Practice Exam 1, ST518/590

Directions: Answer questions as directed. Please show work. At least half credit will be given when *correct expressions* are given, but numerical answers are either incomplete or incorrect. For true/false questions, *circle* either true or false.

1. Summary statistics from the heights, x, and weights, y, of $n = 101$ football players are given below.

$$\begin{array}{rcl}
\overline{x} = 73.5 \text{ inches} & & \overline{y} = 226.0 \text{ lbs} \\
s_x^2 = 5.8 \text{ inches}^2 & \Longrightarrow & \sum(x_i - \overline{x})^2 = ? \\
s_y^2 = 1639.1 \text{ lbs}^2 & \Longrightarrow & SS[Tot] = \sum(y_i - \overline{y})^2 = ? \\
s_{xy} = 62.3 \text{ lb-inches} & &
\end{array}$$

(a) Report r_{xy}, the sample correlation coefficient for weight and height among these $n = 101$ players.

(b) Suppose that these 101 players can be regarded as a random sample from a bivariate population of interest. Briefly specify a simple regression model for $Y_1, \ldots, Y_n$ in which the mean weight of players in the population is linear in height, x_i.

(c) Fill in the spaces after the question marks above with the total sum of squares of the weights, $SS[Tot]$ and the sum of squares for heights, S_{xx}. Use the summary statistics to obtain the least squares (LS) estimate of the slope. Specify the units.

(d) True or false: in SLR, the least squares (LS) estimate of slope satisfies

$$\hat{\beta}_1 = r_{xy}\frac{s_x}{s_y}.$$

(e) True or false: $SS[Reg] = \hat{\beta}^2 \sum(x_i - \overline{x})^2$

(f) True or false: The *fitted* or *predicted* value of the i^{th} observation may be written as follows

$$y_i = \overline{y} + r_{xy}s_y\left(\frac{x_i - \overline{x}}{s_x}\right)$$

(g) True or false: $SS[E] = \sum(\hat{y}_i - \overline{y})^2$.

(h) True or false: under repeated sampling in the SLR model, the estimated slope and intercept will be positively correlated in this context.

(i) Complete the ANOVA table below:

Source	*df*	Sum of squares	Mean square
Regression	1		
Error			
Total			

(j) The observed F-ratio is greater than the $99.99th$ percentile of the F distribution with 1 and 99 numerator and denominator degrees of freedom. With regard to the association of y and x, ...

i. ...this provides strong evidence of a linear association. (True or false)

ii. ...this provides evidence of a strong linear association. (True or false)

(k) Evaluating the least squares regression line at $x = 73.5$ yields

$$\hat{\mu}(73.5) = \hat{\beta}_0 + \hat{\beta}_1(73.5) = 225.5$$

i. Report an estimated standard error for the estimated mean weight among the population of players of height = 73.5 inches.

ii. Estimate the standard deviation among this population of players with $x = 73.5$ inches.

2. The data below taken from Dickey's website correspond to temperatures taken in January, 1992.

	Mon H	Mon L	Tues H	Tues L	Wed H	Wed L
seattle	51	44	52	44	59	47
boston	29	12	32	29	44	28
richmond	47	23	55	40	51	28
louisville	53	30	37	29	46	24
lubbock	46	40	48	42	54	42
omaha	31	26	36	22	47	31
sanfran	56	47	65	49	65	47
philly	36	18	46	27	46	26
cincinnati	50	25	36	29	41	30
phoenix	74	49	75	48	75	48
miami	72	68	77	71	79	72
milwauke	31	23	33	26	35	26
dallas	50	47	53	47	56	44
burlingvt	20	-2	28	03	39	24
buffalo	34	18	34	28	32	26
charlotte	49	38	60	41	59	41
bismark	27	-5	43	15	47	17
elpaso	61	34	64	33	64	32
rapidcity	46	20	62	25	57	41

Consider predicting the high temperature thi for a given day using a linear function of as many as 4 independent variables: the low and high temp from two days ago $(lo2, hi2)$ and yesterday's low and high temp (ylo, yhi). The correlations of thi with the four predictors, along with $p-$ values for tests of 0 correlation are given below:

```
Pearson Correlation Coefficients, N = 19
        Prob > |r| under H0: Rho=0

                        thi

        hi2         0.84769
                     <.0001

        lo2         0.77497
                     <.0001

        yhi         0.94572
                     <.0001

        ylo         0.77438
                     <.0001

        tlo         0.83471
                     <.0001
```

Several multiple regression models were considered. Sums of squares appear below. Parameter estimates are given for the full model.

- Model 1 $\mu(ylo, yhi, lo2, hi2) = \beta_0 + \beta_1 lo2 + \beta_2 hi2 + \beta_3 ylo + \beta_4 yhi$
- Model 2 $\mu(ylo, yhi, lo2, hi2) = \beta_0 + \beta_3 ylo + \beta_4 yhi$
- Model 3 $\mu(ylo, yhi, lo2, hi2) = \beta_0 + \beta_2 hi2 + \beta_4 yhi.$
- Model 4 $\mu(ylo, yhi, lo2, hi2) = \beta_0 + \beta_4 yhi$

Model: MODEL1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2624.48822	656.12206	36.43	<.0001
Error	14	252.14336	18.01024		
Corrected Total	18	2876.63158			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	15.67260	4.25081	3.69	0.0024
yhi	1	0.71275	0.13444	5.30	0.0001
ylo	1	-0.12905	0.20062	-0.64	0.5305
hi2	1	-0.00512	0.17243	-0.03	0.9768
lo2	1	0.21464	0.17800	1.21	0.2479

Model: MODEL2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2590.67200	1295.33600	72.48	<.0001
Error	16	285.95958	17.87247		
Corrected Total	18	2876.63158			

Model: MODEL3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2593.51335	1296.75667	73.28	<.0001
Error	16	283.11823	17.69489		
Corrected Total	18	2876.63158			

Model: MODEL4

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2572.83737	2572.83737	143.97	<.0001
Error	17	303.79421	17.87025		
Corrected Total	18	2876.63158			

(a) True or false: the regression coefficient β_4 has the same interpretation in all four models.

(b) Note that $F(0.05, 2, 16) = 3.63$ and $F(0.05, 1, 18) = 4.41$ and $F(.05, 2, 14) = 3.74$.

- By comparison of a reduced model with Model 1, use an F-ratio to test the hypothesis that today's expected high temperature does not depend on temperatures from two days ago, given the other predictors.

- By comparison of a reduced model with Model 1, use an F-ratio to test the hypothesis that today's expected high temperature does not depend on either of last two low temperatures, given the other predictors.

- Test $H_0 : \beta_{2 \cdot 4} = 0$ using level $\alpha = 0.05$. Briefly describe what's being tested here. Report and interpret the appropriate partial coefficient of determination.

(c) Provide a matrix formulation of Model 1. Make sure to identify X, Y, β and an error vector in the model, specifying the dimension of each. Provide matrix expressions for $\hat{\beta}$ and its estimated variance-covariance matrix $\hat{\Sigma}$.

(d) Consider Model 1. Provide a matrix expression for an estimate of the mean mean high temperature among days where the preceding measurements were $(lo2, hi2, ylo, yhi) = (50, 30, 60, 50)$. Provide an expression for the standard error as well.

(e) True or false: There is no evidence of any linear association between today's high temperature and the high temperature from two days ago.