Stat 328, Summer 2005

Exam #2, 6/16/05

Name (print) Key

UnivID 123-45-6789

I have neither given nor received any unauthorized aid in completing this exam.

Signed Key

Answer each question completely—showing your work where appropriate (for possible partial credit).
Round your answers to **4 decimal places**.

Questions 1 to 7 Many mutual funds compare their performance with that of a benchmark, an index of the returns of all securities of the kind the fund buys. The Vanguard International Growth (VIG) Fund, for example, takes as its benchmark the Morgan Stanley Europe, Australia, Far East (EAFE) index of overseas stock market performance. The data for this analysis are the percent returns for the fund and for the EAFE from 1982 to 2000 (19 years total).

1. Which best describes the relationship between the VIG Fund and the EAFE Index?

D

(A) These two variables have no clear relationship.
(B) In years where the EAFE performs better, the VIG shows poor performance.
(C) The correlation between VIG and EAFE is 0.806906.
(D) These variables have a positive, approximately linear relationship.
(E) None of the above

2. Report the $P-$value for testing $\beta_1 = 0$ vs. $\beta_1 \neq 0$ and the result of this test.

B

(A) $< .0001$, fail to reject $\beta_1 = 0$.
(B) $< .0001$, reject $\beta_1 = 0$.
(C) 0.2038, fail to reject $\beta_1 = 0$.
(D) 0.2038, reject $\beta_1 = 0$.
(E) None of the above

3. If the EAFE Index is zero next year, what would this model predict for the value of the VIG Fund?

B

(A) 16.34526
(B) 3.505144
(C) 0.8278888
(D) 9.461562
(E) None of the above

4. Is the prediction of VIG when EAFE is zero an extrapolation?

B

(A) Yes
(B) No
(C) This can't be determined from the information given.

5. A simple rule some people use to identify outliers in a regression analysis is the "3 times RMSE" rule. The rule is applied as follows: a point whose residual is greater than $3 \times$ RMSE is considered to be an outlier. Apply this rule to the regression model for VIG vs. EAFE.

B

(A) There is at least one outlier.
(B) There are no outliers.
(C) There are too many outliers to count.
(D) This can't be determined from the information given.
(E) None of the above

6. Refer to the regression of VIG vs EAFE (Output Pages 1–2). Calculate an 80% confidence interval for the intercept. Calculate a 99.9% confidence interval for the slope. Be sure to show your work for both. Also, clearly indicate any JMP output values and/or table values you use in your work.

Intercept, β_0

$b_0 \pm t \cdot SE_{b_0}$

$3.5051 \pm (1.333)(2.6519)$
(JMP) (Table D) (JMP)

Slope, β_1

$b_1 \pm t \cdot SE_{b_1}$

$0.8279 \pm (3.965)(0.0982)$
(JMP) (Table D) (JMP)

Table D

$df = n-2 = 19-2 = 17$

80% $\leftrightarrow t = 1.333$ 99.9% $\leftrightarrow t = 3.965$

7. Refer to the regression of VIG vs EAFE (Output Pages 1–2) as well as the quadratic model (see Output Page 3). Comparing 2 or 3 values from the linear model's output and the quadratic model's output, explain why the linear model is preferable over the quadratic model. Be sure to indicate the 2 or 3 values you used from each model as well as how each of these values indicate that the linear model is preferable to the quadratic model.

Compare RMSE: Model #1 Model #2

$s = 9.4616$ $s = 9.7455$

The smaller RMSE of Model #1 indicates this model fits the data better (ie is closer to the data points).

[redacted]

Consider the test for EAFE^2 term in Model #2

This is like comparing Full model (quadratic) vs. Reduced model (linear) with a Partial F-test. The large P-value indicates that EAFE^2 can be removed without significantly diminishing the predictive power of the model.

Questions 8 to 14 The data for these questions are for insured commercial banks by state and other U.S. properties. The variables are as follows:

ASSETS average bank assets for the state or property (billions of dollars)

NUMBER number of commercial banks in the state or property

DEPOSITS average amount on deposit with banks in the state or property (billions of dollars)

We are interested in describing how assets are explained by deposits and the number of banks in a state or property.

8. Model#1 includes the variable Number*Deposits. This is an example of what type of explanatory variable?

D

(A) response
(B) indicator
(C) additive
(D) interaction
(E) None of the above

9. In Model#1, does the $P-$value for Number*Deposits indicate that this variable can be dropped from the model if we desire a simpler model?

B

(A) Yes
(B) No
(C) This can't be determined from the information given.

10. In Model#1, which single $x-$variable appears to be the most significant given the other variables in the model?

C

(A) Intercept
(B) Number
(C) Deposits
(D) Number*Deposits
(E) This can't be determined from the information given.

11. In Model#2, what does the $P-$value for the ANOVA $F-$test indicate about this model?

B

(A) Both of the explanatory variables are useful for predicting Assets.
(B) At least one of the explanatory variables is useful for predicting Assets.
(C) None of the explanatory variables are useful for predicting Assets.
(D) Model#1 is better than Model#2.
(E) None of the above

12. OUTPUT PAGE #7 contains many columns of output related to Model#2. Using this output, identify (by Obs#) all observations that have unusually large studentized residuals and/or unusually large "hat" values (h Assets column) and/or unusually large Cook's D values. Indicate how you decided if an observation had an "unusually large" value for these three quantities.

unusually large

1. studentized resids — not in $(-3, 3)$ range
2. "hats" — larger than $2 \cdot \frac{K+1}{n} = 2 \cdot \frac{2+1}{54} = 0.1111$
3. Cook's D's — "large" compared to other Cook's D values

large student. resids	large "hats"	large Cook's D
#5	#5	#33 (D = 19.7936)
#33	#14	(?) #5 (D = 1.6771)
	#33	
	#34	
	#44	

13. Refer to the insured commercial banks data and models (Output Page 4–6). List the 5 models in order from "best Assets predictor" to "worst Assets predictor" and record the value from each model's output you used to rank the models in this way. Then, beside each model's name, record the value of RSquare Adj from each model's output. What do you notice about the RSquare Adj values as you move down your list? Which model has the best (i.e. largest) RSquare Adj value? Which model has the worst (i.e. smallest) RSquare Adj value?

		RMSE	R^2 Adjusted
Best →	Model #5	20.201	0.986461
	Model #1	20.294	0.986336
	Model #2	21.939	0.984031
	Model #3	26.425	0.976834
Worst →	Model #4	170.135	0.039688

R^2 Adj decreases

Model #5 is best

Model #1 is worst (by R^2 Adj)

14. Using the output from Model#5, calculate a 95% confidence interval for σ. (You will need the χ^2-table from the website.)

$df = n-(k+1) = n-(2+1) = 54-3 = 51$

2.5%-tile $= L = 33.162$

97.5%-tile $= U = 72.616$

$$20.2013\sqrt{\frac{51}{72.616}} \quad \text{to} \quad 20.2013\sqrt{\frac{51}{33.162}}$$

Questions 15 to 16 A process to fill plastic 500ml bottles with water is expected to produce bottles with a mean water content of 505ml and a standard deviation of 12ml. The company plans to use an $\overline{x}-$chart and an $s-$chart to monitor this process. The company plans to measure the water content in a sample of 9 bottles. A sample of 9 bottles will be taken every 2 hours.

15. Calculate the centerline, lower control limit, and upper control limit for the $\overline{x}-$chart.

$$UCL = \mu + 3\frac{\sigma}{\sqrt{n}} = 505 + 3\frac{12}{\sqrt{9}} = 517$$

$$\text{Centerline} = \mu = 505$$

$$LCL = \mu - 3\frac{\sigma}{\sqrt{n}} = 505 - 3\frac{12}{\sqrt{9}} = 493$$

16. Calculate the centerline, lower control limit, and upper control limit for the $s-$chart.

$$UCL = B_6 \cdot \sigma = (1.707)(12) = 20.4840$$

$$\text{Centerline} = C_4 \cdot \sigma = (0.9693)(12) = 11.6316$$

$$LCL = B_5 \cdot \sigma = (0.232)(12) = 2.7840$$