STAT 425 — Section D1U, D1G — Spring 2014

Midterm Exam II

April 23, 2014

Full Name: Key

- This is a 50 minute exam. There are 4 problems for everyone, and 2 additional problems for graduate students only.
- The exam is worth a total of 34 points for undergraduates and 42 points for graduate students.
- You may use *three* pages of personal notes and a standard scientific calculator. (You may *not* share these items with anyone else.)
- *Write all answers in the spaces provided.* If you require more space to write your answer, you may use the back side of the page.
- You are not allowed to communicate with anyone except the instructor or proctors before you submit this exam.

Useful Abbreviations:

CI = confidence interval
PI = prediction interval
se = standard error
E = expected value
var = variance (or variance-covariance)
cov = covariance
SLR = simple linear regression
BLUE = best linear unbiased estimate/estimator
GLS = generalized least squares
WLS = weighted least squares
MSE = mean square error
TSS = total sum of squares
RSS = residual sum of squares
df = degrees of freedom
VIF = variance inflation factor
VST = variance stabilizing transformation
H_0 = the null hypothesis of a test
H_a = the alternative hypothesis of a test

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:16:28 GMT -05:00

https://www.coursehero.com/file/9954387/midterm2soln/

1. For each part below, CIRCLE the ONE BEST answer. [1 pt each]

   (a) How many *blocking factors* are in a Latin Square design?

   0 1 2 3

   (b) A Box-Cox transformation requires the dependent variable to be

   positive negative nonzero none of these

   (c) Initially, all variables are in the model in

   forward selection backward elimination stepwise selection none of these

   (d) The C_p statistic may be used to ___ a model.

   select test H_0 for estimate design

   (e) In a randomized complete block design, the number of blocks must be ___ the number of treatments.

   greater than less than equal to none of these

   (f) In a homogeneity-of-regressions ("ANCOVA") model, a significant interaction between the X variable and the "dummy" variable(s) would indicate that the regression lines are

   parallel not parallel perpendicular none of these

   (g) A VST involves transformation of the

   independent variable(s) dependent variable both neither

   (h) Both adding and dropping variables may be performed in

   forward selection backward elimination stepwise selection none of these

2. You intend to compare two brands of eyedrops (liquid solutions applied to the eyes to treat redness) using a randomized experiment with 20 human subjects. Think of a designed experiment that uses blocking. Answer the following:

   (a) What are the treatments? [1 pt]

   the two brands of eyedrops

   (b) What are the blocks? [1 pt]

   the human subjects

   (c) What are the experimental units? [1 pt]

   eyes

   (d) Briefly describe the structure of the randomization. [2 pts]

   For each human subject, independently, one eye is randomly chosen to receive brand 1, and the other eye receives brand 2.

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:16:28 GMT -05:00

https://www.coursehero.com/file/9954387/midterm2soln/

3. A linear regression of a variable Y on variables X_1 and X_2, including some polynomial terms, is performed and summarized in R as follows:

```
> summary(lm(y ~ x1 + x2 + I(x2^2) + x1*x2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.8981     0.2879  34.383   <2e-16 ***
x1            0.8269     0.3178   2.602   0.0143 *
x2            0.1883     0.3219   0.585   0.5631
I(x2^2)       0.7890     0.6353   1.242   0.2239
x1:x2         2.2568     0.5666   3.983   0.0004 ***
```

(a) Write an expression for the full model equation. (Do not substitute any estimates!) [3 pts]

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \varepsilon$$

(b) Test for interaction between X_1 and X_2. (Give the test statistic, p-value, and conclusion.) [3 pts]

$t = 3.983$

$p = 0.0004$

Reject the null hypothesis of no interaction.

(c) What is the estimated expected change in Y if X_1 increases by one unit? (Hint: It depends on X_2.) [2 pts]

$$\hat{\beta}_1 + \hat{\beta}_{12} X_2 \approx 0.8269 + 2.2568\, X_2$$

(d) What polynomial term appears to be "missing" from this model? [1 pt]

the term for X_1^2

(e) Suppose you are going to perform backward elimination on this model, based on a p-value threshold of $p = 0.05$. Which term (if any) would be removed first? Why? (Hint: Remember hierarchy!) [2 pts]

By hierarchy, only X_2^2 and $X_1 X_2$ are eligible for removal, at the first stage. Since X_2^2 has the larger p-value (0.2239), and it exceeds 0.05, remove X_2^2 first.

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:16:28 GMT -05:00

https://www.coursehero.com/file/9954387/midterm2soln/

4. An experiment is conducted with a completely randomized design and two treatment factors, A and B, each with two levels. The design is balanced, with 5 replications. An ANOVA table from R is

```
> anova(lm(response ~ A*B))
Analysis of Variance Table

Response: response
          Df Sum Sq Mean Sq F value   Pr(>F)
A          1  11.25   11.25   1.125 0.304594
B          1 101.25  101.25  10.125 0.005793 **
A:B        1   1.25    1.25   0.125 0.728289
Residuals 16 160.00   10.00
```

(a) How many observations are there? [1 pt]

20

(b) How many treatments are there? [1 pt]

4

(c) How many blocks are there? [1 pt]

none

(d) Summarize the conclusions you would draw concerning the interactions and main effects. [3 pts]

Interaction: No evidence for it ($p > 0.05$).

Factor A: No evidence for it ($p > 0.05$).

Factor B: There is evidence of effects differing ($p < 0.05$).

(e) Let α_1 and α_2 be the main effects of Factor A. Perform a test of $H_0 : \alpha_1 = \alpha_2$ versus $H_a : \alpha_1 \neq \alpha_2$. (Give test statistic, p-value, and conclusion.) [2 pts]

$F = 1.125 \quad p \approx 0.305$

No evidence against $H_0 : \alpha_1 = \alpha_2$.

(f) Would you need the Tukey multiple comparisons method for comparisons between the levels of Factor A? Explain briefly. [2 pts]

No. Factor A has only two levels, so there would be only one comparison.

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:16:28 GMT -05:00

https://www.coursehero.com/file/9954387/midterm2soln/

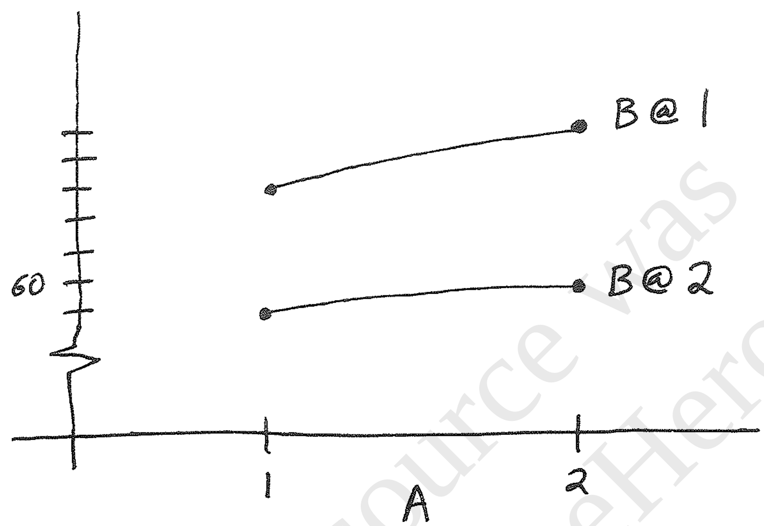
GRADUATE STUDENTS ONLY

5. In the previous problem, the estimated mean responses are

$$\hat{\mu}_{11} = 63 \qquad \hat{\mu}_{12} = 59 \qquad \hat{\mu}_{21} = 65 \qquad \hat{\mu}_{22} = 60$$

where $\hat{\mu}_{ij} = \overline{y}_{ij}$ is the average response for the observations at level i of Factor A ($i = 1, 2$), and level j of Factor B ($j = 1, 2$).

Draw an interaction plot for this situation, with Factor A on the horizontal axis. Make sure it is fully labeled! [4 pts]



6. A scientific model proposes the following relationship between variables X and Y:

$$Y = \gamma X^{\alpha} e^{-\beta X} \epsilon$$

where γ, α, and β are unknown constants ($\gamma > 0$) and $\epsilon > 0$ is a multiplicative error whose distribution is the same for all X.

Transform this to a *linear* regression model: Write the transformed model equation. What are the (transformed) dependent and independent variables? What are the regression parameters? [4 pts]

dependent var. independent vars.

$$\overline{\ln Y} = \underbrace{\ln \gamma} + \alpha \, \overline{\ln X} \underbrace{- \beta} X + \ln \varepsilon$$

Regression params.

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:16:28 GMT -05:00

https://www.coursehero.com/file/9954387/midterm2soln/

STAT 425 — Sections T1U, T1G — Fall 2015

Midterm Exam II

December 9, 2015

Full Name: Key

- This is an 80 minute exam. There are 6 problems, worth a total of 57 points.
- You may use *three* pages of personal notes and a standard scientific calculator. (You may *not* share these or any other items with anyone else.)
- *Write all answers in the spaces provided.* If you require more space to write your answer, you may use the back side of the page.
- You are not allowed to communicate with anyone except the instructor or proctors before you submit this exam.

Useful Abbreviations:

CI = confidence interval	PI = prediction interval
se = standard error	
E = expected value	var = variance (or variance-covariance)
cov = covariance	
SLR = simple linear regression	BLUE = best linear unbiased estimate/estimator
GLS = generalized least squares	WLS = weighted least squares
MSE = mean square error	TSS = total sum of squares
RSS = residual sum of squares	df = degrees of freedom
VIF = variance inflation factor	VST = variance stabilizing transformation
H_0 = the null hypothesis of a test	H_a = the alternative hypothesis of a test

Some selected formulas:

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p' - n$$

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:07:54 GMT -05:00

https://www.coursehero.com/file/14156463/midterm2soln/

This study resource was shared via CourseHero.com

1. In the process of using R to select a regression model of `Y` on `X1` and `X2` based on $\underline{n = 16}$ observations, the following results are obtained for the current model (having `X1` only):

```
> current.model <- lm(Y ~ X1)
> add1(current.model, ~ X1 + X2, test="F")
Single term additions
...
       Df Sum of Sq   RSS     AIC F value    Pr(>F)
<none>               96.00  32.668
X2      1     90.25  5.75 -10.374  204.04 2.519e-09 ***
...
> drop1(current.model, test="F")
Single term deletions
...
       Df Sum of Sq RSS    AIC F value Pr(>F)
<none>               96 32.668
X1      1          9 105 32.102  1.3125 0.2711
```

(a) If this is *forward selection* based on an F-statistic threshold of $F_{\text{in}} = 3$, what would be the next step? Why? [2 pts]

Add X_2, because its "add1" F-statistic is $\approx 204 > 3 = F_{in}$

(b) If this is *backward elimination* based on an F-statistic threshold of $F_{\text{out}} = 3$, what would be the next step? Why? [2 pts]

Drop X_1, because its "drop 1" F-statistic is $\approx 1.3 < 3 = F_{out}$

(c) Compute Mallows' C_p for the full model `Y ~ X1 + X2` and the reduced model `Y ~ X1`. Of these two models, which is better according to C_p? [5 pts]

Full model: $\dfrac{5.75}{5.75/(16-3)} + 2\cdot 3 - 16 = 3$

Model Y~X1: $\dfrac{96}{5.75/(16-3)} + 2\cdot 2 - 16 \approx 205$

C_p prefers the full model Y~X1+X2

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:07:54 GMT -05:00

https://www.coursehero.com/file/14156463/midterm2soln/

2. For each part below, CIRCLE the ONE BEST answer. [1 pt each]

(a) A variance stabilizing transformation is intended to remedy the problem of

heteroscedasticity curvature in the mean non-normality none of these

(b) According to the principle of hierarchy, any sub-model of the two-factor ANOVA model $y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$ that contains the term α_i must also contain the term

β_j $\alpha\beta_{ij}$ both of these neither of these

(c) In a randomized complete block design (with no missing values), the number of experimental units is evenly divisible by the number of

treatments blocks both of these neither of these

(d) If CI_1 and CI_2 are (random) Tukey 95% simultaneous confidence intervals for mean differences δ_1 and δ_2, respectively, then the probability that $\delta_1 \in \text{CI}_1$ is

0.95 ≥ 0.95 ≤ 0.95 0.95^2

(e) In a *balanced* completely randomized design with 3 treatments, the probability that a particular experimental unit ends up in the *first* treatment group

is 1/9 is 1/2 is 1/3 cannot be determined

(f) Which of these *cannot* be a problem for a one-way ANOVA model ($y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$)?

non-normality heteroscedasticity curvature in the mean none of these

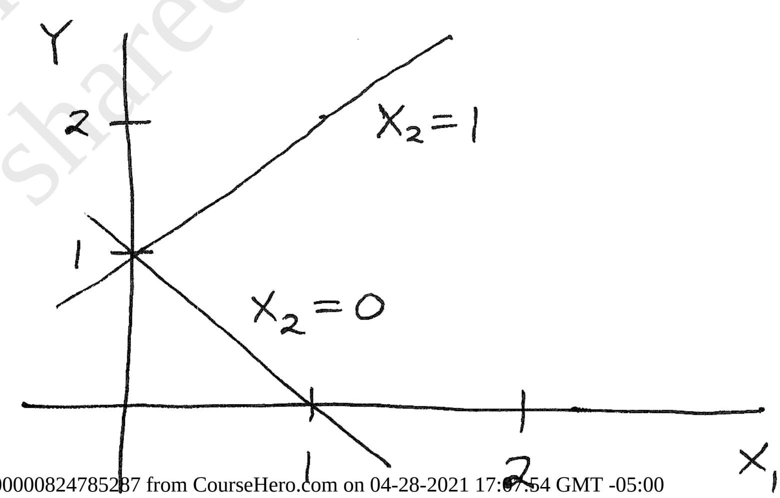
3. Briefly answer the following:

(a) List one *advantage* and one *disadvantage* of using a randomization test. [2 pts]

Advantage: no need for the usual assumptions

Disadvantage: requires special software or programming

(b) Consider homogeneity-of-regressions model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$, where X_1 is quantitative and X_2 is a dummy (0/1) variable. Draw a Y-versus-X_1 least-squares fitted regression plot for the case $\hat{\beta}_0 = 1$, $\hat{\beta}_1 = -1$, $\hat{\beta}_2 = 0$, $\hat{\beta}_3 = 2$. Label which line corresponds to $X_2 = 0$ and which to $X_2 = 1$. [3 pts]



This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:07:54 GMT -05:00

https://www.coursehero.com/file/14156463/midterm2soln/

4. An experiment is designed to compare the mean quarter-mile times of three racehorses (coded A, B, and C). Three jockeys are recruited to ride the horses. Three separate quarter-mile races take place. In each race, all three horses run (once), and each of the three jockeys rides one horse. The assignments of the horses are randomized with the restriction that each horse has a different jockey in each race.

   (a) Name the type of design for this experiment. Then draw a diagram that shows one possible outcome of the randomization. [4 pts]

   Latin square design

   race

jockey	1	2	3
1	A	B	C
2	B	C	A
3	C	A	B

   (b) What are the *treatments* in this experiment? How many? [2 pts]

   the racehorses — three

   (c) What are the *experimental units* in this experiment? How many? [2 pts]

   jockey–race combinations — 9

   (d) What are the *blocks* in this experiment (if any)? How many (in *total*)? [2 pts]

   jockeys (3) + races (3) = 6

   (e) Write out a full model equation for the usual analysis, and identify each term. (Start with $y_{ijk} = \cdots$ where y_{ijk} is the finish time of horse i ridden by jockey j in race k.) [4 pts]

   mean (overall)

   $$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk}$$

   horse effect   jockey effect   race effect   error

   (f) Analysis of the finish times (seconds) using an appropriate model yields this ANOVA:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
race	2	24.667	12.333	12.3333	0.075000	.
jockey	2	290.667	145.333	145.3333	0.006834	**
horse	2	12.667	6.333	6.3333	0.136364	
Residuals	2	2.000	1.000			

   What conclusion should the experimenters make? (There should be only one!) Why? [2 pts]

   No evidence of differences in mean quarter-mile time among the horses, since the p-value = 0.136364 > 0.05.

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:07:54 GMT -05:00

https://www.coursehero.com/file/14156463/midterm2soln/

5. Consider these two models for data with two (crossed) factors:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \tag{1}$$
$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk} \tag{2}$$

where: $i = 1, 2 \qquad j = 1, 2 \qquad k = 1, \ldots, K$

(a) In model (2), which parameters are *main effects*, and which are *interaction effects*? [2 pts]

main: $\alpha_1, \alpha_2, \beta_1, \beta_2$

interaction: $\alpha\beta_{11}, \alpha\beta_{12}, \alpha\beta_{21}, \alpha\beta_{22}$

(b) How many *levels* do the first factor and the second factor have? [2 pts]

both have two

(c) What is the apparent (*not* effective) total number of *mean-related* parameters for model (1)? For model (2)? [2 pts]

model (1): $2 \times 2 = 4$

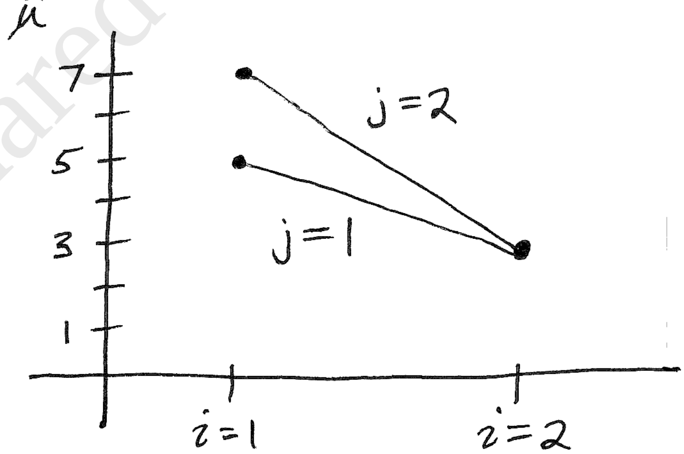
model (2): $1 + 2 + 2 + 4 = 9$

(d) What is the *error degrees of freedom* (that is, $n - p'$) for model (1)? For model (2)? (Your answers should be in terms of K.) [2 pts]

model (1): $4K - 4 = 4(K-1)$

model (2): same

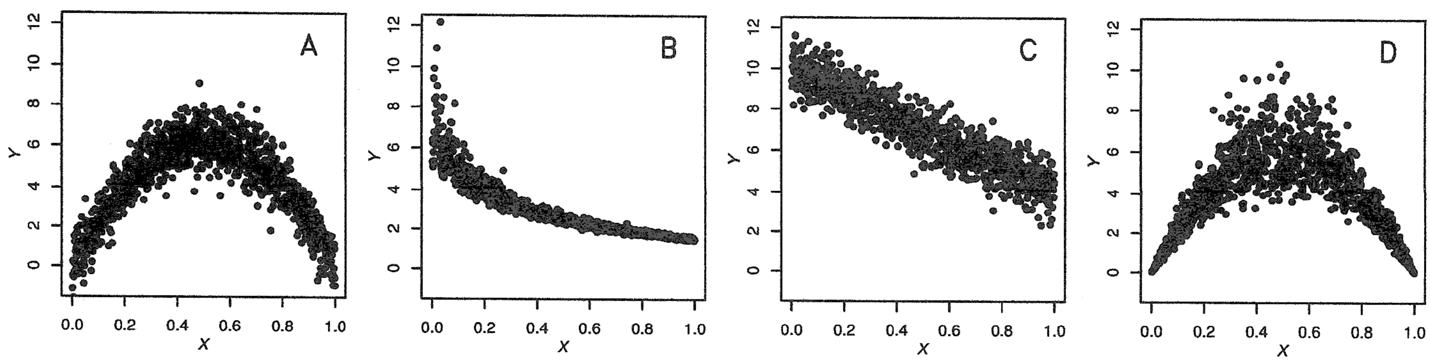
(e) Suppose least squares estimates for model (1) are $\hat{\mu}_{11} = 5$, $\hat{\mu}_{12} = 7$, $\hat{\mu}_{21} = 3$, and $\hat{\mu}_{22} = 3$. Draw an *interaction plot* with the first factor (corresponding to i) on the horizontal axis. [4 pts]



This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:07:54 GMT -05:00

https://www.coursehero.com/file/14156463/midterm2soln/

This study resource was shared via CourseHero.com



6. For each data set plotted above (A, B, C, and D), we seek transformations of X and/or Y so that a linear model satisfying the usual assumptions is appropriate. *Briefly* answer:

(a) Which (if any) do not need any transformation of X or Y? Why? [2 pts]

C only, because it has both a straight line trend and a homogeneous, roughly normal, vertical spread.

(b) For which (if any) might using a polynomial model (in X) be enough by itself? Why? [2 pts]

A (and perhaps C) because it has only the problem of curvature (and not unequal spread or non-normality)

(c) For which (if any) might a single monotone transformation of X be enough by itself? Why? [2 pts]

None (except perhaps C) because B and D also have heteroscedasticity problems, and A (and D) have a non-monotone trend.

(d) Which *one* (if any) would *certainly require* transformations of both X and Y? [1 pt]

D

(e) If the Box-Cox method (based on a SLR) were used for data set C, what would you expect the estimated power $\hat{\lambda}$ to be, approximately? Why? [2 pts]

$\hat{\lambda} \approx 1$ because that corresponds to the identity transformation (i.e. no transformation at all)

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:07:54 GMT -05:00

https://www.coursehero.com/file/14156463/midterm2soln/

STAT 425 — Section 1UG, 1GR — Spring 2016

Midterm Exam II

April 27, 2016

Full Name: Key

- This is an 80 minute exam. There are 6 problems, one of which has a part for the graduate section only.

  The exam is worth a total of 55 points for the undergraduate section and 60 points for the graduate section.

- You may use *three* pages of personal notes and a standard scientific calculator. (You may *not* share these items with anyone else.)

- *Write all answers in the spaces provided.* If you require more space to write your answer, you may use the back side of the page.

- You are not allowed to communicate with anyone except the instructor or proctors before you submit this exam.

Useful Abbreviations:

CI = confidence interval	PI = prediction interval
se = standard error	
E = expected value	Var = variance (or variance-covariance)
Cov = covariance	
SLR = simple linear regression	BLUE = best linear unbiased estimate/estimator
OLS = ordinary least squares	
GLS = generalized least squares	WLS = weighted least squares
MSE = mean square error	TSS = total sum of squares
RSS = residual sum of squares	df = degrees of freedom
VIF = variance inflation factor	VST = variance stabilizing transformation
CRD = completely randomized design	RCBD = randomized complete block design
ML = maximum likelihood	REML = restricted maximum likelihood
LRT = likelihood ratio test	ICC = intraclass correlation coefficient
H_0 = the null hypothesis of a test	H_a = the alternative hypothesis of a test

Some selected formulas:

$$AIC = n\ln(RSS/n) + 2p$$

$$BIC = n\ln(RSS/n) + \ln(n)\,p$$

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p' - n$$

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:14:59 GMT -05:00

https://www.coursehero.com/file/18063741/midterm2soln/

1. A study examined the impact of two methods for teaching sight-singing in 40 4th grade students. Students were evenly randomly assigned to control (Treatment = 0) and experimental (Treatment = 1) groups. Sight-singing test scores were collected before (Pretest) and after the experimental intervention (Posttest). Fitting a linear model with the Posttest score as response yields the following:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.1780	3.4480	1.212	0.233510	
Pretest	1.0061	0.2457	4.094	0.000229	***
Treatment	15.0953	4.7309	3.191	0.002939	**
Pretest:Treatment	-0.6383	0.3048	-2.094	0.043349	*

(a) Write out the linear model equation ($Y = \cdots$) that was apparently used. Let Y be the Posttest score, X_1 the Pretest score, and X_2 the treatment indicator. [3 pts]

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + e$$

(b) Is there evidence of an interaction effect? Support your answer using the output above. [2 pts]

Yes, the Pretest:Treatment term has $p = 0.043349 < 0.05$ indicating mild evidence for interaction.

(c) Compute the estimated *intercepts* of the Posttest-versus-Pretest relationship: one for the control group and one for the experimental group. [2 pts]

control: $\hat{\beta}_0 = 4.1780$

experimental: $\hat{\beta}_0 + \hat{\beta}_2 = 4.1780 + 15.0953 = 19.2733$

(d) Compute the estimated *slopes* of the Posttest-versus-Pretest relationship: one for the control group and one for the experimental group. [2 pts]

control: $\hat{\beta}_1 = 1.0061$

experimental: $\hat{\beta}_1 + \hat{\beta}_{12} = 1.0061 + (-0.6383) = 0.3678$

(e) What is the predicted Posttest score for a student in the experimental group who had a Pretest score of 10? [2 pts]

$$\hat{Y} = 4.1780 + 1.0061 \times 10 + 15.0953 + (-0.6383) \times 10$$

$$= 22.9513$$

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:14:59 GMT -05:00

https://www.coursehero.com/file/18063741/midterm2soln/

This study resource was shared via CourseHero.com

2. For each part below, CIRCLE the ONE BEST answer. [1 pt each]

(a) Which design necessarily has *overlapping* blocks?
Latin square   randomized complete block   split-plot   none of these

(b) In a single 4×4 Latin square design, the number of treatments is
4   8   16   none of these

(c) A randomized paired comparison experiment (like the shoes example from lecture) is a special case of which design?
split-plot   randomized complete block   Latin square   none of these

(d) Methods for estimating the *variance components* in a random effects model include
maximum likelihood   least squares   both   neither

(e) In an interaction plot, the horizontal axis represents different
factor levels   treatment means   treatments   blocks

(f) In a split-plot design, the number of levels of the whole-plot factor must be ____ the number of levels of the split-plot factor.
less than   more than   the same as   none of these

(g) In a one-factor *random* effects model, the random effects generally
sum to zero   have expected value zero   both   neither

(h) Selection of a subset of independent variables may be based on
stepwise algorithms   theoretical considerations   variable selection criteria   any of these

3. Briefly answer the following:

(a) In the context of variable selection for a linear model with many observations, which criterion would tend to choose a model with <u>fewer</u> variables: AIC or BIC? Explain. [2 pts]

BIC, because it has the same penalty for lack of fit as AIC ($n \ln(RSS/n)$) but has larger penalty (coefficient $\ln(n)$ versus 2 for AIC) for number of variables.

(b) Briefly describe how variable selection tends to affect the <u>*bias*</u> and the <u>*variance*</u> of the coefficient estimators for the retained variables. [2 pts]

Generally, variable selection <u>increases</u> the absolute bias of regression coefficient estimators, but <u>decreases</u> their variance.

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:14:59 GMT -05:00

https://www.coursehero.com/file/18063741/midterm2soln/

This study resource was shared via CourseHero.com

4. An experiment involving nine cyclists was conducted to study the effect of caffeine on cycling endurance. On each of four days, each cyclist completed an endurance test. Each day, a cyclist would receive a different dose of caffeine (0, 5, 9, or 13 mg) before the test. The order in which the doses were assigned was randomly determined, independently between cyclists. The response was endurance time (minutes) until exhaustion.

```
                  Df Sum Sq Mean Sq F value    Pr(>F)
factor(Cyclist)    8 5558.0  694.75 13.2159 4.174e-07 ***
factor(Dose)       3  933.1  311.04  5.9168  0.003591 **
Residuals         24 1261.7   52.57
```

(a) Name the design of this experiment. What role do cyclists play in this design? [2 pts]

Randomized complete block design
with cyclists as blocks

(b) In this particular experiment, what advantage might this design offer over complete randomization? [2 pts]

If cyclists have inherently different endurance levels, using them as blocks helps to cancel out those effects, making treatment comparisons more precise.

(c) What are the experimental units? How many are there? [2 pts]

The individual endurance tests (or cyclist/day combinations), of which there are $9 \times 4 = 36$.

(d) From the ANOVA, make a conclusion about the treatments. (State a p-value.) [2 pts]

$p = 0.003591 < 0.05$
so there is evidence for different levels of endurance due to different caffeine levels

(e) Based on the following 95% Tukey intervals, draw a general conclusion about the evidence for how different caffeine doses affect endurance. [2 pts]

```
           diff       lwr       upr       p adj
5-0  11.2366667  1.808030 20.665303 0.015329185
9-0  12.2411111  2.812474 21.669748 0.007661564
13-0 11.7088889  2.280252 21.137526 0.011092908
9-5   1.0044444 -8.424192 10.433081 0.990936901
13-5  0.4722222 -8.956414  9.900859 0.999031318
13-9 -0.5322222 -9.960859  8.896414 0.998616184
```

Every positive caffeine dose (5, 9, 13 mg) leads to greater endurance than no caffeine (0 mg), but there is insufficient evidence of any endurance differences among the positive doses.

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:14:59 GMT -05:00

https://www.coursehero.com/file/18063741/midterm2soln/

This study resource was shared via CourseHero.com

5. Consider the following means model (e.g. for data from a CRD):

$$Y_{ij} \;=\; \mu_i + e_{ij} \qquad i = 1,2,3 \qquad j = 1,\ldots,5$$

Suppose the least squares estimates are $\hat{\mu}_1 = 5$, $\hat{\mu}_2 = 3$, $\hat{\mu}_3 = 4$.

(a) Write out the *row* of the matrix $\boldsymbol{X}$ (in the matrix-vector formulation) that corresponds to the observation $Y_{3,2}$. Also, to which parameter does each column correspond? [2 pts]

$$\begin{array}{ccc} (0 & 0 & 1) \\ \mu_1 & \mu_2 & \mu_3 \end{array}$$

(b) Write the usual equation ($Y_{ij} = \cdots$) for the corresponding *(treatment) effects* model. [2 pts]

$$Y_{ij} = \mu + \tau_i + e_{ij}$$

(c) Compute the least squares estimates of all mean-related parameters in the *(treatment) effects* model, under the (unweighted) *sum-to-zero restriction*. [3 pts]

$$\hat{\mu} = \frac{\hat{\mu}_1 + \hat{\mu}_2 + \hat{\mu}_3}{3} = \frac{5+3+4}{3} = 4$$

$$\hat{\tau}_1 = \hat{\mu}_1 - \hat{\mu} = 5-4 = 1 \qquad \hat{\tau}_2 = 3-4 = -1 \qquad \hat{\tau}_3 = 4-4 = 0$$

(d) Give least squares estimates of all *pairwise (mean) differences*. [2 pts]

$$\hat{\mu}_1 - \hat{\mu}_2 = 5-3 = 2$$
$$\hat{\mu}_1 - \hat{\mu}_3 = 5-4 = 1$$
$$\hat{\mu}_2 - \hat{\mu}_3 = 3-4 = -1$$

(e) For this situation, consider instead a *random-effects* model with a single random factor. Write out an appropriate model equation ($Y_{ij} = \cdots$), along with *all* of the usual conditions that the terms satisfy. [4 pts]

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$

$$\left.\begin{array}{l} \alpha_i \sim N(0, \sigma_\alpha^2) \\ e_{ij} \sim N(0, \sigma^2) \end{array}\right\} \text{all independent}$$

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:14:59 GMT -05:00

https://www.coursehero.com/file/18063741/midterm2soln/

This study resource was shared via CourseHero.com

6. An experiment is conducted to determine the effects of three laundry detergent brands (A, B, C) and also of whether or not a pre-treatment is applied (yes or no). Twelve identical white T-shirts that have been soiled are *completely randomized*, such that exactly two are assigned each brand/pre-treatment combination. Each T-shirt is washed separately in its own wash load. The response, photometric brightness of the garments, is measured after washing, with the following ANOVA results:

```
                 Df  Sum Sq Mean Sq F value   Pr(>F)
deterg            2 28.2676 14.1338 19.6697 0.002318 **
pretreat          1 14.0061 14.0061 19.4921 0.004495 **
deterg:pretreat   2 12.3302  6.1651  8.5799 0.017388 *
Residuals         6  4.3113  0.7186
```

(a) Is the design *balanced*? How do you know? [2 pts]

Yes. Exactly two T-shirts are assigned each treatment (same number for all treatments).

(b) How many *treatments* are there? [1 pt]

6

(c) Is there evidence that detergent brand and pre-treatment interact? State the null and alternative hypotheses, p-value, and conclusion. [4 pts]

H_0: no interaction H_a: factors interact

Since $p = 0.017388 < 0.05$,
there is evidence for interaction.

(d) If appropriate, draw conclusions regarding the presence of the *main effects*. If not, explain why not. [2 pts]

Not appropriate, since main effects are meaningless in the presence of interaction.

(e) **[GRADUATE SECTION ONLY]** Use the other side of this page to answer: Describe an alternative design that uses 12 T-shirts, but only needs 6 wash loads. Name the design, and describe the roles of the two factors, and the T-shirts, and the wash loads. Make sure to describe the randomization.
(Hint: T-shirts in the same wash load must receive the same brand of detergent, but may be differently pre-treated.) [5 pts]

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:14:59 GMT -05:00

https://www.coursehero.com/file/18063741/midterm2soln/

Split-plot design:

Whole-plot factor: detergent brand
Split-plot factor: pre-treatment

Whole plots: wash loads
Split plots: T-shirts

Randomly assign detergent brands among loads (2 loads per brand), and randomly assign the two T-shirts in each load: one to pre-treatment, the other to no pre-treatment (independently between loads).

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:14:59 GMT -05:00

https://www.coursehero.com/file/18063741/midterm2soln/

1. (10 points) For each part below, CIRCLE ALL apropriate answers..

   (a) Which of the following estimator(s) are unbiased for the linear regression coefficients?

   (x) LASSO (xx) Ridge regression (xxx) Least Squares

   (b) How many **free parameters in total** are present in a one-way ANOVA model with 5 levels?

   (x) 5 (xx) 6 (xxx) 7

   (c) How many **free parameters in total** are present in a two-way ANOVA model with 3 levels for each factor?

   (x) 10 (xx) 9 (xxx) 12

   (d) Which of the following transformation(s) are not a by Box-Cox transformation?

   (x) $y \log y$ (xx) $\frac{1}{y^2}$ (xxx) $y + \frac{1}{y}$

   (e) In a factorial experiment, which factor (s) are used as a block?

   (x) The most important factor (xx) Factor causing variability (xxx) Any factor

2. (8 points) Briefly answer the following:

   (a) Define the variable selection criteria AIC and BIC and describe the variance and bias tradeoff they provide.

Criteria AIC&BIC try to find out the model with minimal

$AIC = \frac{RSS_p}{\hat\sigma^2} + 2p'$, $BIC = \frac{RSS_p}{\hat\sigma^2} + (\log n) p'$ respectively.

We know from homework that using small model while the true model is the larger one can introduce bias. However it could be the case that the variance becomes smaller if the true coefficient of those predictors removed are small.

   (b) When there are many predictors that are highly correlated, which estimator is preferred between least squares and ridge regression and why? Which one has larger bias? Which one has larger variance?

Ridge. It is biased while OLS gives unbiased estimators.

However, it will give much smaller variance.

when predictors are highly correlated, the estimates using OLS could be large and so are the variances, so we need to avoid them using ridge regression.

3. (15 points) The data set chickwts contains the results of an experiment with a completely randomized design: n newly hatched chicks were randomly allocated into 6 groups, and each group was given a different feed supplement. We would like to analyze how the weight (grams, after six weeks) of a chick depends on the feed that it was assigned. Use the following R output to answer the questions. Mention exactly which part of the output is being used to answer the questions.

```
> mod=lm(weight~feed)
> summary(mod)

Call:
lm(formula = weight ~ feed)

Residuals:
     Min       1Q   Median       3Q      Max
-123.909  -34.413    1.571   38.170  103.091

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    323.583     15.834  20.436  < 2e-16 ***
feedhorsebean -163.383     23.485  -6.957 2.07e-09 ***
feedlinseed   -104.833     22.393  -4.682 1.49e-05 ***
feedmeatmeal   -46.674     22.896  -2.039 0.045567 *
feedsoybean    -77.155     21.578  -3.576 0.000665 ***
feedsunflower    5.333     22.393   0.238 0.812495
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.85 on 65 degrees of freedom
Multiple R-squared:  0.5417, Adjusted R-squared:  0.5064
F-statistic: 15.36 on 5 and 65 DF,  p-value: 5.936e-10
```

(a) Describe the model being used here along with all the parameters and assumptions.

Treatment effects model

(Reference Cell) $Y_{ik} = \mu + \tau_i + \varepsilon_{ik}$

$i = 1, 2, \ldots, 6$

Assumptions, $\varepsilon_{ik} \overset{iid}{\sim} N(0, \sigma^2)$

Parameters and Estimates

μ 323.583

τ_2 -163.383

$\vdots$

τ_6 5.333

σ 54.85

constraint $\tau_1 = 0$)

(Don't need to write the estimates)

(b) What is the total number of observations n used in this experiment?

$n - p = n - 6 = 65 \Rightarrow n = 71$

(c) Test whether there are any differences among the mean weights of the groups.

Ho: No differences.

$F_{5,65} = 15.36$ $\quad P = 5.936 \times 10^{-16}$

Reject the null. There are differences among the mean weights

(d) What is the estimated mean weight of a chick feeding on "feedsoybean"? Based on the two outputs provided, can you give a valid 95% confidence interval for the difference between the mean weight corresponding to "feedsoybean" and "feedsunflower"?

$\hat{\mu}_5 = \hat{\mu} + \hat{\tau}_5 = 323.583 - 77.155$

$= 246.428$

95% CI for $\hat{\mu}_6 - \mu_5$ $(\tau_6 - \tau_5)$: $(19.126, 145.85)$.

(e) Which pairs of feeds have significantly different means? (Circle the pairs.) Is the confidence level valid simultaneously for all the pairwise comparisons? Explain.

Yes, because it provides CI for mean differences between all pairs.
Tukey's HSD

```
> Tukey=TukeyHSD(aov(weight~feed))
> Tukey
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = weight ~ feed)

$feed
                           diff         lwr       upr     p adj
horsebean-casein    -163.383333 -232.346876 -94.41979 0.0000000
linseed-casein      -104.833333 -170.587491 -39.07918 0.0002100
meatmeal-casein      -46.674242 -113.906207  20.55772 0.3324584
soybean-casein       -77.154762 -140.517054 -13.79247 0.0083653
sunflower-casein       5.333333  -60.420825  71.08749 0.9998902
linseed-horsebean     58.550000  -10.413543 127.51354 0.1413329
meatmeal-horsebean   116.709091   46.335105 187.08308 0.0001062
soybean-horsebean     86.228571   19.541684 152.91546 0.0042167
sunflower-horsebean  168.716667   99.753124 237.68021 0.0000000
meatmeal-linseed      58.159091   -9.072873 125.39106 0.1276965
soybean-linseed       27.678571  -35.683721  91.04086 0.7932853
sunflower-linseed    110.166667   44.412509 175.92082 0.0000884
soybean-meatmeal     -30.480519  -95.375109  34.41407 0.7391356
sunflower-meatmeal    52.007576  -15.224388 119.23954 0.2206962
sunflower-soybean     82.488095   19.125803 145.85039 0.0038845
```

4. (15 points) We are interested in testing the wear of a rubber-covered fabric. There are three types of fabric materials of interest: A, B, and C. The tester used for the experiment has three different postions: 1, 2 and 3. We are interested in conducting the experiment on $n = 27$ experimental units to understand the effects of the fabric materials, the different positions and how they interact.

   (a) Explain which design you recommend and specify a model that can be used to analyze the data produced from the design.

   A two-way means model (two-factor randomized design).

   $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, $i,j = 1,2,3$

   μ overall mean

   α_i, β_j main effects for fabric materials and positions.

   $(\alpha\beta)_{ij}$ interaction term. $\varepsilon_{ijk} \overset{iid}{\sim} N(0,\sigma^2)$ errors.

   (b) Provide the three null hypotheses for testing (i) whether the fabric material has any effect on the wear, (ii) whether the position of the tester has any effect, (iii) whether there is an interaction between the material and the tester.

   (i) H_0: No effect (fabric material) $Y_{ijk} = \mu + \beta_j + \varepsilon_{ijk}$

   (ii) H_0: No effect (position) $Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$

   (iii) H_0: No interaction. $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$

(c) Now suppose that another experimenter is only interested in comparing the fabric materials but would like to use the tester position as a blocking factor to reduce variability. Which experimental design and model (write it explicitly) would you use?

Randomized complete block design.

$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$. $i=1,2,3$, $j=1,2,\cdots 9$.

We need to create 9 blocks, so we have 3 blocks for each position. In each block, we have 3 units randomly assigned to 3 different fabric materials.

5. (12 points) Let us consider the simple linear regression problem: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i = 1, \cdots, n$. The ridge regression estimator $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ that minimizes

$$\ell(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 + \lambda\beta_1^2.$$

(a) Find an explicit expression for $\hat{\beta}$ by taking derivatives with respect to β_0 and β_1.

$\frac{\partial l}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \Rightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$ ①

$\frac{\partial l}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i + 2\lambda\beta_1 = 0$

$\Rightarrow \hat{\beta}_1(\sum x_i^2 + \lambda) + \hat{\beta}_0 \sum x_i - \sum x_i y_i = 0$ ②

Plug ① into ② $\Rightarrow$ $0 = \hat{\beta}_1(\sum x_i^2 + \lambda) + (\bar{y} - \hat{\beta}_1\bar{x})\sum x_i \underbrace{}_{-\sum x_i y_i} = \hat{\beta}_1(\sum x_i^2 + \lambda - n\bar{x}^2) + \sum x_i(\bar{y} - y_i) = 0$

$\Rightarrow \hat{\beta}_1 = \frac{\sum x_i(y_i - \bar{y})}{\sum x_i^2 - n\bar{x}^2 + \lambda} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2 + \lambda} = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2 + \lambda}$ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

(b) Find the bias and variance of $\hat{\beta}_1$.

$Var(\hat{\beta}_1) = \frac{\sum(x_i - \bar{x})^2 Var(y_i)}{[\sum(x_i - \bar{x})^2 + \lambda]^2} = \frac{\sigma^2 S_{xx}}{(S_{xx} + \lambda)^2}$, where $S_{xx} = \sum(x_i - \bar{x})^2$

$E(\hat{\beta}_1) = \frac{\sum(x_i - \bar{x})\beta_1 x_i}{\sum(x_i - \bar{x})^2 + \lambda} = \frac{\beta_1\sum(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2 + \lambda} = \beta_1\left(1 - \frac{\lambda}{\sum(x_i - \bar{x})^2 + \lambda}\right)$

$\Rightarrow bias(\hat{\beta}_1) = E(\hat{\beta}_1) - \beta_1 = -\frac{\lambda\beta_1}{\sum(x_i - \bar{x})^2 + \lambda} = -\frac{\lambda\beta_1}{S_{xx} + \lambda}$

(c) Find the mean squared error (MSE) of $\hat{\beta}_1$.

$MSE(\hat{\beta}_1) = bias^2(\hat{\beta}_1) + Var(\hat{\beta}_1)$

$= \frac{\lambda^2\beta_1^2}{(S_{xx} + \lambda)^2} + \frac{\sigma^2 S_{xx}}{(S_{xx} + \lambda)^2} = \frac{\lambda^2\beta_1^2 + \sigma^2 S_{xx}}{(S_{xx} + \lambda)^2}$

Practice Questions - Stat 425, Spring 2017

1. Consider the linear model $Y = X\beta + \epsilon$ where $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2\Sigma$, for known and invertible Σ.

   (a) Find $Var(\hat{\beta})$ for the ordinary least squares estimator $\hat{\beta} = (X^TX)^{-1}X^TY$. Is $\hat{\beta}$ unbiased for β?

   $E(\hat{\beta}) = E[(X'X)^{-1}X'Y] = (X'X)^{-1}X'\beta = \beta$. Unbiased.

   $Var(\hat{\beta}) = Var[(X'X)^{-1}X'Y] = (X'X)^{-1}X^T Var(Y) \cdot X(X'X)^{-1}$

   $= \sigma^2 \cdot (X'X)^{-1}X^T\Sigma X(X'X)^{-1}$

   (b) If $\Sigma = CC^T$ for a known invertible matrix C, what is the covariance matrix of $\epsilon^* = C^{-1}\epsilon$?

   $Var(\varepsilon^*) = Var(C^{-1}\varepsilon)$

   $= C^{-1} Var(\varepsilon) \cdot (C^T)^{-1}$

   $= \sigma^2 C^{-1} \cdot \Sigma \cdot (C^T)^{-1} = \sigma^2 \cdot I$

   (c) If $Y^* = C^{-1}Y$ and $X^* = C^{-1}X$. Does the linear model $Y^* = X^*\beta + \epsilon^*$ satisfy Gauss-Markov conditions? Show that the least squares estimator for this model is same as the generalized least squares estimator $\hat{\beta}_G = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$.

   YES, because of (b) and $E(\varepsilon^*) = 0$.

   $LSE(Y^*, X^*) = (X^{*T}X^*)^{-1}X^{*T}Y^*$

   $= (X^T(C^T)^{-1} \cdot C^{-1}X)^{-1} \cdot X^T(C^T)^{-1} \cdot C^{-1} \cdot Y$

   $= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$.

   (d) Find $Var(\hat{\beta}_G)$. Is $\hat{\beta}_G$ unbiased for β?

   $E(\hat{\beta}_G) = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1} \cdot X \cdot \beta = \beta$. Unbiased.

   $Var(\hat{\beta}_G) = \sigma^2 (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1} \cdot \Sigma \cdot \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}$

   $= \sigma^2 \cdot (X^T\Sigma^{-1}X)^{-1}$

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:10:36 GMT -05:00

Scanned by CamScanner

https://www.coursehero.com/file/31737964/Exam2-Practice-Solutions1pdf/

2. A sample of 654 youths of ages 3 to 19 was collected in East Boston during the middle to late 1970s. Researchers measured the forced expiratory volume (FEV) of each youth as a measure of lung capacity. Using the response Y = log(FEV), a quadratic model in the variable Age (which has integer values only) was t by least squares with these results:

$$\hat{Y} = -0.55 + 0.21 \times Age - 0.0058 \times (Age)^2 \qquad RSS = 26.036$$

(a) The transformation of FEV was suggested by the Box-Cox procedure. What λ value was apparently chosen?

$\lambda = 0$ because

$$\log(FEV) \simeq \lim_{\lambda \to 0} \frac{(FEV)^{\lambda} - 1}{\lambda}.$$

(b) Determine the vector $\hat{\beta}$ of least squares regression coefficients.

$$\hat{\beta} = \begin{pmatrix} -0.55 \\ 0.21 \\ -0.0058 \end{pmatrix}$$

(c) In the matrix-vector form $Y = X\beta + \epsilon$ for this regression, what would be the dimensions of X?

654×3

(d) Compute the usual (unbiased) estimate of error variance. Show your work.

$$\hat{\sigma}^2 = \frac{RSS}{n-p'} = \frac{26.036}{654-3} \simeq 0.04$$

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:10:36 GMT -05:00

Scanned by CamScanner https://www.coursehero.com/file/31737954/Exam2-Practice-Solutions1pdf/

(e) Predict the FEV of a 10-year-old. (Note: log is the natural logarithm.)

$$\hat{y} = -0.55 + 0.21 \times 10 - 0.0058 \times 100$$

$$\approx 0.97$$

$$FEV = e^{\hat{y}} \approx 2.64$$

(f) The simple linear regression of log(FEV) on Age has a residual sum of squares of 29.316. Compute the F -statistic for testing whether the quadratic term is needed. Also, state a conclusion based on the critical value $F_{0.05,m_1,m_2} \approx 3.856$. Show your work.

$$F = \frac{(29.316 - 26.036)/1}{26.036/651} \approx 82 > 3.856$$

THERE IS EVIDENCE TO SUPPORT PRESENCE OF QUADRATIC TERM.

3. For each part below, CIRCLE the ONE BEST answer..

(a) In a factorial experiment, which factor is used as a **block**?
(x) Treatment factor (xx) Factor causing variability (xxx) Any random factor

(b) Which one below is likely to produce **exact zeroes** for the linear regression estimates?
(x) LASSO (xx) Ridge regression (xxx) Least Squares

(c) In the classroom example of shoe experiment with boys, what are the **blocks**?
(x) Right or Left foot (xx) Shoe material type (xxx) Boys

(d) Which of the following criteria provides a **smaller model**?
(x) Mallow's C_P (xx) AIC (xxx) BIC

(e) How many **blocking factors** are present in a 4 × 4 Latin square design?
(x) 1 (xx) 2 (xxx) 4

4. Briefly answer the following:

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:10:36 GMT -05:00

https://www.coursehero.com/file/31737964/Exam2-Practice-Solutions1pdf/

Scanned by CamScanner

(a) When there are many predictors that are highly correlated, which estimator is preferred between least squares and lasso, and why? Which one has larger bias? Which one has larger variance?

LASSO IS PREFERRED BECAUSE IT INTRODUCES SHRINKAGE WHEREAS LEAST SQUARES SUFFERS WHEN (& SPARSITY) THERE ARE MANY PREDICTORS. LASSO HAS LARGER BIAS WHEREAS OLS IN GENERAL HAS LARGER VARIANCE.

= BE = FS

(b) Describe the methods of best subset selection, backward elimination, and forward selection to perform variable selection. What are some pros and cons of each of these methods?

BEST SUBSET SELECTION FINDS THE BEST COMBINATION OF VARIABLES FROM ALL POSSIBLE COMBINATIONS. BE SEQUENTIALLY, FINDS VARIBLES TO ELIMINATE STARTING FROM THE FULL MODEL. LIKEWISE FS STARTS FROM NULL. BEST SUBSETS HAS GOOD PROPERTIES B/C IT EXPLORES ALL MODELS BUT RESTRICTIVE FOR IMPLEMENTING WHEN # OF VARIABLES LARGE. BE & FS ARE MORE APPEALING IN THAT CASE.

5. We are interested in testing the wear of a rubber-covered fabric. There are three types of fabric materials of interest: A, B, and C. The tester used for the experiment has two factors (1) three positions of the tester, (2) three different times for setting up the tester. We are interested in conducting the experiment on $n = 9$ experimental units.

In each of the below cases, explain which design you recommend along with an example design. Also, specify a model that can be used to analyze the data produced from each design.

(a) It is known that the position of the tester and the different times of testing do not matter for measuring the wear.

WE WOULD SELECT A COMPLETELY RANDOM DESIGN SINCE THE FACTORS POSITIONS & TIMES ARE KNOWN NOT TO HAVE ANY EFFECT. A SAMPLE DESIGN IS

	P_1	P_2	P_3
T_1	A	A	C
T_2	B	C	B
T_3	C	B	A

MODEL: TREATMENT EFFECT MODEL:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \rightarrow \text{errors}$$

τ_i: i^{th} treatment

y_{ij}: Obs. for i^{th} trt, j^{th} unit.

This study source was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:10:36 GMT -05:00

Scanned by CamScanner https://www.coursehero.com/file/31737954/Exam2-Practice-Solutions1pdf/

(b) It is believed that the wear measurement may vary depending on the position of the tester.

IN THIS CASE, WE USE A RANDOMIZED BLOCK DESIGN.
POSITION OF THE TESTER IS BLOCKING FACTOR.

Eg.

	P_1	P_2	P_3
T_1	B	C	B
T_2	C	A	A
T_3	A	B	C

Model :

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

i^{th} trt, j^{th} Position, Trt eff, Block effect, error.

(c) It is believed that the wear measurement may vary based on both the position of the tester and the time of testing.

LATIN SQUARE DESIGN AS BOTH MATTER.

Eg: design

	P_1	P_2	P_3
T_1	A	C	B
T_2	B	A	C
T_3	C	B	A

Model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk}$$

Trt Position Time error

2 Blocking Factors
1 Treatment Factor

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:10:36 GMT -05:00

Scanned by CamScanner https://www.coursehero.com/file/31787854/Exam2-Practice-Solutions1pdf/

STAT 426 — Sections 1GR, 1UG — Spring 2019

Exam 2

March 27, 2019

Full Name: Key ID/Email: ______________

- This is a 50 minute exam. There are 4 problems, one of which is for the graduate section only.

  The exam is worth a total of 35 points for the undergraduate section and 45 points for the graduate section.

- You may use *two* physical pages of personal notes and a standard scientific calculator. (You may *not* share these items with anyone else.)

- *Write all answers in the spaces provided.* If you require more space to write your answer, you may use the back side of the page.

- You are not allowed to communicate with anyone except the instructor or proctors before you submit this exam.

Useful Abbreviations:

CI = confidence interval
SE = standard error
E = expected value var = variance cov = covariance
df = degrees of freedom
ML = maximum likelihood LRT = likelihood ratio test L = log-likelihood
RR = relative risk
GLM = generalized linear model
H_0 = the null hypothesis of a test H_a = the alternative hypothesis of a test

Some selected formulas:

Poisson pdf: $$p(y) = \frac{\mu^y e^{-\mu}}{y!} \qquad y = 0, 1, 2, \ldots$$

binomial pdf: $$p(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y} \qquad y = 0, \ldots, n$$

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\nu(\hat{\mu}_i)}} \qquad d_i = -2\big(L(\hat{\mu}_i; y_i) - L(y_i; y_i)\big) \qquad r_i = \frac{e_i}{\sqrt{1-\hat{h}_i}}$$

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:08:42 GMT -05:00

https://www.coursehero.com/file/40278612/exam2soln1pdf/

1. In the month of May (which always has 31 days), the number of days with at least one US tornado report is modeled as binomial, with probability π satisfying

$$\text{logit}(\pi) \;=\; \alpha + \beta \cdot \text{Year}$$

where Year is the year number (e.g., 2018). For data from 2005 to 2018 (assuming independence between years), the ML estimates (with standard errors) are

$$\hat{\alpha} \;\approx\; 146.73 \;\; (48.90) \qquad\qquad \hat{\beta} \;\approx\; -0.07289 \;\; (0.02431)$$

(a) Briefly state why these counts were *not* modeled with a Poisson distribution. [2 pts]

The counts have a maximum of 31, whereas a Poisson random variable has no upper bound.

(b) What is the link function? Is it canonical? [2 pts]

logit, which is canonical (for a binomial)

(c) Perform a Wald test (level 0.05) for whether there is a year effect. Interpret. [4 pts]

$$z \approx \frac{-0.07289}{0.02431} \approx -3.0 \qquad |z| > z_{0.025} \approx 1.96$$

There appears to be evidence for a year effect: The number of May days with a report appears to decrease, on average.

(d) Estimate the *mean* of the number of days in May 2019 that will have at least one US tornado report. [5 pts]

$$\text{logit}(\hat{\pi}_{2019}) \approx 146.73 + (-0.07289) \cdot 2019$$
$$\approx -0.435$$

$$\hat{\pi}_{2019} \approx \frac{e^{-0.435}}{1 + e^{-0.435}} \approx 0.393$$

So the mean number of May 2019 days is estimated to be

$$31 \cdot \hat{\pi}_{2019} \approx 12.2$$

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:08:42 GMT -05:00

https://www.coursehero.com/file/40278612/exam2soln1pdf/

This study resource was shared via CourseHero.com

2. Let response Y be the number of far-right extremism terrorism incidents in a given year and region. Consider loglinear models having the following linear predictors and deviances:

Model 1:	$\eta = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2$	$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1) \approx 46.782$
Model 2:	$\eta = \alpha + \beta_1 X_1 + \beta_2 X_2$	$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_2) \approx 49.792$

where X_1 is year number (e.g., 2018), and X_2 indicates region (0 = Western Europe, 1 = North America). The data are from 2008 to 2017, for a total of 20 observations.

(a) Which of these two models, if any, is *saturated*? Justify your answer. [2 pts]

Neither, since the saturated model would need 20 parameters (one for each obs.).

(b) Assuming Model 1 is correct, perform a likelihood ratio test (level 0.05) for whether the annual multiplicative change in the mean number of incidents depends on the region. Be sure to state H_0 and your conclusion. [$\chi^2_1(0.05) \approx (1.96)^2$] [5 pts]

$H_0: \beta_{12} = 0$ (annual change does not depend on region)

$$D(y; \hat{\mu}_2) - D(y; \hat{\mu}_1) \approx 49.792 - 46.782$$
$$= 3.01 < 3.84 \approx \chi^2_1(0.05)$$

So we fail to reject H_0.
There is not sufficient evidence that the annual mult. change depends on region.

(c) For a deviance-based goodness of fit test for Model 1, give the values of the chi-squared statistic and the degrees of freedom. [2 pts]

stat: 46.782 df: $20 - 4 = 16$

(d) The sum of the squared Pearson residuals for Model 1 is about 41.4. Estimate the dispersion parameter (in the quasi-likelihood). What does its value suggest? [2 pts]

$\hat{\phi} \approx \frac{41.4}{16} \approx 2.59$ which exceeds 1, suggesting overdispersion

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:08:42 GMT -05:00

https://www.coursehero.com/file/40278612/exam2soln1pdf/

3. A simple binary logistic regression with (success) probability $\pi(x)$ has linear predictor

$$\operatorname{logit}(\pi(x)) \;=\; \alpha + \beta x$$

Suppose

$\pi(2) = 0.5$ and the *odds* at $x = 3$ is 0.25 times the *odds* at $x = 2$.

(a) Determine the ratio of the *odds* at $x = 4$ to the *odds* at $x = 3$. [2 pts]

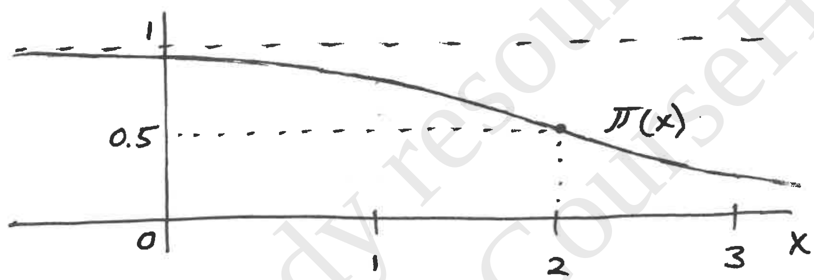
$$\frac{\text{odds}(x=4)}{\text{odds}(x=3)} = e^{\beta} = \frac{\text{odds}(x=3)}{\text{odds}(x=2)} = 0.25$$

(b) Compute the *slope* of $\pi(x)$ at $x = 2$. [2 pts]

$$\frac{d}{dx}\pi(2) = \beta\, \pi(2)(1-\pi(2)) = \ln(0.25)\cdot 0.5 \cdot (1-0.5)$$

$$\approx -0.35$$

(c) Sketch $\pi(x)$ versus x (with reasonable accuracy). [3 pts]



(d) Determine the *median effective level* (of x). [1 pt]

$x = 2$, since $\pi(2) = \frac{1}{2}$

(e) Compute the distance (in x units) between the x values for which $\pi(x) = 0.25$ and $\pi(x) = 0.75$. [3 pts]

Say $\pi(x_1) = 0.25$ and $\pi(x_2) = 0.75$

$$\alpha + \beta x_1 = \operatorname{logit}(0.25) = \ln(1/3)$$
$$\alpha + \beta x_2 = \operatorname{logit}(0.75) = \ln(3)$$

$$\Rightarrow \quad |\beta(x_1 - x_2)| = |\ln(1/3) - \ln(3)|$$

$$\Rightarrow \quad |x_1 - x_2| = \frac{|\ln(1/9)|}{|\ln(1/4)|} \approx 1.585$$

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:08:42 GMT -05:00

https://www.coursehero.com/file/40278612/exam2soln1pdf/

GRADUATE SECTION ONLY

4. For independent counts $Y_1, \ldots, Y_n$, consider a Poisson loglinear *rate* model with the *only* parameter being an intercept α (i.e., no explanatory variables), and with the count means proportional to observed positive constants $t_1, \ldots, t_n$, respectively.

(a) Derive a simplified expression for the log-likelihood. (You may drop terms without α.) [5 pts]

$$E(Y_i) = \mu_i = t_i \lambda_i = t_i e^{\alpha}$$

$$L(\alpha; y) = \ln\left(\prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i}\right)$$

$$= \sum_{i=1}^{n} (y_i \ln \mu_i - \mu_i - \ln y_i!)$$

$$= \sum_{i=1}^{n} (y_i \ln(t_i e^{\alpha}) - t_i e^{\alpha} - \ln y_i!)$$

$$= \sum_{i=1}^{n} (y_i \alpha - t_i e^{\alpha}) + \text{constants}$$

↑ (not depending on α)

(b) Write an expression for the *likelihood equation.* [2 pts]

$$\frac{\partial}{\partial \alpha} L(\alpha; y) = \sum_{i=1}^{n} (y_i - t_i e^{\alpha})$$

so the likelihood equation is this set to zero:

$$\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} t_i\right) e^{\alpha} = 0$$

(c) Find an explicit expression for the ML estimator. When does it exist? [3 pts]

$$\hat{\alpha} = \ln\left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} t_i}\right)$$

which exists when $\sum_{i=1}^{n} y_i > 0$.

This study resource was shared via CourseHero.com

This study source was downloaded by 100000824785287 from CourseHero.com on 04-28-2021 17:08:42 GMT -05:00

https://www.coursehero.com/file/40278612/exam2soln1pdf/

1. (10 points) For each part below, CIRCLE ALL apropriate answers..

   (a) Which of the following estimator(s) are unbiased for the linear regression coefficients?

   (x) LASSO (xx) Ridge regression (xxx) Least Squares

   (b) How many **free parameters in total** are present in a one-way ANOVA model with 5 levels?

   (x) 5 (xx) 6 (xxx) 7

   (c) How many **free parameters in total** are present in a two-way ANOVA model with 3 levels for each factor?

   (x) 10 (xx) 9 (xxx) 12

   (d) Which of the following transformation(s) are not a by Box-Cox transformation?

   (x) $y \log y$ (xx) $\frac{1}{y^2}$ (xxx) $y + \frac{1}{y}$

   (e) In a factorial experiment, which factor (s) are used as a block?

   (x) The most important factor (xx) Factor causing variability (xxx) Any factor

2. (8 points) Briefly answer the following:

   (a) Define the variable selection criteria AIC and BIC and describe the variance and bias tradeoff they provide.

   Criteria AIC&BIC try to find out the model with minimal

   $AIC = \frac{RSS_{p}}{\hat{\sigma}^2} + 2p'$, $BIC = \frac{RSS_p}{\hat{\sigma}^2} + (\log n) p'$ respectively.

   We know from homework that using small model while the true model is the Larger one can introduce bias. However it could be the case that the variance becomes smaller if the true coefficient of those predictors removed are small.

   (b) When there are many predictors that are highly correlated, which estimator is preferred between least squares and ridge regression and why? Which one has larger bias? Which one has larger variance?

   Ridge. It is biased while OLS gives unbiased estimators.

   However, it will give much smaller variance.

   when predictors are highly correlated, the estimates using OLS could be large and so are the variances, so we need to avoid them using ridge regression.

3. (15 points) The data set chickwts contains the results of an experiment with a completely randomized design: n newly hatched chicks were randomly allocated into 6 groups, and each group was given a different feed supplement. We would like to analyze how the weight (grams, after six weeks) of a chick depends on the feed that it was assigned. Use the following R output to answer the questions. Mention exactly which part of the output is being used to answer the questions.

```
> mod=lm(weight~feed)
> summary(mod)

Call:
lm(formula = weight ~ feed)

Residuals:
     Min       1Q   Median       3Q      Max 
-123.909  -34.413    1.571   38.170  103.091 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    323.583     15.834  20.436  < 2e-16 ***
feedhorsebean -163.383     23.485  -6.957 2.07e-09 ***
feedlinseed   -104.833     22.393  -4.682 1.49e-05 ***
feedmeatmeal   -46.674     22.896  -2.039 0.045567 *  
feedsoybean    -77.155     21.578  -3.576 0.000665 ***
feedsunflower    5.333     22.393   0.238 0.812495    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.85 on 65 degrees of freedom
Multiple R-squared:  0.5417,	Adjusted R-squared:  0.5064 
F-statistic: 15.36 on 5 and 65 DF,  p-value: 5.936e-10
```

(a) Describe the model being used here along with all the parameters and assumptions.

Treatment effects model

(Reference Cell) $Y_{ik} = \mu + \tau_i + \varepsilon_{ik}$

$i = 1, 2, \ldots, 6$

Assumptions, $\varepsilon_{ik} \overset{iid}{\sim} N(0, \sigma^2)$

Parameters and Estimates

μ 323.583

τ_2 −163.383

⋮

τ_6 5.333

σ 54.85

constraint $\tau_1 = 0$)

(Don't need to write the estimates)

(b) What is the total number of observations n used in this experiment?

$n - p = n - 6 = 65 \Rightarrow n = 71$

(c) Test whether there are any differences among the mean weights of the groups.

Ho: No differences.

$F_{5,65} = 15.36$ $p = 5.936 \times 10^{-16}$

Reject the null. There are differences among the mean weights

(d) What is the estimated mean weight of a chick feeding on "feedsoybean"? Based on the two outputs provided, can you give a valid 95% confidence interval for the difference between the mean weight corresponding to "feedsoybean" and "feedsunflower"?

$\hat{\mu}_5 = \hat{\mu} + \hat{\tau}_5 = 323.583 - 77.155$

$= 246.428$

95% CI for $\hat{\mu}_6 - \mu_5$ $(\tau_6 - \tau_5)$: $(19.126, 145.85)$.

(e) Which pairs of feeds have significantly different means? (Circle the pairs.) Is the confidence level valid simultaneously for all the pairwise comparisons? Explain.

Yes, because it provides CI for mean differences between all pairs.
Tukey's HSD

```
> Tukey=TukeyHSD(aov(weight~feed));
> Tukey
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = weight ~ feed)

$feed
                           diff        lwr       upr     p adj
horsebean-casein    -163.383333 -232.346876 -94.41979 0.0000000
linseed-casein      -104.833333 -170.587491 -39.07918 0.0002100
meatmeal-casein      -46.674242 -113.906207  20.55772 0.3324584
soybean-casein       -77.154762 -140.517054 -13.79247 0.0083653
sunflower-casein       5.333333  -60.420825  71.08749 0.9998902
linseed-horsebean     58.550000  -10.413543 127.51354 0.1413329
meatmeal-horsebean   116.709091   46.335105 187.08308 0.0001062
soybean-horsebean     86.228571   19.541684 152.91546 0.0042167
sunflower-horsebean  168.716667   99.753124 237.68021 0.0000000
meatmeal-linseed      58.159091   -9.072873 125.39106 0.1276965
soybean-linseed       27.678571  -35.683721  91.04086 0.7932853
sunflower-linseed    110.166667   44.412509 175.92082 0.0000884
soybean-meatmeal     -30.480519  -95.375109  34.41407 0.7391356
sunflower-meatmeal    52.007576  -15.224388 119.23954 0.2206962
sunflower-soybean     82.488095   19.125803 145.85039 0.0038845
```

4. (15 points) We are interested in testing the wear of a rubber-covered fabric. There are three types of fabric materials of interest: A, B, and C. The tester used for the experiment has three different postions: 1, 2 and 3. We are interested in conducting the experiment on $n = 27$ experimental units to understand the effects of the fabric materials, the different positions and how they interact.

   (a) Explain which design you recommend and specify a model that can be used to analyze the data produced from the design.

   A two-way means model (two-factor randomized design).

   $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, $i,j = 1,2,3$

   μ overall mean

   α_i, β_j main effects for fabric materials and positions.

   $(\alpha\beta)_{ij}$ interaction term. $\varepsilon_{ijk} \overset{iid}{\sim} N(0,\sigma^2)$ errors.

   (b) Provide the three null hypotheses for testing (i) whether the fabric material has any effect on the wear, (ii) whether the position of the tester has any effect, (iii) whether there is an interaction between the material and the tester.

   (i) H_0: No effects (fabric material) $Y_{ijk} = \mu + \beta_j + \varepsilon_{ijk}$

   (ii) H_0: No effect (position) $Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$

   (iii) H_0: No interaction. $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$

(c) Now suppose that another experimenter is only interested in comparing the fabric materials but would like to use the tester position as a blocking factor to reduce variability. Which experimental design and model (write it explicitly) would you use?

Randomized complete block design.

$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$. $i=1,2,3$, $j=1,2,\cdots 9$.

We need to create 9 blocks, so we have 3 blocks for each position. In each block, we have 3 units randomly assigned to 3 different fabric materials.

5. (12 points) Let us consider the simple linear regression problem: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i = 1, \cdots, n$. The ridge regression estimator $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ that minimizes

$$\ell(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 + \lambda\beta_1^2.$$

(a) Find an explicit expression for $\hat{\beta}$ by taking derivatives with respect to β_0 and β_1.

$$\frac{\partial l}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \Rightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x} \quad ①$$

$$\frac{\partial l}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i + 2\lambda\beta_1 = 0$$

$$\Rightarrow \hat{\beta}_1(\sum x_i^2 + \lambda) + \hat{\beta}_0\sum x_i - \sum x_i y_i = 0 \quad ②$$

Plug ① into ② $\Rightarrow$ $0 = \hat{\beta}_1(\sum x_i^2 + \lambda) + (\bar{y} - \hat{\beta}_1\bar{x})\sum x_i \underset{-\sum x_i y_i}{} = \hat{\beta}_1(\sum x_i^2 + \lambda - n\bar{x}^2) + \sum x_i(\bar{y} - y_i) = 0$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum x_i(y_i - \bar{y})}{\sum x_i^2 - n\bar{x}^2 + \lambda} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2 + \lambda} = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2 + \lambda} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

(b) Find the bias and variance of $\hat{\beta}_1$.

$$Var(\hat{\beta}_1) = \frac{\sum(x_i - \bar{x})^2 Var(y_i)}{[\sum(x_i - \bar{x})^2 + \lambda]^2} = \frac{\sigma^2 S_{xx}}{(S_{xx} + \lambda)^2}, \text{ where } S_{xx} = \sum(x_i - \bar{x})^2$$

$$E(\hat{\beta}_1) = \frac{\sum(x_i - \bar{x})\beta_1 x_i}{\sum(x_i - \bar{x})^2 + \lambda} = \frac{\beta_1\sum(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2 + \lambda} = \beta_1\left(1 - \frac{\lambda}{\sum(x_i - \bar{x})^2 + \lambda}\right)$$

$$\Rightarrow bias(\hat{\beta}_1) = E(\hat{\beta}_1) - \beta_1 = -\frac{\lambda\beta_1}{\sum(x_i - \bar{x})^2 + \lambda} = -\frac{\lambda\beta_1}{S_{xx} + \lambda}$$

(c) Find the mean squared error (MSE) of $\hat{\beta}_1$.

$$MSE(\hat{\beta}_1) = bias^2(\hat{\beta}_1) + Var(\hat{\beta}_1)$$

$$= \frac{\lambda^2\beta_1^2}{(S_{xx} + \lambda)^2} + \frac{\sigma^2 S_{xx}}{(S_{xx} + \lambda)^2} = \frac{\lambda^2\beta_1^2 + \sigma^2 S_{xx}}{(S_{xx} + \lambda)^2}$$