SURVIVAL ANALYSIS WITH SAS

HANDIN 1

Andreas Kracht Frandsen*

201506176

* Faculty of Mathematics, Aarhus University, andreas.kracht.frandsen@post.au.dk.

Typeset using Palatino Linotype

Contents

List of Figures

List of Tables

Preface

This document answers Exercise 13, 16 and 22 of Handin 1 in the course Survival Analysis with SAS.

To see an interactive HTML version of this document, with the possibility to edit out mistakes or make comments, go to this website afrandsen.rbind.io/bare/h1saws/. It will be updated continuously if I find any mistakes myself through GitHub.

1 Exercise 13

In this exercise we use the SAS data set houses again. Execute the following code.

```
PROC GLM DATA = houses;
CLASS new;
MODEL price = new size;
OUTPUT OUT = outdata PREDICTED = predvalues;
RUN;
QUIT;
```

Use the data set outdata to produce the same graphs as in Figure 5.1 and Figure 5.2 (Pedersen, 2019), noting in particular the ingenious choice of colors and symbols in the first figure. Save as pdf files.

1.1 SAS Code

1.1.1 The Houses Data Set

In this exercise we use the data set `houses` which is obtained from the course website[1]. Table 1.1 shows the variables and the first 10 observations.

Table 1.1: The first 10 observations of houses.txt.

	taxes	beds	baths	new	price	size
1	3104	4	2	0	279.9	2048
2	1173	2	1	0	146.5	912
3	3076	4	2	0	237.7	1654
4	1608	3	2	0	200.0	2068
5	1454	3	3	0	159.9	1477
6	2997	3	2	1	499.9	3153
7	4054	3	2	0	265.5	1355
8	3002	3	2	1	289.9	2075
9	6627	5	4	0	587.0	3990
10	320	3	2	0	70.0	1160

1.1.2 Import of the Houses Data Set

First we start by importing our data set to our SAS 9.4 session.

```
DATA houses;
  INFILE '~/Survival Analysis/Supplementary Notes/houses.txt'
   FIRSTOBS = 2;
  INPUT case taxes beds baths new price size;
RUN;
```

Thus we take use of the `DATA` step. First we pass on the path to our data using the `INFILE` statement. Since our observations start in the second row, we must use the `FIRSTOBS` argument to tell SAS to start reading

[1] Blackboard, Survival Analysis with SAS.

observations from the second row, by setting it to 2. Next we tell SAS which columns and thereby variables in the houses data set it should read. We want every variable even though we aren't going to use all of them. Thus we use the INPUT statement which takes variables as arguments. (SAS Institute Inc., 2013a).

1.1.3 Model Fitting

We want to analyse the following additive model

$$Y_{ij} \sim N\left(\alpha + \beta_i + \gamma s_{ij}, \sigma^2\right),$$

where $i = 0, 1$, $j = 1, \ldots, n_i$, $n_0 = 89$ and $n_1 = 11$. Thus we make the assumption of constant variance. Next we run the GLM procedure as stated in the exercise.

```
PROC GLM DATA = houses;
CLASS new;
MODEL price = new size;
OUTPUT OUT = outdata PREDICTED = predvalues;
RUN;
QUIT;
```

We obtain the new data set outdata which is the same as the houses data set, but with an extra variable predvalues with the predicted prices. This variable can be spotted in Table 1.2 below. (SAS Institute Inc., 2011b).

Table 1.2: The first 10 observations of outdata.sas7bdat.

	taxes	beds	baths	new	price	size	predvalues
1	3104	4	2	0	279.9	2048	197.6
2	1173	2	1	0	146.5	912	65.7
3	3076	4	2	0	237.7	1654	151.9
4	1608	3	2	0	200.0	2068	199.9
5	1454	3	3	0	159.9	1477	131.3
6	2997	3	2	1	499.9	3153	383.7
7	4054	3	2	0	265.5	1355	117.1
8	3002	3	2	1	289.9	2075	258.5
9	6627	5	4	0	587.0	3990	423.1
10	320	3	2	0	70.0	1160	94.5

1.1.4 Creation of Plots

Before creating our plots we need to decide where to save them to and in which format.

```
ODS PDF FILE = '~/Survival Analysis/Supplementary Notes/Graph1.pdf' NOTOC;
OPTIONS NODATE
        NONUMBER
        ORIENTATION = LANDSCAPE;
```

Thus we use the ODS PDF statement. Where ODS is SAS' output delivery system. In this way we tell SAS to open a new pdf file that it can write to. Using the FILE argument we pass the physical path where we would like to save our pdf to. The NOTOC make sure that no table of contents is attached to our pdf file. Next we want to make sure that we get a plain file to work with, without dates and numbers this is handled by the NODATE and NONUMBER arguments of the OPTIONS statement. Lastly we want the orientation of our pdf to be in landscape[2], this is taking care of by the ORIENTATION argument. Luckily enough we don't have to remove

[2] Because of LaTeX reasons.

any graph titles this time, otherwise arguments such as NOPROCTITLE, NOGTITLE, NOBYLINE, TITLE would be helpful[3]. (SAS Institute Inc., 2011c).

Now that our pdf is open and ready to write we can make the plot.

```
ODS GRAPHICS ON / NOBORDER
                  ATTRPRIORITY = NONE
                  HEIGHT = 9IN;

PROC SGPLOT DATA = outdata;
SCATTER X = size Y = price / GROUP = new;

REG X = size Y = predvalues / GROUP = new
                              NOMARKERS
                              LINEATTRS = (PATTERN = SOLID);

STYLEATTRS DATASYMBOLS = (TRIANGLE STAR)
           DATACONTRASTCOLORS = (BLUE ORANGE);

YAXIS LABEL = "Selling price";
XAXIS LABEL = "Size";

KEYLEGEND / NOBORDER DOWN = 2;
RUN;

ODS GRAPHICS / RESET = ALL;
ODS GRAPHICS OFF;

ODS PDF CLOSE;
```

Using the output delivery system again we make sure that our following procedures will produce the plots, this is handled by ODS GRAPHICS ON. We want to make sure that: no border is present, that we can distinguish groups by colors and markers and that we produce a relatively big plot in our pdf. This can be achieved using the options of our ODS GRAPHICS statement namely the arguments NOBORDER, ATTRPRIORITY and HEIGHT or WIDTH. (SAS Institute Inc., 2011c).

Next we create our plot using the procedure SGPLOT which is compatible with ODS GRAPHICS. To make a scatter plot we use the SCATTER statement with the corresponding X and Y axis settings, set accordingly. We want different colors depending on our group variable new thus we set the GROUP argument in the options correspondingly.

We want SAS to make a regression line per group using the predvalues as the response variable. The REG statement creates regression lines on our plot, by setting the Y and X setting correctly we obtain the regression line. To get both of the regression lines, we again take use of the option argument GROUP. REG does also create markers for every observation, but we don't want these so we remove them using NOMARKERS. Lastly we want both lines to be solid, thus we set the LINEATTRS correspondingly.

We want: the old houses to have a star as a marker, orange as it's corresponding color, new houses to have a triangle and blue as it's corresponding color. Thus by using the STYLEATTRS statement with the arguments DATASYMBOLS and DATACONTRASTCOLORS we can set the specific markers and colors.

The KEYLEGEND statement controls the settings regarding our plot legend, the DOWN option argument controls the number of rows in the legend. YAXIS and XAXIS control the axis labels. (SAS Institute Inc., 2013c).

Lastly we make sure to RESET and turn OFF our ODS GRAPHICS such that other plots in the same *session* behave *cleanly*. A lot of headache can arise if we forget to reset the output delivery system.

Figure 1.1 below shows the resulting plot.

The SAS code below is almost identical to the above. I will explain SGPANEL below.

[3] A style template would give us the possibility to remove *any* titles.

```
ODS PDF FILE = '~/Survival Analysis/Supplementary Notes/Graph2.pdf' NOTOC;
OPTIONS NODATE
        NONUMBER
        ORIENTATION = LANDSCAPE;

ODS GRAPHICS ON / NOBORDER HEIGHT = 9IN;

PROC SGPANEL DATA = outdata;
PANELBY new / COLUMNS = 2
              ROWS = 1;

SCATTER X = size Y = price;

REG x = size y = predvalues / NOMARKERS;

ROWAXIS LABEL = "Selling price";
COLAXIS LABEL = "Size";

KEYLEGEND / NOBORDER;
RUN;

ODS GRAPHICS / RESET = ALL;
ODS GRAPHICS OFF;

ODS PDF CLOSE;
```

The SGPANEL procedure splits a plot into a matrix of panels or facets. In Figure 1.1 we had both new and old house observations in the same plot. This time we want to seperate them, thus we use the PANELBY statement and tell SAS to use new as the classification variable. We want the two plots to be located beside each other, thus we set the option argument COLUMNS to 2 and ROWS to 1. The rest of the statements in SGPANEL is similiar to those in SGPLOT above. This time we fortunatly don't need to set the colors or markers specifically. (SAS Institute Inc., 2013b).

Figure 1.2 below shows the resulting plot.

1.2 Plots

1.2.1 The First Figure

Figure 1.1 below show the pdf output achieved from the first ODS PDF statement in the SAS code from Section 1.1.4.

1.2.2 The Second Figure

Figure 1.2 below show the pdf output achieved from the second ODS PDF statement in the SAS code from Section 1.1.4.

We notice that Figure 1.1 and Figure 1.2 are indeed equal to Figure 5.1 and Figure 5.2 (Pedersen, 2019).

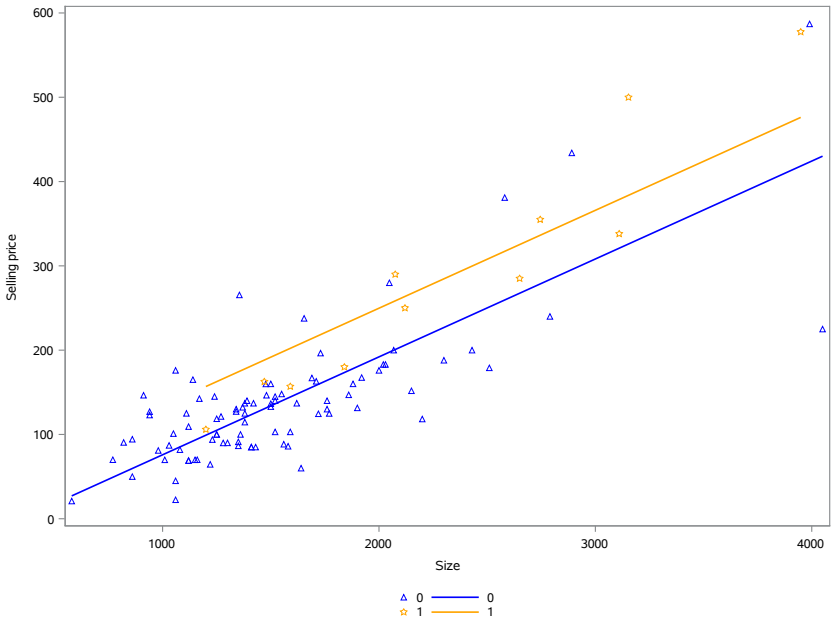

Figure 1.1: Output from the SGPlot Procedure.

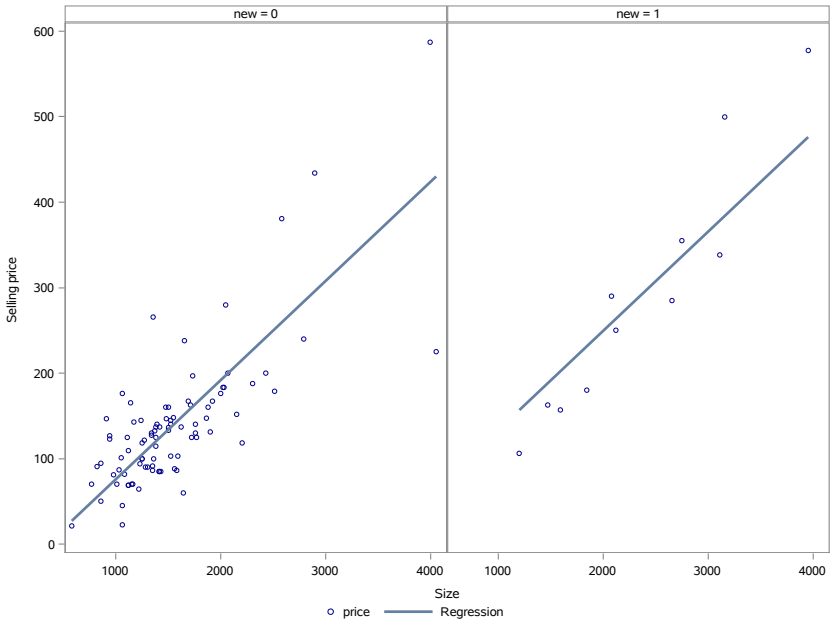

Figure 1.2: Output from the SGPanel Procedure.

2 Exercise 16

Exercise 1.21 (Agresti, 2015) presented a study comparing forced expiratory volume after 1 hour of treatment for three drugs (a, b, and p = placebo), adjusting for a baseline measurement x_1. Table 4.1 (Agresti, 2015) shows the results of fitting some normal GLMs (with identity link, except one with log link) and a GLM assuming a gamma response. Interpret results.

1. *Do all analyses in SAS*
2. *Write up the models mathematically.*
3. *Include graphs to illustrate (lack of) fit.*
4. *Which models (or which model) seem to perform best?*

2.1 Models

In this exercise we use the dataset `FEV` which is obtained from (College of Liberal Arts and Sciences, 2000). Table 2.1 shows the variables and the first 10 observations.

Table 2.1: The first 10 observations of FEV.dat.

	base	fev1	fev2	fev3	fev4	fev5	fev6	fev7	fev8	drug
1	2.46	2.68	2.76	2.50	2.30	2.14	2.40	2.33	2.20	a
2	3.50	3.95	3.65	2.93	2.53	3.04	3.37	3.14	2.62	a
3	1.96	2.28	2.34	2.29	2.43	2.06	2.18	2.28	2.29	a
4	3.44	4.08	3.87	3.79	3.30	3.80	3.24	2.98	2.91	a
5	2.80	4.09	3.90	3.54	3.35	3.15	3.23	3.46	3.27	a
6	2.36	3.79	3.97	3.78	3.69	3.31	2.83	2.72	3.00	a
7	1.77	3.82	3.44	3.46	3.02	2.98	3.10	2.79	2.88	a
8	2.64	3.67	3.47	3.19	2.19	2.85	2.68	2.60	2.73	a
9	2.30	4.12	3.71	3.57	3.49	3.64	3.38	2.28	3.72	a
10	2.27	2.77	2.77	2.75	2.75	2.71	2.75	2.52	2.60	a

We will study the following models:

- Model with base as the only explanatory variable, Section 2.1.1.
- Model with drug as the only explanatory variable, Section 2.1.2.
- Model with base and drug as the explanatory variables, Section 2.1.3.
- Model with base and drug as the explanatory variables and using the Gamma distribution, Section 2.1.4.
- Model with base and drug as the explanatory variables and using log as the link function, Section 2.1.5.
- Model with base, drug and base·drug as the only explanatory variables, Section 2.1.6.

Using SAS to obtain parameter estimates and test values, Section 2.2. Next visual inspection of various regression analysis plots (Cook's Distance, Residuals, Histograms, etc.) and information criterias (AIC, BIC, etc.), will be used to determine which models perform worse than others.

2.1.1 Model – base

Using base as the only explanatory variable we get the following model for our response, fev1:

$$Y_{ij} \sim N\left(\alpha + \gamma b_{ij}, \sigma^2\right),$$

where $i = a, b, p$, $j = 1, \ldots, 24$. I have chosen b_{ij} to be the baseline measurement, instead of x_1. Thus the mean in our model depends linearly on the baseline measurement, with the same slope and intercept for all groups.

The above model is a NLM with $\beta = (\alpha, \gamma)^\mathsf{T}$.

2.1.2 Model – drug

Using drug as the only explanatory variable we get the following model for our response, fev1:

$$Y_{ij} \sim N\left(\alpha + \beta_i, \sigma^2\right),$$

where $i = a, b, p$, $j = 1, \ldots, 24$. Thus the mean is a constant within each group of drug, but with different intercepts.

The above model is a NLM with $\beta = \left(\alpha, \beta_a, \beta_b, \beta_p\right)^\mathsf{T}$.

2.1.3 Model – base+drug

Using base and drug as the explanatory variables we get the following model for our response, fev1:

$$Y_{ij} \sim N\left(\alpha + \gamma b_{ij} + \beta_i, \sigma^2\right),$$

where $i = a, b, p$, $j = 1, \ldots, 24$. The mean in the above model depends linearly on the baseline measurement, with the same slope for all groups and with different intercepts for all groups.

The above model is a NLM with $\beta = \left(\alpha, \gamma, \beta_a, \beta_b, \beta_p\right)^\mathsf{T}$.

2.1.4 Model – base+drug (gamma)

Using base and drug as the explanatory variables in a Gamma distribution we get the following model for our response, fev1:

$$Y_{ij} \sim \Gamma\left(\alpha + \gamma b_{ij} + \beta_i, k\right),$$

where $i = a, b, p$, $j = 1, \ldots, 24$. The mean in this model depends linearly on the baseline measurement, with the same slope for all groups and with different intercepts for all groups.

The above model is a GLM with $\beta = \left(\alpha, \gamma, \beta_a, \beta_b, \beta_p\right)^\mathsf{T}$. The variance in this model is given by

$$\mathbb{V}\left(Y_{ij}\right) = \frac{\mathbb{E}\left(Y_{ij}\right)^2}{k}.$$

Thus the variance increases with the mean of fev1.

2.1.5 Model – base+drug (log link)

Using base and drug as the explanatory variables with a log link function we get the following model for our response, fev1:

$$Y_{ij} \sim N\left(\mu_{ij}, \sigma^2\right),$$

where $i = a, b, p$, $j = 1, \ldots, 24$. The log of the mean of Y_{ij} in this model depends linearly on the baseline measurement, with the same slope for all groups and with different intercepts for all groups.

$$\mu_{ij} = \log(\mathbb{E}(Y_{ij})) = \alpha + \gamma b_{ij} + \beta_i.$$

The above model is a NLM with $\beta = \left(\alpha, \gamma, \beta_a, \beta_b, \beta_p\right)^\mathsf{T}$.

2.1.6 Model – base+drug+base*drug

Using base, drug and base*drug as the explanatory variables we get the following model for our response, fev1:

$$Y_{ij} \sim N\left(\alpha + \gamma b_{ij} + \beta_i + \delta_i b_{ij}, \sigma^2\right),$$

where $i = a, b, p$, $j = 1, \ldots, 24$. This model's mean depends linearly on the baseline measurement, with different slopes for all groups and with different intercepts for all groups.

The above model is a NLM with $\beta = \left(\alpha, \gamma, \beta_a, \beta_b, \beta_p, \delta_a, \delta_b, \delta_p\right)^{\mathsf{T}}$.

2.2 Analysis in SAS

2.2.1 Model – base

The following show the analysis for the first model. The SAS code, SAS outputs and figures for the rest of the models are available in Appendix A.

```
PROC GENMOD DATA = fev PLOTS(UNPACK) = ALL;
CLASS drug;
MODEL fev1 = base / DIST = NORMAL LINK = IDENTITY TYPE3;
RUN;
```

The following listing show some of the output from the above `GENMOD` procedure. (SAS Institute Inc., 2011a).

```
                    Criteria For Assessing Goodness Of Fit

          Criterion                     DF           Value        Value/DF

          Deviance                      70         25.0952          0.3585
          Scaled Deviance               70         72.0000          1.0286
          Pearson Chi-Square            70         25.0952          0.3585
          Scaled Pearson X2             70         72.0000          1.0286
          Log Likelihood                          -64.2200
          Full Log Likelihood                     -64.2200
          AIC (smaller is better)                 134.4400
          AICC (smaller is better)                134.7929
          BIC (smaller is better)                 141.2700


      Algorithm converged.


                Analysis Of Maximum Likelihood Parameter Estimates

                                Standard      Wald 95% Confidence          Wald
Parameter     DF    Estimate       Error            Limits           Chi-Square    Pr > ChiSq

Intercept      1      0.9480      0.3558      0.2506      1.6454          7.10        0.0077
base           1      0.8989      0.1317      0.6408      1.1571         46.58        <.0001
Scale          1      0.5904      0.0492      0.5014      0.6951

NOTE: The scale parameter was estimated by maximum likelihood.

                                The GENMOD Procedure

                        LR Statistics For Type 3 Analysis
```

Source	DF	Chi-Square	Pr > ChiSq
base	1	35.92	<.0001

We notice that DF $= 70$ corresponding to the fact that the dimension of subspace in our model is 2 and there are $n. = 72$ observations. We cannot remove base from our model as there is a significant effect ($\chi^2 \approx 35.92$, $p < 0.0001$), which is as expected.

Under `Estimate` we can find the estimate of the intercept, base coefficient and the variance. Thus we obtain the following estimates of our model.[1]

$$\hat{\alpha} = 0.9480, \quad \hat{\gamma} = 0.8989, \quad s^2 = 0.3485722.$$

The associated AIC of the model is 134.44.

If we take a look at the associated plots provided by `GENMOD` and a similar `GLM` procedure, we quickly notice problems with our model. First we take a look at the QQ plot.

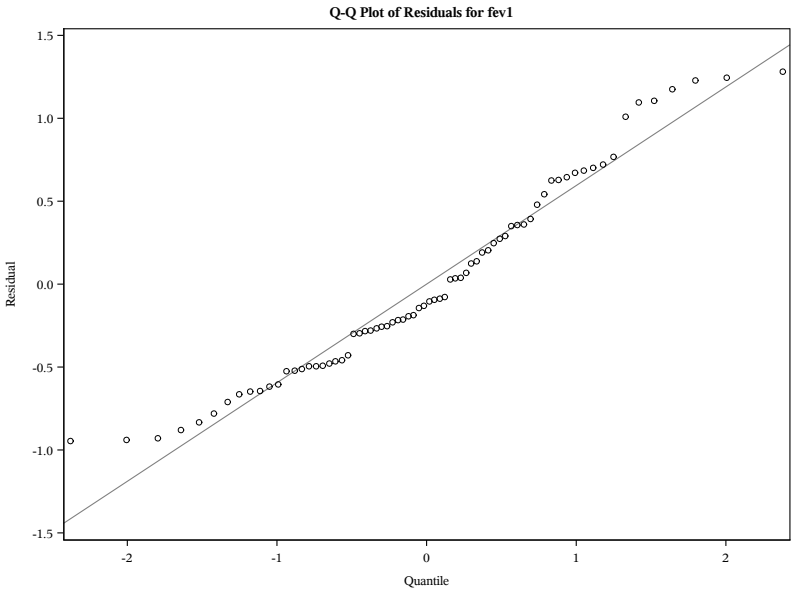

Figure 2.1: QQ Plot of Model – base.

Figure 2.1 show a slight left skew and heavy-tailedness of our model. This can also be confirmed by taking a look at the empirical distribution held up against the normal distribution, as in Figure 2.2.

Lastly Cook's Distance, Figure 2.3, indicate some possible outliers which affects the model estimates more than other observations. These are observations (patients) 7 and 64. A peek in the data set indicates that observation 7 has a very low base value and a relatively high fev1 value. Observation 64 has a lower base value than fev1 value. These two observations could also turn out to create problems in the other models.

[1] In a NLM, `Scale` is the estimate of the standard deviation.

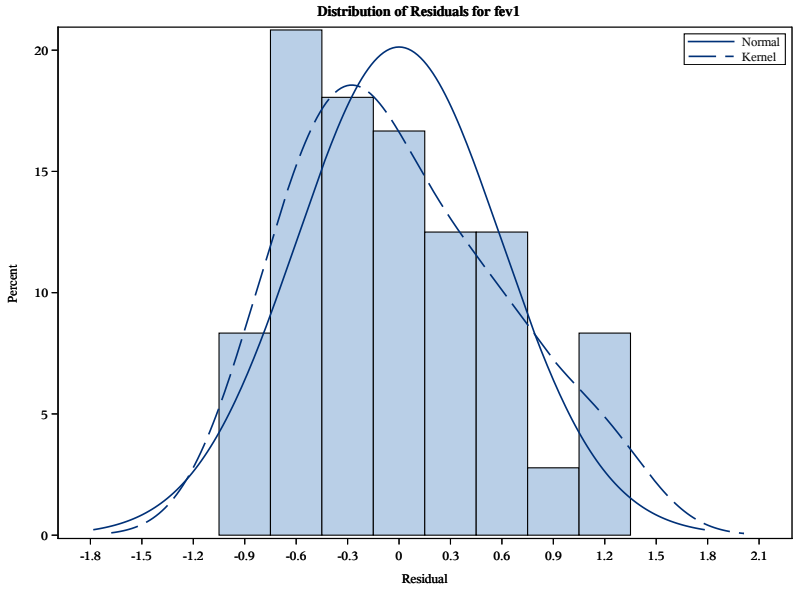

Figure 2.2: Distribtion of Residuals Plot of Model – base.

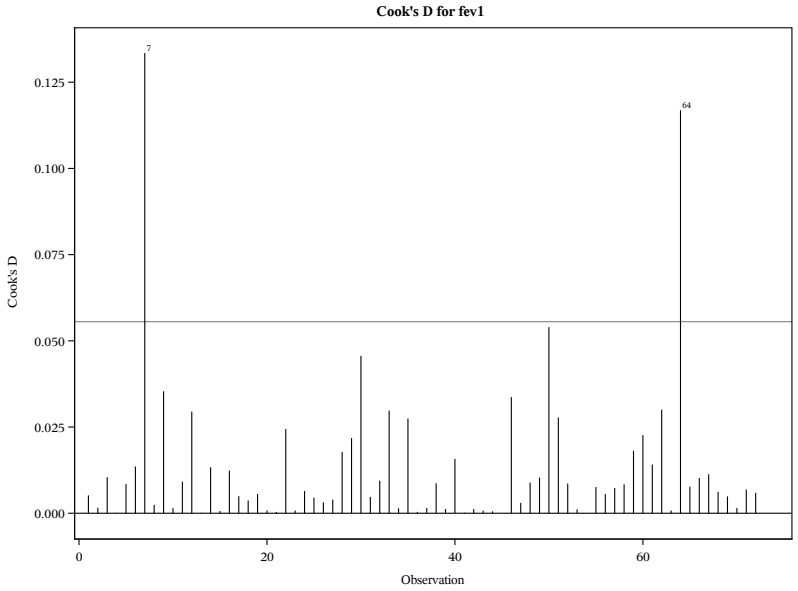

Figure 2.3: Cook's Distance Plot of Model – base.

2.2.2 Model – drug

We obtain the following estimates of our model.

$$\hat{\alpha} = 3.4887, \quad \hat{\beta} = \begin{pmatrix} 0 \\ 0.1963 \\ -0.6737 \end{pmatrix}, \quad s^2 = 0.4352041.$$

In this model (and the rest of the models) we need to set $\beta_a = 0$ before we can estimate the rest of the coefficents. Due to the fact noted in (Pedersen (2019), Section 1.1.3).

In the SAS output A.2.1 we see a significant effect of drug ($\chi^2 \approx 19.93$, $p < 0.0001$). We notice that a Wald Test of $H_0 : \beta_b = 0$ can't be rejected ($W \approx 1.06$, $p \approx 0.3028$). Thus we can't reject that drug b have no effect over drug a.

In Figure A.2 we again notice a skew of the distribution, this time to the right. Fat tailes are again a problem in the model. And under this model several observations seem to have a great effect of the regression as seen in Figure A.3.

The AIC of this model is 152.43, thus worse than the first model.

2.2.3 Model – base+drug

We obtain the following estimates of our model.

$$\hat{\alpha} = 1.1139, \quad \hat{\gamma} = 0.8900, \quad \hat{\beta} = \begin{pmatrix} 0 \\ 0.2181 \\ -0.6448 \end{pmatrix}, \quad s^2 = 0.214369.$$

In the SAS output A.2.2 we see a significant effect of drug ($\chi^2 \approx 35.01$, $p < 0.0001$) and base ($\chi^2 \approx 50.99$, $p < 0.0001$). Thus we cannot reduce the model to the one in 2.1.1 or 2.1.1. This time we again can't reject that $H_0 : \beta_b = 0$ ($W \approx 2.66$, $p \approx 0.1027$).

Figure A.5 suggests a better model compared to the two models above. But the QQ plot in Figure A.4 still show a line that isn't straight. Observation 7 seem to be a problem again.

Figure A.7, showing the standardized Pearson residuals by XBeta, suggests a non constant variance. We observe a possible inverted trumpet shape, and thus heteroskedastic data. On the other hand, we have a lack of observations for high values of the linear predictor.

The AIC of this model is 103.43. This is better than the two previous models.

2.2.4 Model – base+drug (gamma)

We obtain the following estimates of our model.

$$\hat{\alpha} = 0.9302, \quad \hat{\gamma} = 0.9654, \quad \hat{\beta} = \begin{pmatrix} 0 \\ 0.1998 \\ -0.6628 \end{pmatrix}, \quad \hat{k} = 47.3447.$$

For the model with the Gamma distribution we again can't reject that the drug b treatment doesn't have an effect over drug a ($W \approx 1.88$, $p \approx 0.1698$), as seen in A.2.3

Figure A.9 could possibly suggest a non constant variance. Thus we again observe a possible inverted trumpet shape, and thus heteroskedastic data.

The AIC of this model is 106.16, thus not better than drug+base (normal).

2.2.5 Model – base+drug (log link)

We obtain the following estimates of our model.

$$\hat{\alpha} = 0.5598, \quad \hat{\gamma} = 0.2543, \quad \hat{\beta} = \begin{pmatrix} 0 \\ 0.0634 \\ -0.1992 \end{pmatrix}, \quad s^2 = 0.2244864.$$

In the log link model, we see in A.2.4 that we still can't reject the hypothesis $H_0 : \beta_b = 0$ ($W \approx 2.80$, $p \approx 0.0943$).

The Figures in A.2.4 seem to give the same conclusions as in 2.2.4.

The AIC of this model is 106.75.

2.2.6 Model – base+drug+base*drug

We obtain the following estimates of our model.

$$\hat{\alpha} = 1.3316, \quad \hat{\gamma} = 0.8084, \quad \hat{\beta} = \begin{pmatrix} 0 \\ -0.1745 \\ -0.9147 \end{pmatrix}, \quad \hat{\delta} = \begin{pmatrix} 0 \\ 0.1478 \\ 0.1014 \end{pmatrix}, \quad s^2 = 0.2132592.$$

The last model give us an additional interaction term to deal with. From the SAS output in A.2.5 we can't reject that base*drug doesn't have an effect as seen in the Type 3 Analysis ($\chi^2 \approx 0.37$, $p \approx 0.8305$). The same conclusion can be made on drug ($\chi^2 \approx 1.94$, $p \approx 0.3798$). Thus we can't reject the reduction to 2.1.1

This time we have several hypothesis' we can't reject: $H_0 : \beta_b = 0$, $H_0 : \beta_p = 0$, $H_0 : \delta_b = 0$ and $H_0 : \delta_p = 0$.

Again we have the same conclusion regarding Figure A.12 and Figure A.13 as in the two previous models.

The AIC of this model is 107.06.

2.3 Model Conclusion

From a pure AIC perspective the base+drug model performs better than the others, as a sidenote it can be seen that the associated BIC also is smallest in this model. Since AIC/BIC penalizes many parameters they favor simple models. The models with one explanatory variable is too simple though, and cannot explain all the variation in fev1, thus they also have a high AIC. As it can be seen in 2.4 below, the base+drug, base+drug (gamma) and base+drug+base models have very similar lines for the predicted values.

Table 2.2 gives a summary of the estimates.

Table 2.2: Summary of all models.

Explanatory Variables	Fitted Linear Predictor	AIC
base	$0.95 + 0.9 \cdot b_{ij}$	134.4
drug	$3.49 + 0.20_b - 0.67_p$	152.4
base+drug	$1.11 + 0.89 \cdot b_{ij} + 0.22_b - 0.64_p$	103.4
bas+drug (gamma)	$0.93 + 0.97 \cdot b_{ij} + 0.2_b - 0.66_p$	106.2
base+drug (log link)	$0.55 + 0.25 \cdot b_{ij} + 0.06_b - 0.2_p$	106.8
base+drug+base*drug	$1.33 + 0.8 \cdot b_{ij} - 0.17_b - 0.91_p + 0.15_b \cdot b_{ij} + 0.1_p \cdot b_{ij}$	107.1

Figure 2.4 gives a summary of the models.

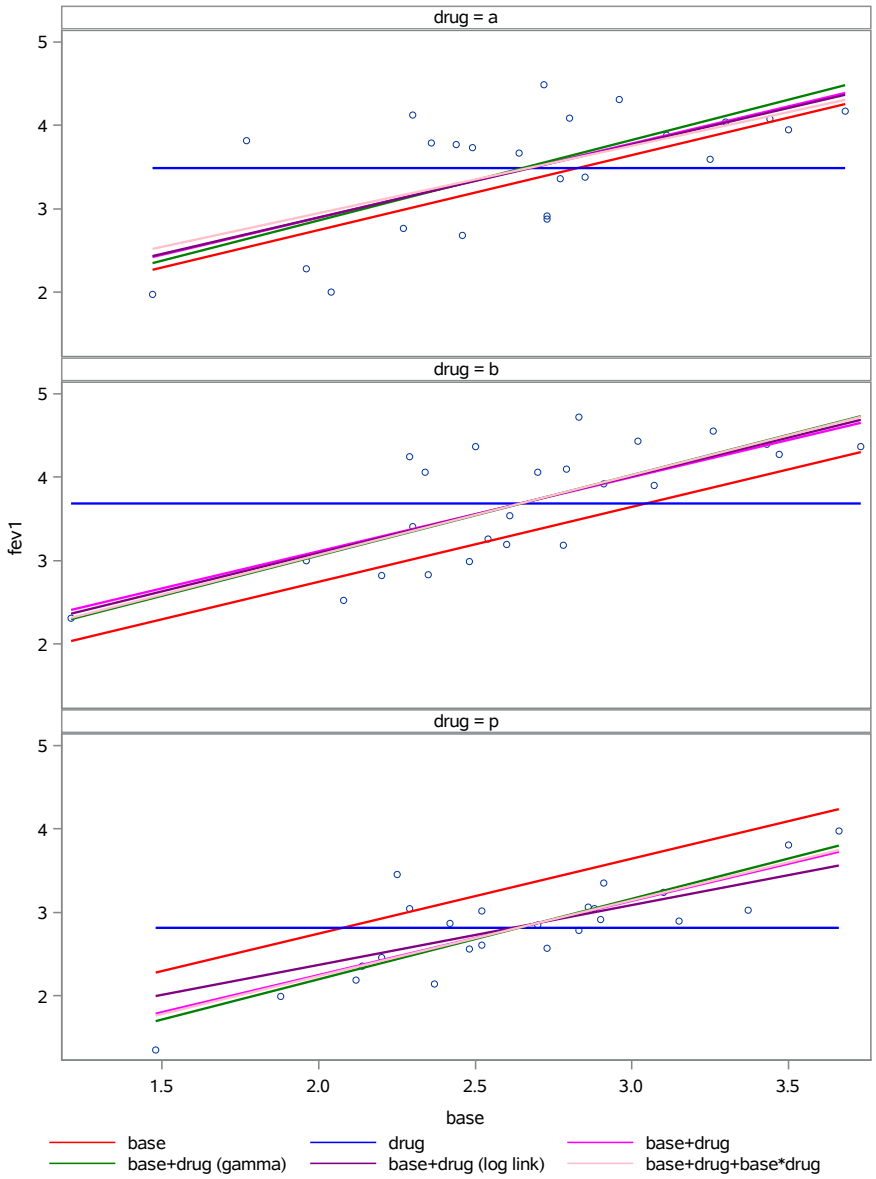

Figure 2.4: Visual summary of all models.

3 Exercise 22

Let X have a Weibull distribution with parameters $\alpha, \lambda > 0$, that is, X has density f as indicated in Table 2.2 in Klein and Moeschberger (2003).

3.1 Question 1

Show that h and S are as indicated in Klein and Moeschberger (2003), Table 2.2. Determine H as well.

3.1.1 Weibull Distributed Random Variable

The random variable $X \sim \text{Weibull}(\alpha, \lambda)$ with $\alpha, \lambda > 0$, has the following probability density function

$$f(x) = \alpha\lambda x^{\alpha-1} \exp(-\lambda x^{\alpha}) \quad x \geq 0. \tag{3.1}$$

3.1.2 Survival Function

For a continuous random variable the survival function is defined as the complement of the cumulative distribution function

$$S(x) = 1 - F(x) = P(X > x) = \int_x^{\infty} f(t)dt. \tag{3.2}$$

Thus for the continouos random variable X we use Equation (3.2).

We insert Equation (3.1) in Equation (3.2) and perform the calculation for $x \geq 0$.

$$\begin{aligned}
S(x) &= \int_x^{\infty} \alpha\lambda t^{\alpha-1} \exp(-\lambda t^{\alpha})dt \\
&= \alpha\lambda \int_x^{\infty} t^{\alpha-1} \exp(-\lambda t^{\alpha})dt \\
&= \alpha\lambda \int_{x^{\alpha}}^{\infty} t^{\alpha-1} \exp(-\lambda u) \frac{1}{\alpha t^{\alpha-1}} du \\
&= \lambda \int_{x^{\alpha}}^{\infty} \exp(-\lambda u)du \\
&= [-\exp(-\lambda u)]_{x^{\alpha}}^{\infty} \\
&= (-\exp(-\lambda \cdot \infty) - (-\exp(-\lambda \cdot x^{\alpha}))) \\
&= (0 - (-\exp(-\lambda \cdot x^{\alpha}))) \\
&= \exp(-\lambda x^{\alpha}).
\end{aligned} \tag{3.3}$$

Where we in the third and fifth equality perform integration by substitution. We see that the above function is indeed equal to the survival function in Table 2.2 (Klein and Moeschberger, 2003).

3.1.3 Hazard Function

For a continuous random variable the hazard function (rate) is defined as

$$h(x) = \frac{f(x)}{S(x)}. \tag{3.4}$$

We insert Equation (3.3) and Equation (3.1) in Equation (3.4) and perform the calculation for $x \geq 0$.

$$\begin{aligned} h(x) &= \frac{f(x)}{S(x)} \\ &= \frac{\alpha\lambda x^{\alpha-1}\exp(-\lambda x^{\alpha})}{\exp(-\lambda x^{\alpha})} \\ &= \alpha\lambda x^{\alpha-1}. \end{aligned} \tag{3.5}$$

We see that the above function is indeed equal to the hazard function in Table 2.2 (Klein and Moeschberger, 2003).

3.1.4 Cumulative Hazard Function

For a continuous random variable the cumulative hazard function is defined as

$$H(x) = \int_0^x h(u)du = -\ln(S(x)). \tag{3.6}$$

We insert Equation (3.3) in Equation (3.6) and perform the calculation for $x \geq 0$.

$$H(x) = -\ln(\exp(-\lambda x^{\alpha})) = \lambda x^{\alpha}. \tag{3.7}$$

We see that the above function is indeed equal to the cumulative hazard function in (Klein and Moeschberger (2003), page 32).

Thus we shown that the functions in Table 2.2 (Klein and Moeschberger, 2003) are correct and we are done.

3.2 Question 2

Determine the distribution of X^γ for $\gamma > 0$.

We calculate the survival function of X^γ. Since the distribution of a random variable is fully described by it's survival function.

$$\begin{aligned}
S_{X^\gamma}(x) &= P\left(X^\gamma > x\right) \\
&= P\left(\ln(X^\gamma) > \ln(x)\right) \\
&= P\left(\gamma \ln(X) > \ln(x)\right) \\
&= P\left(\ln(X) > \frac{\ln(x)}{\gamma}\right) \\
&= P\left(\exp(\ln(X)) > \exp\left(\frac{\ln(x)}{\gamma}\right)\right) \\
&= P\left(X > \exp\left(\frac{\ln(x)}{\gamma}\right)\right) \\
&= S_X\left(\exp\left(\frac{\ln(x)}{\gamma}\right)\right) \\
&= \exp\left(-\lambda\left(\exp\left(\frac{\ln(x)}{\gamma}\right)\right)^\alpha\right) \\
&= \exp\left(-\lambda x^{\frac{\alpha}{\gamma}}\right).
\end{aligned} \tag{3.8}$$

Where we in the eigth equality uses Equation (3.3). Thus from Table 2.2 (Klein and Moeschberger, 2003) we have $X^\gamma \sim \text{Weibull}\left(\frac{\alpha}{\gamma}, \lambda\right)$.

3.3 Question 3

Determine the distribution of λX^α.

We calculate the survival function of λX^α. Since the distribution of a random variable is fully described by it's survival function.

$$\begin{aligned} S_{\lambda X^\alpha}(x) &= P\left(\lambda X^\alpha > x\right) \\ &= P\left(X^\alpha > \frac{x}{\lambda}\right) \\ &= S_{X^\alpha}\left(\tfrac{x}{\lambda}\right) \qquad (3.9) \\ &= \exp\left(-\lambda\left(\frac{x}{\lambda}\right)^{\frac{\alpha}{\alpha}}\right). \\ &= \exp(-x). \qquad (3.10) \end{aligned}$$

Where we in the fourth equality uses Equation (3.8). Thus from Table 2.2 (Klein and Moeschberger, 2003) we have $\lambda X^\alpha \sim \text{Weibull}\,(1,1) \overset{\text{d}}{=} \text{Exp}(1)$.

Figure 3.1 below show the distribution.

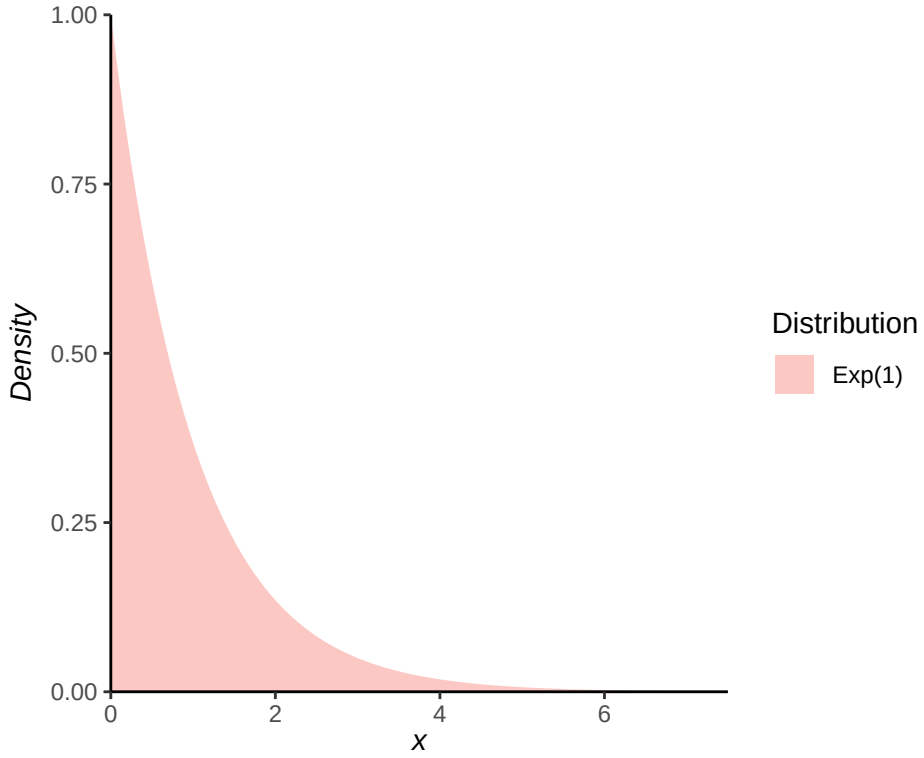

Figure 3.1: Density of a Exp(1) distribution.

3.4 Question 4

Let $n \in \mathbb{N}$ and $X_1, \ldots, X_n$ be i.i.d. and have a Weibull distribution with parameters $\alpha, \lambda > 0$ as common distribution. Determine the distribution of $\min(X_1, \ldots, X_n)$.

The random variable X_i has the survival function

$$S_{X_i}(x) = \exp(-\lambda x^\alpha) \quad x \geq 0,$$

for $i = 1, 2, \ldots, n$. As proved in Section 3.1.2, Equation (3.3).

Let $Y = \min(X_1, \ldots, X_n)$, then the survival function of Y is

$$\begin{aligned} S_Y(y) &= P(Y > y) \\ &= P(\min(X_1, \ldots, X_n) > y) \\ &= P(X_1 > y, X_2 > y, \ldots, X_n > y) \\ &= \Pi_{i=1}^n P(X_i > y) \\ &= \Pi_{i=1}^n S_{X_i}(y) \\ &= \Pi_{i=1}^n (\exp(-\lambda y^\alpha)) \\ &= \exp(-n \cdot \lambda y^\alpha). \end{aligned} \tag{3.11}$$

Where we in the fourth equality take use of the independency of our random variables. Thus from Table 2.2 (Klein and Moeschberger, 2003) we have $Y = \min(X_1, \ldots, X_n) \sim \text{Weibull}(\alpha, n\lambda)$.

A SAS of Exercise 16

A.1 SAS Code

```
DATA fev;
  INFILE '~/Survival Analysis/Supplementary Notes/FEV.dat'
  FIRSTOBS = 2;
  INPUT patient base fev1 fev2 fev3 fev4 fev5 fev6 fev7 fev8 drug $1.;
RUN;

* Model - base;
PROC GENMOD DATA = fev PLOTS = ALL;
CLASS drug;
MODEL fev1 = base / DIST = NORMAL LINK = IDENTITY TYPE3;
RUN;

* Model - drug;
PROC GENMOD DATA = fev PLOTS = ALL;
CLASS drug(REF = 'a');
MODEL fev1 = drug / DIST = NORMAL LINK = IDENTITY TYPE3;
RUN;

* Model - base+drug;
PROC GENMOD DATA = fev PLOTS = ALL;
CLASS drug(REF = 'a');
MODEL fev1 = base drug / DIST = NORMAL LINK = IDENTITY TYPE3;
RUN;

* Model - base+drug (gamma);
PROC GENMOD DATA = fev PLOTS = ALL;
CLASS drug(REF = 'a');
MODEL fev1 = base drug / DIST = GAMMA LINK = IDENTITY TYPE3;
RUN;

* Model - base+drug (log link);
PROC GENMOD DATA = fev PLOTS = ALL;
CLASS drug(REF = 'a');
MODEL fev1 = base drug / DIST = NORMAL LINK = LOG TYPE3;
RUN;

* Model - base+drug+base*drug;
PROC GENMOD DATA = fev PLOTS = ALL;
CLASS drug(REF = 'a');
MODEL fev1 = base drug base*drug / DIST = NORMAL LINK = IDENTITY TYPE3;
RUN;
```

A.2 SAS Output

A.2.1 Model – drug

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	69	31.3347	0.4541
Scaled Deviance	69	72.0000	1.0435
Pearson Chi-Square	69	31.3347	0.4541
Scaled Pearson X2	69	72.0000	1.0435
Log Likelihood		-72.2137	
Full Log Likelihood		-72.2137	
AIC (smaller is better)		152.4274	
AICC (smaller is better)		153.0244	
BIC (smaller is better)		161.5340	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	3.4887	0.1347	3.2248	3.7527	671.21	<.0001
drug	b	1	0.1963	0.1904	-0.1770	0.5695	1.06	0.3028
drug	p	1	-0.6737	0.1904	-1.0470	-0.3005	12.52	0.0004
drug	a	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	0.6597	0.0550	0.5603	0.7767		

NOTE: The scale parameter was estimated by maximum likelihood.

The GENMOD Procedure

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
drug	2	19.93	<.0001

A.2.2 Model – base+drug

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	68	15.4323	0.2269
Scaled Deviance	68	72.0000	1.0588
Pearson Chi-Square	68	15.4323	0.2269
Scaled Pearson X2	68	72.0000	1.0588
Log Likelihood		-46.7162	
Full Log Likelihood		-46.7162	
AIC (smaller is better)		103.4324	
AICC (smaller is better)		104.3415	
BIC (smaller is better)		114.8157	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.1139	0.2915	0.5427	1.6851	14.61	0.0001
base		1	0.8900	0.1033	0.6875	1.0925	74.19	<.0001
drug	b	1	0.2181	0.1337	-0.0439	0.4801	2.66	0.1027
drug	p	1	-0.6448	0.1337	-0.9068	-0.3828	23.26	<.0001
drug	a	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	0.4630	0.0386	0.3932	0.5451		

NOTE: The scale parameter was estimated by maximum likelihood.

The GENMOD Procedure

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
base	1	50.99	<.0001
drug	2	35.01	<.0001

A.2.3 Model – base+drug (gamma)

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	68	1.5261	0.0224
Scaled Deviance	68	72.2535	1.0626
Pearson Chi-Square	68	1.5931	0.0234
Scaled Pearson X2	68	75.4266	1.1092
Log Likelihood		-48.0806	
Full Log Likelihood		-48.0806	
AIC (smaller is better)		106.1612	
AICC (smaller is better)		107.0703	
BIC (smaller is better)		117.5446	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.9302	0.2738	0.3935	1.4669	11.54	0.0007
base		1	0.9654	0.1026	0.7643	1.1666	88.48	<.0001
drug	b	1	0.1998	0.1455	-0.0854	0.4850	1.88	0.1698
drug	p	1	-0.6628	0.1280	-0.9136	-0.4120	26.83	<.0001
drug	a	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	47.3447	7.8632	34.1902	65.5604		

NOTE: The scale parameter was estimated by maximum likelihood.

The GENMOD Procedure

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
base	1	53.53	<.0001
drug	2	36.52	<.0001

A.2.4 Model – base+drug (log link)

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	68	16.1604	0.2377
Scaled Deviance	68	72.0000	1.0588
Pearson Chi-Square	68	16.1604	0.2377
Scaled Pearson X2	68	72.0000	1.0588
Log Likelihood		-48.3759	
Full Log Likelihood		-48.3759	
AIC (smaller is better)		106.7518	
AICC (smaller is better)		107.6609	
BIC (smaller is better)		118.1352	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.5598	0.0922	0.3791	0.7405	36.86	<.0001
base		1	0.2543	0.0312	0.1932	0.3154	66.54	<.0001
drug	b	1	0.0634	0.0379	-0.0109	0.1376	2.80	0.0943
drug	p	1	-0.1992	0.0437	-0.2849	-0.1134	20.73	<.0001
drug	a	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	0.4738	0.0395	0.4024	0.5578		

NOTE: The scale parameter was estimated by maximum likelihood.

The GENMOD Procedure

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
base	1	47.68	<.0001
drug	2	33.14	<.0001

A.2.5 Model – base+drug+base*drug

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	66	15.3529	0.2326
Scaled Deviance	66	72.0000	1.0909
Pearson Chi-Square	66	15.3529	0.2326
Scaled Pearson X2	66	72.0000	1.0909
Log Likelihood		-46.5305	
Full Log Likelihood		-46.5305	
AIC (smaller is better)		107.0610	
AICC (smaller is better)		108.8110	
BIC (smaller is better)		122.9976	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.3316	0.4733	0.4039	2.2593	7.91	0.0049
base		1	0.8084	0.1738	0.4677	1.1491	21.63	<.0001
drug	b	1	-0.1745	0.6711	-1.4899	1.1408	0.07	0.7948
drug	p	1	-0.9147	0.6879	-2.2630	0.4335	1.77	0.1836
drug	a	0	0.0000	0.0000	0.0000	0.0000	.	.
base*drug	b	1	0.1478	0.2477	-0.3376	0.6332	0.36	0.5507
base*drug	p	1	0.1014	0.2546	-0.3975	0.6003	0.16	0.6904
base*drug	a	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	0.4618	0.0385	0.3922	0.5437		

NOTE: The scale parameter was estimated by maximum likelihood.

The GENMOD Procedure

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
base	1	51.18	<.0001
drug	2	1.94	0.3798
base*drug	2	0.37	0.8305

A.3 SAS Plots

A.3.1 Model – drug

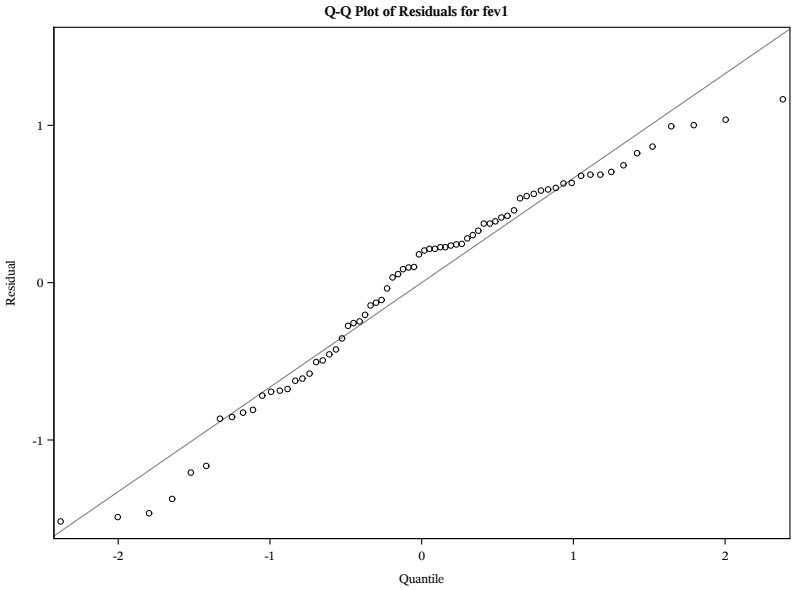

Figure A.1: QQ Plot of Model – drug

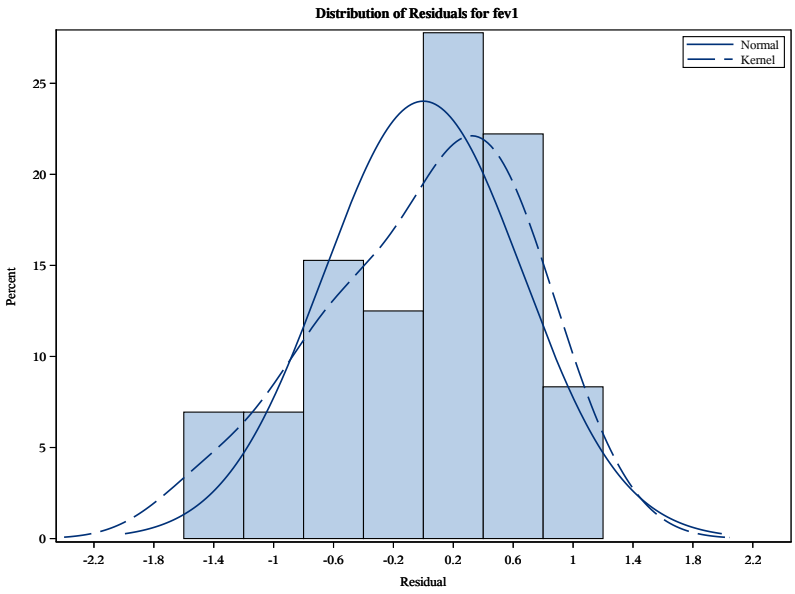

Figure A.2: Distribtion of Residuals Plot of Model – drug.

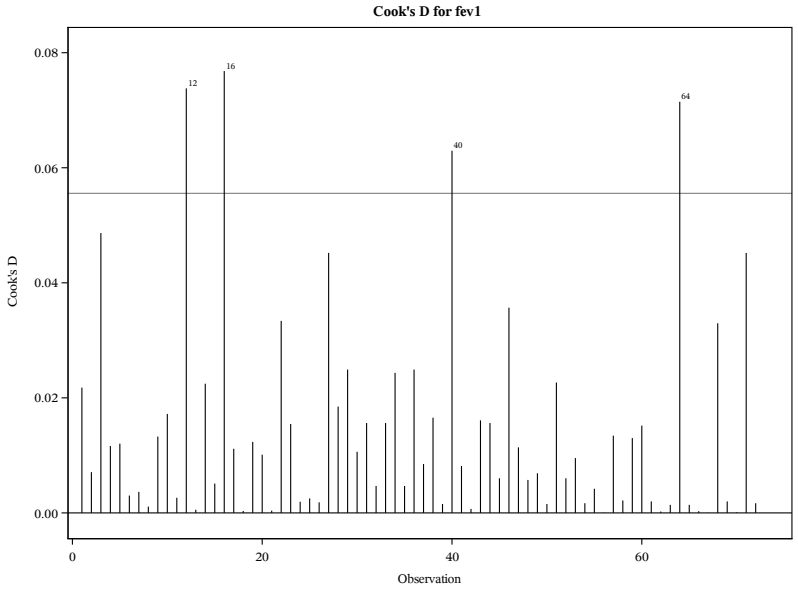

Figure A.3: Cook's Distance Plot of Model – drug

A.3.2 Model – base+drug

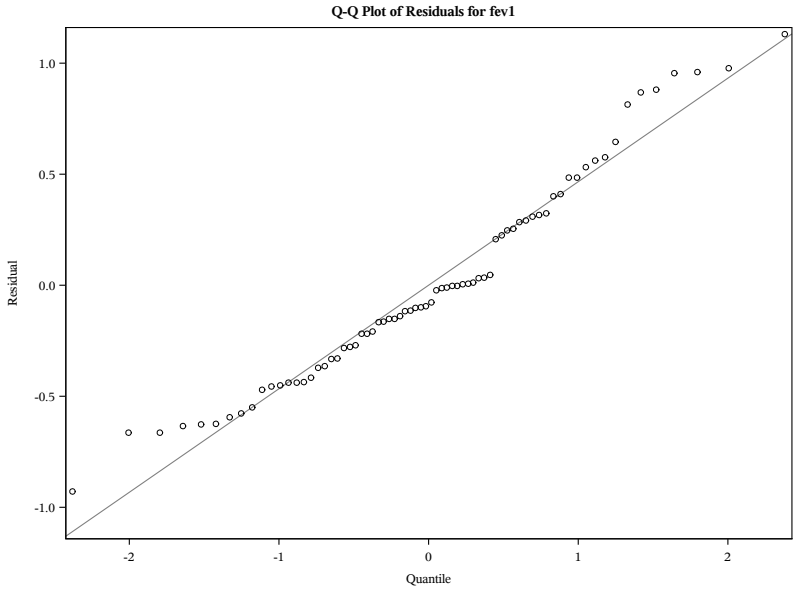

Figure A.4: QQ Plot of Model – base+drug

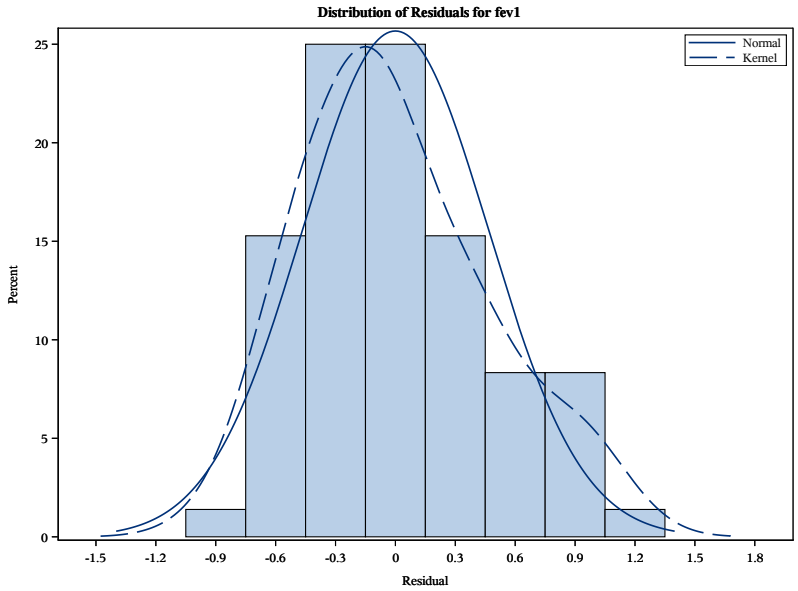

Figure A.5: Distribtion of Residuals Plot of Model – base+drug.

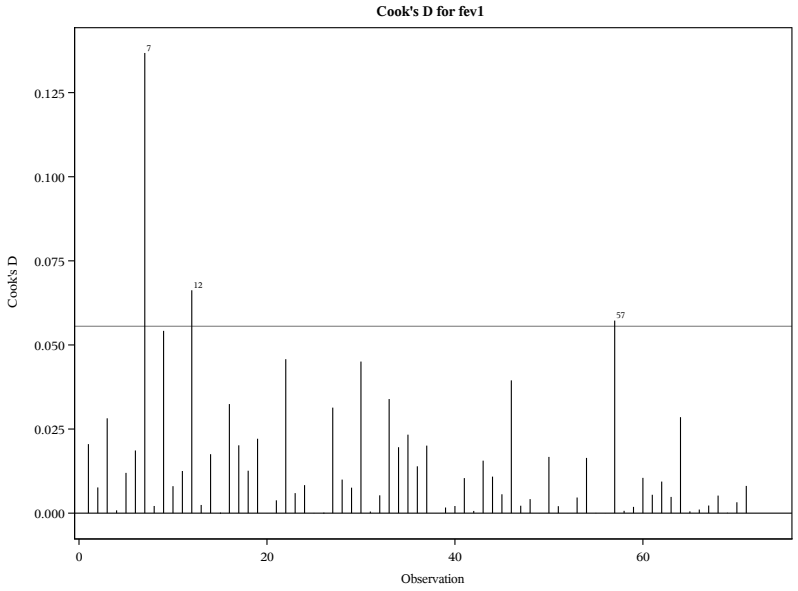

Figure A.6: Cook's Distance Plot of Model – base+drug

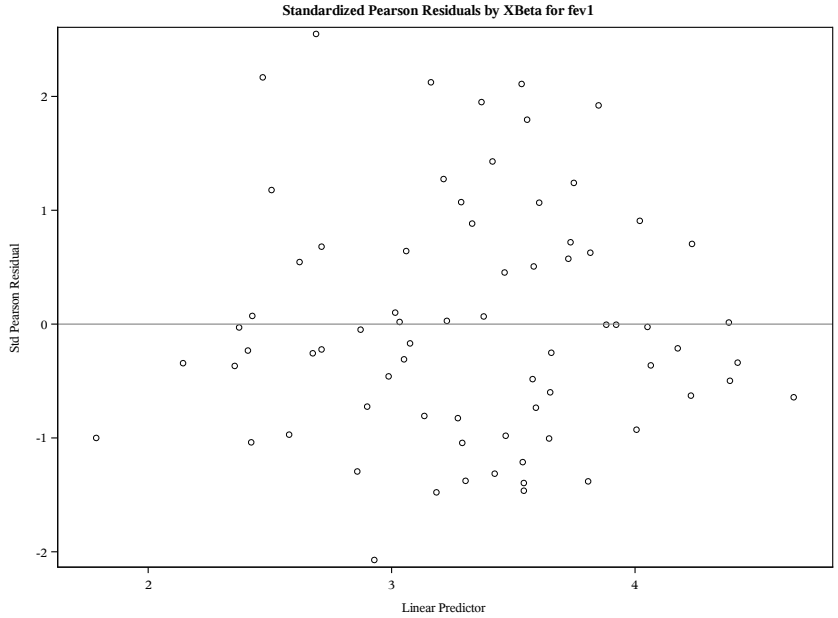

Figure A.7: Standardized Pearson Residuals Plot of Model – base+drug

A.3.3 Model – base+drug (gamma)

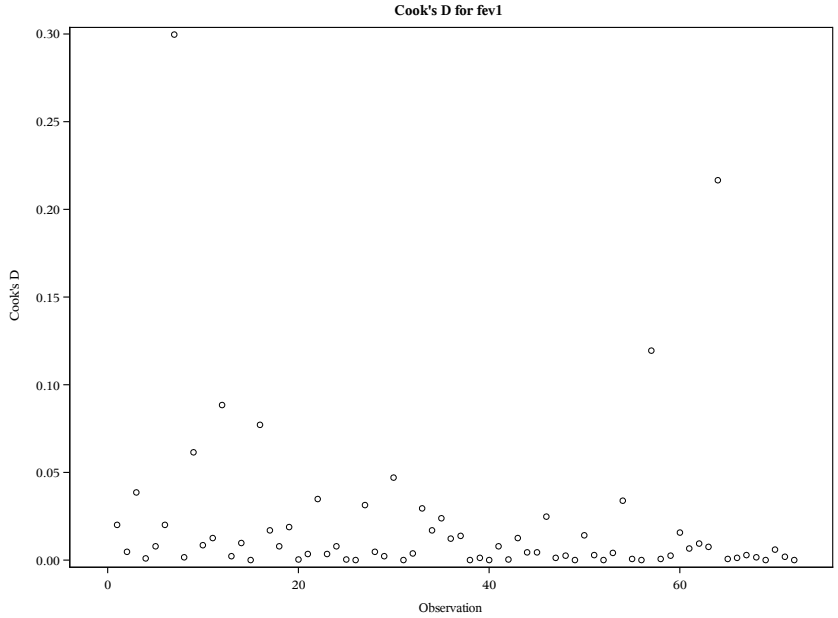

Figure A.8: Cook's Distance Plot of Model – base+drug (gamma)

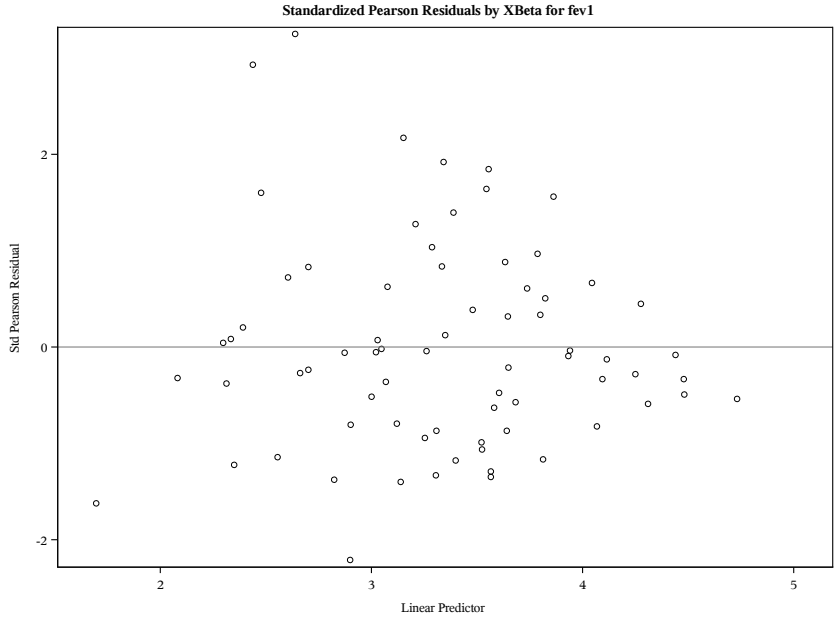

Figure A.9: Standardized Pearson Residuals Plot of Model – base+drug (gamma)

A.3.4 Model – base+drug (log link)

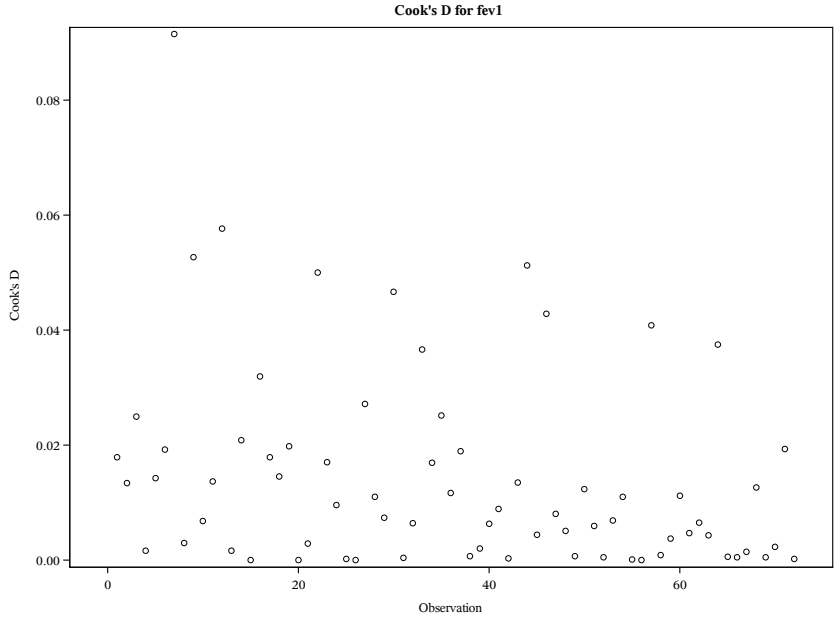

Figure A.10: Cook's Distance Plot of Model – base+drug (log link)

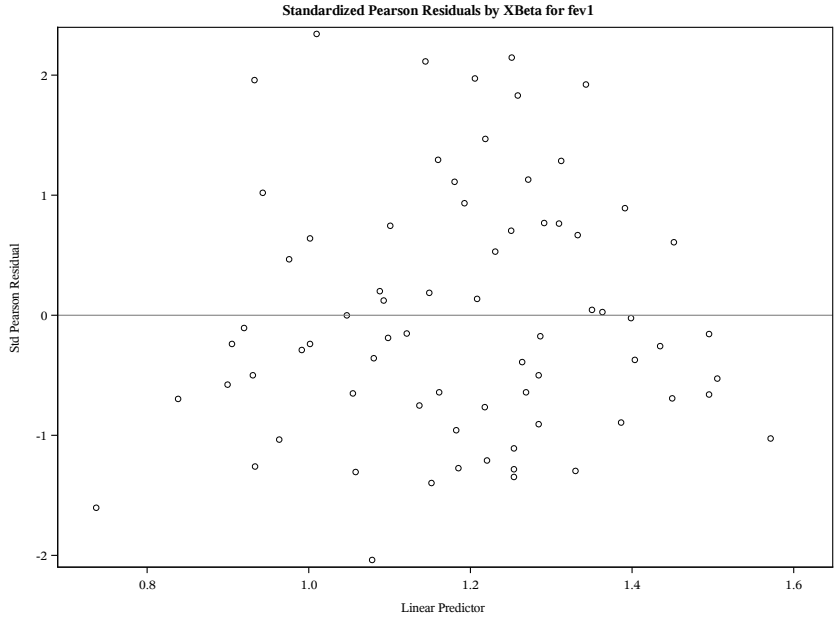

Figure A.11: Standardized Pearson Residuals Plot of Model – base+drug (log link)

A.3.5 Model – base+drug+base*drug

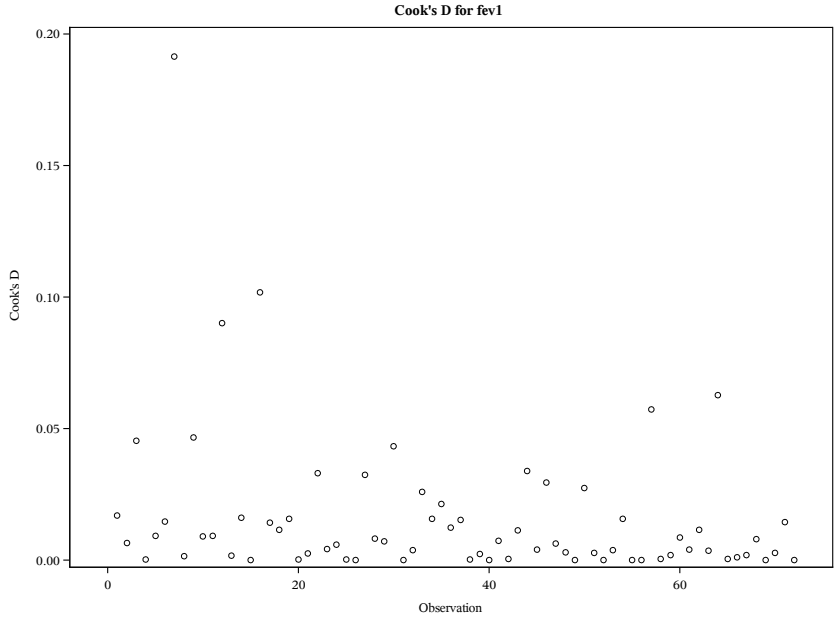

Figure A.12: Cook's Distance Plot of Model – base+drug+base*drug

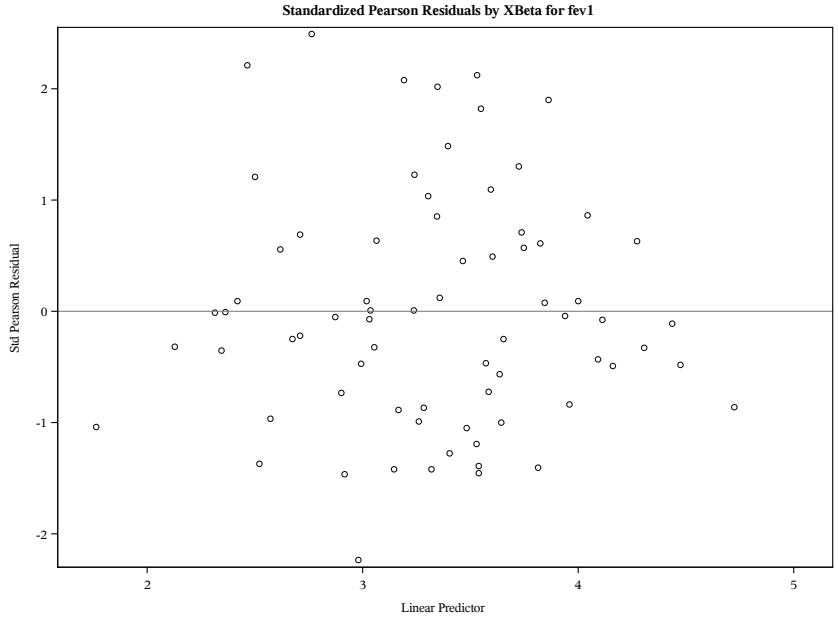

Figure A.13: Standardized Pearson Residuals Plot of Model – base+drug+base*drug

Bibliography

Agresti, A. (2015). *Foundations of linear and generalized linear models*. Wiley series in probability and statistics. John Wiley & Sons Inc, 1 edition.

College of Liberal Arts and Sciences (2000). http://users.stat.ufl.edu/~aa/glm/data/FEV.dat, Last accessed on 2019-10-08.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data*. Springer, 2nd edition.

Pedersen, J. (2019). *Survival Analysis (Supplementary Notes)*. Department of Mathematics, Aarhus University.

SAS Institute Inc. (2011a). SAS 9.3 ® GENMOD Procedure. https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_genmod_sect010.htm, Last accessed on 2019-10-08.

SAS Institute Inc. (2011b). SAS 9.3 ® GLM Procedure. https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#glm_toc.htm, Last accessed on 2019-10-08.

SAS Institute Inc. (2011c). SAS 9.3 ® Output Delivery System. https://documentation.sas.com/?docsetId=odsug&docsetTarget=titlepage.htm&docsetVersion=9.4&locale=en, Last accessed on 2019-10-08.

SAS Institute Inc. (2013a). SAS 9.4 ® DATA Step Statements: Reference. https://documentation.sas.com/?docsetId=lestmtsref&docsetTarget=titlepage.htm&docsetVersion=9.4&locale=en, Last accessed on 2019-10-08.

SAS Institute Inc. (2013b). SAS 9.4 ® SGPANEL Procedure. https://documentation.sas.com/?docsetId=grstatproc&docsetTarget=p121sy0a2jycdfn13zygo90opvra.htm&docsetVersion=9.4&locale=en, Last accessed on 2019-10-08.

SAS Institute Inc. (2013c). SAS 9.4 ® SGPLOT Procedure. https://documentation.sas.com/?docsetId=grstatproc&docsetTarget=n0yjdd910dh59zn1toodgupaj4v9.htm&docsetVersion=9.4&locale=en, Last accessed on 2019-10-08.