One of my quirks is that I prefer to grade anonymously. That means I don't know to see your name until after I finish grading. Put your name on the **back of the last page** of your answers and **only on the back** of that page. Make sure your answer packet is securely stapled.

R code and output and SAS code and output are included at the end of the questions. Unless specifically indicated for a bit of R code, all R lm() or gls() fits use contr.treatment (set first to 0) constraints.

1. Consider a small experiment, with 5 treatments, each with two replicates in a CRD. The treatments are levels of a quantitative variable, X. Here are the X values, the data, and some summary statistics.

	Treatment					
	A	B	C	D	E	overall
X	0	2.5	5	10	20	$\overline{X} = 7.50$
Data	4.14, 2.82	5.83, 4.55	4.73, 7.31	4.80, 5.82	8.81, 10.31	
Means	3.48	5.19	6.02	5.31	9.56	5.91
$\sum_{ij}(y_{ij} - \overline{y}_{i})^2$						6.54
$\sum_{ij}(y_{ij} - \overline{y}_{.})^2$						46.78
$\sum_{i}(\overline{y}_{i} - \overline{y}_{.})^2$						40.25

   (a) Use a model comparison approach to test the null hypothesis that the five treatments have the same mean. Report your test statistic and its distribution under the null hypothesis.

   (b) What are the coefficients for the contrast that evaluates linear trend in the data (i.e. coefficients for the linear orthogonal polynomial defined by X)?

   If you can not answer the previous question, use -3 -2 -1 1 5 as a replacement set of coefficients for the remaining parts of this question. If you wish to use the replacement set, please write "replacement" clearly on your answer sheet.

   (c) A second question of interest to the investigators is the contrast 0 2 -3 1 0. Is this contrast orthogonal to the contrast in part 1b (or the replacement -3 -2 -1 1 5)? Explain why or why not.

   (d) Using your coefficients from part 1b (or the replacement contrast), estimate the value of the contrast and its standard error.

   (e) Calculate the SS associated with the contrast from part 1b (or the replacement) and use that value to construct an F test that the population value of the contrast = 0. Report the SS for the contrast, the F statistic, and the distribution of the F statistic under the null hypothesis.

   (f) Test whether there is evidence for any differences among the means other than that described by the contrast in part 1b (or the replacement contrast), i.e. the "left-over" SS. Report your test statistic and its distribution under the null hypothesis.

   (g) Consider a C matrix with two rows. If you answered part 1b, the rows are:

	Treatment				
Row	A	B	C	D	E
1	your coefficients from part 1b				
2	0	2	-3	1	0

If you are using the "replacement" contrast, the rows are:

	Treatment				
Row	A	B	C	D	E
1	-3	-2	-1	1	5
2	0	2	-3	1	0

Test the null hypothesis that $\boldsymbol{C\beta} = 0$. Report your test statistic and its distribution under the null hypothesis.

2. Folklore, especially from eastern Europe, tells us that people behave differently during a full moon. Think of Count Dracula and werewolves. The data for this problem come from an observational study to assess the association between the phase of the moon and admissions to a mental health clinic. The daily admisions rate (# new patients/day) was calculated for three moon PHASEs (B: immediately Before full moon, D: During full moon, and A: immediately After full moon) for each MONTH from August 1971 to July 1972. These 12 months were arbitrarily chosen.

   Admissions are known to vary monthly for many reasons that are irrelevant to the question of interest. For example, admission rates were expected to be low in August because many doctors took month-long vacations (or longer) in August (these are 1970's data). Although this is an observational study, for the purpose of working out a model and an analysis, you may consider PHASE to be randomly assigned to observations within each month.

   R code and output and SAS code and output are included at the end of the questions.

   (a) For the following three observations:
   1) Write out the three rows of the part of the full-rank X matrix that concerns the moon phase using sum-to-zero constraints.
   2) Write out the three rows of the part of the full-rank X matrix that concerns the interaction of April and moon phase, again using sum-to-zero constraints. April is the first level of the month factor. You only have write the interaction columns that concern April and moon phase.
   (No need to write out the intercept column, the columns for months, or the other interaction columns).

Month	Phase
Apr	A
Apr	B
Apr	D

   (b) Are the data balanced (equal sample sizes for all cells)? Briefly explain why or why not.

   (c) Are there any missing cells? Briefly explain why or why not.

(d) Test the hypothesis that the three PHASEs have the same mean admissions rate. Report your test statistic and state its distribution under the null hypothesis.

(e) Estimate the mean difference in admissions rate between during a full moon and "not a full moon". Not a full moon is defined as the average of the before and after rates. Report your estimate and its s.e.

(f) The investigators plan on continuing this study. They are especially interested in whether there is evidence of a difference between the before (B) and after (A) periods. If there are two observations per month, how many months of data will need to be collected to get 80% power for an $\alpha = 0.05$ two-sided test to detect a difference of 0.75 patients/day. The error variance is presumed to be 4.2 (patients/day)^2. Some T quantiles at the appropriate error df are $T =_{0.5}= 0$, $T_{0.7} = 0.525$, $T_{0.8} = 0.844$, $T_{0.9} = 1.286$, $T_{0.95} = 1.653$, $T_{0.975} = 1.973$, and $T_{0.999} = 2.347$.

(g) When the data from the study in part 2f are analyzed by 2-way ANOVA with all the usual terms in the model, what will be the degrees of freedom for error? To help me understand your answer, list the terms you are including in your model.

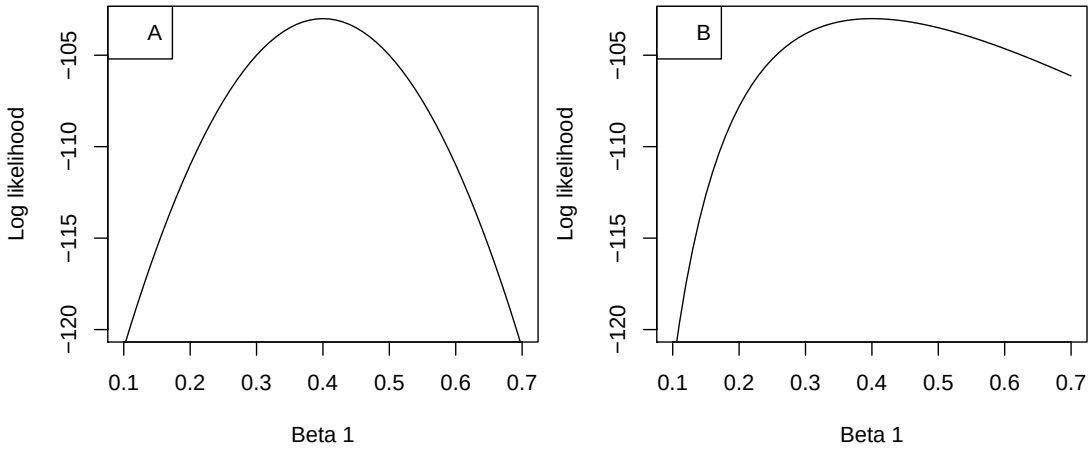
If you don't have an answer for part 2f, use 40 months.

3. Do less expensive houses sell faster? Here we evaluate the association between whether a house sold in less than 3 months (yes / no), considered the response variable, and its price per square foot and its age. The model is

$$\begin{aligned} Y_i &\sim \text{independent Bernoulli}(\pi_i) \\ \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + \beta_1 P_i + \beta_2 A_i \end{aligned} \tag{1}$$

where $Y_i = 1$ if the house sold in less than 3 months and 0 otherwise, π_i is the probability of selling in less than 3 months, P_i is the price per square foot, and A_i is the age. Our focus is on the association of price per square foot and probability (or odds) of selling quickly.

Below are plots of two hypothetical log likelihood functions for β_1 maximized over the other three parameters, i.e. I have plotted $\log L(\beta_1 \mid \boldsymbol{Y}) = \max_{\beta_0,\beta_2,\sigma^2} \log L(\beta_0, \beta_1, \beta_2, \sigma^2 \mid \boldsymbol{Y})$ as a function of β_1.



(a) Is the variance of the MLE of β_1, i.e. $\hat{\beta}_1$, in situation A *larger* or *smaller* than that in situation B? Briefly explain your choice.

(b) In situation A, explain how to calculate the **99%** profile likelihood confidence interval for β_1. Your answer does not need to include specific numbers and you do not have to actually calculate the confidence interval. But, it should explain in sufficient detail so someone with a computer could follow your instructions and calculate that interval.

(c) In situation B, will the endpoints of the Wald confidence interval for $\hat{\beta}_1$ be *similar to* or *quite different from* the endpoints of the profile likelihood confidence interval? Briefly explain your choice.

Data are available for 115 houses sold in one part of Iowa in 2010. Model (1) was fit to these data. R and SAS code and output are at the end of the questions. I have removed some numbers from the output. Using those results:

(d) Test the hypothesis of no association between price per square foot (price) and the odds of selling in less than 3 months for houses of the same age. Report your test statistic and its distribution under the null hypothesis.

(e) Estimate the difference in (or ratio of) the odds of selling in less than 3 months associated with an **decrease** of 10$ per square foot in the price of a house when house age is held constant. Clearly indicate in your answer whether you are calculating a difference in odds or a ratio of odds.

(f) Calculate a Wald 95% two-sided confidence interval for the difference in (or ratio of) odds of selling in less than 3 months associated with an **decrease** of 10$ per square foot in the price of a house when age is held constant. Some potentially relevant quantiles are: $Z_{0.95} = 1.645$, $Z_{0.975} = 1.960$, $T_{0.95,112} = 1.658$, $T_{0.975,112} = 1.981$. If you need a different quantile for your interval, use similar notation to indicate which quantile you need to calculate your confidence interval. Or, if you need something not available in the output I provide to calculate your confidence interval, tell me what you need.

(g) Estimate the probability that a house that is 10 years old with a price per square foot of 90$ sells in less than 3 months.

Remember:
write your name on the back of the last page of your answers

```
fm <- read.table('fullmoon.txt',header=T, as.is=T)

fm$month.f <- factor(fm$month)
fm$phase.f <- factor(fm$phase)


fm.lm1 <- lm(rate ~ phase.f, data=fm)
anova(fm.lm1)
summary(fm.lm1)

with(fm, tapply(rate, phase, mean))

fm.lm2 <- lm(rate ~ month.f + phase.f, data=fm)
anova(fm.lm2)
summary(fm.lm2)

coef(fm.lm2)

C2 <- rbind(c(1,rep(1/12,11),0,0),
            c(1,rep(1/12,11),1,0),
            c(1,rep(1/12,11),0,1) )
C2 %*% coef(fm.lm2)

fm.lm3 <- lm(rate ~ month.f + phase.f + month.f:phase.f, data=fm)
anova(fm.lm3)
summary(fm.lm3)

coef(fm.lm3)

C3 <- rbind(c(1,rep(1/12,11),0,0, rep(0,22)),
            c(1,rep(1/12,11),1,0, rep(1/12,11),rep(0,11)),
            c(1,rep(1/12,11),0,1, rep(0,11),rep(1/12,11) ) )
C3 %*% coef(fm.lm3)

options(contrasts=c('contr.helmert','contr.poly'))

fm.lm2b <- lm(rate ~ month.f + phase.f, data=fm)
drop1(fm.lm2b, ~., test='F')

fm.lm3b <- lm(rate ~ month.f + phase.f + month.f:phase.f, data=fm)
drop1(fm.lm3b, ~., test='F')
```

```
> fm <- read.table('fullmoon.txt',header=T, as.is=T)
>
> fm$month.f <- factor(fm$month)
> fm$phase.f <- factor(fm$phase)
>
>
> fm.lm1 <- lm(rate ~ phase.f, data=fm)
> anova(fm.lm1)
Analysis of Variance Table

Response: rate
          Df Sum Sq Mean Sq F value Pr(>F)
phase.f    2  63.32  31.660  1.6068 0.2135
Residuals 39 768.44  19.703
> summary(fm.lm1)

Call:
lm(formula = rate ~ phase.f, data = fm)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4929 -3.3107  0.4571  2.9946 11.4071

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.7905     0.6849  17.214   <2e-16 ***
phase.f1     -0.2964     0.8389  -0.353   0.7257
phase.f2      0.8512     0.4843   1.757   0.0867 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.439 on 39 degrees of freedom
Multiple R-squared: 0.07613,    Adjusted R-squared: 0.02875
F-statistic: 1.607 on 2 and 39 DF,  p-value: 0.2135

>
> with(fm, tapply(rate, phase, mean))
       A        B        D
11.23571 10.64286 13.49286
>
> fm.lm2 <- lm(rate ~ month.f + phase.f, data=fm)
> anova(fm.lm2)
Analysis of Variance Table

Response: rate
```

```
          Df Sum Sq Mean Sq F value    Pr(>F)
month.f   11 581.90  52.900  8.4957 2.453e-06 ***
phase.f    2  75.51  37.756  6.0636  0.006487 **
Residuals 28 174.35   6.227
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> summary(fm.lm2)

Call:
lm(formula = rate ~ month.f + phase.f, data = fm)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9144 -1.4722  0.0824  1.1542  5.1486

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.91458    0.38903  30.627  < 2e-16 ***
month.f1    -6.79560    0.95543  -7.113 9.72e-08 ***
month.f2    -0.95162    0.52687  -1.806 0.081652 .
month.f3     0.41181    0.40729   1.011 0.320638
month.f4    -0.11292    0.31817  -0.355 0.725326
month.f5     0.89417    0.23321   3.834 0.000654 ***
month.f6     0.01481    0.19580   0.076 0.940225
month.f7     0.31852    0.16924   1.882 0.070266 .
month.f8     0.32454    0.16832   1.928 0.064035 .
month.f9    -0.34037    0.15081  -2.257 0.032003 *
month.f10   -0.35652    0.12066  -2.955 0.006280 **
month.f11   -0.19223    0.12468  -1.542 0.134344
phase.f1    -0.43426    0.48023  -0.904 0.373560
phase.f2     0.93241    0.27726   3.363 0.002247 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.495 on 28 degrees of freedom
Multiple R-squared: 0.7904,     Adjusted R-squared: 0.6931
F-statistic: 8.122 on 13 and 28 DF,  p-value: 2.063e-06

>
> coef(fm.lm2)
(Intercept)    month.f1    month.f2    month.f3    month.f4    month.f5
11.91458333 -6.79560185 -0.95162037  0.41180556 -0.11291667  0.89416667
   month.f6    month.f7    month.f8    month.f9   month.f10   month.f11
 0.01481481  0.31851852  0.32453704 -0.34037037 -0.35651515 -0.19223485
   phase.f1    phase.f2
```

```
-0.43425926  0.93240741
>
> C2 <- rbind(c(1,rep(1/12,11),0,0),
+             c(1,rep(1/12,11),1,0),
+             c(1,rep(1/12,11),0,1) )
> C2 %*% coef(fm.lm2)
         [,1]
[1,] 11.34913
[2,] 10.91487
[3,] 12.28154
>
> fm.lm3 <- lm(rate ~ month.f + phase.f + month.f:phase.f, data=fm)
> anova(fm.lm3)
Analysis of Variance Table

Response: rate
                Df Sum Sq Mean Sq F value   Pr(>F)
month.f         11 581.90  52.900 12.6303 0.002753 **
phase.f          2  75.51  37.756  9.0145 0.015568 *
month.f:phase.f 22 149.22   6.783  1.6194 0.285875
Residuals        6  25.13   4.188
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> summary(fm.lm3)

Call:
lm(formula = rate ~ month.f + phase.f + month.f:phase.f, data = fm)

Residuals:
   Min     1Q Median     3Q    Max
  -2.9    0.0    0.0    0.0    2.9

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        11.95278    0.32657  36.601 2.77e-08 ***
month.f1           -6.48333    0.79993  -8.105 0.000189 ***
month.f2           -1.30556    0.44762  -2.917 0.026748 *
month.f3            0.42222    0.33632   1.255 0.256004
month.f4           -0.10667    0.26200  -0.407 0.698039
month.f5            0.89833    0.19889   4.517 0.004031 **
month.f6            0.03214    0.16762   0.192 0.854255
month.f7            0.30536    0.14491   2.107 0.079674 .
month.f8            0.32639    0.13844   2.358 0.056469 .
month.f9           -0.33889    0.12397  -2.734 0.034022 *
month.f10          -0.30152    0.10331  -2.919 0.026681 *
```

```
month.f11          -0.19571    0.10245  -1.910 0.104675
phase.f1           -0.27917    0.39996  -0.698 0.511319
phase.f2            0.97778    0.23092   4.234 0.005474 **
month.f1:phase.f1   0.22500    1.02327   0.220 0.833252
month.f2:phase.f1  -0.45833    0.53931  -0.850 0.428012
month.f3:phase.f1  -0.15417    0.41337  -0.373 0.722008
month.f4:phase.f1  -0.14250    0.32156  -0.443 0.673185
month.f5:phase.f1   0.17583    0.23384   0.752 0.480534
month.f6:phase.f1   0.09702    0.19643   0.494 0.638903
month.f7:phase.f1   0.10089    0.19208   0.525 0.618237
month.f8:phase.f1   0.20069    0.16965   1.183 0.281570
month.f9:phase.f1   0.06056    0.15190   0.399 0.703954
month.f10:phase.f1 -0.10955    0.12075  -0.907 0.399276
month.f11:phase.f1 -0.07462    0.12548  -0.595 0.573773
month.f1:phase.f2  -1.70833    0.53931  -3.168 0.019377 *
month.f2:phase.f2  -0.84722    0.32158  -2.635 0.038823 *
month.f3:phase.f2  -0.09861    0.23696  -0.416 0.691779
month.f4:phase.f2  -0.26917    0.18487  -1.456 0.195642
month.f5:phase.f2   0.25250    0.14605   1.729 0.134558
month.f6:phase.f2   0.04940    0.12343   0.400 0.702826
month.f7:phase.f2   0.08393    0.09327   0.900 0.402882
month.f8:phase.f2   0.15417    0.09784   1.576 0.166148
month.f9:phase.f2   0.05333    0.08762   0.609 0.565046
month.f10:phase.f2  0.15121    0.07624   1.983 0.094576 .
month.f11:phase.f2  0.05657    0.07244   0.781 0.464595
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.047 on 6 degrees of freedom
Multiple R-squared: 0.9698,     Adjusted R-squared: 0.7935
F-statistic: 5.503 on 35 and 6 DF,  p-value: 0.01997

>
> coef(fm.lm3)
       (Intercept)           month.f1           month.f2           month.f3
       11.95277778       -6.48333333       -1.30555556        0.42222222
          month.f4           month.f5           month.f6           month.f7
       -0.10666667        0.89833333        0.03214286        0.30535714
          month.f8           month.f9          month.f10          month.f11
        0.32638889       -0.33888889       -0.30151515       -0.19570707
          phase.f1           phase.f2  month.f1:phase.f1  month.f2:phase.f1
       -0.27916667        0.97777778        0.22500000       -0.45833333
 month.f3:phase.f1  month.f4:phase.f1  month.f5:phase.f1  month.f6:phase.f1
       -0.15416667       -0.14250000        0.17583333        0.09702381
 month.f7:phase.f1  month.f8:phase.f1  month.f9:phase.f1 month.f10:phase.f1
```

```
        0.10089286          0.20069444          0.06055556         -0.10954545
month.f11:phase.f1  month.f1:phase.f2  month.f2:phase.f2  month.f3:phase.f2
       -0.07462121         -1.70833333         -0.84722222         -0.09861111
 month.f4:phase.f2  month.f5:phase.f2  month.f6:phase.f2  month.f7:phase.f2
       -0.26916667          0.25250000          0.04940476          0.08392857
 month.f8:phase.f2  month.f9:phase.f2 month.f10:phase.f2 month.f11:phase.f2
        0.15416667          0.05333333          0.15121212          0.05656566
>
> C3 <- rbind(c(1,rep(1/12,11),0,0, rep(0,22)),
+             c(1,rep(1/12,11),1,0, rep(1/12,11),rep(0,11)),
+             c(1,rep(1/12,11),0,1, rep(0,11),rep(1/12,11) ) )
> C3 %*% coef(fm.lm3)
          [,1]
[1,] 11.39051
[2,] 11.10475
[3,] 12.19144
>
> options(contrasts=c('contr.helmert','contr.poly'))
>
> fm.lm2b <- lm(rate ~ month.f + phase.f, data=fm)
> drop1(fm.lm2b, ~., test='F')
Single term deletions

Model:
rate ~ month.f + phase.f
         Df Sum of Sq    RSS     AIC F value   Pr(>F)
<none>                174.35  87.782
month.f 11    594.09 768.44 128.081  8.6737    2e-06 ***
phase.f  2     75.51 249.86  98.895  6.0636 0.006487 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
> fm.lm3b <- lm(rate ~ month.f + phase.f + month.f:phase.f, data=fm)
> drop1(fm.lm3b, ~., test='F')
Single term deletions

Model:
rate ~ month.f + phase.f + month.f:phase.f
                Df Sum of Sq    RSS     AIC F value   Pr(>F)
<none>                        25.13  50.428
month.f         11    530.70 555.83 158.478 11.5191 0.003538 **
phase.f          2     77.13 102.26 105.375  9.2082 0.014839 *
month.f:phase.f 22    149.22 174.35  87.782  1.6194 0.285875
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
sell.glm <- glm(sell~price+age, family=binomial)

summary(sell.glm)
anova(sell.glm, test='Chi')
```

```
> sell.glm <- glm(sell~price+age, family=binomial)
> summary(sell.glm)

Call:
glm(formula = sell ~ price + age, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4635  -0.1821   0.2537   0.6472   1.5232

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 13.35337    2.90953   4.590 4.44e-06 ***
price       -0.10068    0.02403  -4.190 2.78e-05 ***
age         -0.05194    0.01175  -4.422 9.78e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 144.480  on 114  degrees of freedom
Residual deviance:  84.922  on 112  degrees of freedom
AIC: 90.922

> anova(sell.glm, test='Chi')
Analysis of Deviance Table
Model: binomial, link: logit
Response: sell

Terms added sequentially (first to last)
      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    114    144.480
price  1   23.820       113    120.660 1.058e-06 ***
age    1   35.738       112     84.922 2.257e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
data moon;
  infile 'fullmoon.txt' firstobs=2;
  input month $ phase $ rate;
  run;

proc glm;
  class phase;
  model rate = phase;
  lsmeans phase /stderr;
  title 'phase only';
  run;

proc glm;
  class month phase;
  model rate = month phase;
  lsmeans phase /stderr;
  title 'additive model';
  run;

proc glm;
  class month phase;
  model rate = month phase month*phase;
  lsmeans phase /stderr;
  title 'with interaction';
  run;
```

```
data sell;
  infile 'sell.csv' dsd firstobs=2;
  input price age sell;
  run;

proc logistic descending;
  model sell = price age ;
  run;
```