Please read this page but

Do not open the exam until I tell you to start.

Please put your name on the back of your answer book.
Do NOT put it on the front. Thanks.

- The exam is closed book, closed notes. Use only the formula sheet and tables I provide today. You may use a calculator.
- Write your answers in your blue book. Ask if you need a second (or third) blue book.
- You have 2 hours (120 minutes) to complete the exam.
  Stop working when the end of the exam is announced.
- Points are indicated for each question. There are 160 total points.
- Important reminders:
  - budget your time. Some parts of each question should be easy; others may be hard. Make sure you do all parts you can.
  - notice that some parts do not require any computations.
  - show your work neatly so you can receive partial credit.
- Good luck!

1. 40 pts – Warranty claims on cars. A major car manufacturer is interested in modeling the number of warranty claims per month for two different car TYPEs, A and B. Their data include:
TIME: the number of months since the type was introduced
CLAIMS: the number of warranty claims for that type that month.

   For each of these two scenarios, described in words and pictures, construct an appropriate model. Briefly explain any additional X variables that you construct. Indicate what parameters (or combinations of parameters) estimate the quantity (or quantities) of interest.

   (a) 10 pts. There are no claims in month 0, but then claims increase linearly with time, perhaps at a different rate for the two car types. The quantity of interest is the difference in claims (between car types) in month 12.

   (b) 15 pts. There is an initial spike in claims in month 1. After that, claims increase linearly. The quantities of interest are:
   a) the difference in slopes from month 1 to 12,
   and b) the difference in claims in month 12.

   The number of warranty claims is associated with many other factors not included in the previous models. The company has over 1500 observations from many months, many car types, and many geographic regions. They want a model to predict monthly warranty claims. The data set includes a total of 28 possible variables, including all the variables you constructed in part 1c. These X variables are named X1, X2, · · ·, X28. The next page includes output from a stepwise regression ($\alpha_{include} = 0.15$ and $\alpha_{leave} = 0.15$) and selected parts of an all subsets regression. The all subsets regression includes the best 10 models using the C_p criterion and a few additional models.

   (d) 10 pts. Which variables would you include in a regression model to predict monthly warranty claims? Briefly explain your choice.

   (e) 5 pts. You present your model from part 1d to the company engineers and find yourself in a heated argument. The engineers very much prefer a model with X2, X7, X15 and X28 because it is based on subject-matter knowledge. You find no problems with autocorrelation, lack of fit, outliers, or influential points in either model. Other information about the "engineer's model" is in the output on the next page. Management wants you to recommend **one** model. Evaluate all the relevant information and make a recommendation. Explain your choice.

```
                      Summary of Stepwise Selection

       Variable     Number   Partial    Model
Step   Entered  Gone  Vars In  R-Square  R-Square    C(p)   F Value   Pr > F

  1    x28             1      0.6183    0.6183   1377.17  482.82   <.0001
  2    x2              2      0.2989    0.9173     68.70 1073.09   <.0001
  3    x15             3      0.0096    0.9268     28.73   38.73   <.0001
  4    x7              4      0.0073    0.9341      9.30   32.74   <.0001
  5    x9              5      0.0023    0.9364      1.33   10.58   0.0013



                           C(p) Selection Method

                      Number of Observations Read         300
                      Number of Observations Used         300

Number in
  Model  C(p)  R-Square    AIC       SBC  Variables in Model

   5    1.33    0.9364   -448.78   -426.56  x3 x7 x9 x15 x28
   6    1.41    0.9369   -449.82   -423.89  x3 x6 x7 x9 x15 x28
   5    1.47    0.9364   -448.40   -426.17  x2 x7 x9 x15 x28
   7    1.58    0.9372   -448.62   -418.99  x3 x6 x7 x9 x11 x15 x28
   6    1.58    0.9368   -448.44   -422.51  x2 x6 x7 x9 x15 x28
   6    1.62    0.9368   -448.39   -422.46  x3 x7 x9 x11 x15 x28
   7    1.68    0.9372   -448.19   -418.56  x2 x6 x7 x9 x11 x15 x28
   6    1.75    0.9367   -448.00   -422.07  x3 x7 x9 x13 x15 x28
   6    1.82    0.9367   -447.95   -422.03  x2 x7 x9 x11 x15 x28
   7    1.87    0.9371   -448.07   -418.44  x3 x6 x7 x8 x9 x15 x28
   3 1356.882   0.6314     74.48     89.29  x7 x15 x28
   3 2840.393   0.2998    266.99    281.81  x2 x7 x15

the engineer's model
   4    1.95    0.9341   -443.77   -425.26  x2 x7 x15 x28
```

2. 60 pts. Strength of mortar.

Mortar is used to 'glue' together bricks in a brick wall or surface of a building. It is made by mixing water and plaster, then letting the mixture cure. A strong mortar is important to hold together the bricks; strength depends on the composition of the plaster, the type of water, and the length of time the mortar is allowed to cure. The effects of type of water (WATER, soft or hard) and length of curing time (CURE, 3 days, 7 days, 28 days) were evaluated in an experiment. Nine BAGS of plaster were available. These are assumed to be randomly chosen from a population of all bags of plaster. A small amount from each of the 9 bags was randomly assigned to each of the six combinations of type of water and curing time. The tensile strength of each of the 54 samples was measured (larger values indicate stronger mortar). Recap: 9 bags, 2 types of water, 3 cure times, 54 samples.

The average strengths for each of the six treatments are:

	Curing Time		
Water type	3 days	7 days	28 days
hard	47	56	59
soft	36	44	58

(a) 5 pts. If the values in the above table were population means, is there an interaction between water type and curing time? Explain why or why not.

(b) 10 pts. Write out the skeleton ANOVA table with sources of variation and degrees of freedom. Indicate which effects are random and which effects are fixed.

(c) 5 pts. The Sums-of-squares for **some** of the lines in the ANOVA table are:

Source	SS
Water	179.62
Water * Cure	1924.06
Error	8812.81

Please test whether the type of water affects the strength, averaged over curing times. You only need to report the test statistic and its degrees of freedom. You do not need to find a p-value. If this test is not possible with the data provided, indicate what additional data is needed.

(d) 5 pts. Construct a 95% confidence interval for the difference between 7 days and 28 days of curing, averaged over water types. If this is not possible with the data provided, indicate what additional data is needed.

(e) 5 pts. Mortar is either made with hard water or soft water. Is your estimate from the previous part a reasonable description of the increase in strength between 7 and 28 days in soft water? Explain why or why not.

After you provide the investigators with the results of your analysis, they tell you about some additional data. Each of the 54 samples was measured twice, so they can now give you 108 observations. The questions on this page concern the analysis of these 108 observations.

(f) 5 pts. Write out the skeleton ANOVA table (sources of variation and d.f.) that is appropriate if each sample is measured twice. Indicate which effects should be considered random and which should be considered fixed.

(g) 5 pts. Sums-of-squares and some additional information for **some** of the rows in the ANOVA table are:

Source	SS	E MS
Water	915.1	$\sigma_o^2 + 2\sigma_s^2 + 36\Sigma\alpha_i^2$
Water*Cure	1907.8	$\sigma_o^2 + 2\sigma_s^2 + 18\Sigma\gamma_{ij}^2$
between samples	16810.0	$\sigma_o^2 + 2\sigma_s^2$
Error	1080.2	σ_o^2

Describe what σ_o^2 and σ_s^2 in the above table represent.

(h) 5 pts. Using the information in part 2g, test whether the type of water affects the strength, averaged over curing times. You only need to report the test statistic and its degrees of freedom.

(i) 5 pts. Estimate σ_s^2.

(j) 5 pts. Test H0: $\sigma_s^2 = 0$. You only need to report the test statistic and its degrees of freedom.

(k) 5 pts. If the investigators repeat this study, would recommend that they use one, two, or perhaps more than two, measurements per sample? Briefly explain your recommendation.

3. 15 pts. Tree graft success. A common practice in horticulture is to graft stems of one type of tree onto root systems of another. If the graft is healthy, the combination tree is much more vigorous than either part alone. Unfortunately, grafts often get diseased. The age of the root systems is suspected to influence the probability of success.

   The table below shows data from an experiment to evaluate graft success in root stocks of four different ages. The experiment started with 160 trees, 40 in each of 4 age groups. After two years, the status of each tree was recorded in one of two categories: healthy graft or not (diseased graft or dead). The contingency table is:

	Age of Tree			
Status	$1-2$	$2-5$	$5-7$	$7-10$
Healthy	1	17	15	3
Diseased/Dead	39	23	25	37

   (a) 5 pts. The Chi-square statistic for this table is 43.57. How many degrees of freedom does it have? What is the null hypothesis being tested?

   (b) 5 pts. Consider just the 1-2 year old and 2-5 year old root stocks.

	Age of Root Stock	
Status	$1-2$	$2-5$
Healthy	1	17
Diseased/Dead	14	16

   Compute the odds ratio that compares the odds of a healthy graft on a 2-5 year old root stock to that on a 1-2 year old root stock.
   Compute a 95% confidence interval for the odds ratio.

   (c) 5 pts. After completing your analysis, the experiments tell you that the experiment was actually conducted at 4 sites, with 10 trees of each age at each site. Some of the sites are much more stressful than others, so P[Dead] varies between sites. Is the Chi-square test in question 3a still appropriate? Explain why or why not.

4. 45 pts. Studies and models

   (a) 15 pts. You have developed a technique to measure an important quantity in your research area. You design a study to estimate the precision of your technique. You are concerned about two different components of precision:
   the variability between measurements made by the same person, and
   the variability of measurements made by different people.

   You recruit 10 PEOPLE to measure a single object; each person makes five separate MEASUREMENTS. Recap: 1 object, 10 people, 50 total measurements.

   Write out the skeleton ANOVA table (sources of variation and d.f.) and indicate whether each term should be considered fixed or random.

   (b) 15 pts. You make some refinements to your technique and decide to repeat your assessment of variability. Your major professor did not like the previous assessment (part 4a) because it only applied to one object. This time you use a different design.

   You have 6 OBJECTs to be measured. You expect each object to have a different mean; you are not interested in estimating their variance. Each object is measured by the **same** 10 PEOPLE. Each person makes two MEASUREMENTS of each object. Recap: 6 objects, 10 people, 120 total measurements.

   Write out the skeleton ANOVA table (sources of variation and d.f.) and indicate whether each term should be considered fixed or random.

   (c) 15 pts. You are studying the loss of topsoil (erosion) in Iowa farm fields. You want to estimate the differences between three farming PRACTICEs (no till, reduced till, and conventional) and two DRAINAGE (tile or no tile). For each of the 6 combinations of practice and drainage, you locate 3 FIELDs, i.e. a total of 18 fields. These fields are scattered randomly across a six county area of Iowa. Within each combination of practice and drainage, fields are expected to differ from each other because of differences in farming history, soil type, and weather. Within each field, you measure soil erosion at three topographic POSITIONs: hill top, mid-slope, and bottom. You have one MEASUREMENT at each position per field. Recap: 3 practices, 2 drainages, and 3 positions, 18 fields and 54 measurements.

   Write out the skeleton ANOVA table (sources of variation and d.f.) and indicate whether each term should be considered fixed or random.

   Also, indicate the appropriate error term for an F test of each main effect.

That's all! Have a great break.