# Faculty of Science
# FINAL EXAMINATION
# MATH-523B
# Generalized Linear Models

Examiner: Professor K.J. Worsley
Associate Examiner: Professor A. Vandal

Date: Tuesday, April 20, 2004
Time: 14:00 - 17:00 hours

INSTRUCTIONS: Answer all questions. Any books, notes or calculators may be brought into the exam. Computer printout and tables are provided at the end of the exam.

Each part of each question is worth approximately equal marks.

This exam comprises this cover, 5 pages of questions, 12 pages of computer printout, 2 pages of figures and 2 pages of tables (22 pages in all).

1. Dr P. J. Solomon of the Australian National Centre in HIV Epidemiology and Clinical Research collected data on 2843 patients diagnosed with AIDS in Australia before 1 July 1991:

   `state`: Grouped state of origin: NSW, Other, QLD or VIC

   `sex`: Sex of patient

   `diag`: (Julian) date of diagnosis (days)

   `death`: (Julian) date of death or end of observation (days)

   `status`: "A" (alive) or "D" (dead) at end of observation

   `T.categ`: Reported transmission category (8 categories)

   `age`: Age (years) at diagnosis.

   The survival time (`time`) was assumed to have an exponential distribution with a log link to a linear model in the regressors. Choose suitable models to decide if the survival time is related to

   (a) i. state
   ii. sex
   iii. transmission category

   (b) i. age
   ii. date of diagnosis.

   Does the survival time increase or decrease with age? with date of diagnosis?

   (c) Choose a *suitable* model to estimate the mean survival time of a 25 year old male patient diagnosed with AIDS in NSW on July 1 2004 (`diag`=16253) who reported transmission by heterosexual contact (`T.categhet`), and the probability that such a patient would survive more than three years (365×3=1095 days). How reliable do you think this estimator is?

2. A breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. He assessed biopsies of breast tumors for 699 patients up to 15 July 1992; each of nine attributes has been scored on a scale of 1 to 10, and the outcome is also known: benign (`Y`=0) or malignant (`Y`=1). This data frame contains the following columns:

   `V1` Clump thickness

   `V2` Uniformity of cell size

   `V3` Uniformity of cell shape

   `V4` Marginal adhesion

   `V5` Single epithelial cell size

   `V6` Bare nuclei (16 values are missing)

`V7` Bland chromatin

`V8` Normal nucleoli

`V9` Mitoses class "benign" or "malignant"

(a) To relate the probability that a tumor is malignant to the first variable, clump thickness, two sets of models were fitted, the first assuming a normal family, the second assuming a binomial family. Choose suitable models to test for a linear effect of `V1`.

(b) A factor `fV1` was created taking the values of `V1` as levels. Use this to test if the effect of clump thickness is linear in `V1` (as opposed to non-linear).

(c) Take a look at Figures 2.1 and 2.2. Why is it that the plots of the fitted values using the model with `V1` (triangles) are different in Figures 2.1 and 2.2, yet the plots of the fitted values using the model with `fV1` (circles) are the same in Figures 2.1 and 2.2? Explain.

(d) Which attributes are related to the malignancy of breast tumors?

(e) Do you think a goodness of fit test for the last model is valid? If so, do it; if not, say why not.

3. The table below gives the frequencies (`freq`) of reported happiness (`happ`) cross-classified by years of schooling (`years`) and number of siblings (`sibs`), analysed by Clogg, C.C. (1982), *Journal of the American Statistical Association*, **77**:803-815.

| Years of school completed | Number of siblings 0-1 | 2-3 | 4-5 | 6-7 | 8+ |
|---|---|---|---|---|---|
| | Not too happy | | | | |
| <12 | 15 | 34 | 36 | 22 | 61 |
| 12 | 31 | 60 | 46 | 25 | 26 |
| 13-16 | 35 | 45 | 30 | 13 | 8 |
| 17+ | 18 | 14 | 3 | 3 | 4 |
| | Pretty happy | | | | |
| <12 | 17 | 53 | 70 | 67 | 79 |
| 12 | 60 | 96 | 45 | 40 | 31 |
| 13-16 | 63 | 74 | 39 | 24 | 7 |
| 17+ | 15 | 15 | 9 | 2 | 1 |
| | Very happy | | | | |
| <12 | 7 | 20 | 23 | 16 | 36 |
| 12 | 5 | 12 | 11 | 12 | 7 |
| 13-16 | 5 | 10 | 4 | 4 | 3 |
| 17+ | 2 | 1 | 2 | 0 | 1 |

Treating years of schooling and number of siblings as factors, choose suitable tests to decide if happiness is related to

(a) number of years of schooling,

(b) number of siblings,

(c) an interaction between the two.

(d) New variables `xyears` and `xsibs` were created, taking the same values as `years` and `sibs`. Interactions of `happ` with `years` and `sibs` were replaced by interactions of `happ` with `xyears` and `xsibs`. Explain exactly why `happ3:xyears` and `happ3:xsibs` are not estimated.

(e) Is there any evidence that the interaction of happiness with years and siblings is non-linear as opposed to linear?

(f) Based on the model in (d), explain how happiness is affected by an increase in years of schooling, or an increase in number of siblings. Who are the happiest people?

(g) Do you think a goodness of fit test for the last model is valid? If so, do it; if not, say why not.

4. In the table below, McCool (1980) gives the failure times (`time`) for hardened steel specimens in a rolling contact fatigue test; 10 independent observations were taken at each of 4 values of contact stress (`stress`:

| Stress ($\text{psi}^2 \times 10^6$) | Failure times |
|---|---|
| 0.87 | 1.67 2.20 2.51 3.00 2.90 4.70 7.53 14.70 27.8 37.4 |
| 0.99 | 0.80 1.00 1.37 2.25 2.95 3.70 6.07 6.65 7.05 7.37 |
| 1.09 | 0.012 0.18 0.20 0.24 0.26 0.32 0.32 0.42 0.44 0.88 |
| 1.18 | 0.073 0.098 0.117 0.135 0.175 0.262 0.270 0.350 0.386 0.456 |

We shall assume that `time` has a gamma distribution.

(a) A plot of log(`time`) against `stress` (Figure 4.1) suggests a log link function with a model that is linear in `stress`. Test that log(E(`time`)) is linearly related to stress.

(b) A factor `fstress` was created with a level for each different value of `stress`, and added to the model. Explain exactly why the 4th level of `fstress` is not estimated.

(c) Test that the relationship is linear in stress, as opposed to non-linear.

(d) Do you think that the data follows an exponential distribution (no formal test required)? If so, how would this affect your answer to (a) and (c)?

(e) The 21st observation $Y_{21} = 0.012$ (the smallest) at stress level 1.09 appears to be rather low in Figure 3. Estimate its mean failure time using the model which is linear in stress.

(f) Assuming that the data follows an exponential distribution, what is the estimated probability that a steel specimen subjected to a stress of 1.09 would fail before 0.012? Do you think the 21st specimen fits the model?

5. Leo Breiman, Department of Statistics, UC Berkeley, collected 45 observations of the apparent crack growth rate, obtained by dividing crack depth by rotor operating time, for disk cracks in US power plants (mostly nuclear). The variables measured were

   `loc`: crack location: 1=bore, 2=web face, 3=keyway, 4=rim attachment,

   `temp`: estimated disk temperature (degrees F),

   `stren`: 0.2% offset yield strength,

   `grow`: apparent crack growth rate.

   (a) From the graphs of `grow` against `temp` (Figure 5.1), and log(`grow`) against `temp` (Figure 5.2), give *two* reasons why it might be better to use log(`grow`) as the dependent variable in a linear model with normal errors.

   (b) Test that log(`grow`) is related simultaneously to `loc`, `temp` and `stren` using an F test.

   (c) Is log(`grow`) related to `temp` allowing for `loc` and `stren`? Is `Y` related to `stren` allowing for `loc` and `temp`?

   (d) Is the effect of `temp` the same for all locations? Is the effect of `stren` the same for all locations?

   (e) Notice that the product of the indicator variable for location 4 and temperature is not estimated, and the product of the indicator variable for location 4 and strength is not estimated. Using the plots of `temp` and `stren` against `location` (Figures 5.3, 5.4), explain exactly why this occurs.

   (f) Do you think the assumption of equal variance is satisfied?

   (g) Test that the observations have a normal distribution.

   (h) Which model, amongst all those fitted, appears to be *best* for predicting crack growth rate? Justify your choice.

6. Carl Morris (see next page) showed that there are only six families of distributions in the exponential family with quadratic variance functions: normal, poisson, gamma, binomial, negative binomial, and a sixth distribution which he called the hyperbolic secant distribution. Its variance function is $V(\mu) = \mu^2 + 1$, it is continuous on $(-\infty, \infty)$ (like the normal distribution), but it is not symmetric. The deviance parameter is $\phi > 0$. (If $m = 1/\phi$ is an integer and $\mu = 0$, then the hyperbolic secant random variable is $Y = (2/\pi) \sum_{i=1}^{m} \log |C_i|$, where $C_1, \ldots, C_m$ are independent Cauchy random variables.)

   (a) Find the canonical link. [Hint: make the substitution $\mu = \tan\theta$]. Is this a good choice for a generalized linear model?

   (b) What is the variance function of the inverse hyperbolic secant distribution?

   (c) Find an expression for the deviance as a function of the observations $Y_1, Y_2, \ldots, Y_n$ and their fitted values $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n$.

(d) Suppose we have 4 observations from this distribution with values 0.2,0.5,0.4,0.9. If the mean $\mu$ and the deviance parameter $\phi$ is the the same for each observation, find the maximum likelihood estimate of $\mu$, and any good estimate of $\phi$.

(e) We suspect that the data in (d) have a hyperbolic secant distribution with $\phi = 0.05$. Do you think a goodness of fit test for this model with $\phi = 0.05$ is valid? If so, do it (approximately); if not, say why not.

*The Annals of Statistics*
1982, Vol. 10, No. 1, 65–80

# NATURAL EXPONENTIAL FAMILIES WITH QUADRATIC VARIANCE FUNCTIONS[1]

BY CARL N. MORRIS

*University of Texas, Austin*

The normal, Poisson, gamma, binomial, and negative binomial distributions are univariate natural exponential families with quadratic variance functions (the variance is at most a quadratic function of the mean). Only one other such family exists. Much theory is unified for these six natural exponential families by appeal to their quadratic variance property, including infinite divisibility, cumulants, orthogonal polynomials, large deviations, and limits in distribution.

**1. Introduction.** The normal, Poisson, gamma, binomial, and negative binomial distributions enjoy wide application and many useful mathematical properties. What makes them so special? This paper says two things: (i) they are *natural exponential families* (*NEFs*); and (ii) they have *quadratic variance functions* (*QVF*), i.e., the variance $V(\mu)$ is, at most, a quadratic function of the mean $\mu$ for each of these distributions.

Section 2 provides background on general exponential families, making two points. First, because of some confusion about the definition of exponential families, the terms "natural exponential families" and "natural observations" are introduced here to specify those exponential families and random variables whose convolutions comprise one exponential family. Second, the "variance function" $V(\mu)$ is introduced as a quantity that characterizes the NEF.

Only six univariate, one-parameter families (and linear functions of them) are natural exponential families having a QVF. The five famous ones are listed in the initial paragraph. The sixth is derived in Section 3 as the NEF generated by the hyperbolic secant distribution. Section 4 shows this sixth family contains infinitely divisible, generally skewed, continuous distributions, with support $(-\infty, \infty)$.

```
> ########################################
> # QUESTION 1
> ########################################
> data(Aids2)
> attach(Aids2)
> time<-death-diag+1
> c<-codes(status)-1
> rate<-c/time
> summary(glm(rate~state+sex+diag+T.categ+age, family=poisson, weight=time))

Call:
glm(formula = rate ~ state + sex + diag + T.categ + age, family = poisson,
    weights = time)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-4.37597  -0.77433   0.04455   0.91472   3.42263

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.6728465  0.4716294  -7.788 6.83e-15 ***
stateOther    -0.0944785  0.0895655  -1.055  0.29149
stateQLD       0.1860238  0.0878128   2.118  0.03414 *
stateVIC      -0.0018092  0.0613208  -0.030  0.97646
sexM          -0.0369529  0.1757609  -0.210  0.83348
diag          -0.0003179  0.0000421  -7.552 4.29e-14 ***
T.categhsid   -0.1211765  0.1520374  -0.797  0.42544
T.categid     -0.3799289  0.2459986  -1.544  0.12248
T.categhet    -0.7307894  0.2652388  -2.755  0.00587 **
T.categhaem    0.3462834  0.1881367   1.841  0.06568 .
T.categblood   0.1393095  0.1374007   1.014  0.31063
T.categmother  0.4603228  0.5893405   0.781  0.43475
T.categother   0.1200160  0.1636915   0.733  0.46345
age            0.0139496  0.0024987   5.583 2.37e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4407.2  on 2842  degrees of freedom
Residual deviance: 4283.0  on 2829  degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 8

> glm(rate~state+sex+diag+age, family=poisson, weight=time)

Degrees of Freedom: 2842 Total (i.e. Null);  2836 Residual
Null Deviance:      4407
Residual Deviance: 4302         AIC: Inf
> glm(rate~sex+diag+T.categ+age, family=poisson, weight=time)

Degrees of Freedom: 2842 Total (i.e. Null);  2832 Residual
Null Deviance:      4407
Residual Deviance: 4289         AIC: Inf
> glm(rate~state+diag+T.categ+age, family=poisson, weight=time)

Degrees of Freedom: 2842 Total (i.e. Null);  2830 Residual
Null Deviance:      4407
```

```
Residual Deviance: 4283         AIC: Inf
> glm(rate~diag+T.categ+age, family=poisson, weight=time)

Degrees of Freedom: 2842 Total (i.e. Null);  2833 Residual
Null Deviance:      4407
Residual Deviance: 4289         AIC: Inf
> glm(rate~sex+diag+age, family=poisson, weight=time)

Degrees of Freedom: 2842 Total (i.e. Null);  2839 Residual
Null Deviance:      4407
Residual Deviance: 4308         AIC: Inf
There were 50 or more warnings (use warnings() to see the first 50)
> glm(rate~state+diag+age, family=poisson, weight=time)

Degrees of Freedom: 2842 Total (i.e. Null);  2837 Residual
Null Deviance:      4407
Residual Deviance: 4303         AIC: Inf
There were 50 or more warnings (use warnings() to see the first 50)
> summary(glm(rate~diag+age, family=poisson, weight=time))

Call:
glm(formula = rate ~ diag + age, family = poisson, weights = time)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-4.19449  -0.77350   0.04316   0.91967   3.40510

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.681e+00  4.352e-01  -8.457  < 2e-16 ***
diag        -3.251e-04  4.122e-05  -7.888 3.07e-15 ***
age          1.521e-02  2.411e-03   6.308 2.83e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4407.2  on 2842  degrees of freedom
Residual deviance: 4309.4  on 2840  degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 8

There were 50 or more warnings (use warnings() to see the first 50)
>
> ##########################################
> # QUESTION 2
> ##########################################
> data(biopsy)
> attach(biopsy)
> Y<-codes(class)-1
> fV1<-factor(V1)
>
> par(mfrow=c(2,2))
> glm0<-glm(Y~V1)
> summary(glm0)

Call:
glm(formula = Y ~ V1)
```

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.77804  -0.17331  -0.01994   0.06859   1.06859

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.189535   0.023395  -8.102 2.43e-15 ***
V1           0.120947   0.004467  27.078  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1104095)

    Null deviance: 157.908  on 698  degrees of freedom
Residual deviance:  76.955  on 697  degrees of freedom
AIC: 447.39

Number of Fisher Scoring iterations: 2

> glm1<-glm(Y~fV1)
> summary(glm1)

Call:
glm(formula = Y ~ fV1)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-9.565e-01  -1.111e-01  -2.069e-02   3.331e-15   9.793e-01

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02069    0.02647   0.782  0.43466
fV12         0.05931    0.05227   1.135  0.25689
fV13         0.09042    0.04051   2.232  0.02593 *
fV14         0.12931    0.04439   2.913  0.00369 **
fV15         0.32546    0.03850   8.455  < 2e-16 ***
fV16         0.50872    0.06073   8.377 3.04e-16 ***
fV17         0.93583    0.07153  13.083  < 2e-16 ***
fV18         0.89235    0.05393  16.546  < 2e-16 ***
fV19         0.97931    0.08920  10.979  < 2e-16 ***
fV110        0.97931    0.04661  21.010  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1015776)

    Null deviance: 157.908  on 698  degrees of freedom
Residual deviance:  69.987  on 689  degrees of freedom
AIC: 397.04

Number of Fisher Scoring iterations: 2

> plot(V1,fitted(glm1))
> points(V1,fitted(glm0),pch=2)
> title('Figure 2.1: Normal family')
>
> glm0<-glm(Y~V1,family=binomial)
> summary(glm0)
```

```
Call:
glm(formula = Y ~ V1, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1986  -0.4261  -0.1704   0.1730   2.9118

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.16017    0.37772  -13.66   <2e-16 ***
V1           0.93546    0.07372   12.69   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 900.53  on 698  degrees of freedom
Residual deviance: 464.05  on 697  degrees of freedom
AIC: 468.05

Number of Fisher Scoring iterations: 5

> glm1<-glm(Y~fV1, family=binomial)
> summary(glm1)

Call:
glm(formula = Y ~ fV1, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.50419  -0.48535  -0.20448   0.01184   2.78500

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.8572     0.5834  -6.611 3.81e-11 ***
fV12          1.4149     0.7824   1.808  0.07054 .
fV13          1.7778     0.6589   2.698  0.00697 **
fV14          2.1226     0.6621   3.206  0.00135 **
fV15          3.2212     0.6119   5.265 1.40e-07 ***
fV16          3.9750     0.6771   5.871 4.34e-09 ***
fV17          6.9483     1.1772   5.902 3.58e-09 ***
fV18          6.2086     0.7837   7.922 2.33e-15 ***
fV19         13.4232    19.3753   0.693  0.48844
fV110        13.4232     8.7430   1.535  0.12471
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 900.53  on 698  degrees of freedom
Residual deviance: 450.21  on 689  degrees of freedom
AIC: 470.21

Number of Fisher Scoring iterations: 8

> plot(V1,fitted(glm1))
> points(V1,fitted(glm0),pch=2)
> title('Figure 2.2: Binomial family')
```

```
>
> summary(glm(Y~V1+V2+V3+V4+V5+V6+V7+V8+V9, family=binomial))

Call:
glm(formula = Y ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9,
    family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.48404  -0.11529  -0.06192   0.02221   2.46983

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.103859   1.170793  -8.630  < 2e-16 ***
V1            0.535008   0.141838   3.772 0.000162 ***
V2           -0.006278   0.208786  -0.030 0.976011
V3            0.322705   0.230224   1.402 0.161005
V4            0.330634   0.123318   2.681 0.007337 **
V5            0.096634   0.156467   0.618 0.536836
V6            0.383024   0.093741   4.086 4.39e-05 ***
V7            0.447184   0.171156   2.613 0.008982 **
V8            0.213030   0.112757   1.889 0.058855 .
V9            0.534817   0.328105   1.630 0.103098
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 884.35  on 682  degrees of freedom
Residual deviance: 102.89  on 673  degrees of freedom
AIC: 122.89

Number of Fisher Scoring iterations: 7

> summary(glm(Y~V1+V4+V6+V7, family=binomial))

Call:
glm(formula = Y ~ V1 + V4 + V6 + V7, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.69637  -0.14510  -0.06093   0.02317   2.44758

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.11370    1.03190  -9.801  < 2e-16 ***
V1            0.81166    0.12579   6.453 1.10e-10 ***
V4            0.43412    0.11399   3.808  0.00014 ***
V6            0.48136    0.08813   5.462 4.72e-08 ***
V7            0.70154    0.15190   4.619 3.87e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 884.35  on 682  degrees of freedom
Residual deviance: 125.77  on 678  degrees of freedom
AIC: 135.77
```

```
Number of Fisher Scoring iterations: 7

>
> ########################################
> # QUESTION 3
> ########################################
>
> freq<-c(scan("C:/keith/teaching/datasets/happy"))
Read 60 items
> (t(matrix(freq,5,12)))
      [,1] [,2] [,3] [,4] [,5]
 [1,]   15   34   36   22   61
 [2,]   31   60   46   25   26
 [3,]   35   45   30   13    8
 [4,]   18   14    3    3    4
 [5,]   17   53   70   67   79
 [6,]   60   96   45   40   31
 [7,]   63   74   39   24    7
 [8,]   15   15    9    2    1
 [9,]    7   20   23   16   36
[10,]    5   12   11   12    7
[11,]    5   10    4    4    3
[12,]    2    1    2    0    1
> happ<-gl(3,20,60)
> years<-gl(4,5,60)
> sibs<-gl(5,1,60)
> glm(freq~happ+years+sibs, family=poisson)

Degrees of Freedom: 59 Total (i.e. Null);  50 Residual
Null Deviance:      1264
Residual Deviance: 323.7        AIC: 612.2
> glm(freq~happ+years+sibs+happ:years, family=poisson)

Degrees of Freedom: 59 Total (i.e. Null);  44 Residual
Null Deviance:      1264
Residual Deviance: 283.6        AIC: 584.2
> glm(freq~happ+years+sibs+happ:sibs, family=poisson)

Degrees of Freedom: 59 Total (i.e. Null);  42 Residual
Null Deviance:      1264
Residual Deviance: 297.4        AIC: 602
> glm(freq~happ+years+sibs+years:sibs, family=poisson)

Degrees of Freedom: 59 Total (i.e. Null);  38 Residual
Null Deviance:      1264
Residual Deviance: 79.68        AIC: 392.2
> glm(freq~happ+years+sibs+happ:years+happ:sibs, family=poisson)

Degrees of Freedom: 59 Total (i.e. Null);  36 Residual
Null Deviance:      1264
Residual Deviance: 257.3        AIC: 573.9
> glm(freq~happ+years+sibs+happ:years+years:sibs, family=poisson)

Degrees of Freedom: 59 Total (i.e. Null);  32 Residual
Null Deviance:      1264
Residual Deviance: 39.62        AIC: 364.2
> glm(freq~happ+years+sibs+happ:sibs+years:sibs, family=poisson)

Degrees of Freedom: 59 Total (i.e. Null);  30 Residual
```

```
Null Deviance:      1264
Residual Deviance: 53.41        AIC: 382
> glm(freq~happ+years+sibs+happ:years+happ:sibs+years:sibs, family=poisson)

Degrees of Freedom: 59 Total (i.e. Null);  24 Residual
Null Deviance:      1264
Residual Deviance: 24.88        AIC: 365.4
> xyears<-codes(years)
> xsibs<-codes(sibs)
> summary(glm(freq~happ+years+sibs+happ:xyears+happ:xsibs+years:sibs, family=poisson))

Call:
glm(formula = freq ~ happ + years + sibs + happ:xyears + happ:xsibs +
    years:sibs, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9874  -0.6371   0.1328   0.6043   1.8400

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.03256    0.26800   7.584 3.35e-14 ***
happ2         0.85653    0.22149   3.867 0.000110 ***
happ3        -0.42920    0.34943  -1.228 0.219347
years2        0.51165    0.21084   2.427 0.015235 *
years3        0.17421    0.26770   0.651 0.515201
years4       -1.32736    0.37480  -3.542 0.000398 ***
sibs2         1.13427    0.19545   5.804 6.49e-09 ***
sibs3         1.44297    0.21393   6.745 1.53e-11 ***
sibs4         1.35530    0.24872   5.449 5.06e-08 ***
sibs5         1.98625    0.27677   7.177 7.15e-13 ***
happ1:xyears  0.51831    0.11236   4.613 3.97e-06 ***
happ2:xyears  0.38959    0.10849   3.591 0.000329 ***
happ1:xsibs  -0.10484    0.06852  -1.530 0.126041
happ2:xsibs  -0.16570    0.06530  -2.538 0.011158 *
years2:sibs2 -0.44499    0.22664  -1.963 0.049591 *
years3:sibs2 -0.77687    0.22907  -3.391 0.000695 ***
years4:sibs2 -1.15496    0.31135  -3.709 0.000208 ***
years2:sibs3 -1.12560    0.23162  -4.860 1.18e-06 ***
years3:sibs3 -1.52466    0.23857  -6.391 1.65e-10 ***
years4:sibs3 -2.09401    0.36553  -5.729 1.01e-08 ***
years2:sibs4 -1.19480    0.24221  -4.933 8.10e-07 ***
years3:sibs4 -1.88580    0.26376  -7.150 8.70e-13 ***
years4:sibs4 -2.90593    0.51412  -5.652 1.58e-08 ***
years2:sibs5 -1.88934    0.23993  -7.875 3.42e-15 ***
years3:sibs5 -3.21421    0.31181 -10.308  < 2e-16 ***
years4:sibs5 -3.22639    0.47720  -6.761 1.37e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1264.198  on 59  degrees of freedom
Residual deviance:   38.855  on 34  degrees of freedom
AIC: 359.42

Number of Fisher Scoring iterations: 4

> glm(freq~happ+years+sibs+happ:xsibs+years:sibs, family=poisson)
```

```
Degrees of Freedom: 59 Total (i.e. Null);  36 Residual
Null Deviance:      1264
Residual Deviance: 61.78        AIC: 378.3
> glm(freq~happ+years+sibs+happ:xyears+years:sibs, family=poisson)

Degrees of Freedom: 59 Total (i.e. Null);  36 Residual
Null Deviance:      1264
Residual Deviance: 45.82        AIC: 362.4
> glm(freq~happ+years+sibs+years:sibs, family=poisson)

Degrees of Freedom: 59 Total (i.e. Null);  38 Residual
Null Deviance:      1264
Residual Deviance: 79.68        AIC: 392.2
>
> ##########################################
> # QUESTION 4
> ##########################################
>
> m<-t(matrix(scan("C:/keith/teaching/datasets/steel"),2,40))
Read 80 items
> stress<-m[,1]
> time<-m[,2]
> ltime<-log(time)
> plot(stress,ltime)
> title('Figure 4.1: Time vs. stress')
> summary(glm(time~stress, family=Gamma(link="log")))

Call:
glm(formula = time ~ stress, family = Gamma(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4658  -0.8694  -0.4579   0.3320   1.3122

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.187      1.250   11.35 8.96e-14 ***
stress       -13.383      1.203  -11.13 1.62e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.7707791)

    Null deviance: 110.033  on 39  degrees of freedom
Residual deviance:  34.226  on 38  degrees of freedom
AIC: 114.15

Number of Fisher Scoring iterations: 5

> fstress<-factor(stress)
> summary(glm(time~stress+fstress, family=Gamma(link="log")))

Call:
glm(formula = time ~ stress + fstress, family = Gamma(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1644  -0.7898  -0.2709   0.3822   1.6162
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.0268     1.2303  10.588 1.32e-12 ***
stress      -12.2771     1.1868 -10.345 2.49e-12 ***
fstress0.99   0.4939     0.3213   1.537   0.1330
fstress1.09  -0.7620     0.3278  -2.324   0.0259 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.6767753)

    Null deviance: 110.033  on 39  degrees of freedom
Residual deviance:  27.407  on 36  degrees of freedom
AIC: 108.18

Number of Fisher Scoring iterations: 5

>
> ##########################################
> # QUESTION 5
> ##########################################
>
> m<-t(matrix(scan("C:/keith/teaching/datasets/turbines"),4,45))
Read 180 items
> loc<-factor(m[,1])
> temp<-m[,2]
> stren<-m[,3]
> grow<-m[,4]
> plot(temp,grow)
> title('Figure 5.1: grow vs. temp')
>
> lgrow<-log(grow)
> plot(temp,lgrow)
> title('Figure 5.2: log(grow) vs. temp')
> summary(glm(lgrow~loc+temp+stren))

Call:
glm(formula = lgrow ~ loc + temp + stren)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8264  -0.2920   0.1705   0.5418   1.3028

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.819121   2.547952  -5.424 3.27e-06 ***
loc2          0.521480   0.783089   0.666   0.5094
loc3          0.381040   0.651127   0.585   0.5618
loc4          1.390096   0.696012   1.997   0.0528 .
temp          0.023630   0.003522   6.710 5.38e-08 ***
stren         0.051197   0.011367   4.504 5.90e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.6502627)

    Null deviance: 59.893  on 44  degrees of freedom
Residual deviance: 25.360  on 39  degrees of freedom
```

```
AIC: 115.90

Number of Fisher Scoring iterations: 2

> summary(glm(lgrow~loc+temp+stren+loc:temp))

Call:
glm(formula = lgrow ~ loc + temp + stren + loc:temp)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4694  -0.4236   0.1705   0.4913   1.2027

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -22.350664   3.064030  -7.295 1.15e-08 ***
loc2         11.641169   3.160552   3.683 0.000732 ***
loc3         12.113633   3.002635   4.034 0.000264 ***
loc4          3.475396   0.796278   4.365 9.84e-05 ***
temp          0.051220   0.007606   6.734 6.43e-08 ***
stren         0.040536   0.010363   3.912 0.000378 ***
loc2:temp    -0.031396   0.009395  -3.342 0.001913 **
loc3:temp    -0.033771   0.008503  -3.972 0.000317 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.4799778)

    Null deviance: 59.893  on 44  degrees of freedom
Residual deviance: 17.759  on 37  degrees of freedom
AIC: 103.87

Number of Fisher Scoring iterations: 2

> summary(glm(lgrow~loc+temp+stren+loc:temp+loc:stren))

Call:
glm(formula = lgrow ~ loc + temp + stren + loc:temp + loc:stren)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-1.404046  -0.407984   0.002867   0.481278   1.276484

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.274658   7.493320  -1.371 0.179048
loc2          4.507530   9.367458   0.481 0.633376
loc3         -1.477930   7.817169  -0.189 0.851136
loc4          2.860366   0.847970   3.373 0.001827 **
temp          0.054244   0.007575   7.161 2.37e-08 ***
stren        -0.057728   0.056870  -1.015 0.317030
loc2:temp    -0.040405   0.011031  -3.663 0.000818 ***
loc3:temp    -0.035371   0.008403  -4.209 0.000170 ***
loc2:stren    0.076142   0.062015   1.228 0.227719
loc3:stren    0.106627   0.057964   1.840 0.074332 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.4514024)
```

```
    Null deviance: 59.893  on 44  degrees of freedom
Residual deviance: 15.799  on 35  degrees of freedom
AIC: 102.60

Number of Fisher Scoring iterations: 2

> plot(codes(loc),temp)
> title('Figure 5.3: temp vs. loc')
> plot(codes(loc),stren)
> title('Figure 5.4: stren vs. loc')
> glm0<-glm(lgrow~loc+temp+stren+loc:temp)
> plot(fitted(glm0),resid(glm0))
> title('Figure 5.5: fitted vs. resid')
>
> vl<-predict.glm(glm0,se.fit=T)$se.fit^2
> sc<-summary(glm0)$dispersion
> r<-resid(glm0)
> z<-r/sqrt(sc-vl)
> df<-glm0$df.residual
> tstat<-z/sqrt((df-z^2)/(df-1))
> cbind(r,z,tstat)
               r             z         tstat
1   4.532054e-01  7.294899e-01  7.247955e-01
2   6.791526e-01  1.073304e+00  1.075576e+00
3   4.913005e-01  7.764302e-01  7.721825e-01
4   2.480867e-01  3.993261e-01  3.947444e-01
5   1.705180e-01  3.480765e-01  3.439041e-01
6  -1.705180e-01 -3.480765e-01 -3.439041e-01
7  -7.549517e-15 -5.850169e-07 -5.770571e-07
8   2.210938e-01  3.552592e-01  3.510247e-01
9  -2.588842e-01 -4.159819e-01 -4.112849e-01
10 -3.125465e-01 -5.022078e-01 -4.970718e-01
11 -6.601056e-01 -1.060675e+00 -1.062522e+00
12  5.342853e-01  7.977819e-01  7.937840e-01
13 -6.253082e-02 -9.336951e-02 -9.210997e-02
14 -3.422210e-01 -5.158707e-01 -5.106916e-01
15 -9.589952e-01 -1.445608e+00 -1.467999e+00
16 -1.266480e+00 -1.909117e+00 -1.983360e+00
17 -1.469444e+00 -2.194139e+00 -2.320511e+00
18  6.332496e-01  9.410566e-01  9.395648e-01
19  5.462382e-01  8.117512e-01  8.079331e-01
20  4.238994e-01  6.299464e-01  6.247346e-01
21 -4.653627e-01 -6.915640e-01 -6.866065e-01
22 -4.653627e-01 -6.915640e-01 -6.866065e-01
23  1.215227e-01  1.821543e-01  1.797565e-01
24 -1.320416e+00 -1.976909e+00 -2.061948e+00
25  1.202692e+00  1.793519e+00  1.851426e+00
26  7.895837e-01  1.177470e+00  1.183841e+00
27  6.556868e-01  9.777957e-01  9.771998e-01
28  4.976299e-01  7.420926e-01  7.375047e-01
29  4.520879e-01  6.741779e-01  6.691275e-01
30 -7.925746e-02 -1.181930e-01 -1.166069e-01
31  6.957840e-01  1.106666e+00  1.110136e+00
32 -4.236339e-01 -6.780891e-01 -6.730581e-01
33 -6.467774e-01 -1.035264e+00 -1.036297e+00
34 -9.344595e-01 -1.495741e+00 -1.522126e+00
35  4.121176e-01  6.310234e-01  6.258142e-01
36  2.384064e-01  3.823720e-01  3.779168e-01
```

```
37  2.384064e-01  3.823720e-01  3.779168e-01
38  3.661216e-01  5.744468e-01  5.691746e-01
39 -1.567381e-02 -2.459230e-02 -2.425789e-02
40 -8.602059e-01 -1.379655e+00 -1.397299e+00
41  6.418068e-01  1.160119e+00  1.165733e+00
42 -8.297673e-02 -1.484039e-01 -1.464283e-01
43  7.423709e-01  1.264778e+00  1.275446e+00
44  2.464709e-01  4.199128e-01  4.151900e-01
45 -9.058651e-01 -1.511841e+00 -1.539582e+00
> z[7]=0
> u<-pnorm(z)
> i<-1:45
> uhat<-(i-0.5)/45
> u<-sort(u)
> u-uhat
          17           24           16           45           34           15
 0.003001594 -0.009307393 -0.027432043 -0.012490633 -0.032639479 -0.048078676
          40           11           33           21           22           32
-0.060597980 -0.022247813 -0.038615955  0.033494477  0.011272254 -0.006697940
          14           10            9            6           42           30
 0.025194570  0.007760666  0.016489405  0.019446913  0.074345334  0.064068477
          13           39            7           23            5            8
 0.051693894  0.056856749  0.044444444  0.094491405  0.136108642  0.116580081
          36           37            4           44           38           20
 0.104462809  0.082240587  0.066284652  0.051614327  0.083833918  0.080079623
          35           29            1           28            3           12
 0.058209603  0.049900890  0.044926759  0.026539945  0.014585836 -0.001387428
          19           18           27            2           31           41
-0.019578325 -0.006671253 -0.019643254 -0.019345243 -0.034219119 -0.045222319
          26           43           25
-0.063948459 -0.069642080 -0.025333837
```

**Figure 2.1: Normal family**

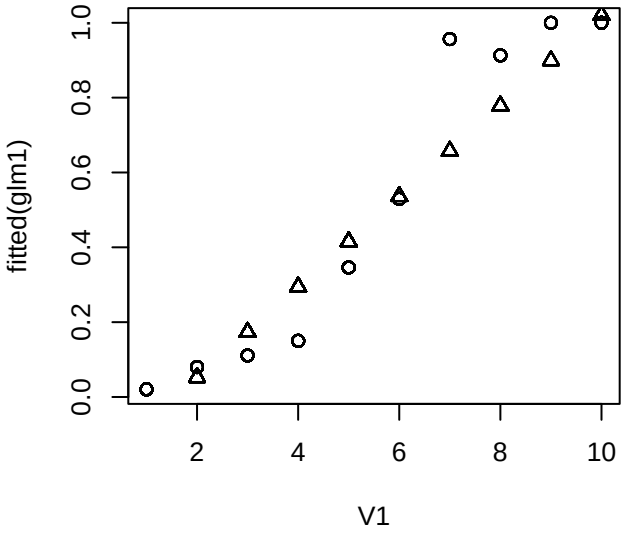


**Figure 2.2: Binomial family**

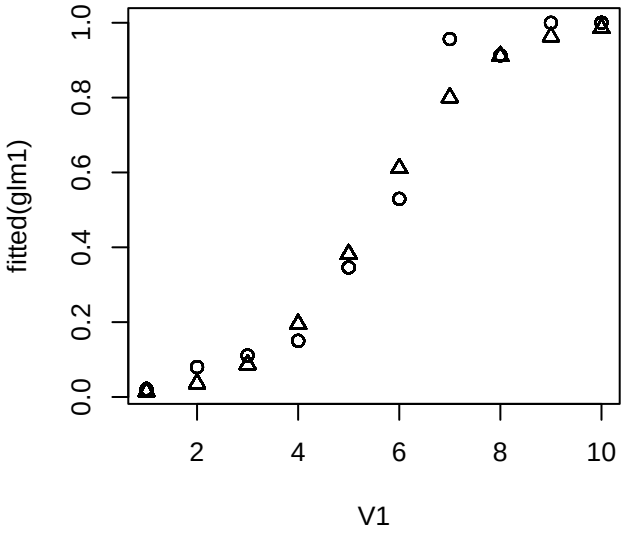


**Figure 4.1: Time vs. stress**

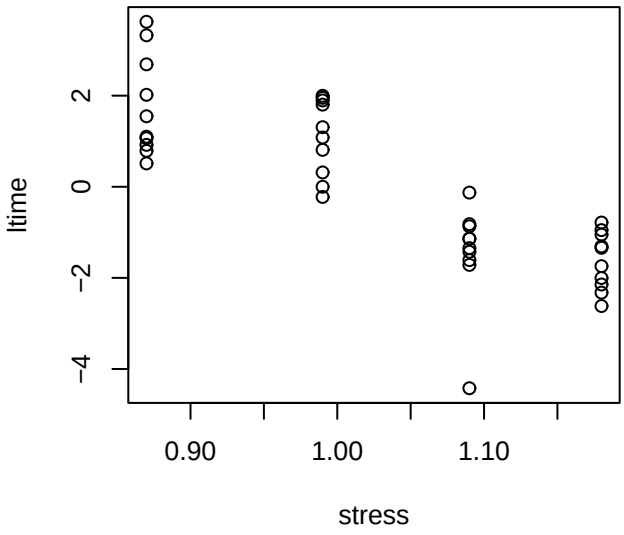


**Figure 5.1: grow vs. temp**

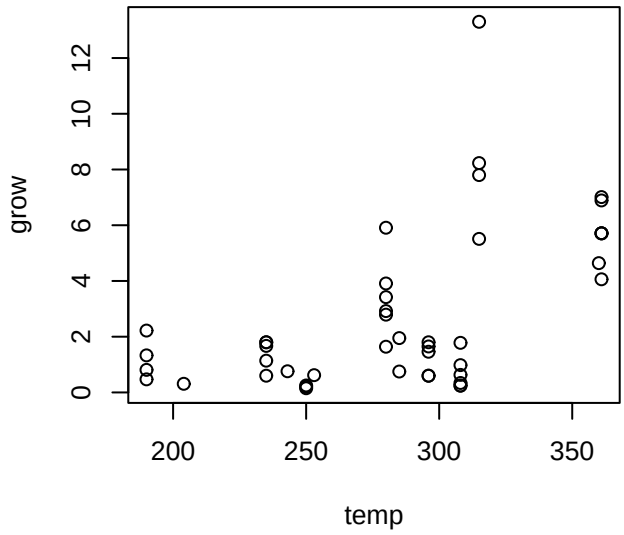

**Figure 5.2: log(grow) vs. temp**

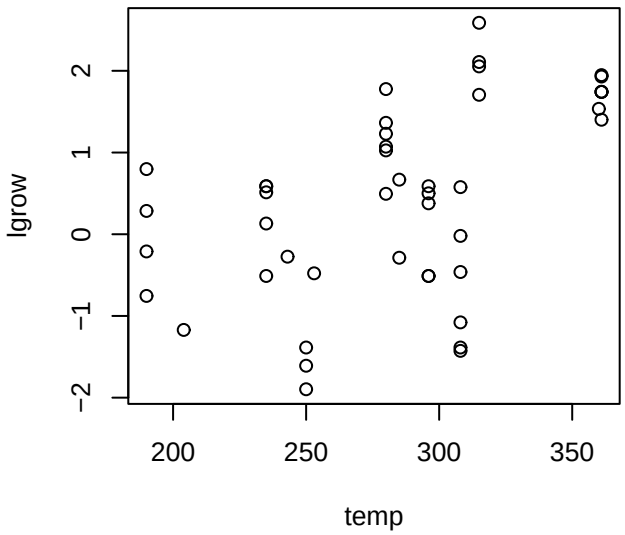


**Figure 5.3: temp vs. loc**

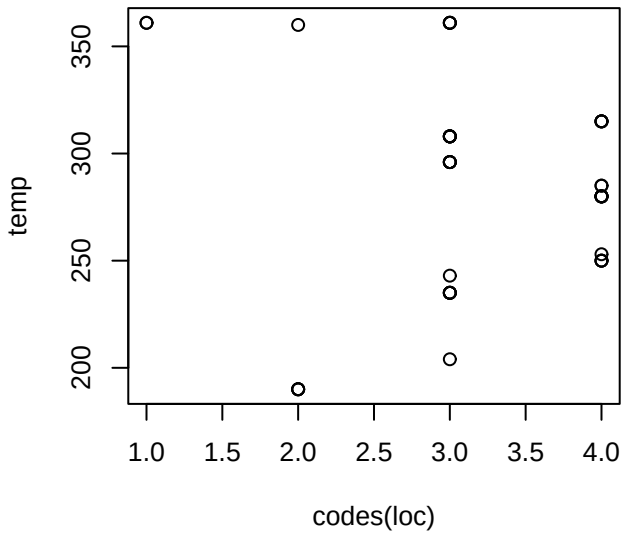


**Figure 5.4: stren vs. loc**

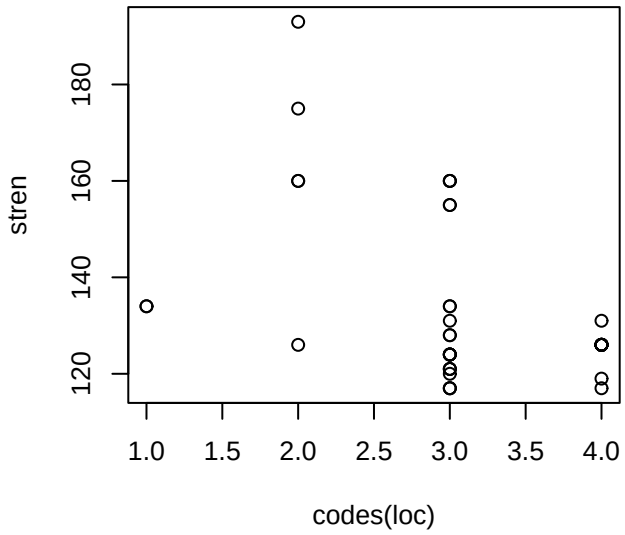


**Figure 5.5: fitted vs. resid**