STOR 590: Spring 2020
Take-home Midterm Exam

Answer all questions.

This is a take-home exam that you are expected to do in your own time and hand in no later than **6pm Friday February 28**. The exam should be submitted via the "Assignments" tab of the course sakai page.

Rules of the Exam. All course resources including text, personal notes and resources available through R or R-Studio are permitted. Your submitted answers should include full verbal answers to the questions, illustrated where appropriate by R code, tables or figures. Very long-winded answers are discouraged; greatest credit will be given for full but concise answers to the questions. Solutions may be submitted in R-Markdown but this is not required. (A fully acceptable alternative is if you submit a Word document into which you cut and paste R output as appropriate; however, I recommend you "save as" a pdf file for the final submission.) Other web resources may be used if fully acknowledged and referenced. Discussion among yourselves or with an outside party is not permitted; you are allowed to email the instructor if you find the question ambiguous or if you think there is an error, but the instructor will not give advice how to solve the problems.

Please acknowledge you accept these conditions by copying out and signing:

PLEDGE: I will neither give nor receive unauthorized aid in this exam.

SIGNED: (A typed signature will be accepted)

1. A number of patients were given a new drug which unfortunately has some undesirable after-effects. Table 1 shows how many patients were given the drug at each of seven doses, and how many of them developed after-effects.

Dose	0.9	1.1	1.8	2.3	3.0	3.3	4.0
Number of patients	46	72	118	96	84	53	38
Number with after-effects	17	22	52	58	56	43	30

Table 1: Data for Question 1

[The data are available in a file Drugs.csv, in the Data folder under resources on sakai. To load it from the file, first copy to a directory on our own machine, then use some command like `Drugs=read.csv('Drugs.csv')`. You may need to insert a directory path in front of the file name.]

   (a) Draw a plot that shows how the probability of after-effects is related to dose. Describe the shape of the plot. [**4 points.**]

   (b) Using the glm command, construct a model for the probability of getting after-effects as a function of dose. Show the parameter estimates, standard errors and deviance. [**6 points.**]

   (c) Does the model fit the data? Use the standard diagnostics such as deviance and Pearson statistics, residual plots, leverage and influence. Summarize your conclusions. [**6 points.**]

(d) Do alternative forms of model fit the data better? Consider, in particular, standard alternatives such as including a quadratic term in the Drug variable, and overdispersion. Summarize your conclusions. [**6 points.**]

(e) Using the model from part (b), calculate the probability of after-effects for a dose of 2.6, with a 95% confidence limit. [**6 points.**]

(f) At what level of dose would you estimate the probability of after-effects to be 0.5? [**5 points.**]

Bonus question: Calculate a 95% confidence interval for your answer to (f).

2. Backache is a common complaint in pregnant women. To investigate this, researchers conducted a survey among mothers who had just given birth. The data file Backache.csv records the results of the survey. (You can read this in R by first downloading the data to your computer, and then a command of the structure `Back=read.csv('.../Backache.csv')`, as for the previous question.) Just for clarification, "Severity" denotes the severity of the perceived backache on a scale of 0,1,2,3, the variables from "Age" through "PrevBackache" are explanatory variables that related to the mother's background, the variables from "Tablets" through "Walking" are factors that are perceived to relieve backache, while variables from "Fatigue" through "Walking2" are factors perceived to aggravate backache. All the variables that essentially have yes-no responses are recorded 0 for a negative response and 1 for a positive response. Thus, for example, the variable "Walking" is 1 if walking is perceived to relieve the pain and "Walking2" is 1 if walking is perceived to aggravate the pain.

(a) Create two new variables as follows: (a) instead of the "Severity" variable as given, create a binary response y which is 0 if Severity=0 or 1, or $y = 1$ if Severity=2 or 3. (This is more convenient for logistic regression.) (b) Instead of treating "Weightstart" and "Weightend" (meaning mother's weight at the beginning and end of the pregnancy) as two separate variables, create a variable "Weightgain" that represents weight gain during the pregnancy, and then use "Weightstart" and "Weightgain" (but not "Weightend") as covariates in the subsequent analysis. [**2 points.**]

(b) Do you think there are potential outliers or even erroneous observations in the data? Construct plots or tables (as appropriate) to illustrate your answer, and make adjustments to the data if required. [**5 points.**]

(c) Use appropriate plots to examine the relationship between backache and (i) age, (ii) weight gain, (iii) number of previous children. Based on visual analyses alone, say which one(s) you think are important. [**5 points.**]

(d) Now conduct a formal logistic regression analysis, using the variables from "Age" through "PrevBackache" as predictors. You should decide which variables are significant using any standard variable selection technique. Summarize your conclusions. [**6 points.**]

(e) For the analysis you did in (d), construct suitable diagnostic plots to judge how well the model fits the data. Also, construct the Hosmer-Lemeshow test to judge overall fit, and summarize your conclusions. [**6 points.**]

(f) Of the eight variables from "Tablets" to "Walking" that are identified as relieving backache, which ones seem to have an effect? Use any appropriate statistical procedure. [**3 points.**]

(g) Of the fifteen variables from “Fatigue” to “Walking2” that are identified as aggravating backache, which ones seem to have an effect? Use any appropriate statistical procedure. **[3 points.]**

(h) Write a short plain-English summary of your conclusions, that would be understandable to a doctor who has never taken a statistics course. **[4 points.]**

3. (See text, question 8.2.) An experiment was conducted as part of an investigation of the effect of certain toxic agents. The survival time of rats depended on the type of poison and the treatment. The data are in `rats` in the Faraway package.

(a) Make plots of the data and comment on the difference between treatments and poisons. **[4 points.]**

(b) Fit a standard (normal-theory)linear model with an interaction between the two predictors. Use a Box-Cox transformation to determine an optimal transformation of the response. Can this optimal transformation be rounded to a more interpretable response? **[6 points.]**

(c) Refit the model using your chosen transformation. Draw suitable plots to determine whether the model is a good fit. **[6 points.]**

(d) Is the interaction statistically significant? Simplify the model if justified and determine which combination of poison and treatment will result in the shortest survival time. **[5 points.]**

(e) Build an inverse Guassian GLM for this data. Based on your previous modeling, select an appropriate link function. Based on this model, which combination of poison and treatment will result in the shortest survival time? **[7 points.]**

(f) Compare the predicted values on the original scale of the response for the two models. Do the two models produce a similar fit? **[5 points.]**

Solutions

1. (a) The scatterplot (Figure 1) shows that the proportion of patients with after-effects appears to increase linearly with the dose.

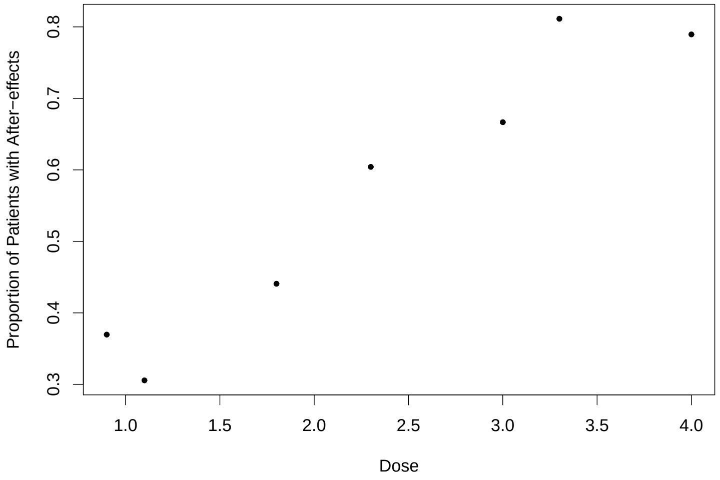

Figure 1: Figure for Question 1

(b) See the following (edited) R output.

```
> g1=glm(cbind(Aftereffects,Patients-Aftereffects)˜Dose,family=binomial,Drugs)
> summary(g1)

Call:
glm(formula = cbind(Aftereffects, Patients - Aftereffects) ˜
    Dose, family = binomial, data = Drugs)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5207     0.2566  -5.927 3.09e-09 ***
Dose          0.7806     0.1107   7.055 1.73e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 60.6850  on 6  degrees of freedom
Residual deviance:  4.3847  on 5  degrees of freedom
```

The estimated parameters are -1.52 and 0.78 with standard errors 0.26 and 0.11. The deviance is 4.3847.

(c) The commands

```
sum(residuals(g1,type=’deviance’)^2)
sum(residuals(g1,type=’pearson’)^2)
```

show that the deviance and Pearson statistics are respectively 4.3847 (as above) and 4.3436. Neither is statistically significant based on the χ^2_5 distribution.

(d)
```
> g2=glm(cbind(Aftereffects,Patients-Aftereffects)~Dose+I(Dose^2),
family=binomial,Drugs)
> summary(g2)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.58122    0.59402  -2.662  0.00777 **
Dose         0.84263    0.55989   1.505  0.13233
I(Dose^2)   -0.01362    0.12042  -0.113  0.90995
...
> anova(g1,g2,test=’Chi’)
Analysis of Deviance Table
..
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         5     4.3847
2         4     4.3719  1 0.012767     0.91
```

There is no evidence of overdispersion based on the deviance and Pearson statistics (both smaller that the degrees of freedom, so any estimated dispersion parameter would be < 1). In the model g2, the added quadratic term is not significant, and an ANOVA test between the two models shows a p-value of 0.91, equivalent to accepting the null hypothesis of the linear model.

(e) You can run a `predict.glm` command as follows:

```
> Dose=2.6
> Patients=100
> Aftereffects=50
> Drugs1=data.frame(Dose,Patients,Aftereffects)
> p1=predict.glm(g1,newdata=Drugs1,se.fit=T)
```

(the values for Patients and Aftereffects are just to define the dataframe and do not play a role in the prediction). This results in a prediction of 0.5088256 with a standard error of 0.1064261, so a 95% confidence intervals would be (0.300,0.717). This is on a logit scale: applying Faraway’s `ilogit` function, on a probability scale the predicted value is 0.625 and a 95% confidence interval is (0.574,0.672). (Minor variations in the answer because of rounding etc. will earn full credit.)

(f) EITHER (i) write the model in the form $\eta_i = \beta_0 + \beta_1 d_i$ where d_i is dose and solve $\hat{\beta}_0 + \hat{\beta}_1 d_i = 0$, which leads to $\hat{d}_i = 1.5207/0.7806 = 1.948$, OR (ii) repeat the above `predict.glm` command and show by trial and error that setting $d_i = 1.95$ (to two decimal places) leads to predicted probability 0.5.

For the bonus question: repeat the predict.glm command on a large set of possible doses. Find the dose for which the upper bound on the confidence interval for probability of after-effect is 0.5. Repeat using the lower bound. The range of doses is (1.68,2.19).

2. (a)
```
Back$y=Back$Severity
Back$y[Back$Severity<2]=0
Back$y[Back$Severity>=2]=1
Back$Weightgain=Back$Weightend-Back$Weightstart
```
or something equivalent.

(b) No specific “right answer” here but there are a number of strange things you could point out. The histograms of height and starting weight (see Figure 2) both have an normal shape but there are some strange gaps in the list of heights (possibly the data were recorded in inches and then converted to centimeters for the analysis). There are two women with a weight gain of zero, which is suspicious, that they failed to understand the question or wrote down the wrong numbers, but it doesn’t prove there was an error. Two of the baby weights were also apparent outliers (5.97 and 6.28, both presumably in kilograms) and there’s one big outlier in weight gain (nearly 40 kg.). However, such numbers are possible,so I didn’t make any adjustment to the data. As general principle, you shouldn’t delete outliers unless there’s a genuine reason to think they are upsetting the analysis.

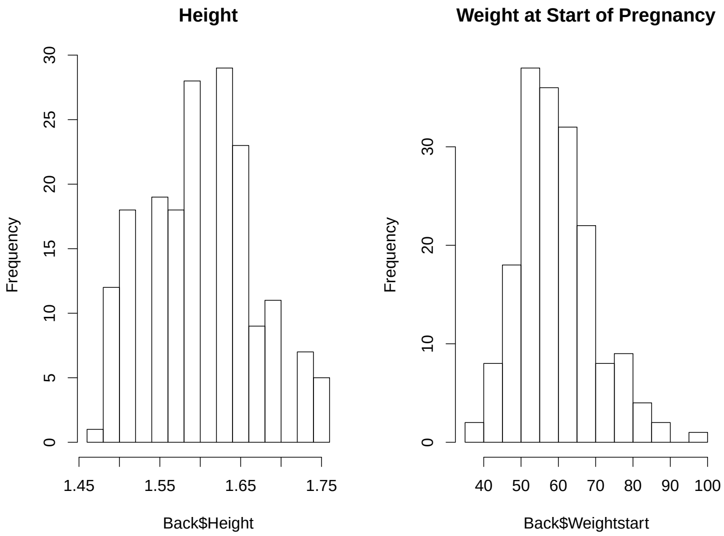

Figure 2: Histograms for Height and Starting Weight

(c) See Figure 3. Some evidence that increasing age, greater weight gain and larger number of previous children are all indicative of back pain, but none of them are clear-cut based on these figures.

(d) An initial call to glm with the stated variables produced the following:

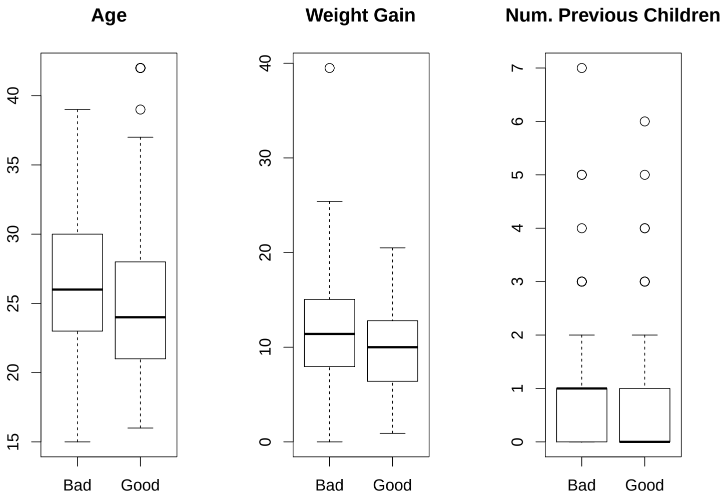

Figure 3: Factors influencing back pain

```
glm(formula = y ~ Age + Height + Weightstart + Weightgain + Weightbaby +
    PrevKids + PrevBackache, family = "binomial", data = Back)
...
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.0760572  4.1534418   1.463   0.1435
Age          -0.0129749  0.0362921  -0.358   0.7207
Height       -6.4468910  2.8618026  -2.253   0.0243 *
Weightstart   0.0409242  0.0183875   2.226   0.0260 *
Weightgain    0.0854778  0.0350612   2.438   0.0148 *
Weightbaby   -0.0185621  0.2978181  -0.062   0.9503
PrevKids      0.0002787  0.1795266   0.002   0.9988
PrevBackache  0.8203859  0.2048231   4.005 6.19e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
    Null deviance: 249.33  on 179  degrees of freedom
Residual deviance: 216.07  on 172  degrees of freedom
AIC: 232.07
```

Backward selection (e.g. the `step` command) reduces it to the following:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.93391    4.13045   1.437  0.15082
Height       -6.56023    2.84956  -2.302  0.02132 *
Weightstart   0.04053    0.01822   2.225  0.02611 *
Weightgain    0.08482    0.03256   2.605  0.00919 **
```

```
PrevBackache  0.79393    0.17318   4.584 4.55e-06 ***
...
Residual deviance: 216.23  on 175  degrees of freedom
AIC: 226.23
```

The conclusion is that height, weight at the start of the pregnancy and weight gain during pregnancy are all significant factors, but previous back pain seems to be the most significant predictor. Of the factors in Figure 3, neither age nor number of previous children seems to have an effect.

We can also try the quasibinomial model to look for overdispersion, but this is not significant (the estimated overdispersion coefficient is only 1.08). In fact, there seems to be a bit of a paradox here, because the Pearson X^2 statistic (187.9021) is much smaller than the deviance (216.2306).

(e) See the following two figures. The standard diagnostic plots (Figure 4) do not show much of interest — as with all logistic regression models, there are clearly separated groups of points according to the binary response. The grouped residual plot (Figure 5, here based on 20 bins) does not show anything of concern. The Hosmer-Lemeshow statistic computed from Figure 5 shows $HL = 23.788$ with $DF = 20$ (not significant); an alternative calculation using the "generalhoslem" package gave X^2=17.35, df=18, p-value=0.4992 (you may get different answers based on exactly how you did the test, but answers within this general ballpark will be accepted).

(f) Various answers are possible here — you could choopse to emphasize any of (i) which of the responses is most common, regardless whether the responded reports backache, (ii) which of the responses is more common among responders who do report backache, (iii) which responses give the most statistically significant result when tested for independence against incidence of backache in a 2×2 table. Any of these leads approximately to the following ordering: Cushion (57 responses total, 42 among those reporting backache, p-value 5×10^{-6}), Lying down, sitting, tablets.

(g) Similar reasoning to (f): most prevalent responses are Standing, Lifting, Ironing and Fatigue.

(h) Previous back pain in pregnancy is the strongest indicator of the likelihood of back pain in current pregnancy. Other factors are height (negative association), weight at the beginning of pregnancy (positive association) and weight gain during pregnancy (positive association). The most effective means of relieving pain is to use a cushion. Aggravating factors include [some or all of] standing, lifting, ironing, fatigue.

3. (a) See Figure 6. Increasing the dose of poison obviously results in shorter survival time. It also looks as though treatments A and C are more effective (i.e. kill faster) than treatments B and D.

(b) Use code such as `l1=lm(time poison*treat,rats)` followed by `l2=boxcox(l1)` (load library `MASS` first). The Box-Cox plot (Figure 7) suggests an optimal λ of about –0.9 but $\lambda = -1$ is within the confidence region and more interpretable (rate of killing).

(c) The cleanest way to do this is to define a new variable `rats$tinv=1/rats$time` and then run `l2=lm(tinv poison*treat,rats)` for the model with interactions and

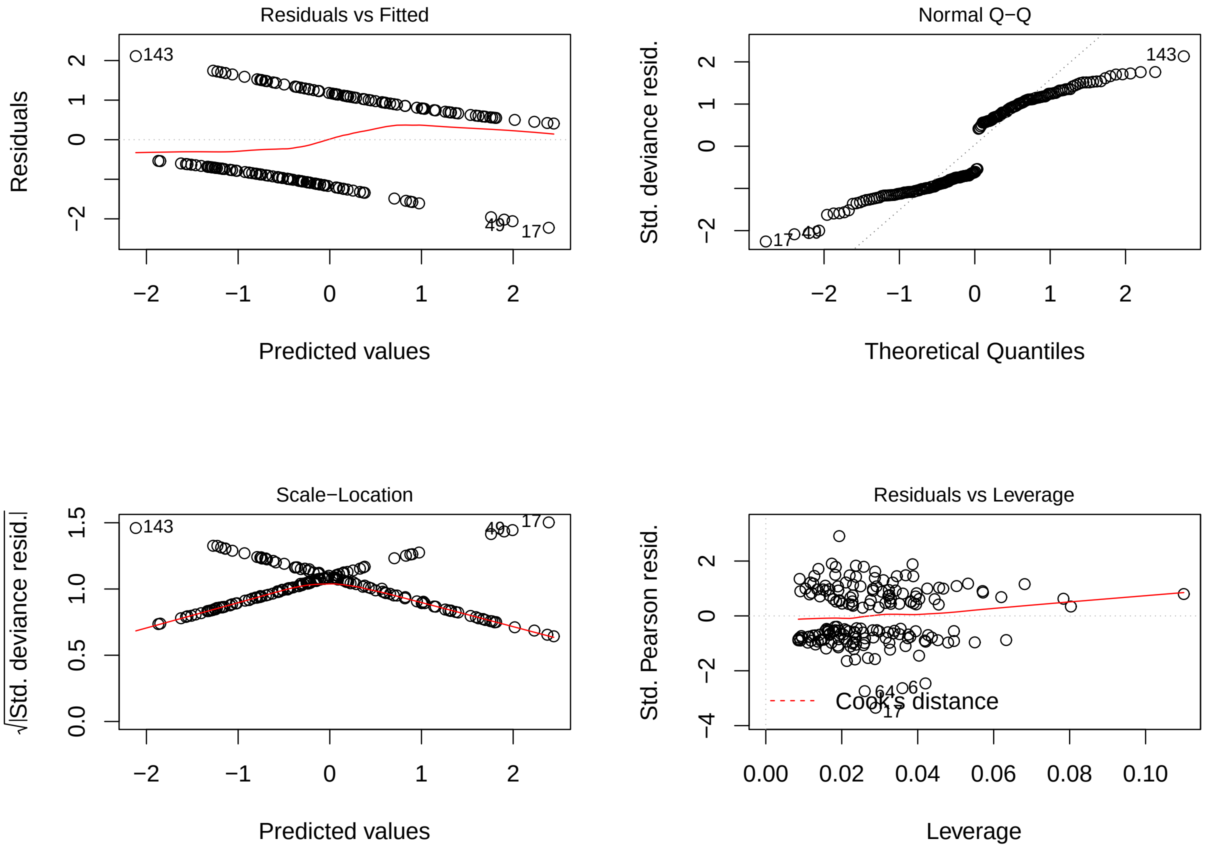

Figure 4: Diagnostic plots for logistic regression model

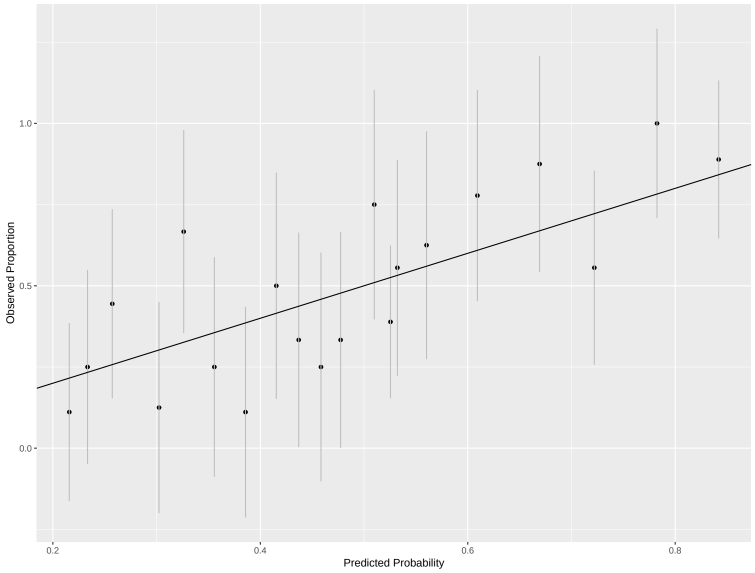

Figure 5: Residual plot based on grouped data

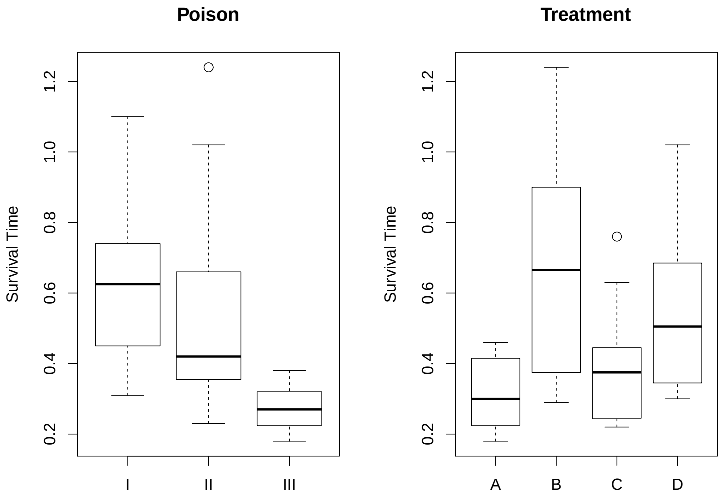

Figure 6: Effects of Treatment and Poison

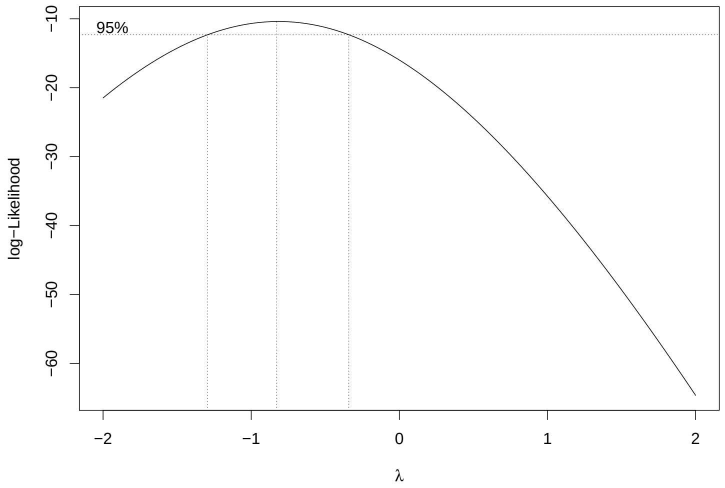

Figure 7: Box-Cox Plot for Rat Survival Model

`l3=lm(tinv poison+treat,rats)` without. The standard model plots (Figure 8) show no problems.

(d) An anova test (`anova(l2,l3,test='F')`) shows that the interaction terms are not statistically significant. Based on that conclusion and Figure 6, the most effective combination is treatment A and poison III. Alternatively, find the observation that maximizes the value of `l2$fitted` (which does take account of the interactions): the max value is at observation 45, which is the same combination.

(e) Previous modeling suggests an inverse link with the inverse Gaussian model (rather than $1/\mu^2$ which is the canonical link). So you can fit models like `g1=glm(time poison*treat,family=inverse.gaussian(link="inverse"),rats)` with interactions or `g2=glm(time poison+treat,family=inverse.gaussian(link="inverse"),rats)` without interactions and a test such as `anova(g1,g2,test='F')` to compare the two (note that the inverse Gaussian model by default includes an estimated dispersion parameter). The test for interactions is again not significant (so use model `g2`) and then a command such as `which.min(g2$fitted)` find the combination with smallest fitted value (fastest death — note that predictions are on the original scale here). This again leads to treatment A and poison III as the fastest killer.

(f) See the plot of predictions from one model against the other, on the original scale. Overall there is little difference.

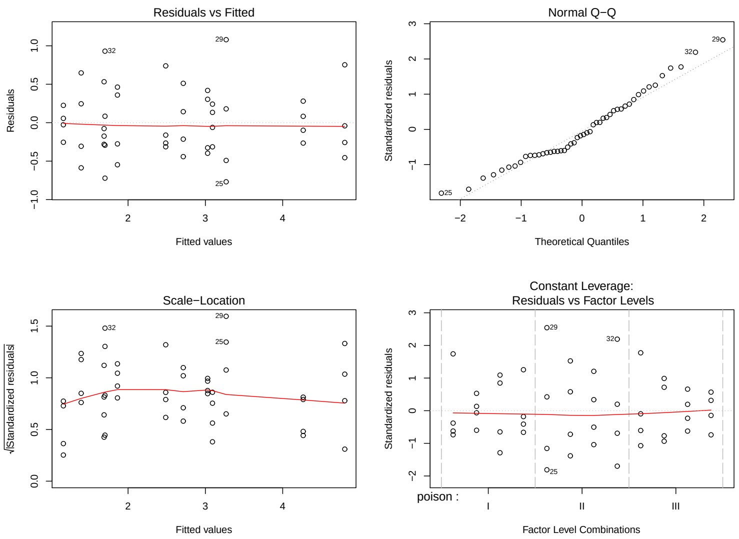

Figure 8: Diagnostic Plots for the Model l2

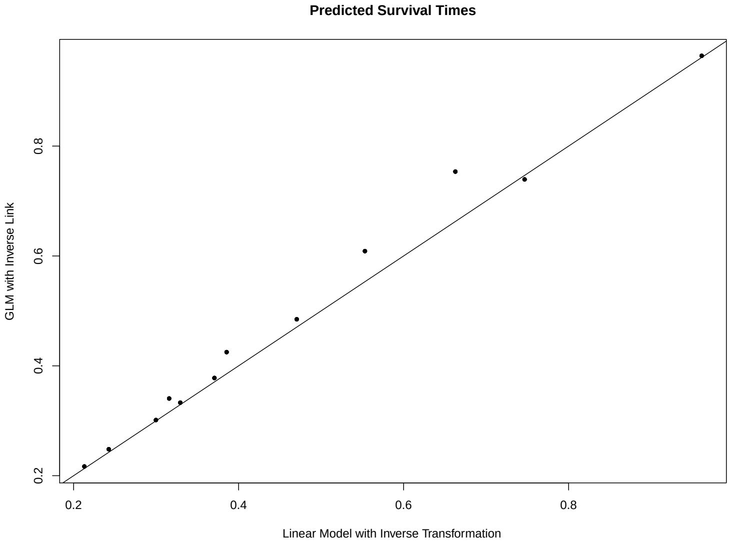

Figure 9: Comparisons of Predictions Under Two Models