Homework 5

Each part of the problems 5 points
Due in class or by email before 9am on Thursday February 24

1. Agresti 10.1 (a) and (b).
2. Agresti 10.3

3. [Methods qualifying exam, ???: use paper and pencil.] Results from a Copenhagen housing condition survey are compiled in an R data frame `housing1` consisting of the following components:

   - **Infl**: Influence of renters on management: Low, Medium, High.
   - **Type**: Type of rental property: Tower, Atrium, Apartment, Terrace.
   - **Cont**: Contact between renters: Low, High.
   - **Sat**: Highly satisfied or not: two columns of counts.

A model is fitted to the data using the following commands.

```r
fit <- glm(Sat~Infl+Type+Cont,family=binomial,data=housing1)
```

Part of the results are summarized below (`summary(fit)`).

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.6551     0.1374  -4.768  1.86e-06 ***
InflMedium  0.5362     0.1213   4.421  9.81e-06 ***
InflHigh    1.3039     0.1387   9.401  < 2e-16 ***
TypeApartment -0.5285   0.1295  -4.081   4.49e-05 ***
TypeAtrium   -0.4872   0.1728  -2.820    0.00480 **
TypeTerrace  -1.1107   0.1765  -6.294   3.10e-10 ***
ContHigh     0.3130     0.1077   2.905    0.00367 **
```

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 166.179 on 23 degrees of freedom
Residual deviance: 27.294 on 17 degrees of freedom
AIC: 146.55
```

(a) Write the assumptions of the model, and the expression of the log-likelihood.
(b) According to the fitted model, what percentage of renters, who have low influence on management, live in apartment, and have high contact between neighbors, are highly satisfied?
(c) Do people who live in apartments have a significantly different probability of satisfaction than people who live in atriums? The correlation between the respective coefficients is 0.494.

(d) Estimate the odds ratio of high satisfaction for groups with high contact among neighbors over groups with low contact among neighbors, using a 95% confidence interval.

4. [Methods qualifying exam, August 2005: use paper and pencil.] A sample of elderly people was given a psychiatric examination to determine whether symptoms of senility were present. Other measurements taken at the same time included the score on a subset of the Wechsler Adult Intelligence Scale (WAIS). The data are shown below.

<table>
<thead>
<tr>
<th>x</th>
<th>9</th>
<th>13</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>4</th>
<th>14</th>
<th>8</th>
<th>11</th>
<th>7</th>
<th>9</th>
<th>7</th>
<th>5</th>
<th>14</th>
<th>13</th>
<th>16</th>
<th>10</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

| x  | 11 | 14  | 15  | 18  | 7  | 16  | 9  | 11  | 13  | 15  | 13  | 10  | 11  | 6   | 17  | 14  | 19  |
|----|----|-----|-----|-----|----|-----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|
| s  | 0  | 0   | 0   | 0   | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   |

<table>
<thead>
<tr>
<th>x</th>
<th>9</th>
<th>11</th>
<th>14</th>
<th>10</th>
<th>16</th>
<th>10</th>
<th>16</th>
<th>14</th>
<th>13</th>
<th>13</th>
<th>9</th>
<th>15</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>4</th>
<th>14</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

A linear logistic regression model is fitted to the data,

\[
\log \frac{p}{1-p} = \alpha + \beta x
\]

where \( p = P(y = 1) \), with \( \hat{\alpha} = 2.404, \hat{\beta} = -0.3235 \), and a deviance of 51.017.

(a) For a person with WAIS score \( x = 10 \), what is the estimated probability that the person has symptoms of senility.

(b) The standard errors of \( \hat{\alpha}, \hat{\beta} \) are given by \( s\{\hat{\alpha}\} = 1.1918, s\{\hat{\beta}\} = 0.1140 \), and the correlation between \( \hat{\alpha}, \hat{\beta} \) is estimated to be -0.96. Obtain an approximate 95% confidence interval for the probability of senility of a person with a WAIS score \( x = 10 \).

(c) Assuming \( \beta = 0 \), fit the constant model by estimating \( \alpha \), and obtain the deviance of the fit.

(d) Test the hypothesis that \( \beta = 0 \) using the likelihood ratio test.

5. [Methods qualifying exam, August 2010: use paper and pencil.] When modeling a binary response in logistic regression, input data can have two forms: (1) individual observations, where the response records the status of “failure” or “success”, and (2) grouped data, where the response records the number of experimental units with the same covariate pattern, and the corresponding number of “successes”.

For the same data set, would a logistic regression analysis of each type of input data yield the same parameter estimates? Would they yield the same deviance and test of goodness-of-fit? Why or why not?

6. Data analysis: Adapted from Faraway, p. 52, Problem 2. The dataset \texttt{wbca} from the library \texttt{Faraway} comes from a study of breast cancer in Wisconsin. There are 681 cases of potentially cancerous tumors, of which 238 are actually malignant. Malignant
tumors are traditionally determined using a surgically invasive procedure. Our goal is
to determine whether a new procedure called the needle aspiration, which draws only
a small sample of tissue, could be effective at determining tumor status.

(a) Split the data into two parts: assign every third observation to a test set, and
the remaining two thirds to the training set. Parts (a) - (i) below will only use
the training set. Use the training set to fit a binomial regression with \texttt{class} as
response, and the other nine variables as predictors.

(b) For both null and residual deviance tests, specify the null and the alternative hy-
potheses, and report the conclusions if possible. (\textit{Hint}: can we use the residual
deviance test in this case?). Use the Hosmer-Lemeshow test, and compare the
results.

(c) Use the full model to check for outliers, and for influential observations. Report
the results.

(d) Use stepwise variable selection combined with the AIC criterion to determine the
best subset of predictors. (\textit{Hint}: use \texttt{step(fullModelFit, direction="both")}).

(e) Provide the definition of overdispersion. Check for evidence of over- or under-
dispersion in this dataset. Compare the model selected in (d) to the full model,
while accounting for over- or under-dispersion (\textit{Hint}: use approximate F test).

(f) Based on the logistic regression fit in (d) without overdispersion, plot the predicted
probability of receiving a flu shot as a function of \texttt{Thick}, while setting the values
of the remaining predictors at the median value observed in the dataset. Overlay
the corresponding confidence interval and interpret the results.

(g) Overlay the confidence interval obtained in presence of overdispersion, and compare
it to the confidence interval in (f).

(h) Fit the probit regression model using the same predictors as in parts (f)-(g). On
the plot, overlay the predicted probabilities based on the probit regression (you do
not need to plot confidence intervals this time). Discuss the differences between
the two sets of curves obtained with the two link functions (if any).

(i) Use the model in (d) to predict the tumor status of the patients in the training set.
Report confusion matrices for probability cutoffs $p = 0.5$ and $p = 0.9$. Discuss the
predictive ability of the model, and the role of the cutoff.

(j) Plot two ROC curves: the ROC curve based on the training set, and the ROC
curve based on the validation set. Report the areas under the curves. Discuss
the difference between the in-sample results and the results on the validation set.
Discuss the ability of the needle aspiration method to determine tumor status.