

This practical contributes 8.5% to your raw BS1 total mark. It is due in by 12 noon Tuesday week 7, Hilary Term 2011, at the Statistics Department reception, 1 South Parks Road. Please include the R-code you write (either as commented code in an Appendix, or in the text).

- (1) The ‘impact’ data from

`http://www.statsci.org/data/general/insulate.html`

give the impact strength (in foot-pounds) of $n = 100$ pieces of insulating material. The material was produced in five lots, and the pieces made by cutting off pieces of material. Two different cuts were used: with and across the grain. We wish to model effects influencing impact strength.

```
> file<-’http://www.stats.ox.ac.uk/~nicholls/bs1a/data/impact.txt’
> imp<-read.table(file,header=TRUE)
> imp
  Lot    Cut Strength
1    1  Cross    0.46
2    1  Cross    0.67
...
99   5 Length    0.67
100  5 Length    0.72
```

In this question you will carry out a Monte Carlo test for goodness of fit for a normal linear regression model.

- What kind of experimental design is in use here? How might this help or hinder an ANOVA analysis?
- Fit the normal linear model $\text{Strength} \sim \text{Cut} * \text{Lot}$ and calculate the fitted values \hat{y} , the residual standard error s , and the studentised residuals, and make a goodness of fit analysis.
- The `rnorm(n)` function simulates n independent standard normal random variables. Explain how to simulate n new independent observations, $Y'_i \sim N(\hat{y}_i, s^2)$, $i = 1, 2, \dots, n$, using R .
- Under the normal linear model in 1(b), the distribution of the studentised residuals should be approximately $N(0, 1)$. Compute the Kolmogorov-Smirnov test statistic for the studentised residuals from the original fit in 1(b). (The `pnorm(z)` function gives the cdf $\Phi(z) = \Pr(Z < z)$ for Z a standard normal random variable.)
- Make a Monte-Carlo KS-test for goodness-of-fit of the studentised residuals to a standard normal distribution: compare the KS test-statistic for the studentised residuals from 1(b) to the KS test-statistics of sets of studentised residuals obtained by fitting simulated data, and compute a Monte-Carlo p -value. Interpret your result.
- When we inspect the Cook’s Distances, we used a threshold $8/(n-2p)$ (where p is the the number of variables in the linear model) in order to identify possible outliers. Estimate the probability to have at least one point exceeding this threshold in the normal linear model in 1(b). How concerned are you about points of high influence in 1(b)? Justify your answer.

- (2) The data in this example were gathered on undergraduates applying to a certain USA graduate school. Student Grade Point Averages, undergraduate university status ('topnotch' or not) and Graduate Record Exam score are recorded. GRE is rounded up to the nearest 50 marks, and GPA up to the nearest fifth. The number of students applying with each combination of GRE, GPA and university status, and the number admitted, are recorded. Thus (row 97) seven students applied from topnotch universities, with GRE 650-700 and GPA 3.8-4. Of these two were admitted.

```
> file<- 'http://www.stats.ox.ac.uk/~nicholls/bs1a/data/admits.txt'
> (adm<-read.table(file,header=T))
  topnotch gre gpa admit applied
1         0 450 2.4     0       1
2         0 400 2.6     0       1
...
97        1 700 4.0     2       7
98        1 750 4.0     0       2
99        1 800 4.0     4       6
```

Consider a logistic regression for `admit`, the number of successes in `applied` trials, with linear predictor $\text{admit} \sim \text{gre} + \text{gpa} + \text{topnotch}$ (in R formula notation). We fit as follows.

```
> formula<-cbind(admit,applied-admit)~gre+gpa+topnotch
> adm.glm<-glm(formula,family=binomial,data=adm)
> summary(adm.glm)
...
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.854244    1.157877  -4.192 2.76e-05 ***
gre          0.002781    0.001087   2.559 0.0105 *
gpa          0.657245    0.334836   1.963 0.0497 *
topnotch     0.436155    0.291899   1.494 0.1351
...
Residual deviance: 122.97  on 95  degrees of freedom
```

In this question you will carry out a Monte Carlo test for goodness of fit for a logistic model. Denote by $x_{i,2}, x_{i,3}$ and $x_{i,4}$ the variables corresponding to intercept, `gre`, `gpa` and `topnotch` in the i th row of the data, and by $\beta_1, \beta_2, \beta_3$ and β_4 the corresponding logistic regression parameters.

- Calculate the MLE $\hat{p}_i = p(\hat{\beta}_1 + \hat{\beta}_2 x_{i,2} + \hat{\beta}_3 x_{i,3} + \hat{\beta}_4 x_{i,4})$ for the probability for acceptance given the covariates in row $i = 1, 2, \dots, n$ (using `predict()`, or otherwise).
- Show how to simulate new values for the `admit` response variable, so that $\text{admit}[i] \sim \text{Binomial}(\hat{p}_i, \text{applied}[i])$.
- Carry out a Monte Carlo test for goodness of fit for the GLM specified in the code above, by comparing its residual deviance to the distribution of the residual deviance under the model $\text{admit}[i] \sim \text{Binomial}(\hat{p}_i, \text{applied}[i])$.
- Explain briefly how to carry out a goodness of fit test for a GLM, when we use a χ^2 approximation for the distribution of the residual deviance under the null. Why is it necessary to use Monte Carlo to test for goodness of fit to the data in this question?