THE UNIVERSITY OF AUCKLAND

SECOND SEMESTER, 2008
Campus: City

STATISTICS

Advanced Statistical Modeling

(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.
- Each correct answer scores 2.4 marks.

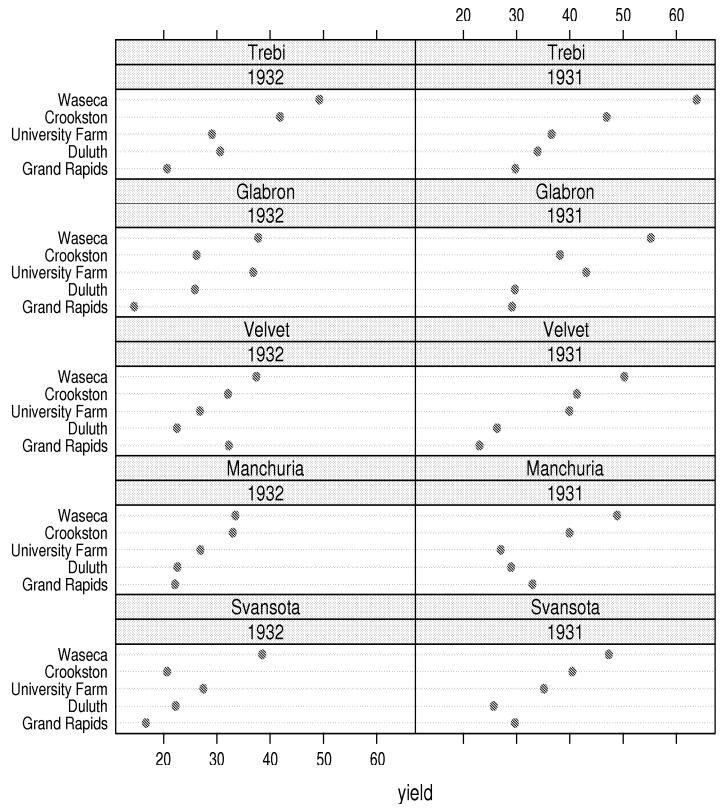
SECTION B (40 marks)

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

Total for both parts: 100 marks

CONTINUED

SECTION A



1. Figure 1: Trellis plot for Question 2.

The data for this question are yields in bushels per acre, of 5 varieties of barley grown at University Farm, St. Paul, and at the five branch experiment stations located at Waseca, Morris, Crookston, Grand Rapids, and Duluth (all in Minnesota). The varieties (Svansota, Manchuria, Velvet, Glabron and Trebi) were grown at each of the five stations during 1931 and 1932, different land being used each year of the test. A trellis plot of these data is shown in Figure 1. Which of the following is **TRUE**?

(zz) The yields at Waseca are higher than the other stations.

(1) The yields in 1932 are higher than in 1931.

(1) Manchuria and Trebi have similar yields.

(1) Grand Rapids always has the worst yield.

(1) Svansota clearly has the highest yield.

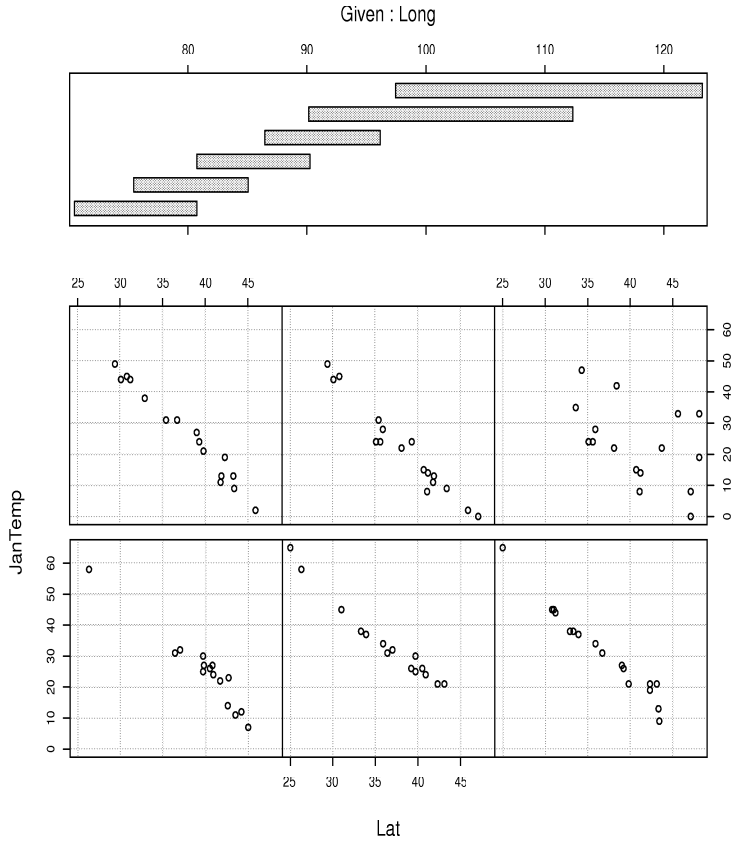


Figure 1: Trellis plot for Question 2.

2. The data for this question consist temperature data for several US cities, together with their latitudes and longitudes. The variables are

   JanTemp: The minimum January temperature,

   Lat: The latitude, (expressed in degrees north of the equator,so that places with higher values are further north),

   Long: The longitude, (expressed in degrees west of Greenwich,so that places with higher values are further west).

   The data are in a data frame called `temperatures.df`. Which of the following is **TRUE**?

   (zz) For most longitudes, the temperatures decrease linearly with increasing latitude.

   (1) The temperatures don't depend on longitude.

   (1) For most latitudes, the temperatures decrease linearly with increasing longitude.

   (1) The relationship between temperature and longitude is weakest in the East.

   (1) The cities having the highest temperatures are in the west.

3. Which of the following R commands produced Figure 2?

   (zz) `coplot(JanTemp~Lat|Long, data=temperatures.df).`

   (1) `coplot(Long~Lat|JanTemp, data=temperatures.df).`

   (1) `coplot(Jantemp~Long|Lat, data=temperatures.df).`

   (1) `bwplot(Jantemp~Lat|Long, data=temperatures.df).`

   (1) `dotplot(Jantemp~Lat|Long, data=temperatures.df).`

4. Which of the following plots would be useful in detecting serial correlation in a regression?

   (zz) An acf plot.

   (1) A plot of residuals versus fitted values.

   (1) A gam plot.

   (1) A leverage-residual plot.

   (1) A normal plot.

5. When fitting a regression model with a continuous response Y and two continuous covariates X and W, which of the following does **NOT** affect the variance of the regression coefficient of X?

   (zz) The correlation between X and the response.

   (1) The sample size.

   (1) The correlation between X and W.

   (1) The error variance.

   (1) The variance of X.

   A regression was fitted to the temperature data in Question 2, with `JanTemp` as the response and the other two variables as explanatories. The following summary was obtained:

6.
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 100.26630    8.69438  11.532 4.56e-16 ***
Lat          -2.17010    0.18345 -11.829  < 2e-16 ***
Long          0.11698    0.06593   1.774   0.0817 .

Residual standard error: 7.24 on 53 degrees of freedom
Multiple R-squared: 0.7253,     Adjusted R-squared: 0.7149
F-statistic: 69.97 on 2 and 53 DF,  p-value: 1.348e-15
```

   Assuming the fitted model is adequate, which of the following is **TRUE**?

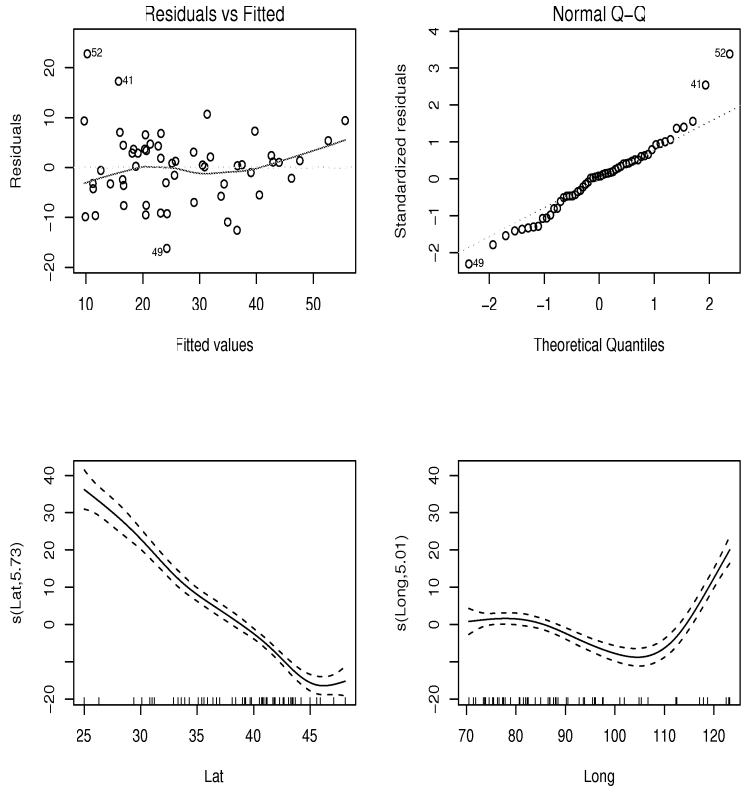
(zz) The temperature goes down about 2 degrees as you go one degree further north.

(1) The temperature goes down about one degree as you go two degrees further south.

(1) The temperature goes up about 1 degree as you go ten degrees further east.

(1) The estimate of error variance is 7.24.

(1) There is no evidence that either of these variables helps explain the response.



7. Figure 3: Trellis plot for Question 7.

In Figure 3, some diagnostic plots are shown. Which is the correct interpretation of these plots?

(zz) The regression surface is non-planar.

(1) There are no problems with this regression.

(1) There is serial correlation in these data.

(1) The variability of the data is increasing with the mean.

(1) Latitude needs to be transformed.

8. Which is the correct remedial action to improve the fit of the model for the temperature data?

   (zz) Try a quadratic in Longitude.

   (1) Try a quadratic in Latitude.

   (1) Remove the outliers.

   (1) Transform the response.

   (1) Do nothing, there are no problems.

9. Part of an influence display (3 out of 56 lines) for the temperature data is shown below.

```
      dfb.1_   dfb.Lat  dfb.Long    dffit cov.r   cook.d    hat inf

41 -0.783554  0.358883  0.769850  0.98216 0.813 2.88e-01 0.1180   *
49  0.334179 -0.063991 -0.461593 -0.58307 0.817 1.04e-01 0.0555   *
52 -1.285911  0.771910  1.031853  1.49236 0.590 5.93e-01 0.1346   *
```

   Note: Points are influential if

   (a) Cook's D is more than $F_{3,52}(0.1) = 0.1940$,

   (b) $|DFBETAS| > 1$,

   (c) $|DFFITS| > \sqrt{\frac{3p}{n-p}}$,

   (d) $|COVRATIO - 1| > 3p/n$,

   and have high leverage if the HMD exceeds $3p/n$. Which is the **CORRECT** interpretation?

   (zz) Point 52 is influencing the estimation of the constant term.

   (1) Point 49 is influencing the estimation of the latitude coefficient.

   (1) Point 52 is has no effect on the standard errors.

   (1) Point 52 is influential because of its very high leverage.

   (1) Point 49 has a small residual.

10. Suppose we have a regression with a continuous response Y, and four explanatory variables X, W, A and B, where X and W are continuous and A and B are factors. Which of the following models would you fit intitially when building a model?

    (zz) `Y~X*A*B + W*A*B`.

    (1) `Y~A*B`.

    (1) `Y~X + W + A + B`.

    (1) `Y~X*A + W*B`.

    (1) `Y~X+W`.

11. Consider another regression with a continuous response Y, and three explanatory variables X, A and B, where X is continuous and A and B are factors, each having two levels, which we denote by 1 (baseline) and 2. We fit the model `Y~X + A*B`, and obtain the following output:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2291     0.1590   7.731 3.86e-11 ***
x             2.3060     0.2111  10.925  < 2e-16 ***
A             0.3402     0.1719   1.978   0.0516 .
B             1.6006     0.1732   9.239 5.24e-14 ***
A:B          -1.7683     0.2426  -7.289 2.66e-10 ***
Residual standard error: 0.5423 on 75 degrees of freedom
Multiple R-squared: 0.7844,     Adjusted R-squared: 0.7729
F-statistic: 68.21 on 4 and 75 DF,  p-value: < 2.2e-16
```

Assuming the model is correct, which of the following is the **INCORRECT** interpretation?

(zz) When A is at its baseline level, and B is at level 2, the relationship between Y and X is linear with slope 2.3060+1.6066 and intercept 1.2291.

(1) When A and B are at their baseline levels, the relationship between Y and X is linear with slope 2.3060 and intercept 1.2291.

(1) There is strong evidence of interaction between A and B.

(1) When A is at level 2, and B is at its baseline level, the relationship between Y and X is linear with slope 2.3060 and intercept 1.2291+0.3402.

(1) Irrespective of the values of A and B, the mean response goes up about 2.3 units with a unit increase in x.

12. The data for Questions 12-14 come from a study to investigate the effect of marijuana smoking on reaction times. The study involved 36 subjects, of whom 12 had never smoked marijuana, 12 were classified as "light users" and 12 "moderate users". Each of these three groups of 12 subjects was divided at random into two groups of six: six were asked to smoke regular cigarettes that tasted and smelled like real marijuana cigarettes (the "placebo"), and six smoked real marijuana cigarettes ('experimental"). After smoking, each subject given a reaction time test, and the reaction times measured. The data were (times in 1/000's of a second)

```
Placebo
  None Light Moderate
1  795   800      790
2  700   705      695
3  648   645      634
4  605   610      600
5  752   757      752
6  710   712      705
```

```
Experimental

  None Light Moderate
1  965   843      815
2  865   765      735
3  811   713      983
4  878   665      635
5  916   810      782
6  840   776      744
```

A model using time as the response and the factors `previous`, (with levels "None","Light","Moderate") and `treat` (with levels "Placebo","Experimental") as explanatory variables was fitted to the data, with the following results:

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        701.667     31.251  22.452  < 2e-16 ***
previousLight                        3.167     44.196   0.072 0.943356
previousModerate                    -5.667     44.196  -0.128 0.898833
treatExperimental                  177.500     44.196   4.016 0.000365 ***
previousLight:treatExperimental   -120.333     62.503  -1.925 0.063725 .
previousModerate:treatExperimental  -91.167     62.503  -1.459 0.155065

Analysis of Variance Table

Response: times
               Df Sum Sq Mean Sq F value    Pr(>F)
previous        2  23634   11817  2.0166 0.1507518
treat           1 103041  103041 17.5842 0.0002239 ***
previous:treat  2  23642   11821  2.0173 0.1506649
Residuals      30 175796    5860

> anova(reaction.sub, reaction.lm)
Analysis of Variance Table

Model 1: times ~ treat
Model 2: times ~ previous * treat
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     34 223073
2     30 175796  4     47277 2.017 0.1174
```

What is the best interpretation of this output?

(zz) There seems to be very little effect of previous smoking on reaction time.

(1) The effect of the treatment depends on the previous smoking history.

(1) For persons with moderate smoking history, the subjects in the treatment group are about 0.177 sec slower than those in the placebo group.

(1) The treatment has no effect on reaction time.

(1) The average reaction time for all subjects is 0.701 sec.

13. Two diagnostic plots from the fit in Question 12 are shown in Figure 4.

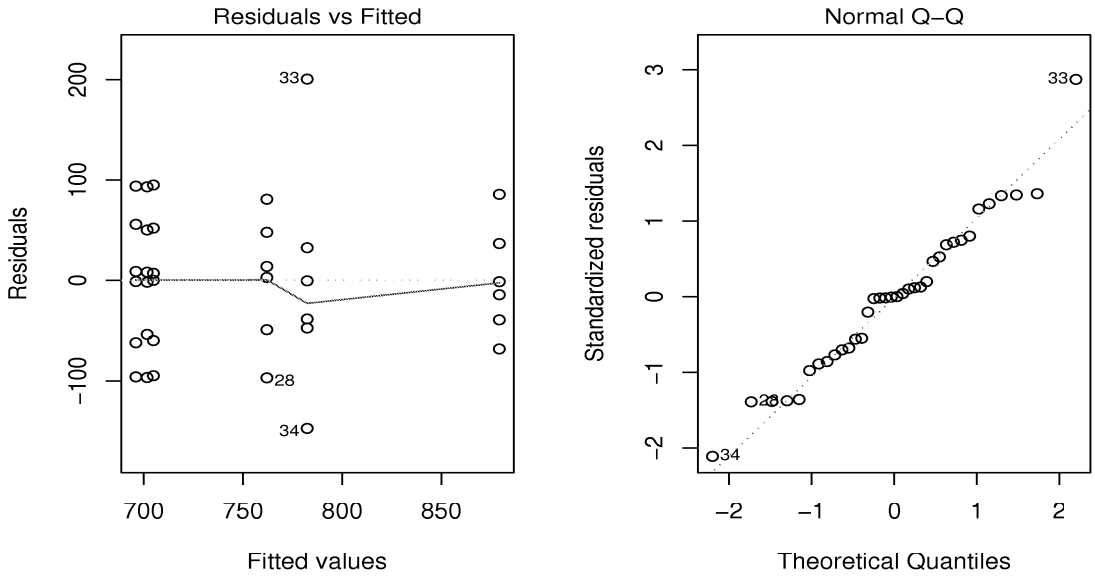


Figure 4: Diagnostic plots for Question 13.

What is the **BEST** interpretation of these plots?

(zz) The plots indicate no serious problems.

(1) The largest standardised residual is suspiciously large.

(1) The six groups seem to have very different variances. variance.

(1) The variances of the groups seem too increase with the mean.

(1) A transformation of the response seems indicated.

14. What is your estimate of the mean response time for all subjects with no previous history of smoking if they are subjected to the treatment?

(zz) 701.667 + 177.500

(1) 701.667 - 177.500

(1) 701.667 + 3.167

(1) 701.667 - 5.667

(1) 701.667 + 177.500 - 5.667 - 91.167

15. Suppose we have a binary response Y (with values 0 and 1) and a continuous explanatory variable X. Which of the following is `TRUE`?

    (zz) The correct R code to fit the model is `glm(Y~x, family=binomial)`.

    (1) The correct R code to fit the model is `lm(Y~x, family=binomial)`.

    (1) The correct R code to fit the model is `glm(Y~x, family=poisson)`.

    (1) The correct R code to fit the model is `glm(Y~x)`.

    (1) The correct R code to fit the model is `lm(x ~Y)`.

16. In the coronary heart disease example studied in class, we fitted a linear logistic model, using `CHD` (0=no CHD, 1=CHD) as the response and `AGE` as the explanatory variable. The regression summary is

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.2784     1.1296  -4.673 2.97e-06 ***
age           0.1103     0.0240   4.596 4.30e-06 ***
```

    What is the correct interpretation of this summary?

    (zz) The log-odds of having CHD goes up 0.1103 with each extra year of age.

    (1) The log-odds of having CHD goes down 5.2784 with each extra year of age.

    (1) The odds of having CHD goes up 0.1103 with each extra year of age.

    (1) The probability of having CHD goes up 0.1103 with each extra year of age.

    (1) The probability of having CHD goes down by a factor of exp(5.2784) with each extra year of age.

17. Suppose we wnat to estimate the age for which the probability of having CHD is 0.1. Which of the following is `TRUE`?

    (zz) The age (to 2 decimal places) is estimated as 27.93.

    (1) The age (to 2 decimal places) is estimated as 56.01.

    (1) The age (to 2 decimal places) is estimated as 48.76.

    (1) The age (to 2 decimal places) is estimated as 67.78.

    (1) The age can't be estimated from the information given.

18. In a logistic regression with ungrouped data, with two continuous explanatory variables, the residual deviance was 131.67 on 97 degress of freedom, and the null deviance was 135.37 on 99 degrees of freedom. The following output was obtained:

```
> 1-pchisq(135.37,99)
[1] 0.008929477
> 1-pchisq(131.67,97)
[1] 0.01104207
> 1-pchisq(3.70,2)
[1] 0.1572372
```

CONTINUED

Which of the following is **TRUE**? (Note that for ungrouped data, we can use the difference in deviances to compare models.)

(zz) There is no evidence that either of the variables helps explain the response.

(1) The residual deviance indicates that the model fits well.

(1) At least one explanatory variable should be retained.

(1) The residual deviance can be used to judge the goodness of fit.

(1) The null deviance indicates that the model fits well.

19. In the Child cancer death rate example studied in class, the the number of deaths (the variable `n`) and the population at risk (the variable `pop`) were given. Suppose we want to model the death rate per 1000 people. The offset used would be

(zz) `log(pop/1000)`.

(1) `log(pop)`.

(1) `log(pop/100000)`.

(1) `log(1000/pop)`.

(1) `log(100000/pop)`.

20. Suppose we want to model the number of deaths per 1000 people in terms of the cytology (either "L" or "M", the place of residence ("Rural" or "Urban"), and age ("0-5", "6-14"). The regression summary from the analysis is

```
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.00543    0.16222  -6.198 5.72e-10 ***
CytologyM                     -2.02815    0.47573  -4.263 2.01e-05 ***
ResidenceU                     0.02502    0.21430   0.117   0.9071
Age6-14                       -1.47811    0.32131  -4.600 4.22e-06 ***
CytologyM:ResidenceU           0.66127    0.56820   1.164   0.2445
CytologyM:Age6-14              1.54264    0.65440   2.357   0.0184 *
ResidenceU:Age6-14             0.75173    0.38714   1.942   0.0522 .
CytologyM:ResidenceU:Age6-14 -0.79095    0.77576  -1.020   0.3079
```

Which of the following is `TRUE`?

(zz) The death rate (deaths per per 1000 persons) to 4 decimal places for rural individuals aged 0-5 having type "L" cytology is 0.3659.

(1) The death rate (deaths per per 1000 persons) to 4 decimal places for urban individuals aged 0-5 having type "L" cytology is 0.3659.

(1) The death rate (deaths per per 1000 persons) to 4 decimal places for rural individuals aged 6-14 having type "L" cytology is 0.3659.

(1) The death rate (deaths per per 1000 persons) to 4 decimal places for rural individuals aged 0-5 having type 'M" cytology is 0.3659.

(1) None of the other alternatives is true.

21. In the analysis of a one-dimensional contingency table with I cells, cell probabilities π_i and cell counts y_i, the log-likelihood is (up to a constant not involving π is

$$\sum_{i=1}^{I} y_i \log \pi_i.$$

Which of the following gives the log-likelihood $\log L_{MAX}$ of the maximal (unrestricted) model?

(zz) Substituting the estimated probabilities y_i/n for π_i , where n is the sum of the counts.

(1) Substituting the probabilities 1/I for π_i.

(1) Substituting probabilities worked out under the assumption of independence for π_i.

(1) Substituting probabilities worked out under the the Poisson assumption for π_i.

(1) Substituting probabilities worked out under the the Binomial assumption for π_i.

22. The data below are from a famous data set, gathered by Sir Earnest Rutherford, which records the number of α-particles emitted by a radioactive substance over $N = 2608$ time intervals of 7.5 seconds each. The data have been grouped for your convenience:

```
Number of particles    0   1   2   3   4   5   6   7  8  9 10 or more
CNumber of intervals  57 203 383 525 532 408 273 139 45 27 16
```

We want to check if these data follow a Poisson distribution. We get the following R output:

```
freqs = c(57, 203,383,525,532,408,273,139,45,27,16)
N=sum(freqs)

p.mean = sum((0:10)*freqs)/sum(freqs)
poisson.probs = c(dpois(0:9, p.mean), 1-ppois(9, p.mean))

logL1  = sum(freqs*log(freqs/N))
logL2 = sum(freqs*log(poisson.probs))
logL3 = sum(freqs*log(1/11))
> logL1
[1] -5326.298
> logL2
[1] -5333.258
> logL3
[1] -6253.711
 1-pchisq(13.92081,9)
[1] 0.1251702
> 1-pchisq(1854.826,10)
[1] 0
```

Which of the following is **FALSE**?

(zz) The residual deviance for the Poisson model is 1854.826.

(1) The test statistic for testing the hypothesis that the true distribution is Poisson is has value 13.92.

(1) The Poisson model fits the data well.

(1) The test statistic for testing the hypothesis that the true distribution is uniform (i.e. the probabilities of 0,1,2,...,10 particles are all the same) is 1854.826.

(1) The null deviance for the Poisson model is 1854.826.

23. The table below is the result of classifying a group of Yanomamo indians according to their lineage ("Sha", "Hor", "Shamatar" or "Other") and their gender ("Male", "female", with Male treated as the baseline).

	Male	Female
Sha	70	59
Hor	82	48
Shamatar	11	31
Other	55	58

The following output was obtained:

```
> indian.df
  counts  lineage    sex
1     70      Sha   Male
2     82      Hor   Male
3     11 Shamatar   Male
4     55    Other   Male
5     59      Sha Female
6     48      Hor Female
7     31 Shamatar Female
8     58    Other Female

> indian.glm = glm(counts~lineage*sex, data=indian.df, family=poisson)

> anova(indian.glm, test="Chisq")

            Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                            7     80.116
lineage      3   60.179         4     19.936 5.382e-13
sex          1    1.170         3     18.767     0.279
lineage:sex  3   18.767         0  1.532e-14 3.055e-04
```

Which of the following is **TRUE**?

(zz) There is strong evidence that sex and lineage are not independent.

(1) It appears that sex is not required in the model.

(1) To decide the question, we need to look at odds ratios.

(1) The model fitted is inappropriate: we should be doing a logistic regression.

(1) Fitting a Poisson regression is inappropriate for grouped data.

24. The regression summary for the model fitted in Question 23 is

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                  4.2485     0.1195  35.545  < 2e-16 ***
lineageHor                   0.1582     0.1627   0.972  0.33089
lineageShamatar             -1.8506     0.3243  -5.706 1.16e-08 ***
lineageOther                -0.2412     0.1802  -1.338  0.18077
sexFemale                   -0.1710     0.1767  -0.967  0.33339
lineageHor:sexFemale        -0.3646     0.2535  -1.438  0.15041
lineageShamatar:sexFemale    1.2070     0.3929   3.072  0.00213 **
lineageOther:sexFemale       0.2241     0.2582   0.868  0.38547
```

Which of the following is **TRUE**?

(zz) To 4 decimal places, a 95% confidence interval for the odds ratio corresponding to lineage "Other" and sex "Female" is [0.7542, 2.0754].

(1) To 4 decimal places, a 95% confidence interval for the odds ratio corresponding to lineage "Other" and sex "Female" is [0.7542, 2.0754].

(1) The odds ratio is not defined for this combination of levels.

(1) The odds ratio can't be calculated from this display.

(1) To 4 decimal places, a 95% confidence interval for the odds ratio corresponding to lineage "Shamatar" and sex "Female" is [0.4369, 1.9771].

25. In a three-dimensional contingency table with factors A, B and C, we want to test the hypothesis that factor A is independent of factors B and C, using an R statement of the form, `anova(model1, model2)`. What should the formulas defining model 1 and model 2 be? The R vector `count` contains the cell counts.

(zz) Model 1: `count ~ A + B*C,` Model 2: `count ~A*B*C.`

(z) Model 1: `count ~ A + B + C,` Model 2: `count ~A*B*C.`

(1) Model 1: `count ~ A*B + A*C,` Model 2: `count ~A + B*C.`

(1) Model 1: `count ~ 1,` Model 2: `count ~A*B*C.`

(1) Model 1: `count ~ A + B + C,` Model 2: `count ~A + B*C.`