Stat 511 Homework 1 Spring 2005

Maximum score is 20 points, due date is Friday, Jan 21st 12pm. You can either hand in the solution electronically with WebCT or on paper during class.
Use `R` or SPlus for your computation. Whenever your solution involves either one, hand in the **relevant** output.

1 Generalized Inverse, cf. Rencher 2.46

Let $A = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{pmatrix}$

(a) Find a symmetric generalized inverse for A,

(b) find a non-symmetric generalized inverse for A.

(c) Load the library MASS in `R` and use the command `ginv()` to find a generalized inverse for A. Does the result match either one of the matrices you found by hand?

2 Generalized Inverse, cf Koehler

Consider the matrices $A = \begin{pmatrix} 4 & 4.001 \\ 4.001 & 4.002 \end{pmatrix}$ and $B = \begin{pmatrix} 4 & 4.001 \\ 4.001 & 4.002001 \end{pmatrix}$.
Obviously, these matrices are nearly identical. Use `R` and compute the determinants and inverses of these matrices. Note that $A^{-1} \approx 3B^{-1}$ even though the original matrices are nearly the same. This shows that small changes in the elements of nearly singular matrices can have big effects on some matrix operations.

3 Eigenvalues, cf. Rencher 2.72

Let $A = \begin{pmatrix} 1 & 1 & -2 \\ -1 & 2 & 1 \\ 0 & 1 & -1 \end{pmatrix}$

(a) Find the eigenvalues and associated normalized eigenvectors.

(b) Find $\mathrm{tr}(A)$ and determinant of A and verify that $\mathrm{tr}(A) = \sum_{i=1}^{3} \lambda_i$ and $\det(A) = \prod_{i=1}^{3} \lambda_i$.

4 Matrices, cf. Rencher 2.79

Let

$$A = \begin{pmatrix} \frac{2}{3} & 0 & \frac{1}{3}\sqrt{2} \\ 0 & 1 & 0 \\ \frac{1}{3}\sqrt{2} & 0 & \frac{1}{3} \end{pmatrix}$$

(a) Find the rank of A.

(b) Show that A is idempotent.

(c) Show that $I - A$ is idempotent.

(d) Show that $A(I - A) = 0$.

(e) Find $\text{tr}(A)$.

(f) Find the eigenvalues of A.

5 Multivariate Normal Distribution, cf Rencher 3.19

Let $y = (y_1, y_2, y_3)'$ be a random vector with mean vector and covariance matrix

$$\mu = \begin{pmatrix} 1 \\ -1 \\ 3 \end{pmatrix}, \text{ and } \Sigma = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 3 \\ 0 & 3 & 10 \end{pmatrix}$$

(a) Let $z = 2y_1 - 3y_2 + y_3$. Find $E[z]$ and $Var[z]$.

(b) Let $z_1 = y_1 + y_2 + y_3$ and $z_2 = 3y_1 + y_2 - 2y_3$. Find $E[z]$ and $cov[z]$, where $z = (z_1, z_2)'$.

6 Blood Coagulation

Blood Coagulation times (in seconds) for blood samples from six different rats were taken. Each rat was fed one of three diets.

Diet 1	Diet 2	Diet 3
62	71	72
60		68
		67

(a) Write one-way anova models (in matrix notation) for both the "means" and the "effects" model.

(b) Check that for both versions $C(X)$ are the same. Find an orthogonal projection matrix P_X, that projects onto $C(X)$.

7 Temperatures in the US

The data at `http://www.public.iastate.edu/ hofmann/stat511/data/temperature.txt` gives the normal average January minimum temperature in degrees Fahrenheit with the latitude and longitude of 56 U.S. cities. (For each year from 1931 to 1960, the daily minimum temperatures in January were added together and divided by 31. Then, the averages for each year were averaged over the 30 years.)

(a) Load the data into `R`.

(b) Draw a scatterplot of the relationship between longitude and temperature in January. Describe the relationship between the variables.

(c) Compute the least squares estimates for the model

$$JanTemp = \alpha + \beta Long + \epsilon$$

Draw a scatterplot of the residuals versus longitude. Describe the relationship between the variables.

(d) Write out a linear model (in matrix notation form) of the relationship you found in (b).

(e) Using `R` find the least squares estimate for the model.

Stat 511 Homework 2 Spring 2005

Maximum score is 20 points, due date is Wednesday, Feb 2nd 12pm. You can either hand in the solution electronically with WebCT or on paper during class.
Use R or SPlus for your computation. Whenever your solution involves either one, hand in the **relevant** output.

8 Estimability in Effects Model, cf. Rencher 11.10

In the model

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \text{ with } i = 1, 2, ..., k, \text{ and } j = 1, ..., n$$

show that $\sum_{i=1}^k c_i \tau_i$ is estimable if and only if $\sum_{i=1}^k c_i = 0$ (as suggested following the example 11.2.2(b) in Rencher).
Use the following two approaches:

(a) In $\lambda' \beta = \sum_{i=1}^k c_i \tau_i$ express λ' as a linear combination of the rows of X.

(b) Express $\sum_{i=1}^k c_i \tau_i$ as a linear combination of the elements of $E[Y] = X\beta$.

9 Anova Model

Consider the (non-full-rank) effects model for the 2×2 factorial model (with 2 observations per cell) as given in the notes on page 4 (in the framework of the Pizza Delivery Data)

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon ijk, \text{ with } i = 1, 2, j = 1, 2, k = 1, 2$$

(a) Determine which of the parametric functions below are estimable.

$$\alpha_1, \alpha_1 - \alpha_2, \mu + \alpha_1 + \beta_1, \mu + \alpha_1 + \beta_1 + \alpha\beta_{11}, \alpha\beta_{11}, \alpha\beta_{12} - \alpha\beta_{11} - (\alpha\beta_{22} - \alpha\beta_{21})$$

For those that are estimable, find the 9×1 row vector $c'(X'X)^- X'$ that when multiplied by Y produces the ordinary least squares estimate of $c'\beta$.

(b) For the parameter vector $\beta = (\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \alpha\beta_{11}, \alpha\beta_{12}, \alpha\beta_{21}, \alpha\beta_{22})'$, consider the hypothesis $H_0 : C\beta = 0$, for

$$C = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \end{pmatrix}$$

Is this hypothesis testable? Explain.

10 House Prices, cf. Dr Vardeman's Notes

On the Web page http://www.public.iastate.edu/~hofmann/stat511/data/ you will find the file homes.txt. We're going to do some statistical analysis on this set of home sale price data obtained from the Ames City Assessors Office. Data on sales December 2003 through January 2005 of 1 1/2 and 2 story homes built 1945 and before. $n = 86$ different homes fitting this description were sold in Ames during this period. For each home, the value of the response variable recorded sales price of the home = *Price* and the values of 14 potential explanatory variables were obtained. These variables are

- *Size*, the floor area of the home above grade in sq ft

- *Land*, the area of the lot the home occupies in sq ft
- *BedRooms*, a count of the number in the home
- *Central Air*, a binary variable that is “yes” if the home has central air conditioning and is “no” if it does not
- *Fireplace*, a count of the number in the home
- *Full Bath*, a count of the number of full bathrooms above grade
- *Half Bath*, a count of the number of half bathrooms above grade
- *Basement*, the floor area of the homes basement (including both finished and unfinished parts) in sq ft
- *FinishedBsmt*, the percentage of area of any finished part of the homes basement
- *Bsmt Bath*, a dummy variable that is 1 if there is a bathroom of any sort (full or half) in the homes basement and is 0 otherwise
- *Garage*, a dummy variable that is 1 if the home has a garage of any sort and is 0 otherwise
- *Style* (2 Story), a dummy variable that is 1 if the home is a 2 story home (or bigger) and is 0 otherwise
- *Zone* , a dummy variable that is 1 if the home is in an area zoned as Urban Core Medium Density and 0 otherwise

The first row of the file has the variable names in it. (You might open this file by double clicking on the link to have a look at it.)

```
> homes <- read.table("http://www.public.iastate.edu/~hofmann/stat511/data/home-04.txt",
+ header=T,sep="\t")
```

Use the command

```
> homes
```

to view the data frame. It should have 14 columns and 86 rows. You can check that by using the command

```
> dim(homes)
```

Now create two subsets of the data that will be used to fit a regression model. Type

```
> Y<-homes[,1]
> X<-homes[,c(2,4,9,10,3)]
```

Note the use of [] to select columns from the data frame. To add a column of ones to the data frame, type

```
> X0<-rep(1,length(Y))
> X<-cbind(X0,X)
```

Make a scatterplot matrix for $y, x_1, x_2, ..., x_5$. To do this, first load the lattice package. You can do that with the command

```
> library(lattice)
```

(Look under the ”Packages” heading on the R GUI to see all available packages)
Then type

```
 > splom(~homes[,c(1,2,4,9,10,3)],aspect="fill")
```

If you had to guess based on this plot, which single predictor do you think is probably the best predictor of Price? Do you see any evidence of multicollinearity (correlation among the predictors) in this graphic? Also compute a sample correlation matrix for $y, x_1, x_2, x_3, x_4, x_5$. You may compute the matrix using the `cor()` function and round the printed values to four places using the `round()` function as

```
> round (cor(homes[c(1,2,4,9,10,3)]),4)
```

Use the `qr()` function to find the rank of X. Use R matrix operations on the X matrix and Y vector to find the estimated regression coefficient vector b_{OLS}, the estimated mean vector $\hat{Y}$, and the vector of residuals $e = Y - \hat{Y}$.
For the remainder of this question, you will need to transform X and Y into a matrix format. This is done by

```
> Y <- as.matrix(Y)
> X <- as.matrix(X)
```

Plot the residuals against the fitted means. After loading the MASS package, this can be done using the following code.

```
> b<-solve(t(X)%*%X)%*%t(X)%*%Y
> yhat<-X%*%b
> e<-Y-yhat
> par(fin=c(6.0,6.0),pch=18,cex=1.5,mar=c(5,5,4,2))
> plot(yhat,e,xlab="Predicted Y",ylab="Residual",main="Residual Plot")
```

Type `> help(par)` to see the list of parameters that may be set on a graphic. What does the first specification above do, i.e. what does `fin=c(6.0,6.0)` do? Plot the residuals against home size. You may use the following code.

```
> plot(homes$Size,e,xlab="Size",ylab="Residual",main="Residual  Plot")
```

And you can add a smooth trend line to the plot by typing

```
> lines(loess.smooth(homes$Size,e,0.90))
```

What happens when you type `> lines(loess.smooth(homes$Size,e,0.50))` ? (The values 0.90 and 0.50 are values of a ”smoothing parameter.” You could have discovered this (and more) about the loess.smooth function by typing `> help(loess.smooth)`) Now plot the residuals against each of x_2, x_3, x_4, and x_5.
Create a normal plot from the values in the residual vector. You can do so by typing

```
> qqnorm(e,main="Normal Probability Plot")
> qqline(e)
```

Now compute the sum of squared residuals and the corresponding estimate of σ^2, namely

$$\hat{\sigma}^2 = \frac{(Y - \hat{Y})'(Y - \hat{Y})}{n - \text{rank}(X)}$$

Use this and compute an estimate of the covariance matrix for b_{OLS} , namely

$$\hat{\sigma}^2(X'X)^{-1}$$

Sometimes you may want to write a summary matrix out to a file. This can be done as follows. First prepare the row and columns labels and round all entries to 4 places using the code

```
 > case<-1:86
 > temp<-cbind(case,homes[,c(2,4,9,10,3)],Y,yhat,e)
 > round(temp,4)
```

Then with the MASS package loaded (in order to make the `write.matrix` function available). The code

```
> write.matrix(temp,file="c:/temp/regoutput.out")
```

will then write output to the file `c:/temp/regoutput.out` (you may choose another name and destination for this file). Modify the above to create a matrix that has b_{OLS} in the first column and a vector of corresponding standard errors (square roots of diagonal entries of the estimated covariance matrix for b_{OLS}) in the second. Label the rows and columns of your matrix and write it out to a file. Submit a listing of that file and your R code.
It is, of course, possible to do the linear model calculations in R "automatically" by calling the right function. After loading the homes data frame, type

```
> lm(formula = Price ~ Size + BedRooms + Basement + FinishedBsmt + Land, data = homes)
```

Compare the values printed out with things you computed more painfully before. There is not much detail in what is printed out. Try instead typing

```
> summary(lm(Price ~ Size + BedRooms + Basement + FinishedBsmt + Land, data = homes))
```

and notice that there is more detail provided.

Stat 511 Homework 3 Spring 2005

Maximum score is 20 points, due date is Wednesday, Feb 9th 12pm. You can either hand in the solution electronically with WebCT or on paper during class.
Use `R` or SPlus for your computation. Whenever your solution involves either one, hand in the **relevant** output.

11 Anova Model

Consider the one-way ANOVA model:

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

Suppose, that there are 4 treatments (groups) ($i = 1, 2, 3, 4$) and the sample sizes are respectively 2,1,1,2 for treatments 1 through 4.
Let $Y' = (2, 1, 5, 4, 3, 6)$ and

$$C = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}, \qquad V_1 = \text{diag}(1, 4, 4, 1, 2, 2), \text{ and}$$

$$V_2 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 2 \end{pmatrix}.$$

For both covariance structures V_1 and V_2:

(a) Find the ordinary least squares estimate of EY and $C\beta$.

(b) Use linear algebra in R to find weighted generalized least squares estimates for EY and $C\beta$ (You will find the R command `eigen()` helpful - type `> help(eigen)` for more details).

(c) Compare the covariance matrices for generalized least squares estimators to covariance matrices for OLS estimators of EY and the $C\beta$.

(d) The `lm` function in R allows one to do weighted least squares, i.e. minimize $\sum w_i(y_i - \hat{y}_i)^2$ for positive weights w_i. For the V_1 case of the Aitken model, find the BLUEs of the 4 treatment means using `lm` and an appropriate vector of weights. (Type `> help(lm)` in R in order to get help with the syntax.)

12 Chemical Process (cf. Ken Koehler's notes)

Data were collected to study the effect of temperature on the yield of a chemical process. Two different catalysts A and B, were used in the study. Yields were measured under 5 different temperatures for each catalyst. The data are as follows:

Run	Yield (grams)	Temperature (C)	Catalyst	Run	Yield (grams)	Temperature (C)	Catalyst
3	20	90	A	9	25	90	B
10	24	95	A	2	29	95	B
4	27	100	A	6	32	100	B
8	33	105	A	1	37	105	B
5	38	110	A	7	41	110	B

Each run can be considered as an independent observation. The order in which the runs were made was randomized.
Consider the linear model

$$y_{ij} = \mu + \alpha_i + \beta(T_{ij} - 100) + \epsilon_{ij}, \text{ for } i = 1, 2 \text{ and } j = 1, 2, ..., 5$$

where Y_{ij} the observed yield for the run using the i-th catalyst and the j-th temperature level.
α_i corresponds to the ith catalyst
T_{ij} the temperature under which the process was run

(a) For this linear model the vector of mean responses can be written as $E[Y] = Xb$. Write out the model matrix X corresponding to the parameter vector $b = (\mu, \alpha_1, \alpha_2, \beta)$.

(b) Determine which of the following quantities are estimable:

$$\mu, \mu + \alpha_2, \beta, \alpha_1 - \alpha_2, \mu + \beta T, \mu + \alpha_1 + \beta(T - 100),$$

where T is any specified temperature.

(c) The data are posted in the file data/process.txt on the course web page. This file has five columns. The first two columns match the first two columns in the table shown above. The third and fourth column use dummy variables to indicate which catalyst was used. The third column is coded 1 when catalyst A was used and coded 0 when the other catalyst was used. The fourth column is coded 1 when catalyst B was used and coded 0 when the other catalyst is used. The fifth column contains the temperature values minus 100. Use the command

```
process <- read.table("http://www.public.iastate.edu/~hofmann/data/process.txt",
header= T)
```

to enter these data into a data frame in R.

Use the command

```
Y <- as.matrix(process[ ,2 ])
```

to create a vector of observed responses. Use the command

```
X  <- as.matrix(cbind(rep(1,length(Y)),process[ ,3:5]))
```

to construct the model matrix for the model in part (a).

Use R to compute the vector $(X'X)^{-}X'Y$. Store the result in object `b`.

For all estimable quantities in (b) compute OLS estimates using object `b`. Report their values.

(d) To visually check if the proposed model is reasonable for these data create a scatter plot of yield (Y) versus temperature (T). Use an open circle for the 5 observations with catalyst A and a filled circle for the five observations with catalyst B. Also include two parallel lines on the plot corresponding to the least squares estimates of the models for catalysts A and B. This can be done with the following code:

```
# read data in a data frame
process <- read.table("http://www.public.iastate.edu/~hofmann/stat511/data/process.txt",
header=T)

# create response
Y <- process[,2]
```

```
# construct vector of temperatures
temp <- process$temp

# construct a factor to represent the groups
groups <- as.factor(process$B + 1)

# fit the model
lmfit <- lm(Y~groups+temp)

# Draw the plot
# initially, only draw the frame and title
plot(c(min(temp),max(temp)), c(min(Y),max(Y)), xlab='Temperature-100',ylab='Yield',
type="n", main="Problem 2 on HW 3")

# fill in observations
points(temp[groups==1],Y[groups==1],pch=1)
points(temp[groups==2],Y[groups==2],pch=16)

# fill in fitted lines
mu <- lmfit$coefficients[1]
alpha1 <- 0
alpha2 <- lmfit$coefficients[2]
beta <- lmfit$coefficients[3]

abline(c(mu+alpha1, beta))
abline(c(mu+alpha2, beta))
```

Submit your plot and comment on the results.

Stat 511 Homework 4 Spring 2005

Maximum score is 20 points, due date is Wednesday, Feb 16th 12pm. You can either hand in the solution electronically with WebCT or on paper during class.
Use R or SPlus for your computation. Whenever your solution involves either one, hand in the **relevant** output.

13 Distribution of Quadratic Forms

For the following questions use the theorems we have shown in class.

(a) Let $y \sim N(\mu, \sigma^2 I_{n\times n})$. Find the distribution of $\frac{1}{\sigma^2}y'y$.

(b) Let $y_i \sim N(\mu, \sigma^2)$ i.i.d for $i = 1, ..., n$. Let s^2 and $\bar{y}$ be sample variance and sample mean respectively, i.e.

$$\bar{y} = \frac{1}{n}\sum_i y_i \qquad \text{and} \qquad s^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2$$

i. Find the distribution of $(n-1)s^2/\sigma^2$.

ii. Why are $\bar{y}$ and s^2 independent?

iii. Find a $(1-\alpha)100\%$ C.I. for σ^2.

14 Testable Hypothesis

Consider the linear model

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij},$$

where $j = 1, 2, 3$ the treatment, and $i = 2, 4, 2$ the number of respective replicates for treatment j. Assume that Gauss-Markov assumptions hold.

(a) Determine which of the following hypothesis are testable. If possible, write the hypothesis in the form $H_0 : C\beta = d$:

i. $H_0 : \alpha_1 = \alpha_3$

ii. $H_0 : \alpha_1 - 2\alpha_2 + \alpha_3 = 0$

iii. $H_0 : \alpha_2 = 0$

iv. $H_0 : \mu = 0$

v. $H_0 : \alpha_1 = \alpha_3$ and $\alpha_1 - 2\alpha_2 + \alpha_3 = 0$

vi. $H_0 : \alpha_1 = \alpha 2 = \alpha_3$ and $2\alpha_1 - \alpha_2 - \alpha_3 = 0$

(b) For the hypothesis

$$H_0 : \begin{pmatrix} 0 & 1 & 0 & -1 \\ 0 & 1 & -2 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

construct a test against the alternative

$$H_a : \alpha_1 \neq \alpha_2 \text{ or } \alpha_1 - 2\alpha_2 + \alpha_3 \neq 0$$

Cover the following parts:

i. Show that SSE/σ^2 has a central χ^2 distribution with 5 degrees of freedom - this will give the denominator of your F statistic.

ii. Express the numerator of your test statistic as quadratic form. Show that it is independent of SSE. Show that it has a non central chi^2 distribution.

iii. Show that you can choose your test statistic such that it has a non-central F distribution. Report the degrees of freedom and non-centrality parameter as a function of α_1, α_2 and α_3.

iv. Show that under the null hypothesis your test statistic has a central F distribution.

15 Chemical Process - revisited

For this question, we are going to use the data on a chemical process
(at http://www.public.iastate.edu/~hofmann/stat511/data/process.txt) again.

(a) Consider the linear model

$$y_{ij} = \mu + \alpha_i + \beta(T_{ij} - 100) + \epsilon_{ij}, \text{ for } i = 1, 2 \text{ and } j = 1, 2, ..., 5$$

where Y_{ij} the observed yield for the run using the i-th catalyst and the j-th temperature level.
α_i corresponds to the ith catalyst
T_{ij} the temperature under which the process was run, and
$\epsilon_{ij} \sim MVN(0, \sigma^2 I)$.

i. Find 90% two-sided C.I for σ, $\mu + \alpha_1$, β and $\alpha_1 - \alpha_2$.

ii. Find a p-value for testing the null hypothesis $H_0 : \alpha_1 - \alpha_2 = 0$ vs $H_a : \alpha_1 - \alpha_2 \neq 0$.

iii. Find 90% two-sided prediction limits for the sample mean of four future observations using catalyst A at temperature level 105.

(b) Let $Y \sim MVN(W\gamma, \sigma^2 I)$ with

$$W\gamma = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \\ -10 & -5 & 0 & 5 & 10 & -10 & -5 & 0 & 5 & 10 \end{pmatrix}' \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix}$$

i. Show that this model is a reparameterization of the model in (a)

ii. Give an interpretation of the parameters γ_1, γ_2, and γ_3 with respect to yield, catalyst and temperature.

iii. Use R to find OLS estimates for γ and EY. Plot the vector of residuals versus Y and comment.

16 F Distribution

Use the R function `pf(x, df1, df2, ncp=0,log = FALSE)` to plot on the same set of axes the F probability distribution function for degrees of freedom 3 and 5 and non-centrality parameter 0,1,2,5,10. Describe the result.

Stat 511 Homework 5 Spring 2005

Maximum score is 20 points, due date is Wednesday, Mar 7th 12pm. You can either hand in the solution electronically with WebCT or on paper during class.
Use `R` or SPlus for your computation. Whenever your solution involves either one, hand in the **relevant** output.

17 Particle Data, cf. Dr Vardeman

The data set `particle.txt` located

`http:\\www.public.iastate.edu/~hofmann/stat511/data/particle.txt`

. Given are (coded) values of y = *strength*, x_1 = *temperature*, x_2 = *time* from an experiment run to find good values of x_1 and x_2 (ones that produce large y) for an industrial process used to make particle boards. A "response surface analysis of these data is based on a multivariate quadratic regression. Use R and appropriate matrix calculations to do the following.

(a) Fit the (linear in the parameters and quadratic in the predictors) model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{2i}^2 + \beta_5 x_{1i} x_{2i}$$

to these data. Then compute and plot standardized residuals.

(b) In the model from a), test $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ Report a p-value. Does quadratic curvature in response (as a function of the x's) appear to be statistically detectable? (If the null hypothesis is true, the response is planar as a function of the x's.)

(c) Use some multivariate calculus on the fitted quadratic equation and find the location (x_1, x_2) of an absolute maximum. Use R matrix calculations to find 90% two-sided confidence limits for the mean response here. Then find 90% two-sided prediction limits for a new response from this set of conditions.

18 Diagnostic Tools

Read chapter 9 in Rencher and identify for the model in question 1 points, that

(a) might be outliers,

(b) are influential.

Confirm your results with appropriate scatterplots.

19 Protein Diet

The file `protein.txt` at

`http:\\www.public.iastate.edu/~hofmann/stat511/data/protein.txt`

contains data on the weight gain of 47 rats, who were exposed to different diets (cf. Rencher, table 14.9).

(a) Load the data from the file into R and create vector `gain`.

(b) Create two variables `protein` and `meat`, both of length 47, to encode under which diet combination a rat is in. (You might find the command `rep` useful). Make sure that both `protein` and `meat` are factors in R (use command `as.factor`).

(c) Make boxplots of weight gain within each diet combinations:

```
> boxplot(gain~protein+meat, col=2)
```

Comment on the result.

(d) Using the command `means <- tapply(gain, list(meat,protein),FUN=mean); means` compute and print out the cell means.

(e) Make a crude interaction plot by doing the following. First type

```
> x.axis<-as.integer(unique(d$meat))
```

to set up horizontal plotting positions for the sample means. Then make a “matrix plot” with lines connecting points by issuing the commands

```
> matplot(c(1,3),c(70,110),type="n",xlab="Meat",ylab="Mean  Response",main="Weight Gain")
> matlines(x.axis,means,type="b")
```

The first of these commands sets up the axes and makes a dummy plot with invisible points ”plotted” at (1,70) and (3,110). The second puts the lines and identifying protein levels (as plotting symbols) on the plot. Comment on the result.

(f) Set the default for the restriction used to create a full rank model matrix, run the linear models routine and find both sets of “Type I” sums of squares by issuing the following commands

```
> options(contrasts=c("contr.sum","contr.sum"))
> lm.out1<-lm(gain~protein*meat)
> anova(lm.out1)
> lm.out2<-lm(gain~meat*protein)
> anova(lm.out2)
```

Load library `car` and use the command `Anova` to compute Type II and Type III sum of squares. Describe briefly the corresponding hypotheses for these sum of squares.

(g) Start over with this problem, doing the calculations “from scratch using your basic linear models knowledge and matrix calculations in R. Compute all of Type I, Type II and Type III sums of squares here, using the sum restriction in the first two cases (and the order of factors A, B). Then compute Type I and Type II sums of squares using the SAS baseline restriction. Did you expect the results to be different? Explain.

(h) Now suppose that by some misfortune some rats escaped from their cages and all of the observations corresponding to cell (protein=low, meat =pork) got lost - leading to an incomplete design. Test the hypothesis that at least for the cells where one has data, there are no interactions.

Stat 511 Homework 6 Spring 2005

Maximum score is 20 points, due date is Wednesday, Mar 23rd 12pm. You can either hand in the solution electronically with WebCT or on paper during class.
Use R or SPlus for your computation. Whenever your solution involves either one, hand in the **relevant** output.

20 Nonlinear Regression, cf. Dr Vardeman

Consider the following small data set:
y is reaction velocity (in counts/min^2) for an enzymatic reaction and x is substrate concentration (in ppm) for untreated enzyme and enzyme treated with Puromycin.

x		0.02	0.06	0.11	0.22	0.56	1.10
y	Untreated	67, 51	84, 86	98, 115	131, 124	144, 158	160
	Treated	76, 47	97, 107	123, 139	159, 152	191, 201	207, 200

A standard model here (for either the untreated enzyme or for the treated enzyme) is the “Michaelis-Menten model :

$$y_i = f(x_i, \theta) + \epsilon_i \qquad \text{with } f(x_i, \theta) = \frac{\theta_1 x_i}{\theta_2 + x_i} \tag{1}$$

Note that in this model, 1) the mean of y is 0 when $x = 0$,
2) the limiting (large x) mean of y is θ_1, and
3) the mean of y reaches half of its limiting value when $x = \theta_2$.

Begin by considering only the Treated part of the data set (and an $\epsilon_i \sim N(0, \sigma^2)$ iid version of the model). Read in vectors y and x (of length 12).

(a) Plot x vs y and make“eye-estimates of the parameters based on your plot and the interpretations of the parameters offered above.

(b) Add the `stats` package to your R environment (it contains function `nls`). Then issue the command

```
> REACT.fm<nls(formula=y~theta1*x/(theta2+x),start=c(theta1=#,theta2=##),trace=T)
```

where in place of # and ## you enter your eye-estimates from a). This will fit the nonlinear model (**??**) via least squares. What are the least squares estimate of the parameter vector and the “deviance (error sum of squares)

$$\hat{\theta}_{OLS} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \qquad \text{and} \qquad SSE = \sum_{i=1}^{12} \left(y_i - f(x_i, \hat{\theta}_{OLS}) \right)^2$$

(c) Re-plot the original data with a superimposed plot of the fitted equation. To do this, you may type

```
> conc<-seq(0,1.5,.05)
> velocity<-coef(REACT.fm)[1]*conc/(coef(REACT.fm)[2]+conc)
> plot(c(0,1.5),c(0,250),type="n",xlab="Conc (ppm)",ylab="Vel  (counts/sqmin)")
> points(x,y)
> lines(conc,velocity,col=3)
```

(d) Get more complete information on the fit by typing

```
> summary(REACT.fm)
> vcov(REACT.fm)
```

Verify that the output of this last call is $MSE\left(\hat{D}'\hat{D}\right)^{-1}$ and that the standard errors produced by the first are square roots of diagonal elements of this matrix. Then use the information produced here and make an approximate 95% prediction interval for one additional reaction velocity, for substrate concentration .50 ppm.

(e) The concentration, say x_{100}, at which mean reaction velocity is 100 counts/min^2 is a function of θ_1 and θ_2. Find a sensible point estimate of x_{100} and a standard error (estimated standard deviation) for your estimate.

(f) As a means of visualizing what function the R routine `nls` minimized in order to find the least squares coefficients, do the following. First set up a grid of (θ_1, θ_2) pairs as follows. Type

```
> theta<-coef(REACT.fm)
> se<-sqrt(diag(vcov(REACT.fm)))
> dv<-deviance(REACT.fm)
> gsize<-101
> th1<-theta[1]+seq(-4*se[1],4*se[1],length=gsize)
> th2<-theta[2]+seq(-4*se[2],4*se[2],length=gsize)
> th<-expand.grid(th1,th2)
```

Then create a function to evaluate the sums of squares

```
> ss<-function(t)
+ {
+ sum((y-t[1]*x/(t[2]+x))^2)
+ }
```

As a check to see that you have it programmed correctly, evaluate this function at $\hat{\theta}_{OLS}$ for the data in hand, and verify that you get the deviance. Then evaluate the error sum of squares over the grid of parameter vectors θ set up earlier and produce a contour plot using

```
> SumofSquares<-apply(th,1,ss)
> SumofSquares<-matrix(SumofSquares,gsize,gsize)
> plot(th1,th2,type="n",main="Error Sum of Squares Contours")
> contour(th1,th2,SumofSquares,levels=c(seq(1000,4000,200)))
```

What contour on this plot corresponds to an approximately 90% approximate confidence region for the parameter vector θ?

(g) Now add to the contour plotting, by placing two additional contours on the plot using the following code.

```
> plot(th1,th2,type="n",main="Error Sum of Squares Contours")
> contour(th1,th2,SumofSquares,levels=dv*c((1+.1*qf(.95,1,10)),
+ (1+.2*qf(.95,2,10))),add=T,col=3)
```

Identify on this plot an approximately 95% (joint) confidence region for θ and individual 95% confidence regions for θ_1 and θ_2.

(h) Use the standard errors for the estimates of the coefficients produced by the routine `nls()` and make 95% t intervals for θ_1 and θ_2. How much different are these from your intervals in g)? (Notice that the sample size in this problem is small and reliance on any version of large sample theory to support inferences is tenuous. One should take any of these inferences above as very approximate. We will later discuss the possibility of using "bootstrap" calculations as an alternative method of inference.)

(i) Make two different approximate 95% confidence intervals for σ. Base one on carrying over the linear model result that $SSE/\sigma^2 \sim \chi^2_{n-k}$. Base the other on the profile likelihood material.

(j) Use the R function `confint()` to get 95% intervals for θ_1 and θ_2. (You can find `confint` in the `MASS` package) Then type

```
> confint(REACT.fm, level=.95)
```

How do these intervals compare to the ones you found in part g)?

(k) Apparently, scientific theory suggests that treated enzyme will have the same value of θ_2 as does untreated enzyme, but that θ_1 may change with treatment. That is, if

$$z_i = \begin{cases} 0 & \text{if treated (Puromycin is used)} \\ 1 & \text{otherwise} \end{cases}$$

a possible model is

$$y_i = \frac{(\theta_1 + \theta_3 z_i)\, x_i}{\theta_2 + x_i} + \epsilon_i$$

and the parameter θ_3 then measures the effect of the treatment. Go back to the data table and now do a fit of the (3 parameter) nonlinear model including a possible Puromycin effect using all 23 data points. Make 2 different approximately 95% confidence intervals for θ_3. Interpret these. (Do they indicate a statistically detectable effect? If so, what does the sign say about how treatment affects the relationship between x and y?) Plot on the same set of axes the curves

$$y = \frac{\hat\theta_1 x}{\hat\theta_2 + x} \qquad \text{and} \qquad y = \frac{\left(\hat\theta_1 + \hat\theta_3\right) x}{\hat\theta_2 + x} \qquad \text{for } 0 < x < 2$$

21 Gauss-Newton Algorithm

For the Michaelis Menten Model as given in (**??**), write a function `mmGN` in `R`, which provides one step of a Gauss Newton algorithm, i.e. write a function

```
mmGN <- function (x,y, theta1, theta2) {
   .
   .
   .
   print (c(theta1new, theta2new, sse, maxtheta))
   return (c(theta1new, theta2new))
}
```

where `sse` is the deviance in this step, i.e.

$$SSE = \sum_i (y_i - f(x_i, \theta))^2$$

and `maxtheta` is the maximum relative difference between the old and the new `theta`:

$$\max(\left|\theta_1^{r+1} - \theta_1^r\right| / |\theta_1^r + EPS|\,, \left|\theta_2^{r+1} - \theta_2^r\right| / |\theta_2^r + EPS|),$$

where you can pick EPS as some small number.
Use the following code to test your routine:

```
theta <- c(#, ##)
for (i in 1:10) {
   theta <- mmGN (x,y, theta[1], theta[2])
}
theta
```

where # and ## are your starting values for θ_1 and θ_2 as in question 1 (b) Comment on the values of `sse` and `maxtheta`.

Stat 511 Homework 7 Spring 2005

Maximum score is 20 points, due date is Wednesday, Apr 6th 12pm. You can either hand in the solution electronically with WebCT or on paper during class.
Use R or SPlus for your computation. Whenever your solution involves either one, hand in the **relevant** output.

22 MS Exam

Do Part III of the Stat 511 question from the May 2003 Statistics MS Exam. A pdf of the questions is posted on the course's webpage

23 Calibration, cf. Dr Vardeman

The following is a "calibration scenario met in an engineering lab. Several resistance temperature devices (RTDs) from a large batch of the devices (that can be used to provide cheap temperature measurements in terms of resistance they register) were placed in a water bath with an extremely high quality thermometer (that we will assume produces the "true temperature of the bath). At some regular intervals, both the "true temperature and measured resistances were recorded. Let y_{ij} be the measured resistance produced by RTD i at time$j$ and suppose that these are linear functions of the true temperatures plus random error, i.e. assume that for $t_j =$ the jth measured temperature it is appropriate to model y_{ij} as

$$y_{ij} = \alpha_i + \beta_i t_j + \epsilon_{ij}$$

where α_i and β_i are intercept and slope specific to the particular RTD.
Further suppose that α_i and β_i can be described as

$$\alpha_i = \alpha + \gamma_i \quad \text{and} \quad \beta_i = \beta + \delta_i$$

where α and β are known constants and γ_i and δ_i are unobservable random effects. Assume that $E\gamma_i = E\delta_i = E\epsilon_{ij} = 0$ for all i and j. Assume further that variances are

$$Var\gamma = \sigma^2_\gamma I, Var\delta = \sigma^2_\delta I \quad \text{and} \quad Var\epsilon = \sigma^2 I$$

and they are pairwise independent.
We then have a model with the 5 parameter vectors $\alpha, \beta, \sigma^2_\gamma, \sigma^2_\delta$ and σ^2. The first two of these are "fixed effects and the last three are "variance components. Suppose that there are 3 RTDs and only 3 different measured temperatures, with respective (coded) values $t_1 = 0, t_2 = 1$ and $t_3 = 4$.

(a) Write out in matrix form the mixed linear model for Y

(b) What is EY? Write out and simplify as much as you can the covariance matrix, $VarY$.

(c) Suppose that in fact $Y = (99.8, 108.1, 136.0, 100.3, 109.5, 137.7, 98.3, 110.1, 142.2)'$ and that it is known somehow that $\sigma^2_\gamma = 1, \sigma^2_\delta = 1$ and $\sigma^2 = 0.25$. We can then use generalized least squares to estimate the fixed effects vector α, β. Do this. Does the answer change if we know only that $\sigma^2_\gamma = \sigma^2_\gamma = 4\sigma^2$? Explain. (Indeed, can I even get a BLUE for α and β with only this knowledge?)

(d) Suppose that I know that $\sigma^2_\gamma = 1, \sigma^2_\delta = 1$ and $\sigma^2 = 1$. Use again the data vector from part c). What is the BLUE of (α, β) under these circumstances?

(e) Suppose that it is your job to estimate the variance components in this model. One thing you might consider is maximum likelihood under a normal distribution assumption. This involves maximizing the likelihood as a function of the 5 parameters and its clear how to get the profile likelihood for the variance components alone. That is, for a fixed set of variance components $(\sigma^2_\gamma, \sigma^2_\delta, \sigma^2)$ one knows $VarY$, and may use generalized least squares to estimate EY and plug that into the likelihood function in order to get the profile likelihood for $(\sigma^2_\gamma, \sigma^2_\delta, \sigma^2)$. Consider the two vectors of variance components used in parts c) and d) of this problem and the data vector from c). Which set has the larger value of profile likelihood (or profile loglikelihood)?

24 Acidity in Plants, cf. Dr Koehler

Four plants of the same variety were randomly sampled from a large field of plants. Three leaves were randomly selected from each plant and three determinations of the concentration of a certain acid were made on each leaf using each of three different methods. These methods are labeled as method A, method B, and method C. Larger values correspond to higher concentrations of the acid. You can find the data in file `macid.txt` at `http://www.public.iastate.edu/h̃ofmann/stat511/data`
Consider the model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{jk} + \epsilon_{ijk}$$

where y_{ijk} is the observed acid concentration made by the i-th method on the k-th leaf of the j-th plant and

$$\beta_j \sim N(0, \sigma^2_\beta) \quad \gamma_{jk} \sim N(0, \sigma^2_\gamma \quad \epsilon_{ijk} \sim N(0, \sigma^2)$$

and all random effects are independent of each other. In this model, β_j represents variation in acid concentrations among plants, γ_{jk} represents variation in acid concentrations among leaves within plants, and ϵ_{ijk} represents and remaining random variation.

(a) Plot the data: plot measurements of acidity by plant and method, and by plant and leaf. Describe the different sources of variations - based on the plots, are the model assumptions made justified?

(b) Using this model, obtain an ANOVA table for the observed data.

(c) Report formulas for expectations of mean squares.

(d) Obtain REML estimates of the variance components $\sigma^2_\beta, \sigma^2_\gamma$ and σ^2. What is the largest source of random variation in this study?

(e) Construct a 95% confidence interval for the mean acid concentration for the population of plants as measured by method C.

(f) Examine differences in estimates of acid concentration means for all of the methods. State your conclusions.

(g) Estimate the correlation between determinations of acid concentrations taken from two different leaves of the same plant, compare this to the estimate of the correlation between determinations of acid concentrations taken from the same leaf.

Stat 511 Homework 8 Spring 2005

Maximum score is 20 points, due date is Friday, Apr 22nd 12pm. You can either hand in the solution electronically with WebCT or on paper during class.
Use `R` or SPlus for your computation. Whenever your solution involves either one, hand in the **relevant** output.

25 Bootstraps

(a) **Heart Transplants**

i. The file posted as heart1.txt on the course web page contains survival times (in days) for 69 patients who received heart transplants. There is one line of data for each patient and each line contains a survival time. Compute the sample median of the survival times. Use bootstrap methods to obtain a standard error for the sample median and 95% percentile confidence limits for the population median. Use 5000 bootstrap samples.

ii. Repeat part (i) for the survival times for a similar set of 34 patients that did not receive heart transplants these data are posted as heart2.txt.

iii. Use bootstrap methods to test the null hypothesis that the median survival time for patients who receive heart transplants is equal to the median survival time for similar patients who do not receive heart transplants. State your conclusion.

(b) **Chemical Process**
The yield of a chemical process depends on temperature X_1 and pressure X_2. In one study, process yield Y was measured for $n = 18$ runs of the process under various temperature and pressure conditions. The data are posted in the file pnonlin.txt with one line of data for each run.

Consider the following model you fit to the chemical process data:

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} + \epsilon$$

Use bootstrap methods to find 95% confidence intervals for β_0, β_1, and β_2. Use 1000 bootstrap samples.

26 Failure of O-Rings

The file challenger.txt contains information on 23 (out of 24) pre-Challenger space shuttle flights. (On one flight, the solid rocket motors were lost at sea and so no data are available.) Provided are launch temperatures, T (in Fahrenheit), and a 0-1 response, Y, indicating whether there was postlaunch evidence of a field joint primary O-ring incident. (O-ring failure was apparently responsible for the tragedy.) $Y = 1$ indicates that at least one of the 6 primary O-rings showed evidence of erosion.

(a) Draw a scatterplot of O-ring failure versus temperature. Summarize the relationship between these variables.

(b) Assume, O-ring failures can be described by a binomial distribution $B_{1,p}$. Using `glm` fit an appropriate generalized linear model, using a logit link.

Notice that the case $\beta_1 < 0$ is the case where low temperature launches are more dangerous than warm day launches. NASA managers ordered the launch after arguing that these and other data data showed no relationship between temperature and O-ring failure. Was their claim correct? Explain.

(c) `fit` and `se.fit` of your fitted object contains estimated means and corresponding standard errors.

Plot estimated means versus the temperature.

Connect those with line segments to get a rough plot of the estimated relationship. Plot "2 standard error" bands around that response function as a rough indication of the precision with which the relationship between and tp could be known from the pre-Challenger data.

The temperature at Cape Canaveral for the last Challenger launch was 31 F. What does your analysis here say might have been expected in terms of O-ring performance at that temperature? Use `predict.glm` to get ab estimated 31 F mean and standard error.