

Examination #1, Stat 8061, Fall, 2000

This is a closed book exam. You may use one page of notes and a calculator. Put your answers in the blue book provided. There are 10 parts, worth 10 points each.

1. Figure 1 gives a scatterplot of $y = \text{age}$ versus $x = \text{shell length}$ for a sample of 500 abalone (which are edible marine snails).

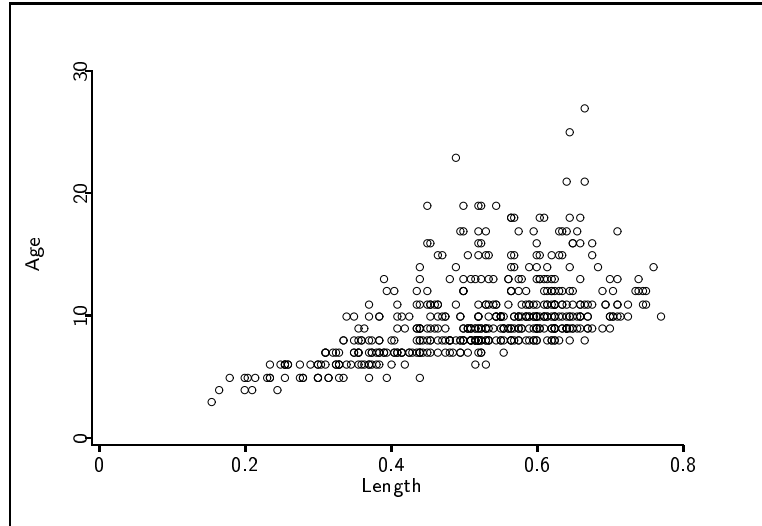


Figure 1: Figure for Problem 1.

- (a) Give up to three reasons for examining a scatterplot like Figure 1. Then, summarize the information in Figure 1 in up to three sentences.
- (b) Explain how a smoother like lowess or the slice smoother estimates $E(y|x)$. Drawing a graph might help you with this problem.
- (c) Abalone can be classified using a grouping variable G as either $G = 1$ for male, $G = 2$ for female or $G = 0$ if sexually immature (even older abalone may be sexually immature). Suppose it were hypothesized that, for $g = 0, 1, 2$,

$$\begin{aligned} E(y|x, G = g) &= \eta_{0g} + \eta_{1g}x \\ \text{Var}(y|x, G = g) &= \sigma^2 \end{aligned}$$

Could Figure 1 be consistent with this hypothesis or does it contradict the hypothesis? Why? (HINT: If you have trouble with this one, assume G has only 2 levels, not 3, and then solve the problem.)

2. For the abalone data, using the variables as defined in Problem 1, the following regression output was obtained for a particular fitted model (values have been rounded to make calculations easier):

```

Data set = Abalone, Name of Fit = L3
Normal Regression
Kernel mean function = Identity
Response      = Age
Terms         = ({F}G Length)
Coefficient Estimates
Label      Estimate      Std. Error      t-value      p-value
Constant   3.84              0.56            6.876        0.0000
{F}G[1]    1.24              0.37            3.345        0.0009
{F}G[2]    0.82              0.34            2.435        0.0152
Length     12.33             1.21            10.160       0.0000

R Squared:
Sigma hat:
Number of cases:          500
Degrees of freedom:      496

Summary Analysis of Variance Table
Source      df      SS      MS      F      p-value
Regression   3     1675     558     71.15   0.0000
Residual    496    3892     7.85
Lack of fit  194    1329     6.85     0.81   0.9473
Pure Error  302    2563     8.49

```

- (a) The estimated coefficient for $\{F\}G[1]$ is 1.24. Explain what this number means.
3. R^2 , the proportion of variability explained, is often used as a summary of a regression problem. Give a formula for computing R^2 , and give the most general conditions you can for which R^2 is indeed a useful summary of a regression problem.

4. A linear regression with two predictors x_1 and x_2 has been computed as follows (results have been rounded to simplify computations):

```
Data set = Prob4, Name of Model = L1
Normal Regression Model
Mean function = Identity
Response      = Y
Predictors    = (X1 X2)
Coefficient Estimates
Label      Estimate      Std. Error      t-value
Constant   0.20                0.10            1.99
X1         -0.24                0.12            -1.93
X2          0.90                0.047           19.23

R Squared:          0.93
Sigma hat:          0.094
Number of cases:    53
Degrees of freedom  50
```

```
Summary Analysis of Variance Table
Source      df      SS      MS      F      p-value
Regression  2      5.52    2.76    311.21  0.0000
Residual    50     0.44    0.0089
```

- (a) Suppose you were to draw added-variable plots for both x_1 and x_2 . Which is more likely to show a strong linear trend? Why? Are there circumstances that would suggest that the other added-variable plot would show a stronger linear trend? What are these circumstances?
- (b) Define $z_1 = x_1 + x_2$ and $z_2 = x_1$. Suppose we then compute ols regression using the mean function

$$E(y|z_1, z_2) = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2$$

assuming constant variance σ^2 . Give values for the following statistics, based on the output for this problem. **If you don't have enough information to give the value, tell what else you need.**

- i. $\hat{\gamma}_0$
- ii. $\hat{\gamma}_1$
- iii. $\hat{\gamma}_2$
- iv. $\text{se}(\hat{\gamma}_2|z_1, z_2)$.
- v. Test of $\gamma_1 = \gamma_2 = 0$ against a general alternative.

5. Figure 2 is a scatterplot matrix for the response (*Age*) and two predictors *Height* of the shell and the *Shell* weight. Suppose we fit the linear regression of *Age* on *Height* and *Shell*.

- (a) Explain why the inverse response-plot of the fitted values from this regression versus *Age* is not likely to help select a transformation of *Age*.
- (b) Figure 3 is the output from the "Choose response transform" item for the regression suggested in this problem. Explain what this plot shows.

6. The plot shown in Figure 4 is labeled as a residual plot, where x is one of the terms in the fitted linear model. Is this possible? Why or why not?

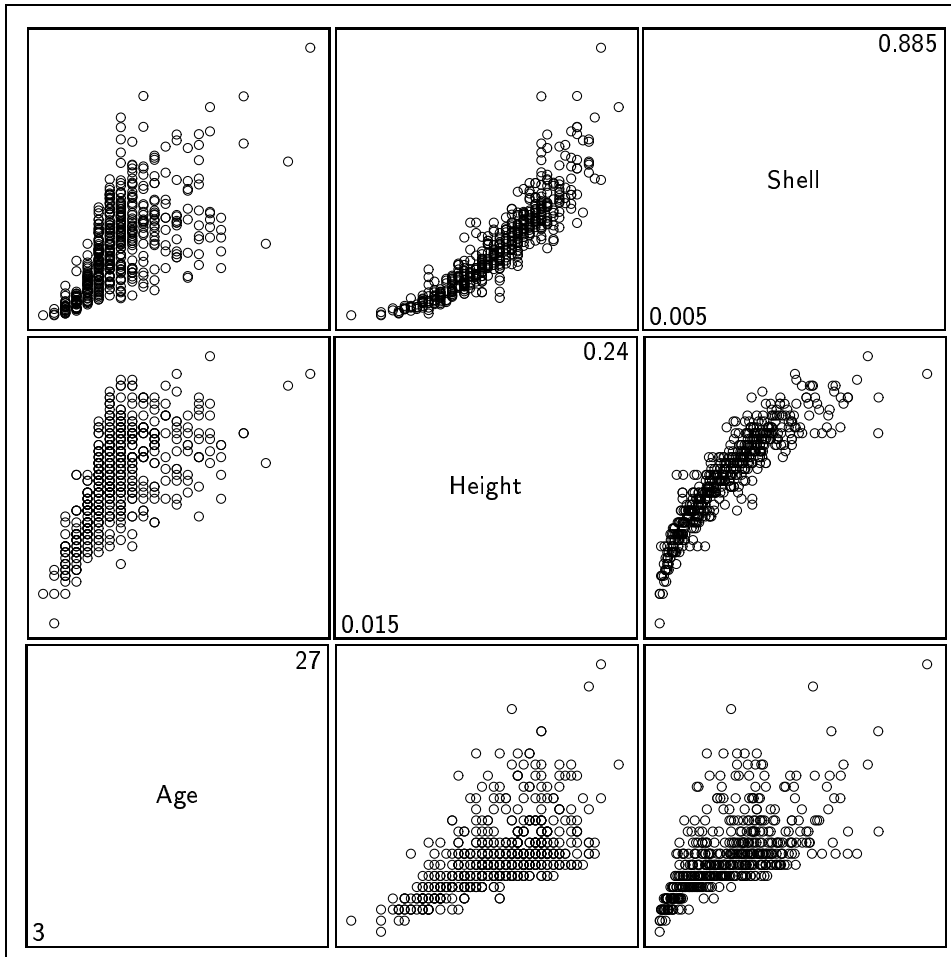


Figure 2: Scatterplot matrix for abalone data and Problem 5.

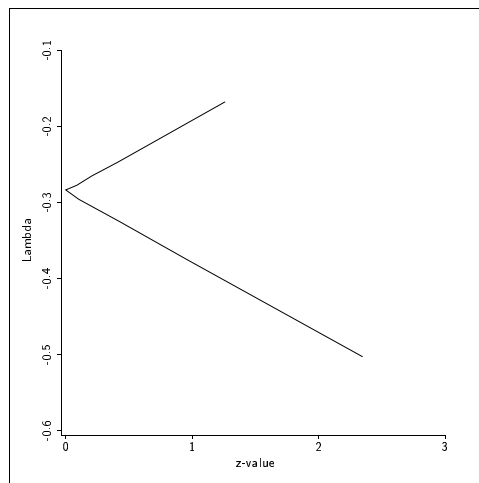


Figure 3: Graphical output from “Choose response transform” item in the regression menu for Problem 5b.

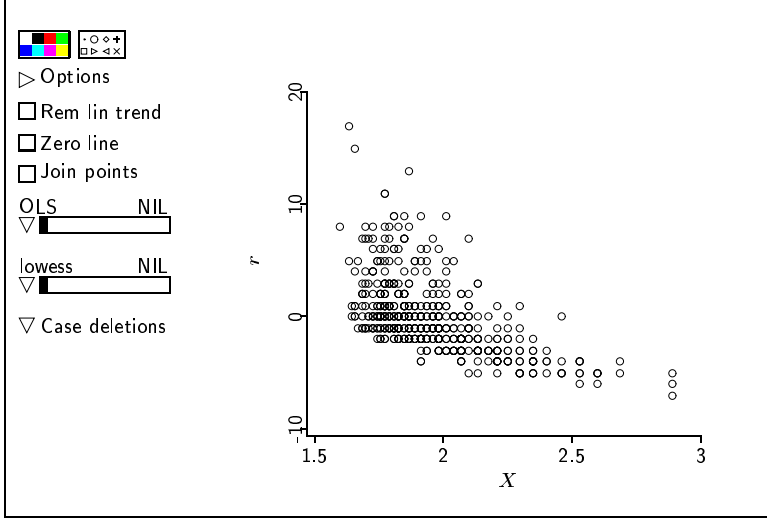


Figure 4: Graph for Problem 6.