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

DO NOT START until I tell you to. You are welcome to read this front page.

- 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 120 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. 45 points. Are baseball free agents 'worth it'? After a professional baseball's player contract expires, that player becomes a 'free agent' and is permitted to sell his services to the highest bidder. The salaries paid to free agents are often much higher than those paid to 'regular' players. Do players perform better when paid higher salaries? The batting average is one measure of performance of a baseball player. Higher batting averages indicate better performance.

   Data were collected on 132 players who became free agents between 1976 through 1989. For each player, two batting averages were computed. PRE is the batting average the year before the player became a free agent. POST is the batting average the next year, after the player became a free agent. We want to make inferences about the mean difference in batting average.

   (a) These data are an example of: (write the numbers of all appropriate terms on your answer sheet)
   1) Block design
   2) Paired data
   3) Randomized experiment
   4) Observational data
   5) Latin Square design

   (b) You want to test the null hypothesis that the difference (POST - PRE) = 0. Here are the means and standard deviations for four quantities that might be useful. Please calculate the appropriate t-statistic. **Note:** you do not need to provide a p-value.

Quantity	Average	s.d.
Pre Free Agent	0.258	0.0337
Post Free Agent	0.249	0.0418
Difference	-0.00929	0.0459
Player average	0.253	0.0302

   (c) There are 132 players and 264 observations in the data set. What are the d.f. associated the T statistic in part 1b?

   (d) Fill in the missing d.f. and SS. in the ANOVA table. The variable PLAYER is the player name; the variable PERIOD has two values: PRE or POST.
   Note: you do not need to compute MS or F values.

Source	d.f.	SS
Player		0.23902
Period		0.00569
Error		
Total	263	0.38316

   (e) The s.d. for the pre values is smaller than that for the post values. Does this invalidate your test in part 1b or part 1d? Explain why or why not.

   (f) One assumption of this analysis is additivity. Briefly explain, using this study as an example, the meaning of additivity.

   (g) This study will be repeated, but the choice of design is not clear. Is it better to use the same design or something different? The correlation between two observations on the same player is estimated to be 0.26. What design would you recommend? Justify your choice.

(h) Below is the residual vs. predicted value plot from the ANOVA in part 1d. Does this plot indicate any problems or concerns about the analysis?

2. 20 pts. Pharmaceuticals usually have a shelf life. One reason is that the concentration of active ingredient decreases over time. One common model for this is the exponential decay model

$$Y_i = C_0 e^{-\lambda X_i} \times \epsilon_i,$$

where Y_i is the concentration at time i, C_0 is the initial concentration (at time of manufacture), X_i is the time since manufacture, and λ is the decay rate. This model can be fit using linear regression by log transforming both sides to get

$$Y_i^* = \log Y_i = \beta_0 + \lambda X_i + \varepsilon_i.$$

Data were collected from a randomized experiment. Twenty packages of a particular drug were randomly assigned to be sampled at one of five times, 0 months, 2 months, 4 months, 8 months, and 12 months after manufacture. 4 packages were assigned to each sampling time. Each package was measured once and only once, at the assigned sampling time. The data are shown below.

(a) Four different models were fit to the data. Some unimportant details of each model are left out. The Error sums of squares for each are:

Model	# parameters	Error SS
$Y_i^* = \beta_0$	1	10.020
$Y_i^* = \beta_0 + \lambda X_i$	2	0.5051
$Y_i^* = \beta_0 + \lambda X_i + \beta_2 X_i^2$	3	0.5050
$Y_i^* = \mu + \tau_i$	5	0.2550

Please test whether $\lambda = 0$. Report the test statistic, the appropriate d.f. and an approximate p-value. If this is not possible from the information provided, indicate what additional information you need.

(b) Construct the most appropriate test of lack of fit of the linear regression, using one or more of the sums of squares in part 2a. Report the test statistic, its d.f., an approximate p-value and a short conclusion.

(c) The drug company has a large batch of the drug (many packages) that has been in storage for four months. They will use the regression to predict the mean concentration of drug in this batch. They also want to report the precision of that estimate. There are three possible formulae that could be used:

1) $\sqrt{MSE\,(\frac{1}{n} + \frac{(4-\overline{x})^2}{\sum(x_i-\overline{x})^2})}$
2) $\sqrt{MSE\,(1 + \frac{1}{n} + \frac{(4-\overline{x})^2}{\sum(x_i-\overline{x})^2})}$
3) $\sqrt{MSE/4}$

Which is the most appropriate formula to calculate standard error? Explain (briefly) why your choice is the most appropriate.

The next problem starts on the next page.

3. 55 pts. Mercury is a naturally occuring contaminant in fish in freshwater lakes. In some lakes, the mercury concentration is sufficiently high that warnings are posted to limit the consumption of fish from that lake. The mercury concentration is influenced by many characteristics of lakes. The state of Maine collected information from a random sample of 110 lakes. The variables of interest included:

   hg: mercury concentration in fish.
   elev: elevation of the lake
   area: area of lake
   drain: size of the drainage basin for the lake
   runoff: annual runoff into lake

   Here are the parameter estimates and standard errors when a multiple linear regression is fit to the data, and an ANOVA table:

parameter	estimate	s.e.	VIF
β_0	0.758	0.144	0
β_{elev}	-0.211	0.061	1.07
β_{area}	-0.014	0.016	1.82
β_{drain}	0.0137	0.029	1.85
β_{runoff}	-0.292	0.262	1.02

Source	d.f.	SS
Model		1.137
Error		7.515
Total		8.652

   (a) Fill in the degrees of freedom in the above ANOVA table.

   (b) An undergraduate student calculated the slope for runoff in a simple linear regression as -0.527. When they see your results (β_{runoff} = -0.292) they are concerned that they made a calculation error. Should the two estimates be the same? If not, explain why they are likely to be different.

   (c) Construct a test of H0: $\beta_{elev} = 0$, $\beta_{area} = 0$, $\beta_{drain} = 0$, $\beta_{runoff} = 0$, if that is possible from the data provided. If not, indicate how you would construct the desired test.

   (d) Construct a test of H0: $\beta_{drain} = 0$, $\beta_{runoff} = 0$, if that is possible from the data provided. If not, indicate how you would construct the desired test.

   (e) Construct a test of H0: $\beta_{area} + \beta_{drain} = 0$, if that is possible from the data provided. If not, indicate how you would construct the desired test.

   (f) Construct a test of H0: $\beta_{drain} = 0$, if that is possible from the data provided. If not, indicate how you would construct the desired test.

   (g) The next two pages provide a variety of diagnostic plots. In order, they are a plot of h_{ii} for each observation, $DFFITS_i$ for each observation, Cook's distance for each observation, externally studentized residual for each observation, and a plot of externally studentized residuals against predicted values. Also, VIF values for each parameter are in the table at the top of the page. The investigators want to use this model to predict mercury concentrations in unsampled lakes (there are a lot of lakes in Maine!).

   Do you have any concerns about the fitted model? If so, describe your concerns.

   (h) Would it be appropriate to use a Durbin-Watson test here? Explain why or why not.

   (i) Would it be appropriate to use a Breusch-Pagan test here? Explain why or why not.