THE UNIVERSITY OF AUCKLAND

SECOND SEMESTER, 2006
Campus: City

STATISTICS

Advanced Statistical Modeling/Special Topic in Regression
(Time allowed: THREE hours)

INSTRUCTIONS

SECTION A: Multiple Choice (60 marks)

• Answer ALL 25 questions on the answer sheet provided.
• All questions have a single correct answer and carry the same mark value.
• If you give more than one answer to any question you will receive zero marks for that question.
• Incorrect answers are not penalized.

SECTION B (40 marks)

• Answer 2 out of 3 questions. Each is worth 20 marks.

Total for both parts: 100 marks

CONTINUED
1. Suppose we have a continuous response $Y$ and several explanatory variables, all of which are factors. Which of the following plots is the most useful for quickly assessing which factors have an effect on the response, and how the effect depends on the factor levels?

(1) A plot produced by the `plot.design` function.
(2) A plot produced by the `plot` function.
(3) A plot produced by the Trellis function `bwplot`.
(4) A plot produced by the `boxcoxplot` function.
(5) A plot produced by the Trellis function `xyplot`.

2. The data for this question come from a study concerning soldering electrical components onto printed circuit boards. The aim of the study was to identify factors causing “solder skips” or gaps in the soldering. Each board comprises three panels, and counts of solder skips were made on each panel of each board. Some additional data on each panel were also collected.

The resulting data were assembled into a data frame `solder.df` with the following variables:

**Opening**: The amount of clearance around the mounting pad, (small, medium or large);

**Mask**: Type and thickness of the solder mask (A3, A6, B3, B6)

**Panel**: Each board was divided into 3 panels, this variable refers to the panel (1,2,3).

**skips**: The number of solder skips in the panel.

A Trellis plot of these data is shown in Figure 1. Which of the following R commands produced this graph?

(1) `bw(sqrt(skips)~Mask|Opening*Panel, data=solder.df)`.
(2) `dotplot(sqrt(skips)~Mask|Opening*Panel, data=solder.df)`.
(3) `dotplot(sqrt(skips)~Panel|Opening*Mask, data=solder.df)`.
(4) `dotplot(Opening~sqrt(skips)|Panel*Mask, data=solder.df)`.
(5) `dotplot(sqrt(skips)~Opening|Opening*Mask, data=solder.df)`.

CONTINUED
3. A Trellis display of the data in Question 2 is shown in Figure 1. Which of the following is FALSE?

(1) For small amounts of clearance, B3 has more skips than A3.

(2) B3 tends to have more skips than A3.

(3) When there is a small amount of clearance around the mounting pad there are more skips.

(4) B6 has more skips than A3.

(5) There are more skips on Panel 3 than Panel 1.
4. In an ecological study, mussels were harvested at random and the following variables measured:

- **H**: The height of the mussel shell,
- **L**: The length of the mussel shell,
- **S**: The weight of the mussel shell,
- **W**: The width of the mussel shell,
- **M**: The weight of the edible mussel meat.

Suppose we want to fit a model that explains the meat weight $M$ in terms of the length and width of the shell. A coplot of these three variables is shown in Figure 2.

Which of the following is **FALSE**?

1. From the coplot, it appears that the data are not planar.
2. From the coplot, it appears that outliers will not be a serious problem in fitting this regression.
3. From the coplot, it appears that the data are not normally distributed.
4. From the coplot, it appears that the meat weight increases with the width and length of the shell.
5. From the coplot, it appears that the equal variance assumption is not satisfied.

Figure 2: Coplot for Question 4.
5. Suppose we fit a regression model using all the variables in the mussel data set, using the meat weight as the response. The diagnostic plots shown in Figure 3 were obtained. Of the following actions, which would you take first?

(1) Do nothing, the regression looks OK.
(2) You should transform the explanatory variables.
(3) Points 16, 21 and 39 should be removed.
(4) You should do a Weisberg-Bingham test to check the normality.
(5) You should transform the response, since the equal-variance assumption seems violated.

Figure 3: Diagnostic plots for Question 5.
6. In a regression analysis with a continuous response, which of the following is **FALSE**?

(1) If all the regression assumptions are true, the ratio of the coefficient to its standard error has a normal distribution.

(2) The standard error of an estimated coefficient measures the variability of the estimated regression coefficient.

(3) $R^2$ is the square of the correlation between the observations and the fitted values.

(4) If all the regression assumptions are true, the estimated regression coefficients are normally distributed.

(5) A regression coefficient measures the increase in the mean response associated with a unit increase in the covariate, for fixed values of the other covariates.

7. In the course we discussed several types of influence diagnostics. Which of the following statements about these diagnostics is **FALSE**?

(1) The COVRATIO measures the change in the standard errors when a point is deleted.

(2) The hat matrix diagonals measure how much of an outlier a point is.

(3) The hat matrix diagonals measure how much leverage a point has.

(4) Cook’s distances measure the change in the regression coefficients when a point is deleted.

(5) The average hat matrix diagonal is $p/n$, where $n$ is the sample size and $p$ is the number of regression coefficients.

8. Below we show the result of running an “all possible regressions” on the mussel data. Which model should we fit?

<table>
<thead>
<tr>
<th>rssp</th>
<th>sigma2</th>
<th>adjRsq</th>
<th>Cp</th>
<th>AIC</th>
<th>BIC</th>
<th>CV</th>
<th>H</th>
<th>L</th>
<th>S</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1461.668</td>
<td>18.271</td>
<td>0.867</td>
<td>11.409</td>
<td>93.409</td>
<td>98.222</td>
<td>157.236</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1285.487</td>
<td>16.272</td>
<td>0.882</td>
<td>2.632</td>
<td>84.632</td>
<td>91.852</td>
<td>141.434</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1263.498</td>
<td>16.199</td>
<td>0.882</td>
<td>3.287</td>
<td>85.287</td>
<td>94.914</td>
<td>143.917</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1258.804</td>
<td>16.348</td>
<td>0.881</td>
<td>5.000</td>
<td>87.000</td>
<td>99.034</td>
<td>148.381</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

(1) $M \sim W+L$.

(2) $M \sim H+L+S+W$.

(3) $M \sim H+S+W$.

(4) $M \sim H+S$.

(5) None of these models fit very well.
9. In a US study, researchers for *Consumer Reports* analysed the sodium content of 54 hot dogs. There were three types of hot dog: Beef, “Meat” (a mixture of pork and beef, with up to 15% poultry), and Poultry. The researchers also measured the number of calories in each hot dog. There are three variables in the data set:

**Type:** The type of hot dog (Beef, Meat, Poultry)

**Calories:** The number of calories in each hot dog,

**Sodium:** The amount of sodium in each hot dog, in mg.

If we ignore the calories, and just fit the variable Type, we get the output shown on the next page:

Call:
```r
lm(formula = Sodium ~ Type, data = hotdog.df)
```

Coefficients:
```
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 401.15      21.13   18.98  <2e-16 ***
TypeMeat    17.38       31.17    0.56   0.5795 
TypePoultry 57.85       31.17    1.86   0.0692 .
```

Residual standard error: 94.48 on 51 degrees of freedom
Multiple R-Squared: 0.06517, Adjusted R-squared: 0.02851
F-statistic: 1.778 on 2 and 51 DF, p-value: 0.1793

Which of the following is **FALSE**?

(1) The average poultry hot dog contains about 58mg of sodium.
(2) On average, meat hot dogs have more sodium than beef.
(3) The $R^2$ is poor because the sodium content of hot dogs is highly variable.
(4) There is weak evidence that there is a difference in the mean sodium content of poultry and beef.
(5) A 95% confidence interval for the average sodium content of beef hot dogs is $401mg \pm 42mg$ sodium.

10. If we reanalyse the data including the variable Calories, we get

Call: `lm(formula = Sodium ~ Type + Calories, data = hotdog.df)`

Coefficients:
```
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -113.28      53.30  -2.13   0.0385 *  
TypeMeat    11.29       18.28   0.62   0.5395  
TypePoultry 182.76      22.19   8.24 7.15e-11 ***
Calories     3.28       0.33    9.92 2.09e-13 ***
```

Residual standard error: 55.38 on 50 degrees of freedom
Multiple R-Squared: 0.6851, Adjusted R-squared: 0.6663
F-statistic: 36.27 on 3 and 50 DF, p-value: 1.356e-12

CONTINUED
Which of the following is FALSE?

(1) There is a strong relationship between calories and sodium content.
(2) The residual standard error is much smaller than in Question 9 because the variable Calories is explaining some of the variation in the sodium content.
(3) There is strong evidence that, for a given amount of calories, the types of hot dog differ in their sodium content.
(4) There is no evidence that the sodium content of Meat and Beef hot dogs differ.
(5) The regression line for meat hot dogs has intercept 11.29.

11. Further analysis of the hot-dog data yielded the following output:

```r
> model2 = lm(Sodium~Type*Calories, data=hotdog.df)
> anova(model2)
Analysis of Variance Table
Response: Sodium
Df Sum Sq Mean Sq  F value Pr(>F)
Type 2 31739 15869 5.3294 0.008124 **
Calories 1 301917 301917 101.3927 2.019e-13 ***
Type:Calories 2 10402 5201 1.7466 0.185267
Residuals 48 142930 2978
```

Which of the following is FALSE?

(1) The p-value 0.008124 is comparing the “parallel lines” model to the “all lines the same model”.
(2) The p-value 0.185267 is testing the hypothesis that the regression lines for the three types of hot dog are parallel.
(3) The p-value 2.019e-13 is testing the hypothesis that, given type is in the model, there is no relationship between Calories and Sodium.
(4) The mean square 2978 is estimating the variance of hot dogs of a fixed type and level of calories.
(5) The p-value 2.019e-13 indicates that the variable Calories should be included in the model.

12. In a chemical process, the yield of the process is thought to depend on both the temperature and pressure at which the reaction takes place. An experiment was set up to study the relationship between these variables, here called yield, temp and pressure. There were 3 levels of temperature: Low, Medium and High, and three levels of pressure 200 psi, 215 psi and 230 psi. (Note in the output below, “Low” is the baseline level for temperature.) Each of the possible 9 treatment combinations was used twice in the experiment, generating 18 observations.

The model yield ~ temp*pressure was fitted to the data, and a fragment of the resulting output is shown on the next page.
Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>90.300</td>
</tr>
<tr>
<td>tempMedium</td>
<td>-0.100</td>
</tr>
<tr>
<td>tempHigh</td>
<td>0.300</td>
</tr>
<tr>
<td>pressure215.psi</td>
<td>0.350</td>
</tr>
<tr>
<td>pressure230psi</td>
<td>0.000</td>
</tr>
<tr>
<td>tempMedium:pressure215.psi</td>
<td>0.000</td>
</tr>
<tr>
<td>tempHigh:pressure215.psi</td>
<td>-0.100</td>
</tr>
<tr>
<td>tempMedium:pressure230psi</td>
<td>-0.200</td>
</tr>
<tr>
<td>tempHigh:pressure230psi</td>
<td>-0.350</td>
</tr>
</tbody>
</table>

Which of the following is TRUE?

(1) The mean yield at high temperature and pressure 230 psi is estimated as 90.100.
(2) The mean yield at medium temperature and pressure 215 psi is estimated as 90.550.
(3) The mean yield at medium temperature is estimated as 90.200.
(4) The mean yield at high temperature and pressure 200 psi is estimated as 0.300.
(5) The mean yield at pressure 230 psi is estimated as 0.000.

13. In the experiment described in Question 12, the following ANOVA table was obtained:

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>temp</td>
<td>2</td>
<td>0.30111</td>
<td>0.15056</td>
<td>8.4687</td>
<td>0.0085392**</td>
</tr>
<tr>
<td>pressure</td>
<td>2</td>
<td>0.76778</td>
<td>0.38389</td>
<td>21.5937</td>
<td>0.0003673***</td>
</tr>
<tr>
<td>temp:pressure</td>
<td>4</td>
<td>0.06889</td>
<td>0.01722</td>
<td>0.9687</td>
<td>0.4700058</td>
</tr>
<tr>
<td>Residuals</td>
<td>9</td>
<td>0.16000</td>
<td>0.01778</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Which of the following is FALSE?

(1) There is no effect of changing pressure from 200 psi to 230 psi.
(2) There is strong evidence of interaction in these data.
(3) The effect of changing temperature from medium to low is to increase the yield by 0.100.
(4) The effect of changing temperature from medium to high is to increase the yield by 0.400.
(5) The effect of changing pressure from 215 psi to 230 psi is to decrease the yield by 0.350.
14. In a logistic regression model, the explanatory variable $X$ had a regression coefficient of 0.5. Which is the correct interpretation?

(1) If the other variables are held constant, a unit increase in $X$ should increase the probability of a “success” by 0.5.
(2) If the other variables are held constant, a unit increase in $X$ should increase the log-odds of a “success” by 0.5.
(3) If the other variables are held constant, a unit increase in $X$ should increase the log of the mean by 0.5.
(4) If the other variables are held constant, a unit increase in $X$ should increase the mean by 0.5.
(5) If the other variables are held constant, a unit increase in $X$ should increase the odds of a “success” by 0.5.

15. In a logistic regression, with response $Y = 0$ (failure) and $Y = 1$ (success) and two continuous explanatory variables $X$ and $W$, the following coefficients were obtained:

<table>
<thead>
<tr>
<th>Estimate</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.2</td>
</tr>
<tr>
<td>$X$</td>
<td>0.3</td>
</tr>
<tr>
<td>$W$</td>
<td>-0.5</td>
</tr>
</tbody>
</table>

Which of the following is TRUE?

When $X = 2$ and $W = 5.3$,

(1) To 4 decimal places, the estimated log-odds of a success are 0.0388.
(2) To 4 decimal places, the estimated probability of a success is 0.0373.
(3) The estimated log-odds of a failure are -3.25.
(4) To 4 decimal places, the estimated mean number of successes is 0.0388.
(5) The estimated odds of a success are -3.25.

16. In a survey to study the factors that affect psychotropic drug consumption, the following variables were measured:

**sex:** Gender (0=male, 1=female);

**agegroup:** Age group (with levels 16-29, 30-44, 45-64, 65-74, > 74);

**GHQ:** Result of General Health Questionnaire (0=Poor health, 1=Good health);

**taking:** Number (out of total) taking psychotropic drugs;

**total:** Total number having the covariate pattern.

The data are contained in the data frame `psycho.df` in grouped form. A model was fitted and the following output obtained:

CONTINUED
Call:
glm(formula = cbind(taking, total - taking) ~ sex + agegroup + GHQ, family = binomial, data = psycho.df)
Deviance Residuals:
          Min       1Q     Median       3Q      Max
-1.69383 -0.38364  0.03628  0.40832  1.45093
Coefficients:  
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.00536   0.15066 -26.586  < 2e-16 ***
   sex          0.62780   0.09554   6.571  4.98e-11 ***
agegroup30-44  0.76807   0.16106   4.769  1.85e-06 ***
agegroup45-64  1.31152   0.14771   8.879  < 2e-16 ***
agegroup65-74  1.73636   0.16225  10.702  < 2e-16 ***
agegroup>74    1.70073   0.19004   8.949  < 2e-16 ***
     GHQ        1.41364   0.09047  15.626  < 2e-16 ***
Null deviance: ****** on 19 degrees of freedom
Residual deviance: ****** on 13 degrees of freedom
AIC: 125.59
> 1-pchisq(ResidualDeviance,13)
[1] 0.3524482
> 1-pchisq(NullDeviance,19)
[1] 0

In this output, the null and residual deviances have been replaced by ******, but the corresponding p-values are 0 and 0.3524482 respectively.

Which of the following is FALSE?

1. The p-value of 0 corresponding to the null deviance indicates that at least one of the factors sex, agegroup and GHQ is having an effect on the response.
2. Other things being equal, people with good health have a higher probability of taking the drug.
3. Other things being equal, females are less likely to take the drug than males.
4. Other things being equal, the probability a person will be taking these drugs increases with age (up to age 74).
5. The residual deviance (p-value 0.3524482) indicates the model fits well.

17. The following R-code refers to the data in Question 16.

> r = psycho.df$taking
> n = psycho.df$total
> sum(r*log(r/n)+(n-r)*log(1-r/n))
[1] -1760.499
\begin{verbatim}
> pi.hat = predict(model2, type="response")
> sum(r*log(pi.hat)+(n-r)*log(1-pi.hat))
[1] -1767.653

> pi.null = sum(r)/sum(n)
> sum(r*log(pi.null)+(n-r)*log(1-pi.hat))
[1] -1965.256

Which of the following is \textbf{TRUE}? To 3 decimal places:

1. The null deviance is 204.757.
2. The null deviance is 14.308.
3. The residual deviance is 14.308.
4. The residual deviance is 7.154.
5. The maximum value of the saturated model log-likelihood is -1965.256.

18. For the psychotropic drug example, suppose we want to predict if a 17 year-old male with a high GHS score will be a taker of psychotropic drugs. We get the output

\begin{verbatim}
> predict(model2, data.frame(sex=0, agegroup = "16-29",GHQ=1), se=T)
$fit
[1] -2.591728
$se.fit
[1] 0.1489626
$residual.scale
[1] 1
\end{verbatim}

Which of the following is \textbf{TRUE}? To 3 decimal places:

1. For this individual, a confidence interval for the estimated log-odds is -2.592 ± 0.292.
2. For this individual, a confidence interval for the estimated odds is (0.053, 0.091).
3. For this individual, a confidence interval for the estimated probability is (0.056, 0.100).
4. For this individual, a confidence interval for the estimated log-odds is (0.056, 0.100).
5. For this individual, a confidence interval for the estimated odds is 2.592 ± 0.292.
\end{verbatim}
19. Which of the following is FALSE?

(1) Over-dispersion causes us to underestimate the standard errors.
(2) In logistic regression, we can allow for over-dispersion by using the family=quasibinomial argument.
(3) Over-dispersion causes us to overestimate the significance of coefficients.
(4) In logistic regression, over-dispersion means that the variance of the sample proportion is more than the mean of the sample proportion.
(5) In a logistic regression with grouped data, over-dispersion can be caused by correlation between the responses of individuals having the same covariate pattern.

20. In a Poisson regression, which is the correct interpretation?

(1) The scale factor is always more than one.
(2) The mean is a linear function of the covariates.
(3) The regression coefficient measures the increase in the mean response for a unit increase in the covariate.
(4) The log of the mean is a linear function of the covariates.
(5) The residual deviance cannot be used to measure goodness of fit.

21. The data in Table 1 arose from a 1988 survey in the US. Respondents were asked “Should the Federal Government pay the medical costs of AIDS patients?” The respondents were classified by (a) whether they agreed with this proposition, and (b) their gender.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Aids proposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agree</td>
<td>Disagree</td>
</tr>
<tr>
<td>Male</td>
<td>82</td>
</tr>
<tr>
<td>Female</td>
<td>125</td>
</tr>
</tbody>
</table>

Some R output is shown below:

```r
> aids.df<-data.frame(expand.grid(gender=c("Male", "Female"), AIDS =c("Agree", "Disagree")), count=c(82, 125, 185, 229))
> aids glm<-glm(count ~ gender*AIDS, family=poisson, data=aids.df)
> summary(aids glm)
Coefficients:

                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         4.4067     0.1104  39.905  < 2e-16 ***
genderFemale        0.4216     0.1421   2.967   0.00301 **
AIDSDisagree       0.8136     0.1327   6.133  8.63e-10 ***
genderFemale:AIDSDisagree -0.2082    0.1731  -1.203   0.22903
```

CONTINUED
Which of the following is **FALSE**?

1. A 95% confidence interval for the odds ratio for this table is (0.578, 1.140).
2. The residual deviance for this model is 0.
3. There is no evidence of a relationship between gender and the response to the survey question.
4. The log-odds ratio for this table is -0.2082.
5. If we swap the rows of the table the log-odds stay the same.

22. In addition to the AIDS question in Question 21, a second question was asked: “Do you think the Government should promote safe sex practices?” The results are shown in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Educ</strong></td>
<td>Agree</td>
<td>Disagree</td>
</tr>
<tr>
<td><strong>AIDS</strong></td>
<td>Agree</td>
<td>Disagree</td>
</tr>
<tr>
<td><strong>Govt Promote</strong></td>
<td>Agree</td>
<td>Disagree</td>
</tr>
<tr>
<td>Agree</td>
<td>76</td>
<td>114</td>
</tr>
<tr>
<td>Disagree</td>
<td>160</td>
<td>181</td>
</tr>
<tr>
<td>Agree</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>Disagree</td>
<td>25</td>
<td>48</td>
</tr>
</tbody>
</table>

An analysis of these data was performed, resulting in the following (edited) R output:

```r
> aids2.df<-data.frame(expand.grid( Educ=c("Agree", "Disagree"),
                                    gender=c("Male", "Female"), AIDS =c("Agree", "Disagree")),
                                    count=c(76,6,114,11,160,25,181,48))
> aids2.glm<-glm(count~Educ*AIDS*gender, family=poisson, data=aids2.df)
> anova(aids2.glm, test="Chisq")
Analysis of Deviance Table
Model: poisson, link: log
Response: count

| Df | Deviance | Resid. Df | Dev | P(>|Chi|) |
|----|----------|-----------|-----|---------|
| NULL | 7  | 445.82 |
| Educ | 1  | 346.94 | 6  | 98.88   | 1.967e-77 |
| AIDS | 1  | 10.74  | 3  | 5.58    | 1.050e-03 |
| gender | 1  | 3.20  | 2  | 2.38    | 0.09     |
| Educ:AIDS | 1  | 2.08  | 1  | 0.30    | 0.018    |
| Educ:gender | 1  | 3.00  | 0  | 2.154e-14 | 0.58    |
```

```r
> aids3.glm<-glm(count~Educ*AIDS+gender, family=poisson, data=aids2.df)
> anova(aids3.glm,aids2.glm, test="Chisq")
CONTINUED
```
Analysis of Deviance Table

Model 1: count ~ Educ * AIDS + gender
Model 2: count ~ Educ * AIDS * gender

| Resid. Df | Resid. Dev | Df | Deviance | P(>|Chi|) |
|-----------|------------|----|----------|---------|
| 1         | 3          | 5.5810      | 2          | 2.154e-14 | 3 | 5.5810 | 0.1339 |

Which model is indicated by this output?

(1) A model where Educ and AIDS are conditionally independent, given gender.
(2) A model where Educ and AIDS are independent of gender.
(3) The homogeneous association model.
(4) A model where Educ and AIDS are independent.
(5) The saturated model.

23. A suitable model was fitted to the data. A fragment of computer output relating to this model is:

Coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | 4.40294 | 0.08601 | 51.189 | < 2e-16 *** |
| EducDisagree | -2.41381 | 0.25315 | -9.535 | < 2e-16 *** |
| AIDSDisagree | 0.58486 | 0.09053 | 6.460 | 1.04e-10 *** |
| genderFemale | 0.28205 | 0.08106 | 3.480 | 0.000502 *** |
| EducDisagree:AIDSDisagree | 0.87239 | 0.28411 | 3.071 | 0.002136 ** |

Which of the following is TRUE? To 3 decimal places:

(1) A 95% confidence interval for the conditional log odds ratio between Educ and gender given AIDS is (0.316, 1.429).
(2) A 95% confidence interval for the log odds ratio between Educ and AIDS is (1.371, 4.176)
(3) A 95% confidence interval for the odds ratio between Educ and AIDS is (1.371, 4.176)
(4) Since the confidence interval for the log odds ratio between Educ and AIDS contains 1, Educ and AIDS are independent.
(5) A 95% confidence interval for the conditional odds ratio between between Educ and gender given AIDS is (1.371, 4.176).
24. Which of the following independence graphs represents the best model for the AIDS survey data?

![Independence Graphs](image)

Figure 4: Models for Question 24.

(1) (b)  
(2) (d)  
(3) (e)  
(4) (a)  
(5) (c)  

25. In 1895, 106 male skulls were found in the City of London, during the construction of a whisky store. It is thought that the owners of these skulls perished in the Great Plague of 1665. The mean and standard deviation cranial widths of these 106 skulls are 141.77 mm and 5.41 mm respectively. We wanted to see if the distribution of these cranial widths is normal. To do this, we divided up the skulls into 5 cranial width classes as follows:

<table>
<thead>
<tr>
<th>Class</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than or equal to 135.5 mm</td>
<td>10</td>
</tr>
<tr>
<td>Greater than 135.5 mm and less than or equal to 139.5 mm</td>
<td>21</td>
</tr>
<tr>
<td>Greater than 139.5 mm and less than or equal to 143.5 mm</td>
<td>41</td>
</tr>
<tr>
<td>Greater than 143.5 mm and less than or equal to 147.5 mm</td>
<td>19</td>
</tr>
<tr>
<td>Greater than 147.5 mm</td>
<td>15</td>
</tr>
</tbody>
</table>

An analysis using R produced the following output:

```r
> cuts = c(135.5, 139.5, 143.5, 147.5)
> prob=numeric(5)
> prob[1] = pnorm(cuts[1], mean= 141.77, sd=5.41)
```

CONTINUED
\[
\text{prob}[2] = \text{pnorm}(\text{cuts}[2], \text{mean}= 141.77, \text{sd}=5.41) - \\
\text{pnorm}(\text{cuts}[1], \text{mean}= 141.77, \text{sd}=5.41)
\]
\[
\text{prob}[3] = \text{pnorm}(\text{cuts}[3], \text{mean}= 141.77, \text{sd}=5.41) - \\
\text{pnorm}(\text{cuts}[2], \text{mean}= 141.77, \text{sd}=5.41)
\]
\[
\text{prob}[4] = \text{pnorm}(\text{cuts}[4], \text{mean}= 141.77, \text{sd}=5.41) - \\
\text{pnorm}(\text{cuts}[3], \text{mean}= 141.77, \text{sd}=5.41)
\]
\[
\text{prob}[5] = 1 - \text{pnorm}(\text{cuts}[4], \text{mean}= 141.77, \text{sd}=5.41)
\]
\[
y = c(10,21,41,19,15)
\]
\[
\text{log.Lmax} = \text{sum}(y*\log(y/\text{sum}(y)))
\]
\[
\text{log.Lmax}
\]

\[
[1] -158.5422
\]
\[
\text{log.Lmod} = \text{sum}(y*\log(\text{prob}))
\]
\[
\text{log.Lmod}
\]

\[
[1] -161.2592
\]
\[
\text{log.Lnull} = \text{sum}(y*\log(1/5))
\]
\[
\text{log.Lnull}
\]

\[
[1] -170.6004
\]
\[
\text{D}=2*(\text{log.Lmax} -\text{log.Lmod})
\]
\[
\text{D}
\]

\[
[1] 5.434104
\]
\[
1-\text{pchisq}(\text{D},1)
\]

\[
[1] 0.01974722
\]
\[
1-\text{pchisq}(\text{D},2)
\]

\[
[1] 0.06606925
\]
\[
1-\text{pchisq}(\text{D},3)
\]

\[
[1] 0.1426335
\]
\[
1-\text{pchisq}(\text{D},4)
\]

\[
[1] 0.2455828
\]
\[
1-\text{pchisq}(\text{D},5)
\]

\[
[1] 0.3652258
\]

(Note that the function \text{pnorm}(x, \text{mean}=m, \text{sd}=s) calculates the probability that \(X \leq x\) where \(X\) is Normally distributed with mean \(m\) and standard deviation \(s\).)

Which if the following is \textbf{FALSE}?

(1) The residual deviance for the normal model is about 5.4341.
(2) The residual degrees of freedom are 2.
(3) The null deviance is about 24.1164.
(4) The normal model seems more plausible than the null model.
(5) The normal model is a very poor fit to these data.
1. (a) In class, we discussed the various assumptions behind the multiple regression model. One of these was the assumption of equal variances. Briefly describe how we might detect any departures from this assumption, and then outline two possible ways of dealing with this problem. What would be the main consequence if this assumption did not hold and no corrective action was taken? [6 marks]

(b) Briefly describe and compare the different “leave one out” diagnostics we can use to identify influential points in a regression. In your discussion, specify in which aspects of the regression (coefficients, standard errors and so on) the various diagnostics can detect a change. [7 marks]

(c) Alfalfa is an important cattle food. It is thought that in areas where there is a high density of cows, land planted in alfalfa might be relatively more expensive to rent than land used for other agricultural purposes. In addition, it is thought than in areas where liming is required, the rents might be relatively less, because of the expense involved. To assess the impact of these two factors (cow density and liming) on rents for alfalfa land, data was collected on each one of the 67 counties in Minnesota that have appreciable rented farmland.

Figure 5: Diagnostic plots for QB1(c).
Figure 6: Influence plots for QB1(c).
For each county, the following were measured:

**rent**: The ratio of average alfalfa rents to average tillable rents (response);

**cow.den**: The density of cows in numbers per square mile;

**lime**: A dummy variable having value 0 if no liming is required in the county, and 1 if liming is required.

**prop**: The proportion of farmland used as pasture.

In Figures 5 and 6, we present some residual and influence plots. Give a careful discussion of points 19, 21, 33, 42 and 66, describing the effect their removal will have on the regression. Should any be removed from the regression? What else might you do to decide whether or not to remove any points? [7 marks]
2. In 1912, the liner Titanic struck an iceberg in the North Atlantic and sank with heavy loss of life. The data frame `titanic.df` contains data on 663 of the passengers. The variables are

**age.group**: The age group of the passenger (0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60+), treated as a factor;

**av.age**: The average age of persons in the age group, treated as a continuous variable;

**survival**: 0 = died, 1 = survived;

**pclass**: The passenger class (1st, 2nd, 3rd), treated as a factor;

**sex**: The gender of the passenger.

The data are not in grouped form, but if we group the data, we get the following results for \( r \) (survivors in a group) and \( n \) (size of a group)

Results for \( r \)

<table>
<thead>
<tr>
<th></th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>females</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-9</td>
<td>0</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>10-19</td>
<td>13</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>20-29</td>
<td>20</td>
<td>23</td>
<td>4</td>
</tr>
<tr>
<td>30-39</td>
<td>19</td>
<td>19</td>
<td>8</td>
</tr>
<tr>
<td>40-49</td>
<td>19</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>50-59</td>
<td>18</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>60+</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>males</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-9</td>
<td>3</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>10-19</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>20-29</td>
<td>10</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>30-39</td>
<td>12</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>40-49</td>
<td>10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>50-59</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>60+</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Results for \( n \)

<table>
<thead>
<tr>
<th></th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>females</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-9</td>
<td>1</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10-19</td>
<td>13</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>20-29</td>
<td>21</td>
<td>27</td>
<td>12</td>
</tr>
<tr>
<td>30-39</td>
<td>20</td>
<td>22</td>
<td>13</td>
</tr>
<tr>
<td>40-49</td>
<td>19</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>50-59</td>
<td>19</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>60+</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

CONTINUED
males

<table>
<thead>
<tr>
<th>Age Group</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-9</td>
<td>3</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>10-19</td>
<td>5</td>
<td>11</td>
<td>23</td>
</tr>
<tr>
<td>20-29</td>
<td>20</td>
<td>47</td>
<td>59</td>
</tr>
<tr>
<td>30-39</td>
<td>30</td>
<td>34</td>
<td>26</td>
</tr>
<tr>
<td>40-49</td>
<td>32</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>50-59</td>
<td>20</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>60+</td>
<td>15</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

(a) A logistic model including `av.age` but excluding `age.group` was fitted first. The following output was obtained:

```r
Call:
glm(formula = survived ~ av.age * pclass * sex, family = binomial, 
     data = titanic.df)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.694032   1.213975  2.219  0.0265 *
av.age        0.007081   0.031096  0.228  0.8199 
pclass2nd     0.076663   1.506380  0.051  0.9594 
pclass3rd    -2.475243   1.347298 -1.837  0.0662 .
sexmale       -1.075660   1.362284 -0.790  0.4298 
 av.age:pclass2nd -0.032895   0.041022 -0.802  0.4226 
av.age:pclass3rd -0.018260   0.038641 -0.473  0.6365 
av.age:sexmale  -0.065378   0.034787 -1.879  0.0602 .
pclass2nd:sexmale -0.103701   1.769926 -0.059  0.9533 
pclass3rd:sexmale  0.403625   1.593845  0.253  0.8001 
av.age:pclass2nd:sexmale -0.043849   0.053343 -0.822  0.4111 
av.age:pclass3rd:sexmale  0.013655   0.048850  0.280  0.7798

---

Null deviance: 869.54 on 632 degrees of freedom
Residual deviance: 506.65 on 621 degrees of freedom
AIC: 530.65

> 1-pchisq(506.65,621)
[1] 0.9997189

Analysis of Deviance Table

| Term                | Df | Deviance | Resid. Df | Resid. Dev | P(>|Chi|) |
|---------------------|----|----------|------------|------------|---------|
| NULL                | 632| 869.54    |            |            |         |
| av.age              | 1  | 4.19     | 631        | 865.36     | 0.04    |
| pclass              | 2  | 95.43    | 629        | 769.92     | 1.892e-21 |
| sex                 | 1  | 229.35   | 628        | 540.57     | 8.265e-52 |
| av.age:pclass       | 2  | 1.69     | 626        | 538.88     | 0.43    |
| av.age:sex          | 1  | 20.67    | 625        | 518.22     | 5.467e-06 |
| pclass:sex          | 2  | 10.31    | 623        | 507.90     | 0.01    |
| av.age:pclass:sex   | 2  | 1.25     | 621        | 506.65     | 0.53    |

CONTINUED
**Output of stepwise regression****

Call: glm(formula = survived ~ av.age + pclass + sex + av.age:pclass + av.age:sex + pclass:sex, family = binomial, data = titanic.df)

Coefficients:

(Intercept) av.age pclass2nd
2.48940 0.01283 0.82760
pclass3rd sexmale av.age:pclass2nd
-2.47826 -0.81700 -0.05581
av.age:pclass3rd av.age:sexmale pclass2nd:sexmale
-0.01487 -0.07257 -1.31882
pclass3rd:sexmale
0.58948

Degrees of Freedom: 632 Total (i.e. Null); 623 Residual
Null Deviance: 869.5
Residual Deviance: 507.9 AIC: 527.9

Do you think the model survived ~ av.age * pclass * sex fits well? Would a sub-model be as good? [5 marks]

(b) Write down a prediction equation for the probability that a first class female passenger will survive. Calculate the probability for a first class female passenger in age group 30-39. You may need the following table of average ages:

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Average Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-9</td>
<td>4.28260</td>
</tr>
<tr>
<td>10-19</td>
<td>16.89024</td>
</tr>
<tr>
<td>20-29</td>
<td>24.18817</td>
</tr>
<tr>
<td>30-39</td>
<td>33.84137</td>
</tr>
<tr>
<td>40-49</td>
<td>44.51578</td>
</tr>
<tr>
<td>50-59</td>
<td>53.94339</td>
</tr>
<tr>
<td>60+</td>
<td>64.07692</td>
</tr>
</tbody>
</table>

[5 marks]

(c) The original model fitted above was refitted treating av.age as a factor, using the code

```
model1 = glm(survived ~ factor(av.age)*pclass*sex, data=titanic.df, family=binomial)
```

Describe in detail the differences between these two models. Why is the model treating av.age as a factor more general than the model treating av.age as a continuous variable? [5 marks]
Figure 7: Plot for QB2(d).

(d) In fact a test comparing these two models has a p-value of 0.05, so there is some evidence that the “factor version” of the model is better. In Figure 6 we show a plot of the sample logits plotted against average age, for the 6 sex/class groups. What do you think might be the reason for the “non-factor” version proving inadequate?[5 marks]
3. (a) Suppose we divide voters into two sub-populations of males and females. We take a separate random sample from each, and classify the individuals by how they voted (V) in the last election, say by Labour, National, Other, Didn’t vote. Describe how we could use Poisson regression to decide if the male and female sub-populations differ in their voting behaviour. [5 marks]

(b) Suppose we also measured if the respondents favour increased immigration (I) using a question with three possible responses (Yes, No and Don’t know/won’t answer). How could we test if the joint distribution of V and I was different in the male and female sub-populations? [5 marks]

(c) The data shown on the next page arose from a US political survey. Separate samples of 500 male and 600 female voters were taken, and their political affiliation (Democrat, Republican, Independent) was measured.

<table>
<thead>
<tr>
<th>count</th>
<th>affil</th>
<th>sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>119</td>
<td>R</td>
<td>F</td>
</tr>
<tr>
<td>203</td>
<td>D</td>
<td>F</td>
</tr>
<tr>
<td>178</td>
<td>I</td>
<td>F</td>
</tr>
<tr>
<td>159</td>
<td>R</td>
<td>M</td>
</tr>
<tr>
<td>236</td>
<td>D</td>
<td>M</td>
</tr>
<tr>
<td>205</td>
<td>I</td>
<td>M</td>
</tr>
</tbody>
</table>

Do you think the party preference differs between the sexes? If so, how? Use the output below. [5 marks]

```
> anova(glm(count~affil*sex, family=poisson, data=US.df), test="Chisq")
Analysis of Deviance Table
Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL 5 47.709
affil 2 37.546 3 10.163 7.032e-09
sex 1 9.103 2 1.060 0.003
affil:sex 2 1.060 0 2.220e-15 0.589
```

(d) In another survey, a population was divided into 4 sub-populations according to their sex and socio-economic status (having values low and not low.) A sample of 250 from each subpopulation was taken and the sampled individuals asked their opinion of legalised abortion (support/don’t support). The results were as follows:

<table>
<thead>
<tr>
<th>count</th>
<th>status</th>
<th>sex</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>171</td>
<td>low</td>
<td>F</td>
<td>S</td>
</tr>
<tr>
<td>138</td>
<td>not low</td>
<td>F</td>
<td>S</td>
</tr>
<tr>
<td>152</td>
<td>low</td>
<td>M</td>
<td>S</td>
</tr>
<tr>
<td>167</td>
<td>not low</td>
<td>M</td>
<td>S</td>
</tr>
<tr>
<td>79</td>
<td>low</td>
<td>F</td>
<td>DS</td>
</tr>
<tr>
<td>112</td>
<td>not low</td>
<td>F</td>
<td>DS</td>
</tr>
<tr>
<td>148</td>
<td>low</td>
<td>M</td>
<td>DS</td>
</tr>
<tr>
<td>133</td>
<td>not low</td>
<td>M</td>
<td>DS</td>
</tr>
</tbody>
</table>

CONTINUED
Do the 4 sub-populations differ in their opinion on legalised abortion? Give a reason for your answer, based on the following (edited) output. [5 marks]

```r
> anova(glm(count~status*sex*support, family=poisson, data=survey.df), test="Chisq")
Analysis of Deviance Table

Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL 7 50.392
status 1 2.842e-14 6 50.392 1.000
sex 1 9.103 5 41.289 0.003
support 1 22.198 4 19.090 2.459e-06
status:sex 1 1.421e-14 3 19.090 1.000
status:support 1 1.203 2 17.888 0.273
sex:support 1 8.331 1 9.556 0.004
status:sex:support 1 9.556 0 1.377e-14 0.002
```

```r
> summary(glm(count~status*sex*support, family=poisson, data=survey.df))
> 1-pchisq(19.090,3)
[1] 0.0002619304

> summary(glm(count~status*sex*support, family=poisson, data=survey.df))
Residual deviance: 10.768 on 3 degrees of freedom
> 1-pchisq(10.768,3)
[1] 0.01304887

> summary(glm(count~status*sex*support, family=poisson, data=survey.df))
Residual deviance: 19.090 on 4 degrees of freedom
> 1-pchisq(19.090,4)
[1] 0.0007545808

> summary(glm(count~status*support+sex, family=poisson, data=survey.df))
Residual deviance: 17.888 on 3 degrees of freedom
> 1-pchisq(17.888,3)
[1] 0.0004638738
```

ANSWER SHEET FOLLOWS

CONTINUED