**STOR 557: Fall 2023**
**Midterm Exam 1: September 21, 2023**

Open book exam. Course text and all personal notes allowed. Calculator is allowed but no computer (except to read e-edition of text and personal notes).

Statistical tables (6 pages) are provided — please return at end of exam.

Time allowed: One hour.

The data in Table 1 are taken from a study that measured eight variables (some continuous, some descriptive) on 118 female patients. The raw data table is shown for reference on the last page of this exam. Of the 118 individuals in the study, 31 have at least one variable missing (written "NA" in R notation) and the bulk of the analysis is based on the 87 patients for whom full data are available.

Although there are many ways we could possibly choose to analyze this dataset, we are treating it here as trying to analyze "depression" as a function of the other seven variables. Also, some of the factor variables that are given with three or four levels would be better treated (for this analysis) as binary variables. With these objectives in mind, we recode as follows:

1. The `y` variable is defined to be 0 if Dep is 1, 1 if Dep is $> 1$
2. `anx` is defined to be 0 if Anx is 1 or 2, 1 if Anx is 3 or 4
3. `sle` is defined to be 0 if Sle is 1, 1 if Sle is 2 or 3.

There are various tabulations of the data we could do, but here is a cross-tabulation of `y` and `life` within the data frame which we have called `D`:

```
> xtabs(~D$y+D$life)
   D$life
D$y  1  2
  0 26  0
  1 25 58
```

Answer each of the following questions. Lengthy explanations are not required, but you should try to summarize the conclusion from each analysis in plain English. Each of the five parts is worth 20 points.

(a) An initial analysis was fitted as follows:

```
glm1=glm(y~age+iq+anx+sle+sex+life+wt,family=binomial,D)
summary(glm1)
```

The output (edited) was as follows:

```
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
glm(formula = y ~ age + iq + anx + sle + sex + life + wt, family = binomial,
```

```
    data = D)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -66.4603 10311.2611  -0.006   0.9949
age           -0.2816     0.1681  -1.675   0.0939 .
iq             0.1583     0.1393   1.136   0.2558
anx            1.7440     1.2704   1.373   0.1698
sle           21.0171  4547.7725   0.005   0.9963
sex            0.9557     1.2787   0.747   0.4548
life          39.8886  6653.9622   0.006   0.9952
wt             0.1065     0.2954   0.361   0.7184
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 96.164  on 86  degrees of freedom
Residual deviance: 29.405  on 79  degrees of freedom
  (31 observations deleted due to missingness)
AIC: 45.405

Number of Fisher Scoring iterations: 21
```

Write a brief summary of this model and your conclusion. How do you interpret the message "fitted probabilities numerically 0 or 1 occurred"?

(b) For a second analysis, we first eliminated all the rows with missing data, refitted the model omitting `life`, and performed a variable reduction with the `step` function, as follows:

```
D1=na.omit(D)
glm2=glm(y~age+iq+anx+sle+sex+wt,family=binomial,D1)
glm3=step(glm2)
summary(glm3)
```

Part of the output appears as follows:

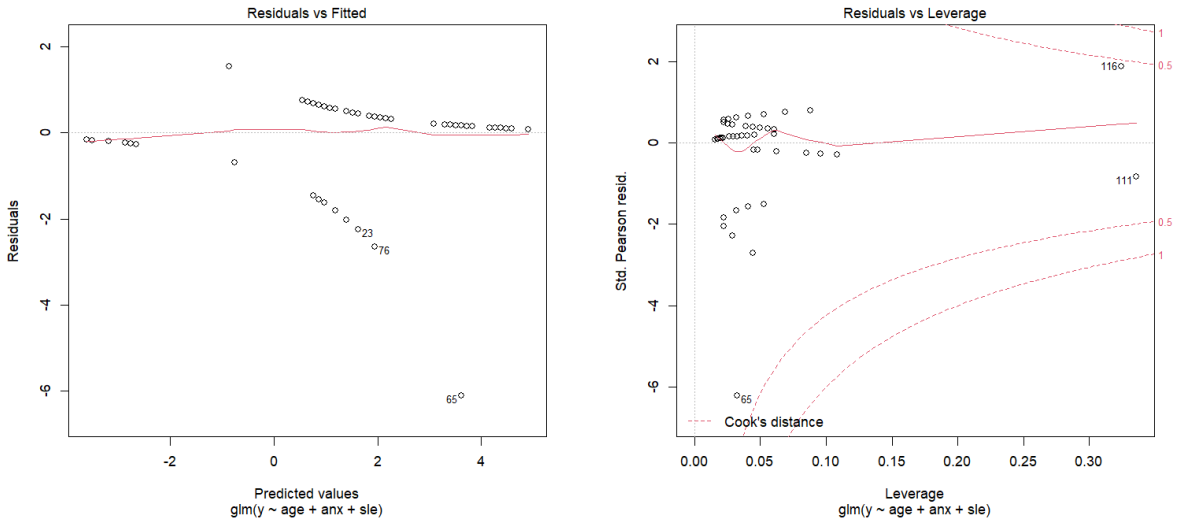
```
glm(formula = y ~ age + anx + sle, family = binomial, data = D1)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.90262    3.24187   0.278 0.780686
age         -0.10778    0.07944  -1.357 0.174837
anx          2.64489    1.16638   2.268 0.023353 *
sle          4.48864    1.29433   3.468 0.000524 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
    Null deviance: 96.164  on 86  degrees of freedom
Residual deviance: 59.701  on 83  degrees of freedom
AIC: 67.701

Number of Fisher Scoring iterations: 6
```

Write a brief explanation of this analysis. With specific reference to the rows showing `Null deviance` and `Residual deviance`, what do these rows tell you about how well the model fits the data?

(c) It looks as though the sleep variable `sle` is highly significant. What is the odds ratio for depression in an individual with `sle=1` compared to `sle=0`? State a 95% confidence interval for that odds ratio.

(d) When we do `plot(glm3)`, two of the plots look like this:




How do you interpret those plots — in particular, what do you conclude about the possibly anomalous behavior of patients 65, 111 and 116?

(e) Further inspection of the data shows that we actually have 104 patients (rather than 87) for whom the variables `y, age, anx, sle` are available. If we refit the model to this expanded dataset we get results

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.88919    2.75066  -0.323   0.7465
age         -0.05648    0.06787  -0.832   0.4053
anx          1.79123    0.86240   2.077   0.0378 *
sle          4.45095    1.12573   3.954 7.69e-05 ***
    Null deviance: 114.717  on 103  degrees of freedom
Residual deviance:  71.734  on 100  degrees of freedom
AIC: 79.734

Number of Fisher Scoring iterations: 5
```

Briefly summarize how this model differs from the earlier one based on 87 observations.

| Num | Age | IQ | Anx | Dep | Sle | Sex | Life | Wt | Num | Age | IQ | Anx | Dep | Sle | Sex | Life | Wt |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 39 | 94 | 2 | 2 | 2 | 2 | 2 | 4.9 | 60 | 41 | 89 | 2 | 1 | 2 | 1 | 1 | 3.2 |
| 2 | 41 | 89 | 2 | 2 | 2 | 2 | 2 | 2.2 | 61 | 41 | 89 | 3 | 2 | 2 | 2 | 2 | 2.1 |
| 3 | 42 | 83 | 3 | 3 | 3 | 2 | 2 | 4 | 62 | 44 | 98 | 3 | 2 | 2 | 2 | 2 | 3.8 |
| 4 | 30 | 99 | 2 | 2 | 2 | 2 | 2 | -2.6 | 63 | 35 | 98 | 2 | 2 | 2 | 2 | 1 | -2.4 |
| 5 | 35 | 94 | 2 | 1 | 1 | 2 | 1 | -0.3 | 64 | 41 | 103 | 2 | 2 | 2 | 2 | 2 | -0.8 |
| 6 | 44 | 90 | NA | 1 | 2 | 1 | 1 | 0.9 | 65 | 41 | 91 | 3 | 1 | 2 | 2 | 1 | 5.8 |
| 7 | 31 | 94 | 2 | 2 | NA | 2 | 2 | -1.5 | 66 | 42 | 91 | 4 | 3 | NA | NA | 2 | 2.5 |
| 8 | 39 | 87 | 3 | 2 | 2 | 2 | 1 | 3.5 | 67 | 33 | 94 | 2 | 2 | 2 | 2 | 1 | -1.8 |
| 9 | 35 | NA | 3 | 2 | 2 | 2 | 2 | -1.2 | 68 | 41 | 91 | 2 | 1 | 2 | 2 | 1 | 4.3 |
| 10 | 33 | 92 | 2 | 2 | 2 | 2 | 2 | 0.8 | 69 | 43 | 85 | 2 | 2 | 2 | 1 | 1 | NA |
| 11 | 38 | 92 | 2 | 1 | 1 | 1 | 1 | -1.9 | 70 | 37 | 92 | 1 | 1 | 2 | 2 | 1 | 1 |
| 12 | 31 | 94 | 2 | 2 | 2 | NA | 1 | 5.5 | 71 | 36 | 96 | 3 | 3 | 2 | 2 | 2 | 3.5 |
| 13 | 40 | 91 | 3 | 2 | 2 | 2 | 1 | 2.7 | 72 | 44 | 90 | 2 | NA | 2 | 2 | 2 | 3.3 |
| 14 | 44 | 86 | 2 | 2 | 2 | 2 | 2 | 4.4 | 73 | 42 | 87 | 2 | 2 | 2 | 1 | 2 | -0.7 |
| 15 | 43 | 90 | 3 | 2 | 2 | 2 | 2 | 3.2 | 74 | 31 | 95 | 2 | 3 | 2 | 2 | 2 | -1.6 |
| 16 | 32 | NA | 1 | 1 | 1 | 2 | 1 | -1.5 | 75 | 29 | 95 | 3 | 3 | 2 | 2 | 2 | -0.2 |
| 17 | 32 | 91 | 1 | 2 | 2 | NA | 1 | -1.9 | 76 | 32 | 87 | 1 | 1 | 2 | 2 | 1 | -3.7 |
| 18 | 43 | 82 | 4 | 3 | 2 | 2 | 2 | 8.3 | 77 | 35 | 95 | 2 | 2 | 2 | 2 | 2 | 3.8 |
| 19 | 46 | 86 | 3 | 2 | 2 | 2 | 2 | 3.6 | 78 | 42 | 88 | 1 | 1 | 1 | 2 | 1 | -1 |
| 20 | 30 | 88 | 2 | 2 | 2 | 2 | 1 | 1.4 | 79 | 32 | 94 | 2 | 2 | 2 | 2 | 1 | 4.7 |
| 21 | 34 | 97 | 3 | 3 | NA | 2 | 2 | NA | 80 | 39 | NA | 3 | 2 | 2 | 2 | 2 | -4.9 |
| 22 | 37 | 96 | 3 | 2 | 2 | 2 | 1 | NA | 81 | 34 | NA | 3 | NA | 2 | 2 | 1 | NA |
| 23 | 35 | 95 | 2 | 1 | 2 | 2 | 1 | -1 | 82 | 34 | 87 | 3 | 3 | 2 | 2 | 1 | 2.2 |
| 24 | 45 | 87 | 2 | 2 | 2 | 2 | 2 | 6.5 | 83 | 42 | 92 | 1 | 1 | 2 | 1 | 1 | 5 |
| 25 | 35 | 103 | 2 | 2 | 2 | 2 | 1 | -2.1 | 84 | 43 | 86 | 2 | 3 | 2 | 2 | 2 | 0.4 |
| 26 | 31 | NA | 2 | 2 | 2 | 2 | 1 | -0.4 | 85 | 31 | 93 | NA | 2 | 2 | 2 | 2 | -4.2 |
| 27 | 32 | 91 | 2 | 2 | 2 | 2 | 1 | -1.9 | 86 | 31 | 92 | 2 | 2 | 2 | 2 | 1 | -1.1 |
| 28 | 44 | 87 | 2 | 2 | 2 | 2 | 2 | 3.7 | 87 | 36 | 106 | 2 | 2 | 2 | 1 | 2 | -1 |
| 29 | 40 | 91 | 3 | 3 | 2 | 2 | 2 | 4.5 | 88 | 37 | 93 | 2 | 2 | 2 | 2 | 2 | 4.2 |
| 30 | 42 | 89 | 3 | 3 | 2 | 2 | 2 | 4.2 | 89 | 43 | 95 | 2 | 2 | 2 | 2 | 1 | 2.4 |
| 31 | 36 | 92 | 3 | NA | 2 | 2 | 2 | NA | 90 | 32 | 95 | 3 | 2 | 2 | 2 | 2 | 4.9 |
| 32 | 42 | 84 | 3 | 3 | 2 | 2 | 2 | 1.7 | 91 | 32 | 92 | NA | NA | NA | 2 | 2 | 3 |
| 33 | 46 | 94 | 2 | NA | 2 | 2 | 2 | 4.8 | 92 | 32 | 98 | 2 | 2 | 2 | 2 | 2 | -0.3 |
| 34 | 41 | 92 | 2 | 1 | 2 | 2 | 1 | 1.7 | 93 | 43 | 92 | 2 | 2 | 2 | 2 | 2 | 1.2 |
| 35 | 30 | 96 | NA | 2 | 2 | 2 | 2 | -3 | 94 | 41 | 88 | 2 | 2 | 2 | 2 | 1 | 2.6 |
| 36 | 39 | 96 | 2 | 2 | 2 | 1 | 1 | 0.8 | 95 | 43 | 85 | 1 | 1 | 2 | 2 | 1 | 1.9 |
| 37 | 40 | 86 | 2 | 3 | 2 | 2 | 2 | 1.5 | 96 | 39 | 92 | 2 | 2 | 2 | 2 | 1 | 3.5 |
| 38 | 42 | 92 | 3 | 2 | 2 | 2 | 1 | 1.3 | 97 | 41 | 84 | 2 | 2 | 2 | 2 | 2 | -0.6 |
| 39 | 35 | 102 | 2 | 2 | 2 | 2 | 2 | 3 | 98 | 41 | 92 | 2 | 1 | 2 | 2 | 1 | 1.4 |
| 40 | 31 | 82 | 2 | 2 | 2 | 2 | 1 | 1 | 99 | 32 | 91 | 2 | 2 | 2 | 2 | 2 | 5.7 |
| 41 | 33 | 92 | 3 | 3 | 2 | 2 | 2 | 1.5 | 100 | 44 | 86 | 3 | 2 | 2 | 2 | 2 | 4.6 |
| 42 | 43 | 90 | NA | NA | 2 | 2 | 2 | 3.4 | 101 | 42 | 92 | 3 | 2 | 2 | 2 | 1 | NA |
| 43 | 37 | 92 | 2 | 1 | 1 | 1 | 1 | NA | 102 | 39 | 89 | 2 | 2 | 2 | 2 | 1 | 2 |
| 44 | 32 | 88 | 4 | 2 | 2 | 2 | 1 | NA | 103 | 45 | NA | 2 | 2 | 2 | 2 | 2 | 0.6 |
| 45 | 34 | 98 | 2 | 2 | 2 | 2 | NA | 0.6 | 104 | 39 | 96 | 3 | NA | 2 | 2 | 2 | NA |
| 46 | 34 | 93 | 3 | 2 | 2 | 2 | 2 | 0.6 | 105 | 31 | 97 | 2 | NA | NA | NA | 2 | 2.8 |
| 47 | 42 | 90 | 2 | 1 | 1 | 2 | 1 | 3.3 | 106 | 34 | 92 | 3 | 2 | 2 | 2 | 2 | -2.1 |
| 48 | 41 | 91 | 2 | 1 | 1 | 1 | 1 | 4.8 | 107 | 41 | 92 | 2 | 2 | 2 | 2 | 2 | -2.5 |
| 49 | 31 | NA | 3 | 1 | 2 | 2 | 1 | -2.2 | 108 | 33 | 98 | 3 | 2 | 2 | 2 | 2 | 2.5 |
| 50 | 32 | 92 | 3 | 2 | 2 | 2 | 2 | 1 | 109 | 34 | 91 | 2 | 1 | 1 | 2 | 1 | 5.7 |
| 51 | 29 | 92 | 2 | 2 | 2 | 1 | 2 | -1.2 | 110 | 42 | 91 | 3 | 3 | 2 | 2 | 2 | 2.4 |
| 52 | 41 | 91 | 2 | 2 | 2 | 2 | 2 | 4 | 111 | 40 | 89 | 3 | 1 | 1 | 1 | 1 | 1.5 |
| 53 | 39 | 91 | 2 | 2 | 2 | 2 | 2 | 5.9 | 112 | 35 | 94 | 3 | 3 | 2 | 2 | 2 | 1.7 |
| 54 | 41 | 86 | 2 | 1 | 1 | 2 | 1 | 0.2 | 113 | 41 | 90 | 3 | 2 | 2 | 2 | 2 | 2.5 |
| 55 | 34 | 95 | 2 | 1 | 1 | 2 | 1 | 3.5 | 114 | 32 | 96 | 2 | 1 | 1 | 2 | 1 | NA |
| 56 | 39 | 91 | 1 | 1 | 2 | 1 | 1 | 2.9 | 115 | 39 | 87 | 2 | 2 | 2 | 1 | 2 | NA |
| 57 | 35 | 96 | 3 | 2 | 2 | 1 | 1 | -0.6 | 116 | 41 | 86 | 3 | 2 | 1 | 1 | 2 | -1 |
| 58 | 31 | 100 | 2 | 2 | 2 | 2 | 2 | -0.6 | 117 | 33 | 89 | 1 | 1 | 1 | 1 | 1 | 6.5 |
| 59 | 32 | 99 | 4 | 3 | 2 | 2 | 2 | -2.5 | 118 | 42 | NA | 3 | 2 | 2 | 2 | 2 | 4.9 |

Table 1: Data for Depression Study. Variables are: Age; IQ; anxiety (Anx, four-point scale, 1=none, 2=mild, 3=moderate, 4=severe); depression (Dep, four-point scale, same as anxiety); quality of sleep (Sle, 1=good, 2=bad); interest in sex (Sex, 1=yes, 2=no); whether the patient has ever contemplated taking her own life (Life, 1=no, 2=yes); weight change over the last 6 months (Wt, in pounds, negative values mean a decrease). Data from a paper by B.S. Everitt, Institute of Psychiatry, London.

## SKETCH SOLUTIONS

The following are not intended as full and complete answers, but as outlines of the main points. I am expecting that your own answers will differ quite a bit from these, but I am still expecting you to identify the main points, however you choose to word them.

(a) Based on this analysis, none of the variables seems statistically significant but we note especially that `sle` and `life` have enormous standard errors and very low z-values, suggesting that there is some problem estimating these parameters. The cross-tabulation table of `y` and `life` shows a 0 in the cell for `y=0,life=0` — in plain English, every patient who had considered taking her own life was depressed. Therefore, it is natural to expect that a statistical model will predict that any individual with `life=2` will have an estimated probability of depression near 1. This is the most natural interpretation of the warning message. The solution would either be to try a bias reduction analysis (as in one of the homeworks) or just to use a different model. [There is a similar issue with the "sleep" variable, but I didn't give you the cross-tabulation for that. However, the later analysis shows that this is not such a problem once "life" is dropped from the analysis.]

(b) In this analysis, the variables `anx` and `sle` are statistically significant but `age` is not, despite being included in the stepwise analysis. It would be reasonable to drop `age` from the equation. Regarding the deviances, the difference between the null and residual deviances is 96.164-59.701=36.463 with 86-83=3 DF and this is highly significant (from the tables, the 0.001 right tail point of a $\chi^2_3$ is 16.27, well under 36.463). The interpretation is that the three variables together (`age, anx, sle`) are highly significant against the null model in which there is just an intercept and no covariates. You could also refer to the residual deviance itself: 59.701 is less than the DF (83) which suggests that this model does fit the data.

(c) If $p$ is the probability of depression, the coefficient of `sle` in the model for $\log \frac{p}{1-p}$ is 4.45095 with a standard error of 1.12573. Based on the normal distribution, a 95% confidence interval for the coefficient is $4.45095 \pm 1.96 \times 1.12573 = (2.244519, 6.657381)$. This coefficient represents the logarithm of the odds ratio. For the odds ratio itself, we take exponentials: the estimated odds ratio is 85.7 with a confidence interval from 9.4 to 778.5. (Exact numbers are not required; any reasonable roundings will be accepted.)

Comment on this: it seems clear that there is a relationship between poor sleep and depression, but quantifying the relationship in terms of an odds ratio is hard, based on this data.

(d) 65 is an outlier; it's clearly at the bottom end of the residuals v. fitted values plot and also shows up as the most extreme residual in the residuals v. leverage plot. However, it's not a point of high leverage. Observations 111 and 116 are high leverage and 116 may be an outlier as well (it has the largest residual amongst any positive value). As for influence, the only value with a large Cook's D statistic is number 116 which lies just *inside* the $D = 0.5$ contour: it does not meet the technical definition of an influential value, but it's close.

As for explanations, patient 65 had the unusual combination of Anx=3, Sle=2 (both of which would be predictive of depression) but Dep=1 (no depression). So this was an unusual response. Patients 111 and 116 both had Anx=3, Sle=1 and (this is harder to spot) they are in fact the only two patients with that combination (and in fact, one of them had depression

and one of them didn't). So these were individuals who had an unusual combination of predictor variables, which is what leverage is measuring. None of the three had an unusual value for age (they were all either 40 or 41) and the other variables are irrelevant because they were not used in the analysis.

There is no real reason to consider dropping these three individuals from the analysis, but it might be of interest to refit the model without them and see how much it changes.

(e) All four coefficients (including the intercept) change in the larger dataset but the changes are all less than one standard error, which implies that the difference are not statistically significant. The deviances are larger, as might be expected given the larger number of observations, but the difference between the null and residual deviances is 114.717-71.734=42.983, definitely significant by the same chi-squared analysis as earlier. The other thing you might point out is that all four standard errors are smaller with this dataset than in the earlier analysis, reflecting the obvious fact that if you use more data, you expect to get more accurate estimates.

Summary: it seems like a good idea to use the larger dataset, but there is no evidence that the qualitative conclusions from the study are changed as a result.