Exercises on Chapter 3: Inferences in Regression Analysis:

Summary:

TABLE 3.2 ANOVA table for simple regression

Source of Variation	*SS*	*df*	*MS*	*E(MS)*
Regression	$SSR = \sum (\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$\sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$
Error	$SSE = \sum (Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	σ^2
Total	$SSTO = \sum (Y_i - \bar{Y})^2$	$n-1$		

TABLE 3.3 ANOVA table for Westwood Company example

Source of Variation	*SS*	*df*	*MS*
Regression	13,600	1	13,600
Error	60	8	7.5
Total	13,660	9	

TABLE 3.4 Modified ANOVA table for simple regression and results for Westwood Company example

(a) General

Source of Variation	*SS*	*df*	*MS*
Regression	$SSR = \sum (\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum (Y_i - \hat{Y}_i)^2$	$n - 2$	$MSE = \frac{SSE}{n-2}$
Total	$SSTO = \sum (Y_i - \bar{Y})^2$	$n - 1$	
Correction for mean	$SS(\text{correction for mean}) = n\bar{Y}^2$	1	
Total, uncorrected	$SSTOU = \sum Y_i^2$	n	

(b) Westwood Company Example

Source of Variation	*SS*	*df*	*MS*
Regression	13,600	1	13,600
Error	60	8	7.5
Total	13,660	9	
Correction for mean	121,000	1	
Total, uncorrected	134,660	10	

PROBLEMS

3.1. A student, working on a summer internship in the economic research office of a large corporation, studied the relation between sales of a product *(Y,* in million dollars) and population *(X,* in million persons) in the firm's 50 marketing districts. Regression model (3.1) was employed. The student first wished to test whether or not a linear association between *Y* and *X* existed. Using a time-sharing computer service available to the firm, the student accessed an interactive simple linear regression program and obtained the following information on the regression coefficients:

Parameter	*Estimated Value*	*95 Percent Confidence Limits*	
Intercept	7.43119	-1.18518	16.0476
Slope	.755048	.452886	1.05721

a. The student concluded from these results that there is a linear association between *Y* and X. Is the conclusion warranted? What is the implied level of significance?
b. Someone questioned the negative lower confidence limit for the intercept, pointing out that dollar sales cannot be negative even if the population in a district is zero. Discuss.

3.2. In a test of the alternatives $H_0: \beta_1 \leq 0$ versus $H_a: \beta_1 > 0$, an analyst concluded H_0. Does this conclusion imply that there is no linear association between X and *Y?* Explain.

3.3. A member of a student team playing an interactive marketing game received the following computer output when studying the relation between advertising expenditures *(X)* and sales *(Y)* for one of the team's products:

Estimated regression equation: $\hat{Y} = 350.7 - 0.18\,X$
Two-sided P-value for estimated slope: 0.91

The student stated: "The message I get here is that the more we spend on advertising this product, the fewer units we sell!". Comment.

3.4. Refer to **Grade point average** Problem 2.15. Some additional results are:
$b_0 = -1.700$, $s(b_0) = 0.7267$, $b_1 = 0.8399$, $s(b_1) = 0.144$, $MSE = 0.1892$.

a.Obtain a 99 percent confidence interval for β_1. Interpret your confidence interval. Does it include zero? Why might the director of admissions be interested in whether the confidence interval includes zero?

b.Test, using the test statistic t^*, whether or not a linear association exists between student's entrance test score *(X)* and GPA at the end of the freshman year *(Y)*. Use a level of significance of .01. State the alternatives, decision rule, and conclusion.

c.What is the P-value of your test in part (b)? How does it support the conclusion reached in part (b)?

3.5. Refer to **Calculator maintenance** Problem 2.16. Some additional results are: b_0 = -2.3221, $s(b_0)$ = 2.564, b_1 = 14.738, $s(b_1)$ = .519, *MSE* = 20.086.

a. Estimate the change in the mean service time when the number of machines serviced increases by one. Use a 90 percent confidence interval. Interpret your confidence interval.

b. Conduct a *t* test to determine whether or not there is a linear association between X and *Y* here; control the α risk at .10. State the alternatives, decision rule, and conclusion. What is the *P-value* of your test?

c. Are your results in parts (a) and (b) consistent? Explain.

d. The manufacturer has suggested that the mean required time should not increase by more than 14 minutes for each additional machine that is serviced on a service call. Conduct a test to decide whether this standard is being satisfied by Tri-City. Control the risk of a Type I error at .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

e. Does b_0 give any relevant information here about the "start-up" time on calls- i.e., about the time required before service work is begun on the machines at a customer location?

3.6. Refer to **Airfreight breakage** Problem 2.17.

a. Estimate β_1 with a 95 percent confidence interval. Interpret your interval estimate.

b. Conduct a *t* test to decide whether or not there is a linear association between number of times a carton is transferred *(Y)* and number of broken ampules *(X)*. Use a level of significance of .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

c. β_0 represents here the mean number of ampules broken when no transfers of the shipment are made - i.e., when *X*=0. Obtain a 95 percent confidence interval for β_0 and interpret it.

d. A consultant has suggested, based on previous experience, that the mean number of broken ampules should not exceed 9 when no transfers are made. Conduct an appropriate test using α = .025. State the alternatives, decision rule, and conclusion. What is the P-value of the test?
e. Obtain the power of your test in part (b) if actually β_1 = 2.0. Assume $\sigma(b_1) = .50$. Also obtain the power of your test in part (d) if actually $\beta_0 = 11$. Assume $\sigma(b_1) = .75$.

3.7. Refer to **Plastic hardness** Problem 2. 18.

a. Estimate the change in the mean hardness when the elapsed time increases by one hour. Use a 99 percent confidence interval. Interpret your interval estimate.
b. The plastic manufacturer has stated that the mean hardness should increase by 2 Brinell units per hour. Conduct a two-sided test to decide whether this standard is being satisfied; use α = .0I. State the alternatives, decision rule, and conclusion. What is the P-value of the test?
c. Obtain the power of your test in part (b) if the standard actually is being exceeded by .5 Brinell units per hour. Assume $\sigma(b_1) = .16$.

3.8. Refer to Figure 3.7 for the Westwood Company example. A consultant has advised that an increase of one unit in lot size should require an increase of 1.8 in the expected number of man-hours for the given production item.

a. Conduct a test to decide whether or not the increase in the expected number of man-hours in the Westwood Company equals this standard. Use α= .05. State the alternatives, decision rule, and conclusion.
b. Obtain the power of your test in part (a) if the consultant's standard actually is being exceeded by .1 hour. Assume $\sigma(b_1) = .05$
c. Why is *F** = 1813.333, given in the printout, not relevant for the test in part (a)?

3.9. Refer to Figure 3.7. A student, noting that *sib,* is furnished in the printout, asks why $s(\hat{Y}_h)$ is not also given. Discuss.

3.10. For each of the following questions, explain whether a confidence interval for a mean response or a prediction interval for a new observation is appropriate.

a. What will be the humidity level in this greenhouse tomorrow when we set the temperature level at 31^0 C?
b. How much do families whose disposable income is $23,500 spend, on the average, for meals away from home?
c. How many kilowatt-hours of electricity will be consumed next month by commercial and industrial users in the Twin Cities service area, given that the index of business activity for the area remains at its present level?

3.11. A person asks if there is a difference between the "mean response at X = X_h" and the "mean of *m* new observations at X = X_h". Reply.

3.12. Can $\sigma^2(Y_{h(new)})$ in (3.36) be brought increasingly close to 0 as *n* becomes large? Is this also the case for $\sigma^2(\hat{Y}_h)$ in (3.28b)? What is the implication of this difference?

3.13. Refer to **Grade point average** Problems 2.15 and 3.4.

a. Obtain a 95 percent interval estimate of the mean freshman GPA for students whose entrance test score is 4.7. Interpret your confidence interval.
b. Mary Jones obtained a score of 4.7 on the entrance test. Predict her freshman GPA using a 95 percent prediction interval. Interpret your prediction interval.
c. Is the prediction interval in part (b) wider than the confidence interval in part (a)? Should it be?

3.14. Refer to **Calculator maintenance** Problems 2.16 and 3.5.

a. Obtain a 90 percent confidence interval for the mean service time on calls in which six machines are serviced. Interpret your confidence interval.
b. Obtain a 90 percent prediction interval for the service time on the next call in which six machines are serviced. Is your prediction interval wider than the corresponding confidence interval in part (a)? Should it be?
c. Suppose that management wishes to estimate the expected service time *per machine* on calls in which six machines are serviced. Obtain an appropriate confidence interval by converting the interval obtained in part (a). Interpret the converted confidence interval.

3.15. Refer to **Airfreight breakage** Problem 2.17.

a. Because of changes in airline routes, shipments may have to be transferred more frequently than in the past. Estimate the mean breakage for the following

numbers of transfers: $X = 2, 4$. Use separate 99 percent confidence intervals. Interpret your results.

b. The next shipment will entail two transfers. Obtain a 99 percent prediction interval for the number of broken ampules for this shipment. Interpret your prediction interval.

c. In the next several days, three independent shipments will be made, each entailing two transfers. Obtain a 99 percent prediction interval for the mean number of ampules broken in the three shipments. Convert this interval into a 99 percent prediction interval for the total number of ampules broken in the three shipments.

3.16. Refer to **Plastic hardness** Problem 2.18.

a. Obtain a 98 percent confidence interval for the mean hardness of molded items with an elapsed time of 60 hours. Interpret your confidence interval.

b. Obtain a 98 percent prediction interval for the hardness of a newly molded test item with an elapsed time of 60 hours.

c. Obtain a 98 percent prediction interval for the mean hardness of 10 newly molded test items, each with an elapsed time of 60 hours.

d. Is the prediction interval in part (c) narrower than the one in part (b)? Should it be?

3.17. An analyst fitted regression model (3.1) and conducted an F test of $\beta_1 = 0$ versus $\beta_1 \neq 0$. The P-value of the test was .033, and the analyst concluded $H_a: \beta_1 \neq 0$. Was the α level used by the analyst greater than or smaller than .033? If the α level had been .0 I, what would have been the appropriate conclusion?

3.18. For conducting statistical tests concerning the parameter β_1, why is the t test more versatile than the F test?

3.19. When testing whether or not $\beta_1 = 0$, why is the F test a one-sided test even though H_a includes both $\beta_1 < 0$ and $\beta_1 > 0$? *Hint,* Refer to (3.57)

3.20. A student asks whether r^2 is a point estimator of any parameter in regression model (3.1). Respond.

3.21. A value of r^2 near 1 is sometimes interpreted to imply that the relation between Y and X is sufficiently close so that suitably precise predictions of Y can be made from knowledge of X. Is this implication a necessary consequence of the definition of r^2?

3.22. Using regression model (3.1) in an engineering safety experiment, a researcher found for the first 10 observations that r^2 was zero. Is it possible that for the complete set of 30 observations r^2 will not be zero? Could r^2 not be zero for the first 10 observations, yet equal zero for all 30 observations? Explain.

3.23. Refer to **Grade point average** Problems 2.15 and 3.4. Some additional calculation results are: $SSE = 3.406$, $SSR = .6.434$

a. Set up the ANOVA table.

b. What is estimated by *MSR* in your ANOVA table? By *MSE?* Under what condition do *MSR* and *MSE* estimate the same quantity!

c. Conduct an *F* test of whether or not $\beta_1 = 0$. Control the ex risk at .0 I. State the alternatives, decision rule, and conclusion.

d. What is the absolute magnitude of the reduction in the variation of *Y* when *X* is introduced into the regression model? What is the relative reduction? What is the name of the latter measure?

e. Obtain *r* and attach the appropriate sign.

f. Which measure, r^2 or *r,* has the more clear-cut operational interpretation? Explain.

3.24. Refer to **Calculator maintenance** Problems 2.16 and 3.5. Some additional calculation results are: $SSE = 321.396$, $SSR = .16,182.604$

a. Set up the basic ANOVA table in the format of Table 3.2. Which elements of your table are additive? Also set up the ANOVA table in the format of Table 3.4. How do the two tables differ?

b. Conduct an *F* test to determine whether or not there is a linear association between time spent and number of machines serviced; use $\alpha = .10$. State the alternatives, decision rule, and conclusion.

c. By how much, relatively, is the total variation in number of minutes spent on a call reduced when the number of machines serviced is introduced into the analysis? Is this a relatively small or large reduction? What is the name of this measure?

d. Calculate *r* and attach the appropriate sign.

e. Which measure, r^2 or *r*, has the more clear-cut operational interpretation?

3.25. Refer to **Airfreight breakage** Problem 2.17.

a. Set up the ANOVA table. Which elements are additive?

b. Conduct an *F* test to decide whether or not there is a linear association between the number of times a carton is transferred and the number of broken ampules; control the α risk at .05. State the alternatives, decision rule, and conclusion.

c.Obtain the *t** statistic for the test in part (b) and demonstrate its equivalence to the *F** statistic obtained in part (b).

d.Calculate r^2 and *r.* What proportion of the variation in *Y* is accounted for by introducing *X* into the regression model?

3.26. Refer to **Plastic hardness** Problem 2.18.

a.Set up the ANOVA table.

b.Test by means of an *F* test whether or not there is a linear association between the hardness of the plastic and the elapsed time. Use α =.01. State the alternatives, decision rule, and conclusion.

*c.*Plot the deviations $Y_i - \hat{Y}_i$ against X_i on a graph. Plot the deviations $\hat{Y}_i - \bar{Y}$ against X_i on another graph. From your two graphs, does *SSE* or *SSR* appear to be the larger component of *SSTO?*

*d.*Calculate *r2* and *r.*

3.27. Refer to **Muscle mass** Problem 2.23.

a.Conduct a test to decide whether or not there is a negative linear association between amount of muscle mass and age. Control the risk of Type I error at .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

b.The two-sided P-value for *ha* is 0+. Can it now be concluded that *ha* provides relevant information on the amount of muscle mass at birth for a female child?

c.Estimate with a 95 percent confidence interval the difference in expected muscle mass for women whose ages differ by one year. Why is it not necessary to know the specific ages to make this estimate?

3.28. Refer to **Muscle mass** Problem 2.23.

a.Obtain a 95 percent confidence interval for the mean muscle mass for women of age 60. Interpret your confidence interval.

b. Obtain a 95 percent prediction interval for the muscle mass of a woman whose age is 60. Is the prediction interval relatively precise?

3.29. Refer to **Muscle mass** Problem 2.23.

a. Plot the deviations $Y_i - \hat{Y}_i$ against X_i on one graph. Plot the deviations $\hat{Y}_i - \bar{Y}$ against X_i on another graph. From your graphs, does *SSE* or *SSR* appear to be the larger component of *SSTO?*

b. Set up the ANOVA table.

c. Test whether or not $\beta_1 = 0$ using an *F* test with α=.10. State the alternatives, decision rule, and conclusion.

d. What proportion of the total variation in muscle mass remains "unexplained" when age is introduced into the analysis? Is this proportion relatively small or large?
e. Obtain r^2 and r.

Exercises:

3.36 Show that b_0 as defined in (3.19) is an unbiased estimator of β_0.

3.37. Derive the expression in (3.20b) for the variance of b_0 making use of theorem (3.29). Also explain how variance (3.20b) is a special case of variance (3.28b).

3.39. Suppose that the normal error regression model (3. I) is applicable except that the error variance is not constant; rather the variance is larger, the larger is X. Does β_1=0 still imply that there is no linear association between X and Y? That there is no association between X and *Y?* Explain.

3.40. Derive the expression for *SSR* in (3.50b).

3.41. In a small-scale regression study, five observations on *Y* were obtained corresponding to X = 1,4, 10, 11, and 14. Assume that σ = .6, β_0 = 5, and β_1 = 3.
a. What are the expected values of *MSR* and *MSE* here?
b. For purposes of determining whether or not a regression relation exists, would it have been better or worse to have made the five observations at X = 6, 7, 8, 9, and 10? Why? Would the same answer apply if the principal purpose were to estimate the mean response for X = 8? Discuss.

3.42. The simple linear regression model (3.1) is assumed to be applicable.

a. When testing H_0: $\beta_1 = 5$ versus H_a: $\beta_1 \neq 5$ by means of a general linear test, what is the reduced model? df_R?

b. When testing H_0: $\beta_0 = 2$, $\beta_1 = 5$ versus H_a: $\beta_0 \neq 2, \beta_1 \neq 5$, what is the reduced model? df_R?