THE UNIVERSITY OF AUCKLAND

SECOND SEMESTER, 2010
Campus: City

STATISTICS

Statistical Modelling

(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 question is worth 20 marks.

Total for both parts: 100 marks

SECTION A

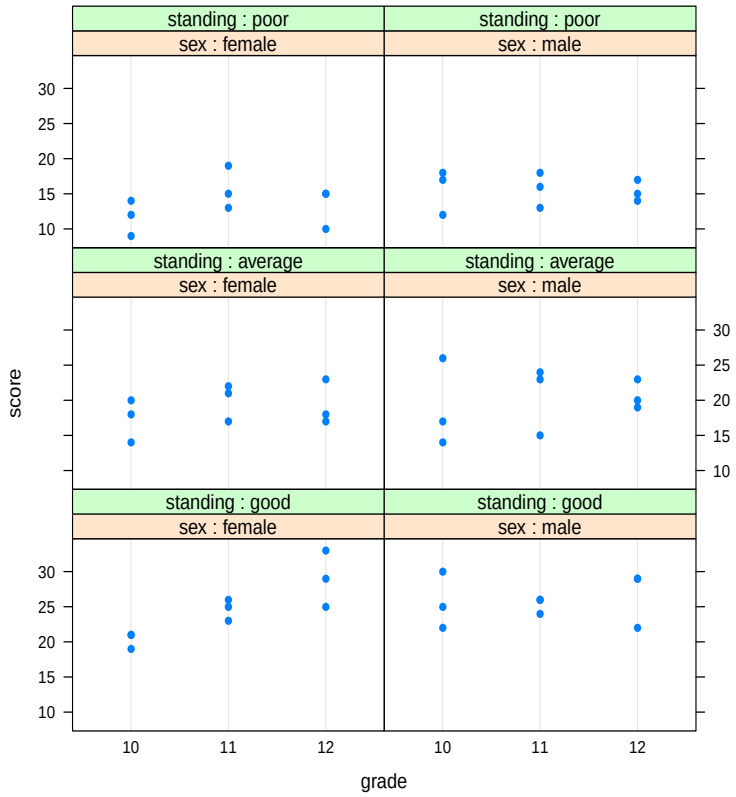

Figure 1: Trellis plot for Question 1.

1. The data for this question are exam scores (variable `score`) of students, who are classified according to the following factors:

   sex : female/male,

   standing : Scholastic standing, either good, average, poor,

   grade : either 10, 11, 12.

   A trellis plot of these data is shown in Figure 1. Which of the following is **TRUE**?

   (1) Scores are uniformly better in the 12th grade than in the 10th.

   (2) Females in the 10th grade have the best scores.

   (3) For males, the scores in 12th grade are better than 10th grade.

   (4) As scholastic standing goes from poor to average to good, the score goes up.

   (5) Females consistently do better than males.

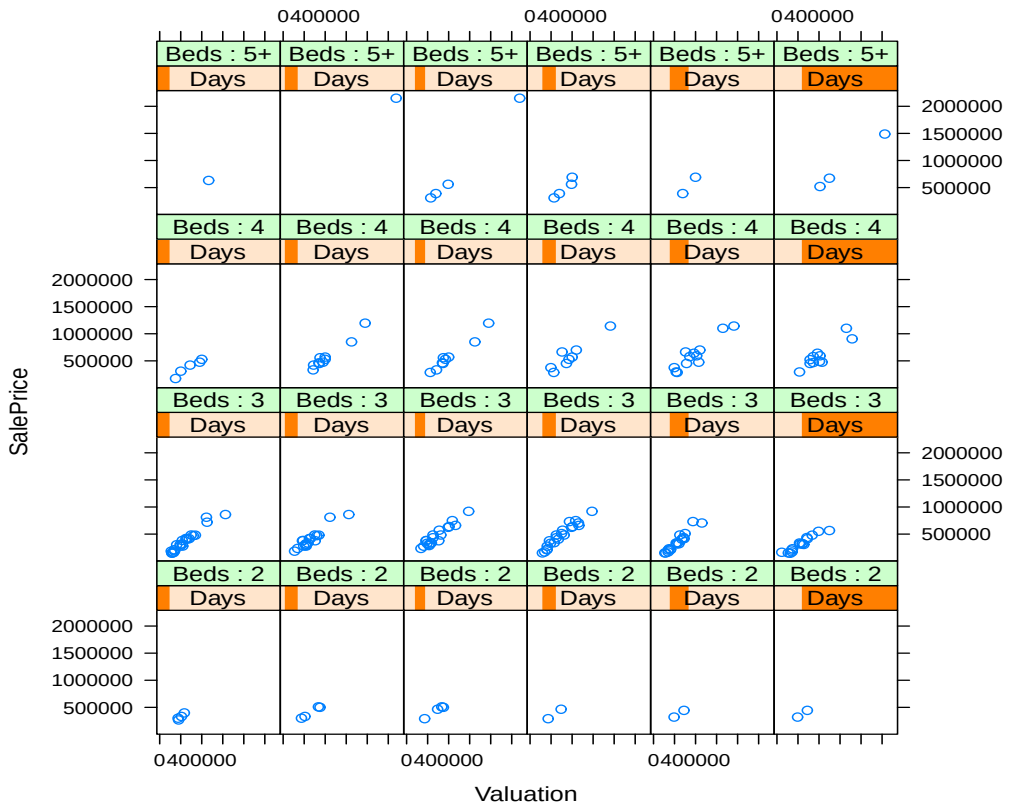

Figure 2: Trellis plot for Question 2.

2. The data for this question consist of the sale prices (variable`SalePrice`) of 113 houses, together with the variables

   Valuation : the valuation of the house in 2002

   DaysOnMarket : Number of days the house was on the market before selling

   Bedrooms : The number of bedrooms, treated as a numeric variable.

   Beds : The number of bedrooms, treated as a factor with levels 2,3,4,5+.

   These variables are in a data frame `houses.df`. The idea is to explore the relationship between the valuation and the sale price, and how this is affected by the number of bedrooms and the days on the market.

Which of the following is **FALSE**?

(1) The data seem planar.

(2) There are not many 2-bedroom houses in the data set.

(3) There is a strong relationship between the valuation and the sale price.

(4) Days on market seems to have an effect on the slopes.

(5) The most expensive house has 5+ bedrooms.

3. Which of the following R commands produced Figure 2? Note: the variable Days was created by the line

```
Days = equal.count(houses.df$DaysOnMarket)
```

(1) `xyplot(SalePrice~Days|Valuation*Beds, data=houses.df)`

(2) `xyplot(SalePrice~Valuation|Days*Beds, data=houses.df)`

(3) `dotplot(SalePrice~Valuation|Days*Beds, data=houses.df)`

(4) `bwplot(SalePrice~Valuation|Days*Beds, data=houses.df)`

(5) `xyplot(SalePrice~Beds|Days*Valuation, data=houses.df)`

4. In a regression with two continuous covariates X and W, which of the following plots would be most useful in detecting high-leverage points in a regression?

(1) A scatter plot of X versus W.

(2) An acf plot.

(3) A gam plot.

(4) A plot of residuals versus fitted values.

(5) A normal plot.

5. Suppose that we fit a regression model with a continuous response Y and two continuous covariates X and W. Which of the following statements is **FALSE**?

(1) If the correlation between X and W is very close to 0, the VIF for W will be close to its minimum value.

(2) A large p-value for X does not necessarily mean that there is no relationship between X and the response Y.

(3) The VIF for X depends only on the values of X.

(4) If the correlation between X and W is very close to 1, the standard error for the coefficient of X will be large.

(5) If the correlation between X and W is zero, the VIF of X is 1.

6. A regression was fitted to the housing data described in Question 2, with `SalePrice` as the response and the other three variables as explanatories. The following summary was obtained:

```
Model1.lm = lm(SalePrice~Valuation + DaysOnMarket + Bedrooms, data=houses.df)
> summary(houses.lm)

Call:
lm(formula = SalePrice ~ Valuation + DaysOnMarket + Bedrooms,
    data = houses.df)

Residuals:
    Min      1Q  Median      3Q     Max
-190326  -34597   -5860   28334  460125

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.552e+03  3.046e+04   0.314  0.75444
Valuation     1.596e+00  4.967e-02  32.122  < 2e-16 ***
DaysOnMarket -7.063e+02  2.500e+02  -2.825  0.00563 **
Bedrooms     -3.718e+03  1.002e+04  -0.371  0.71124
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 79000 on 109 degrees of freedom
Multiple R-squared: 0.9244,     Adjusted R-squared: 0.9223
F-statistic: 444.1 on 3 and 109 DF,  p-value: < 2.2e-16
```

Assuming the fitted model is adequate, which of the following is **NOT** a sensible interpretation?

(1) Valuations tend to over-value houses with more bedrooms.

(2) Based on the R^2, the fit seems to be very good.

(3) A house that stays a long time on the market seems to have a lower price, given the valuation and the number of bedrooms.

(4) At least some of the explanatory variables help explain the response.

(5) The estimate of standard deviation is 79000.

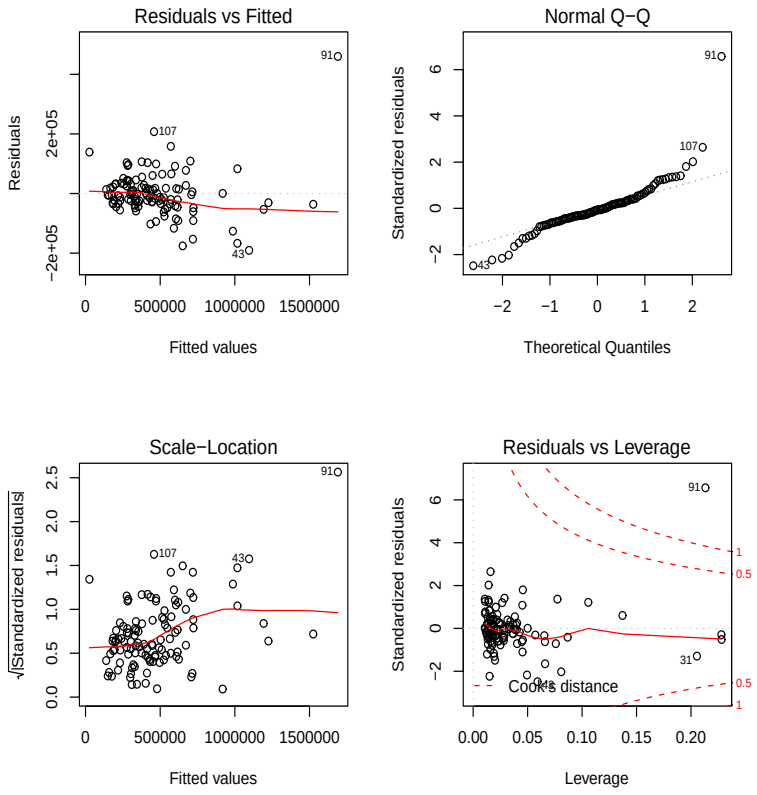

Figure 3: Trellis plot for Question 7.

7. In Figure 3, some diagnostic plots of the house price regression are shown. Which is of the following is **NOT** a correct interpretation of these plots?

   (1) There are some high-leverage points present.

   (2) There are some high-influence points present.

   (3) Point 91 is probably inflating the standard errors of the coefficients.

   (4) The regression surface is non-planar.

   (5) There is a hint that the variability of the data is increasing with the mean.

8. On the basis of Figure 3, which of the following actions is **NOT** indicated?

   (1) Use weighted least squares.

   (2) Remove one or more outliers.

   (3) Try transforming the response.

   (4) Draw an influence plot.

   (5) Try a quadratic in Valuation.

9. Part of an influence display of the house price data is shown below. Recall that there are 113 house in the data set.

	dfb.1_	dfb.Vltn	dfb.DyOM	dfb.Bdrm	dffit	cov.r	cook.d	hat	inf
3	-0.086457	-0.221853	-4.85e-01	2.59e-01	-0.61046	0.968	9.05e-02	0.0809	*
31	0.509362	0.154707	1.98e-01	-6.12e-01	-0.65580	1.228	1.07e-01	0.2054	*
43	0.106928	-0.536669	-7.32e-02	1.14e-01	-0.63820	0.874	9.70e-02	0.0591	*
64	0.099795	-0.175753	-1.55e-01	1.77e-02	-0.28119	1.331	1.99e-02	0.2281	*
88	0.032118	0.017043	-1.53e-01	-5.06e-03	-0.15796	1.340	6.29e-03	0.2279	*
91	-2.173637	3.134664	-1.42e+00	1.08e+00	4.37574	0.176	2.92e+00	0.2132	*
92	0.010614	-0.112367	4.20e-02	2.27e-02	-0.12561	1.129	3.97e-03	0.0868	*
95	0.013069	-0.049521	2.32e-01	-3.41e-02	0.24179	1.186	1.47e-02	0.1371	*
107	-0.122824	-0.113425	-3.39e-02	2.29e-01	0.34537	0.809	2.82e-02	0.0159	*
109	0.102890	-0.073510	-2.44e-02	-1.01e-01	-0.28199	0.873	1.91e-02	0.0151	*

Note: Points are influential if

- Cook's D is more than 1,
- $|DFBETAS| > 1$,
- $|DFFITS| > \sqrt{\frac{3p}{n-p}}$,
- $|COVRATIO - 1| > 3p/n$,

and have high leverage if the HMD exceeds $3p/n$. Which is **NOT** a correct interpretation?

(1) Several points have high leverage.

(2) Point 91 is influencing the estimation of the standard errors.

(3) Point 31 is having an effect on the standard errors.

(4) Point 88 has a large residual.

(5) Point 91 is influencing the estimation of the constant term.

10. Now consider fitting the same model to the house price data, except that we now use the variable Beds (which is a factor) rather than the numerical variable Bedrooms. Call this new model Model 2. Consider also a model, Model 3, which expresses the idea that the number of bedrooms does not effect the response, given the valuation and the days on the market.

```
> Model2.lm = lm(SalePrice~Valuation + DaysOnMarket + Beds, data=houses.df)
> Model3.lm = lm(SalePrice~Valuation + DaysOnMarket, data=houses.df)
> anova(Model2.lm)
Analysis of Variance Table

Response: SalePrice
              Df     Sum Sq    Mean Sq   F value    Pr(>F)
Valuation      1 8.2629e+12 8.2629e+12 1354.3792 < 2.2e-16 ***
DaysOnMarket   1 5.1182e+10 5.1182e+10    8.3892  0.004578 **
Beds           3 2.8363e+10 9.4542e+09    1.5496  0.205969
Residuals    107 6.5280e+11 6.1009e+09
> summary(houses3.lm)
Coefficients:
               Estimate Std. Error   t value Pr(>|t|)
(Intercept)  -1.301e+00  1.623e+04 -8.02e-05  0.99994
Valuation     1.587e+00  4.372e-02    36.300  < 2e-16 ***
DaysOnMarket -7.137e+02  2.482e+02    ******  0.00485 **
```

Note that in the Model3 summary, the t-value for the coefficient of DaysOnMarket has been replaced by *****. Which of the following is **FALSE**?

(1) Model 1 is assuming the effect on price is proportional to the number of bedrooms.

(2) Under Model 3, the t-value for the coefficient of DaysOnMarket is -2.876 (to 3 decimal places).

(3) Under Model 3, a house with 3 bedrooms and a valuation of $320,000 which has been on the market for 10 days is estimated to sell for just over $477,000.

(4) Under Model 3, a house with 2 bedrooms and a valuation of $300,000 which has been on the market for 10 days is estimated to sell just over $469,000.

(5) Model 3 seems adequate.

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. Which of the following statements is **INCORRECT**?

    (1) The model where the linear relationships between Y and X have arbitrary slopes and intercepts depending on A and B is expressed in R as `Y~X * A * B`.

    (2) The model where the linear relationships between Y and X have a slope depending on A but not B and arbitrary intercepts depending on A and B is expressed in R as `Y~A*B + A*X`.

    (3) The model where the linear relationships between Y and X have a common slope and intercepts depending on A but not B is expressed in R as `Y~A+ X`.

    (4) The model where the variable X has no effect on the response is expressed in R as `Y~A * B`.

    (5) The model where the linear relationship between Y and X does not depend on A nor B is expressed in R as `Y~A*B + X`.

The data for Questions 12-14 come from an experiment to compare different designs for the display panels used in air traffic control. There are three designs (denoted by I, II, III), and each design was tested under four types of simulated emergency (labelled 1,2,3,4). Two air traffic controllers were assigned to each of the 12 possible panel/emergency combinations, 24 controllers in all. The response was the time in seconds to resolve the emergency situation.

12. A model using time as the response and the factors `panel`, (with levels I,II,III) and `emergec` (with levels 1,2,3,4) as explanatory variables was fitted to the data, with the following results:

```
Coefficients:
                       Estimate Std. Error  t value Pr(>|t|)
(Intercept)          1.550e+01  1.702e+00    9.108 9.73e-07 ***
panelII             -2.000e+00  2.407e+00   -0.831 0.422172
panelIII             7.000e+00  2.407e+00    2.909 0.013113 *
emergency2           9.000e+00  2.407e+00    3.740 0.002823 **
emergency3           1.200e+01  2.407e+00    4.986 0.000316 ***
emergency4          -2.000e+00  2.407e+00   -0.831 0.422172
panelII:emergency2  -2.000e+00  3.403e+00   -0.588 0.567666
panelIII:emergency2 -3.000e+00  3.403e+00   -0.881 0.395380
panelII:emergency3   4.000e+00  3.403e+00    1.175 0.262674
panelIII:emergency3  *********  *********  ******* ********
panelII:emergency4  -2.000e+00  3.403e+00   -0.588 0.567666
panelIII:emergency4 -3.500e+00  3.403e+00   -1.028 0.324057

Residual standard error: 2.407 on 12 degrees of freedom
Multiple R-squared: 0.9498,     Adjusted R-squared: 0.9037
F-statistic: 20.63 on 11 and 12 DF,  p-value: 4.162e-06

Analysis of Variance Table

Response: time
                Df  Sum Sq Mean Sq F value    Pr(>F)
panel            2  232.75  116.37 20.0935 0.0001478 ***
emergency        3 1052.46  350.82 60.5731 1.612e-07 ***
panel:emergency  6   28.92    4.82  0.8321 0.5675015
Residuals       12   69.50    5.79
```

What is the **BEST** interpretation of this output?

(1) The average time for the two observations taken using panel I and emergency 2 was 9 seconds.

(2) The effect of the type of panel depends on the emergency.

(3) Both the type of emergency and the panel type have an effect on the time.

(4) Panel type III seems to have the shortest time.

(5) The first type of emergency seems to have the shortest time.

13. In the output above, the line for `panelIII:emergenc3` has been blanked out. The mean of the two times for for panel III and emergency 3 is 34.5. The estimate of the interaction for panel III and emergency 3 is

    (1) none of these
    (2) 7
    (3) 19
    (4) 12
    (5) 0

14. What is the **BEST** interpretation of the output above?

    (1) The difference in means between emergency 1 and emergency 2 depends on the panel.
    (2) An interaction plot would show non-parallel traces.
    (3) There is no significant evidence of interaction.
    (4) The difference in means between panel I and panel II depends on the emergency.
    (5) Because there are only two observations per factor level combination we cannot make a decision about the interactions.

15. In the logistic regression model relating a binary response Y (with values 0 and 1) to a continuous covariate X, with $\text{logit}P[Y = 1] = \alpha + \beta x$, which one of the following is **TRUE**?

    (1) It will always be possible to estimate α and β using maximum likelihood.
    (2) If β is large and positive, the probability that $Y = 1$ changes slowly from 0 to 1 as x increases.
    (3) If $\alpha < 0$, the probability that $Y = 1$ goes up as x goes up.
    (4) If α is very large in magnitude and negative, and β is positive and has moderate magnitude, the probability that $Y = 1$ for moderate x is very large.
    (5) If $\beta > 0$, the probability that $Y = 1$ goes up as x goes up.

Questions 16-18 refer to the following data: When patients have been diagnosed as having prostate cancer, an important question when deciding on treatment is whether or not the cancer has spread to the neigbouring lymph nodes. To study the risk factors for this, a study was performed in which the following explanatory variables were measured on 53 patients :

X.ray : A binary variable, 0=less serious, 1=serious,

stage : the stage of the tumour, 0=less serious, 1=serious,

acid.ph : a continuous variable giving the level of serum acid phosphotase.

The response is recorded in the variable `Y`, a binary variable where 1 indicates a nodal involvement, and 0 no nodal involvement.

16. A logistic model `Y~ acid.ph + X.ray +stage` was fitted, and the following summary obtained:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.57565    1.18115  -3.027  0.00247 **
acid.ph      0.02063    0.01265   1.631  0.10291
X.ray        2.06179    0.77767   2.651  0.00802 **
stage        1.75556    0.73902   2.376  0.01752 *

Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 50.660  on 49  degrees of freedom
AIC: 58.66
```

What is the **CORRECT** interpretation of this summary?

(1) The odds of having nodal involvement goes up 0.02063 if the acid ph increases by one unit.

(2) The log-odds of having nodal involvement goes up by a factor of 5.786687 if the stage is serious, other things being equal.

(3) The probability of having nodal involvement goes up over seven times if the X-ray is serious, other things being equal.

(4) The probability of having nodal involvement goes up 1.75556 if the stage is serious, other things being equal.

(5) The log-odds of having nodal involvement goes up by 2.06179 if the X-ray is serious, other things being equal.

17. Suppose that we want to we want to estimate the probability that a person having an acid phosphotase level of 62, a serious X-ray and a non-serious stage will have nodal involvement. We get the following output (`lymph.glm` contains the results of fitting the model above)

```
> predict(lymph.glm, newdata=data.frame(acid.ph=62, X.ray = 1,  stage=0),
+  type="response", se=TRUE)
$fit
        1
0.4415589

$se.fit
        1
0.1905992

> qnorm(0.975)
[1] 1.959964
> qnorm(0.995)
[1] 2.575829
```

Which of the following is **TRUE**? To 4 decimal places:

(1) A 95% confidence interval for the odds is [0.0680, 0.8151].

(2) A confidence interval for the probability can't be calculated from the information given.

(3) A 95% confidence interval for the probability is [0.0680, 0.8151].

(4) A 99% confidence interval for the probability is [0.0680, 0.8151].

(5) A 95% confidence interval for the log-odds is [1.0704, 2.2595].

18. 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

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

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

(3) `log(pop)`.

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

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

19. Suppose that 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", "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**?

(1) None of the other alternatives is true.

(2) The death rate (deaths per 1000 persons) to 4 decimal places for rural individuals aged 6-14 having type "L" cytology is 0.2281.

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

(4) The death rate (deaths per 1000 persons) to 4 decimal places for urban individuals aged 6-14 having type "M" cytology is 0.4534.

(5) The death rate (deaths per 1000 persons) to 4 decimal places for urban individuals aged 0-5 having type "L" cytology is 1.0253.

20. In the analysis of a one-dimensional contingency table using Poisson regression, which is the null model?

(1) The model assuming independence of the cells.

(2) The model assuming the cell counts have Poisson distribution.

(3) The model with no restrictions on the cell probabilities.

(4) The model assuming the cell counts have multinomial distribution.

(5) The model with all cell probabilities equal.

21. In the "death by falling" data studied in class, classifying 16,976 fatal falls by month of occurrence, we can fit the model using the code

```
deaths = c(1688, 1407, 1370, 1309, 1341, 1388, 1406,
          1446, 1322, 1363, 1410, 1526)
months = c("Jan","Feb","Mar","Apr","May","Jun","Jul",
         "Aug","Sep","Oct","Nov","Dec")
Months = factor(months, levels=months)
deaths.glm=glm(deaths~Months, family=poisson)
```

The null deviance is 81.095 on 11 degrees of freedom, and the probability that a Chi-squared variable having 11 degrees of freedom exceeds 81.095 is effectively 0. The factor `Months` has baseline "Jan". What is the **CORRECT** interpretation?

(1) The fitted intercept is 1688.

(2) The large null deviance says the Poisson regression model fits well.

(3) The maximal model deviance is 81.095.

(4) The hypothesis that falls are equally likely to fall in any month is not supported by the data.

(5) The p-value is effectively one.

22. In tables of numbers, such as mathematical tables or stock market prices, it seems reasonable to assume that the leading non-zero digit would be equally likely to be one of 1, 2, 3, ..., 9. However, often this is not true, and low digits tend to be more frequent than the higher digits. It has been suggested that the distribution of digits follows Benford's Law, which states that the probability that a leading non-zero digit is i is $\log_{10}(i/(i+1)), i = 1, 2, \ldots, 9$. The data below classify 2524 stock market returns according to their leading digits.

```
leading digit:    1   2   3   4   5   6   7   8   9
count:          735 432 273 266 200 175 169 148 126
```

We want to check if these data follow Benford's law. We get the following R output (one line has been blanked out with *****'s)

```
> Return = c(735, 432, 273, 266, 200, 175, 169, 148, 126)
> Digits=1:9
> bed.probs = log10((1+Digits)/Digits)
> Return.props = Return/sum(Return)
> equal.props=1/9
> L1 = sum(Return*log(Return.props))
> L2 = sum(Return*log(bed.probs))
> L3 = sum(Return*log(equal.props))
> L1
[1] -5103.534
> L2
[1] -5111.564
> L3
[1] -5545.795
> G2 = ******* # a line calculating the residual deviance
> 1-pchisq(G2,8)
[1] 0.04153191
> 1-pchisq(G2,9)
[1] 0.06564292
```

Which of the following is **FALSE**?

(1) The residual deviance for the saturated (maximal) model is 8.03.

(2) As expected, the uniform model (all digits equally likely) model does not fit the data well.

(3) The Benford model fits better than the uniform model.

(4) The test statistic for testing the hypothesis that the data follow Benford's law has value 16.06.

(5) The test statistic for testing the hypothesis that the true distribution is uniform (i.e. the probabilities of digits 1,2,...,9 are all the same) is 884.522.

23. The table below is the result of classifying a group of patients with abnormal cervical smear results according to their age group (< 20, 20-24, 25-29, 30-39, 40-49, 50-59, 60-64 ,65+) and their abnormality type (A: severe dyskaryosis, B: mild or moderate dyskaryosis or C: borderline change).

```
  <20 20-24 25-29 30-39 40-49 50-59 60-64 65+
A   6    77   313   940   412   201    84 175
B  54   300   370   510   199    63    10   3
C 139   533   497   792   475   225    47  66
```

The following output was obtained:

```
> counts = c(6,54,139,77,300,533,313,370,497,940,510,792,412,
+    199,475,201,63,225,84,10,47,175,3,66)
> grade = factor(rep(c("A","B","C"), 8))
> age.group = factor(rep(c("<20","20-24","25-29","30-39","40-49",
+    "50-59","60-64","65+"), each=3)))
> cancer.glm = glm(counts ~ grade*age.group, family=poisson)
> anova(cancer.glm)
Analysis of Deviance Table
               Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                              23     5387.9
grade           2    380.6        21     5007.2 < 2.2e-16 ***
age.group       7   4149.0        14      858.3 < 2.2e-16 ***
grade:age.group 14   858.3         0        0.0 < 2.2e-16 ***
```

Which of the following is **TRUE**?

(1) There is no evidence of interaction, the residual deviance is zero.

(2) The second line of the anova table is comparing the additive model to the full model.

(3) The first line of the anova table is testing that the intercept is zero.

(4) There is strong evidence that the grade of cancer depends on the age group.

(5) The deviance of the full model is 858.3.

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

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.7918     0.4082   4.389 1.14e-05 ***
gradeB                  2.1972     0.4303   5.106 3.29e-07 ***
gradeC                  3.1427     0.4170   7.537 4.81e-14 ***
age.group20-24          2.5520     0.4239   6.021 1.73e-09 ***
age.group25-29          3.9544     0.4121   9.595  < 2e-16 ***
.....................................................
gradeB:age.group20-24  -0.8372     0.4489  -1.865  0.06216 .
gradeC:age.group20-24  -1.2080     0.4344  -2.781  0.00542 **
.....................................................
```

Which of the following is **FALSE**?

(1) The fitted mean corresponding to grade A and age group 20-24 is 77.

(2) The odds ratio corresponding to grade A and age group 20-24 is 1 by definition.

(3) The residuals from the model are all zero.

(4) To 4 decimal places, a 95% confidence interval for the odds ratio corresponding to grade B and age group 20-24 is [-1.7170, 0.0426].

(5) To 4 decimal places, a 95% confidence interval for the log-odds ratio corresponding to grade C and age group 20-24 is [-2.0594 -0.3566].

25. In a three-dimensional contingency table with factors A, B and C, we want to test the hypothesis that factor A is conditionally independent of the factor B, given the value of 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.

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

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

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

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

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

SECTION B

26. (a) What do we mean by "overfitting" a statistical model? What is the consequence of overfitting? [4 marks]

(b) We discussed two different approaches to variable selection, namely all possible regressions (APR) and stepwise regression. Briefly describe the difference between them. In APR, we use different criteria of "model goodness", including the adjusted R^2, AIC and BIC. Describe these three criteria and briefly discuss the motivation behind each of them. [6 marks]

(c) What do we mean by cross validation? How can we apply this to select a subset of regression variables? [6 marks]

(d) The output below uses atmospheric data for 41 US cities. For each city the variables measured were

SO2 : SO_2 content of air in mcg/cubic metre, (the response),

temp : Average annual temperature, degrees F,

mfgfirms : number of manufacturing firms employing at least 20 people,

popn : population, in 000's,

wind : average annual wind speed,in mph,

precip :average annual rainfall, inches.

It is desired to build a model using `SO2` as the response and some of the other variables as explanatory variables. Part of some APR output is shown below. What model is indicated by this output? Give a reason. [4 marks]

```
        CV temp mfgfirms popn wind precip raindays
1 1354.228    0        1    0    0      0        0
2 1025.272    0        1    1    0      0        0
3 1010.102    0        1    1    0      0        1
4 1030.473    1        1    1    0      1        0
5  973.463    1        1    1    1      1        0
6 1036.259    1        1    1    1      1        1
```

27. (a) In our discussion of logistic regression, we focused on two different types of residual, deviance residuals and Poisson residuals. Give a brief definition of each. [6 marks]

(b) Under what circumstances is a plot of deviance residuals versus the fitted values not a useful diagnostic tool? [4 marks]

(c) Suppose that we have two logistic regression models, Model 1 and Model 2. Model 2 is a submodel of Model 1. Model 2 has a deviance of 30.4 based on 30 degrees of freedom, while Model 1 has a deviance of 21.7 based on 23 degrees of freedom. How could we test if Model 2 is adequate? What would the value of the test statistic be? [6 marks]

(d) The data for this part relate to 94 Broadway shows. The Broadway equivalent of the Oscars are the Tony awards, of which there are 6 major awards every year. The response variable `Tony.awards` is the number awards won, (out of a possible 6) won by a particular show in a particular year. The covariates are

Revival : Is the show a revival (Yes/No),

NYT : The New York Times rating, a continuous variable,

Show.run : How long the show has been running for, either < 6 months, 6-12 months, 1-2 years, ≥ 3 years.

Some output is shown below. Based on this on this, do you think that whether or not a show is a revival has an effect on the number of Tony awards won? [4 marks]

```
> broadway.glm = glm(cbind(Tony.awards, 6-Tony.awards)~
+     Revival*Show.run*NYT , family=binomial, data=broadway.df)
> summary(broadway.glm)
Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -5.67443    1.92067  -2.954  0.00313 **
RevivalYes                        0.97996    2.29988   0.426  0.67004
Show.run1-2yrs                    2.56071    2.24518   1.141  0.25406
Show.run6-12mo                   -7.80426    8.99420  -0.868  0.38556
Show.run>=3 yrs                 -30.99827 5646.02210  -0.005  0.99562
NYT                               0.35284    0.59582   0.592  0.55372
RevivalYes:Show.run1-2yrs        55.40951 8468.96589   0.007  0.99478
RevivalYes:Show.run6-12mo       -20.17707   15.74335  -1.282  0.19997
RevivalYes:Show.run>=3 yrs       36.38589 5646.02367   0.006  0.99486
RevivalYes:NYT                    0.12360    0.67558   0.183  0.85483
Show.run1-2yrs:NYT                0.09575    0.68554   0.140  0.88892
Show.run6-12mo:NYT                2.20427    1.91135   1.153  0.24881
Show.run>=3 yrs:NYT              10.44104 1693.80839   0.006  0.99508
RevivalYes:Show.run1-2yrs:NYT   -27.55664 4234.48266  -0.007  0.99481
RevivalYes:Show.run6-12mo:NYT     4.04766    3.31324   1.222  0.22184
RevivalYes:Show.run>=3 yrs:NYT  -10.91748 1693.80867  -0.006  0.99486

Null deviance: 159.39  on 91  degrees of freedom
Residual deviance:  64.96  on 76  degrees of freedom
AIC: 137.81
```

```
> broadway2.glm = glm(cbind(Tony.awards, 6-Tony.awards)~
+     Show.run*NYT.rating, family=binomial, data=broadway.df)
> summary(broadway2.glm)
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -5.5438     1.0451  -5.305 1.13e-07 ***
Show.run1-2yrs        2.8791     1.4549   1.979   0.0478 *
Show.run6-12mo      -14.4414     6.9262  -2.085   0.0371 *
Show.run>=3 yrs       1.1528     2.2451   0.513   0.6076
NYT                   0.5855     0.2729   2.145   0.0319 *
Show.run1-2yrs:NYT   -0.2581     0.4185  -0.617   0.5374
Show.run6-12mo:NYT    3.3866     1.4440   2.345   0.0190 *
Show.run>=3 yrs:NYT   0.5541     0.5913   0.937   0.3487

Null deviance: 159.39  on 91  degrees of freedom
Residual deviance:  77.86  on 84  degrees of freedom
AIC: 134.71
> pchisq(12.9,8)
[1] 0.884663
> pchisq(64.96,76)
[1] 0.1872378
> pchisq(77.39,84)
[1] 0.3183954
```

28. (a) What is an association graph? How is it used to represent the models you can fit to a 3-way contingency table? [4 marks]

(b) Describe two methods for choosing a model to fit to a 3-dimensional contingency table. [4 marks]

(c) The data in the table below come from an Australian study that classified 211,105 births according to the following factors:

agegroup : the age of the baby at birth, one of ≤ 24 weeks, 25-28 weeks, 29-32 weeks, 33-36 weeks, 37-41 weeks. Baseline is ≤ 24 weeks.

stillbirth : Stillbirth indicator, either Still or Live. Baseline is Still.

gender : the baby's gender, Female or Male. Baseline is Female.

	Female		Male	
agegroup	Still	Live	Still	Live
≤ 24 weeks	167	107	171	121
25-28 weeks	100	314	109	358
29-32 weeks	78	727	95	944
33-36 weeks	92	422	112	5155
37-41 weeks	209	96,077	169	101,776

Below are the results of fitting several models to these data. Which model fits best? Give an explanation in practical terms of what your chosen model means (e.g. that being stillborn is independent of age and gender). Draw the association graph for your chosen model. [4 marks]

```
> counts = c(167, 100, 78, 92, 209, 107, 314, 727, 4224, 96077,
   171, 109, 95, 112, 169, 121, 358, 944, 5155, 101776)
> stillbirth = factor(rep(c("Still", "Live","Still", "Live"), c(5,5,5,5)),
+   levels=c("Still", "Live")),
> gender=factor(rep(c("Female", "Male"), c(10,10)))
> agegroup = factor(rep(c("<=24 weeks","25-28","29-32","33-36","37-41"), 4),
+   levels=c("<=24 weeks","25-28","29-32","33-36","37-41"))
> stillbirth.glm = glm(counts~agegroup*stillbirth*gender,family=poisson)
> anova(stillbirth.glm, test="Chi")
Analysis of Deviance Table
                          Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                         19     844077
agegroup                   4   561466        15     282610 < 2.2e-16 ***
stillbirth                 1   276807        14       5803 < 2.2e-16 ***
gender                     1      227        13       5576 < 2.2e-16 ***
agegroup:stillbirth        4     5506         9         71 < 2.2e-16 ***
agegroup:gender            4       63         5          7 6.494e-13 ***
stillbirth:gender          1        4         4          3   0.03837 *
agegroup:stillbirth:gender 4        3         0          0   0.53593
```

```
> AIC(glm(counts~agegroup*stillbirth*gender,family=poisson))
[1] 199.3927
> AIC(glm(counts~agegroup*stillbirth*gender-agegroup:stillbirth:gender,
        family=poisson))
[1] 194.5250
> AIC(glm(counts~agegroup*stillbirth+agegroup*gender,family=poisson))
[1] 196.8134
> AIC(glm(counts~agegroup*stillbirth+gender*stillbirth,family=poisson))
[1] 253.08
> AIC(glm(counts~agegroup*gender+gender*stillbirth,family=poisson))
[1] 5695.95
> AIC(glm(counts~agegroup*stillbirth+agegroup,family=poisson))
[1] 476.454
> AIC(glm(counts~agegroup*stillbirth+gender,family=poisson))
[1] 251.9043
> AIC(glm(counts~agegroup*gender+stillbirth,family=poisson))
[1] 5694.775
> AIC(glm(counts~agegroup+gender+stillbirth,family=poisson))
[1] 5749.866
```

(d) The result of fitting the full (maximal) model is shown below. Give a point estimate of the conditional odds ratio between gender and being stillborn for age group 37-41. Repeat for the other age groups. Give an interpretation of the odds ratios in terms of the relationship between live births and gender. [4 marks]

```
Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                              5.11799    0.07738  66.139  < 2e-16 ***
agegroup25-28                           -0.51282    0.12644  -4.056 5.00e-05 ***
agegroup29-32                           -0.76128    0.13714  -5.551 2.84e-08 ***
agegroup33-36                           -0.59621    0.12984  -4.592 4.39e-06 ***
agegroup37-41                            0.22434    0.10379   2.161 0.030661 *
stillbirthLive                          -0.44516    0.12383  -3.595 0.000324 ***
genderMale                               0.02367    0.10879   0.218 0.827767
agegroup25-28:stillbirthLive             1.58939    0.16887   9.412  < 2e-16 ***
agegroup29-32:stillbirthLive             2.67738    0.17184  15.580  < 2e-16 ***
agegroup33-36:stillbirthLive             4.27191    0.16260  26.272  < 2e-16 ***
agegroup37-41:stillbirthLive             6.57574    0.14188  46.348  < 2e-16 ***
agegroup25-28:genderMale                 0.06251    0.17610   0.355 0.722618
agegroup29-32:genderMale                 0.17350    0.18757   0.925 0.354979
agegroup33-36:genderMale                 0.17304    0.17786   0.973 0.330601
agegroup37-41:genderMale                -0.23611    0.15013  -1.573 0.115785
stillbirthLive:genderMale                0.09929    0.17160   0.579 0.562840
agegroup25-28:stillbirthLive:genderMale -0.05433    0.23366  -0.233 0.816140
agegroup29-32:stillbirthLive:genderMale -0.03526    0.23501  -0.150 0.880733
agegroup33-36:stillbirthLive:genderMale -0.09682    0.22288  -0.434 0.664001
agegroup37-41:stillbirthLive:genderMale  0.17077    0.20042   0.852 0.394187
```

(e) Can you think of a better approach to estimating this conditional odds ratio? Write down a line or two of R-code that would produce the necessary output, and explain how you would get the estimate. [4 marks]

ANSWER SHEET FOLLOWS

Surname: _____ First Names: _____ ID No: _____