

# THE UNIVERSITY OF AUCKLAND

---

SECOND SEMESTER, 2007

Campus: City

---

## STATISTICS

Advanced Statistical Modeling

(Time allowed: **THREE** hours)

### INSTRUCTIONS

#### SECTION A: Multiple Choice (60 marks)

- Answer **ALL 25** questions on the answer sheet provided.
- All questions have a single correct answer and carry the same mark value.
- If you give more than one answer to any question you will receive zero marks for that question.
- Each correct answer scores 2.4 marks.

#### SECTION B (40 marks)

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

**Total for both parts:** 100 marks

CONTINUED

# SECTION A

1. The data frame `Prestige` contains data collected in 1971 on 102 different occupations in Canada. Each line in the data frame represents an occupation. The variables are
  - prestige:** A score reflecting the prestige of the occupation (the bigger the score, the higher the prestige);
  - income:** The average annual income of persons having this occupation;
  - education:** The average number of years of education of persons having this occupation;
  - women:** Female participation i.e. the percentage of persons having this occupation who are women;
  - type:** The type of occupation, one of `bc` (blue collar), `prof` (professional) or `wc` (white collar). All the variables except `type` are continuous.

Suppose we want to see if the relationship between prestige and type of occupation depends on the level of income. Which of the following R commands would produce the most informative graph?

- (1) `dotplot(prestige~type|income, data=Prestige)`
  - (2) `plot(prestige~type, data=Prestige)`
  - (3) `dotplot(type~income|prestige, data=Prestige)`
  - (4) `dotplot(prestige~income|type, data=Prestige)`
  - (5) `dotplot(prestige~type|equal.count(income), data=Prestige)`
2. A trellis plot of `prestige` versus `education`, conditional on `women` and `income` is shown in Figure 1. Which of the following is an **incorrect** interpretation of this graph?
    - (1) As income goes up, prestige goes up.
    - (2) There is an increasing relationship between education and prestige at all levels of income and female participation.
    - (3) The most prestigious occupations are associated with high education and high income.
    - (4) The data appear planar.
    - (5) The relationship between prestige and education depends on the level of female participation.

CONTINUED

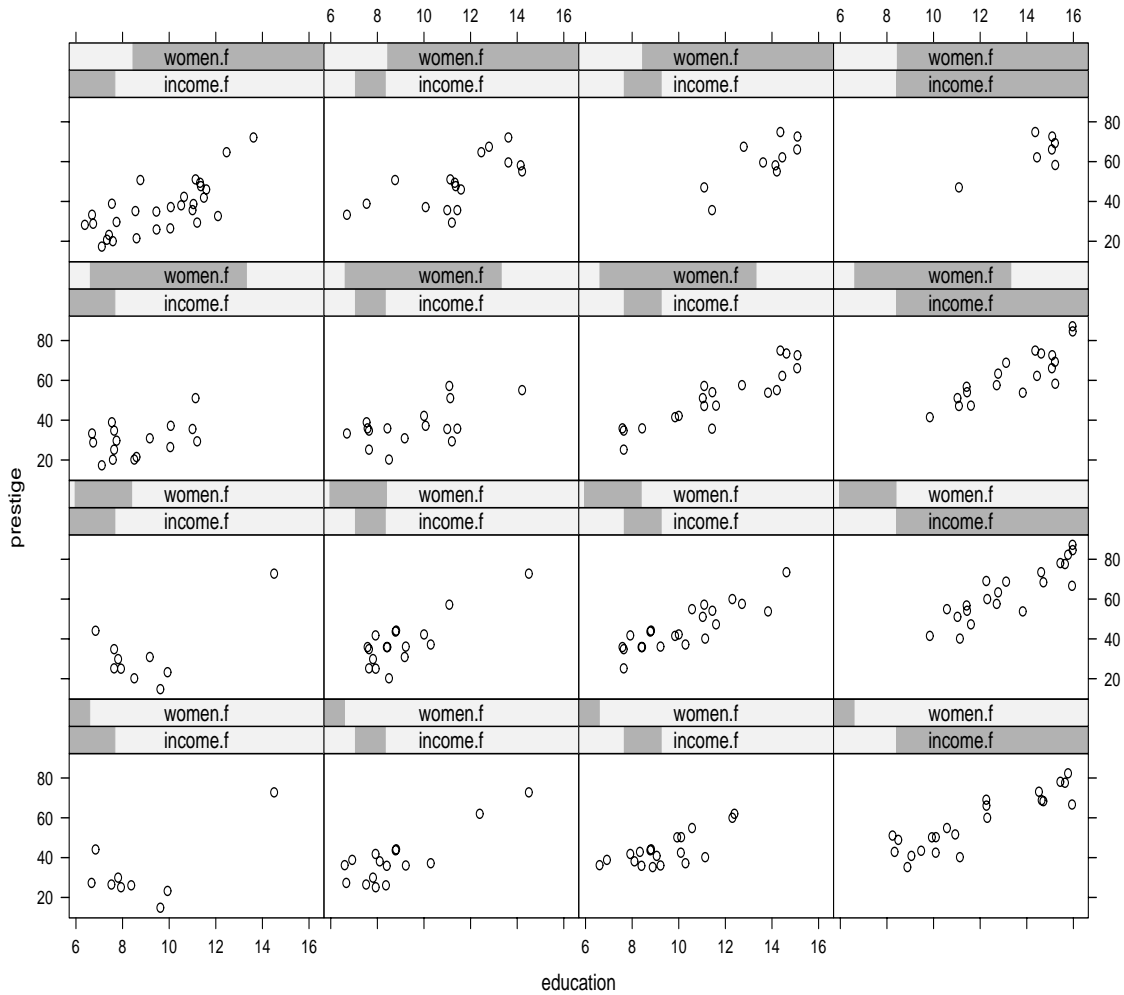


Figure 1: Trellis plot for Question 2.

3. In a psychology experiment, 45 subjects were asked to express opinions on 40 different issues. Each subject then interacted with a partner, who attempted to get the subject to change his or her mind on each of the 40 issues. Each partner was graded as either low status or high status, and more or less authoritarian, on a three point scale (low/medium/high). The number of issues (out of 40) on which the subjects changes their minds were recorded in the variable `conformity`, along with the two attributes of their partner in the variables `status` and `authoritarian`. The aim of the experiment was to see how the attributes of the partner affected the conformity.

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

- (1) Conformity is lowest for medium levels of authoritarianism and low status.
- (2) Overall, conformity is higher for high-status partners.
- (3) For low-status partners, conformity stays roughly constant as authoritarianism goes up, apart from two subjects.
- (4) Conformity is highest for low-status, low authoritarianism partners.
- (5) For high-status partners, conformity goes down as authoritarianism goes up.

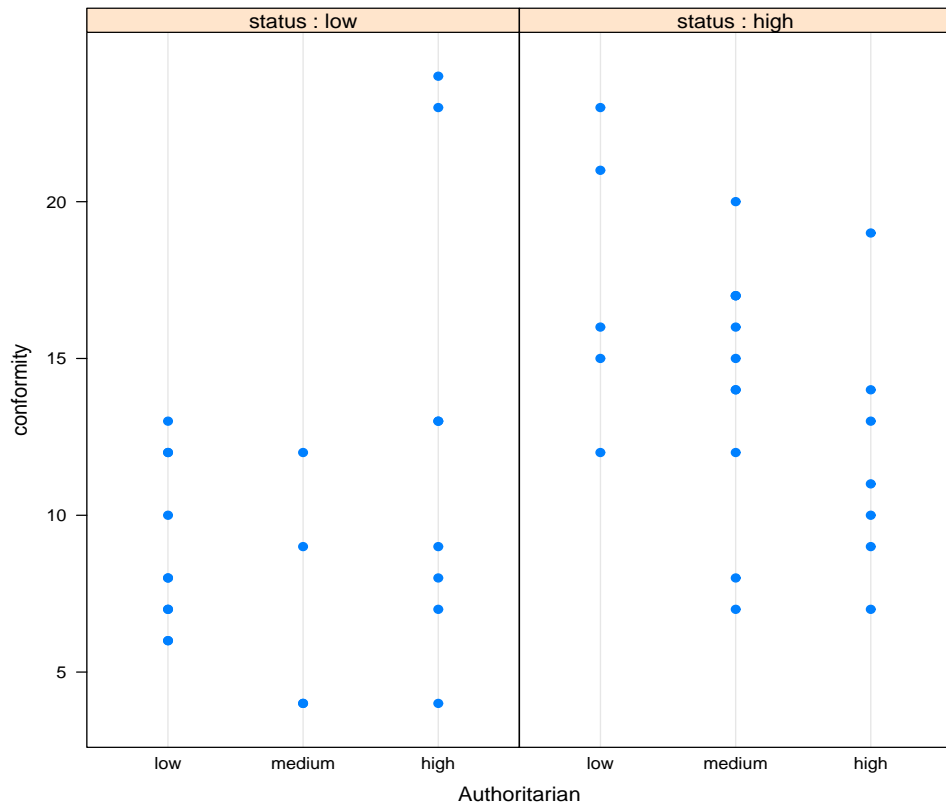


Figure 2: Trellis plot for for Question 3.

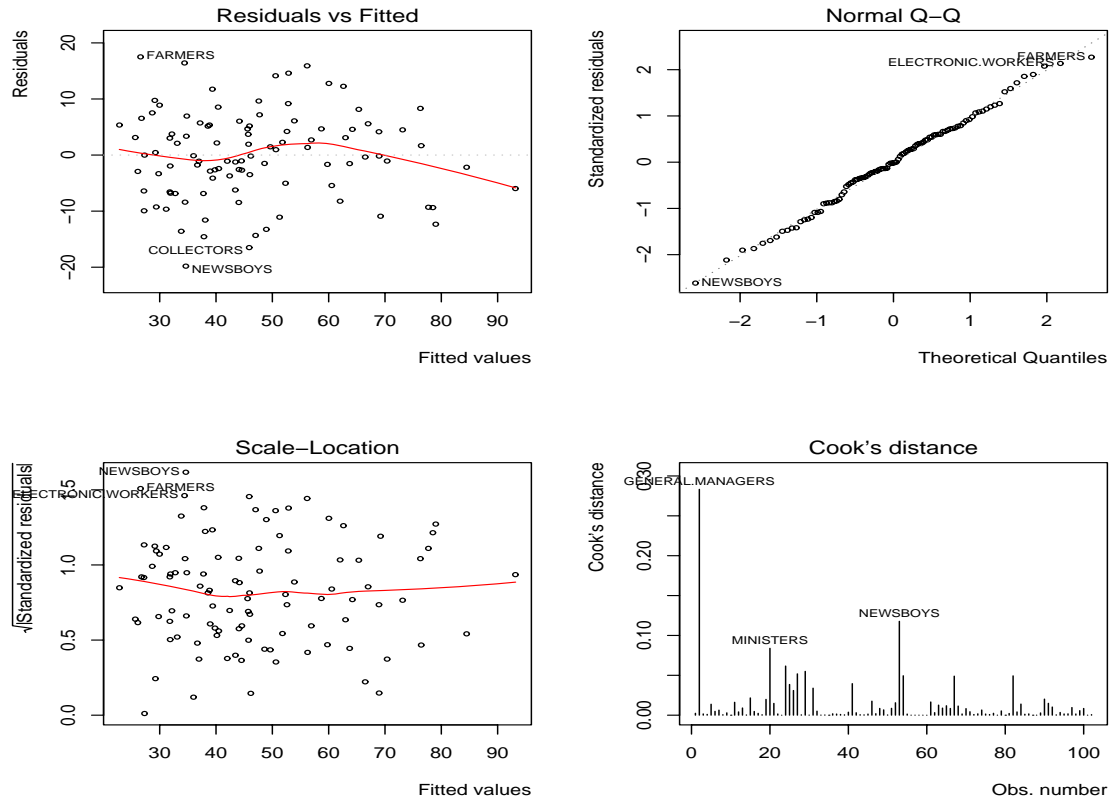


Figure 3: Diagnostic plots for Question 4.

4. Suppose that we fit a regression model using the continuous variables in the Prestige data set, using `prestige` as the response. The diagnostic plots shown in Figure 3 were obtained. The following output may be useful:

```
> dim(Prestige)
[1] 102  5
> qf(0.1,4,98)
[1] 0.2646549
> qf(0.9,4,98)
[1] 2.003122
```

Which of the following actions is indicated by the plots?

- (1) You should transform the explanatory variables.
- (2) You should transform the response, since the equal-variance assumption seems violated.
- (3) The effect of GENERAL MANAGERS on the regression should be explored.
- (4) Do nothing, the regression looks OK.
- (5) You should do a Weisberg-Bingham test to check the normality.

CONTINUED

5. In a regression analysis with a continuous response  $y$  and two explanatory variables  $x$  and  $w$ , the coefficient of  $x$  is estimated as 2.3. Which of the following is **false**?
- (1) If  $x$  and  $w$  are uncorrelated, adding  $w$  to the regression does not change the estimated coefficient of  $x$ .
  - (2) A unit increase in  $x$  produces an increase of 2.3 in the mean response, provided that  $w$  is held constant.
  - (3) The slopes in a coplot of  $y$  versus  $x$ , conditional on  $w$  are about 2.3.
  - (4) The slope in a scatterplot of  $y$  versus  $x$  is about 2.3.
  - (5) The variance of the estimated coefficient of  $x$  is a minimum if  $x$  and  $w$  are uncorrelated.
6. In a regression with several explanatory variables, several factors affect the standard errors of the regression coefficients. Suppose  $\hat{\beta}$  is the estimated coefficient corresponding to a variable  $x$ . Which of the following is **false**?
- (1) Suppose that we regress  $x$  on the other explanatory variables. The bigger the  $R^2$  from this regression, the smaller the standard error of  $\hat{\beta}$  will be.
  - (2) In general, the standard error of  $\hat{\beta}$  depends on all the explanatory variables, not just  $x$ .
  - (3) The standard error of  $\hat{\beta}$  goes down as the sample size goes up.
  - (4) The bigger the variance of  $x$ , the smaller the standard error of  $\hat{\beta}$  will be.
  - (5) The smaller the error variance, the smaller the standard error of  $\hat{\beta}$  will be.
7. Suppose that we fit a regression model to the occupation data discussed in Q1, using `prestige` as the response. We get the following output:

Call:

```
lm(formula = prestige ~ education + income + women, data = Prestige)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-6.7943342	3.2390886	-2.098	0.0385 *
education	4.1866373	0.3887013	10.771	< 2e-16 ***
income	0.0013136	0.0002778	4.729	7.58e-06 ***
women	-0.0089052	0.0304071	-0.293	0.7702

---

Residual standard error: 7.846 on 98 degrees of freedom

Multiple R-Squared: 0.7982, Adjusted R-squared: 0.792

F-statistic: 129.2 on 3 and 98 DF, p-value: < 2.2e-16

CONTINUED

Which of the following is **false**?

- (1) Other things being equal, for each extra dollar of income, the mean prestige score goes up by 0.0013136.
  - (2) Other things being equal, the mean prestige score goes down for each 1% decrease in female participation.
  - (3) The estimate of the error variance is 61.55972.
  - (4) At least one of the variables in the regression has a significant relationship with the response.
  - (5) Other things being equal, for each extra year of education, the mean prestige score goes up by 4.1866373
8. Consider the diagnostic plots of the occupation data shown in Figure 3. Which of the following is **false**?
- (1) According to the model, collectors have more prestige than newsboys.
  - (2) There seems to be no problem with normality.
  - (3) Newsboys have a lower prestige than that predicted by the model.
  - (4) According to the model, farmers have low prestige.
  - (5) The high value of Cooks D for general managers is caused by a big residual.
9. Suppose we have another occupation, not included in the original data set, that has `education = 8.55`, `income = 3617`, `women = 70.87`. We get the following output:

```
> predict(prestige.lm, data.frame(education=8.55,
                                income=3617, women=70.87), se=T)
$fit
[1] 33.12145
$se.fit
[1] 1.527865
$df
[1] 98
$residual.scale
[1] 7.846467
```

Which piece of R code gives the correct 95% prediction interval?

- (1)  $33.12145 + c(-1,+1)*qt(0.975, 98)*7.846467$
- (2)  $33.12145 + c(-1,+1)*qt(0.975, 98)*1.527865$
- (3)  $33.12145 + c(-1,+1)*qt(0.95, 98)*\sqrt{7.846467^2+1.527865^2}$
- (4)  $33.12145 + c(-1,+1)*qt(0.975, 98)*\sqrt{7.846467^2+1.527865^2}$
- (5)  $33.12145 + c(-1,+1)*qt(0.95, 98)*1.527865$

CONTINUED

10. Below we show the result of running an “all possible regressions” on the prestige data.

```
all.poss.regs(prestige~education+income+women, data=Prestige, best=2)
```

	rssp	sigma2	adjRsq	Cp	AIC	BIC	CV	education	income	women
1	8286.990	82.870	0.720	36.601	138.601	143.851	849.919	1	0	0
1	14616.170	146.162	0.506	139.403	241.403	246.652	1534.524	0	1	0
2	6038.851	60.998	0.794	2.086	104.086	111.961	640.593	1	1	0
2	7410.281	74.851	0.747	24.361	126.361	134.236	774.294	1	0	1
3	6033.570	61.567	0.792	4.000	106.000	116.500	661.042	1	1	1

Which of the following is **true**?

- (1) The model with all three variables is best since it has the smallest residual sum of squares.
  - (2) The best model is the model using education and income as the only explanatory variables.
  - (3) The best one variable model uses income as an explanatory variable.
  - (4) For big sample sizes, BIC is sometimes smaller than AIC.
  - (5) The adjusted  $R^2$  is usually bigger than the  $R^2$ .
11. Suppose in a regression, there is strong positive serial correlation in the errors. What is **not** likely to happen?
- (1) A plot of residuals versus previous residuals will show a linear trend.
  - (2) The value of the Durbin-Watson statistic will be close to 0.
  - (3) There will be long runs of positive residuals.
  - (4) The standard errors will be incorrectly estimated by the `lm` function.
  - (5) The acf plot will cut off sharply after lag zero.
12. The output below relates to the psychology experiment described in Question 3. The data are in a data frame `Moore`.

```
> Moore.lm=lm(conformity~authority*status, data=Moore)
> means=tapply(Moore$conformity, list(Moore$authoritarian,Moore$status), mean)
> means
```

	low	high
low	8.900	17.40000
medium	7.250	14.27273
high	12.625	11.85714

CONTINUED

```

> summary(Moore.lm)
Call:
lm(formula = conformity ~ authoritarian * status, data = Moore)
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    8.900      1.448    6.146 3.24e-07 ***
authoritarianmedium            -1.650      2.709   -0.609  0.54601
authoritarianhigh              3.725      2.172    1.715  0.09429 .
statushigh                     8.500      2.508    3.389  0.00162 **
authoritarianmedium:statushigh **1**                -0.403  0.68917
authoritarianhigh:statushigh  **2**                -2.686  0.01057 *

Residual standard error: 4.579 on 39 degrees of freedom
Multiple R-Squared: 0.3237,    Adjusted R-squared: 0.237
F-statistic: 3.734 on 5 and 39 DF,  p-value: 0.007397

```

In the summary output, two figures have been removed and two figures have been replaced by **\*\*1\*\*** and **\*\*2\*\***. Which of the following is **true**?

- (1) The value of **\*\*1\*\*** is -19.2772
  - (2) The value of **\*\*1\*\*** is -1.47727
  - (3) The value of **\*\*2\*\*** is 7.73214
  - (4) The value of **\*\*2\*\*** is -27.0678
  - (5) The value of **\*\*2\*\*** is -1.81785
13. Which of the following is **true**?
- (1) Authoritarianism has no significant effect on conformity.
  - (2) The estimated error standard deviation is 20.96724.
  - (3) The effect on conformity of changing the value of the variable **authoritarian** is different for low status and high status.
  - (4) Since the residual sum of squares is so large, no conclusion can be drawn.
  - (5) Status has no significant effect on conformity.
14. The output below is relates to a study to explore the connection between a company's sales and various explanatory variables. Data from 25 sales districts in three regions were recorded. The variables are
- SALES:** 1999 sales, in thousands of dollars ;
- ADV:** Amount spent on advertising (hundreds of dollars);
- BONUS:** Total amount of bonuses (hundreds of dollars);
- SALES.REGION:** One of 1, 2 or 3 (a factor).

A model was fitted with sales as the response and the other variables as explanatory variables, and the following output obtained:

CONTINUED

```
> model1.lm=lm(SALES ~ ADV+BONUS+SALES.REGION, data=sales.df)
> summary(model1.lm)
```

Call:

```
lm(formula = SALES ~ ADV + BONUS + SALES.REGION, data = sales.df)
```

Residuals:

Min	1Q	Median	3Q	Max
-116.984	-24.494	-1.104	35.929	101.955

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	177.2073	170.1159	1.042	0.3100
ADV	1.3678	0.2622	5.216	4.19e-05 ***
BONUS	0.9752	0.4808	2.028	0.0561 .
SALES.REGION2	48.1459	32.8013	1.468	0.1577
SALES.REGION3	257.8916	48.4129	5.327	3.26e-05 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57.63 on 20 degrees of freedom

Multiple R-Squared: 0.9468, Adjusted R-squared: 0.9362

F-statistic: 89.03 on 4 and 20 DF, p-value: 1.892e-12

```
> model2.lm=lm(SALES ~ ADV*SALES.REGION+BONUS*SALES.REGION, data=sales.df)
```

```
> anova(model1.lm, model2.lm)
```

Analysis of Variance Table

Model 1: SALES ~ ADV + BONUS + SALES.REGION

Model 2: SALES ~ ADV \* SALES.REGION + BONUS \* SALES.REGION

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	20	66414				
2	16	46487	4	19927	1.7146	0.1959

Which of the following is **true**?

- (1) Sales in region 2 are about \$48,000 less than in region 1, for comparable bonuses and advertising.
- (2) Sales in region 3 are about \$258,000 more than in region 2, for comparable bonuses and advertising.
- (3) For a district in region 2 that has advertising of \$50,000 and bonuses of \$20,000, the expected 1999 sales are about \$930,000.
- (4) The effect of increasing advertising is different for all sales regions.
- (5) The effect of increasing bonuses is the same for all sales regions.

CONTINUED

15. Geometrically, the **best** model in Q14 consists of
- (1) Three parallel planes, one for each region.
  - (2) Three parallel lines, one for each region.
  - (3) Three non-parallel planes, one for each region.
  - (4) Three non-parallel lines, one for each region.
  - (5) A curved three-dimensional surface.
16. In a logistic regression, with response  $Y = 0$  (failure) and  $Y = 1$  (success) and two continuous explanatory variables  $X$  and  $W$ , the following coefficients were obtained:

	Estimate
(Intercept)	-1.0
$X$	0.5
$W$	-0.3

Which of the following is **true**?

When  $X = 2$  and  $W = 5$ ,

- (1) To 4 decimal places, the estimated probability of a success is 0.1824.
  - (2) The estimated log-odds of a failure are -1.5.
  - (3) The estimated odds of a success are -1.5.
  - (4) The estimated mean number of successes is -1.5.
  - (5) To 4 decimal places, the estimated log-odds of a success are 0.1824.
17. The data for the next few questions were gathered in a study to investigate the factors which determine participation by married women in the labour force. The following variables were measured:
- lfp**: Wife's labour-force participation (factor with levels no, yes);
- k5**: Number of children aged 5 and under;
- age**: Wife's age in years;
- wc**: Attended college? (no, yes);
- lwg**: Log of estimated wage rate (estimate of earning ability);
- inc**: Family income in \$000, excluding wife's income if any.

The data are contained in the data frame `labour.df` in ungrouped form. There are 753 cases.

CONTINUED

A model was fitted and the following output obtained:

```
> labour.glm = glm(lfp~ k5+age+wc+lwg+inc , data=labour.df,
  family=binomial)
> summary(labour.glm)
Call:
glm(formula = lfp ~ k5 + age + wc + lwg + inc, family = binomial,
  data = labour.df)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.90193    0.54290   5.345 9.03e-08 ***
k5           -1.43180    0.19320  -7.411 1.25e-13 ***
age          -0.05853    0.01142  -5.127 2.94e-07 ***
wcyes        0.87237    0.20639   4.227 2.37e-05 ***
lwg          0.61568    0.15014   4.101 4.12e-05 ***
inc          -0.03367    0.00780  -4.317 1.58e-05 ***
Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  906.46  on 747  degrees of freedom
AIC: 918.46
```

Which of the following is an **incorrect** interpretation?

- (1) Other things being equal, women with higher earning power have greater participation in the labour force.
  - (2) Other things being equal, older women have lower participation in the labour force.
  - (3) Other things being equal, college educated women have greater participation in the labour force.
  - (4) Other things being equal, having young children discourages participation in the labour force.
  - (5) Other things being equal, the bigger the family income, the greater the participation.
18. The diagnostic plots shown in Figure 4 were obtained for the data in Q17. Which of the following is **false**?
- (1) Three women with large positive residuals participate in the labour force, even though the model suggests they don't.
  - (2) Because the data are not grouped, we can't use the deviance as a goodness of fit measure.
  - (3) Three women have relatively large Cook's distances.
  - (4) The normal plot indicates something is very wrong with the regression.
  - (5) There a small set of three women whose participation is atypical of women with similar covariates.

CONTINUED

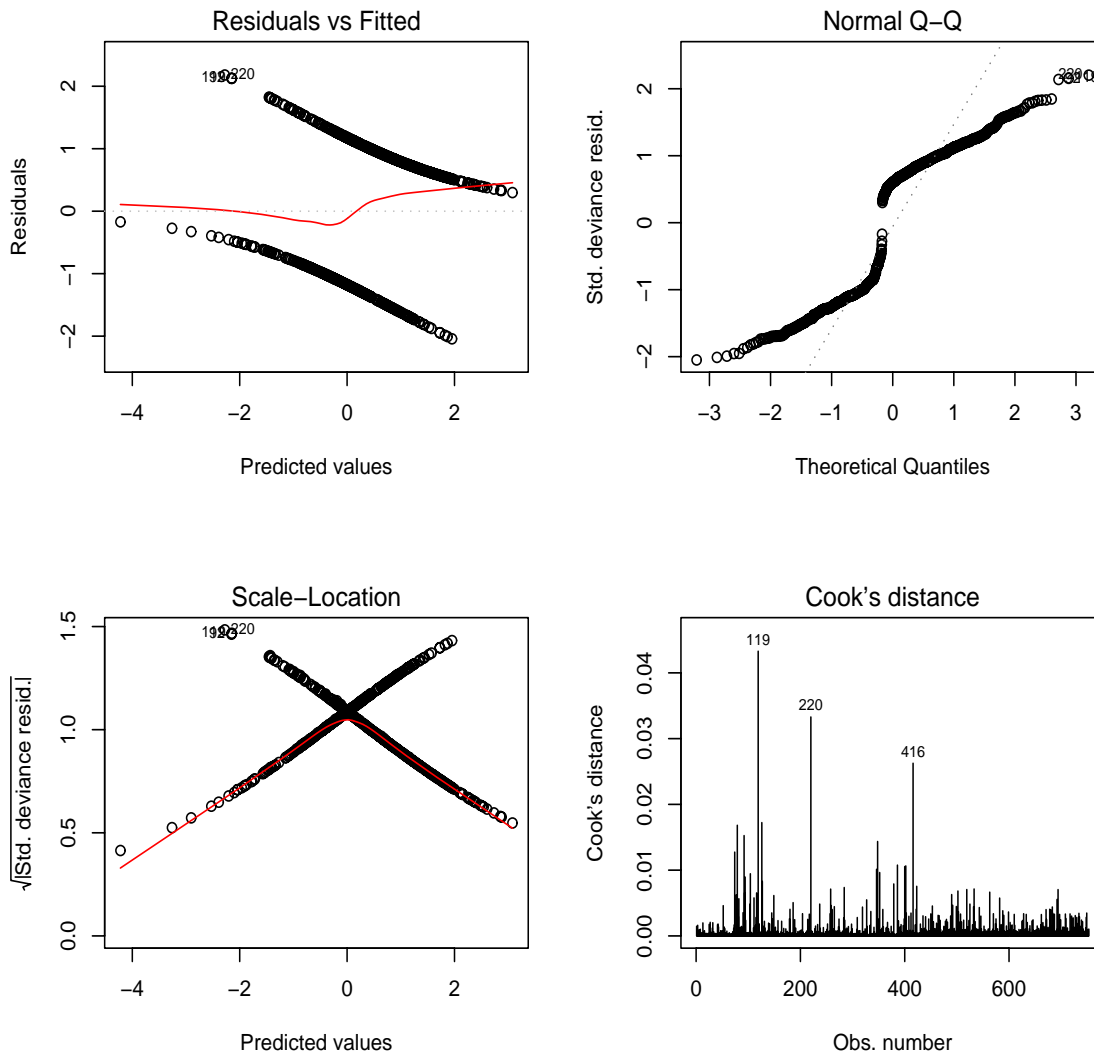


Figure 4: Diagnostic plots for Question 17.

19. The following refer to the data of Q17. Which of the following is **true**?

- (1) Other things being equal, each child 5 or under multiplies the odds of participation by a factor of 0.2389 (to 4dp).
- (2) Other things being equal, each year of age reduces the probability of participation by 0.0586 (to 4dp).
- (3) Other things being equal, each dollar of income reduces the odds of participation by 0.0337 (to 4dp).
- (4) Other things being equal, attendance at college reduces the log-odds of participation by 0.8724 (to 4dp).
- (5) Other things being equal, each child 5 or under subtracts 0.2389 off the odds of participation (to 4dp).

CONTINUED

20. In a Poisson regression, which is the **correct** interpretation?
- (1) The residual deviance cannot be used to measure goodness of fit.
  - (2) The regression coefficient measures the increase in the mean response for a unit increase in the covariate.
  - (3) The scale factor is always more than one.
  - (4) The mean is a linear function of the covariates.
  - (5) The log of the mean is a linear function of the covariates.
21. Suppose we want to model the effect of certain covariates on the death rate (expressed as deaths per 100,000 population) for a certain disease. Data for several health districts are available for the 2004 calendar year. Among the variables measured for each district are **pop** (the average population for 2004) and **deaths**, (the number of deaths due to the disease in 2004). To model the rates, we use Poisson regression with an offset. How do we express the offset in the glm function?
- (1) `offset = pop/100000`.
  - (2) `offset = deaths`.
  - (3) `offset = log(pop)`.
  - (4) `offset = log(pop/100000)`.
  - (5) `offset = pop`.
22. The data in Table 1 arose from a very large insurance company party. After an outbreak of food poisoning, a random sample of 305 attendees were asked three questions: (i) Did you feel ill after the party? (ii) Did you eat the crab? (iii) Did you eat the potato salad?

Table 1. Data for Question 22.

	Crab			
	Not eaten		Eaten	
	Potato salad		Potato salad	
	Not eaten	Eaten	Not eaten	Eaten
Not Ill	23	24	31	80
Ill	1	22	4	120

Some R output is shown below:

```
> counts = c(23,1,24,22,31,4,80,120)
> party.df = data.frame(counts=counts, expand.grid(ill=c("No", "Yes"),
  Potato=c("NotEaten", "Eaten"), Crab=c("NotEaten", "Eaten")))
> party.glm = glm(counts~ill*Potato + Potato*Crab + ill*Crab,
  family=poisson, data=party.df)
> summary(part.glm)
Call:
glm(formula = counts ~ ill*Potato + Potato*Crab + ill*Crab,
```

CONTINUED

```

family = poisson,
data = party.df)

```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	3.11490	0.20656	15.080	< 2e-16	***
IllYes	-2.73048	0.51694	-5.282	1.28e-07	***
PotatoEaten	0.08249	0.28093	0.294	0.7690	
CrabEaten	0.33409	0.26666	1.253	0.2103	
IllYes:PotatoEaten	2.60259	0.48747	5.339	9.35e-08	***
IllYes:CrabEaten	0.54313	0.31304	1.735	0.0827	.
PotatoEaten:CrabEaten	0.84466	0.33908	2.491	0.0127	*
Null deviance: 285.97418 on 7 degrees of freedom					
Residual deviance: 0.26879 on 1 degrees of freedom					

On the basis of this output, which of the following models is the most appropriate?

- (1)  $\text{counts} \sim \text{Ill} * \text{Potato} * \text{Crab}$ .
- (2)  $\text{counts} \sim \text{Potato} * \text{Crab}$ .
- (3)  $\text{Ill} \sim \text{Potato} * \text{Crab}$ .
- (4)  $\text{counts} \sim \text{Ill} * \text{Potato} + \text{Potato} * \text{Crab} + \text{Ill} * \text{Crab}$ .
- (5)  $\text{counts} \sim \text{Ill} + \text{Potato} + \text{Crab}$ .

23. On the basis of the output above, which of the following is **false**?

- (1) People who ate crab tended not to eat potato salad, for both the Ill and Not ill groups.
- (2) There is weak evidence that eating crab made people ill, for the people who ate potato salad and those who did not.
- (3) The residual deviance indicates that the model fits well.
- (4) The population odds ratio for being ill and eating potato salad is the same for those who ate crab and those who did not.
- (5) There is very strong evidence that eating potato salad made people ill, for the people who ate crab and those who did not.

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

- (1) A 95% confidence interval for the conditional log odds ratio between illness and eating potato salad, given that crab was eaten, is (5.192, 35.094).
- (2) A 95% confidence interval for the conditional odds ratio between between illness and eating crab, given that potato salad was eaten, is (0.932, 3.179).
- (3) A 95% confidence interval for the odds ratio between illness and eating potato salad is (5.192, 35.094).
- (4) Since 1 is contained in the confidence interval for the conditional odds ratio between being ill and eating crab, given potato salad was eaten, eating crab and being ill are independent.

CONTINUED

- (5) A 95% confidence interval for the log odds ratio between eating crab and eating potato salad is (0.180, 1.509).
25. A set of 5 coins was tossed 3590 times, and the number of heads (either 0,1,2,3,4 or 5) was recorded each time. The results are shown in Table 2.

Number of heads	Frequency
0	100
1	524
2	1080
3	1126
4	655
5	105

Theory predicts that the number of heads will have a binomial distribution with parameters  $n = 5$  and  $p = 0.5$ . An analysis using R produced the following output:

```
> counts = c(100,524,1080,1126,655,105)
> phat=sum((0:5)*counts)/(sum(counts)*5)
> L.sat=sum(counts* log(counts/sum(counts)))
> L.mod1 = sum(counts* log(dbinom(0:5,5, phat)))
> L.mod2 = sum(counts* log(dbinom(0:5,5, 0.5)))
> D1 = 2*(L.sat - L.mod1)
> D2 = 2*(L.sat - L.mod2)
> D3 = 2*(L.mod1-L.mod2)
> 1-pchisq(D1,4)
[1] 0.06380235
> 1-pchisq(D2,5)
[1] 0.0008498923
> 1-pchisq(D3,1)
[1] 0.0005332796
```

(Note that the function `dbinom(x, n, p)` calculates the probability of getting  $x$  heads when  $n$  coins are tossed, and the probability of a head on each toss is  $p$ .)

Which if the following is **true**?

- (1) The p-value for the test that the binomial model (with  $p$  unspecified) is appropriate is 0.0008498923.
- (2) The binomial model (with  $p$  unspecified) is a barely acceptable fit to these data.
- (3) The p-value for the test that the binomial model with  $p = 0.5$  is appropriate is 0.06380235.
- (4) The binomial model with  $p = 0.5$  is a very good fit to these data.
- (5) The p-value for the test that the binomial model (with  $p$  unspecified) is appropriate is 0.0005332796.

CONTINUED

## SECTION B

26. (a) What is meant by the term “collinearity” in linear regression? If present, what effect does collinearity have on the estimated regression coefficients? [5 marks]
- (b) What is meant by the “variance inflation factor” in linear regression? Describe two methods of calculating it. [6 marks]

The rest of Question 26 refers to the following situation: Aerobic fitness (measured by the ability to consume oxygen) is an important aspect of athletic performance. However, it is difficult to measure and expensive equipment is required. One alternative is to develop a regression model to predict fitness based on simple exercise tests rather than on expensive and cumbersome oxygen consumption measurements.

To fit such a model, measurements were made on 31 men involved in a physical fitness course at North Carolina State University. The variables are **Age** (years), **Weight** (kg), **Oxygen** (oxygen intake rate in ml per kg body weight per minute, the response), **Runtime** (time to run 1.5 miles in minutes), **RestPulse** (heart rate while resting), **RunPulse** (heart rate while running) and **MaxPulse** (maximum heart rate recorded while running).

A regression model was fitted and the following output was obtained.

Call:

```
lm(formula = Oxygen ~ Age + Weight + RunTime + RestPulse + RunPulse +
    MaxPulse, data = fitness.df)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	102.93448	12.40326	8.299	1.64e-08	***
Age	-0.22697	0.09984	-2.273	0.03224	*
Weight	-0.07418	0.05459	-1.359	0.18687	
RunTime	-2.62865	0.38456	-6.835	4.54e-07	***
RestPulse	-0.02153	0.06605	-0.326	0.74725	
RunPulse	-0.36963	0.11985	-3.084	0.00508	**
MaxPulse	0.30322	0.13650	2.221	0.03601	*

Residual standard error: 2.317 on 24 degrees of freedom  
 Multiple R-Squared: 0.8487, Adjusted R-squared: 0.8108  
 F-statistic: 22.43 on 6 and 24 DF, p-value: 9.715e-09

Correlations

	Age	Weight	Oxygen	RunTime	RestPulse	RunPulse	MaxPulse
Age	1.000	-0.234	-0.305	0.189	-0.164	-0.338	-0.433
Weight	-0.234	1.000	-0.163	0.144	0.044	0.182	0.249
Oxygen	-0.305	-0.163	1.000	-0.862	-0.399	-0.398	-0.237
RunTime	0.189	0.144	-0.862	1.000	0.450	0.314	0.226
RestPulse	-0.164	0.044	-0.399	0.450	1.000	0.352	0.305
RunPulse	-0.338	0.182	-0.398	0.314	0.352	1.000	0.930
MaxPulse	-0.433	0.249	-0.237	0.226	0.305	0.930	1.000

Variance inflation factors

CONTINUED

	Age	Weight	RunTime	RestPulse	RunPulse	MaxPulse
	1.512836	1.155329	1.590868	1.415589	8.437274	8.743848

- (c) The VIF's for RunPulse and Maxpulse are quite high. What is causing this? [3 marks]
- (d) The p-value for RestPulse is high. What is causing this? [3 marks]
- (e) Some further output is shown below. Which model would you use to predict the oxygen uptake? Give reasons. [3 marks]

	rssp	sigma2	adjRsq	Cp	AIC	BIC	CV	Age	Weight	RunTime	RestPulse	RunPulse	MaxPulse
1	218.481	7.534	0.735	13.699	44.699	47.567	24.328	0	0	1	0	0	0
2	200.716	7.168	0.747	12.389	43.389	47.691	24.217	1	0	1	0	0	0
3	160.831	5.957	0.790	6.960	37.960	43.696	19.934	1	0	1	0	1	0
4	138.930	5.343	0.812	4.880	35.880	43.050	18.415	1	0	1	0	1	1
5	129.408	5.176	0.818	5.106	36.106	44.710	17.827	1	1	1	0	1	1
6	128.838	5.368	0.811	7.000	38.000	48.038	19.044	1	1	1	1	1	1

27. (a) Define the two kinds of residuals we have discussed in connection with the logistic regression model. [5 marks]
- (b) Describe how you would assess the fit of a logistic regression model in the case of (i) grouped data, and (ii) ungrouped data. Your discussion should include an account of the role of the deviance and of the residuals. [5 marks]

The data (which are in ungrouped form) for the next two parts of Question 27 come from a study of the analgesic effects of 3 treatments on 60 elderly patients with neuralgia. Two test treatments (A and B) and a placebo (P) are compared. The response variable (Pain) is whether the patient reported pain or not after treatment, with levels "No" and "Yes". Researchers recorded the age and gender of the patients and the duration of the complaint before the treatment began. A logistic model was fitted and the following output obtained, as well as the plots in Figure 5.

```
> pain.glm = glm(Pain~Treatment+Sex+Age+Duration,
                 family=binomial, data=pain.df)
> summary(pain.glm)
Call:
glm(formula = Pain ~ Treatment + Sex + Age + Duration,
     family = binomial, data = pain.df)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-20.588282	7.102883	-2.899	0.00375 **
TreatmentB	-0.526853	0.937025	-0.562	0.57394
TreatmentP	3.181690	1.016021	3.132	0.00174 **
SexM	1.832202	0.796206	2.301	0.02138 *
Age	0.262093	0.097012	2.702	0.00690 **
Duration	-0.005859	0.032992	-0.178	0.85905

Null deviance: 81.503 on 59 degrees of freedom  
Residual deviance: 48.736 on 54 degrees of freedom

CONTINUED

```

> HLstat(pain.glm)
Value of HL statistic = 8.498
P-value = 0.386

```

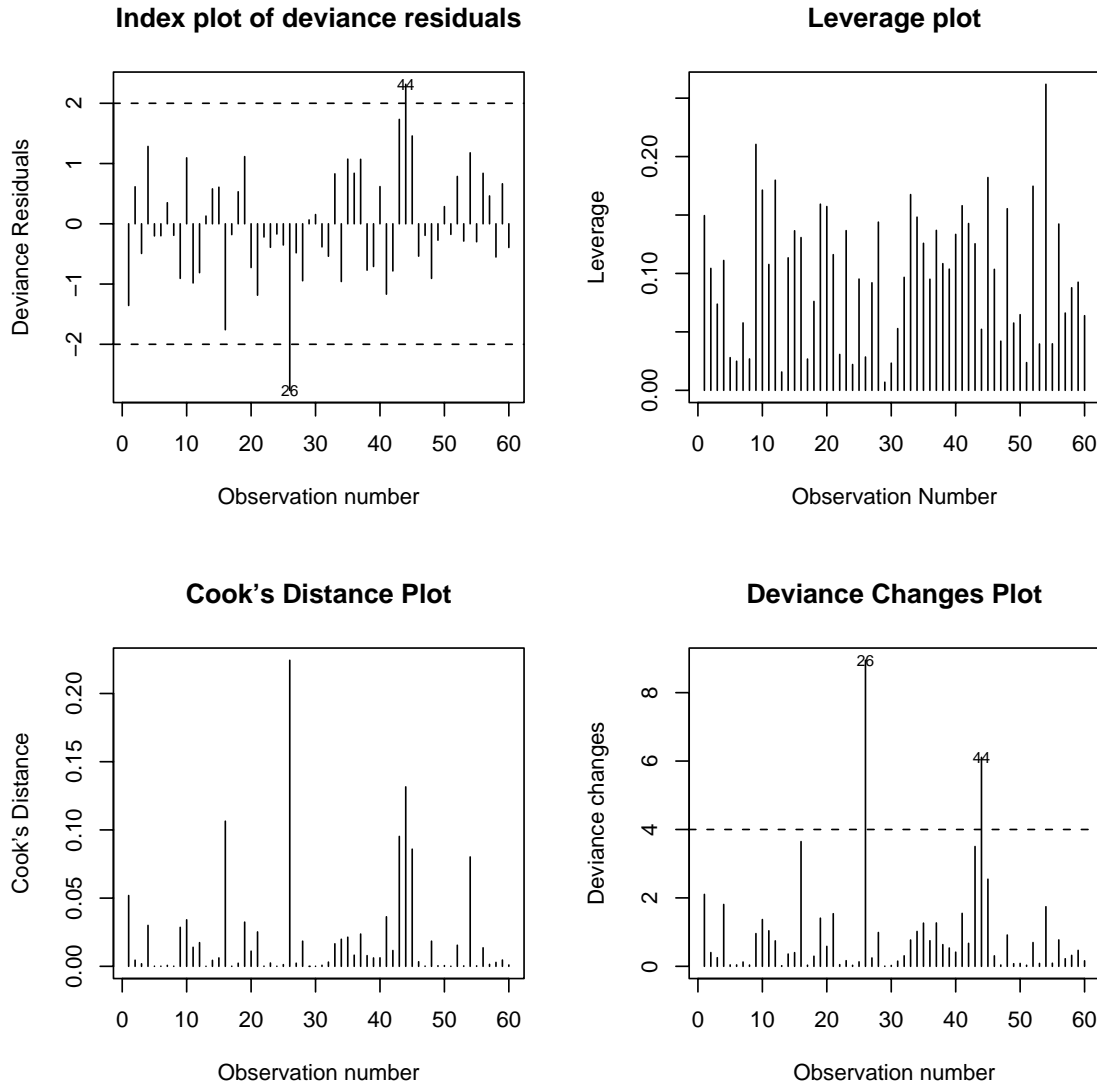


Figure 5: Plot for Q27.

- (c) Do you think this model fits well? Give reasons for your answer. [5 marks]
- (d) Is there any evidence that the two treatments A and B are any better than the placebo? Is B better than A? Give reasons. [5 marks]
28. (a) Describe the Poisson regression model, clearly outlining the assumptions behind the model.
- (b) Suppose we have a contingency table with factors A, B and C. State how you would use a Poisson regression model to check the following: (i) that A is independent of B and C, (ii) that A and B are conditionally independent, given C. [6 marks]

The rest of Question 28 concerns the following: Suppose you have collected marketing research data to examine the relationship between a prospect's likelihood

CONTINUED

of buying your product and their education and income. Specifically, the variables are as follows:

Variable	Levels	Interpretation
Education	low, high	Prospect's education level
Income	low, high	Prospect's income level
Purchase	no,yes	Did prospect purchase product?

The data shown in the data frame `purchase.df` below were collected on 234 prospects:

```
> purchase.df
  Education Income Purchase Count
1     high   high     yes     54
2     high   high     no     23
3     high   low     yes     41
4     high   low     no     12
5     low   high     yes     35
6     low   high     no     42
7     low   low     yes     19
8     low   low     no      8
```

A model was fitted to the data and the following output obtained:

```
> purchase.glm =glm(Count~Education*Income+ Education*Purchase
+Income*Purchase , family=poisson, data=purchase.df)
> anova(purchase.glm, 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|)
              Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                                7      67.280
Education              1      2.895              6      64.386      0.089
Income                 1     23.808              5      40.577 1.064e-06
Purchase               1     17.729              4      22.848 2.547e-05
Education:Income       1      5.711              3      17.137      0.017
Education:Purchase     1     11.195              2       5.943      0.001
Income:Purchase        1      4.803              1       1.140      0.028
```

```
> summary(purchase.glm)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.27220    0.27577   8.239 < 2e-16 ***
Educationhigh    0.05987    0.31342   0.191  0.84851
Incomehigh      1.42413    0.29374   4.848 1.25e-06 ***
Purchaseyes     0.57846    0.30828   1.876  0.06060 .
Educationhigh:Incomehigh -0.54937    0.29356  -1.871  0.06129 .
Educationhigh:Purchaseyes  0.84368    0.28274   2.984  0.00285 **
Incomehigh:Purchaseyes -0.67200    0.31238  -2.151  0.03146 *
```

CONTINUED

```
Null deviance: 67.2804 on 7 degrees of freedom
Residual deviance: 1.1400 on 1 degrees of freedom
> 1-pchisq(1.1400,1)
[1] 0.2856523
```

- (c) Which model would you fit to this contingency table? Give reasons. [3 marks]
  - (d) Draw the association graph for the model you select. [3 marks]
  - (e) Are both the variables Income and Education associated with purchasing the product? If so, how? [3 marks]
-