THE UNIVERSITY OF AUCKLAND

SECOND SEMESTER, 2012
Campus: City

STATISTICS

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.
- 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

SECTION A

1. A data set consists of measurements on three variables X, Y and Z. The variables X and Y are categorical and Z is continuous. Which of the following plots would you expect to give the **best picture** of the relationship between the variables?

   (1) A trellis plot consisting of panels corresponding to values of X, and each panel containing a dot plot.

   (2) A barchart, with bars corresponding to the frequencies of X and Y.

   (3) A coplot corresponding to the formula `X ~ Y | Z`.

   (4) A scatterplot of X versus Y, with the value of Z shown by a colour coding.

   (5) A coplot corresponding to the formula `Y ~ X | Z`.

2. The data for this question come from a study involving 200 men and women who were asked to guess their height and weight. These were then compared to the actual height and weight. The resulting data are graphed in Figure 1. Lines at 45 degrees have been added.

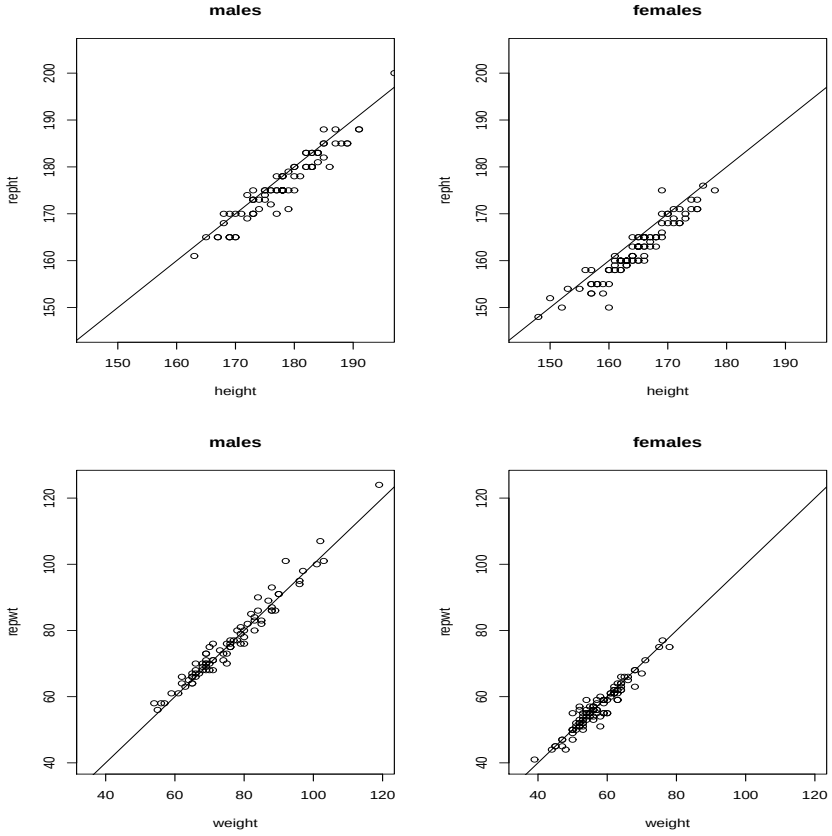

Figure 1: Plots for Question 2. Vertical axis: reported (guessed) heights and weights. Horizontal axis: actual heights and weights.

Which of the following is **FALSE**?

(1) The tallest females have approximately the same height as the average male.

(2) Males estimate their weights quite well.

(3) Females overestimate their heights.

(4) Females estimate their weights better than their heights.

(5) Males slightly underestimate their heights.

3. Which of the following alternative displays would be the **MOST** helpful for answering Question 2?

(1) Two histograms of the difference between weight and estimated height, one for each sex.

(2) A trellis plot plotting estimated weight versus weight in one panel and estimated height versus height in the other.

(3) Two histograms of the difference between weight and estimated weight, one for each sex, plus two histograms of the difference between height and estimated height, one for each sex.

(4) Two histograms of the difference between height and estimated weight, one for each sex.

(5) A coplot of estimated height versus height, conditioning on sex and weight.

4. In a regression analysis, which of the following is **FALSE**?

(1) The residual sum of squares is zero if and only if the R^2 is 1.

(2) If we use R^2 as a goodness-of-fit index, the bigger R^2 is, the better the fit.

(3) If the residual sum of squares is a small number, the fit must be good.

(4) If all of the estimated regression coefficients other than the constant term are zero, the regression sum of squares is zero.

(5) In linear regression, the "analysis of variance identity" expresses the "total sum of squares" as the sum of the "regression sum of squares" and the "residual sum of squares".

5. Which of the following plots might be useful in diagnosing possible non-independence in a regression analysis where the data were collected sequentially in time?

(1) A leverage-residual plot.

(2) A normal plot.

(3) A gam plot.

(4) A plot of residuals versus fitted values.

(5) An autocorrelation plot of residuals.

6. The data for this question are taken from a Californian hydrological study. The dataset contains 43 years worth of precipitation measurements taken at six sites in the Owens Valley (labeled APMAM, APSAB, APSLAKE, OPBPC, OPRC, and OPSLAKE), and stream runoff volume at a site near Bishop, California. Each year corresponds to an observation.

   Year: Collection year

   APRAM: Snowfall in inches at the APMAM site,

   APSAB: Snowfall in inches at the APSAB site,

   APSLAKE: Snowfall in inches at the APSLAKE site,

   OPBPC: Snowfall in inches at the OPBPC site,

   OPRC: Snowfall in inches at the OPRC site,

   OPSLAKE: Snowfall in inches at the OPSLAKE site,

   BSAAM: Stream runoff near Bishop, CA, in acre-feet.

   A regression model with `BSAAM` as response and the other variables as explanatories was fitted with the following results:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -227814.8   197920.2  -1.151  0.25752
Year            123.9      100.6   1.232  0.22621
APMAM           143.4      715.2   0.200  0.84228
APSAB          -546.0     1515.1  -0.360  0.72074
APSLAKE        1885.0     1368.1   1.378  0.17699
OPBPC            76.6      458.4   0.167  0.86827
OPRC           2081.5      650.7   3.199  0.00293 **
OPSLAKE        2055.0      758.1   2.711  0.01033 *
---
Residual standard error: 7503 on 35 degrees of freedom
Multiple R-squared: 0.928,      Adjusted R-squared: 0.9136
F-statistic:  64.4 on 7 and 35 DF,  p-value: < 2.2e-16

>  diag(solve(cor(X)))
     Year     APMAM     APSAB   APSLAKE     OPBPC      OPRC   OPSLAKE
 1.190464  3.661340  7.212829  7.120946  9.267469  7.987377 17.461359
> round(cor(X),2)
        Year APMAM APSAB APSLAKE OPBPC OPRC OPSLAKE
Year    1.00  0.00  0.05    0.17  0.12 0.02    0.14
APMAM   0.00  1.00  0.83    0.82  0.12 0.15    0.11
APSAB   0.05  0.83  1.00    0.90  0.04 0.11    0.03
APSLAKE 0.17  0.82  0.90    1.00  0.09 0.11    0.10
OPBPC   0.12  0.12  0.04    0.09  1.00 0.86    0.94
OPRC    0.02  0.15  0.11    0.11  0.86 1.00    0.92
OPSLAKE 0.14  0.11  0.03    0.10  0.94 0.92    1.00
```

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

(1) The VIF's indicate a high degree of multicolinearity in these data.

(2) If it snows heavily at the APSAB site the stream run-off at Bishop is reduced.

(3) Runoff at Bishop has been increasing.

(4) The residual sum of squares for this fit is 7503 x 35 = 262,605.

(5) Only the variables OPRC and OPSLAKE are related to the response.

7. Referring to the outputs in Question 6, which of the following is **NOT** indicated by this output?

(1) There are high correlations between APMAM, APSAB and APSLAKE.

(2) The coefficient of year is estimated quite well.

(3) Snowfall in the APMAM, APSAB, APSLAKE regions seems unrelated to snowfall in the AOPBPC OPRC OPSLAKE regions.

(4) A model to predict stream run-off should include all of the variables.

(5) The coefficient of OPSLAKE is not being estimated well.

8. A model was fitted to the data in Question 6, and stored in the R object `modelQ8`. We want to predict the stream runoff near Bishop for a new year. Prior to the thaw, we get values of the snowfall at the six sites, which are entered into a data frame `newdata`. Using the following output, which of the following is **TRUE**?

```
> predict(modelQ8, newdata, se=TRUE)
$fit
       1
59718.26
$se.fit
[1] 2695.466
$df
[1] 35
$residual.scale
[1] 7503.143
> qt(0.95, 35)
[1] 1.689572
> qt(0.975, 35)
[1] 2.030108
```

(1) A 95% prediction interval for the stream runoff is (46247.94, 73188.58).

(2) A 95% prediction interval for the stream runoff is (54246.17, 65190.35).

(3) A 95% prediction interval for the stream runoff is (47041.16, 72395.36).

(4) A 95% prediction interval for the stream runoff is (43532.98, 75903.54).

(5) A 95% prediction interval for the stream runoff is (44486.07, 74950.45).

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

   (1) The COVRATIO measures the overall change in the regression coefficients when a point is deleted.

   (2) Cook's distances measure the change in the standard errors when a point is deleted.

   (3) The DFFITS measures the change in R^2 when a point is deleted.

   (4) The hat matrix diagonal measures the leverage of a data point.

   (5) Points are considered influential if the DFBETAS have negative values.

10. A biologist has approached you for statistical advice. She is interested in the effect that various trace elements in the soil have on the growth of a species of marsh grass, and has data on 45 plots of ground, on which the following variables are measured:

    Bio: the above-ground biomass of the marsh grass growing on the plot (grams per square metre);

    H2S: Free sulphide (moles);

    Sal: Salinity (%);

    Eh7: Redox potential at pH 7;

    pH: acidity of water (pH);

    BUF: Buffer acidity at pH 6.6 (meg/100 cm3);

    P: Phosphorus concentration (ppm)

    K: Potassium concentration (ppm)

    Ca: Calcium concentration (ppm)

    Mg: Magnesium concentration (ppm)

    Na: Sodium concentration (ppm)

    Mn: Manganese concentration (ppm)

    Zn: Zinc concentration (ppm)

    Cu: Copper concentration (ppm)

    NH4: Ammonium concentration (ppm)

    The biologist is interested in using the other measurements to predict the biomass. An "all possible regressions" is run on her data with the following results:

	rssp	sigma2	adjRsq	Cp	AIC	BIC	CV
1	7680575	178618.0	0.590	21.406	66.406	70.019	750883.8
2	6527175	155408.9	0.643	14.034	59.034	64.454	683127.8
3	5256940	128218.1	0.706	5.713	50.713	57.940	573511.2
4	4797151	119928.8	0.725	3.978	48.978	58.011	540067.7
5	4490750	115147.4	0.736	3.488	48.488	59.328	532266.9
6	4114075	108265.1	0.752	2.428	47.428	60.074	503297.1
7	3949194	106735.0	0.755	3.088	48.088	62.541	509181.3
8	3898463	108290.6	0.751	4.676	49.676	65.936	601297.3
9	3765134	107575.2	0.753	5.592	50.592	68.659	564406.6
10	3711740	109168.8	0.749	7.158	52.158	72.032	614259.3
11	3701406	112163.8	0.743	9.075	54.075	75.754	633321.1
12	3694788	115462.1	0.735	11.021	56.021	79.507	669787.7
13	3692616	119116.6	0.727	13.003	58.003	83.296	704246.3
14	3692233	123074.4	0.718	15.000	60.000	87.100	737431.5

	H2S	Sal	Eh7	pH	BUF	P	K	Ca	Mg	Na	Mn	Zn	Cu	NH4
1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	0	1	0	0	0	0	0
3	0	0	0	1	0	0	0	1	1	0	0	0	0	0
4	0	0	0	1	0	0	0	1	1	0	0	0	1	0
5	0	1	0	0	0	0	1	0	0	0	0	1	1	1
6	0	1	1	0	0	0	1	0	0	0	0	1	1	1
7	0	1	1	0	0	0	1	0	1	0	0	1	1	1
8	0	0	1	1	0	1	1	1	1	0	0	0	1	1
9	0	1	1	1	0	0	1	1	1	0	0	1	1	1
10	0	1	1	1	0	1	1	1	1	0	0	1	1	1
11	0	1	1	1	0	1	1	1	1	0	1	1	1	1
12	0	1	1	1	0	1	1	1	1	1	1	1	1	1
13	1	1	1	1	0	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Which of the following is the **BEST** interpretation?

(1) The one-variable model should be used since it has the smallest adjusted R^2.

(2) The six-variable model should be used since it has the smallest CV.

(3) The full model should always be used for prediction.

(4) The full model should be used since it has the smallest residual sum of squares.

(5) The three-variable model should be used since it has the smallest BIC.

11. The data for this question are taken from an experiment which investigated the ascorbic acid content of cabbages. Two variables are thought to influence the ascorbic acid content:

    - the genetic line or cultivar (there were two lines, 39 and 52, recorded as the factor `Line`)
    - the planting date (three dates were considered, labeled as 16, 20 and 21, recorded as the factor `Date`).

    In the data set, there are 10 observations for each of the six factor level combinations, for a total of 60 observations. The response variable is `Ascorbic`, the ascorbic acid content of the cabbage head. The following table of means was obtained:

```
         Line
date    39   52
16    50.3 62.5
20    49.4 58.9
21    54.8 71.8
```

    Which of the following is **FALSE**?

    (1) The baseline mean is 50.3.

    (2) The interaction for date= 16, Line=52 is 0.

    (3) The interaction for date= 21, Line=52 is 4.8.

    (4) The row effect for date=20 is -0.9.

    (5) The main effect for Line=52 is -12.2.

12. In addition the variables Ascorbic, Line and Date, the weight of the cabbage heads (measured by the variable HeadWt) was also measured. An analysis of variance including this covariate was performed on the data in Question 11, with the following results:

```
> cabbage.lm<-lm(Ascorbic~factor(Date)*factor(Line)*HeadWt,
                                                    data=cabbage.df)
> cabbage2.lm<-lm(Ascorbic~factor(Date)*factor(Line) + HeadWt,
                                                    data=cabbage.df)
> anova(cabbage2.lm, cabbage.lm)
Analysis of Variance Table

Model 1: Ascorbic ~ factor(Date) * factor(Line) + HeadWt
Model 2: Ascorbic ~ factor(Date) * factor(Line) * HeadWt
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     53 1975.05
2     48 1847.24  5    127.82 0.6643 0.6523

> anova(cabbage2.lm)
```

```
Analysis of Variance Table

Response: Ascorbic
                           Df  Sum Sq Mean Sq F value    Pr(>F)
factor(Date)                2  909.30  454.65 12.2004 4.381e-05 ***
factor(Line)                1 2496.15 2496.15 66.9835 5.687e-11 ***
HeadWt                      1  629.61  629.61 16.8955 0.0001379 ***
factor(Date):factor(Line)   2   30.73   15.37  0.4124 0.6641800
Residuals                  53 1975.05   37.27
```

Which of the following is **FALSE**?

(1) An estimate of the error standard deviation is 6.10.

(2) No submodel of the model `Ascorbic~factor(Date) * factor(Line) + HeadWt` seems appropriate.

(3) The model `Ascorbic~factor(Date) * factor(Line) + HeadWt` fits 6 parallel lines.

(4) There is no evidence that the factors Date and Line interact.

(5) The model `Ascorbic~factor(Date) * factor(Line) * HeadWt` does not seem appropriate.

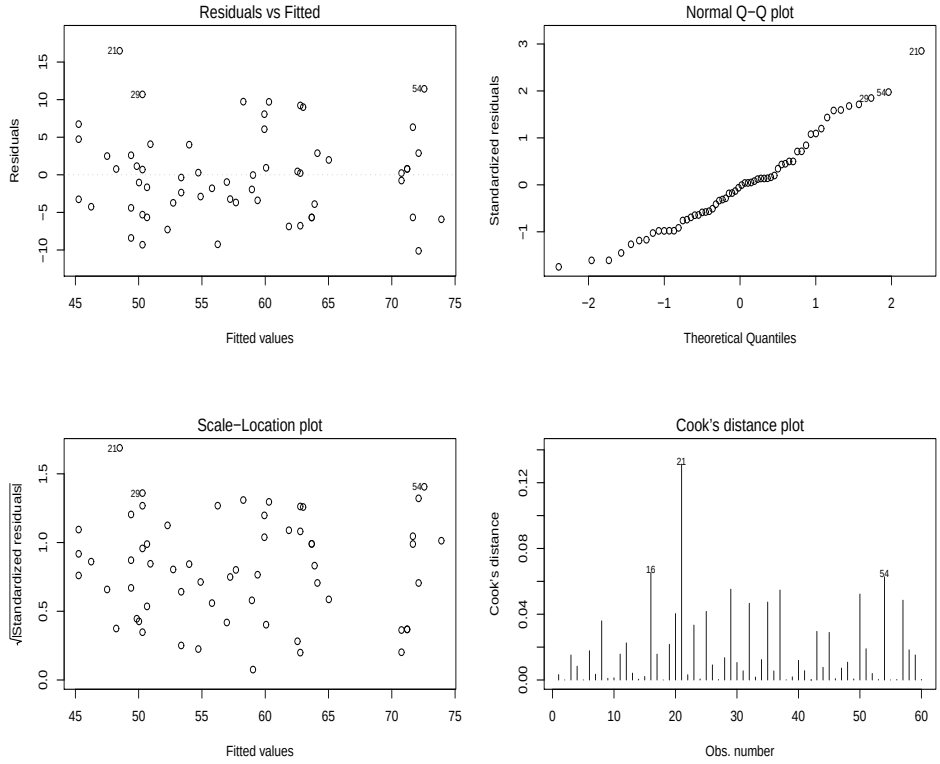

Figure 2: Diagnostic plots for Question 13.

13. The plots in Figure 2 are diagnostic plots from fitting the model `Ascorbic factor(Date) * factor(Line) + HeadWt` to the cabbage data. Which of the following is **TRUE**? The following output may be helpful:

```
qf(0.1,7, 53)
[1] 0.3968593
> qf(0.1,5,7)
[1] 0.2969210
> max(abs(rnorm(60)))
[1] 2.934225
```

(1) The plots suggest that point 21 is a high-leverage point.

(2) The plots do not suggest any violation of the regression assumptions.

(3) The plots suggest that point 21 is an outlier.

(4) The plots suggest that the data are not independent.

(5) The plots suggest that the data are not planar.

14. In the standard logistic regression model, which of the following is **NOT** part of the assumptions?

    (1) The responses have a binomial distribution.

    (2) The data can be grouped or ungrouped.

    (3) The log-odds are a linear function of the covariates.

    (4) The error variances have to be the same.

    (5) The responses are independent.

15. In logistic regression where the cases have few if any repeated covariate patterns, which of the following is **FALSE**?

    (1) High-leverage points will show up in a leverage-residual plot.

    (2) We can't interpret the residual deviance as a goodness of fit measure.

    (3) The residual deviance has approximately a chisquared distribution.

    (4) The plot of residuals versus fitted probabilities shows two curves.

    (5) The normal plot of residuals will not be straight.

16. The data for this question consist of measurements on 173 female horseshoe crabs. Female horseshoe crabs share a nest with a male partner. In some cases, additional males, called satellites, reside nearby. A biologist is interested in what attributes of the female are associated with the presence of satellites.

    The variables in the data set are

    colour: colour of the crab (1=light medium, 2=medium, 3=dark medium, 4=dark),

    spine: Spine condition (1=both good, 2=one broken, 3-both broken),

    width: Width of the carpace (shell) in cm,

    weight: weight of the crab (grams),

    satelite: presence of satellite crabs (0=absent, 1=present).

    A logistic model was fitted with the variable satellite as the response and treating spine and width as factors the following results:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.1452834  1.1750936  -1.826  0.06791 .
colour2     -0.3332333  0.7668349  -0.435  0.66388
colour3     -0.7798928  0.8345703  -0.934  0.35005
colour4     -1.8975821  0.9150676  -2.074  0.03811 *
spine2      -0.2013201  0.6753444  -0.298  0.76563
spine3       0.5569163  0.4808754   1.158  0.24681
weight       0.0012552  0.0003589   3.497  0.00047 ***

Null deviance: 226.90  on 172  degrees of freedom
Residual deviance: 196.31  on 166  degrees of freedom
AIC: 210.31

Analysis of Deviance Table

Model: binomial, link: logit
Response: satelite
Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     172     226.90
colour  3  12.9379       169     213.96 0.0047728 **
spine   2   3.4851       167     210.47 0.1750728
weight  1  14.1637       166     196.31 0.0001676 ***
```

Which of the following is **FALSE**?

(1) The p-value 0.1750728 relates to adding the variable `spine` to the model `colour + weight`

(2) There seems to be little point adding the variable `spine` to the model.

(3) The p-value 0.0001676 relates to adding the variable `weight` to the model `colour + spine`

(4) The estimated log-odds of having a satellite is approximately 1.9 less for dark crabs than light-medium crabs, for crabs with the same values of weight and spine.

(5) The model fitted above contains no interactions.

17. Which of the following is **FALSE**?

(1) The log-odds that a light-medium crab weighing 2000 grams and with both spines good will have a satellite is about 0.37.

(2) The probability that a dark crab weighing 2000 grams and with both spines good will have a satellite is about 0.18.

(3) The probability that a dark crab weighing 2000 grams and with both spines broken will have a satellite is about 0.38.

(4) The odds that a light-medium crab weighing 2000 grams and with both spines broken will have a satellite is about 2.51.

(5) The probability that a dark crab weighing 2000 grams and with both spines broken will have a satellite is about 0.27.

18. Some diagnostic plots for this regression are shown in Figures 3 and 4 overleaf. Some additional information is shown below. What is **NOT** a correct interpretation of these plots?

	colour	spine	weight	satelite
22	1	2	2300	1
131	1	2	1950	1
141	2	1	5200	?

(1) Removing point 141 changes the deviance by about 10.

(2) Points 22 and 131 have high leverage, but don't appear to be affecting the regression.

(3) Point 141 has a small estimated probability.

(4) Point 141 could be affecting the regression, so we should explore the changes if it is removed.

(5) The value of the variable `satellite` for point 141 is 0.

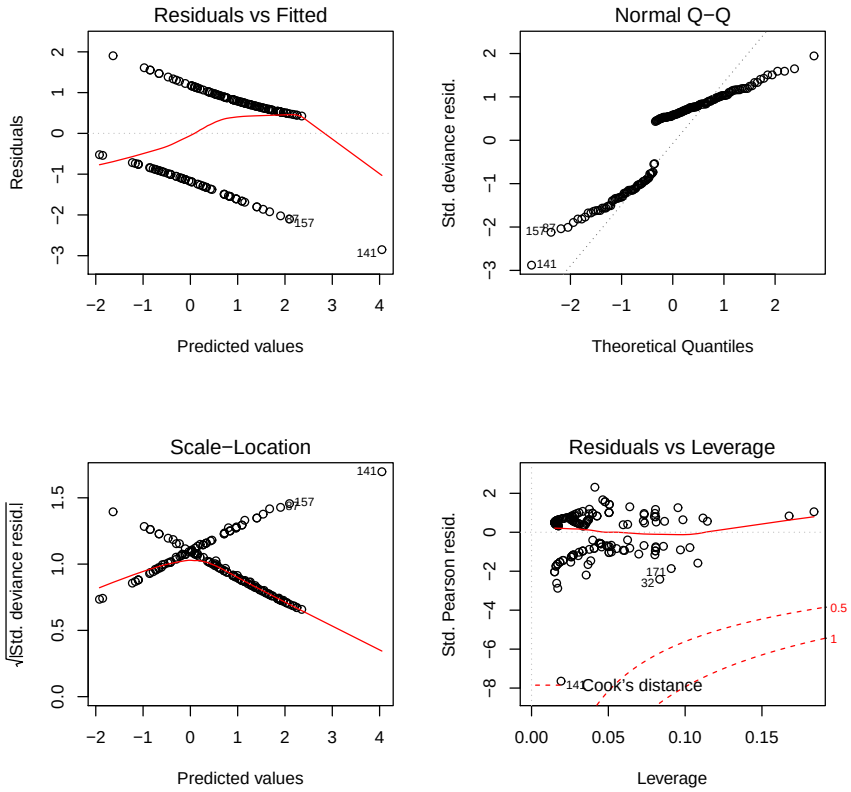

Figure 3: Diagnostic plots for Question 18.

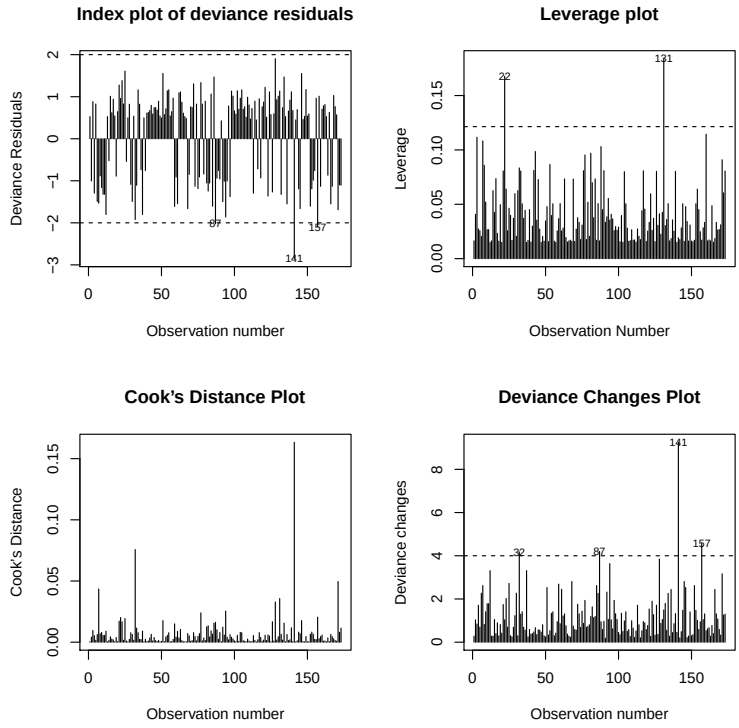

Figure 4: Further diagnostic plots for Question 18.

CONTINUED

19. Which of the following statements is **NOT** correct?

    (1) If the area under the ROC curve is 0.90, then the predictor is a good one.

    (2) Random guessing gives an area under the ROC curve of 0.5.

    (3) Specificity is the probability of a false positive.

    (4) The area under the ROC curve is always less than or equal to 1.

    (5) Sensitivity is the probability of a true positive.

20. In a Poisson regression, a variable X has a regression coefficient of 0.6. Which is the **CORRECT** interpretation?

    (1) If the other explanatory variables are held fixed, a unit increase in X is associated with a 82% increase the mean response.

    (2) If the other explanatory variables are held fixed, a unit decrease in X is associated with a 82% decrease the mean response.

    (3) Averaged over the other variables, a unit increase in X is associated with an increase of 0.6 in the log-odds ratio.

    (4) Averaged over the other variables, a unit increase in X is associated with an increase of 0.6 in the mean response.

    (5) If the other explanatory variables are held fixed, a unit increase in X is associated with an increase of 0.6 in the mean response.

21. The data in Table 1 are taken from a classic British study on smoking and mortality.

Table 1. Data for Question 21.

	Person Years		Coronary Deaths	
Age	Non-smokers	Smokers	Non-smokers	Smokers
35-44	18793	52407	2	32
45-54	10673	43248	12	104
55-64	5710	28612	28	206
65-74	2585	12663	28	186
75-84	1462	5317	31	102

The study investigated the effect of smoking on coronary death rates (measured as deaths per 100,000 person-years).

Which of the following lines of R code will produce the most appropriate analysis, assuming the data is has been stored as an R data frame with variables `Smoke` (Yes/No), `Age` (one of 35-44, 45-54, 55-64, 65-74,75-84), `PersonYears` and `CoronaryDeaths`?

(1) `smoke.glm=glm(CoronaryDeaths~Smoker*Age, offset = PersonYears/100000, family=poisson, data=smoke.df)`

(2) `smoke.glm=glm(CoronaryDeaths~Smoker*Age, family=binomial, data=smoke.df)`

(3) `smoke.glm=glm(CoronaryDeaths~Smoker*Age, offset = log(PersonYears), family=poisson, data=smoke.df)`

(4) `smoke.glm=glm(CoronaryDeaths~Smoker*Age, offset = log(PersonYears/100000), family=poisson, data=smoke.df)`

(5) `smoke.glm=glm(CoronaryDeaths~Smoker*Age, family=poisson, data=smoke.df)`

22. The correct analysis was run with the following results:

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)          2.3648     0.7071   3.344 0.000825 ***
SmokerYes            1.7470     0.7289   2.397 0.016534 *
Age45-54             2.3575     0.7638   3.087 0.002024 **
Age55-64             3.8303     0.7319   5.233 1.67e-07 ***
Age65-74             4.6228     0.7319   6.316 2.68e-10 ***
Age75-84             5.2945     0.7296   7.257 3.95e-13 ***
SmokerYes:Age45-54  -0.9868     0.7901  -1.249 0.211667
SmokerYes:Age55-64  -0.1623     0.7562  -0.215 0.830057
SmokerYes:Age65-74  -1.4424     0.7565  -1.907 0.056564 .
SmokerYes:Age75-84  -1.8472     0.7572  -2.440 0.014706 *

Null deviance: 1.2590e+03  on 9  degrees of freedom
Residual deviance: 5.8176e-14  on 0  degrees of freedom
AIC: 75.068
```

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

(1) The death rate for non-smokers aged 35-44 is a bit over 2 per 100,000 person years.

(2) The death rate for smokers aged 55-64 is about 2800 per 100,000 person years.

(3) The death rate for smokers aged 35-44 is about 61 per 100,000 person years.

(4) The death rate for non-smokers aged 55-64 is over 500 per 100,000 person years.

(5) The death rate for smokers aged 45-54 is about 5.5 per 100,000 person years.

23. For the study in Questions 21-22, the model `CoronaryDeaths~Smoker+Age` was fitted. The null deviance was 1259.048 on 9 degrees of freedom and the residual deviance was 42.172 on 4 degrees of freedom. Which of the statements below is **NOT** correct? The following may be helpful:

```
> pchisq(1259.048,9)
[1] 1
> pchisq(42.172,4)
[1] 1
```

(1) There are four less parameters in the model `CoronaryDeaths~Smoker+Age` than in the model fitted in Question 22.

(2) The model fitted in Question 22 puts no restrictions on the death rates.

(3) The p-value associated with a deviance of 42.172 can be calculated using a Chi-square distribution.

(4) The model `CoronaryDeaths~Smoker+Age` seems an adequate model.

(5) The effect of age on the death rate is different for non-smokers and smokers.

24. In an experiment to study whether birds of prey could detect levels of parasitic infection in the fish they eat, 141 fish infected with different levels of infection were offered to the birds at random. The fish are categorized by their level of parasitic infection, either uninfected, lightly infected, or highly infected. The numbers eaten and not eaten are shown in Table 2.

Table 2. Data for Question 24.

	Uninfected	Lightly Infected	Highly Infected	Total
Eaten	1	10	37	48
Not eaten	49	35	9	93
Total	50	45	46	141

The data were assembled into a data frame with variables `count` (containing the counts), `infected` (Not, Lighly, Highly) and `eaten` (Yes/No). The following R output was obtained:

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               4.677e-11  1.000e+00   0.000 1.000000
infectedLightly           2.303e+00  1.049e+00   2.195 0.028132 *
infectedHighly            3.611e+00  1.013e+00   3.563 0.000366 ***
eatenNo                   3.892e+00  1.010e+00   3.853 0.000117 ***
infectedLightly:eatenNo -2.639e+00  1.072e+00  -2.462 0.013815 *
infectedHighly:eatenNo  -5.306e+00  1.076e+00  -4.929 8.26e-07 ***

Null deviance: 9.2808e+01  on 5  degrees of freedom
Residual deviance: 9.7700e-15  on 0  degrees of freedom

Analysis of Deviance Table

              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                5     92.808
infected       2    0.295         3     92.513 0.8627290
eaten          1   14.616         2     77.897 0.0001318 ***
infected:eaten 2   77.897         0      0.000 < 2.2e-16 ***
```

Which of the following is **TRUE**?

(1) There is very strong evidence that the birds are not choosing the fish to be eaten at random.

(2) The odds ratio corresponding to not eating fish and being lightly infected is 2.639.

(3) The residual deviance is suspiciously small.

(4) The logit of the probability of not being eaten is 3.892.

(5) The odds ratio corresponding to not eating fish and being uninfected is 3.892.

25. In the first round of the 2010 World Cup, there were 54 matches played in the first round. Thus there are 108 scores (number of goals), one for each team playing in each match. Table 3 below gives the distribution of these 108 scores:

Table 3. Data for Question 25.

Number of goals	0	1	2	3	4	5	6	7
Frequency	35	35	18	5	2	0	0	1

We want see if the number of goals per team per match follows a Poisson distribution. Recall that the log-likelihood for a one-dimensional contingency table is $\sum_{i=0}^{7} y_i \log(\pi_i)$ where π_i is the probability a team will score i goals. Calculations in R give the mean number of goals as 1.052. If we substitute the relative frequencies into this log-likelihood we get -127.8429, if we substitute the fitted Poisson probabilities we get a log-likelihood of -132.0501, and if we substitute 1/8 we get a log-likelihood of -199.6264 Some additional information follows.

```
> pchisq(4.207219,6)
[1] 0.3513432
> pchisq(4.207219,7)
[1] 0.2443755
> pchisq(4.207219,8)
[1] 0.1620399
> pchisq(8.414438,6)
[1] 0.790715
> pchisq(8.414438,7)
[1] 0.7025295
> pchisq(8.414438,8)
[1] 0.605932
```

Which if the following is **FALSE**?

(1) The null deviance is 143.6 (to one decimal place).

(2) The null deviance has 7 degrees of freedom.

(3) The residual deviance for the Poisson model is about 8.4.

(4) The Poisson model is a not good fit to these data.

(5) The calculation of the null deviance does not involve the Poisson probabilities.

SECTION B

26. (a) Suppose we have a set of data consisting of a continuous response Y, two continuous explanatory variables X and W and a categorical explanatory variable A having three levels. What is the model you would initially fit to these data? Write a line of R code that would fit the model. Describe the model in geometrical terms. What simplification of the model might be possible? How would you test if this is in fact indicated by the data? Illustrate your answer with R code. [5 marks]

(b) In a tutorial, we looked at a set of data relating the average daily weight gain (ADG) of 23 calves to the ADG (as calves) of their dams (i.e. their mothers). The variables in the data set are

breed: the breed of calf (a factor with levels 1,2,3)

adg: the ADG of the calf,

dadg: the ADG of the dam.

An analysis was run with the following results:

```
Call:
lm(formula = adg ~ breed * dadg, data = calf.df)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.1754     0.5913   5.370 5.09e-05 ***
breed2       -1.6858     0.6214  -2.713   0.0148 *
breed3       -0.1331     0.6710  -0.198   0.8452
dadg         -0.0920     0.2889  -0.318   0.7540
breed2:dadg   0.6119     0.3034   2.017   0.0597 .
breed3:dadg   0.5088     0.3111   1.635   0.1204
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1668 on 17 degrees of freedom
Multiple R-squared: 0.9664,     Adjusted R-squared: 0.9565
F-statistic:  97.7 on 5 and 17 DF,  p-value: 6.56e-12

Analysis of Variance Table

Response: adg
           Df  Sum Sq Mean Sq  F value    Pr(>F)
breed       2 12.3467  6.1733 221.9562 6.577e-13 ***
dadg        1  1.1238  1.1238  40.4068 7.162e-06 ***
breed:dadg  2  0.1160  0.0580   2.0847     0.155
Residuals  17  0.4728  0.0278
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Another model was fitted, yielding

```
Call:
lm(formula = adg ~ breed + dadg, data = calf.df)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.08165    0.16683  12.477 1.34e-10 ***
breed2      -0.44675    0.09522  -4.692 0.000159 ***
breed3       0.88217    0.10509   8.394 8.15e-08 ***
dadg         0.44591    0.07405   6.022 8.57e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.176 on 19 degrees of freedom
Multiple R-squared: 0.9581,     Adjusted R-squared: 0.9515
F-statistic: 144.9 on 3 and 19 DF,  p-value: 2.874e-13
```

Do you think that there is a relationship between the ADG of the calf and that of its dam? If so, does this depend on the breed? In what respect? Discuss, giving full reasons and quoting from the output above. [8 marks]

(c) Describe Cooks D and explain how it can be used to identify influential points [2 marks]

(d) The model fitted in (b) above was checked for influential points using the output below. For each starred point, indicate why it has been flagged, and what effect it is having on the regression fit. [5 marks]

```
Influence measures of
        lm(formula = adg ~ breed * dadg, data = calf.df) :

      dfb.1_  dfb.brd2  dfb.brd3  dfb.dadg  dfb.br2.  dfb.br3.    dffit  cov.r   cook.d   hat inf
1   1.30e+00 -1.24e+00 -1.15e+00 -1.41e+00  1.35e+00  1.31e+00 -1.68091 1.5799 4.42e-01 0.570   *
2  -9.47e-01  9.01e-01  8.34e-01  1.14e+00 -1.09e+00 -1.06e+00  1.97466 0.0713 3.99e-01 0.250   *
3   7.45e-01 -7.09e-01 -6.57e-01 -6.94e-01  6.60e-01  6.44e-01  0.84942 2.2084 1.22e-01 0.500   *
4   1.20e-01 -1.14e-01 -1.06e-01 -2.05e-01  1.95e-01  1.90e-01 -0.75320 0.7145 8.65e-02 0.180
5  -7.85e-01  7.47e-01  6.92e-01  7.07e-01 -6.74e-01 -6.57e-01 -1.00554 1.0407 1.59e-01 0.330
6   6.55e-02 -6.23e-02 -5.77e-02 -3.62e-02  3.45e-02  3.36e-02  0.25863 1.5355 1.16e-02 0.170
7   2.70e-17 -6.92e-03 -2.14e-17 -2.48e-17 -1.44e-02  2.20e-17 -0.22379 1.4630 8.69e-03 0.131
8   2.02e-16 -3.58e-01 -7.55e-17 -1.25e-16  3.08e-01  8.09e-17 -1.20907 1.1676 2.29e-01 0.411
9   2.51e-16  8.96e-02 -2.24e-16 -2.64e-16 -5.24e-02  2.42e-16  0.44958 1.0997 3.33e-02 0.146
10 -2.83e-19  9.16e-05  2.60e-19  2.63e-19  2.12e-05 -2.59e-19  0.00118 1.6451 2.47e-07 0.125
11 -1.44e-17 -3.24e-02  1.78e-17  3.96e-17  5.49e-02 -2.96e-17  0.27833 1.6473 1.35e-02 0.214
12 -2.54e-16  2.07e-01  1.88e-16  2.22e-16 -1.74e-01 -1.95e-16  0.70910 1.5971 8.42e-02 0.355
13 -2.49e-17  2.82e-02  1.12e-17  1.47e-17 -3.67e-02 -9.51e-18 -0.14112 2.6279 3.52e-03 0.457   *
14  3.41e-17  1.28e-02 -3.13e-17 -4.45e-17 -3.42e-02  3.83e-17 -0.23918 1.5324 9.94e-03 0.160
15 -2.87e-16  2.75e-16 -1.87e-01  1.43e-16 -1.37e-16  1.23e-01 -0.51812 1.1646 4.43e-02 0.188
16  1.87e-16 -1.72e-16  1.22e-01 -1.28e-16  1.23e-16 -8.71e-02  0.28446 2.0804 1.42e-02 0.347   *
17  2.44e-16 -2.37e-16  1.97e-01 -9.98e-17  9.81e-17 -1.30e-01  0.54555 1.1146 4.88e-02 0.188
18  1.40e-18 -1.19e-18  9.09e-04 -9.47e-19  8.11e-19 -6.20e-04  0.00231 1.8716 9.48e-07 0.231
19  9.83e-17 -8.59e-17  6.74e-02 -5.89e-17  5.70e-17 -6.49e-02 -0.24237 1.7403 1.03e-02 0.231
20 -7.23e-17  6.69e-17 -7.52e-02  2.63e-17 -1.90e-17  7.25e-02  0.27067 1.7095 1.28e-02 0.231
21  1.66e-17 -1.22e-17 -2.74e-02 -3.40e-17  2.59e-17 -8.35e-18 -0.33108 1.1751 1.84e-02 0.111
22  2.23e-17 -2.85e-17 -4.34e-02 -4.56e-17  4.72e-17  4.04e-02  0.13939 1.9734 3.43e-03 0.284
23 -3.84e-17  3.33e-17 -3.32e-02  2.12e-17 -1.89e-17  3.37e-02  0.14171 1.7153 3.54e-03 0.188
```

27. (a) Carefully define the terms *null deviance* and *residual deviance* as applied to a logistic regression model for grouped data. In the same context, what is meant by a *saturated model*? Why must a saturated model have zero deviance?

[6 marks]

(b) The data below show the free-throw results obtained by the Los Angeles Lakers player Shaq O'Neal in 23 NBA playoff games in the year 2000. (For those unfamiliar with basketball, a free throw is when a player is allowed to take an unopposed shot at the basket from the free-throw line. Thus in game 1, O'Neal attempted 5 free throws, 4 of which were successful.)

```
> freethrows.df
    r  n game
1   4  5    1
2   5 11    2
3   5 14    3
4   5 12    4
5   2  7    5
6   7 10    6
7   6 14    7
8   9 15    8
9   4 12    9
10  1  4   10
11 13 27   11
12  5 17   12
13  6 12   13
14  8  9   14
15  7 12   15
16  3 10   16
17  8 12   17
18  1  6   18
19 18 39   19
20  3 13   20
21 10 17   21
22  1  6   22
23  3 12   23
```

Below is some R output from an analysis of these results.

```
> freethrows.glm<-glm(cbind(r,n-r)~game, family=binomial,
         data=freethrows.df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.3863     1.1180   1.240   0.2150
game2        -1.5686     1.2715  -1.234   0.2173
game3        -1.9741     1.2494  -1.580   0.1141
game4        -1.7228     1.2621  -1.365   0.1722
game5        -2.3026     1.3964  -1.649   0.0992 .
game6        -0.5390     1.3138  -0.410   0.6816
game7        -1.6740     1.2416  -1.348   0.1776
game8        -0.9808     1.2360  -0.794   0.4275
game9        -2.0794     1.2748  -1.631   0.1028
game10       -2.4849     1.6073  -1.546   0.1221
game11       -1.4604     1.1825  -1.235   0.2168
game12       -2.2618     1.2383  -1.827   0.0678 .
game13       -1.3863     1.2583  -1.102   0.2706
```

```
game14        0.6931     1.5411   0.450   0.6529
game15       -1.0498     1.2621  -0.832   0.4055
game16       -2.2336     1.3138  -1.700   0.0891 .
game17       -0.6931     1.2748  -0.544   0.5866
game18       -2.9957     1.5652  -1.914   0.0556 .
game19       -1.5404     1.1633  -1.324   0.1854
game20       -2.5903     1.2974  -1.996   0.0459 *
game21       -1.0296     1.2218  -0.843   0.3994
game22       -2.9957     1.5652  -1.914   0.0556 .
game23       -2.4849     1.3017  -1.909   0.0563 .
---

Null deviance: 3.3376e+01  on 22  degrees of freedom
Residual deviance: 8.8818e-16  on  0  degrees of freedom
AIC: 109.17

> 1-pchisq(33.376, 22)
[1] 0.056777
```

Press reports of these games criticised O'Neal for his inconsistent performance. Is this justified? Give a reason for your answer. What statistical model are you using to arrive at your conclusion? [8 marks]

(c) What is meant by "over-dispersion" and "under-dispersion" in this context? Do you think either could apply here? What effect would it have on the analysis? [6 marks]

28. (a) Suppose we have a three-dimensional contingency table with factors A, B and C. Carefully describe what we mean by the conditional odds ratios between A and B, given C. What are the values of these odds ratios if A and B are conditionally independent, given C? [6 marks]

(b) What is Simpson's paradox? Give an example from class. [4 marks]

(c) The data in Table 4 below relate to motor vehicle accidents in Florida in 1988. All persons injured in motor accidents in Florida in 1988 were classified according to three factors, namely

Injury: either Nonfatal or Fatal;

Seatbelt: either Yes if a seatbelt was worn, or No if a seatbelt was not worn;

Ejected: either Yes if the person was ejected from the vehicle, or No if not.

Table 4: Florida accident data.

		Injury	
		Nonfatal	Fatal
Seatbelt: Yes	Ejected: Yes	1,105	14
	Ejected: No	411,111	483
Seatbelt: No	Ejected: Yes	4,624	497
	Ejected: No	157,342	1,008

A log-linear model was fitted to the counts, and the following output was obtained:

```
Analysis of Deviance Table

                        Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                        7    1624865
Ejected                  1   729871         6     894994  < 2e-16 ***
Seatbelt                 1   111458         5     783536  < 2e-16 ***
Injury                   1   772092         4      11444  < 2e-16 ***
Ejected:Seatbelt         1     7877         3       3568  < 2e-16 ***
Ejected:Injury           1     2423         2       1145  < 2e-16 ***
Seatbelt:Injury          1     1142         1          3  < 2e-16 ***
Ejected:Seatbelt:Injury  1        3         0          0  0.09115
```

What model is indicated by this output? Describe the model in terms of conditional odds ratios. [5 marks]

(d) The output below resulted from fitting the model suggested by the anova table above. Use the output to calculate a confidence interval for the conditional odds ratio between wearing a seatbelt and injury type, given ejection status. [5 marks]

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  6.92251    0.03110  222.56   <2e-16 ***
EjectedYes                  -0.72784    0.05345  -13.62   <2e-16 ***
SeatbeltYes                 -0.75682    0.05394  -14.03   <2e-16 ***
InjuryNonfatal               5.04362    0.03120  161.65   <2e-16 ***
EjectedYes:SeatbeltYes      -2.39964    0.03334  -71.97   <2e-16 ***
EjectedYes:InjuryNonfatal   -2.79779    0.05526  -50.63   <2e-16 ***
SeatbeltYes:InjuryNonfatal   1.71732    0.05402   31.79   <2e-16 ***
```