Sample Final

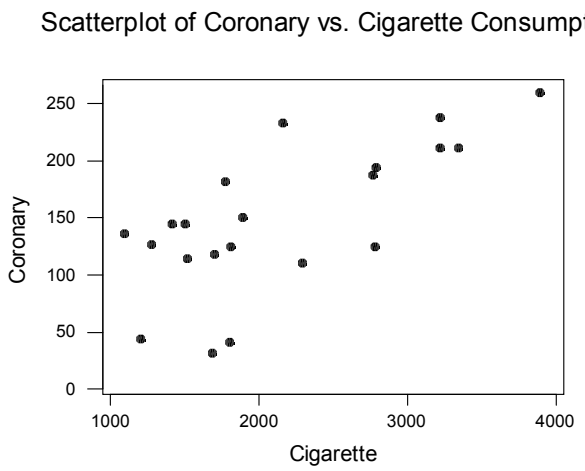
COR1-GB.1305 – Statistics and Data Analysis

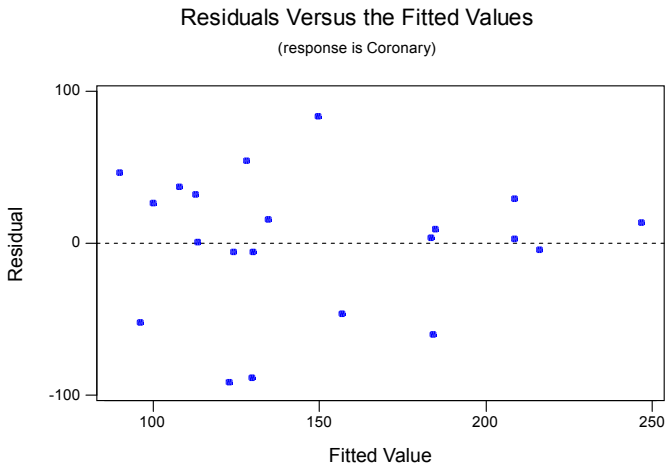
The final is open book and open note. Your are also permitted use of a calculator. Multiple choice problems are worth 5 points each. It is possible to get partial credit for an incorrect multiple choice problem answer, but only if you show your work or provide an explanation for your answer.

Problem 1 (25 points)

The following table presents data collected in the 1960s for 21 countries on X = Annual Per Capita Cigarette Consumption ("Cigarette"), and Y = Deaths from Coronary Heart Disease per 100,000 persons of age 35–64 ("Coronary").

Country	Cigarette	Coronary
United States	3900	259.9
Canada	3350	211.6
Australia	3220	238.1
New Zealand	3220	211.8
United Kingdom	2790	194.1
Switzerland	2780	124.5
Ireland	2770	187.3
Iceland	2290	110.5
Finland	2160	233.1
West Germany	1890	150.3
Netherlands	1810	124.7
Greece	1800	41.2
Austria	1770	182.1
Belgium	1700	118.1
Mexico	1680	31.9
Italy	1510	114.3
Denmark	1500	144.9
France	1410	144.9
Sweden	1270	126.9
Spain	1200	43.9
Norway	1090	136.3





Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	40484.2	40484.2	18.68	0.000
Cigarette	1	40484.2	40484.2	18.68	0.000
Error	19	41181.5	2167.4		
Lack-of-Fit	18	40835.6	2268.6	6.56	0.299
Pure Error	1	345.8	345.8		
Total	20	81665.6			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
46.5558	49.57%	46.92%	41.31%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	29.5	29.5	1.00	0.330	
Cigarette	0.0557	0.0129	4.32	0.000	1.00

Regression Equation

Coronary = 29.5 + 0.0557 Cigarette

(a) Based on the scatterplot of Y versus X, does there appear to be a linear relationship between cigarette consumption and heart disease? If so, does the relationship appear to be negative or positive?

(b) What patterns or problems, if any, do you see in the residuals versus fitted values plot? Would you feel reasonably comfortable in fitting a simple linear regression model to this data set?

(c) Write the equation for the fitted model.

(d) Give an interpretation of the fitted slope, $\hat{\beta}_1$.

(e) How much natural variability is associated with $\hat{\beta}_0$? (In other words, approximately what is the standard deviation of the random variable $\hat{\beta}_0$?)

. . . . . .

Problem 2 (25 points)

For the situation described in Problem 1, answer these questions.

(a) Based on the Minitab output, is it plausible that the true intercept β_0 is zero? Explain. What would be the practical interpretation of the result that $\beta_0 = 0$? Is there any contradiction here?

(b) Do you think that natural variability alone could account for such a large value of $\hat{\beta}_1$ as actually found here? Explain.

(c) Using the Minitab output, determine whether sufficient statistical evidence exists to conclude that there is a linear relationship between X and Y at the 1% level of significance.

(d) Based on R^2, assess the strength of the linear relationship between X and Y.

(e) Do the p-value for $\hat{\beta}_1$ and the value of R^2 provide contradictory evidence on the strength of the linear relationship between smoking and heart disease? Explain.

. . . . . .

Problem 3 (10 points)

The weights of ten \$100 casino chips (selected at random from a large batch of new \$100 chips at the Trump Castle Casino) averaged 0.8 ounces, with a sample standard deviation of 0.03 ounces.

(a) Assuming that the weights of the chips in the batch are normally distributed, construct a 95% confidence interval for the mean weight of the entire batch.

(b) Does the interval you got in part (a) have a 95% chance of containing the mean weight of the entire batch? Explain.

. . . . . .

Problem 4 (5 points)

For the situation described in Problem 3, if μ is the mean weight for the entire batch, test $H_0 : \mu = .83$ versus $H_a : \mu \neq .83$ at level .05.

. . . . . .

Problem 5 (25 points)

One hundred randomly selected milk cows were observed for one week and then given a genetically engineered drug designed to increase milk production. The increase in milk production (second week minus first week) averaged to 11 gallons with a sample standard deviation of 50 gallons.

(a) State the appropriate null and alternative hypotheses for this problem, in terms of μ.

(b) What is the meaning of μ (in terms of cows)?

(c) What do the null and alternative hypotheses imply about the effectiveness of the drug?

(d) Give all values of α at which the null hypothesis can be rejected.

(e) Suppose the drug had no effect. Then out of 1000 random samples of 100 cows, how many samples would be expected to yield an increase in milk production at least as large as what was found in our sample?

. . . . . .

Questions 6–9 concern the following situation. A random sample of 50 adults were asked how much they spend on lottery tickets, and were interviewed about various socioeconomic variables. The variables are

PercLott = Percentage of total household income spent on the lottery. (This is Y).
YrsEdu = Number of years of education,
Age = The persons Age,
Kids = Number of Children,
Income = Personal income (Thousands of Dollars).

Here is the Minitab regression output:

```
Analysis of Variance

Source     DF  Adj SS  Adj MS  F-Value  P-Value
Regression  4  404.42  101.10    17.72    0.000
  YrsEdu    1   60.68   60.68    10.63    0.002
  Age       1    0.21    0.21     0.04    0.850
  Kids      1    0.55    0.55     0.10    0.761
  Income    1   23.30   23.30     4.08    0.050
Error      45  256.80    5.71
Total      49  661.22

Model Summary

    S    R-Sq  R-Sq(adj)  R-sq(pred)
2.389  61.16%     57.71%      52.16%

Coefficients

Term        Coef   SE Coef  T-Value  P-Value   VIF
Constant  15.070     2.444     6.17    0.000
YrsEdu   -0.5911    0.1813    -3.26    0.002  1.47
Age      0.00647   0.03395     0.19    0.850  2.81
Kids      0.0816    0.2665     0.31    0.761  1.93
Income  -0.06663   0.03305    -2.02    0.050  1.58

Regression Equation

PercLott = 15.1 - 0.591 YrsEdu + 0.0065 Age + 0.082 Kids - 0.0666 Income
```

Problem 6

Based on the output, is there statistical evidence to suggest that relatively educated people spend a different amount on lotteries than relatively uneducated people?

(a) Yes

(b) No

.....

Problem 7

The results of the F test imply that, beyond a reasonable doubt:

(a) All of the true slope coefficients in the model are nonzero

(b) At least one of the true slope coefficients in the model is nonzero

(c) None of the true slope coefficients in the model is nonzero

(d) All of the estimated slope coefficients are nonzero

(e) At least one of the estimated slope coefficients is nonzero

. . . . . .

Problem 8

The 95% confidence interval for the true coefficient of YrsEdu is

(a) (-2.12, 3.14)

(b) (-0.5911, 0.5911)

(c) (-1,1)

(d) (-0.956, -0.226)

(e) (-1.06, -0.124).

. . . . . .

Problem 9

Performing a two-tailed hypotesis test for the null hypothesis that the true coefficient of YrsEdu is -1, at the 5% level of significance, we:

(a) Reject the null hypothesis

(b) Do not reject the null hypothesis

. . . . . .

Problem 10

Let's return to the simple regression described in Problem 1. The residual for Greece is:

(a) 1800

(b) 29.45

(c) 31.74

(d) 1768.26

(e) -88.474

. . . . . .

Problem 11

A sample of size 100 is going to be taken from a population with mean 3 and variance 25. The probability that the sample mean will exceed 4 is approximately:

(a) .0456

(b) .4207

(c) .0793

(d) .3446

(e) .0228

. . . . . .

Problem 12

Suppose that X and Y are independent random variables with $P(X > 4) = 0.8$ and $P(Y > 5) = 0.6$. The probability that X exceeds 4 and Y exceeds 5 is

(a) 1.4

(b) 0.92

(c) 0

(d) 0.48

(e) Not enough information to determine

. . . . . .