Final Homework

Math 424 Kafai by Jonathan Dombrowski

Q1.

Consider the data in q1data.txt. All subjects are asthmatics. For the model with Forced Expiratory Volume (FEV) as the response and Height, Weight, and Age as the predictors,

a) Examine a plot of the studentized or jackknife residuals versus the predicted values. Are any regression assumption violations apparent? If so, suggest possible remedies.

b) Examine numerical descriptive statistics, histograms, box-and-whisker plots, and normal probability plots of jackknife residuals. Is the normality assumption violated? If so, suggest possible remedies.

c) Examine outlier diagnostics, including Cook's distance, leverage statistics, and jackknife residuals, and identify any potential outliers. What course of action, if any, should be taken when outliers are identified?

d) Examine variance inflation factors, condition i.Ildices (unadjusted and adjusted for the intercept), and variance proportions. Are there any important collinearity problems? If so, suggest possible remedies.

Work

loading in data

```r
df <- read.table(file = "q1data.txt", header = TRUE)
df
```

```
##    age sex Height Weight FEV
## 1   yr m/f     cm     kg   L
## 2   24   M    175   78.0 4.7
## 3   36   M    172   67.6 4.3
## 4   28   F    171   98.0 3.5
## 5   25   M    166   65.5 4.0
## 6   26   F    166   65.0 3.2
## 7   22   M    176   65.5 4.7
## 8   27   M    185   85.5 4.3
## 9   27   M    171   76.3 4.7
## 10  36   M    185   79.0 5.2
## 11  24   M    182   88.2 4.2
## 12  26   M    180   70.5 3.5
## 13  29   M    163   75.0 3.2
## 14  33   F    180   68.0 2.6
## 15  31   M    180   65.0 2.0
## 16  30   M    180   70.4 4.0
## 17  22   M    168   63.0 3.9
## 18  27   M    168   91.2 3.0
## 19  46   M    178   67.0 4.5
## 20  36   M    173   62.0 2.4
```

Q2.

A random sample of data was collected on residential sales in a large city. The data in q2data.txt shows the selling price (Y, in \$1,000s), area (x1, in hundreds of square feet), number of bedrooms (X2), total number of rooms (X3), house age (X4, in years), and location (Z = 0 for in-town and inner suburbs, Z=1for outer suburbs). In parts a through c, use variables X1, X2, X3, X4, and Z as the predictor variables.

a) Use the all possible regressions procedure to suggest a best model.

b) Use the stepwise regression algorithm to suggest a best model.

c) Use the backward elimination algorithm to suggest a best model.

d) Which of the models selected in a, b, and c seems to be the best model, and why?

Work

loading in data

```r
df <- read.table(file = "q2data.txt", header = TRUE)
df
```

```
##    House    y   x1 x2 x3 x4 Z
## 1      1 84.0 13.8  3  7 10 0
## 2      2 93.0 19.0  2  7 22 1
## 3      3 83.1 10.0  2  7 15 1
## 4      4 85.2 15.0  3  7 12 1
## 5      5 85.2 12.0  3  7  8 1
## 6      6 85.2 15.0  3  7 12 1
## 7      7 85.2 12.0  3  7  8 1
## 8      8 63.3  9.1  3  6  2 1
## 9      9 84.3 12.5  3  7 11 1
## 10    10 84.3 12.5  3  7 11 1
## 11    11 77.4 12.0  3  7  5 0
## 12    12 92.4 17.9  3  7 18 0
## 13    13 92.4 17.9  3  7 18 0
## 14    14 61.5  9.5  2  5  8 0
## 15    15 88.5 16.0  3  7 11 0
## 16    16 88.5 16.0  3  7 11 0
## 17    17 40.6  8.0  2  5  5 0
## 18    18 81.6 11.8  3  7  8 1
## 19    19 86.7 16.0  3  7  9 0
## 20    20 89.7 16.8  2  7 12 0
## 21    21 86.7 16.0  3  7  9 0
## 22    22 89.7 16.8  2  7 12 0
## 23    23 75.9  9.5  3  6  6 1
## 24    24 78.9 10.0  3  6 11 0
## 25    25 87.9 16.5  3  7 15 0
## 26    26 91.0 15.1  3  7  8 1
## 27    27 92.0 17.9  3  8 13 1
## 28    28 87.9 16.5  3  7 15 0
## 29    29 90.9 15.0  3  7  8 1
## 30    30 91.9 17.8  3  8 13 1
```

Q3.

The data listed in q3data.txt relate to a study by Reiter and others concerning the effects of injecting triethyl-tin (TET) into rats once at age 5 days. The animals were injected with 0, 3, or 6 mg per kilogram of body weight. The response was the log of the activity count, log (ac), for 1 hour, recorded at 21 days of age. The rat was left to move about freely in a figure 8 maze. Analysis of other studies with this type of activity count confirms that log counts should yield Gaussian errors if the model is correct.

a) Conduct a two-way ANOVA with SEX and DOSAGE as factors.

b) Using = .05, report your conclusions based on the ANOVA.

c) Which, if any, families of means should be followed up with multiple-comparison tests? What type of comparisons would you recommend?

Work

loading in data

```r
df <- read.table(file = "q3data.txt", header = TRUE)
df
```

```
##    log.ac. Dosage    Sex
## 1    2.636      0   Male
## 2    2.736      0   Male
## 3    2.775      0   Male
## 4    2.672      0   Male
## 5    2.653      0   Male
## 6    2.569      0   Male
## 7    2.737      0   Male
## 8    2.588      0   Male
## 9    2.735      0   Male
## 10   2.444      3   Male
## 11   2.744      3   Male
## 12   2.207      3   Male
## 13   2.851      3   Male
## 14   2.533      3   Male
## 15   2.630      3   Male
## 16   2.688      3   Male
## 17   2.665      3   Male
## 18   2.517      3   Male
## 19   2.769      3   Male
## 20   2.694      6   Male
## 21   2.845      6   Male
## 22   2.865      6   Male
## 23   3.001      6   Male
## 24   3.043      6   Male
## 25   3.066      6   Male
## 26   2.747      6   Male
## 27   2.894      6   Male
## 28   1.851      6   Male
## 29   2.489      6   Male
## 30   2.494      0 Female
## 31   2.723      0 Female
## 32   2.841      0 Female
## 33   2.620      0 Female
```

```
## 34   2.682       0 Female
## 35   2.644       0 Female
## 36   2.684       0 Female
## 37   2.607       0 Female
## 38   2.591       0 Female
## 39   2.737       0 Female
## 40   2.220       3 Female
## 41   2.371       3 Female
## 42   2.679       3 Female
## 43   2.591       3 Female
## 44   2.942       3 Female
## 45   2.473       3 Female
## 46   2.814       3 Female
## 47   2.622       3 Female
## 48   2.730       3 Female
## 49   2.955       3 Female
## 50   2.540       6 Female
## 51   3.113       6 Female
## 52   2.468       6 Female
## 53   2.606       6 Female
## 54   2.764       6 Female
## 55   2.859       6 Female
## 56   2.763       6 Female
## 57   3.000       6 Female
## 58   3.111       6 Female
## 59   2.858       6 Female
```

Q4.

The data in q4data.txt is the record of coronary artery disease (ca, 0=no, 1= yes), age, ECG (0, 1, and 2 based on the reading of ST segment depression), and sex (0=male, 1=female). Based on this model

a) What is the estimated logistic regression model for the relationship between ca and age, ECG, sex?

b) What is a 30-year-old male, ECG=2 predicted probability of having coronary artery disease?

c) Estimate the odds ratio comparing a 30-year-old male, ECG=2 to a 31-year-old male, ECG=2. Interpret this estimated odds ratio.

d) Find a 95% confidence interval for the population odds ratio being estimated in part (c).

Work

loading in data

```r
df <- read.table(file = "q4data.txt", header = TRUE)
df
```

```
##    sex ecg age ca
## 1    0   0  28  0
## 2    1   0  42  1
## 3    0   1  46  0
## 4    1   1  45  0
## 5    0   0  34  0
## 6    1   0  44  1
```

```
## 7     0    1   48   1
## 8     1    1   45   1
## 9     0    0   38   0
## 10    1    0   45   0
## 11    0    1   49   0
## 12    1    1   45   1
## 13    0    0   41   1
## 14    1    0   46   0
## 15    0    1   49   0
## 16    1    1   46   1
## 17    0    0   44   0
## 18    1    0   48   0
## 19    0    1   52   0
## 20    1    1   48   1
## 21    0    0   45   1
## 22    1    0   50   0
## 23    0    1   53   1
## 24    1    1   57   1
## 25    0    0   46   0
## 26    1    0   52   1
## 27    0    1   54   1
## 28    1    1   57   1
## 29    0    0   47   0
## 30    1    0   52   1
## 31    0    1   55   0
## 32    1    1   59   1
## 33    0    0   50   0
## 34    1    0   54   0
## 35    0    1   57   1
## 36    1    1   60   1
## 37    0    0   51   0
## 38    1    0   55   0
## 39    0    2   46   1
## 40    1    1   63   1
## 41    0    0   51   0
## 42    1    0   59   1
## 43    0    2   48   0
## 44    1    2   35   0
## 45    0    0   53   0
## 46    1    0   59   1
## 47    0    2   57   1
## 48    1    2   37   1
## 49    0    0   55   1
## 50    1    1   32   0
## 51    0    2   60   1
## 52    1    2   43   1
## 53    0    0   59   0
## 54    1    1   37   0
## 55    1    0   30   0
## 56    1    2   47   1
## 57    0    0   60   1
## 58    1    1   38   1
## 59    1    0   34   0
## 60    1    2   48   1
```

```
## 61    0    1  32  1
## 62    1    1  38  1
## 63    1    0  36  1
## 64    1    2  49  0
## 65    0    1  33  0
## 66    1    1  42  1
## 67    1    0  38  1
## 68    1    2  58  1
## 69    0    1  35  0
## 70    1    1  43  0
## 71    1    0  39  0
## 72    1    2  59  1
## 73    0    1  39  0
## 74    1    1  43  1
## 75    1    0  42  0
## 76    1    2  60  1
## 77    0    1  40  0
## 78    1    1  44  1
```