

ST 512-Practice Exam I - Osborne

Directions: Answer questions as directed. For true/false questions, circle either true or false.

1. A study was carried out to examine the relationship between the number of moths caught overnight in a light trap y and two weather variables, $x_1 =$ maximum temperature on preceding day and $x_2 =$ avg. wind speed. A total of $n = 63$ trivariate observations were made. Three regression functions $\mu(x_1, x_2) = E(Y|x_1, x_2)$ are being considered

$$\text{Model 1: } \mu(x_1, x_2) = \beta_0 + \beta_1 x_1$$

$$\text{Model 2: } \mu(x_1, x_2) = \beta_0 + \beta_2 x_2$$

$$\text{Model 3: } \mu(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Some sums of squares computed from these data and these models appear below:

$$SS[E]_3 = 2.365$$

$$SS[R]_1 = 0.305$$

$$R(\beta_2|\beta_0, \beta_1) = 0.488$$

$$R(\beta_1|\beta_0, \beta_2) = 0.595$$

Some critical values from F distributions which might be useful here are

$$F(0.95, 1, 60) = 4.001, \quad F(0.95, 2, 60) = 3.150, \quad F(0.95, 1, 62) = 3.996$$

An incomplete ANOVA table which might be useful appears below:

Source	SS	df	MS
Reg			
Error	2.365		
Total			

- (a) Compute and interpret $MS[E]_3$.
- (b) Compute the sum of squares for regression for the full model, $SS[R]_3$
- (c) Test $H_0 : \beta_1 = \beta_2 = 0$ using level $\alpha = 0.05$:
 - i. Compute the appropriate F -ratio:
 - ii. Compare it to the appropriate critical value given above and draw a conclusion.
- (d) Is there anything to be gained by adding β_2 into the model with β_1 already in it? Specify the nested and full models from the candidates above which would be appropriate to address this question. Compute the appropriate F -ratio for model selection and compare it to the appropriate critical value (using $\alpha = 0.05$) and draw a conclusion.
- (e) Compute the sample coefficient of determination for model 3.
- (f) Compute the partial coefficient of determination, $r_{yx_2 \cdot x_1}^2$.
- (g) Compute and interpret the partial correlation coefficient $r_{yx_2 \cdot x_1}$.

2. Data from a study to determine biological maturity and effect of harvesting quality of an Indian grain are given in the 2nd and 3rd columns in the table below. The response variable y is yield (kg/ha) of paddy, a grain farmed in India and x is the number of days after flowering at which harvesting took place.

Obs	x	y	yhat	residual
1	16	2508	3099	-591
2	18	2518	3124	-606
3	20	3304	?	?
4	22	3423	3173	250
5	24	3057	3197	-140
6	26	3190	3222	-32
7	28	3500	3246	254
8	30	3883	3271	612
9	32	3823	3295	528
10	34	3646	3320	326
11	36	3708	3344	364
12	38	3333	?	-36
13	40	3517	3393	124
14	42	3241	3418	?
15	44	3103	3443	-340
16	46	2776	3467	-691

Consider a linear model for the data, $Y_i = \beta_0 + \beta_1 x_i + E_i$ where E_i are i.i.d. $N(0, \sigma^2)$ random variables and $\beta_0, \beta_1, \sigma^2$ are unknown parameters.

Part of the ANOVA table for the simple linear regression of yield y on days after flowering for harvest x is given below:

Source	DF	Sum of Squares	Mean Square	$E(MS)$	F Value	p -value
Model	1	204526	204526			0.2951
Error						
Corrected Total		2625168				

- (a) Complete the ANOVA table by filling in the following:
- SSE
 - MSE and its expected value under the linear model
 - The F -ratio for testing $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$
- (b) Estimate the standard deviation of yields if number of days after flowering is held fixed.

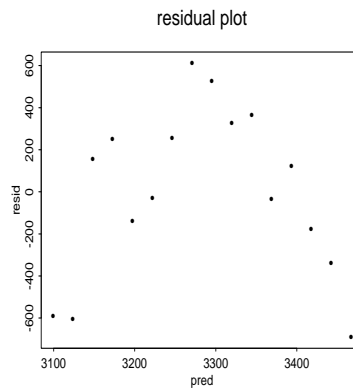
(c) The least squares estimates of β_0 and β_1 are given in the SAS output below:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	2902.96471	364.66796	7.96	<.0001
x	1	12.26324	11.27540	1.09	0.2951

Use this output to

- i. Obtain a 95% confidence interval for β_1
 - ii. Fill in the predicted values and residuals that were left out in the table with the data on page 1.
- (d) What is your conclusion about the linear association between the mean yield and time until harvest? Report a p -value for $H_0 : \beta_1 = 0$
- (e) True or false: the p -value represents the probability that $\beta_1 \neq 0$.
- (f) Calculate r^2 . What is this called and what does it mean in this instance?
- (g) True or false: the result of the test from part d) establishes the independence of yield y and time until harvest x .
- (h) Inspect a plot of the residuals versus the predicted values for the model below.
- i. Circle the points corresponding to observations #3, 12, 14 for which you completed the table on the preceding page.
 - ii. Use this plot to guide you in specifying another regression function below (hint: see next question)

$$E(Y|x_i) = \beta_0 + \beta_1 x_i$$



(i) If you chose a quadratic model (hint!), you get the following ANOVA table:

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	2084779	1042390	25.08
Error	13	540388	41568	
Corrected Total	15	2625168		

- (j) What is the coefficient of determination for this model?
- (k) What is your estimate of the standard deviation of yields when days after flowering is held fixed now?

- (l) Use the estimates given below for a quadratic model to predict the yield when harvesting 30 days after harvesting. (Complete the table below).

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-1070.39769	617.25268	-1.73	0.1065
x	1	293.48295	42.17764	6.96	<.0001
x2	1	-4.53580	0.67442	-6.73	<.0001

Obs	x	y	yhat	residual
1	16	2508	2464	44
2	18	2518	2743	-225
3	20	3304	2985	319
4	22	3423	3191	232
5	24	3057	3361	-304
6	26	3190	3494	-304
7	28	3500	3591	-91
8	30	3883	.	.
9	32	3823	3676	147
10	34	3646	3665	-19
11	36	3708	3617	91
12	38	3333	3532	-199
13	40	3517	3412	105
14	42	3241	3255	-14
15	44	3103	3062	41
16	46	2776	2832	-56

- (m) Using the fact that

$$MSE * (1, 30, 900)(X'X)^{-1} \begin{pmatrix} 1 \\ 30 \\ 900 \end{pmatrix} = 5840,$$

obtain a 95% confidence interval for mean yield when harvest occurs after 30 days.

3. A researcher was interested in how several anatomical factors influence the specific gravity of wood from a particular pine species. Cross-sections were prepared from 20 samples of mature wood from this species. For each sample, the specific gravity (Y) was determined and the following anatomical measurements were also recorded:

- x_1 : number of fibers/mm² of springwood
- x_2 : percent springwood
- x_3 : percent light absorption of summerwood

PROC GLM in SAS was used to fit the following MLR model

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + E_i \quad \text{for } i = 1, \dots, 20$$

Note some potentially useful critical values from the F distribution:

$$F(0.95, 1, 16) = 4.49, \quad F(0.95, 2, 16) = 3.63, \quad F(0.95, 1, 17) = 3.59.$$

Answer the following questions, using the SAS output from this procedure where needed:

- (a) Give the degrees of freedom for the model, error and (corrected) total sums of squares, (SS[R], SS[E] and SS[Tot]) respectively.)
- (b) Compute the MS[E].
- (c) Report the SS which measures variation in specific gravity y explained by percent springwood x_1 alone (i.e. ignoring x_2 and x_3 .)
- (d) Report the SS which measures variation in specific gravity y explained by percent springwood x_1 after accounting for variation in specific gravity explained by x_2 and x_3 .
- (e) Test the reduced model involving only x_2 and x_3 versus the full model at level $\alpha = 0.05$. Report the F ratio, the critical value and your conclusion.
- (f) Test the reduced model involving only x_1 with the full model at level $\alpha = 0.05$. Report the F ratio, the critical value and your conclusion.
- (g) Report the coefficient of determination for the model you think fits best.

Obs	x1	x2	x3	y
1	49.6	735	92	0.690
2	46.9	768	92	0.724
3	50.2	555	87	0.679
4	46.0	651	87	0.689
5	48.5	626	85	0.691
6	57.1	698	88	0.687
7	45.3	521	94	0.751
8	56.1	594	87	0.642
9	49.6	574	94	0.698
10	48.3	689	94	0.726
11	38.3	498	88	0.762
12	47.7	630	93	0.708
13	42.7	696	91	0.740
14	50.7	508	87	0.680
15	53.8	536	90	0.720
16	58.0	574	86	0.635
17	45.9	613	92	0.700
18	54.3	658	91	0.690
19	47.3	520	92	0.738
20	45.0	703	87	0.726

Number of observations 20

The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model		0.01493538			
Error		0.00566182			
Corrected Total		0.02059720			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x1	1	0.01284319	0.01284319	36.29	<.0001
x2	1	0.00000805	0.00000805	0.02	0.8820
x3	1	0.00208414	0.00208414	5.89	0.0274

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x1	1	0.00956371			
x2	1	0.00009650			
x3	1	0.00208414			

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	0.6211773276	0.15196796		
x1	-.0046907088	0.00090228		
x2	-.0000281843	0.00005397		
x3	0.0036747008	0.00151418		