

STAT 511 Final Exam Spring 2012

1. A total of 20 sows (mother pigs) were raised under identical conditions. The 20 sows consisted of 10 sows of breed *A* and 10 sows of breed *B*. Each sow was impregnated, and the litter size (here defined as the number of piglets born alive) was recorded for each sow. Scientists are interested in determining if one breed tends to produce larger litters than the other. Use the following R code and output to complete this problem.

```
> d
  Breed LitterSize
1     A           9
2     A           4
.     .           .
.     .           .
.     .           .
10    A           4
11    B           8
12    B           5
.     .           .
.     .           .
.     .           .
20    B           7

> o=glm(LitterSize~Breed,data=d,family=poisson(link=log))
> summary(o)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7047     0.1348  12.643  <2e-16 ***
BreedB         0.3747     0.1752   2.139   0.0324 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 41.681  on 19  degrees of freedom
Residual deviance: 37.025  on 18  degrees of freedom
```

- (a) Fill in the blank in the following sentence. (Do not write your answer on this sheet. Just write what goes in the blank on your answer sheet.) (4 pts.)

Based on the model fit to these data, the mean litter size for breed *B* is estimated to be _____ times the mean litter size for breed *A*.

- (b) Based on the model fit to these data, provide the value of the likelihood ratio test statistic ($-2 \log \Lambda$) used for testing the null hypothesis that the mean litter size for breed A is the same as the mean litter size for breed B against the alternative hypothesis that each breed has a different mean litter size. (4 pts.)
- (c) Suppose we were to conduct a quasi-likelihood analysis with a dispersion parameter ϕ . Determine the value of an estimator of ϕ . (4 pts.)
- (d) Now assume that there is significant evidence of overdispersion. Again consider testing the null hypothesis that the mean litter size for breed A is the same as the mean litter size for breed B against the alternative hypothesis that each breed has a different mean litter size. Compute a test statistic used in a quasi-likelihood approach that is approximately t -distributed under the null hypothesis. (4 pts.)
- (e) State the degrees of freedom associated with the statistic provided in part (d). (4 pts.)
2. Consider an experiment designed to compare two treatments (oatmeal breakfast vs. egg breakfast) on cholesterol levels in adult humans. A total of 200 subjects (100 male and 100 female) participated in the experiment. From these 200 subjects, 50 males and 50 females were randomly assigned to eat oatmeal for breakfast every day for one year. The other 50 males and the other 50 females were assigned to eat an egg-based breakfast every day for one year. A 6-dimensional response vector was measured for each subject. For $k = 1, \dots, 6$, the k th component of the response vector contained a subject's LDL cholesterol level $2k$ months after the beginning of the experiment.

Two general linear models were fit to the data. Each model has the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Both models used the same mean structure for the data. The design matrix \mathbf{X} (which has 1200 rows) is given by the R command

```
model.matrix(~LDL0+treatment*sex*time),
```

where `LDL0` is a quantitative variable that contains the LDL cholesterol of each subject at the beginning of the experiment and `treatment`, `sex`, and `time` are categorical factors with 2, 2, and 6 levels, respectively.

In both models, $\boldsymbol{\Sigma}$ is assumed to be block diagonal with one 6-by-6 block for each subject that is the same for all subjects. Under Model 1, each block is assumed to have an AR(1) structure with unknown parameters. Under Model 2, each block is assumed to be an unknown and unstructured positive definite variance matrix. Suppose the maximized log likelihood for each model is given in the table below.

Model	Variance Structure	Maximized Log Likelihood
1	AR(1)	-700
2	Unstructured	-675

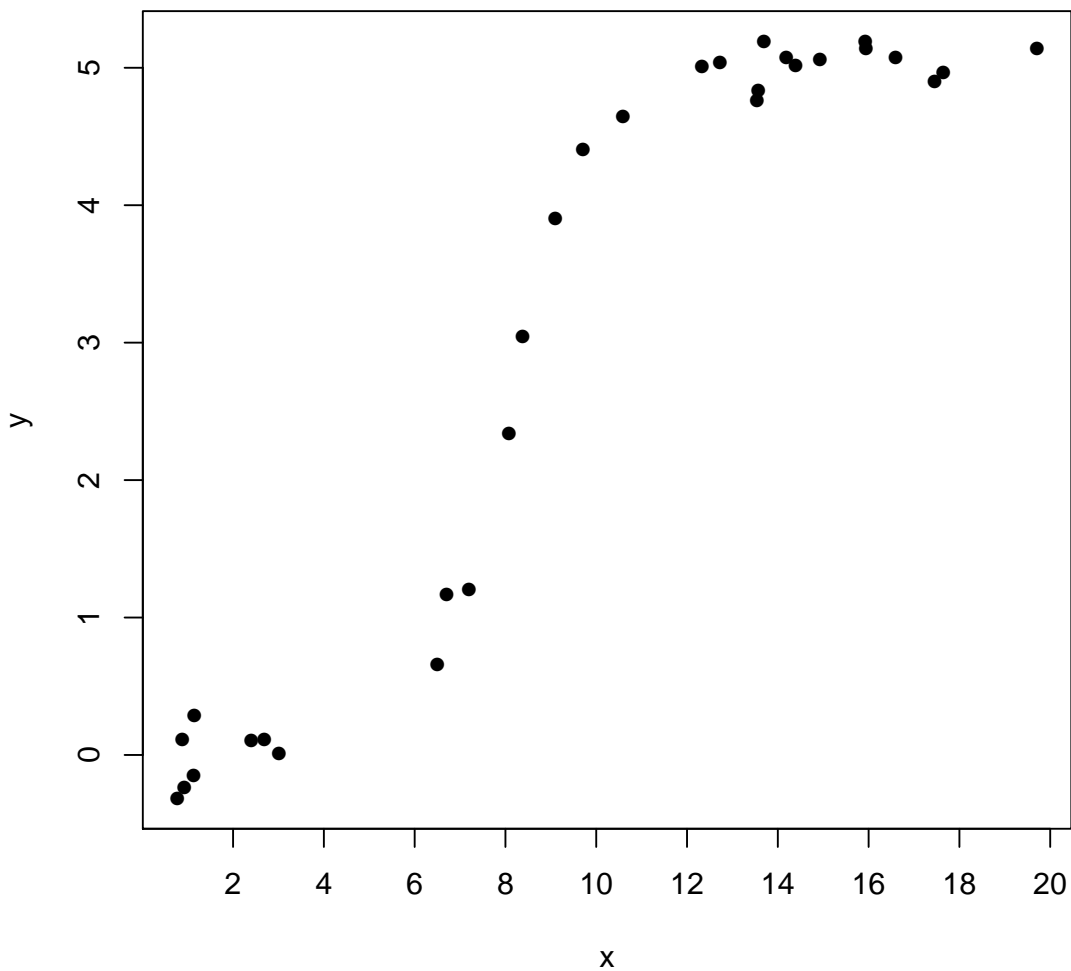
- (a) Compute AIC for each model. (8 pts.)

(b) Which model is preferred according to the AIC criterion? (4 pts.)

3. Suppose for $i = 1, \dots, 30$,

$$y_i = \frac{\beta_1 \exp(x_i - \beta_2)}{1 + \exp(x_i - \beta_2)} + \varepsilon_i,$$

where y_i is the value of the response variable for experimental unit i , x_i is the value of the explanatory variable for experimental unit i , β_1 and β_2 are unknown parameters in \mathbb{R} , and $\varepsilon_1, \dots, \varepsilon_{30} \stackrel{iid}{\sim} N(0, \sigma^2)$ for some unknown $\sigma^2 > 0$. Based on the figure below, provide good starting values for β_1 and β_2 that can be used with the Gauss-Newton Algorithm to find least squares estimates of β_1 and β_2 . Briefly explain the reasoning that you used to obtain your starting values. (8 pts.)



4. Consider a two-factor experiment with a completely randomized design. Suppose factor A has 2 levels ($A1, A2$) and factor B has 3 levels ($B1, B2, B3$). Suppose the experiment is not balanced and the sample sizes corresponding to the treatments are as provided in the table below.

	$B1$	$B2$	$B3$
$A1$	2	4	2
$A2$	3	1	4

Suppose that the values of the response variable are stored in R in a vector y . Suppose A and B are factors in R that correspond to the factors A and B , respectively. Consider the following code and output.

```
> o=lm(y~A+B+A:B)
> summary(o)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	11.000	1.871	5.880	0.000155	***
A2	-9.000	2.415	-3.726	0.003934	**
B2	-5.000	2.291	-2.182	0.054046	.
B3	10.000	2.646	3.780	0.003604	**
A2:B2	12.000	3.819	3.142	0.010468	*
A2:B3	3.000	3.329	0.901	0.388708	

- (a) Determine the lsmean for level $A2$ of factor A . (8 pts.)
- (b) Provide the standard error corresponding to the lsmean for level $A2$ of factor A . (7 pts.)
5. Suppose a response variable y is observed once at each of the explanatory variable values $x = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19$, and 21 . To describe the relationship between y and x , suppose a researcher would like to use ordinary least squares to fit a continuous piecewise linear function with potential slope changes at $x = 6$ and $x = 14$.
- (a) Provide a design matrix \mathbf{X} so that $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ will provide ordinary least squares estimates of parameters needed to construct the piecewise linear function. Here \mathbf{y} is the vector of responses. You are not given that vector, so you cannot calculate $\hat{\beta}$. You are simply being asked to provide an appropriate design matrix \mathbf{X} that could in principle be used to calculate $\hat{\beta}$. (5 pts.)
- (b) Let $\hat{\beta}_1, \dots, \hat{\beta}_p$ denote the components of $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ from part (a). In terms of $\hat{\beta}_1, \dots, \hat{\beta}_p$, write the estimated slope of each piece the best fitting (based on the ordinary least squares criterion) continuous piecewise linear function. (Your answer should give the slope as a function of $\hat{\beta}_1, \dots, \hat{\beta}_p$ for $x < 6$, $x \in (6, 14)$, and $x > 14$.) (5 pts.)

(c) Is the following statement true or false? Please briefly explain your answer. (5 pts.)

The penalized spline approach that we discussed in class will tend to make the estimated slope of each piece of the piecewise linear function closer to zero than the unpenalized (ordinary) least squares estimates.

6. Let w_1 denote the weight (in kg) gained by a Holstein calf from birth to one week of age. Let w_2 denote the weight (in kg) gained by a Holstein calf from one week of age to 12 weeks of age. Suppose it is known that

$$\text{Var}(w_1) = \text{Var}(w_2) = 4$$

and that the correlation between w_1 and w_2 when measured for the same calf is 0.5. Suppose weight gained by any one calf is independent of the weight gained by any other calf. Suppose the following information is available for three randomly selected calves. (Note that the periods in the table denote data that is missing completely at random.)

Calf	w_1	w_2
1	51	54
2	48	.
3	52	.

Determine the best linear unbiased estimator of the expected total weight of a Holstein calf at age 12 weeks; i.e., find the BLUE of $E(w_1 + w_2)$ by carrying out the following steps. (15 pts.)

- (a) Assign specific values to vectors and matrices that you are free to define as you wish.
- (b) Write down an expression for the BLUE that is a function of the vectors and matrices defined in part (a).

Although it is possible to compute a number for the BLUE with paper and pencil in this case, that is not necessary. All you need to do is write down an expression that can be easily computed with R by creating the vectors and matrices that you defined in part (a) and evaluating the expression that you defined in part (b). You do not need to carry out any matrix inversions by hand.

7. Suppose that $\mathbf{y} = \mathbf{1}\mu + \boldsymbol{\epsilon}$, where $\text{Var}(\boldsymbol{\epsilon}) = \sigma_1^2\mathbf{I} + \sigma_2^2\mathbf{1}\mathbf{1}'$ for some unknown $\sigma_1^2, \sigma_2^2 > 0$. Without inverting any matrices, prove that the BLUE of μ is \bar{y} , the average of the n entries in the \mathbf{y} vector. One way to do this is to follow the same strategy we used to prove the Gauss Markov Theorem in our course notes. (15 pts.)