DEPARTMENT OF STATISTICS

STATS 762: Topics in Regression Modelling

Term Test Friday October 12, 2007

INSTRUCTIONS: Answer **ALL 10** questions

1. An experiment was conducted to determine the joint effects of temperature and concentration of herbicide on the absorption of two herbicides A and B on a commercial charcoal material. Two temperatures (10 deg 55 deg) and 5 concentrations (20, 40, 60 80 and 100 g/l) were used twice on each herbicide, so that there were 40 observations in all. Shown below are parts of an analysis of these data. There are three categorical variables `herb`, `temp` and `conc`, and a continuous response `absorb`.

```
> q14.lm<-lm(absorb~herb*temp*conc, data=q14.df)
> summary(q14.lm)
..... some output omitted...

Residual standard error: 0.007204 on 20 degrees of freedom
Multiple R-Squared: 0.9975,     Adjusted R-squared: 0.9952
F-statistic: 425.3 on 19 and 20 DF,  p-value: < 2.2e-16

> anova(q14.lm)
Analysis of Variance Table
Response: absorb
                     Df   Sum Sq  Mean Sq   F value    Pr(>F)
herb                  1 0.137593 0.137593 2651.1156 < 2.2e-16 ***
temp                  1 0.002958 0.002958   57.0019 2.818e-07 ***
conc                  4 0.267424 0.066856 1288.1708 < 2.2e-16 ***
herb:temp             1 0.006452 0.006452  124.3083 4.928e-10 ***
herb:conc             4 0.003104 0.000776   14.9511 8.301e-06 ***
temp:conc             4 0.000914 0.000228    4.4020   0.01028 *
herb:temp:conc        4 0.000955 0.000239    4.5985   0.00852 **
Residuals            20 0.001038 0.000052
```

What hypothesis is being tested by the p-value 0.00852? Express your answer in terms of the effects of changing levels of the factors.

2. In a regression of a continuous response Y on a factor A (having 3 levels) and a continuous covariate X, the following output was obtained:

```
Call:  lm(formula = y ~ A * X)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.99608    0.03481  28.613  < 2e-16 ***
A2           1.05492    0.05017  21.027  < 2e-16 ***
A3           0.99608    0.04668  43.468  < 2e-16 ***
X           -0.48346    0.06686  -7.231 1.27e-10 ***
A2:X        -0.34873    0.09368  -3.722 0.000336 ***
A3:X        -0.57873    0.08592  -6.736 1.29e-09 ***
```

Describe the effect on the mean response when X is increased by one unit.

3. In a Poisson regression of a count response Y on a continuous explanatory variable X, using the R code `glm(Y~X, family=poisson)`, the regression coefficient of X is 0.3. What is the effect on the mean response if X is increased by one?

4. An experiment has been done to test the effect of an insecticide on a certain species of beetle. For each of eight prespecified dosages, a number (n) of beetles were exposed to the insecticide and the number dying (r) was recorded. The following data were obtained:

logdose	r	n
1.6907	5	59
1.7242	11	60
1.7552	31	62
1.7842	37	56
1.8113	48	63
1.8369	55	59
1.8610	61	62
1.8839	59	60

A logistic model of the form

$$\log \frac{\pi}{1-\pi} = \alpha + \beta \times \text{logdose}$$

(where π is the probability a beetle will die) was fitted to the data with the following results:

```
> q4.glm = cbind(r, n - r) ~ logdose, family = binomial)
Call: glm(formula = cbind(r, n - r) ~ logdose, family = binomial)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -57.375      5.018  -11.43   <2e-16 ***
logdose       32.588      2.834   11.50   <2e-16 ***
---
    Null deviance: 253.1354  on 7  degrees of freedom
Residual deviance:   4.3618  on 6  degrees of freedom
AIC: 36.985
```

Calculate the predicted probability that a beetle subjected to a logdose of 1.6907 will die.

5. In the beetle example in Question 4, the following additional output was obtained:

```
> predict(q4.glm, data.frame(logdose=1.8113), se=T)
$fit
[1] 1.65089
$se.fit
[1] 0.1747074
> qnorm(0.975)
[1] 1.959964
```

Calculate a 95% confidence interval for the log-odds of death at log dose 1.8113.

6. The beetle experiment was repeated, using a modified form of the insecticide (treatment 1) as well as the previous form (treatment2), but keeping the same dosages. We now have data

```
 logdose  r  n treat
  1.6907  7 59     1
  1.7242 17 60     1
  1.7552 21 62     1
  1.7842 41 56     1
  1.8113 52 63     1
  1.8369 57 59     1
  1.8610 61 62     1
  1.8839 59 60     1
  1.6907  9 64     2
  1.7242 21 59     2
  1.7552 41 63     2
  1.7842 48 64     2
  1.8113 55 59     2
  1.8369 54 58     2
  1.8610 57 57     2
  1.8839 64 64     2
```

We obtained the following "anova".

```
Analysis of Deviance Table
Model: binomial, link: logit
Response: cbind(r, n - r)
Terms added sequentially (first to last)
              Df   Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                                15     511.23
logdose        1     487.62        14      23.61 4.694e-108
treat          1       9.39        13      14.22  2.184e-03
logdose:treat  1  3.073e-03        12      14.21       0.96
```

Is the effect of increasing the dose the same for the two treatments? Give a reason for your answer.

7. In a logistic regression for ungrouped data, with 3 explanatory variables, the residual deviance was 134.986 on 96 degrees of freedom, while the null deviance was 137.989 on 99 degrees of freedom. The following additional output was obtained:

```
> 1-pchisq(137.989,99)
[1] 0.005884545
>  1-pchisq(134.986,96)
[1] 0.005384848
>  1-pchisq(3.003,3)
[1] 0.3911629
```

Do any of the explanatory variables have a relationship with the response? Give a reason for your answer.

8. In a three-dimensional contingency table with factors A, B and C, write down the R model formula which expresses the idea that the conditional odds ratios for factors A and B are the same for all levels of factor C.

9. In the Florida murder data discussed in class, defendents convicted of murder were classified according to three factors; namely race of the defendent (`defendent`), race of the victim (`victim`), and whether or not the death penalty was imposed (`dp`). The following anova was obtained:

```
> murder.glm<-glm(count~defendant*dp*victim,family=poisson,
   data=murder.df)
> anova(murder.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|)
NULL                                    7     774.73
defendant            1     0.22         6     774.51      0.64
dp                   1   443.51         5     331.00 1.861e-98
victim               1    64.10         4     266.90 1.183e-15
defendant:dp         1     0.43         3     266.47      0.51
defendant:victim     1   254.15         2      12.32 3.230e-57
dp:victim            1    12.30         1       0.02 4.535e-04
defendant:dp:victim  1     0.02         0  3.997e-15      0.89
```

```
> summary(murder.glm)
Call:
glm(formula = count ~ defendant * dp * victim, family = poisson,
    data = murder.df)
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)               2.5649     0.2774   9.248  < 2e-16 ***
defendantw               -2.5649     1.0377  -2.472  0.01345 *
dpno                      2.7081     0.2864   9.454  < 2e-16 ***
victimw                   0.5705     0.3470   1.644  0.10012
defendantw:dpno           0.2364     1.0652   0.222  0.82438
defendantw:victimw        3.0930     1.0705   2.889  0.00386 **
dpno:victimw             -1.1896     0.3675  -3.237  0.00121 **
defendantw:dpno:victimw   0.1613     1.1032   0.146  0.88375
---
Null deviance: 7.7473e+02  on 7  degrees of freedom
Residual deviance: 3.9968e-15  on 0  degrees of freedom
```

Which model is indicated by this output? Give a reason.

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