Written Comprehensive Examination-Methods
Department of Statistics, University of California, Irvine
Monday, June 20, 2016, 9:00am to noon

- There are 4 questions on the examination, each with multiple parts. Select any 3 of them to solve.

- Problems 1 and 2 cover the material from Stat 210 and are the same for everyone. You will be given the appropriate Problems 3 and 4 based on the courses you took:

  - Problems 3 and 4 covering Stat 202-203 OR
  - Problems 3 and 4-covering Stat 211.:212

- Make sure you choose the appropriate problems 3 and 4 for the courses you took. The problems are labeled with both the problem number and the course numbers.

- Your solutions to each of the 3 problems you solve should be written on separate sheets · of paper. Label *each sheet* with your student identification number, the problem number, and the page for that problem written in the upper right hand comer. For Problems 3 and 4, include the appropriate course number as well. For example, the labeling on a. page might be:

ID# 912345678
Problem 3 (202-203), Page 2

- You have three hours to complete your solution. Please be prepared to turn in your exam at 12:00 noon.

Methods Problem 1

The data used for this problem were collected by someone who wears a personal fitness device, which measures how many steps he takes and how many miles he walks each day. He can set the device to register separate readings for easy, moderate and brisk walking. The unit of analysis for all parts of this scenario is one day, and he has provided data for 42 days.
The response variable for all parts is:
Y = total *Miles* walked that day at all speeds (easy, moderate, brisk)
The following explanatory variables are available, with the variable or factor name in bold italics:
The number of steps taken at a *Moderate* pace that day
The number of *Min* (minutes) spent walking at a moderate pace that day
Factor A: Whether or not it was a *Weekend* day (Saturday or Sunday)
Factor B: Whether or not there was *Rain* that day

1. For this question (parts a to c), only-the two factors *Weekendand Rain*_will be used, arrdnot the quantitative explanatory variables. The table below shows the means number of miles walked per day for each combination.

	Rain	Shine
Weekday	3.6	2.6
Weekend	2.7	2.6

a. Create a cell means (interaction) plot. Put the *"Weekend'* factor on the horizontal axis.

b. Discuss whether or not there is a Factor A (Weeke9d) effect, a Factor B (Rain) effect, and an interaction effect in the sample data. Support your answer with specifics about the plot.

c. The person wearing the device has not looked at the data, and asks you ifthere is a difference in how far he walks on weekends and weekdays. What would you tell him? Give the mest complete story available from the data shown in the table of means and interaction plot.

For Questions 2 to 7, only the two quantitative explanatory variables of *"Moderate"* and *"Min"* will be used, and not the Factors. Two models were run in R, and the output is on the separate "First output page." Model I included *Moderate* only, while Model 2 included *Moderate* and *Min.*

2. Use Model I to predict how many *Miles* the person walks on a day that he walks 2000 steps at a moderate pace. (I.e., *Moderate* = 2000.)

3. Give the value of the *Intercept* for the regression equation in Model I.Does the intercept have a useful interpretation in this case? Ifso, provide the interpretation. Ifnot, explain why not.

4. For Model 2, the variance inflation factor (VIP) for *Moderate* is 559.08. What is the VIP for *Min?* Explain how you know. You can do this in words or using the formula (or a. combination of both).

5. Define X1 = *Moderate* and X_2 = *Min.* Using standard notation and X's instead of variable names, write the full and reduced models being tested by the following test statistics and *p*-values taken from the output.

   a. For Model 1, the test shown with *F* = 157.2,p = 1.932e-15.

   b. For Model 2, the test in the "Moderate" row of the first table, with *t* = -0.874, *p* = 0.387

   c. For Model 2, the test in the "Moderate" row of the anova table, with *F* = 161.1551,p = 1.983e-15

6. In Model 1, the testfor whether "Moderate" needs to be in the model results in F = 157.2,p = l.932e-15. In Model 2, the test for "Moderate" shown in the anova table gives slightly different values: *F* = 161.1551, p = 1.983e-15. Explain why the two tests give different results. In other words, what is different in the computation of the F statistic for the two versions?

7. In the first table shown for Model 2, neither variable (Moderate and Min) is statistically significant, with p-values of 0.387 and 0.166, respectively. Yet the overall F test has a *p*-value of 1.162e-14, indicating that at least one of the variables is certainly needed in the model. Explain this discrepancy.

8. For this question (parts a to c) use the Minitab output on the separate "Second output page" with results from running a "best subsets" analysis with the following variables:

   Min: The number of minutes spent walking at a moderate pace that day
   RainRain: Indicator variable = 1 if it rained that day, 0 if it did not
   Rain xMin: The product of Min and RainRain

   a. Which variable(s) would you include in the "best" model? On what did you base your decision?

   ·b. On the same output page as the best subsets analysis there is a scatterplot of Min vs Miles. On the plot on that page draw a line or lines showing what the model you chose in (a) would look like. Include the page when you tum in your exam, and be sure to write your ID number on it.

   c. Based on the results shown in the best subsets output only, give a reason why the model with *"Min"* only is notan acceptable model.

FIRST OUTPUT PAGE FOR METHODS PROBLEM 1

OUTPUT FOR QUESTIONS 2 TO 7

MODEL 1: Model with "Moderate" as the only predictor

```
> Modell<-lm(MilesModerate)
Coefficients:
               Estimate Std. Error t value Pr(>ltl )
(Intercept)   1.9119105  0.1184585 ·  16.14  < 2e-16 ***
Moderate     0.0005555  0.0000443    12.54 l.93e-15 ***

Residual standard error: 0.4525 on 40 degrees of freedom
Multiple R-squared:  0.7972,    Adjusted R-squared:  0.7921
F-statistic: 157.2 on 1 and 40 DF,  p-value: 1.932e-15
```

MODEL 2: Model with "Moderate" and "Min"

```
> Model2<-lm(MilesModerate+Min)
> summary(Mode12)
Coefficients:
               Estimate Std. Error  t value Pr(>ltl )
(Intercept)   1.9103041  0.1170179  16.325   <2e-16 ***
Moderate    -0.0009045  0.0010346  -0.874    0.387
Min          0.1684050  0.1192346   1.412    0.166

Residual standard error: 0.447 on 39 degrees of freedom
Multiple R-squared:  0.8071,    Adjusted R-squared:  0.7972
F-statistic: 81.57 on 2 and 39 DF,  p-value: 1.162e-14

> an.ova(Mode12)
Analysis of Variance Table
Response: Miles
          Df Sum Sq Mean Sq  F value    Pr(>F)
Moderate   1 32.194  32.194 161.1551 1.983e-15 ***
Min        1  0.399   0.399   1.9948    0.1658
Residuals 39  7.791   0.200
```

SECOND OUTPUT PAGE FOR METHODS PROBLEM 1

OUTPUT FOR QUESTION 8:

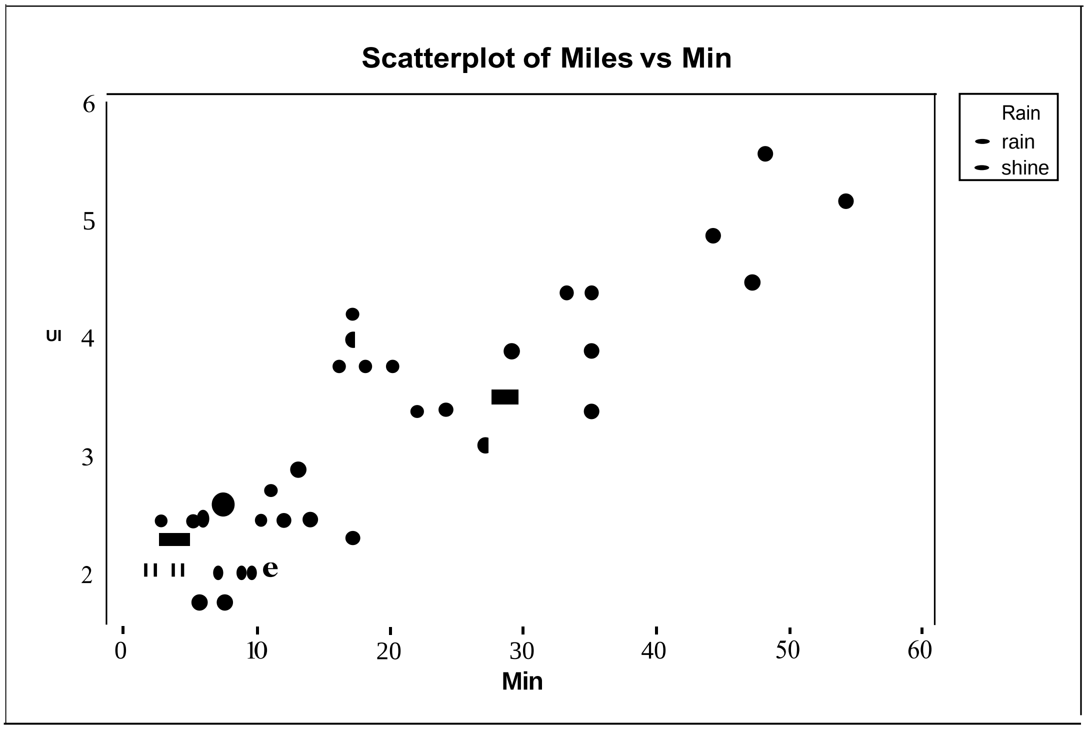
Best Subsets Regression: Miles versus Min, Rain_rain, RainxMin

```
Response is Miles

                                                       R
                                                       a R
                                                       i a
                                                       n i
                                                         n
                                                       r x
                                                     M a M
                               Mallows               i i i

Vars  R-Sq  R-Sq (adj)         Cp         S          n n n
   1  80.3        79.8        6.8   0.44564          X
   1  61.6        60.6       49.5   0.62264              X
   2  83..2       82.4        2.2   0.41648          X   X
   2  81.7        80.8        5.6   0.43501          X X
   3  83.3        82.0        4.0   0.42101          X X X
```

PLOT FOR METHODS PROBLEM 1, QUESTION 8b:



Methods Problem 2

People tend to get more exercise during the summer, when the weather is nicer and they have more time to be outside. A study is being designed to see if on average people weigh less in August then they do in February, and if so, to predict their August weight from their February weight. Fifty adult males and 50 adult females will provide data on their weights at those times.

Let *Yilnibe* the weight for person *i* (1 to 100) in Month *m* (F for February or A for August) for sexj (0 for females and 1 for males). The data can be represented as follows:

Female Feb *(YiFo)*	Female Aug (YiAo)	Male Feb (YiF1)	Male Aug (YiA1)
Y1Fo	Y1AO	Ysm	Ys1A1
Y2Fo	Y2AO	Ys2P1	Ys2A1
J3po	Y2AO	Ys3p1	Ys3Al
YsoPo	YsoAo	Y100P1	Y100A1

Everyone is measured in both February and August, so the data are paired and thus correlated within sexes, but the male and female values are independent of each other.

There are 4 researchers involved in the study, and they each have a different idea for how the data should be analyzed. The four methods proposed are described below, with questions inserted.

Method 1:
The 1st researcher thinks that .Yimj N(µmj, cr2), where the subscript *i* is the individual (1 to 50 for females and 51 to 100 for males), *m* is F or A for the month, and*j* is 0 or 1 for female/male.
He proposes to do three separate tests - a paired t-test for females to see if the mean difference is 0, a paired t-test for males to see if the mean difference is 0, with the alternative hypothesis in both cases that February weights are on average higher than August weights, and a two-sample t-test to test whether the seasonal differences for males and females are the same.

This method can be modeled using the following combinations of means:
Feb -Aug for females: µpo - µAo = *ko*
Feb -Aug for males: µp1 - µA1 = *ki*
Difference of differences: *ko* - *ki* = *D*

Question 1: Write the null and alternative hypotheses for the 3 tests using the symbols defined above. (Throughout this problem please remember that the subscript F stands for February, not for female.)

Method 2:
The 2nd researcher proposes to do simultaneous linear tests using the format **cp = h** to test all 3 hypotheses in Method 1together, where pT = [µpo µAo µp1 µA1].

Question 2: Write the matrix C and the vectors p and h using numbers and the μ notation, as appropriate.

Question 3:Discuss the difference in the two methods proposed so far, including an advantage and a disadvantage of each method.

Method 3:
The 3rd researcher proposes to use the February weight to predict the August weight, taking sex into account. He proposes the model:

$$E(Y_{iAj} \mid Y_{iFj}) = \beta_0 + \beta_1 Y_{iFj} + \beta_2 Z_i$$

where Zi=0 for females and 1 for males.

Question 4: Interpret each of the three coefficients.

Question_5:Using notation introduced so far, write the vector Y,the matrix X and the vector p that would be used to write this-model in the form Y=Xp +E.

Questfon 6: Explain how the parameters in this model are related to the model used in Methods 1 and 2.

Question 7: Write the 3 sets of hypotheses tested in Method 1 using this new parameterization.

Method 4:
The 4th researcher proposes to do something similar to Method 3, but she extends the model to be:

$$E(Y_{iAj} \mid Y_{iFj}) = \beta_0 + \beta_1 Y_{iFj} + \beta_2 Z_i + \beta_3 Y_{iFj} Z_i$$

Question 8:Write the matrix X and the vector p that would be used to write this model in the form Y= XP + E.

Question 9: Interpret the coefficient 3·

Question 10: Draw graphs with February weight on the x-axis and August weight on the y-axis. Draw a separate graph for each of Methods 1, 3 and 4, and illustrate the relationship between the February and August weights proposed by the models in those methods. Label the lines well, including what the equations of the fines are in terms of the parameters defined for that model.

Question 11:Explain which method you would recommend, and why you would do so. Or if you prefer not to choose one, discuss advantages and disadvantages of any of the methods you do think are appropriate.

Problem 3 (202-203)

The Del Norte Salamander *(plethod on elongates)* is a small (5-7 cm) salamander (lizard-like amphibian) found among rock rubble and rock outcrops in a narrow range of northwest California. Researchers interested in studying the habitat of these salamanders randomly selected 47 locations from plausible salamander habitat in national forest and parkland. At each location, randomly chosen grid points were searched for the presence of a site with suitable rocky habitat. At each suitable site a 7 meter by 7 meter (49 m^2 area) was examined for the number of salamanders it contained (Salamanders). Researchers also measured the percentage of forest canopy cover (PctCover) and the age of the forest in hundreds of years (ForestAge). The first few lines of the data set are below:

```
> head (dat )
 _ Si:te  Salamanders  Pct-cm.rer -Forest-Age
1   1           13        85        3.16
2   2           11        86        0.88
3   3           11        90        5.48
4   4            9        88        0.64
5   5            8        89        0.43
6   6            7        83        3.68
```

Plots, descriptive statistics, and R output for the data are included as an Appendix.

(a) Describe the purpose of the link function in a generalized linear modeL Define the identity link and explain why it may not be appropriate for a Poisson distributed response variable.

(b) Consider using a Poisson regression model to predict salamander counts from percent canopy cover and forest age.

i. Do we need to include an offset term when fitting the model? If yes, state what variable we would use in the offset term. If no, explain why not.

ii. Describe one method you could use to check for overdispersion.

For parts (c)-(f), consider the Poisson regression model modi in the R output.

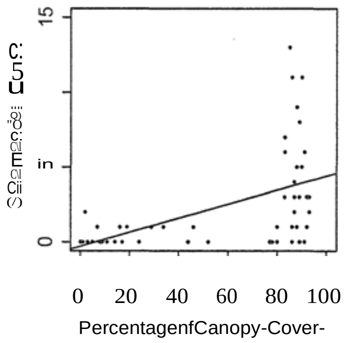
(c) Write out the equation of the Poisson regression model for modi (the assumed model, not the fitted model). Clearly define any variables and parameters used. State all model assumptfons.

(d) Write a sentence interpreting the estimated intercept, in the context of the problem.

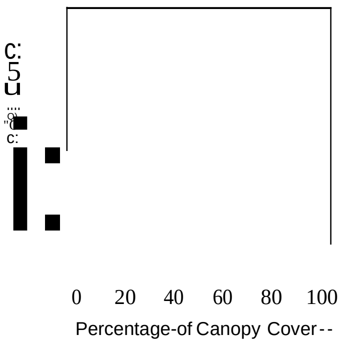
(e) Write a sentence interpreting the estimated coefficient for percent canopy cover, in the context of the problem.

(f) If you were to plot the predicted mean salamander count on the scatterplot of Salamander Count vs. 3 Canopy Cover, which of the following plots (A, B, or C) would you see?

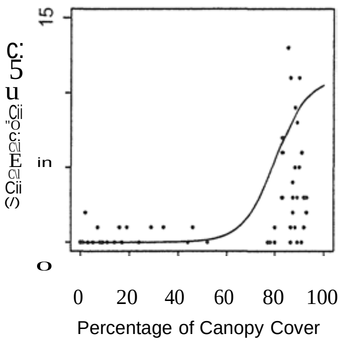
Plot A



Plot B



Plot C



Two distinct canopy conditions were evident: closed canopy (percentage canopy cover > 703) and open canopy (percentage canopy cover < 703). The salamander counts seem like they may follow different distributions for these two canopy conditions. Thus, researchers defined a new variable, CoverType, as "Closed" if PctCover is greater than 70, and as "Open" if PctCover is less than or equal to 70.

(g) Draw well-labeled side-by-side boxplots of salamander counts for the two canopy conditions. Note that summary statistics are given in the Appendix.

For parts (h)-(k), consider the quasi-Poisson regression model mod2 in the R output.

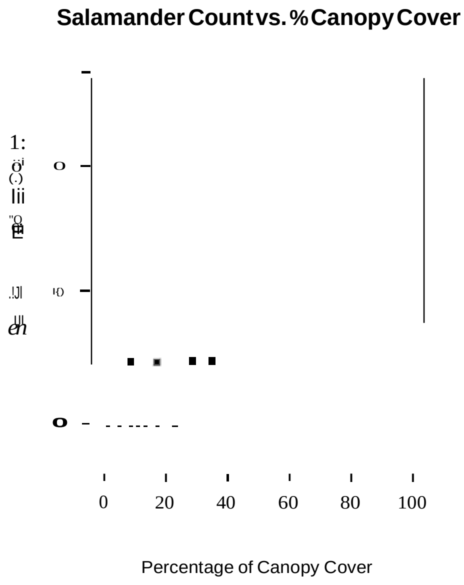
(h) Write the equation of the fitted model for mod2. Clearly define any variables and symbols used.

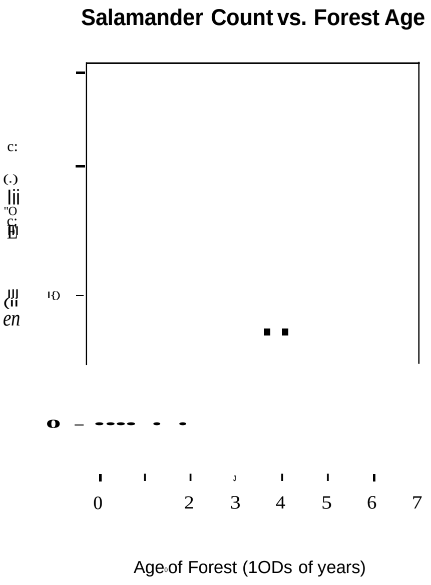
(i) Calculate an approximate 953 confidence interval for the ForestAge coefficient. Interpret this interval in the context of the problem.

(j) Calculate an approximate 953 confidence interval for the salamander rate ratio of Open canopy conditions to Closed canopy conditions for 100-year-old forests. Interpret this interval in the context of the problem.

(k) Is there significant statistical evidence that the effect of forest age on the mean salamander count changes with the canopy condition? Support your answer with an appropriate statistical test. Report the null and alternative hypotheses, value of the test statistic, p-value, decision, and conclusion.

Appendix for Problem 3 (202-203): Plots and R output


Salamander Count vs. % Canopy Cover


Salamander Count vs. Forest Age

```
> summary(dat[,2:4])
 Salamanders        PctCover       ForestAge
 Min.     0.000   Min.   : 0.00   Min.   :0.000
 1st Qu.: 0.000   1st Qu.:18.00   1st Qu.:0.295
 Median : 1.000   Median :83.00   Median :0.640
 Mean   2.468    Mean    :58.98   Mean   :1.688
 3rd Qu.: 3.000   3rd Qu.:88.00   3rd Qu.:2.665
 Max.   :13.000   Max.   :93.00   Max.   :6.750


 > tapply(dat$Salamanders, dat$CoverType, summary)
$Closed
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   3.000   3.857   6.000  13.000
$Open
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.4211  1.0000  2.0000

> mod1 = glm(Salamanders - PctCover, family="poisson", data=dat)
> summary(mod1)
Coefficients:
             Estimate Std. Error z value Pr(>lzl)
(Intercept) -1.481957   0.455780  -3.251  0.00115
PctCover     0.032409   0.005391   6.011 1.84e-09
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 190.22  on 46 degrees of freedom
Residual deviance: 121.31  on 45 degrees of freedom
AIC: 210.36
```

```
> mod2 = glm(Salamanders - CoverType*ForestAge, family="quasipoisson", data=dat)
> summary(mod2)
Coefficients:
                         Estimate Std. Error t value Pr(>itl)
(Intercept)              1.331200  0.269469      ?? 1.23e-05
CoverTypeOpen           -1.977252  0.819433      ??   0.0202
ForestAge                0.006971  0.081158      ??   0.9320
CoverTypeOpen:ForestAge -1.052731  2.792214      ??   0.7080

(Dispersion parameter for quasipoisson family taken to be 2.658398)

    Null deviance: 190.22 on 46 degrees of freedom
Residual deviance: 121.65 on 43 degrees of freedom
AIC: NA

> round(vcov(mod2 ,5)
                        (Intercept) CoverTypeOpen ForestAge CoverTypeOpen:ForestAge
(Intercept)                 0.07261      -0.07261  -0.01778                 0.01778
CoverTypeOpen              -0.07261       0.67147   0.01778                -1.45884
ForestAge                  -0.01778       0.01778   0.00659                -0.00659
CoverTypeOpen:ForestAge     0.01778      -1.45884  -0.00659                 7.79646
```

Problem 4 (202-203)

Mille was collected weekly from 79 Australian cows and analysed for its protein content. The cows were maintained on one of three diets: barley, a mixture of barley and lupins, or lupins alone. Repeated measurements were taken on each animal each week starting with calving (so week zero corresponds to the calving date for a particular cow) up to 19 weeks after calving. The objective of the study is to determine how diet affects the protein in milk over time. The variables in the data set are as follows:

Variable Name	Description
Cow	Cow ID number
Diet	Type of diet (1 = barley, 2 = mixture of barley and lupins, 3 = lupins)
Week	Week since calving
Week.s	$\text{Week.s} = \begin{cases} \text{Week} - 3 & \text{if Week} > 3 \\ 0 & \text{if Week} \leq 3 \end{cases}$
Protein	Percentage protein content in milk sample

Plots, descriptive statistics, and R output for the data are included as an Appendix.

(a) Refer to the linear mixed effects model modi R output for the following questions (i.-iv.).

i. Suppose a cow in the study is on the lupins diet (diet 3), and is measured at weeks 1, 2, and 4. Write the equation of the linear mixed effects model for this cow in atrix terms (the assumed model, not the fitted model). Clearly define any variables used, and write out the elements of each vector or matrix in the model. Identify which terms in the model are fixed, and which are random. State all model assumptions.

ii. Write a sentence interpreting the estimated standard deviation of the random intercepts in context of the problem.

iii. Draw a well-labeled plot with estimated (marginal) mean protein content on the y-axis and weeks since calving on the x-axis, and draw the three estimated (marginal) mean protein contents for the three diets. Include a legend.

iv. Suppose a cow in the study is on the lupins diet (diet 3), and is measured at weeks I, 2, and 4. What is the estimated 3 x 3 (marginal) covariance matrix for this cow?

(b) Refer to the linear mixed effects model mod2 R output for the following questions (i.-v.). Note that this model now includes a linear spline with a knot at week 3.

i. Write a sentence interpreting the estimated standard deviation of the random effect for Week in context of the problem.

ii. Calculate an approximate 953 confidence interval for the `Week` coefficient. Write a sentence interpreting this interval in context.

iii. What is the estimated equation of the (marginal) mean protein content for cows on the mixture diet (diet 2) after week 3? (Write your answer in the form $Y = a + bt$, where t is time, and *a* and *b are* numerical values.)

iv. ·Cow number 30 is on the mixture diet (diet 2), and its predicted random effects are:

```
> ranef(mod2)[30,]
   (Intercept)      Week      Week.s
30     0.02783    0.02399    -0.06342
```

What is the estimated equation of the conditional mean protein content for cow number 30 after week 3? (Write your answer in the form $Y = a + bt$, where *t* is time, and *a* and *b* are numerical values.)

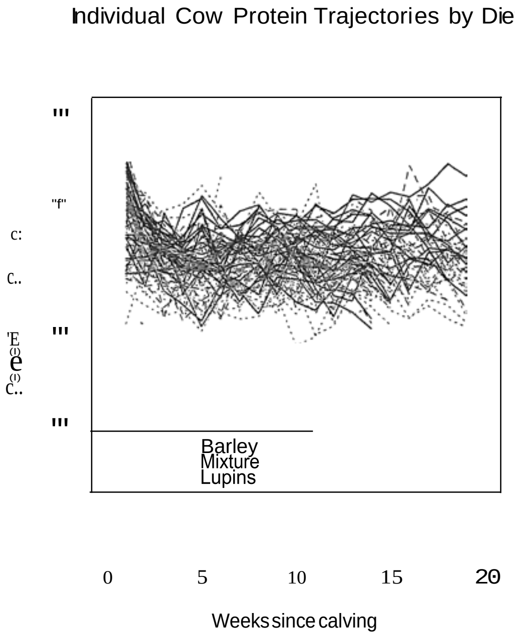
v. In the context of the problem, explain what it means to·includeca Week xDiet interaction in this model.

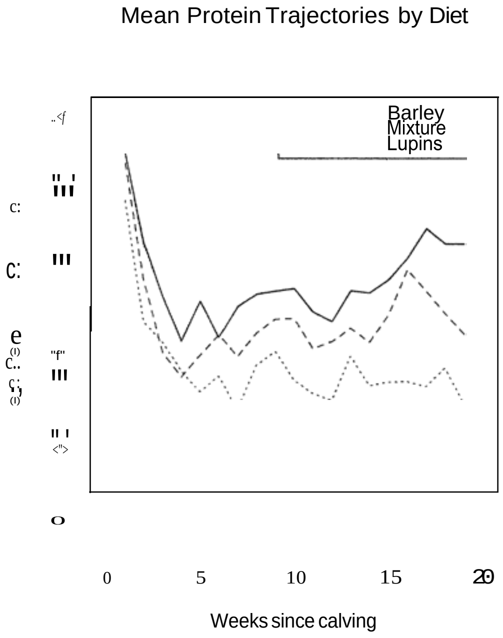
(c) From the R output provided, are you able to determine if we have significant statistical evidence that the type of diet has an effect on the mean protein content (not adjusting for week)? If yes, indicate which part of the R output allows you to make this determination, state the null and alternative hypotheses, the p-value, and your conclusion. If no, state the null and alternative hypotheses that we need to test. In either case, each null and alternative hypothesis should be a clearly defined model equation.

(d) Researchers believe that the type of diet has an effect on the changes in mean protein content over time. State an appropriate set of hypotheses that could be used to test this claim. Hypotheses should be stated in terms of clearly defined model equations.

(e) Time is measured in weeks since calving, and the study was terminated 19 weeks after the earliest calving. Thus, about half of the 79 sequences of milk protein measurements are incomplete. (A table showing the number of observations at each week is shown in the R output.) Calving date may well be associated, directly or indirectly, with the physiological processes that also determine protein content. If a cow iis missing measurements from weeks 15-19 after calving, should these measurements be considered missing completely at random (MCAR), missing at random (MAR), or not missing at random (NMAR)? Explain.

Appendix for Problem 4 (202-203): Plots and R output


Individual Cow Protein Trajectories by Diet


Mean Protein Trajectories by Diet

```
> ## Number of observations per week
> with(milk, table(Week))
Week
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
79 78 79 79 78 79 77 77 77 78 78 79 78 79 59 50 46 46 41

> modi = lme(Protein - Diet, random = - 11Cow, data=milk)
> summary(mod1)
Linear mixed-effects model fit by REML
 Data: millt
      AIC      BIC    logLik
  522.214  548.1937  -256.107

Random effects:
 Formula: -1 l Cow
        (Intercept) Residual
StdDev:    0.167339  0.2752585

Fixed effects: Protein - Diet
                 Value  Std.Error _ DF  t-value p-value
(Intercept)   3.526320 0.03607473 1258 97.75044  0.0000
Diet2        -0.095999 0.05006097   76 -1.91764  0.0589
Diet3        -0.203794 0.05009054   76 -4.06852  0.0001
```

```
> mod2 = lme(Protein - (Week + Week.s)*Diet,
     random = - 1+Week+Week.s1Cow, data=milk, method="ML")
> summary(mod2)
Linear mixed-effects model fit by maximum likelihood
 Data: milk
       AIC      BIC    logLik
  95.74637 178.9173 -31.87318

Random effects:
 Formula: -1 + Week + Week.s | Cow
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 0.4138388 (Intr) Week
Week        0.1454604 -0.848

-Weak.:s    0.1589985  0.784 -0.988
- Residual  0.2127301

Fixed effects: Protein - (Week + Week.s)* Diet
                 Value  Std.Error   DF  t-value p-value
(Intercept)   4.082085 0.10232599 1252 39.89294  0.0000

Week         -0.208098 0.03700470 1252 -5.62357  0.0000
Week.s        0.209664 0.03993271 1252  5.25044  0.0000
Diet2        -0.007420 0.14192764   76 -0.05228  0.9584
Diet3        -0.167465 0.14194571   76 -1.17978  0.2418
Week:Diet2   -0.032344 0.05134297 1252 -0.62996  0.5288
Week:Diet3    0.008152 0.05137106 1252  0.15869  0.8739
Week.s:Diet2  0.032903 0.05541605 1252  0.59375  0.5528
Week.s:Diet3 -0.017511 0.05545517 1252 -0.31576  0.7522
```

Methods E;x:am. Problem 3 (Statistics 211). To study the association between body mass index (BMI) and high blood pressure among adults between 30 – 40 years old, a cardiologist recruited n1= 100 males and *n2* = 100 females in her study. The following variables were recorded.

- The binary response variable *Yi* which takes the value 1 if both conditions are satisfied: systolic blood pressure exceeds 150 mm Hg and diastolic exceeds 90 mm Hg; and 0 otherwise.
- The variable BMI *xi* iri kg/m2.
- The gender indicators: Gli for the male group and *G2i* for the female group.

Part I. Consider this logistic regression model:

- *Yilxi* , Gli, *G2i* ,._, independent Bernoulli(p(xi, Gli, G2i))
- log $\frac{P\ xi,Gi;G2i}{l\text{-}p(xi,Gli,G2i\cdot}$ = *[f3&1* + *f3f* Xi] * G1i + [f3tf + *f3f* xi] * G2i

1. UsinK the logistic regr.essi0n model stated above derive an expression for the probability that a female subject with BMI x = 25 has high blood pressure.
2. Derive an expression for the difference in the log odds of having high blood pressure between males and females with low BMI $x = 20$. Compare this difference between male and females with high BMI $x = 30$.
3. For the remaining questions in Part I, use the following numerical results:

$$\left| \begin{matrix} \widehat{\beta}_0^{M} \\ \widehat{\beta}_1^{M} \\ \widehat{\beta}_0^{F} \\ \widehat{\beta}_1^{F} \end{matrix} \right| \quad \left(\begin{matrix} -2.00 \\ 0.10 \\ -2.10 \\ 0.08 \end{matrix} \right)' \quad \text{<C-((3) = ov –} \begin{pmatrix} 1.00 & 0.10 & 0 & 0. \\ 0.10 & 1 & 0 & 0 \\ 0 & 0 & 1.00 & 0.15 \\ 0 & 0 & 0.15 & 2.00 \end{pmatrix}.$$

   (a) Give an approximate 953 confidence interval for the probability that a male subject with BMI $x = 20$ has high blood pressure.

   (b) Give an approximate 953 confidence interval for the odds ratio of having high blood pressure (males vs females) among those with BMI $x = 20$.

   (c) Test the hypothesis that the log odds of having high SBP is higher for males than females among those with BMI $x = 30$. You may use approximate cbnfidence intervals to address this question.

Part II. In further refining the logistic regression model, the cardiologist takes the following into consideration.

- Based on physiology and empirical evidence, the cardiologist believes that the logit function (In [C::]) is constant, for both men and women, in the low BMI group: $x_i < a$ where a is the threshold parameter that also needs to be estimated. In other words, the log odds of having high blood pressure does not change for men and women with BMI less than a kg/m2. However, note that the log odds for men might actually differ from that of the women.

e The cardiologist has a-priori knowledge that the BMI threshold (change-point) a is in the interval $20 - 25$ kg/m^2.

• For men and women with BMI greater than a, the log odds of having high blood pressure increases as BMI increases. However, the logit function tapers off and thus it makes sense to express the relationship between the log odds and BMI in terms of the log BMI. Moreover the rate at which the log odds increases may differ between the men and women.

"" We need to constrain the logit function to be continuous at all points so one must be careful when specifying the logit function at the BMI threshold a.

Answer the following questions.

1. Write the logit function separately for the men and the women.
2. Write the total likelihood function. For simplicity, assume that the first 100 subjects were men (Gli $= 1$, G2i $= 0$ for i$= 1, \ldots 100$) and the last 100 subjects were women.
3. If the BMI threshold a was known, how would you estimate the logit when $BMI < a$. Note that you need separate estimates for the men and women.
4. Describe how to estimate the parameters in the logistic regression model. Here, it will be sufficient to obtain the estimator of the BMI threshold a in this discretized set of values $\{20, 21, \ldots, 25\}$. Give your algorithm.

Methods Exam. Problem 4 (Statistics 212). Let *Yij* be the response variable for unit iat time *tj.* Each subject iwas randomly assigned either treatment *A* or *B* at the beginning of the study. In this dataset, subjects i= 1, . . . , 25 are given treatment *A* and subjects i= 26, . . . , 40 are given treatment *B.* Each response variable is recorded at weeks *tj* = 1, . . . , 20.

Denote the treatment indicator *Ai* = 1if the i-th unit was given treatment *A* and 0 otherwise. The treatment *B* indicator is defined similarly.

Suppose that the appropriate model for this dataset is the linear splines model with a knot at *TI* = 5:

Yij = *,BoAAi* + *,BoBBi* + *boi* + *,81tj* + *(aAAi* + *aBBi)(tj* – T1)+ + *Eij*

where boi's are iid N(O, *a-)* and Ei/s are iid N(O, *a-).* Assume here that the b's and the c.'s are independent.

1. Suppose that the crucial time point is at 10 weeks. What is the expected response of any subject given treatment *A* (denotethis by *µA(lO)?* For treatment *B* (denoted by *µs(lO)?* What is the difference in the expected responses between treatments A_and *B* at 10 weeks (denoted .0i.(10) = J.LA(lO) – *µB (lO)?* Develop a procedure for testing the difference between the mean responses for treatments A vs. B on week 10.

2. Another important aspect of evaluating the treatment is the slope of the mean trend at around 9 weeks. What is the slope of each treatment and what is the difference in the slopes of each treatment at around 9 weeks. Develop a confidence interval estimator for this difference in the slopes at week 9.

3. What is the expected trend for a particular subject igiven treatment *A?*

4. Derive the conditional mean function for unit i, denoted *JE(Yij\boi).*

5. Derive the conditional (within-unit i) variance *<Cov(Yij , YiF \boi)·*

6. Derive the unconditional mean function JE(Yij).

7. Give the interpretation for Var boi = 0. How would you test the hypothesis Ho : Var boi = 0.

8. Let us stack up all the subject-specific response vectors <u>*Yi*</u> to form the vector <u>*Y*</u> = *[Y , . . . , Yjy.}'.* Consider now the matrix formulation

   <u>*Y*</u> = *X ,B* + *Zb* + f·

   For each of the vectors/matrices above specify the elements of each matrix or vector and also the dimension of each. Identify the random vectors or matrices and specify their distributions (including mean, variances, covariances, etc).

9. Denote <Cov b = *Eb* and <Covf = EE. Suppose for now that the variance components parameters are known. The goal here is to jointly estimate the fixed parameter vector *,B* and predict the random components *b.* Using the penalized weighted least squares criterion below (that treats *b* as a fixed quantity)

   $$S(\underline{\beta}, \boldsymbol{b}) = [\underline{Y} - (\boldsymbol{X}\underline{\beta} + \boldsymbol{Z}\boldsymbol{b})]'\Sigma_{\epsilon}^{-1}[\underline{Y} - (\boldsymbol{X}\underline{\beta} + \boldsymbol{Z}\boldsymbol{b})] + \boldsymbol{b}'\Sigma_{\boldsymbol{b}}^{-1}\boldsymbol{b},$$

   derive the normal equations for solving the estimate and predictor *b* that optimize the above criterion.

10. Derive Cov($[2\_^{1}$ *b′*]' | 2_, *b)* and give an approximate 95% confidence interval for JE(Yi|boi)·