Statistics GIDP
Ph.D. Qualifying Exam
Methodology
May 26
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/graphs 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. Three ovens in a metal working shop are used to heat metal specimens. All the ovens are supposed to operate at the same temperature, although it is suspected that this may not be true. The data collected (dataset: "oven.csv") are as follows:

Oven	Temperature					
1	491.50	498.30	498.10	493.50	493.60	
2	488.50	484.65	479.90	477.35		
3	490.10	484.80	488.25	473.00	471.85	478.65

(a) What type of design is this?

Completely randomized design.

(b) State the statistical model and associated assumptions.

$y_{ij} = \mu + \tau_i + \epsilon_{ij}$ and $\sum \tau_i = 0, \epsilon_{ij} \sim N(0, \sigma^2)$

(c) Construct the ANOVA table. Include the value of the F-statistic and its p-value.

Source	DF	SS	MS	F	P
Oven	2	594.53	297.27	8.62	0.005
Error	12	413.81	34.48		
Total	14	1008.34			

(d) For each oven, construct a point-wise 95% confidence interval for its mean temperature.

$$\bar{y}_{i.} - t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}} \le \mu_i \le \bar{y}_{i.} + t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}}$$

oven 1: 495+/- 2.179*sqrt(34.48/5)=[489.2779, 500.7221]
oven 2: 482.6+/- 2.179*sqrt(34.48/4)=[476.203, 488.997]
oven 3: 481.1083+/- 2.179*sqrt(34.48/6)=[475.8848, 486.3318]

(e) View the three ovens as a random sample from a population of metal shop ovens. Would there be any change in terms of the statistical model with respect to assumptions, hypothesis, and conclusion of homogeneity across oven temperatures? Explain.

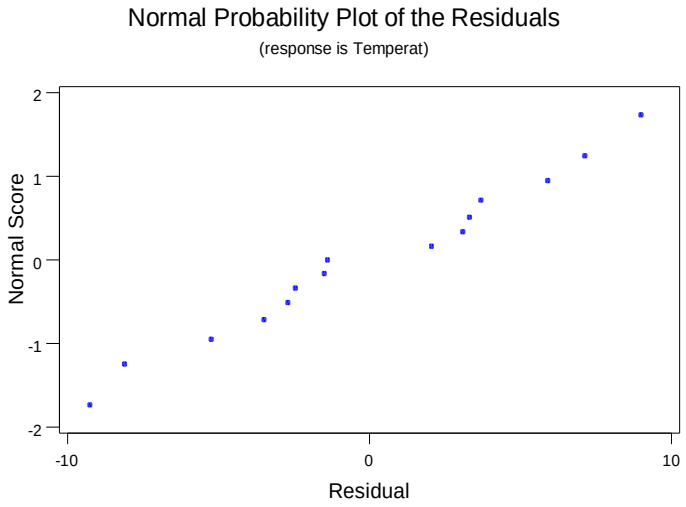
The model and assumptions:
$y_{ij} = \mu + \tau_i + \epsilon_{ij}$ and $\tau_i \sim N(0, \sigma_\tau^2), \epsilon_{ij} \sim N(0, \sigma^2)$
The hypotheses test becomes: $H0$: $\sigma_\tau^2 = 0$ vs. $H1$: $\sigma_\tau^2 > 0$
But the final conclusion of significance stays same since the F-value is same.

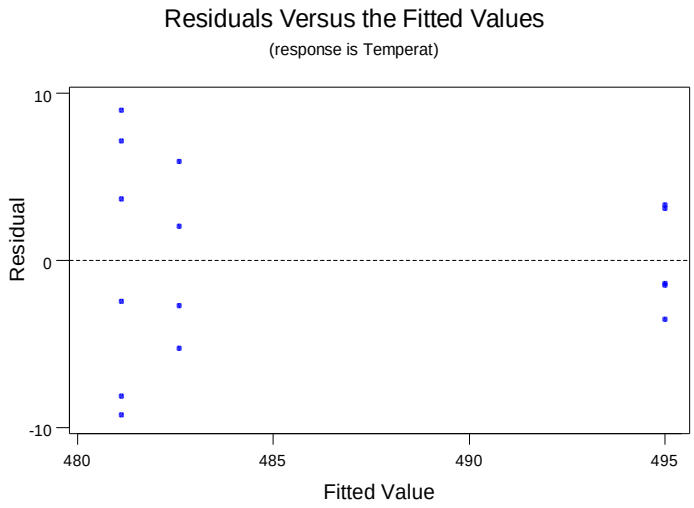
(f) Following (e), estimate the experimental error and the parameter(s) in the hypothesis.

Components of variance: $\hat{\sigma}_\tau^2 = 53.3$ and $\hat{\sigma}^2 = 34.8$

(g) Analyze the residuals from this experiment. Draw conclusions about model adequacy.

There is a funnel shaped appearance in the plot of residuals versus predicted value indicating a possible non-constant variance.





2. A soybean trial involved 30 varieties grown at 10 locations in Indiana. There were 3 replicates of each variety at each location for a total of 30 x 10 x 3 = 900 observations. The researchers considered both variety and location to be random factors in their analysis of yield. Use the SAS output below to answer the following questions.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	299	73900.66757	247.15942	56.64	<.0001
Error	600	2618.21805	4.36370		
Total	899	76518.88562			

	R-Square	Coeff Var	Root MSE	yield Mean
	0.965783	5.470280	2.088946	38.18719

Source	DF	Type III SS	Mean Square	F Value	Pr > F
var	29	10329.35772	356.18475	81.62	<.0001
loc	9	50083.57842	5564.84205	1275.26	<.0001
loc*var	261	13487.73143	51.67713	11.84	<.0001

(a) What design is this?

Factorial design with random factors

(b) State the statistical model and assumptions.

$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}$ and $\tau_i \sim N(0, \sigma_\tau^2), \beta_j \sim N(0, \sigma_\beta^2)$, $(\tau\beta)_{ij} \sim N(0, \sigma_{\tau\beta}^2), \epsilon_{ijk} \sim N(0, \sigma^2)$

(c) What are the estimated variances?

$$\hat{\sigma}^2 = MSE = 4.3637$$
$$\hat{\sigma}^2_{\tau\beta} = \frac{MMS_{loc*var} - MSE}{3} = 15.7711$$
$$\hat{\sigma}^2_{\tau} = \frac{MS_{var} - MS_{loc*var}}{10 * 3} = 10.15025$$
$$\hat{\sigma}^2_{\beta} = \frac{MS_{loc} - MS_{loc*var}}{30 * 3} = 61.2574$$

(d) What is the F-value for testing significance of variety?

$F = \frac{MS_{var}}{MS_{loc*var}} = 6.8936$

(e) Calculate a 95% confidence interval for the average soybean yield.

Standard error = sqrt(4.3637+15.7711+10.15025+61.2574)/30=0.3189
C.I.=38.1872+/-1.962*0.3189=[37.5615, 38.8129]

3. An article by L.B. Hare ("In the Soup: A Case Study to Identify Contributors to Filling Variability", Journal of Quality Technology, Vol. 20, pp. 36-43) describes a factorial experiment used to study the filling variability of dry soup mix packages. The factors are A = number of mixing ports through which the vegetable oil was added (1, 2), B = temperature surrounding the mixer (cooled, ambient), C = mixing time (60, 80 sec), D = batch weight (1500, 2000 lb), and E = number of days between mixing and packaging (1, 7). The standard deviation of package weight is used as the response variable. 16 runs are performed in this experiment (see the table below) and the resulting data are given in the file "soup.csv".

	A: Mixer Ports	B: Temp	C: Time	D: Batch Weight	E: Delay	y: Std Dev
1	-1	-1	-1	-1	-1	1.13
2	1	-1	-1	-1	1	1.25
3	-1	1	-1	-1	1	0.97
4	1	1	-1	-1	-1	1.70
5	-1	-1	1	-1	1	1.47
6	1	-1	1	-1	-1	1.28
7	-1	1	1	-1	-1	1.18
8	1	1	1	-1	1	0.98
9	-1	-1	-1	1	1	0.78
10	1	-1	-1	1	-1	1.36
11	-1	1	-1	1	-1	1.85
12	1	1	-1	1	1	0.62
13	-1	-1	1	1	-1	1.09
14	1	-1	1	1	1	1.10
15	-1	1	1	1	1	0.76

16	1	1	1	1	-1	2.10

(a) What type of design is used?

Fractional factorial design.

(b) Identify the defining relation and the alias relationships.

The defining relation is $I = -ABCDE$ and the alias relations are:

$A=$	$-BCDE$
$B=$	$-ACDE$
$C=$	$-ABDE$
$D=$	$-ABCE$
$E=$	$-ABCD$
$AB=$	$-CDE$
$AC=$	$-BDE$
$AD=$	$-BCE$
$AE=$	$-BCD$
$BC=$	$-ADE$
$BD=$	$-ACE$
$BE=$	$-ACD$
$CD=$	$-ABE$
$CE=$	$-ABD$
$DE=$	$-ABC$

(c) What is the resolution of this design?

This design is Resolution *V.*

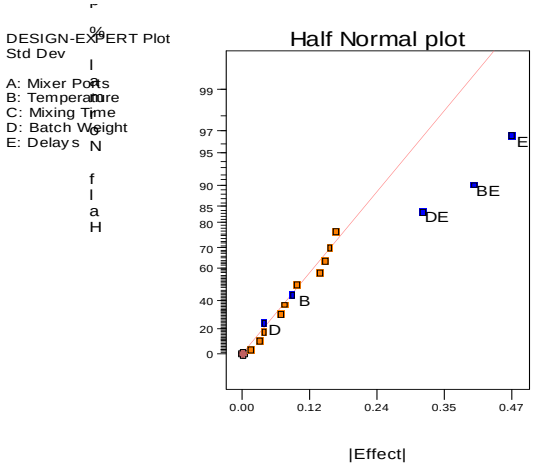
(d) Estimate the factor effects. Which effects are large?

Term	Effect
A	0.145
B	0.0875
C	0.0375
D	-0.0375
E	-0.47
AB	0.015
AC	0.095
AD	0.03
AE	-0.1525
BC	-0.0675
BD	0.1625
BE	-0.405
CD	0.0725
CE	0.135
DE	-0.315

E, BE, and DE seem large.

(e) Use a plot to check the conclusion of (d).

Factor *E* and the two factor interactions *BE* and *DE* appear to be significant in the normal plot.



(f) In order to perform an appropriate statistical analysis to test the hypothesis that the factors identified in part above have a significant effect, which terms should you include in the model, and why?

Besides *E, BE* and *DE*, the factors *B*, and *D* should also be included in the model in order to satisfy the model hierarchy.

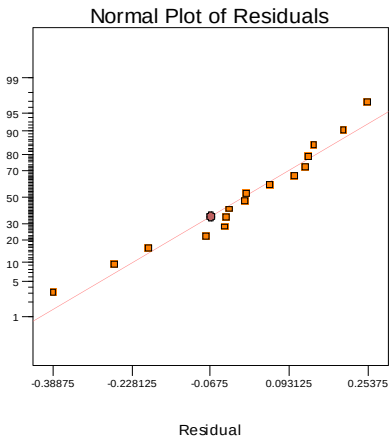
(g) Fit a model that could be used to predict the standard deviation of package weight. Give the estimated prediction equation.

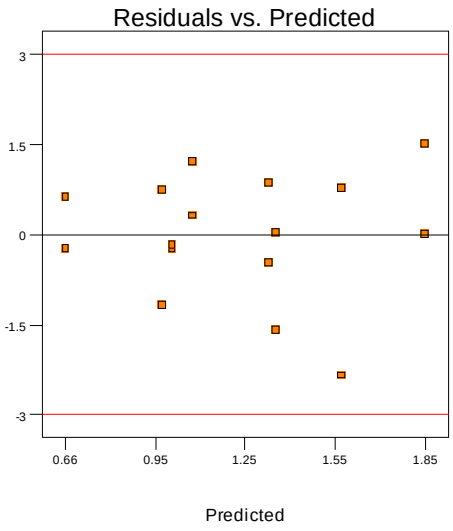
Final Equation in Terms of Coded Factors:

Std Dev	=
+1.23	
+0.044	* B
-0.019	* D
-0.24	* E
-0.20	* B * E
-0.16	* D * E

(h) Does a residual analysis of the model in (g) indicate any problems with the underlying assumptions? Explain.

Often a transformation such as the natural log is required for the standard deviation response; however, the following residuals appear to be acceptable without the transformation.





(i) Make a recommendation about the levels of the factors in this filling process.

The lowest standard deviation can be achieved with the Temperature at ambient, Batch Weight at 2000 lbs, and a Delay of 7 days.

4. The following data represent serum cholesterol (mg/100mℓ) in Midwestern women, along with the age of each woman at the time of the cholesterol testing. (The data are also found in the file "cholesterol.csv").

State = Iowa

age	cholest.	age	cholest.
33	201	54	259
39	182	58	189
41	112	65	249
46	181	71	224
49	121	76	339
52	228		

State = Nebraska

age	cholest.	age	cholest.
18	137	47	196
19	189	51	225
21	191	56	197
30	140	58	262
31	159	58	257
33	177	63	337
42	214	67	356
43	223	70	261
44	190	78	241
44	173		

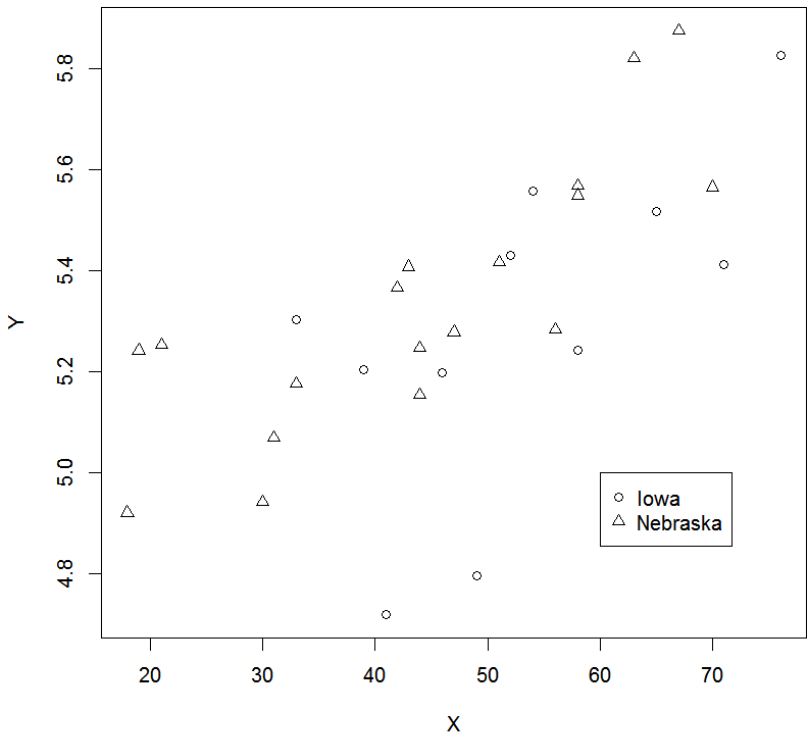
(a) Serum levels such as blood cholesterol are notoriously skewed, so operate with $\log_e$(cholesterol) as your response variable. Test whether log serum cholesterol levels in Nebraska women are significantly different from those in Iowa women. Throughout your analysis, set your pointwise false-positive rate equal to 5%.

(b) Assess the quality of the fit from the eventual model you chose to make the inference in part (a). Are any concerns raised in your analysis?

---

4. (a) Always plot the data! Sample R code:

```
cholesterol.df = read.csv( file.choose() )
attach( cholesterol.df )
X = age; Y = log(cholesterol); State = State
plot( Y ~ X, pch=(as.numeric(State)) )
legend(60,5,legend=c('Iowa','Nebraska'),pch=c(unique(as.numeric(State))))
```



Apparent differences emerge between the states. To continue with the analysis, for an unequal-slopes ANCOVA model:

```
cholestFM.lm = lm( Y ~ X + factor(State) + X*factor(State) )
anova( cholestFM.lm )
```

The output indicates that at $\alpha = 0.05$, no significant difference in slopes is evident ($P = 0.7444 > 0.05$):

```
Analysis of Variance Table
Response: Y
                Df Sum Sq Mean Sq F value    Pr(>F)
X                1 1.0240 1.02401 25.2733 3.473e-05
factor(State)    1 0.1860 0.18600  4.5906   0.04207
X:factor(State)  1 0.0044 0.00440  0.1087   0.74436
Residuals       25 1.0129 0.04052
```

So, re-fit the data with an equal-slopes ANCOVA model use a partial t-test or F-test to assess differences between states:

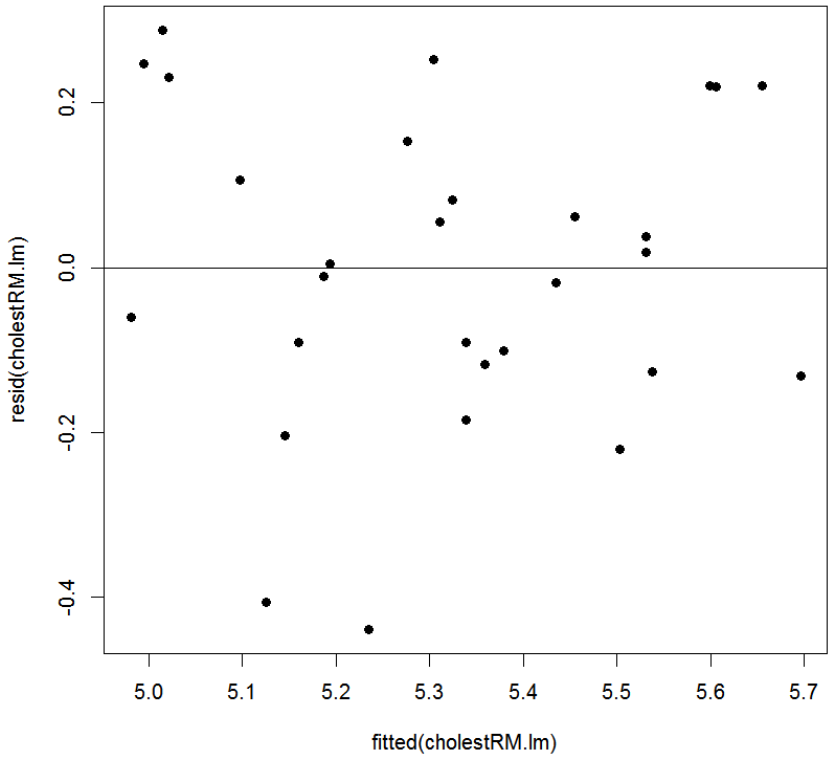
```
cholestRM.lm = lm( Y ~ X + factor(State) )
anova( cholestRM.lm )
```

```
Analysis of Variance Table
Response: Y
                Df Sum Sq Mean Sq F value    Pr(>F)
X                1 1.0240 1.02401 26.1704 2.477e-05
factor(State)    1 0.1860 0.18600  4.7536   0.03848
Residuals       26 1.0173 0.03913
```

The output indicates that at $\alpha = 0.05$, the difference between the two states' cholesterol levels is significant ($P = 0.0385 < 0.05$).

(b) Begin with a residual plot (using the reduced model):

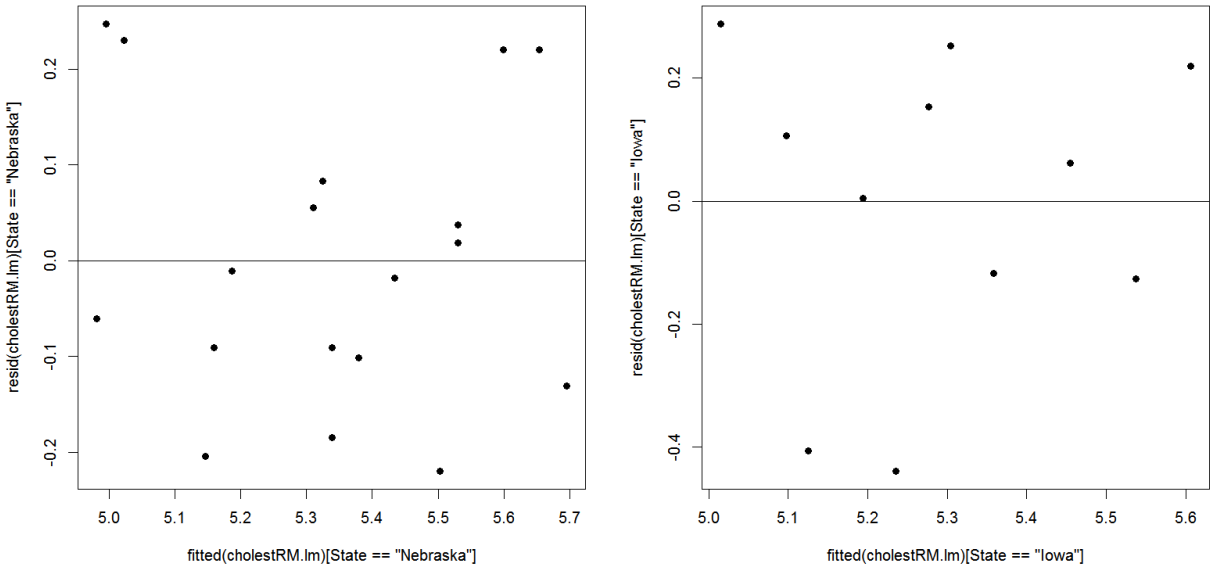
```
plot( resid(cholestRM.lm)~ fitted(cholestRM.lm), pch=19 )
abline( h=0 )
```



No gross concerns are evident. A careful analysis, however, would check the residual pattern for each State separately:

```
par( mfrow=c(1,2) )
plot( resid(cholestRM.lm)[State=='Nebraska'] ~
      fitted(cholestRM.lm)[State=='Nebraska'], pch=19 )
abline( h=0 )
```

```
plot( resid(cholestRM.lm)[State=='Iowa'] ~
      fitted(cholestRM.lm)[State=='Iowa'], pch=19 )
abline( h=0 )
```



The individual patterns are also acceptable. The Nebraska pattern -- left panel above -- presents a slightly curvilinear pattern. Adding a quadratic term for just the Nebraska data does not produce a significant improvement in the fit, however. (Use, e.g.,

```
lm(Y ~ X +
   c(rep(0,length(X[State!="Nebraska"])),I(X[State=='Nebraska']^2)) +
       factor(State))
```

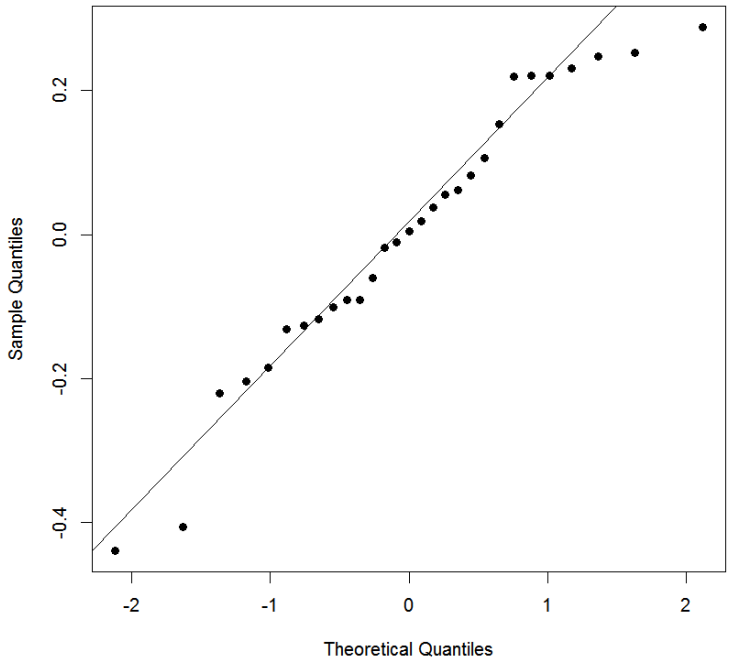
and compare to the larger model in `cholestRM.lm` via, e.g.,

```
anova( cholestRM.lm,cholestRMQ.lm )
```

The consequent improvement in the model fit is insignificant: P = 0.9357; output not shown.)

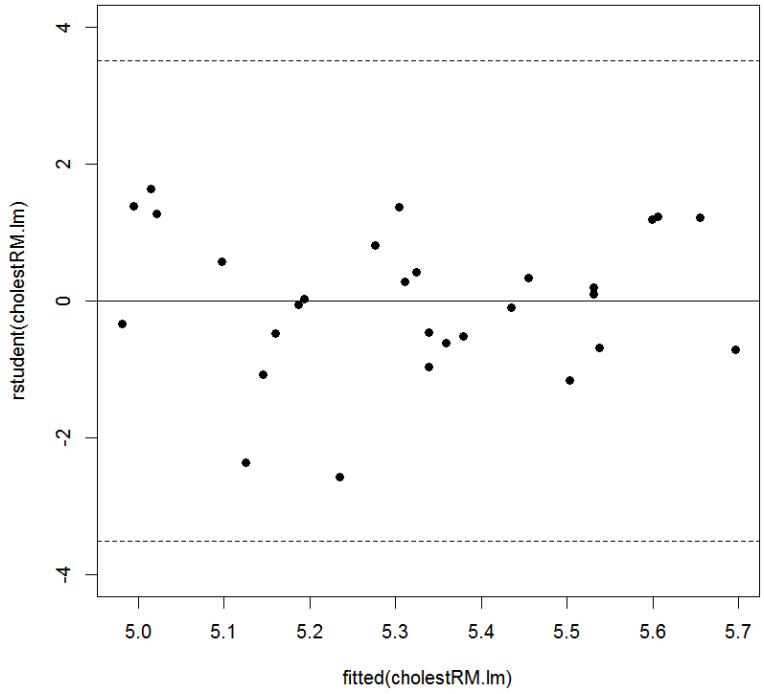
A Normal probability plot of the full set of raw residuals shows rough agreement with the normality assumption (stratifying by state here would leave too few observations for reliable graphics), although there may be a hint of heaviness in just the upper tail, which suggests a possible left skew:

```
qqnorm(resid(cholestRM.lm),pch=19,main='')
qqline(resid(cholestRM.lm))
```



Next, move to an outlier analysis via Studentized deleted residuals (again, do not stratify by state and continue to operate at experiment-wise $\alpha = 0.05$ throughout):

```
plot(rstudent(cholestRM.lm)~fitted(cholestRM.lm), pch=19, ylim=c(-4,4))
abline( h=0 )
n = length(Y); p = length(coef(cholestRM.lm))
tcrit = qt( 1-.5*(.05/n), n-p-1 )
abline( h=tcrit, lty=2 ); abline( h=-tcrit, lty=2 )
```



No Studentized deleted residuals cross above (top) or below (bottom) the exceedance limits, so we identify no potential outliers.

Lastly, apply various influence diagnostics; fastest is via

```
influence.measures( cholestRM.lm )
```

with output (edited)

```
Influence measures of
         lm(formula = Y ~ X + factor(State)) :
      dfb.1_    dfb.X dfb.f.S.    dffit cov.r   cook.d    hat
1   0.631133 -0.45026 -0.53256  0.70058 0.981 1.54e-01 0.1549
2   0.173140 -0.10751 -0.16842  0.21201 1.233 1.54e-02 0.1224
3  -0.661110  0.38302  0.68141 -0.84988 0.692 2.05e-01 0.1141
4   0.005454 -0.00237 -0.00672  0.00835 1.248 2.42e-05 0.0989
5  -0.461593  0.13908  0.65468 -0.82602 0.614 1.87e-01 0.0936
6   0.115497 -0.01162 -0.19612  0.25538 1.146 2.20e-02 0.0911
7   0.162487  0.01638 -0.32154  0.43177 0.998 6.01e-02 0.0910
8  -0.043651 -0.03988  0.13584 -0.19860 1.188 1.35e-02 0.0947
9  -0.004440  0.05248 -0.06500  0.11787 1.252 4.80e-03 0.1134
10  0.059818 -0.16540  0.12062 -0.27619 1.241 2.60e-02 0.1417
11 -0.188716  0.39180 -0.19926  0.56681 1.140 1.05e-01 0.1741
12 -0.108177  0.11867 -0.01675 -0.14592 1.328 7.35e-03 0.1641
13  0.435432 -0.47769  0.07545  0.59537 1.068 1.14e-01 0.1559
14  0.364596 -0.39998  0.07835  0.51426 1.085 8.61e-02 0.1406
15 -0.184317  0.20220 -0.10096 -0.33510 1.074 3.72e-02 0.0874
```

```
16 -0.074240  0.08144 -0.04549 -0.14158 1.195 6.89e-03 0.0830
17 -0.007383  0.00810 -0.00574 -0.01581 1.216 8.67e-05 0.0753
18  0.007234 -0.00794  0.03829  0.06904 1.181 1.65e-03 0.0563
19  0.005837 -0.00640  0.05892  0.10296 1.166 3.65e-03 0.0558
20 -0.000919  0.00101 -0.06665 -0.11327 1.160 4.41e-03 0.0556
21 -0.001890  0.00207 -0.13706 -0.23293 1.068 1.81e-02 0.0556
22  0.017410 -0.01910 -0.07996 -0.12764 1.155 5.59e-03 0.0568
23  0.007616 -0.00835 -0.01588 -0.02437 1.199 2.06e-04 0.0630
24  0.164906 -0.18091 -0.22010 -0.33860 1.040 3.77e-02 0.0777
25 -0.032409  0.03555  0.03845  0.05983 1.225 1.24e-03 0.0859
26 -0.015799  0.01733  0.01874  0.02917 1.229 2.95e-04 0.0859
27 -0.273544  0.30009  0.26082  0.42314 1.073 5.87e-02 0.1118
28 -0.341834  0.37501  0.28734  0.48495 1.100 7.70e-02 0.1382
29  0.232338 -0.25488 -0.18096 -0.31478 1.262 3.37e-02 0.1613
```

Running through the various influence measure of interest:
(i) no observations show any $|DFBETA_j| > 1$ for any variable X_j
(ii) no observations show any $|DFFIT| > 1$
(iii) no observations show any Cook's Distance measure with $P[F(p,n-p) \leq D_5]$ above ½; see, e.g.

```
CookD = influence.measures( cholestRM.lm )$infmat[,6]
pf(q=CookD,df1=n,df2=n-p,lower=T)
```

(iv) no observations show a hat matrix diagonal value is h_{ii} above the leverage rule-of-thumb of $2p/n = 0.2069$.
Thus we conclude there are no observations with unusually high influence.

5. Consider the usual multiple linear regression model, written in matrix notation as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, for $\boldsymbol{\varepsilon} \sim N_p(0,\sigma^2\mathbf{I})$. Assume that $\mathbf{X}$ has full rank. Recall that the various sums of squares from the ANOVA table for this model have the following forms:

SSRegr = $\mathbf{Y}'(\mathbf{H} - n^{-1}\mathbf{J})\mathbf{Y}$

SSE = $\mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$

SSTot = $\mathbf{Y}'(\mathbf{I} - n^{-1}\mathbf{J})\mathbf{Y}$

where the hat matrix is $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{J} = \mathbf{1}\mathbf{1}'$ for $\mathbf{1}' = [1\ 1\ \cdots\ 1]$. As is well-known, these sums of squares are quadratic forms. Show that in each case, the matrix of the quadratic from is symmetric and idempotent.

(*Hint*: where necessary, you may assume that the design matrix can be partitioned as $\mathbf{X} = [\mathbf{1}\ \ \mathbf{X}^*]$, where $\mathbf{X}^*$ is an n×(p–1) submatrix made up of columns that are the individual p–1 predictor variables. What then is $\mathbf{HX}$?)

---

5. For symmetry the solutions are obvious, since
(1) the transpose of a difference is the difference of the transpose, and
(2) we know $\mathbf{H}$ is symmetric, $\mathbf{I}$ is symmetric, and since the constant n^{-1} does not affect the transpose operation, $n^{-1}\mathbf{J}' = n^{-1}(\mathbf{1}\mathbf{1}')' = n^{-1}((\mathbf{1}')'\mathbf{1}') = n^{-1}(\mathbf{1}'\mathbf{1}') = n^{-1}\mathbf{J}$.

For idempotence, we need to show that squaring each matrix returns the original. Recall that $\mathbf{H}$ is idempotent. Start from the 'bottom' and go up:
(i) From SSTot take $(\mathbf{I} - n^{-1}\mathbf{J})^2 = \mathbf{I}^2 - n^{-1}\mathbf{J} - n^{-1}\mathbf{J} + n^{-2}\mathbf{J}^2$. But since $\mathbf{J} = \mathbf{1}\mathbf{1}'$, we see $\mathbf{J}^2 = (\mathbf{1}\mathbf{1}')^2 = \mathbf{1}\mathbf{1}'\mathbf{1}\mathbf{1}' = \mathbf{1}(n)\mathbf{1}' = n\mathbf{1}\mathbf{1}'$. Thus $n^{-2}\mathbf{J}^2 = n^{-2}\, n\mathbf{1}\mathbf{1}' = n^{-1}\mathbf{J}$. Thus

$$(\mathbf{I} - n^{-1}\mathbf{J})^2 = \mathbf{I}^2 - n^{-1}\mathbf{J} - n^{-1}\mathbf{J} + n^{-1}\mathbf{J} = \mathbf{I} - n^{-1}\mathbf{J}$$

which shows the SSTot matrix is idempotent.

(ii) Next, from SSE take $(\mathbf{I} - \mathbf{H})^2 = \mathbf{I}^2 - 2\mathbf{H} + \mathbf{H}^2$. Since $\mathbf{I}$ and $\mathbf{H}$ are idempotent, we quickly see this matrix is idempotent: $(\mathbf{I} - \mathbf{H})^2 = \mathbf{I}^2 - 2\mathbf{H} + \mathbf{H}^2 = \mathbf{I} - 2\mathbf{H} + \mathbf{H} = \mathbf{I} - \mathbf{H}$.

(iii) Lastly, from SSRegr take $(\mathbf{H} - n^{-1}\mathbf{J})^2 = \mathbf{H}^2 - n^{-1}\mathbf{H}\mathbf{J} - n^{-1}\mathbf{J}\mathbf{H} + n^{-2}\mathbf{J}^2$. We know:

(a) $\mathbf{H}^2 = \mathbf{H}$, and

(b) $n^{-2}\mathbf{J}^2 = n^{-1}\mathbf{J}$.

Further, from the hint, let $\mathbf{X} = [\mathbf{1}\ \ \mathbf{X}^*]$ so that $\mathbf{HX} = \mathbf{H}[\mathbf{1}\ \ \mathbf{X}^*] = [\mathbf{H1}\ \ \mathbf{HX}^*]$. But $\mathbf{HX} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{XI} = \mathbf{X}$. So we see $\mathbf{HX} = [\mathbf{H1}\ \ \mathbf{HX}^*] = \mathbf{X} = [\mathbf{1}\ \ \mathbf{X}^*]$. Since the partitioned components in this equality have the same orders, we can therefore conclude from the first partitioned component that $\mathbf{H1} = \mathbf{1}$. Then, $\mathbf{HJ} = \mathbf{H}\mathbf{1}\mathbf{1}' = \mathbf{1}\mathbf{1}' = \mathbf{J}$. A similar argument applied to $\mathbf{X}' = [\mathbf{1}'\ \ \mathbf{X}^{*\prime}]$ and $\mathbf{X}'\mathbf{H}$ yields $\mathbf{1}'\mathbf{H} = \mathbf{1}'$, so that $\mathbf{1}'\mathbf{H} = \mathbf{1}'$ and $\mathbf{JH} = \mathbf{1}\mathbf{1}'\mathbf{H} = \mathbf{1}\mathbf{1}' = \mathbf{J}$. Combining these various results together gives

$$(\mathbf{H} - n^{-1}\mathbf{J})^2 = \mathbf{H}^2 - n^{-1}\mathbf{H}\mathbf{J} - n^{-1}\mathbf{J}\mathbf{H} + n^{-2}\mathbf{J}^2 = \mathbf{H} - n^{-1}\mathbf{J} - n^{-1}\mathbf{J} + n^{-1}\mathbf{J} = \mathbf{H} - n^{-1}\mathbf{J},$$

which establishes the idempotence of the SSRegr matrix.

6. Assume a simple linear regression-through-the-origin model: $Y_i \sim$ indep. $N(\beta_1 X_i, \sigma^2)$, $i = 1, ..., n$. The maximum likelihood estimator (MLE) of β_1 is

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

(a) Derive the standard error, $s\{b_1\}$, of b_1. You may assume that $MSE = \sum_{i=1}^{n}(Y_i - b_1 X_i)^2/(n-2)$ is an unbiased estimator of σ^2.

(b) What is the MLE, $\hat{Y}_h$, of $E[Y_h]$ at any X_h?

(c) Find an equation for the standard error of $\hat{Y}_h$ at any X_h.

---

6. (a) Notice that b_1 can be written in the form $b_1 = \sum_{i=1}^{n} k_i Y_i$, for

$$k_i = \frac{X_i}{\sum_{j=1}^{n} X_j^2}.$$

Thus the variance of b_1 is $Var\{b_1\} = Var\{\sum_{i=1}^{n} k_i Y_i\} = \sum_{i=1}^{n} Var\{k_i Y_i\}$, since the Y_is are independent. But then $Var\{b_1\} = \sum_{i=1}^{n} Var\{k_i Y_i\} = \sum_{i=1}^{n} k_i^2 Var\{Y_i\} = \sigma^2 \sum_{i=1}^{n} k_i^2$, since $Var\{Y_i\} = \sigma^2$. Thus we see

$$Var\{b_1\} = \sigma^2 \frac{\sum_{i=1}^{n} X_i^2}{\left(\sum_{j=1}^{n} X_j^2\right)^2} = \frac{\sigma^2}{\sum_{j=1}^{n} X_j^2}$$

and from this the standard error is

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

where the unbiased estimator MSE has replaced σ^2.

(b) Appealing to MLE invariance, the MLE of $E[Y_h]$ at any X_h clearly has the form $\hat{Y}_h = b_1 X_h$.

(c) Clearly $Var\{\hat{Y}_h\} = Var\{b_1 X_h\} = X_h^2 Var\{b_1\}$, so $s\{\hat{Y}_h\} = \sqrt{Var\{\hat{Y}_h\}} = \sqrt{X_h^2 Var\{b_1\}} = |X_h|\, s\{b_1\}$.