Statistics GIDP
Ph.D. Qualifying Exam
Methodology
January 10, 9:00am-1:00pm

Instructions: Put your ID (not name) on each sheet. Complete exactly 5 of 6 problems; turn in only those sheets you wish to have graded. Each question, but not necessarily each part, is equally weighted. Provide answers on the supplied pads of paper and/or use a Microsoft word document or equivalent to report your software code and outputs. Number each problem. *You may turn in only one electronic document. Embed relevant code and output/graphics into your word document.* Write on only one side of each sheet if you use paper. You may use the computer and/or a calculator. Stay calm and do your best. Good luck!

1. A manufacturer suspects that the batches of raw material furnished by her supplier differ significantly in antibiotic content. There are a large number of batches currently in the warehouse. Five of these are randomly selected for study. A biologist makes five determinations on each batch and obtains the following data:

Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
23.46	23.59	23.51	23.28	23.29
23.48	23.46	23.64	23.40	23.46
23.56	23.42	23.46	23.37	23.37
23.39	23.49	23.52	23.46	23.32
23.40	23.50	23.49	23.39	23.38

A portion of pertinent computer output follows:

```
Analysis of Variance for Content
Source      DF         SS         MS       F      P
Batch        4   0.096976   0.024244    5.54  0.004
Error       20   0.087600   0.004380
Total       24   0.184576
```

(a) What type of design/experiment is this? Why?
(b) State the statistical model and corresponding assumptions.
(c) Write the hypothesis in mathematical notation for testing the equality of the batches.
(d) Are the batches equal in output? Use $\alpha = 0.05$.
(e) Estimate the variability between batches.
(f) Estimate the experimental error variance.

2. A consumer products company relies on direct mail marketing pieces as a major component of its advertising campaigns. The company has three different designs for a new brochure and wants to evaluate their effectiveness, as there are substantial differences in costs between the three designs. The company decides to test the three designs by mailing 5,000 samples of each to potential customers in four different regions of the country. Note that there may be regional differences in the customer base. The number of responses to each mailing is shown below (find the dataset in the file "**brochure.csv**").

	Region			
Design	NE	NW	SE	SW
1	225	350	220	375
2	400	525	390	550
3	275	340	200	320

(a) What design is this? Why?

(b) Write the statistical model and the corresponding assumptions.

(c) Analyze the dataset and draw a conclusion at $\alpha = 0.05$.

(d) Can you check whether there exists an interaction between the Design factor and the Region factor? If no, explain why not. If yes, write your statistical model, state the hypotheses, analyze the data, and report your finding

(e) Attach all SAS/R code.

3. An engineer is interested in the effects of cutting speed (A), tool geometry (B), and cutting angle on the life (in hours) of a machine tool. Two levels of each factor are chosen, and three replicates are run for each combination. The results are as follows (find the dataset in the file "**life.csv**")

			Treatment	Replicate		
A	*B*	*C*	Combination	I	II	III
-	-	-	(1)	22	31	25
+	-	-	*a*	32	43	29
-	+	-	*b*	35	34	50
+	+	-	*ab*	55	47	46
-	-	+	*c*	44	45	38
+	-	+	*ac*	40	37	36
-	+	+	*bc*	60	50	54
+	+	+	*abc*	39	41	47

(a) What design is this?

(b) Estimate the factor effects. Which effects appear to be large?

(c) Use the analysis of variance to confirm your informal conclusions from part (b).

(d) Write down a regression model for predicting tool life (in hours) based on the results of this experiment.

(e) What levels of A, B, and C would you recommend using? Justify your answer.

(f) Attach your SAS/R code.

4. A major state university took a sample of n = 705 students and collected the outcome variable Y = {Freshman Grade Point Average}, along with predictor variables X_1 = {High school class rank (as a percentile; higher percentiles indicate higher rank)} and X_2 = {ACT score}. It was felt that the interaction cross-product $X_3 = X_1X_2$ was likely to also be an important predictor. The data are found in the file "**college.csv**."

(a) Examine the two predictor variables X_1 and X_2 to determine if multicollinearity is a concern with these data.

(b) To assuage any concerns over multicollinearity, apply a ridge regression analysis to these data. Begin with a ridge trace plot and use it to identify a plausible value for the ridge tuning parameter *c*.
[*Hint*: if your plot covers sufficient values of *c*, the stock **`traceplot()`** function in the R *genridge* package will draw vertical lines at plausible values, known as the modified Hoerl-Kennard-Baldwin (HKB) and Lawless-Wang (LW) estimators. Decide if either could be a useful choice here and if so, select it.]

(c) Using the tuning parameter, *c*, selected in part (b), calculate the ridge-regression predicted values. From these find the raw residuals and construct a residual plot. Comment on any patterns that may appear.

5. Consider the simple linear regression model: $Y_i \sim$ indep. $N(\beta_0+\beta_1 X_i, \sigma^2)$, $i = 1, ..., n$, where X is a non-stochastic predictor variable. Suppose interest exists in estimating the X value(s) at which $E[Y] = 0$. Let this target parameter be ξ.

(a) What is ξ in terms of the original regression parameters?

(b) State the maximum likelihood estimator (MLE) of ξ. Call this $\hat{\xi}$.

(c) Recall from Section 5.5.4 in Casella & Berger that the *Delta Method* can be used to determine the asymptotic features of a function of random variables. In particular, for a bivariate random vector $[U_1\ U_2]^T$ and a differentiable function $g(u_1,u_2)$, where $E[U_j] = \theta_j$ $(j = 1,2)$,

$E[g(U_1,U_2)] \approx g(\theta_1,\theta_2) + \sum_{j=1}^{2} \frac{\partial}{\partial\theta_j} g(\theta_1,\theta_2)\, E(U_i - \theta_i)$

and

$\mathrm{Var}[g(U_1,U_2)] \approx \sum_{j=1}^{2} \left\{\frac{\partial}{\partial\theta_j} g(\theta_1,\theta_2)\right\}^2 \mathrm{Var}(U_i) + 2\left\{\frac{\partial}{\partial\theta_1} g(\theta_1,\theta_2)\right\}\left\{\frac{\partial}{\partial\theta_2} g(\theta_1,\theta_2)\right\}\mathrm{Cov}(U_1,U_2)$

From these results:
(i) show that $E[\hat{\xi}] \approx \xi$, and
(ii) find an approximation for $\mathrm{Var}[\hat{\xi}]$ under this regression setting.

6. Assume a simple linear regression-through-the-origin model: $Y_i \sim$ indep. $N(\beta_1 X_i, \sigma^2)$, $i = 1, \ldots, n$. Equation (4.14) in Kutner et al. shows that the maximum likelihood estimator (MLE) of β_1 is

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{j=1}^{n} X_j^2}$$

with standard error

$$s\{b_1\} = \frac{\sqrt{MSE}}{\sum_{k=1}^{n} X_k^2},$$

where $MSE = \sum_{i=1}^{n} (Y_i - b_1 X_i)^2/(n-2)$ is an unbiased estimator of σ^2. Starting from the basic principles of the Working-Hotelling-Scheffé (WHS) construction, derive the $1-\alpha$ WHS simultaneous confidence band for the mean response $E[Y_i]$ with this model. What is the "shape" of this band?

Statistics GIDP
Ph.D. Qualifying Exam
Methodology
January 10, 9:00am-1:00pm

Instructions: Put your ID (not name) on each sheet. Complete exactly 5 of 6 problems; turn in only those sheets you wish to have graded. Each question, but not necessarily each part, is equally weighted. Provide answers on the supplied pads of paper and/or use a Microsoft word document or equivalent to report your software code and outputs. Number each problem. *You may turn in only one electronic document. Embed relevant code and output/graphics into your word document.* Write on only one side of each sheet if you use paper. You may use the computer and/or a calculator. Stay calm and do your best. Good luck!

1. A manufacturer suspects that the batches of raw material furnished by her supplier differ significantly in antibiotic content. There are a large number of batches currently in the warehouse. Five of these are randomly selected for study. A biologist makes five determinations on each batch and obtains the following data:

Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
23.46	23.59	23.51	23.28	23.29
23.48	23.46	23.64	23.40	23.46
23.56	23.42	23.46	23.37	23.37
23.39	23.49	23.52	23.46	23.32
23.40	23.50	23.49	23.39	23.38

A portion of pertinent computer output follows:

```
Analysis of Variance for Content

Source      DF         SS         MS       F      P
Batch        4   0.096976   0.024244    5.54  0.004
Error       20   0.087600   0.004380
Total       24   0.184576
```

(a) What type of design/experiment is this? Why?
This is a one-factor random (effect) design. We are interested in the entire population of batches and batches are randomly chosen.

(b) State the statistical model and corresponding assumptions.
We assume $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, i=1, …, 5, j=1, …, 5, with $\tau_{ij} \sim N(0,\sigma_\tau^2)$ independent of $\varepsilon_{ij} \sim N(0,\sigma^2)$.

(c) Write the hypothesis in mathematical notation for testing the equality of the batches.

H_0: $\sigma_\tau^2 = 0$

H_1: $\sigma_\tau^2 > 0$

(d) Are the batches equal in output? Use $\alpha = 0.05$.

Since the P-value is given as P = 0.004, and this is less than 0.05, we reject the null hypothesis and conclude that the batches differ significantly.

(e) Estimate the variability between batches.

$$\hat{\sigma}_\tau^2 = \frac{MS_{trt} - MSE}{n} = \frac{0.024244 - 0.004380}{5} = 0.003972$$

(f) Estimate the experimental error variance.

$\hat{\sigma}^2 = 0.004380$

2. A consumer products company relies on direct mail marketing pieces as a major component of its advertising campaigns. The company has three different designs for a new brochure and wants to evaluate their effectiveness, as there are substantial differences in costs between the three designs. The company decides to test the three designs by mailing 5,000 samples of each to potential customers in four different regions of the country. Note that there may be regional differences in the customer base. The number of responses to each mailing is shown below (find the dataset in the file "**brochure.csv**").

	Region			
Design	NE	NW	SE	SW
1	225	350	220	375
2	400	525	390	550
3	275	340	200	320

(a) What design is this? Why?

This is a RCBD design (randomized complete block design) with Region as a blocking factor.

(b) Write the statistical model and the corresponding assumptions.

$y_{ij} = \mu + \tau_i + \alpha_j + \varepsilon_{ij}$, i=1,2,3, j=1, …, 4, and where $\sum\alpha_j = 0$, $\sum\tau_i = 0$, and $\varepsilon_{ij} \sim$ i.i.d.$N(0,\sigma^2)$.

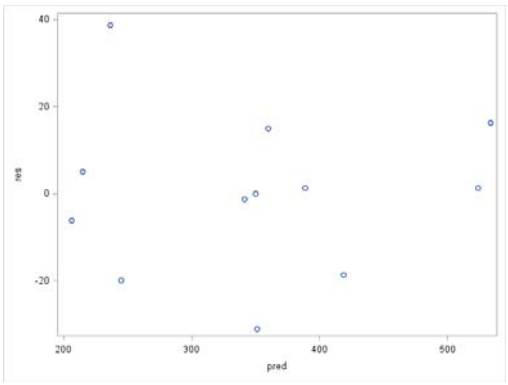
(c) Analyze the dataset and draw a conclusion at $\alpha = 0.05$.

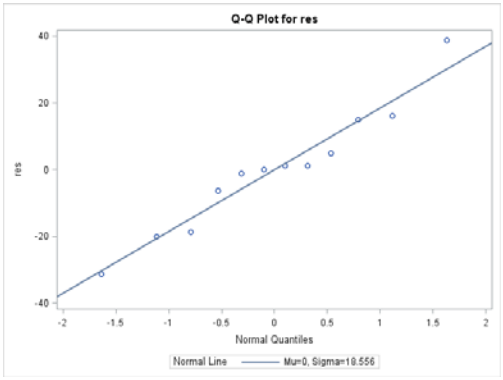
From the SAS output (below) the "design" factor and "region" block factor are both significant at $\alpha = 0.05$.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	133137.5000	26627.5000	42.18	0.0001
Error	6	3787.5000	631.2500		
Corrected Total	11	136925.0000			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
design	2	84762.50000	42381.25000	67.14	<.0001
region	3	48375.00000	16125.00000	25.54	0.0008

The raw residual plot and QQ-plot (below) do not suggest any unusual patterns. Thus the normality assumption is reasonable.





(d) Can you check whether there exists an interaction between the Design factor and the Region factor? If no, explain why not. If yes, write your statistical model, state the hypotheses, analyze the data, and report your finding

Yes. Use Tukey's 1-degree of freedom test, with model $y_{ij} = \mu + \tau_i + \alpha_j + \gamma\tau_i\alpha_j + \varepsilon_{ij}$. The hypotheses are

H_0: $\gamma = 0$

H_1: $\gamma \neq 0$

The following SAS output shows that the interaction is not significant (P-value = 0.4741). Therefore the additive model from above is appropriate.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
design	2	821.240880	410.620440	0.61	0.5808
region	3	1099.093874	366.364625	0.54	0.6745
q	1	404.852146	404.852146	0.60	0.4741

(e) Attach all SAS/R code.

```
data brochure ;
input design region $ response;
datalines;

1 NE    225
1 NW    350
1 SE    220
1 SW    375
2 NE    400
2 NW    525
2 SE    390
2 SW    550
3 NE    275
3 NW    340
3 SE    200
3 SW    320
;


proc glm data=brochure;
class design region;
model response=design region;
```

```
output out=myout p=pred r=res;
run;

proc sgplot data=myout;
scatter x=pred y=res;
run;

proc univariate data=myout normal;
var res;
qqplot res/normal(mu=est sigma=est color=red L=1);
run;


data two;
set myout;
q=pred*pred;


proc glm data=two;
class design region;
model response=design region q/ss3;
run;
```

3. An engineer is interested in the effects of cutting speed (A), tool geometry (B), and cutting angle on the life (in hours) of a machine tool. Two levels of each factor are chosen, and three replicates are run for each combination. The results are as follows (find the dataset in the file "**life.csv**")

			Treatment	Replicate		
A	*B*	*C*	Combination	I	II	III
-	-	-	(1)	22	31	25
+	-	-	*a*	32	43	29
-	+	-	*b*	35	34	50
+	+	-	*ab*	55	47	46
-	-	+	*c*	44	45	38
+	-	+	*ac*	40	37	36
-	+	+	*bc*	60	50	54
+	+	+	*abc*	39	41	47

(a) What design is this?
A 2^3 factorial design.

(b) Estimate the factor effects. Which effects appear to be large?
It seems that the factors B and C and interaction AC are large, they are 11.33, 6.83 and -8.83 respectively (just double the coefficients below obtained through a regression model):

Intercept	**A**	**B**	**C**	**AB**	**AC**	**BC**	**ABC**
40.8333	0.16667	5.66667	3.41667	-0.83333	-4.41667	-1.41667	-1.08333

(c) Use the analysis of variance to confirm your informal conclusions from part (b).
The ANOVA table below confirms the significance of factors B, C and the interaction AC: their P-values are 0.001, 0.0077, and 0.0012, respectively

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	1	0.6666667	0.6666667	0.02	0.8837
B	1	770.6666667	770.6666667	25.55	0.0001
A*B	1	16.6666667	16.6666667	0.55	0.4681
C	1	280.1666667	280.1666667	9.29	0.0077
A*C	1	468.1666667	468.1666667	15.52	0.0012
B*C	1	48.1666667	48.1666667	1.60	0.2245
A*B*C	1	28.1666667	28.1666667	0.93	0.3483

(d) Write down a regression model for predicting tool life (in hours) based on the results of this experiment.
Since the interaction AC is significant we need to include Factor A into the model, though itself, it is not significant. Then, the prediction equation becomes
$\text{Life} = 40.8333 + 0.16667X_A + 5.66667X_B + 3.41667X_C - 4.41667X_{AC}$

(e) What levels of A, B, and C would you recommend using? Justify your answer.

For a maximum of the tool life, we would recommend high levels for factors B and C, and low level for factor A. This is because

(i) all three main factors have a positive effect,

(ii) the interaction AC has a negative effect, and

(iii) the coefficient of factor C is larger than that for factor A.

(f) Attach your SAS/R code.

```
data Q3 ;
input A B C rep life;
datalines;
-1   -1 -1 1 22
-1   -1 -1 2 31
-1   -1 -1 3 25
 1   -1 -1 1 32
 1   -1 -1 2 43
 1   -1 -1 3 29
-1    1 -1 1 35
-1    1 -1 2 34
-1    1 -1 3 50
 1    1 -1 1 55
 1    1 -1 2 47
 1    1 -1 3 46
-1   -1    1 1 44
-1   -1    1 2 45
-1   -1    1 3 38
 1   -1  1 1 40
 1   -1  1 2 37
 1   -1  1 3 36
-1    1    1 1 60
-1    1    1 2 50
-1    1    1 3 54
 1    1    1 1 39
 1    1    1 2 41
 1    1    1 3 47
;

data inter;
set Q3;
AB=A*B; AC=A*C; BC=B*C; ABC=A*B*C;
run;

proc print data=inter;
run;

proc reg outest=effect data=inter;
model life=A B C AB AC BC ABC;
run;
/* or from the estimates in the glm proc */
proc glm data=inter;
class A B C AB AC BC ABC;
model life=A B C AB AC BC ABC;
```

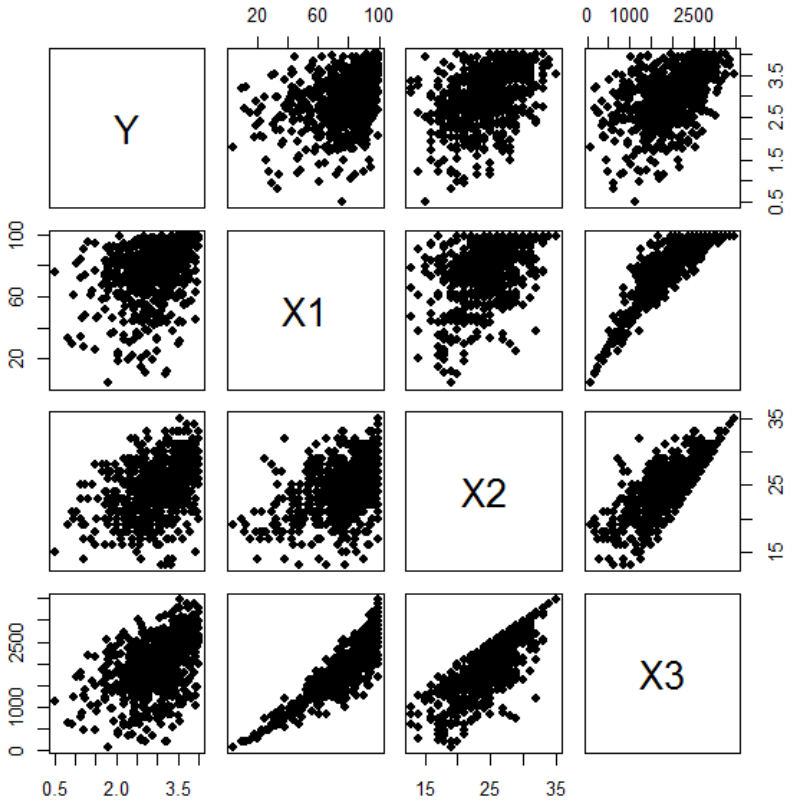
```
estimate 'A' A -1 1;
estimate 'B' B -1 1;
estimate 'C' C -1 1;
estimate 'AB' AB -1 1;
estimate 'AC' AC -1 1;
estimate 'BC' BC -1 1;
estimate 'ABC' ABC -1 1;
run;

/* or through a more complicated procedure (skip) */
proc print data=effect;
run;

proc glm data=Q3;
class A B C;
model life=A|B|C;
output out=myout p=pred r=res;
run;
```

4. A major state university took a sample of n = 705 students and collected the outcome variable Y = {Freshman Grade Point Average}, along with predictor variables X_1 = {High school class rank (as a percentile; higher percentiles indicate higher rank)} and X_2 = {ACT score}. It was felt that the interaction cross-product $X_3 = X_1X_2$ was likely to also be an important predictor. The data are found in the file "**college.csv**."

(a) Examine the two predictor variables X_1 and X_2 to determine if multicollinearity is a concern with these data.

Always plot the data! Sample R code:

```
college.df = read.csv( file.choose() )
attach( college.df )
X1 = class.rank; X2 = ACT; X3=X1*X2
Y = GPA
pairs( Y~X1+X2+X3, pch=19 )
```



We see some multicollinearity may exist between X1 & X3 and between X2 & X3 (no surprise, in either case). A further check is available via the sample correlations among the predictor variables:

```
cor( cbind(X1,X2,X3) )
```

```
          X1        X2        X3
X1 1.0000000 0.4425075 0.8883073
X2 0.4425075 1.0000000 0.7890032
X3 0.8883073 0.7890032 1.0000000
```

where we see large correlations with the interaction term X3.

While informative, however, all the above analysis is subjective. A final, objective, definitive check employs the VIFs:

```
require( car )
vif( lm(Y ~ X1+X2+X3) )
      X1         X2         X3
29.35675   16.40282   62.54291
```

where clearly $\max\{VIF_k\} > 10$. (Also, the mean VIF is 36.1008, which is above the target bound of 6.) Multicollinearity is a clear problem with these data.

(b) To assuage any concerns over multicollinearity, apply a ridge regression analysis to these data. Begin with a ridge trace plot and use it to identify a plausible value for the ridge tuning parameter *c*.

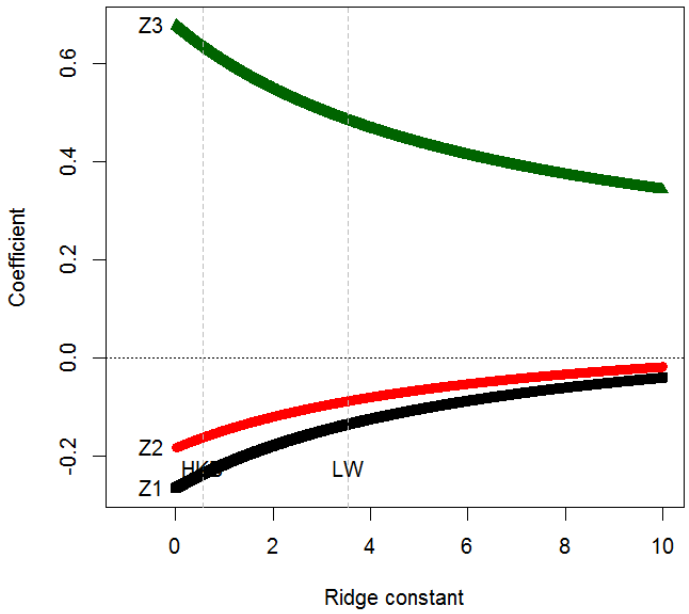
[*Hint*: if your plot covers sufficient values of *c*, the stock **`traceplot()`** function in the R *genridge* package will draw vertical lines at plausible values, known as the modified Hoerl-Kennard-Baldwin (HKB) and Lawless-Wang (LW) estimators. Decide if either could be a useful choice here and if so, select it.]

For a ridge regression fit, start by centering the response variable to $U_i = Y_i - \bar{Y}$ and standardizing the predictors to Z_1, Z_2, and Z_3, each with zero mean and unit variance:

```
U = scale( Y, scale=F )
Z1 = scale( X1 )
Z2 = scale( X2 )
Z3 = scale( X3 )
```

Now load the *genridge* package, choose a range for *c* (here $0 < c < 10$) and construct the traceplot using the **`ridge()`** function:

```
require( genridge )
c = seq( from=.01,to=10,by=.01 )
traceplot( ridge(U~Z1+Z2+Z3, lambda=c) )
```



The stock traceplot suggests that the Lawless-Wang (LW) estimate for c is far enough along to where the traces start to flatten and stabilize. So choose c_{LW}. The exact value can be found as the **`$kLW`** attribute in the **`ridge()`** object:

```
print( ridge(U~Z1+Z2+Z3,lambda=c)$kLW )
[1] 3.531847
```

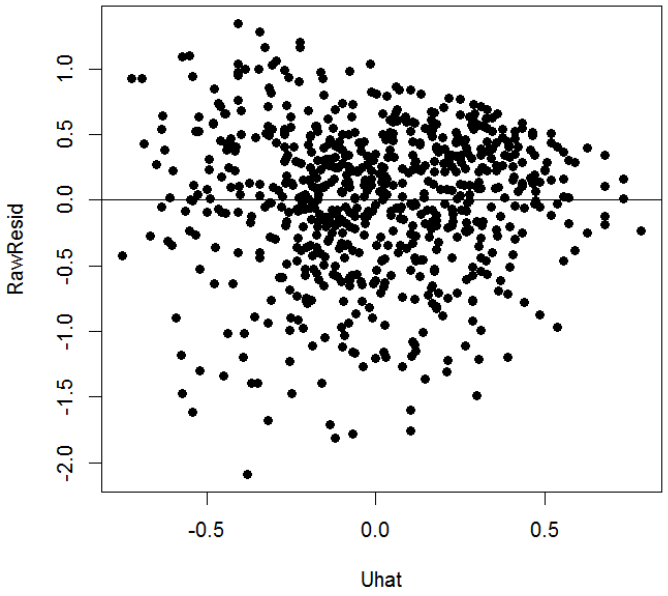
(c) Using the tuning parameter, c, selected in part (b), calculate the ridge-regression predicted values. From these find the raw residuals and construct a residual plot. Comment on any patterns that may appear.

The LW estimate is $c = 3.531847$. The final ridge regression object is then constructed via:

```
cLW = ridge(U~Z1+Z2+Z3,lambda=c)$kLW
college.ridge = ridge( U~Z1+Z2+Z3, lambda=cLW )
```

Fitted values and residuals can be found via direct calculation. First identify the ridge regr. coefficients (use the **`$coef`** attribute in the **`ridge()`** object, but select its components carefully!) and then proceed from there:

```
Zmtx = as.matrix( cbind(Z1,Z2,Z3) )
Uhat = Zmtx %*% college.ridge$coef[,1:3]
RawResid = U - Uhat
plot( RawResid ~ Uhat, pch=19 ); abline( h=0 )
```



The plot shows possible decreasing variability with increasing response (i.e. variance heterogeneity), or perhaps a decreasing trend suggestive of a possible hidden/unidentified predictor variable. Either way, deeper diagnostic study of these data is warranted.

5. Consider the simple linear regression model: $Y_i \sim$ indep. $N(\beta_0+\beta_1 X_i, \sigma^2)$, $i = 1, ..., n$, where X is a non-stochastic predictor variable. Suppose interest exists in estimating the X value(s) at which $E[Y] = 0$. Let this target parameter be ξ.

(a) What is ξ in terms of the original regression parameters?

This is essentially an inverse regression problem. We have $E[Y] = \beta_0+\beta_1 X$.
Clearly, at $E[Y] = 0$ we have $0 = \beta_0+\beta_1 X$, so solving for X produces $\xi = -\beta_0/\beta_1$.

(b) State the maximum likelihood estimator (MLE) of ξ. Call this $\hat{\xi}$.

Under ML invariance, $\hat{\xi} = -\hat{\beta}_0/\hat{\beta}_1$, where $\hat{\beta}_j$ is the ML estimator of β_j (j=0,1).

(c) Recall from Section 5.5.4 in Casella & Berger that the *Delta Method* can be used to determine the asymptotic features of a function of random variables. In particular, for a bivariate random vector $[U_1\ U_2]^T$ and a differentiable function $g(u_1,u_2)$, where $E[U_j] = \theta_j$ (j = 1,2),

$E[g(U_1,U_2)] \approx g(\theta_1,\theta_2) + \sum_{j=1}^{2} \frac{\partial}{\partial\theta_j} g(\theta_1,\theta_2)\, E(U_i - \theta_i)$

and

$\mathrm{Var}[g(U_1,U_2)] \approx \sum_{j=1}^{2} \left\{\frac{\partial}{\partial\theta_j} g(\theta_1,\theta_2)\right\}^2 \mathrm{Var}(U_i) + 2\left\{\frac{\partial}{\partial\theta_1} g(\theta_1,\theta_2)\right\}\left\{\frac{\partial}{\partial\theta_2} g(\theta_1,\theta_2)\right\}\mathrm{Cov}(U_1,U_2)$

From these results:
(i) show that $E[\hat{\xi}] \approx \xi$, and
(ii) find an approximation for $\mathrm{Var}[\hat{\xi}]$ under this regression setting.

Let $g(\beta_1,\beta_2) = \xi = -\beta_0/\beta_1$. We know that the MLEs for β_j are unbiased such that $E[\hat{\beta}_j] = \beta_j$ (j=0,1). Then from the Delta Method we see

(i) $E[\hat{\xi}] = E[-\hat{\beta}_0/\hat{\beta}_1] \approx g(\beta_1,\beta_2) + \sum_{j=0}^{1} \frac{\partial}{\partial\beta_i} g(\beta_0,\beta_1)\, E(\hat{\beta}_i - \beta_i) = -\beta_0/\beta_1 + \sum_{j=0}^{1} \frac{\partial}{\partial\beta_i} g(\beta_0,\beta_1)(0)$

$= -\beta_0/\beta_1 = \xi.$

(This is actually true in general: under the conditions of the Delta Method, $E[\hat{\theta}_0/\hat{\theta}_1] \approx \theta_0/\theta_1$. See Example 5.5.27 in Casella & Berger.)

(ii) From the Delta Method,

$$\mathrm{Var}[\hat{\xi}] = \mathrm{Var}[-\hat{\beta}_0/\hat{\beta}_1] = (-1)^2\mathrm{Var}[\hat{\beta}_0/\hat{\beta}_1] \approx \left(\frac{\beta_0}{\beta_1}\right)^2 \left(\frac{\mathrm{Var}[\hat{\beta}_0]}{\beta_0^2} + \frac{\mathrm{Var}[\hat{\beta}_1]}{\beta_1^2} - 2\frac{\mathrm{Cov}[\hat{\beta}_0;\hat{\beta}_1]}{\beta_0\beta_1}\right).$$

Now let $SS_x = \sum(X_i - \bar{X})^2$. We know $\mathrm{Var}[\hat{\beta}_0] = \sigma^2(n^{-1}SS_x + \bar{X}^2)/SS_x$, $\mathrm{Var}[\hat{\beta}_1] = \sigma^2/SS_x$, and $\mathrm{Cov}[\hat{\beta}_0,\hat{\beta}_1] = -\sigma^2\bar{X}/SS_x$. This allows for some simplification:

$$\mathrm{Var}[\hat{\xi}] \approx \frac{\sigma^2}{\beta_1^2 SS_x}\left(\frac{SS_x}{n} + \bar{X}^2 + \frac{\beta_0^2}{\beta_1^2} + 2\frac{\bar{X}\beta_0}{\beta_1}\right) = \frac{\sigma^2}{\beta_1^2 SS_x}\left(\frac{SS_x}{n} + \bar{X}^2 + \xi^2 - 2\bar{X}\xi\right).$$

6. Assume a simple linear regression-through-the-origin model: $Y_i \sim$ indep. $N(\beta_1 X_i, \sigma^2)$, $i = 1, \ldots, n$. Equation (4.14) in Kutner et al. shows that the maximum likelihood estimator (MLE) of β_1 is

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{j=1}^{n} X_j^2}$$

with standard error

$$s\{b_1\} = \frac{\sqrt{MSE}}{\sum_{k=1}^{n} X_k^2},$$

where $MSE = \sum_{i=1}^{n} (Y_i - b_1 X_i)^2/(n-2)$ is an unbiased estimator of σ^2. Starting from the basic principles of the Working-Hotelling-Scheffé (WHS) construction, derive the $1-\alpha$ WHS simultaneous confidence band for the mean response $E[Y_i]$ with this model. What is the "shape" of this band?

The WHS band will take the form $\hat{Y}_h \pm W\, s\{\hat{Y}_h\}$, where the WHS critical point is based on $W^2 = pF(1-\alpha; p, n-p)$. Here $p=1$, while $\hat{Y}_h = b_1 X_h$ with standard error

$$s\{\hat{Y}_h\} = \sqrt{Var\{\hat{Y}_h\}} = \sqrt{X_h^2 Var\{b_1\}} = |X_h|\, s\{b_1\}$$

for <u>all</u> X_h. Thus the band will be $b_1 X_h \pm W\, |X_h|\, s\{b_1\}$, where $W^2 = F(1-\alpha;1,n-1)$. But recall that $F(1-\alpha;1,n-1) = t^2(1-\{\tfrac{1}{2}\alpha\};n-1)$, so this simplifies to

$$b_1 X_h \pm t(1-\{\tfrac{1}{2}\alpha\};n-1)\, s\{b_1\}\, |X_h|,$$

for <u>all</u> X_h. Note the form of this "band": a pair of straight lines that intersect at the origin.