistics

## PRACTICE FINAL EXAMINATION

Questions updated from previous exam papers in 2016

## STAT3015/STAT4030/STAT7030 Generalised Linear Models

**Examination/Writing Time Duration:** 180 minutes
**Reading Time:** 15 minutes

**Exam Conditions:**
Central Examination. This examination paper is not available to the ANU Library archives.
Students must return the examination paper at the end of the examination.

**Materials permitted in the exam venue: (No electronic aids are permitted e.g. laptops, phones)**
*Unannotated paper-based dictionary (no approval required),*
*One A4 page with notes on both side, Calculator*

**Materials to be supplied to Students:**
*Scribble Paper*

**Instructions to Students:**

1. This examination paper comprises a total of twenty-five (25) pages and there is a separate handout of R output which has a total of twenty-one (21) pages. During the reading time preceding the exam, please check that both documents have the correct number of pages.
2. All answers are to be written on this exam paper, which is to be handed in at the end of the exam. You may make notes on scribble paper (or on the R handout) during the reading time, but **do NOT write on this exam paper until after the start of the writing time.** If you need additional space, use the rear of the previous page and clearly indicate the part of the question that your answer refers to. The R handout and any scribble paper will collected at the end of the examination and destroyed, they will not be marked.
3. There are a total of six questions, worth a total of 65 marks. The parts of each question are of unequal value, with the marks indicated for each part. **You should attempt to answer all parts of either Q1 or Q1A, and each and every part of the other four questions.** This examination counts towards 65% of your final assessment.
4. **Please write your student number in the space provided at the top of this page.**
5. **Include a clear statement of the formulae you use to answer each question.**
6. Statistical tables (generated using R) are provided on pages 20 and 21 at the end of the handout of R output. Unless otherwise indicated, use a significance level of 5% and log x refers to the natural logarithm of x.

| | Q1 | Q1A | Q2 | Q3 | Q4 | Q5 | Total |
|---|---|---|---|---|---|---|---|
| Pages | 2 to 7 | 8 to 10 | 11 to 15 | 16 to 18 | 19 to 22 | 23 to 25 | |
| **Marks** | **13** | **13** | **13** | **13** | **13** | **13** | **65** |
| **Score** | | | | | | | |

**Question 1** **(13 marks)**

Ryan, B.F., Joiner, B.L. and Ryan, T.A. in their text, *MINITAB Handbook* (PWS-Kent: Boston, 1985, p.206), report an experiment conducted to assess the effect of the use of a supplement and the amount of whey on the quality of pancakes. Eight different batches of pancake mixture were made to different recipes both including and not including the **supplement** and using four levels of **whey** (0%, 10%, 20% and 30%). Three pancakes were baked from each batch and each pancake was given a **rating** by an expert, with higher ratings indicating better quality pancakes.

**(a)** Given the above description of the research question, what are the response and the explanatory variables for this experiment? What are the experimental treatments? How many replications were there for each treatment?

response: ratings
explan. vars supplement (2 levels)
whey (4 levels)
exp. trt: 8 possible combinations
rep : 3 each trt.

**(2 marks)**

**Question 1 continued**

*R* was used to fit an initial linear model `pancake.lm1` to the pancake data and the residual plot for this initial model is shown on page 1 of the `R` output.

**(b)** What problem with the initial model does the above residual plot suggest? What possible solution(s) could you investigate to fix this problem?

heteroscedasticity

fix: ① transform on response or

② weighted linear regression or

③ glm with normal error structure

**(1 mark)**

**Question 1 continued**

After examining the residual plot for the initial model, a monotonically increasing transformation was applied to the pancake rating values and a final linear model was chosen for the transformed data. Pages 2 to 4 of the R output shows some residual plots and summary output for this model (pancake.lm):

**(c)** Do the residuals plots suggest any problems with the underlying assumptions?

Are there any problem(s) shown on the "Residuals vs Fitted" plot on page 2?
If so describe the problem(s):

better than initial model, but still heteroscedasticity.

Are there any problem(s) shown on the "Normal Q-Q" plot on page 2?
If so describe the problem(s):

no obvious violation,
only 1 out of -2,2 range.

Are there any problem(s) shown on the "Cook's distance" plot on page 2?
If so describe the problem(s):

no.
no large values.
no relatively extreme values as well.

What is your overall assessment? (select just ONE of the following options)

- ☐ Residuals are not independent (obvious pattern)
- ☐ Residuals do not have constant variance (heteroscedasticity)
- ☐ Residuals are not normally distributed
- ☐ There are possible outliers and/or influential observations
- ☐ More than one of the above problems
- ☑ No obvious problems

**(2 marks – 0.5 for each section)**

**Question 1 continued**

**(d)** Discuss the following features of the model pancake.lm:

Would the model pancake.lm be best described as an ANOVA model or an ANCOVA model? Which of the explanatory variables have been treated as a factor and which, if any, are covariates?

Whey is fitted as factor

2-way ANOVA with supplement, factor(whey) and one interaction term.

If whey is continuous

$\Rightarrow$ ANCOVA (equivalent to 2 simple linear regression, one for obs. where supplement = 0, obs. where supplement = 1)

Specify the algebraic formula for this model. Clearly indicate how each of the variables have been included in the model, including any constraints applied to the parameters and the assumptions regarding the error distribution.

$$\underbrace{\ln(\text{rating})_{ijk}}_{y} = \beta_0 + \tau_j + \delta_k + \gamma_{jk} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0,1)$$

$y$: trans. rating

$i$th obs.

$j \in \{0, 1\}$ supplement

$k \in \{0\%, 10\%, 20\%, 30\%\}$

$i = 1, 2, 3$ for all $j, k$.

constraint:

$\tau_0 = 0$

$\delta_1 = \delta_{0\%} = 0$

$\gamma_{jk} = 0$ when $j = 0$ or $k = 0\%$ i.e. $k = 1$ (factor)

**(3 marks – 1 for the first part and 2 for the second part)**

**Question 1 continued**

**(e)** Which of the explanatory variables in the model pancake.lm have a significant effect on the response variable? Which statistics in the R output are important in this question?

interaction is the main story. (sig.)

main effects factor(cukey) (sig.)

supplement (sig.)

$\Rightarrow$ F-stats & p-value

↑ large ↓ small.

**(1 mark)**

**(f)** Which hypotheses can be tested using the F-statistic shown at the end of the summary output and what do you conclude as a result?

overall F-test

$H_0$: all $\tau_i = \gamma = \delta = 0$

$H_a$: not all $= 0$.

reject $H_0$ in favour of $H_a$.

at at least one main effect is significant.

**(1 mark)**

**Question 1 continued**

**(g)** In the model pancake.lm, which combination of **supplement** and **whey** appears to produce the pancakes with the highest **rating**? Is this combination significantly better than all other combinations of supplement and whey?

supplement & 30% seems to have the best rating.

95% CI

$$(\bar{y}_1 - \bar{y}_2) \mp t_{res\,df}(0.975)\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$\bar{y}_1 = 1.8333$

$\bar{y}_2 = 1.5333$

$t_{res.df}(0.975) = t_{16}(0.975) \approx 2.1199$

$s^2$ = est. error variance (MSE) = 0.03

$n_1 = n_2 = 3$

$\mp 0.3$

So no CI contains 0.

$\Rightarrow$ all differences significant.

(no need to do Bonferroni actually)

**(3 marks)**

**Question 1A (mixed effects model)** **(13 marks)**

Cook R.D. and Weisberg, S. in their text, *Applied Regression including Computing and Graphics*, (Wiley, New York, 1999, problem 16.8 on page 395), report an experiment conducted to compare the rubber yield of seven varieties of guayule, a desert shrub that contains rubber. The experimental area consisted of 35 plots arranged in a 5 x 7 grid; the rows of the grid forming 5 randomized complete **Block**s, each containing 1 plot of each **Variety**. The **Yield** is the total grams **P1** + **P2** of rubber for two selected plants from each plot.

Pages 5 to 7 of the R output show the results of fitting a linear mixed effects model (rubber.lme) to these data.

**(a)** What are the response and the explanatory variables for this experiment? For each explanatory variable, describe whether it should be treated as a fixed or random effect. How many experimental treatments are there? How many replications are there of each treatment?

Yield ($P_1 + P_2$) : response
explan : fixed factor ~~&~~ variety,
random effect block.
trt : 7 different varieties of guayule,
replica : 5 times for each trt,
within one Block.

**(2 marks)**

**(b)** Specify the algebraic formula for this model. Clearly indicate how each of the variables have been included in the model, including any constraints applied to the parameters and the assumptions regarding the error distribution.

$Y_{ijk} = \mu + \eta_j + \lambda_k + \varepsilon_{ijk}$

Y is Yield, ~~i indicate ith obs.~~
j indicates the level of Variety $= \{1, \dots, 7\}$
k indicate the level of Block $= \{1, 2, 3, 4, 5\}$
& only 1 observation in,
for each combination of j & k.
$2 \times 5 \times 7 = 35$ obs in total,
constraint: variety $\sum \eta_j = 0$

Assumptions: Blocking random effect $\sim N(0, \sigma^2_\lambda)$
which are also independent of the residual variation $\sim N(0, \sigma^2_\varepsilon)$

**(3 marks)**

**Question 1A continued**

**(c)** A residual plot for the model rubber.lme is shown on page 6 of the R output. Does the residual plot suggest there are any problems with the fitted model? What other diagnostics could you examine to decide whether or not there really is a problem with the model?

① gap in the fitted values (strange)
② no obvious problem vertically
(no dependency & no heteroscedasticity)
③ one out of (-2, 2) but not necessarily an outlier
④ need normal-qq, CD to further investigate.

**(2 marks)**

**(d)** Does the rubber **Yield** vary depending on the **Variety** of guayule? Present an appropriate hypothesis test to support your conclusion.

$H_0: \eta_1 = \cdots = \eta_7 = 0$
$H_a$: not all $\eta = 0$
~~blah~~
$\Rightarrow F_{6,24} = 3.6612, p = 0.0101$

So, reject $H_0$,
sig. differences between variety

**(1 mark)**

**Question 1A continued**

**(e)** Rank the seven varieties in order from largest to smallest average **Yield**. Which of the varieties had average **Yield** that differed significantly from the overall average?

$\eta_7 = -(\eta_1 + \cdots + \eta_6) = 0.534286$

$\Rightarrow$ 4, 2, 7, 5, 3, 1, 6

above overall mean | below

2 & 6 sig diff from overall mean

(7 has p-value between 2 & 5)

the only is 4 above average.

$\eta_4 = \bar{N}_4 - \bar{N} = 2.6$, $t_{24} = 2.47$, $p = 0.022$

**(3 marks)**

**(f)** Has blocking been effect in this instance? What proportion of the overall variability is due to variation between **Block**s?

$\hat{\sigma}_\lambda = 0.5886959$ est. sd between Blocks

$\hat{\sigma}_\varepsilon = 2.538402$ est. res. sd within Blocks

$$\frac{\hat{\sigma}_\lambda^2}{\hat{\sigma}_\lambda^2 + \hat{\sigma}_\varepsilon^2} = 5.1\%.$$

small proportion

$\Rightarrow$ still enough to make accounting for difference between Blocks an important part.

**(2 marks)**

**Question 2** **(13 marks)**

A survey of attitudes towards the introduction of a non means-tested age pension resulted in the following data:

| | Age of Respondent | | | | |
|---|---|---|---|---|---|
| | 20 | 30 | 40 | 50 | 60 |
| Sample Size | 12 | 12 | 12 | 12 | 12 |
| Number in favour | 4 | 6 | 9 | 11 | 12 |

Pages 5 to 7 of the R output some the results of fitting a series of logisitic regression models to the above data, with the proportion in favour of a non means-tested pension as the response variable and treating age as a continuous explanatory variable.

**(a)** Use the model pension.glm, summary output for which is shown on page 8 of the R output, to estimate the age at which 50% of respondents are in favour of a non means-tested age pension.

$E(\text{logit}(\text{proptn})) = -3.15770 + 0.11125\,\text{Age}$

when proptn $= 0.5$

$\log\left(\frac{0.5}{1-0.5}\right) = \log 1 = 0$

age $= 28.385$ years.

**(1 mark)**

**(b)** Is age a significant predictor of the response in the model pension.glm? Which of the above statistics in the R output are important in answering this question?

The t-stat 3.59% > theoretical value $t_3(0.975)$

with p-value $0.00323 < 0.05$

reject $H_0$: so age is significant.

OR use drop-in deviance

$19.983 > \chi^2_1(0.975)$

with p-value $= 7.815 \times 10^{-6} < 0.05$

**(2 marks)**

**Question 2 continued**

Residual diagnostic plots for the model pension.glm are given on page 9 of the R output.

**(c)** Do the residuals plots suggest any problems with the underlying assumptions?

Are there any problem(s) shown on the "Residuals vs Fitted" plot on page 9?
If so describe the problem(s):

curvature.
independent violation

Are there any problem(s) shown on the "Normal Q-Q" plot on page 9?
If so describe the problem(s):

No violation
but the sample size is notably small.

Are there any problem(s) shown on the "Cook's distance" plot on page 9?
If so describe the problem(s):

Obs 1 has large CD comparatively,
but real problem is the lack of independence
Not determined (need to fix independence first)

What is your overall assessment? (select just ONE of the following options)

- ☑ Residuals are not independent (obvious pattern)
- ☐ Residuals do not display constant deviance/dispersion
- ☐ Residuals are not normally distributed
- ☐ There are possible outliers and/or influential observations
- ☒ More than one of the above problems
- ☐ No obvious problems

**(2 marks – 0.5 for each section)**

**Question 2 continued**

**(d)** Is there evidence of under-dispersion or over-dispersion with the model pension.glm?

$\frac{\text{res. deviance}}{\text{res. df}} = \frac{1.0292}{3} = 0.34 < 1$

(If asked for significant over/under dispersion)

$0.34 > 1 - 3\sqrt{\frac{2}{3}} = -1.45$

Better approach: scaled residual deviance

$\chi^2_3(0.025) = 0.2158$

$\chi^2_3(0.975) = 9.3484$

1.0292 is within so no problem

**(2 marks)**

**(e)** Find a 95% confidence interval for the coefficient of age in the above model. What does this confidence interval imply about the relationship between the proportion in favour and age?

$0.11125 \pm t_3(0.975) \times SE(0.11125)$

$= 0.11125 \pm 3.1824 \times 0.0309298$

$= (0.01282, 0.20968)$

age ↑ by 1, logit(proption) increase by c $> 0$.

positively correlated.

**(2 marks)**

**Question 2 continued**

**(f)** One suggestion for improving the model pension.glm is to try including a quadratic term in age in the model. This results in the model pension.glm1, summary output for which is shown on page 10 of the R output. Does this output suggest that this model is a significant improvement on the earlier model?

① t-stats $0.802 < t_2(0.975) = 4.3027$

② drop-in deviance

0.75209 ◎ p-value $> 0.05$

not significant addition

pattern still exists

prob by need some other solutions:

transformation on explan. vars,

link function

etc.

possible reason:

small # of samples.

**(2 marks)**

**Question 2 continued**

Residual diagnostic plots for the model `pension.glm1` are given on page 11 of the `R` output.

**(g)** Do the residuals plots suggest any problems with the underlying assumptions?

Are there any problem(s) shown on the "Residuals vs Fitted" plot on page 11?
If so describe the problem(s):

pattern exists.

Are there any problem(s) shown on the "Normal Q-Q" plot on page 11?
If so describe the problem(s):

not normal

Are there any problem(s) shown on the "Cook's distance" plot on page 11?
If so describe the problem(s):

Obs. relatively higher CD

What is your overall assessment? (select just ONE of the following options)

- ☐ Residuals are not independent (obvious pattern)
- ☐ Residuals do not display constant deviance/dispersion
- ☐ Residuals are not normally distributed
- ☐ There are possible outliers and/or influential observations
- ☑ More than one of the above problems
- ☐ No obvious problems

**(2 marks – 0.5 for each section)**

**Question 3** **(14 marks)**

Myers, R.H., Mongomery, D.C. and Vining, G.G. (2002) *Generalized Linear Models with Applications in Engineering and the Sciences*, Wiley: New York, which was one of the texts in the recommended reading list for this course, includes as an exercise (4.10, p.154) some data collected for a student project. The student was examining the impact of popping temperature, amount of oil and the popping time on counts of the number of inedible kernels of popcorn.

The data for this experiment is shown on pages 12 of the R output, followed by summary output and graphs for a series of models.

**(a)** There are some obvious problems with the residual plot for the model popcorn.glm shown on page 13 of the R output. Identify the data points that are causing these problems. What can be done to further investigate these problems?

Suspicious 2 ~~influential~~ pts { obs #7 with 4.719/1.304 ~~Cook's D~~
obs #2 with sd res 4.361/5.572.

further investigate:
how influential these pts are in determing the model.
examining the deletion residuals,
~~(calculating the CD for all~~
If ~~CD~~ leverage $> \frac{2p}{n} = \frac{2\times 7}{15}$

if ..... then.

Exfurb, 4/15 ~~#~~ of obs. are out of (-2,2) range, maybe we change model, use stronger link function.

**(3 marks)**

**Question 3 continued**

**(b)** Summary output for the model popcorn.glm is shown on pages 13 and 14 of the R output. Apart from the 0 degrees of freedom associated with the three-way interaction term suggesting that there is insufficient data to sensibly fit such a term, what does the ANOVA table and the table of coefficients suggest are the significant terms in this initial model?

3-way interaction ⇒ no ~~insufficient~~ data (not fit in)

temp:time sig

oil:time sig.

temp:oil insig.

**(3 marks)**

**(c)** Is there evidence of significant over-dispersion? What does this suggest about the adequacy of this model?

$\frac{\text{res. deviance}}{\text{res df}} = \frac{36.715}{8} = 4.58937 >$

$1 + 3\sqrt{\frac{2}{8}} = 2.5$

over-dispersion ✓

Better, formal hypothesis test

$H_0: \phi = 1$, $H_A: \phi \neq 1$

$(\chi^2_8(0.025) = 2.1797, \chi^2_8(0.975) = 17.5345)$

Scaled residual deviance (36.715) not in the interval.

**(3 marks)**

**Question 3 continued**

**(d)** Summary output for another model popcorn.glm1 is shown at the bottom of page 14 of the R output. This is an attempt to refine the initial model. Is this model new an adequate one for the data? Would you suggest any further refinements to the model?

new test for over-dispersion

($\chi^2_9(0.025) = 2.7004$, $\chi^2_9(0.975) = 19.0228$)

39.9847 out of this interval

remaining interactions are still significant.

& these inter involve all 3 main effects, all 3 should be included in the model.

further refinement:

deleting is not a good idea.

we need to use other approaches like transformation etc.

**(4 marks)**

**Question 4** **(13 marks)**

A paper by R.E. Chapman, titled "Degradation study of a Photographic Develop to Determine Shelf Life" (*Quality Engineering* (10), 1997-98, pp.137-140) presents the results of an experiment designed to study the relationship between the shelf life and the density of a photographic developer. Density is considered a good indicator of overall developer performance.

In the experiment, the shelf life of 21 batches of the photographic developer were accelerated by subjecting each batch to a set number of hours at a high temperature and the resulting maximum density of each batch was recorded:

| | | | | | | | | temperature |
|---|---|---|---|---|---|---|---|---|
| life time (hours) | 72 | 144 | 216 | 288 | 360 | 432 | 504 | 72°C |
| maximum density | 3.55 | 3.27 | 2.89 | 2.55 | 2.34 | 2.14 | 1.77 | |
| life time (hours) | 48 | 96 | 144 | 192 | 240 | 288 | 336 | 82°C |
| maximum density | 3.52 | 3.35 | 2.50 | 2.10 | 1.90 | 1.47 | 1.19 | |
| life time (hours) | 24 | 48 | 72 | 96 | 120 | 144 | 168 | 92°C |
| maximum density | 3.46 | 2.91 | 2.27 | 1.49 | 1.20 | 1.04 | 0.65 | |

When the experiment was designed, it was suggested that the life times might follow an exponential distribution, as might the resulting maximum densities. Based on this suggestion R was used to fit the model photo.glm. Summary output for this model is shown on page 15 of the R output; page 16 shows a plot of the data with the fitted values from the model superimposed; and page 17 shows a residual plot for the model.

**(a)** For a Gamma GLM, there are two simple ways to estimate the dispersion factor. Use the summary output for the model photo.glm to find these two estimates and compare them. Is there any evidence of under or over-dispersion?

① Coefficient of Variation (CV)

$0.00379838 < 1$.

② $\frac{\text{res deviance}}{\text{res df}} = \frac{0.056144}{15} = 0.003742933$

③ $H_0: \phi = \phi_{CV}$, $H_A: \phi \neq \phi_{CV}$

$\chi^2_{15}(0.025) = 6.2621$, $\chi^2_{15}(0.975) = 27.4884$

scaled deviance $\frac{0.056144}{0.00379838} = 14.78$

is inside.

NO DISPERSION PROBLEM

**(2 marks)**

**Question 4 continued**

**(b)** How does fitting a Gamma GLM help to implement the suggestion that the life times and the maximum densities follow an exponential distribution? Do the estimated dispersion factors from part (a) support the suggestion that the error distribution is exponential?

An exponential distribution is a special case of Gamma$(1, \beta)$.

The dispersion parameter for GLM with an error distribution ~ Gamma family should be $\frac{1}{\alpha} = 1$

But in part (a) 1

$\frac{1}{0.0037984} = 263$

$\frac{1}{0.0037429} = 267$

$> 1 \Rightarrow$ not exponential.

**(2 marks)**

**(c)** Does the analysis in parts (a) and (b) and the plots shown on pages 16 and 17 of the R output suggest that the model is an appropriate one for the data?

① first plot generally no problem

~~③~~ good fit

② no ~~di~~ violation of assumptions

but just note than error structure is not exponential.

**(2 marks)**

**(d)** Based on the table of coefficients for the model, which terms in the model are significant predictors of the maximum density (and therefore of the overall performance of the developer)?

$t_{15}(0.975) = 2.131$

So lifetime ⇒ sig.

92°C & 72°C sig

but 82°C insig.

Sig. interactions.

~~no need to do~~

✓

chosen model should allow for different slopes & different intercepts for the different levels of temperature.

**(4 marks)**

**Question 4 continued**

**(e)** Based on the analysis of deviance table, which of the explanatory variables are significant predictors of the maximum density? Are these conclusions consistent with your answers to part (d)?

$\chi^2_1(0.95) = 3.845$

lifetime: $\frac{0.45327}{0.00379838} = 119.331$

sig.

$\chi^2_2(0.95) = 5.9915$

temp: $\frac{1.81813}{0.00379838} = 478.653$

sig. for some differences (not all!) consistent with part (d).

interaction:

$\frac{1.45322}{0.00379838} = 382.594$

sig.

**(3 marks)**

**Question 5** **(13 marks)**

568 male and female residents in Australian metropolitan and rural areas were surveyed about their preference for locally manufactured or imported motor vehicles, with the following results:

| Car Preference | Males | | Females | |
|---|---|---|---|---|
| | City Residents | Country Residents | City Residents | Country Residents |
| Imported | 68 | 12 | 16 | 24 |
| Local | 168 | 32 | 84 | 164 |

Pages 18 and 19 of the R output show some analyses of these data.

**(a)** Use the observed and expected values shown on page 18 of the R output to find, for males and females separately, the Pearson $\chi^2$ statistics for tests of independence between car preference and residence. Do these statistics suggest that there is a significant association between car preference and residence for either males or females?

Pearson $\chi^2$ stats.

$$\sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Plug in by gender

male: 0.043 <

female: 0.574 <

df: $(2-1)\times(2-1) = 1$

$\chi^2_1(0.95) = 3.845$

so cannot reject $H_0$.

$\Rightarrow$ ~~si~~ independent

**(3 marks)**

$\Rightarrow$ no sig association between either car preference & res for male or female

| | city | country |
|---|---|---|
| import | 84 | 36 |
| local | 252 | 196 |

**Question 5 continued**

**(b)** The top of page 19 of the R output shows the data table collapsed so that the effects of sex are ignored. Find the expected values for the collapsed data assuming that car preference and residence are independent and then find the deviance statistic for a test of independence (also called the likelihood ratio $\chi^2$ statistic). Is there a significant association between car preference and residence ignoring the effects of sex?

| | City | Country | Total |
|---|---|---|---|
| Imported | $120\times\frac{336}{568}$ | $120\times\frac{232}{568}$ | 120 |
| Local | $448\times\frac{336}{568}$ | $448\times\frac{232}{568}$ | 448 |
| Total | 336 | 232 | 568 |

Likelihood Ratio $\chi^2$ stats:

$$2\sum_{i}\sum_{j} O_{ij}\ln\left(\frac{O_{ij}}{E_{ij}}\right)$$

$$=2\left(84\ln\left(\frac{84}{70.986}\right)+36\ln\left(\frac{36}{49.014}\right)+252\ln\left(\frac{252}{265.014}\right)+196\ln\left(\frac{196}{182.986}\right)\right)$$

$$=7.62>3.8415$$ reject $H_0$

Here is association!

**(4 marks)**

**Question 5 continued**

**(c)** Does the analysis on page 19 of the R output suggest that there are significant associations between sex and either car preference or residence? If so, what do these associations suggest about the relationships between sex and the other variables?

yes sex by car preference

yes sex by residence

**(4 marks)**

**(d)** The results of part (a) and (b) appear to contradict each other, though the results of part (c) may help to explain this apparent contradiction. Is this an example of Simpson's paradox and if so, how might the "paradox" be explained in this instance?

Yes. as at one level of aggregation, when we consider 2 sexes separately, no association between car & residence. But if aggregate, ∃ association.

CG: Both residence & car associated to sex. So, ∃ is no association between even ~~sex~~ car & residence (accounting for sex), when sex is ignored, association becomes apparent.

**(2 marks)**

**END OF EXAMINATION**