Statistics 149 – Midterm Solutions **Tuesday, March 21, 2006**

1. (10 points) The data below represent a comparison of two media for culturing Mycobacterium tuberculosis. Fifty suspect sputum specimens were plated up on **both media** and the following results were obtained:

	Medium B	
Medium A	Growth	No Growth
Growth	20	12
No Growth	2	16

   (a) (6 points) Let π_A and π_B be the probabilities of growth on media A and B. Estimate $\pi_A - \pi_B$ and give a 95% confidence interval for this quantity.
   Since this is paired data,

$$\hat{\pi}_A - \hat{\pi}_B = \frac{32}{50} - \frac{22}{50} = \frac{10}{50} = 0.2$$

$$SE(\hat{\pi}_A - \hat{\pi}_B) = \frac{\sqrt{12+2}}{50} = 0.0748$$

$$CI = 0.2 \pm 1.96 \times 0.0748 = 0.2 \pm 0.147 = (0.053, 0.347)$$

   (b) (4 points) Construct a hypothesis test examining whether the probability of growth is the same for the two media at the 5% level.

$$X^2 = \frac{(12-2)^2}{12+2}\frac{100}{14} = 7.14$$

$$p\text{-value} = 2P[Z \geq \sqrt{7.14}] = 2P[Z \geq 2.67] = 0.0075$$

   Since the p-value < 0.05 (or equivalently $X^2 > 3.84$), there is a statistically significant difference in the growth probabilities.

2. (40 points) A survey conducted in 1974 and 1975 by the National Opinion Research Center at the University of Chicago investigated the relationship of education and gender to attitudes towards the role of women in society. Each respondent was asked if they agreed or disagreed with the statement "Women should take care of running their homes and leave running the country up to men." Of the 1566 women in the study, 555 agreed with the statement, while 465 of the 1305 men also agreed.

(a) (5 points) Calculate a 95% confidence interval for the difference of the proportions of men and women who agree with the statement.

$$\hat{\pi}_m - \hat{\pi}_w = \frac{465}{1305} - \frac{555}{1566} = 0.0019$$

$$SE(\hat{\pi}_m - \hat{\pi}_w) = \sqrt{\frac{0.356 \times 0.644}{1305} + \frac{0.354 \times 0.646}{1566}} = 0.0179$$

$$CI = 0.0019 \pm 1.96 \times 0.0179 = 0.0019 \pm 0.0351 = (-0.0332, 0.0370)$$

(b) (5 points) Construct a hypothesis test to investigate whether the odds that a man agrees with the statement is different than the odds that a women agrees with the statement.

$$\log \hat{\phi} = \log \frac{465}{840} - \log \frac{555}{1011} = 0.0084$$

$$\hat{\pi}_c = \frac{465 + 555}{1305 + 1566} = 0.355$$

$$SE(\log \hat{\phi}) = \sqrt{\frac{1}{1305 \times 0.355 \times 0.645} + \frac{1}{1566 \times 0.355 \times 0.645}} = 0.0783$$

$$z = \frac{0.0084}{0.0783} = 0.107$$

p-value = 0.9.

Valid answers for this question would also be the z test looking at $\pi_m = \pi_w$ or the Chi-square test on the 2×2 table, or a logistic regression as they all have equivalent null and alternative hypotheses.

(c) (4 points) To examine the effect of education on opinions about the statement, a logistic regression was run with the model

$$\text{logit}(\pi_{Agreement}) = \beta_0 + \beta_1 Education + \beta_2 Gender + \beta_3 Education * Gender$$

where Education is measured in years. Edited **R** output for this analysis follows at the end of this question and is the basis for the rest of this question.

Based on this output, estimate the probability that a woman with 10 years of education would agree with the statement.

$$\hat{\pi}(10) = \frac{e^{3.00 + -0.315 \times 10}}{1 + e^{3.00 + -0.315 \times 10}} = 0.462$$

(d) (5 points) Is there any evidence that the chance of agreement with the statement for men and women are different, after the main effect of education are accounted for? Construct a hypothesis test to examine this question at the 5% level.

$$X^2 = 64.03 - 57.10 = 6.93$$

Since $X^2 > 5.99$ (df=2), the response rates are different for men and women.

(e) (4 points) Construct a hypothesis test to examine whether Education and Gender interact at the 5% level.

$$z = \frac{0.081}{0.031} = 2.62$$

p-value = 0.009 which implies education and gender interact.

(f) (5 points) Estimate the odds ratio for agreement when Education increases by one year for women. Note that in the logistic regression fit, the indicator variable for Gender has the women coded as 0 and the men coded as 1. Also give a 95% confidence interval for this odds ratio.

$$e^{\hat{\beta}_1} = e^{-0.315} = 0.729$$

$$CI(\beta_1) = -0.315 \pm 1.96 \times 0.0237 = -0.315 \pm 0.046 = (-0.361, -0.268)$$

$$CI(e^{\beta_1}) = (e^{-0.361}, e^{-0.268}) = (0.697, 0.765)$$

(g) (4 points) Estimate the odds ratio for agreement for when Education increases by one year for men?

$$e^{\hat{\beta}_1 + \hat{\beta}_3} = e^{-0.315+0.081} = e^{-0.234} = 0.791$$

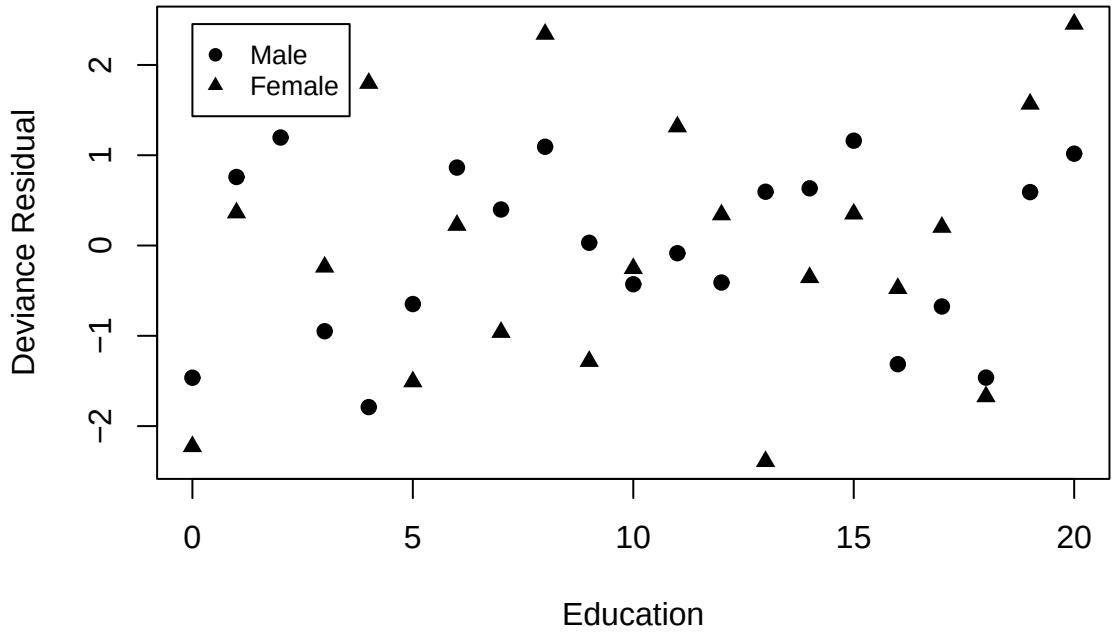
(h) (4 points) Discuss why the question asked in part (b) cannot be answered with the logistic model fit with **R**.

The analysis in part b) only looks at the main effect of gender whereas the logistic regression includes the interaction effect of gender and education. It is not possible to get at the main gender effect from the regression output given. Since there is a significant interaction effect, it also implies that the analysis in b) doesn't really get a question of interest.

(i) (4 points) The following figure shows the deviance residuals plotted against the years of education for men and women. Does this figure suggest any problems with the model? Explain briefly.

The is no suggestion that the linear effects of education for men and women fit are invalid. One problem that is suggested in the plot is with the variance assumption. All the most extreme residuals in the plot are from observations from women. This suggests

that the variance assumption $\text{Var}(\pi_i(1-\pi_i))$ being the same for men and women may not be valid.




```
> summary(opinion.glm)

Call:
glm(formula = opinion ~ Education + Gender + Education:Gender,
    family = binomial())

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.39097  -0.94911   0.03065   0.75927   2.45262

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          3.00294    0.27238  XXXXXX  XXXXXXX
Education           -0.31541    0.02365  XXXXXX  XXXXXXX
GenderMale          -0.90474    0.36007  XXXXXX  XXXXXXX
Education:GenderMale  0.08138    0.03109  XXXXXX  XXXXXXX
```

```
    Null deviance: 451.722  on 40  degrees of freedom
Residual deviance:  57.103  on 37  degrees of freedom
AIC: 203.16

> anova(opinion.glm, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: opinion

Terms added sequentially (first to last)


                 Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                40     451.72
Education         1   XXXXXX        39      64.03   XXXXXXX
Gender            1   XXXXXX        38      64.01   XXXXXXX
Education:Gender  1   XXXXXX        37      57.10   XXXXXXX
```

3. (15 points) A soft drink bottler is analyzing the vending machine service routes in his distribution system. He is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. This service includes stocking the machines with beverage products and minor maintenance or housekeeping. It has been suggested that the two most important predictor variables should be the number of cases stocked and the distance walked by the route driver at an outlet. A data set of 25 observations was collected and an additive linear model was fit in **R** (output follows). One concern was whether this was a reasonable model and whether any observations had a strong influence on the fit.

```
> summary(soda.lm)

Call:
lm(formula = Time ~ Cases + Distance, data = soda)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7880 -0.6629  0.4364  1.1566  7.4197

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.341231   1.096730   2.135 0.044170 *
Cases       1.615907   0.170735   9.464 3.25e-09 ***
Distance    0.014385   0.003613   3.981 0.000631 ***

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-Squared: 0.9596,     Adjusted R-squared: 0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

Following are 3 figures examining the data. The first shows plots of the data, the second shows residual plots based on the additive linear model, and the third shows plots of influence measures, again based on the additive linear model.

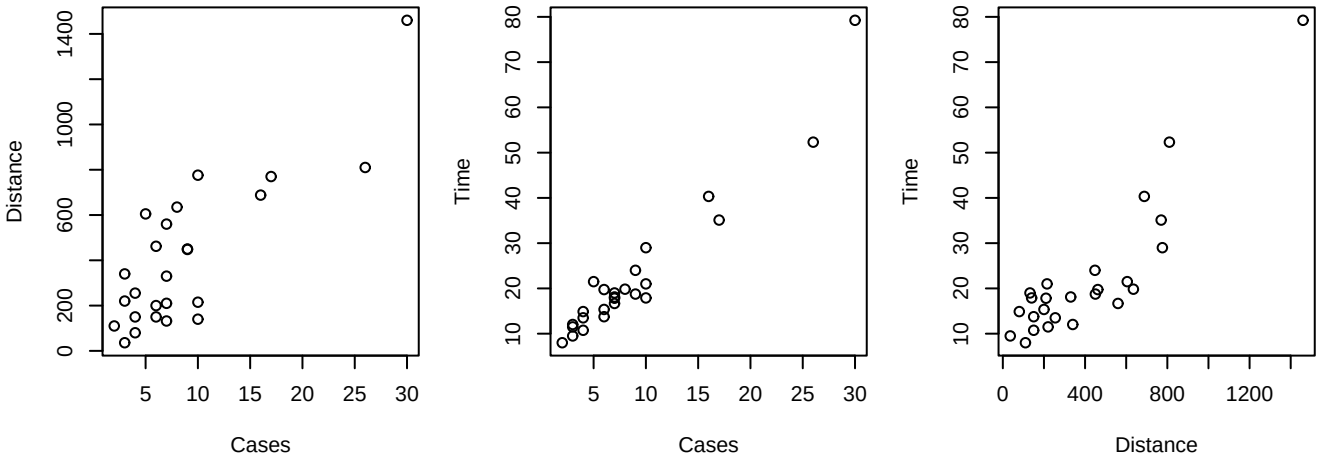


Figure 1: Plots of the data

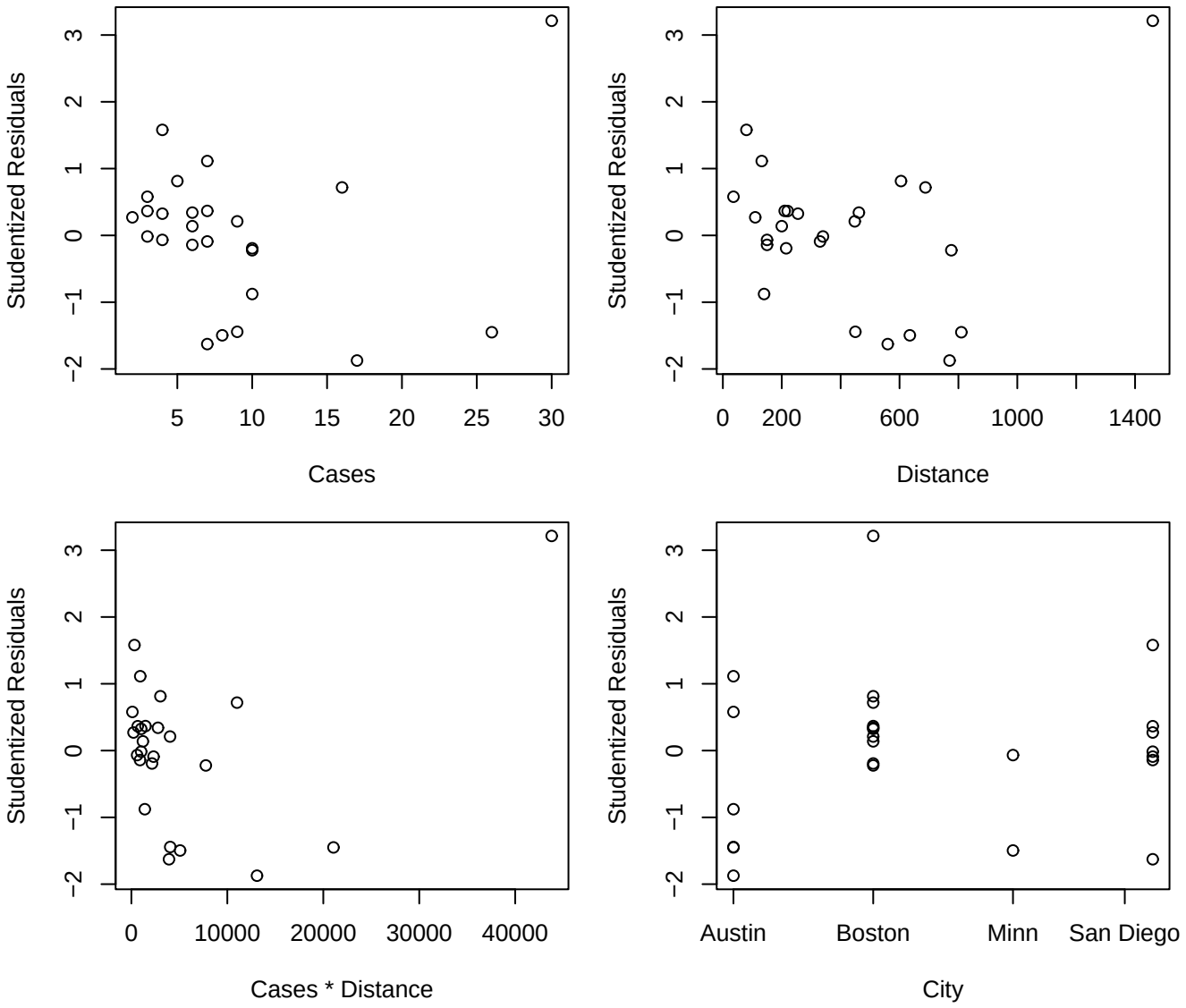


Figure 2: Plots of studentized residuals. In the lower left plot, the residuals are plotted against `Cases*Distance`, the product of the two predictor variables. In the lower right is a plot showing the residuals against the city that the data was collected in. Observations 1-7 are from San Diego, 8-17 are from Boston, 18-23 are from Austin, and 24-25 are from Minneapolis (= Minn)

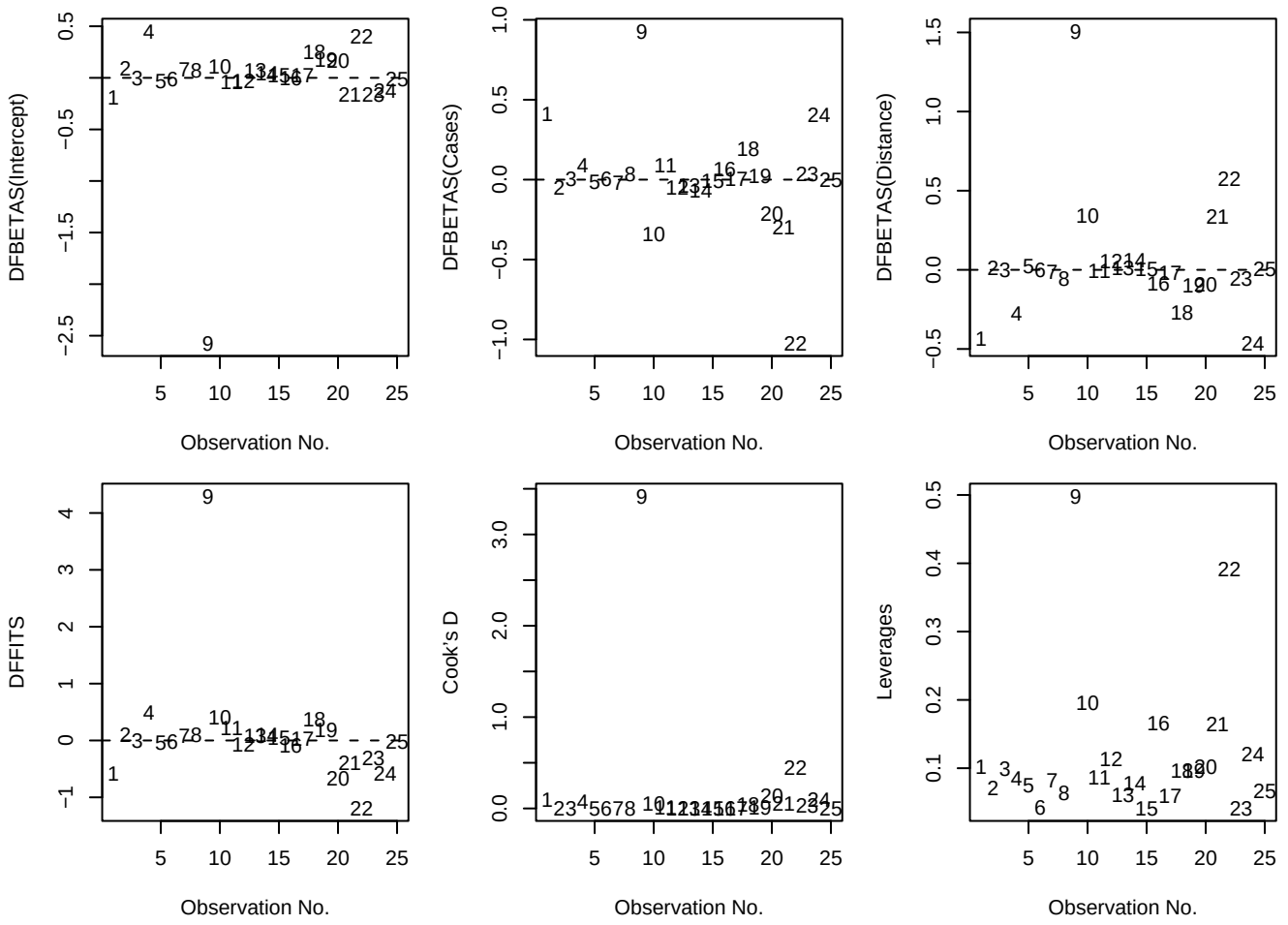


Figure 3: Plots of influence measures. The plotting symbols are the observation (row) numbers in the **R** dataframe.

(a) (5 points) Is there any evidence that the model fit to the data could be improved? If so, how?, If not, why not?

The plot of residuals against city suggests that location may have an effect on the time and possibly should be included in the model. Not other problems with the model stand out, except for the one outlier, which happens to come from Boston.

(b) (5 points) Are there any potential outliers in the data set? If so, which observations (give approximate `Cases` and `Distance` values) are they?

There is one fairly strong potential outlier in the in the dataset. It is observation 9, which is 30 cases and a distance a bit over 1400 feet.

(c) (5 points) Are there any influential points in the data set? If so, which ones (again give approximate `Cases` and `Distance`) values plus the observation number) and describe what effect they appear to have on the fit of the model?

There appear to be 2 influential points in the data set. Observation 9 mentioned earlier, which has influence on the fitted βs, its own fit (DFFITS), and the fits of others (Cook's D). The other observation is observation 22 (cases = 27, distance $\approx$ 900). It appears to have an effect on its own fit and the estimate of the effect of the number of cases.