NAME:

STAT 350:

Final Exam,

Instructions: This is an open book test. You should have a total of **11** pages of questions including an extra page in case you run out of space. Separately you will be given two handouts containing data and statistics. Handout 1 refers to Part A; Handout 2 refers to Part B. You may use notes, text, other books and a calculator. Your presentations of statistical analysis will be marked for clarity of explanation. I expect you to explain what assumptions you are making and to comment if those assumptions seem unreasonable. The exam is out of **50**. You should feel free to write on the back of pages; I have brought extra pieces of blank paper which can be stapled to your exam if you need more space. Please do not hand in the Handouts; I won't be marking them. Your exam has your name preprinted on it. Please make sure you have the right exam.

PART A

The first 4 questions refer to a single data set which is in Handout 1. Here is a description of the data set:

The percentage of a person's body made of fat is a useful indicator of risk of heart disease. It is difficult to measure this quantity directly. The data set accompanying this exam is a set of cases from an experiment in which the density of subjects is measured by a water displacement method. This density is connected to the variable PERCENT in the table of data by an old formula called Siri's equation the details of which are unimportant for this exam.

For each subject in the data set we have measurements of DENSITY, PERCENT, AGE, HEIGHT, WEIGHT and the circumferences of various features of the body. DENSITY is measured in grams per cubic centimetre, AGE in years, WEIGHT in pounds, HEIGHT in inches and all circumferences in centimetres. PERCENT is computed by weight.

The original data set had 252 cases but for question 1 I have selected a subset of 50 cases on which to run the regression and produce diagnostics. Questions 2, 3 and 4 use all 252 cases.

NAME:

1. Handout 1 presents a table of diagnostics. The table shows DFFITS, Cook's distance, case deleted (externally standardized) residuals and leverages for a model in which PERCENT is the response and HEIGHT, WEIGHT, and all the circumferences are used as predictors. Examine the table and identify any cases deserving further scrutiny. Look at the data for the cases you identify and comment on any unusual values for these individuals which might account for the diagnostic values. [5 marks]

NAME:

2. In this question we consider prediction of HEIGHT using AGE (X_1), ABDOMEN(X_2), CHEST (X_3), FOREARM (X_4) and NECK (X_5). A table in Handout 1 presents Error Sums of Squares for all 32 possible submodels of the full model. Carry out backwards selection to pick a final model. Use a 5% level to remove variables. [10 marks]

NAME:

3. The appendix presents output for a model predicting HEIGHT from CHEST, ABDOMEN and HIP circumferences. From the information provided give a 99% confidence interval for the average height of people whose Chest, Abdomen and Hip circumferences are 92, 61 and 92 centimetres respectively. [6 marks]

NAME:

4. The information provided in the appendix can be used to predict HIP from Abdominal circumference. Give the formula for the regression of HIP on ABDOMEN. [5 marks]

NAME:

PART B

A statistics professor is teaching 2 sections of STAT 350, a hypothetical linear models course; one section meets in the morning and one in the evening. He has an idea which he hopes might improve marks on the final exam. He hopes that if he has each student come to his office after the first midterm to spend 15 minutes going over the student's answers to the midterm with him the students will be better able to use the midterm to improve their performances on the final.

He plans to carry out an experiment to see if the idea works. He considers two designs:
Design A: He picks one section at random by tossing a coin. For each student in the selected section he tries his idea. In the other section he just posts solutions as usual. (All students can see the solutions.)
Design B: For each student in both sections he tosses a coin: Heads means that he tries the idea on that student and Tails means that he doesn't. (So if there are 50 students in the two sections he will have to toss the coin 50 times.) Solutions are again posted.

In either case the prof records for student i: the final exam score (Y_i), the midterm score (x_i), the section the student was in and u_i which is 1 if the idea was tried on student i and 0 if not.

Data are shown for 50 hypothetical students in Handout 2. Also in Handout 2 are various regression results and summary statistics. (The last two variables are called Section and Treatment in the handout.)

5. Why do I think Design B is better? [2 marks]

NAME: .

6. Carry out a 2 sample t test of the null hypothesis that my idea does nothing against the alternative that it works. [4 marks]

7. Use a model in which the treatment (my idea) and midterm result act additively to test the same null hypothesis as in the previous part. Why is this a better test? [5 marks]

NAME:

PART C

Consider data Y_i, $i = 1, \ldots, n$ and the heteroscedastic linear model

$$Y_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \epsilon_i$$

where ϵ_i has a $N(0, \sigma^2/w_i)$ distribution and the ϵ_i are independent. The w_i are known constants. In class I showed that the weighted least squares estimate (which I will call $\tilde{\beta}$) of β is

$$\tilde{\beta} = \left(X^T W X\right)^{-1} X^T W Y$$

where W is the diagonal matrix

$$W = \begin{bmatrix} W_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{bmatrix}.$$

This model may be written in matrix form as

$$Y = X\beta + \epsilon$$

where $\epsilon \sim MVN(0, \Sigma)$ for some matrix Σ.

8. Give the simplest formula possible for Σ. [1 mark]

9. What is Var(Y)? [1 mark]

NAME:

10. Suppose you use the ordinary least squares estimate

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

for data following the model above. Show that $\hat{\beta}$ is unbiassed. [2 marks]

11. What is $\mathrm{Var}(\hat{\beta})$? [2 marks]

NAME:

Now consider the following example with $n = 2$:

	$i = 1$	$i = 2$
Y	Y_1	Y_2
x	1	2
w	1	4

12. Show that

$$\tilde{\beta} = \frac{Y_1 + 8Y_2}{17}$$

[2 marks]

13. Find $\mathrm{Var}(\tilde{\beta})$ in terms of σ. [2 marks]

14. Find $\mathrm{Var}(\hat{\beta})$ in terms of σ. [2 marks]

15. Which has the smaller standard error, $\tilde{\beta}$ or $\hat{\beta}$? [1 mark]

NAME:

Extra Page