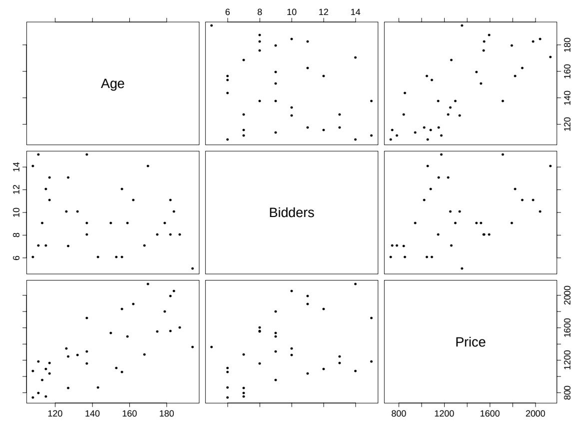
Statistics 328 Examination 2 Name ______________________

Fall 1996 **You must show all of your work**

When asked to explain something, or to provide an interpretation for a quantity, provide an explanation that could be understood by someone who does not have formal training in statistical methods.

1. Data were collected on the auction sale prices of 32 antique clocks. For each observation in the data set, the price in dollars (Y), the age of the clock in years (x_1), and the number of bidders (x_2) was recorded. The following figure is a "pairs plot" showing all pair-wise scatter plots.



Three models were fit to the data

$$\begin{aligned}\text{Model 1:}\quad y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \\ \text{Model 2:}\quad y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \\ \text{Model 3:}\quad y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \epsilon\end{aligned}$$

where $\epsilon \sim \text{nid}(0, \sigma^2)$. The following tables gives a summary of the results:

	Model 1	Model 2	Model 3
β_0	-1336.7221	322.7544	-340.033
β_1	12.7362	0.8733	3.4144
β_2	85.8151	-93.4099	13.6289
β_3		1.2979	1.1234
β_4			-0.0037
β_5			1.1234
$SSE = \sum_{i=1}^{32}(y_i - \widehat{y}_i)^2$	514035	218646	190610
$SS_{yy} = \sum_{i=1}^{32}(y_i - \bar{y})^2$	4791195	4791195	4791195

(a) List some of the things that you see in the pairs plot.

(b) Briefly explain why $\sum_1^{32}(y_i - \bar{y})^2$ is the same for all three models.

(c) What is an estimate for the change in expected price of an auctioned clock for an additional year of age for Model 1? For Model 2? Compute a numerical value if possible. If you cannot compute a numerical value, give a formula and explain how you would do the computation if you had all of the needed information.

(d) One of the assumptions used in the fitted models for the clock data is that the residuals have a standard deviation σ that does not depend on the levels of the explanatory variables (i.e., on age or the number of bidders). The residual (or error) degrees of freedom can be viewed as the effective sample size for estimating σ. Compute the residual degrees of freedom for Model 1, Model 2, and for Model 3.

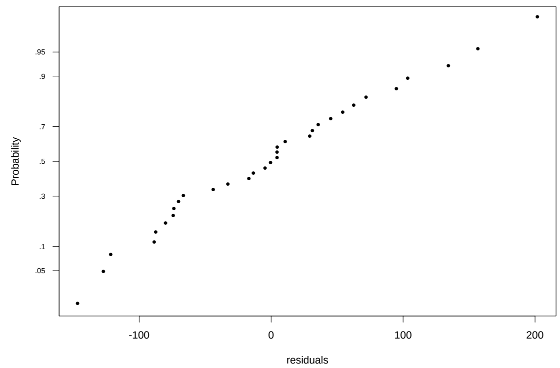
(e) Compute an estimate of σ from Model 2. Briefly explain the *practical* interpretation of this quantity.

(f) Model 2 has an interaction term and Model 1 does not. Briefly explain, in terms of this particular example, the practical difference between Model 1 and Model 2.

(g) Use an F-test to compare Model 1 and Model 2. Use $\alpha = .05$. What do you conclude?

(h) Model 3 includes quadratic terms. Do the data provide sufficient evidence to include these terms in the model? Perform the appropriate test, using $\alpha = .05$.

(i) Normal probability plots (and the closely related normal Q-Q plots) are useful tools for data analysis. The following figure is a normal probability plot for Model 2 of the Clock data.



Briefly explain the general purpose of a normal probability plot how you would interpret this particular plot.

2. Briefly explain the danger involved in using stepwise regression? What is the useful purpose of stepwise regression?

3. Briefly explain the difference between quantitative and qualitative variables and give two examples of each.

4. Briefly explain why it is *not* a good idea to use R^2 as a criterion for deciding how many explanatory variables to use in a regression model.

5. In a simple model to describe the change in housing prices from one year to the next (Y), a regression model was used in which the percent increase is a function of the amount of floor space in the house (x_1) and the size of the lot (x_2). Write down the "full quadratic model" for this example. Briefly explain, in simple terms, the shape of the response surface corresponding to this model.

6. A test was conducted to compare the life times of two different light bulbs designs—the current design, and a new design. Ten bulbs for each brand were tested until failure with the number of hours until failure being recorded for each bulb. A dummy variable regression model was used to analyze the data. The model fitted is

$$\text{Hours} = \beta_0 + \beta_1 X_1 + \epsilon$$

with $X1 = 0$ for the old design and $X1 = 1$ for the new design. Management asked for the life test so that they could decide whether to switch production to the new design or not. The following is some of the Splus output for these data.

```
> attach(bulb.frame)
> bulb.fit1 <- lm(Hours ~ X1)
> summary(bulb.fit1)
Call: lm(formula = Hours ~ X1)
Residuals:
    Min     1Q Median    3Q   Max
 -271.3 -112.5  -14.6 119.5 218.1
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept) 951.9000  45.8659    20.7540   0.0000
         X1 146.4000  64.8642     2.2570   0.0367
Residual standard error: 145 on 18 degrees of freedom
Multiple R-Squared: 0.2206
F-statistic: 5.094 on 1 and 18 degrees of freedom, the p-value is 0.03667
> anova(bulb.fit1)
Analysis of Variance Table
Response: Hours
Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value      Pr(F)
X1         1  107164.8 107164.8 5.094151 0.03667392
Residuals 18  378663.0  21036.8
> other.X1 <- list(X1=c(0,1))
> predict(bulb.fit1,newdata=other.X1,se.fit=T)
$fit:
     1      2
 951.9 1098.3
$se.fit:
        1        2
 45.86593 45.86593
$residual.scale:
[1] 145.0408
$df:
[1] 18
```

(a) Compute a confidence interval for the mean life for the old design. What is the interpretation of this interval?

(b) Compute a confidence interval for the difference in mean life between the new and the old designs.

(c) Compute a prediction interval for the life of a light bulb from the new design.

(d) Is there evidence that the mean life of the new design is longer than that of the old design?