THE UNIVERSITY OF AUCKLAND

SECOND SEMESTER, 2011
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

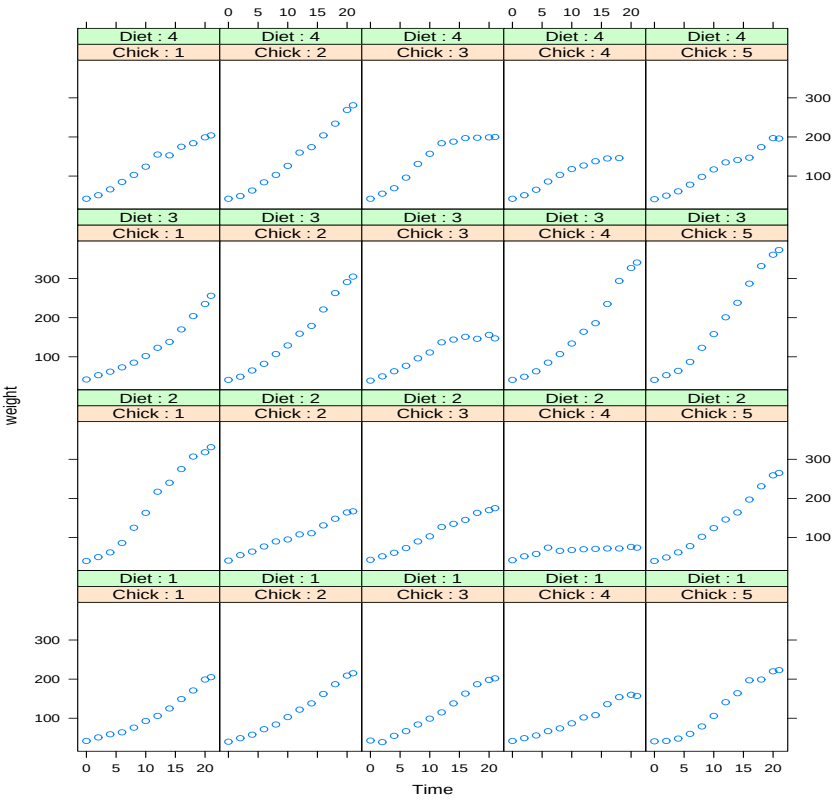

Figure 1: Trellis plot for Question 1.

1. The data for this question are taken from an experiment to monitor the growth of 20 chicks fed one of four different diets. The variables are

   Weight: The weight of the chicken in grams,

   Time: The age of the chick in days,

   Chick: A factor containing an identifier for the chick, taking values 1,2,3,4,5 within each diet,

   Diet: A factor with levels 1,...,4 indicating which experimental diet the chick received.

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

   (1) All the chicks are growing at the same rate.

   (2) The biggest chick weighed approximately 400 grams after 20 days.

   (3) Some chicks stopped growing at about 10 days old.

   (4) The smallest chick weighed less than 100 grams after 20 days.

   (5) All the chicks had approximately the same weight at birth.

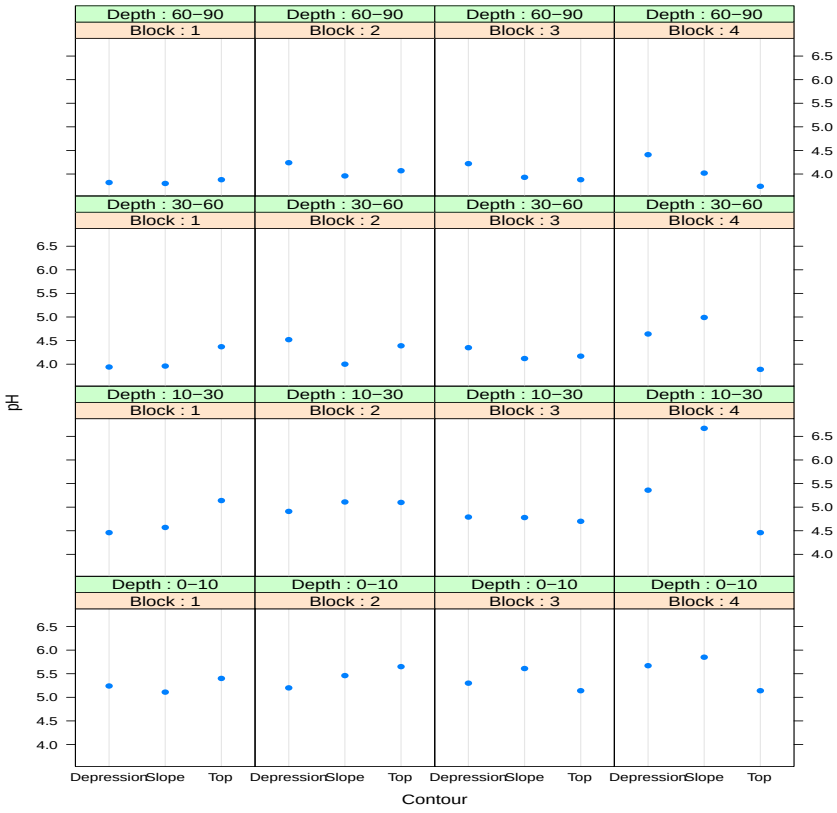

Figure 2: Trellis plot for Question 2.

2. The data for Question 2 come from an experiment relating the pH of soil in a agricultural station to various other factors. In the experiment, 48 soil samples were taken. The variables are

   Contour: The topography of the ground, a factor with 3 levels: Depression, Slope, Top.

   Depth: Depth of the soil sample, a factor with 4 levels: 0-10cm, 10-30cm, 30-60cm, 60-90cm

   Block: A factor with levels 1, 2, 3, 4 giving the four locations where samples were taken from.

   pH: The pH of the sample.

   A trellis plot of the data is shown in Figure 2. Which of the following is **FALSE**?

   (1) Contour does not seem to have a clear systematic effect on pH.

   (2) Ph increases with increasing depth.

   (3) At each depth, Block 4 had the most variable pH readings.

   (4) In block 4, the top contour has the lowest pH at each depth.

   (5) In block 2, the highest pH is a little more than 5.5.

3. Which of the following bits of R code produced Figure 2?

   (1) `dotplot(pH∼Contour|Block*Depth, data=Soils,xlab = "Contour", strip=function(...)strip.default(..., strip.names=TRUE))`

   (2) `dotplot(Block∼Depth|Contour*pH, data=Soils,xlab = "Contour", strip=function(...)strip.default(..., strip.names=TRUE))`

   (3) `xyplot(pH∼Contour|Block*Depth, data=Soils,xlab = "Contour", strip=function(...)strip.default(..., strip.names=TRUE))`

   (4) `bwplot(pH∼Contour|Block*Depth, data=Soils,xlab = "Contour", strip=function(...)strip.default(..., strip.names=TRUE))`

   (5) `dotplot(pH∼Block|Coutour*Depth, data=Soils,xlab = "Contour", strip=function(...)strip.default(..., strip.names=TRUE))`

4. Which of the following would **NOT** enhance the plot in Figure 1?

   (1) Adding units to the axis labels.

   (2) Colour-coding the different chicks with different colours.

   (3) Adding a loess smoother to each panel.

   (4) Joining up the points with lines.

   (5) Adding a least squares line to each panel.

5. Which of the following statements is **FALSE**?

   (1) If the R^2 in a regression equals zero,all the estimated regression coefficients except the intercept are zero.

   (2) If the R^2 in a regression equals one, all the points lie on the fitted plane.

   (3) In a regression, adding variables cannot decrease the R^2.

   (4) If the R^2 in a regression is low, the model must be incorrect.

   (5) In a regression, the R^2 is never less than the adjusted R^2.

6. In a regression where observations are taken sequentially in time, which of the following is **NOT** useful in diagnosing lack of independence?

   (1) An ACF plot.

   (2) A plot of residuals versus the previous residual.

   (3) The Durbin-Watson statistic.

   (4) A plot of residuals versus fitted values.

   (5) A time series plot.

7. The data for this question are taken from the Los Angeles Heart Study. The following variables were measured on 60 men:

   wt: Weight in pounds.

   age: Age in years.

   sbp: Systolic blood pressure in mm of mercury.

   chl: Cholesterol level in mg per dl.

   ht: Height in inches.

   A regression model with `sbp` as response was fitted with the following results:

```
Call:
lm(formula = sbp ~ age + chl + ht + wt, data = heartstudy.df)
Coefficients:
            Estimate  Std. Error t value Pr(>|t|)
(Intercept) 44.58519    66.09363   0.675  0.50277
age          0.53218     0.19459   2.735  0.00838 **
chl          0.03285     0.03508   0.936  0.35318
ht           0.13451     0.96248   0.140  0.88936
wt           0.20222     0.09360   2.160  0.03512 *
Residual standard error: 15.41 on 55 degrees of freedom
Multiple R-Squared: 0.2265, Adjusted R-squared: 0.1702
F-statistic: 4.026 on 4 and 55 DF, p-value: 0.006194
```

   Which of the following is **FALSE**?

   (1) If all other variables are held constant, blood pressure tends to go up with increasing age.

   (2) This fitted model could be used to make a precise prediction of systolic blood pressure.

   (3) The output indicates that cholesterol could be deleted from the model.

   (4) The estimate of the error variance is about 237.47.

   (5) The p-value of 0.006194 arises when comparing the model fitted above to the null model.

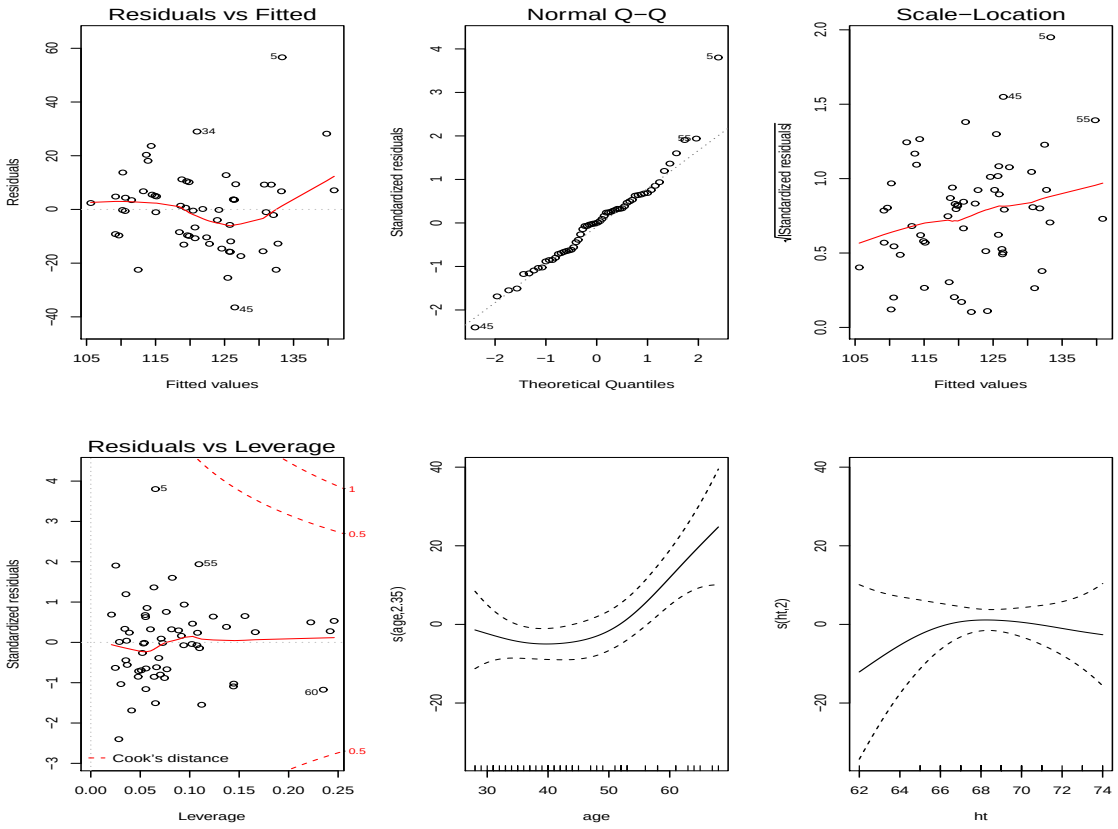

Figure 3: Diagnostic plots for Question 8.

8. Some diagnostic plots from fitting the regression are shown in Figure 3. Which of the following is **FALSE**?

   (1) There is a hint that the variances might increase with the mean.

   (2) There is a hint that the data are not planar.

   (3) There is one large standardised residual.

   (4) Some points are having an undue effect on the regression coefficients.

   (5) Transforming age and height might improve the fit.

9. The display below gives the results of an all possible regressions run on the data from Q7.

```
> all.poss.regs(formula = sbp ~ age + chl + ht + wt, data = heartstudy.df)
      rssp  sigma2 adjRsq    Cp    AIC    BIC       CV age chl ht wt
1 14595.41 251.645  0.121 5.441 65.441 69.629 1611.909   1   0  0  0
2 13276.01 232.913  0.186 1.886 61.886 68.169 1508.040   1   0  0  1
3 13070.08 233.394  0.185 3.020 63.020 71.397 1506.301   1   1  0  1
4 13065.44 237.553  0.170 5.000 65.000 75.472 1535.228   1   1  1  1
```

   Which of the following statements is **TRUE**?

(1) AIC and BIC indicate different models should be chosen.

(2) The best model is `sbp`$\sim$`age + chl + ht + wt`.

(3) The `sigma2` criterion indicates that the best model is `sbp`$\sim$`age`.

(4) The residual sum of squares is the best criterion to use.

(5) The model indicated by the CV criterion is `sbp`$\sim$`age + chl + wt`

10. In the course we discussed the concept of collinearity. Which of the following statements concerning a regression of a continuous response variable Y on two continuous explanatory variables X and Z is **FALSE**?

(1) The variance inflation factor for X is $1/(1-r^2)$ where r is the correlation between X and Z.

(2) If X and Z are highly correlated, they are likely to have non-significant p-values in the regression summary.

(3) The bigger the error variance, the bigger the standard errors of the regression coefficients.

(4) The variance inflation factors are the diagonal elements of the inverse of the correlation matrix of X and Z.

(5) The bigger the correlation between X and Z, the smaller the standard errors of the regression coefficients.

11. Mr Derek Whiteside of the UK Building Research Station recorded the weekly gas consumption and average external temperature at his own house in south-east England for two heating seasons, one of 26 weeks before, and one of 30 weeks after cavity-wall insulation was installed. The object of the exercise was to assess the effect of the insulation on gas consumption. The variables are

Insul: A factor, before or after insulation,

Temp: Average weekly minimum temperature, in degrees C,

Gas: The weekly gas consumption in 1000's of cubic feet.

A model was fitted and the following R output obtained.

```
Call:
lm(formula = Gas ~ Temp * Insul, data = whiteside)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.85383    0.13596  50.409  < 2e-16 ***
Temp            -0.39324    0.02249 -17.487  < 2e-16 ***
InsulAfter      -2.12998    0.18009 -11.827 2.32e-16 ***
Temp:InsulAfter  0.11530    0.03211   3.591  0.00073 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 0.323 on 52 degrees of freedom
Multiple R-squared: 0.9277,     Adjusted R-squared: 0.9235
F-statistic: 222.3 on 3 and 52 DF,  p-value: < 2.2e-16
```

Which of the following is **FALSE**?

(1) There is clear evidence that two lines are required.

(2) The slope of the "after insulation" line is less than the slope of the "before insulation" line.

(3) The slope of the "before insulation" line is -0.39324.

(4) There is clear evidence that two lines are not parallel.

(5) The intercept of the "after insulation" is 4.72385.

12. Use the information below to select the **CORRECT** answer.

```
> insul = lm(Gas~Temp*Insul, data=whiteside)
> predict(insul, newdata = data.frame(Temp=6, Insul="After"), se.fit=TRUE)
$fit
      1
3.05624
$se.fit
[1] 0.06869251
$df
[1] 52
$residual.scale
[1] 0.3230042
> qt(0.975,52)
[1] 2.006647
> qt(0.975,4)
[1] 2.776445
```

(1) To the nearest whole number, a 95% prediction interval for gas consumption after insulation when the outside temperature is 6 degrees is (2408, 3704) cubic feet.

(2) The estimate of error variance is 0.3230042.

(3) To the nearest whole number, a 95% prediction interval for gas consumption after insulation when the outside temperature is 6 degrees is (2918, 3194)cubic feet.

(4) To the nearest whole number, a 95% prediction interval for gas consumption after insulation when the outside temperature is 6 degrees is $(2394, 3719)$ cubic feet.

(5) The estimated prediction error is 0.06869251.

13. Based on the output below, which of the following statements is **FALSE**?

```
> anova(insul)
Analysis of Variance Table

Response: Gas
           Df Sum Sq Mean Sq F value    Pr(>F)
Temp        1 35.019  35.019 335.655 < 2.2e-16 ***
Insul       1 33.224  33.224 318.451 < 2.2e-16 ***
Temp:Insul  1  1.345   1.345  12.893 0.0007307 ***
Residuals  52  5.425   0.104
```

(1) The F-value 12.893 is testing if the interaction is zero.

(2) The F-value 335.655 is comparing the "single line" and null models.

(3) The F-value 318.451 is comparing the "parallel line" and "non-parallel line" models.

(4) The figure 0.104 is the estimate of the error variance.

(5) The R-code `anova(lm(formula = Gas ~Insul*Temp, data = whiteside))` would have produced different output.

14. Using the data for Q2, and ignoring the variable `contour`, we got the following output: (Note that one coefficient has been replaced by *******)

```
Call:
lm(formula = pH ~ Depth * Block, data = Soils)


Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.25000    0.21795  24.088  < 2e-16 ***
Depth10-30        -0.52667    0.30822  -1.709 0.097187 .
Depth30-60         *******    0.30822  -3.764 0.000677 ***
Depth60-90        -1.41667    0.30822  -4.596  6.4e-05 ***
Block2             0.18667    0.30822   0.606 0.549037
Block3             0.10000    0.30822   0.324 0.747716
Block4             0.30333    0.30822   0.984 0.332431
Depth10-30:Block2  0.13000    0.43589   0.298 0.767449
Depth30-60:Block2  0.02667    0.43589   0.061 0.951599
Depth60-90:Block2  0.07000    0.43589   0.161 0.873426
Depth10-30:Block3 -0.06667    0.43589  -0.153 0.879404
Depth30-60:Block3  0.02333    0.43589   0.054 0.957642
Depth60-90:Block3  0.07667    0.43589   0.176 0.861493
Depth10-30:Block4  0.47000    0.43589   1.078 0.288985
Depth30-60:Block4  0.11333    0.43589   0.260 0.796527
Depth60-90:Block4 -0.08000    0.43589  -0.184 0.855539
```

Note that there were 3 soil samples taken for each depth/block combination. Which of the following is **FALSE**?

(1) The mean of the observations in block 4 where the depth is 60-90 is 4.05667.

(2) The mean of the observations in block 4 where the depth is 0-10 is 5.55333.

(3) The mean of the observations in block 3 where the depth is 10-30 is 4.75666.

(4) The mean of the observations in block 1 where the depth is 0-10 is 6.66667.

(5) The mean of the observations in block 1 where the depth is 60-90 is 3.83333.

15. The mean of the three observations in block 1 where the depth is 30-60 is 4.090000. What is the main effect of Depth corresponding to level 30-60?

(1) 9.34000.

(2) 1.94334.

(3) -0.52667.

(4) -1.16000.

(5) 4.090000.

16. Suppose we have a regression with a response variable Y, two continuous variables X and Z, and categorical variables A and B. Which of the following terms should **NOT** appear in a model for these data?

(1) $A * B$.

(2) $X * A$.

(3) $X * B$.

(4) $X * Z$.

(5) $A * Z$.

17. Suppose we have a logistic regression model logit $P(Y = 1) = \alpha + \beta x$ with a single explanatory variable x and a response Y. The estimated regression coefficients are $\hat{\alpha} = -3.25$ and $\hat{\beta} = 0.5$. Which of the following statements is **TRUE**?

(1) To three decimal places, the predicted probability that $Y = 1$ when $x = 10$ is 0.148.

(2) To three decimal places, the predicted probability that $Y = 0$ when $x = 0$ is 0.037.

(3) To three decimal places, the predicted odds that $Y = 1$ when $x = 10$ is 0.174.

(4) To three decimal places, the predicted probability that $Y = 1$ when $x = 10$ is 0.852.

(5) To three decimal places, the predicted log-odds that $Y = 1$ when $x = 10$ is -1.750.

18. The data for this question come from the Panel Study of Income Dynamics (PSID). One of the objectives of the study was to understand the factors having a bearing on women's participation in the labour force. Each of the 753 observations relates to a married woman. The variables are

    lfp: Labour-force participation; a factor with levels: "no", "yes.

    k5: Number of children 5 years old or younger.

    k618: Number of children 6 to 18 years old.

    age: Age in years.

    wc: Wife's college attendance; a factor with levels: "no", "yes.

    hc: Husband's college attendance; a factor with levels: "no", "yes.

    lwg: Log expected wage rate; for women in the labor force, the actual wage rate; for women not in the labor force, an imputed value based on the regression of lwg on the other variables.

    inc: Family income exclusive of wife's income.

    The following output was obtained:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.182140   0.644375   4.938 7.88e-07 ***
k5          -1.462913   0.197001  -7.426 1.12e-13 ***
k618        -0.064571   0.068001  -0.950 0.342337
age         -0.062871   0.012783  -4.918 8.73e-07 ***
wcyes        0.807274   0.229980   3.510 0.000448 ***
hcyes        0.111734   0.206040   0.542 0.587618
lwg          0.604693   0.150818   4.009 6.09e-05 ***
inc         -0.034446   0.008208  -4.196 2.71e-05 ***
```

    Which of the following is **NOT** a correct interpretation?

    (1) More educated women have a higher probability of participating in the labour force.

    (2) Women with more earning potential have a higher probability of participating in the labour force.

    (3) Women from poorer homes have a higher probability of participating in the labour force.

    (4) Women with young children have a lower probability of participating in the labour force.

    (5) Older women have a higher probability of participating in the labour force.

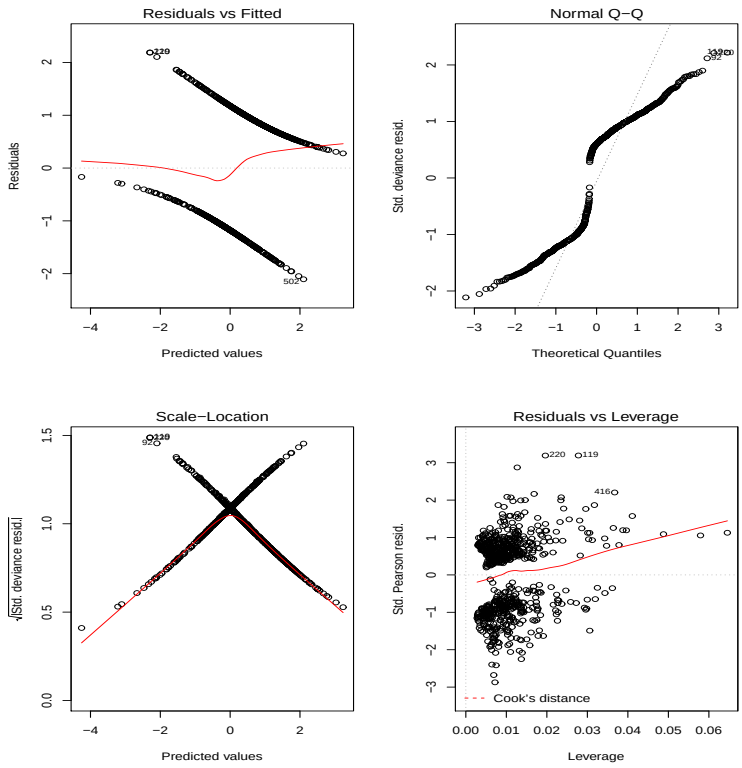

Figure 4: Diagnostic plots for Question 19.

19. Some diagnostics from fitting this model are shown in Figure 4. Which is the **CORRECT** interpretation of these diagnostics?

    (1) None of the points is having an undue effect on the estimated regression coefficients.

    (2) The strange shape of the residual versus fitted value plot shows that something is seriously wrong with the regression.

    (3) The increasing pattern in the scale-location plot shows that the data are over-dispersed.

    (4) The normal plot shows that the normality assumption is violated.

    (5) There are no high-leverage points in the data.

20. When using logistic regression to predict a future observation, with"success" corresponding to $Y = 1$, which if the following is **CORRECT**?

    (1) Sensitivity is the probability of correctly predicting a failure.

    (2) The use of the training set to calculate the error rate usually over-estimates the error.

    (3) Leave-one-out cross-validation is better than 10-fold cross-validation.

    (4) The ROC curve is a plot of sensitivity versus 1-specificity.

    (5) Specificity is the probability of incorrectly predicting a success.

21. The analysis below refers to a cross-classification of 1999 patients by hypertension status (variable `hyper`, with levels "yes" and "no") and the average number of daily alcoholic drinks consumed (variable `alcohol`, with levels 0, 1-2, 3-5 and 6+). The following output was obtained:

```
> two.way = glm(count~hyper*alcohol, family=poisson, data=two.way.df)
> anova(two.way, test="Chisq")
Analysis of Deviance Table
Model: poisson, link: log
Response: count
Terms added sequentially (first to last)
              Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                             7     476.56
hyper          1   434.83         6      41.72 < 2.2e-16 ***
alcohol        3     7.60         3      34.12   0.05505 .
hyper:alcohol  3    34.12         0       0.00 1.865e-07 ***
```

Which of the following statements is **CORRECT**?

(1) The 'alcohol" line of the anova table is testing the hypothesis that an additive model is appropriate.

(2) The variable alcohol is not required in the model.

(3) The "NULL" line of the anova table is testing the hypothesis that the constant term is zero.

(4) The fact that the last line of the anova table has zero degrees of freedom means that the model is not appropriate.

(5) There is an association between hypertension and alcohol consumption.

22. The summary table from fitting the model above is

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)          4.4659     0.1072  41.655  < 2e-16 ***
hyperNo              1.5054     0.1185  12.702  < 2e-16 ***
alcohol1-2           0.3544     0.1399   2.534   0.0113 *
alcohol3-5           0.5839     0.1338   4.364 1.28e-05 ***
alcohol6+            0.6989     0.1312   5.328 9.95e-08 ***
hyperNo:alcohol1-2  -0.4967     0.1583  -3.138   0.0017 **
hyperNo:alcohol3-5  -0.5942     0.1518  -3.915 9.03e-05 ***
hyperNo:alcohol6+   -0.8501     0.1508  -5.639 1.71e-08 ***
```

A 95% confidence interval for the odds ratio corresponding to having 1 to 2 drinks and no hypertension is, to four decimal places: (Take the normal percentage point for a 95% confidence interval to be 1.960)

(1) (0.4462, 0.8299).

(2) (-0.8069, -0.1865).

(3) (1.0836, 1.8748).

(4) (0.0803, 0.6285).

(5) (3.5718, 5.6839).

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

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

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

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

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

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

24. In fact, the table in Q21 was a marginal table obtained from a three-dimensional table cross-classifying `hyper` and `alcohol` with a third factor `obesity`, with levels "Low", "Av" and "High". Some R output from fitting the model `count~hyper*alcohol*obesity` is shown below:

```
> anova(three.way,test="Chisq")

                     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                    23     549.56
hyper                 1   434.83        22     114.73 < 2.2e-16 ***
alcohol               3     7.60        19     107.13  0.055055 .
obesity               2     0.24        17     106.89  0.886148
hyper:alcohol         3    34.12        14      72.76 1.865e-07 ***
hyper:obesity         2    44.95        12      27.81 1.731e-10 ***
alcohol:obesity       6    22.21         6       5.59  0.001107 **
hyper:alcohol:obesity 6     5.59         0       0.00  0.470325
```

Which of the following statements is supported by this analysis?

(1) The three factors alcohol, hypertension and obesity are independent.

(2) Alcohol consumption is independent of hypertension, given obesity.

(3) Alcohol consumption is independent of hypertension and obesity.

(4) The population odds ratios between alcohol and hypertension are the same at every level of obesity.

(5) Alcohol consumption is independent of obesity, given hypertension.

25. In class we looked at the flying bomb data set, which recorded the number of areas of London hit by 0, 1, 2, 3, 4 and 5+ flying bombs. The frequencies of these outcomes were 229, 211, 93, 35, 7 and 1. Study the following R output and pick the BEST interpretation.

```
> no.of.squares=c(0,1,2,3,4,5)
> no.of.hits=c(229,211,93,35,7,1)
> mean.hits=sum(no.of.squares*no.of.hits)/
+ sum(no.of.hits)
> poisson.probs=dpois(0:4,mean.hits)
> poisson.probs = c(poisson.probs, 1-sum(poisson.probs))
> rel.freqs=no.of.hits/576
> Loglik1=sum(no.of.hits*log(rel.freqs))
> Loglik2 = sum(no.of.hits*log(poisson.probs))
> Loglik3 = sum(no.of.hits*log(1/6))
> d1=2*(Loglik1-Loglik2)
> d1
[1] 1.171815
> d2=2*(Loglik1-Loglik3)
> d2
[1] 608.1803
> d3=2*(Loglik2-Loglik3)
> d3
[1] 607.0085
> 1-pchisq(d1, 4)
[1] 0.882717
> 1-pchisq(d2, 5)
[1] 0
> 1-pchisq(d3, 6)
[1] 0
```

Which of the following is **FALSE**?

(1) The residual deviance for the saturated (maximal) model is 608.1803.

(2) The p-value for testing the hypothesis that the data follow a Poisson model has value 0.882717.

(3) As expected, the uniform model (all probabilities equal) does not fit the data well.

(4) The test statistic for testing the hypothesis that the probabilities are all the same is 608.1803.

(5) The test statistic for testing the hypothesis that the data follow a Poisson model has value 1.171815.

SECTION B

26. (a) What is the difference between a high leverage point and an influential point in a regression? [4 marks]

(b) Describe the measures we use to assess the affect individual points are having on a regression. You should make clear what aspect of the regression is being captured by each measure. [6 marks]

(c) Discuss the thresholds that we apply to these methods to decide if points are having an effect on the regression. [6 marks]

(d) The display below refers to the data that was described in Question A7. What points (if any) of those below are having an effect on the regression? If any points are having an effect, what is the nature of these effects? [4 marks]

```
Influence measures of
         lm(formula = sbp ~ age + chl + ht + wt, data = heartstudy.df) :

      dfb.1_   dfb.age  dfb.chl    dfb.ht    dfb.wt    dffit cov.r   cook.d    hat
5  -0.048961  0.859794 -0.19302 -0.097636  0.359529  1.16028 0.255 2.02e-01 0.0653
16 -0.224894  0.177171  0.01618  0.220475 -0.081681  0.26416 1.378 1.42e-02 0.2224
19  0.021543  0.019696 -0.02883 -0.084210  0.284930  0.30206 1.417 1.85e-02 0.2460
38 -0.037014  0.017248  0.09368  0.029941 -0.041891  0.11228 1.307 2.57e-03 0.1663
41 -0.007067 -0.044949  0.14783 -0.013519  0.011653  0.15553 1.436 4.92e-03 0.2419
45 -0.002280 -0.240274 -0.04357  0.028266  0.011275 -0.42937 0.649 3.36e-02 0.0283
```

27. (a) In our discussion of logistic regression, we introduced the idea of deviance. Define the deviance of a model, and state under what circumstances the deviance can be used the assess the fit of the logistic model. [6 marks]

(b) Suppose the circumstances in (i) hold, and the residual deviance is 13.362 with 12 degrees of freedom. The model has 6 parameters. Using the output below, do you think the model fits well? [4 marks]

```
> 1-pchisq(13.362,12)
[1] 0.3432871
> 1-pchisq(13.362,6)
[1] 0.03763428
```

(c) Under what circumstances can we use deviances to compare a model to a sub-model? Briefly describe how this comparison is carried out. [4 marks]

(d) The data for this part are the CHD data studied in class, using the grouped form of the data. The data have a reasonable degree of grouping. The variables are

g.age : The age of the group,

n : The number of individuals in the group,

r : The number of individuals in the group having CHD.

Some output is shown below. Based on this on this, give a careful discussion of how well the logistic model fits. Do you think the variable `g.age` should be transformed? Give reasons. [6 marks]

```
Call:
glm(formula = cbind(r, n - r) ~ g.age,
  family = binomial, data = chd.group.df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.27844    1.13053  -4.669 3.03e-06 ***
g.age        0.11032    0.02402   4.593 4.36e-06 ***

Null deviance: 63.958  on 42  degrees of freedom
Residual deviance: 34.976  on 41  degrees of freedom
> 1-pchisq(34.976,41)
[1] 0.7344371
> 1-pchisq(63.958,42)
[1] 0.01607959
```

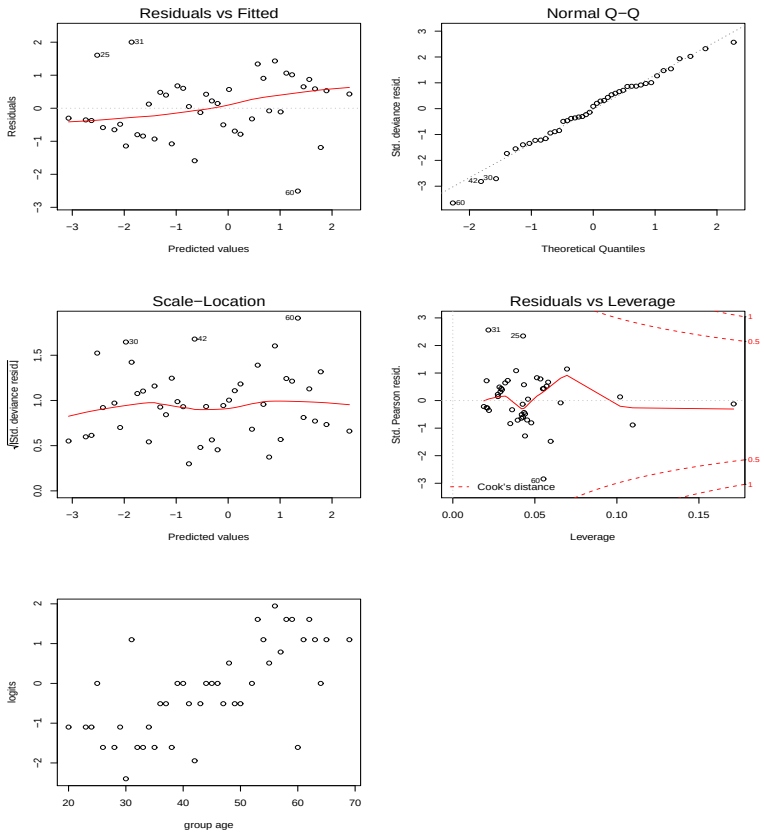

Figure 5: Diagnostic plots for Question 27(d).

28. (a) Define what it means for two factors A and B to be independent. If a number of individuals are cross-classified according to A and B into a contingency table, describe how we can test the independence by fitting a Poisson regression model to the cell counts. [5 marks]

(b) The following table below classifies 1314 women by smoking status (variable smoker with levels "yes" and "no") and 20-year survival (variable `dead`, also with levels "yes" and "no"). Use the output below to decide if smoking and survival are independent. [5 marks]

```
> xtabs(y~smoker+dead, data=femsmoke)
      dead
smoker yes  no
   yes 139 443
   no  230 502
> summary(glm(y~smoker*dead, data=femsmoke, family=poisson))

Call:
glm(formula = y ~ smoker * dead, family = poisson, data = femsmoke)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.98856    0.08482  35.235  < 2e-16 ***
smokerno         0.50361    0.10743   4.688 2.76e-06 ***
deadno           1.15910    0.09722  11.922  < 2e-16 ***
smokerno:deadno -0.37858    0.12566  -3.013  0.00259 **

Null deviance: 1193.9  on 27  degrees of freedom
Residual deviance:  906.3  on 24  degrees of freedom
AIC: 1048.5

Number of Fisher Scoring iterations: 5
```

(c) Interpret the `smokerno:deadno` interaction in terms of smoking and survival. Does this seem surprising? [5 marks]

(d) In fact the table above is a marginal table of a three dimensional table where the third factor is `age`. The three-dimensional table is

```
, , dead = yes

       smoker
age     yes  no
  18-24   2   1
  25-34   3   5
  35-44  14   7
  45-54  27  12
  55-64  51  40
  65-74  29 101
  75+    13  64
```

```
, , dead = no

       smoker
age     yes  no
  18-24  53  61
  25-34 121 152
  35-44  95 114
  45-54 103  66
  55-64  64  81
  65-74   7  28
  75+     0   0
```

The smoking and age was recorded in an initial study and the survival information recorded in a follow-up study 20 years later. An analysis of the three-dimensional table suggested that the homogeneous association model was a good fit. The summary from this model is shown below. Does this new information shed any light on your answer to part(c)? [5 marks]

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.54284    0.58736   0.924 0.355384
age25-34             0.92902    0.68381   1.359 0.174273
age35-44             1.94048    0.62486   3.105 0.001900 **
age45-54             2.76845    0.60657   4.564 5.02e-06 ***
age55-64             3.37507    0.59550   5.668 1.45e-08 ***
age65-74             2.86586    0.60894   4.706 2.52e-06 ***
age75+               2.02211    0.64955   3.113 0.001851 **
smokerno            -0.29666    0.25324  -1.171 0.241401
deadno               3.43271    0.59014   5.817 6.00e-09 ***
age25-34:smokerno    0.11752    0.22091   0.532 0.594749
age35-44:smokerno    0.01268    0.22800   0.056 0.955654
age45-54:smokerno   -0.56538    0.23585  -2.397 0.016522 *
age55-64:smokerno    0.08512    0.23573   0.361 0.718030
age65-74:smokerno    1.49088    0.30039   4.963 6.93e-07 ***
age75+:smokerno      1.89060    0.39582   4.776 1.78e-06 ***
age25-34:deadno     -0.12006    0.68655  -0.175 0.861178
age35-44:deadno     -1.34112    0.62857  -2.134 0.032874 *
age45-54:deadno     -2.11336    0.61210  -3.453 0.000555 ***
age55-64:deadno     -3.18077    0.60057  -5.296 1.18e-07 ***
age65-74:deadno     -5.08798    0.61951  -8.213  < 2e-16 ***
age75+:deadno      -27.31727 8839.01146  -0.003 0.997534
smokerno:deadno      0.42741    0.17703   2.414 0.015762 *
```

ANSWER SHEET FOLLOWS