PBHS 32410 / STAT 22401 Problem Set 2 Winter 2025

Possible points 92.

1. [24 pts total] Adapted from Chaterjee & Hadi text

The data in file assessments.txt on the website shows the scores for a definitive assessment F and the scores in the two preliminary assessments P_1 and P_2 for 22 individuals.

(a) Display the relationships among the three variables and comment [4 pts]

(b) [4 pts] Fit each of the following models to the data:

$$\begin{aligned}
\text{Model 1}: &\quad F = \beta_0 + \beta_1 P_1 \qquad\qquad + \epsilon; \\
\text{Model 2}: &\quad F = \beta_0 \qquad\qquad + \beta_2 P_2 + \epsilon; \\
\text{Model 3}: &\quad F = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \epsilon.
\end{aligned}$$

(c) [4 pts] Describe the results from each of the univariate (SLR) models.

(d) [4 pts] Which variable individually, P_1 or P_2, is a better predictor of F? In other words, which variable explains more variation in F?

(e) [4 pts] What is the interpretation of the coefficient for β_2 in the multiple regression model?

(f) [4 pts] For an individual with scores 78 and 85 on the first and second preliminary assessments, what is the predicted final assessment score?

2. [16 pts] Regression Sudoku

The table below (adapted from C&H) shows the regression output, with some numbers erased, when a simple regression model relating a response variable Y to a predictor variable X_1 is fitted based on 20 observations. Complete the missing numbers and identify Var(Y). **Hint:** starting with observed t-statistic, several other quantities in SLR can be derived to get started.

ANOVA Table

Source	Sum of Squares	df	Mean Square	F-Test
Model	1848.76	$*$	$*$	$*$
Residuals	$*$	$*$	$*$	
Total	$*$	$*$	$*$	

Coefficients Table

3 Variable	Coefficient	std.error	t-test	p-value
Constant	-23.4325	12.74	$*$	0.0824
X_1	$*$	0.1528	8.32	<0.0001
$n = *$	$R^2 = *$	$R^2_{adjusted} = *$	$\hat{\sigma} = *$ (Root MSE)	

Var(Y) =

3. [20 pts total]

The SPRM text provides data (page 84) on body mass index (BMI) and physiologic measures to which it may be associated. BMI is a measure of body fat (an inexpensive but significantly flawed proxy measure for actual fat composition) and as such, might be predicted by other factors associated with nutrition, metabolic function, etc, in addition to age. The data in in BMIdat.txt. Complete the following (note: there are no 'great' models here, so don't be surprised by that).

(a) Evaluate the correlation of BMI predictors with each other and with BMI.

(b) Use simple linear regression to evaluate each candidate predictor alone. Briefly summarize the findings.

(c) Use multiple linear regression to evaluate all predictors. Briefly summarize the findings.

(d) Reduce the model to a suitable 2-variable (or one-variable) model, with a goal of the best R^2 and significant or near significant ($p < .15$) predictors.

(e) Obtain the predicted values ($\hat{y}$) and plot these against the BMI values. Does the plot look fairly random?

4. [32 pts, 4 pts each] Adapted from Exercise 3.15 of Chaterjee & Hadi text

Cigarette Consumption Data: A national insurance organization wanted to study the consumption pattern of cigarettes in all 50 states and the District of Columbia. The variables chosen for the study are described below and given below and the data in cigarettesales.txt on the website.

variable	**definition**
State	U.S. State/Terr
Age	median resident age
HS	percentage high school grads
Income	income measure
AA	percentage African-American
Female	percentage female
Price	price measure
Sales	packs per capita sold - the outcome variable

In (a)-(c) below, specify the null and alternative hypotheses, the test used, and your conclusion using a 5% level of significance.

(a) Test the hypothesis that, among the six (numeric vars, leaving out State) candidate predictors of Sales available, there are at least some useful predictors.

(b) Test the hypothesis that the variable Female is not needed in the regression equation relating outcome Sales to the six predictor variables.

(c) Test the hypothesis that the variables Female and HS (both together) are not needed in the above regression equation Conclude what to do with these predictors.

(d) Identify the 95% confidence interval for the regression coefficient of the variable Income based on the above model and interpret it.

(e) What percentage of the variation in Sales can be accounted for after Income is removed from the regression equation with Female and HS already removed?

(f) What percentage of the variation in Sales can be accounted for by the three predictor variables: Price, Age and Income?

(g) What percentage of the variation in Sales can be accounted for by the variable Income alone?

(h) From the model in (a), omit predictors as warranted to arrive at a final model. Describe the effects of the remaining predictors (direction and magnitude).