Sample Final

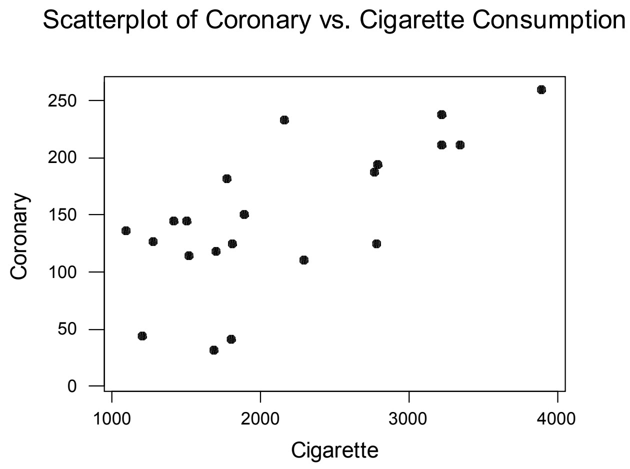
STAT-UB.0003 – Regression and Forecasting

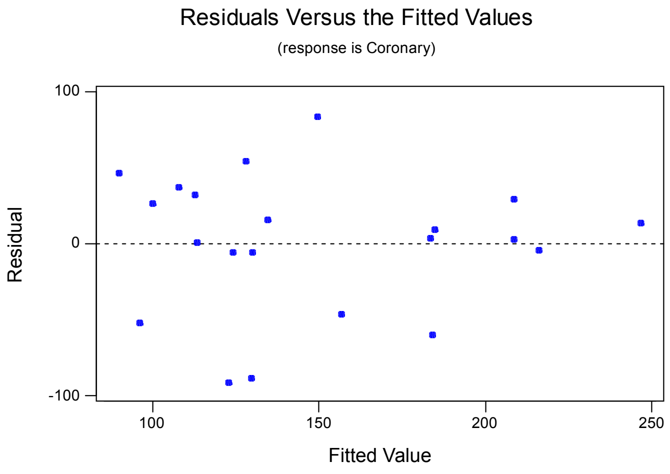
The final is open book and open note. Your are also permitted use of a calculator. Multiple choice problems (3–10) are worth 5 points each. It is possible to get partial credit for an incorrect multiple choice problem answer, but only if you show your work or provide an explanation for your answer.

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





```
The regression equation is
Coronary = 29.5 + 0.0557 Cigarette

Predictor        Coef     SE Coef         T        P
Constant        29.45       29.48      1.00    0.330
Cigarett      0.05568     0.01288      4.32    0.000

S = 46.56       R-Sq = 49.6%     R-Sq(adj) = 46.9%

Analysis of Variance

Source            DF         SS          MS         F        P
Regression         1      40484       40484     18.68    0.000
Residual Error    19      41181        2167
Total             20      81666
```

(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.

. . . .

Questions 3–6 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:

```
The regression equation is
PercLott = 15.1 - 0.591 YrsEdu + 0.0065 Age + 0.082 Kids - 0.0666 Income

Predictor       Coef     SE Coef          T        P
Constant      15.070       2.444       6.17    0.000
YrsEdu       -0.5911      0.1813      -3.26    0.002
Age          0.00647     0.03395       0.19    0.850
Kids          0.0816      0.2665       0.31    0.761
Income      -0.06663     0.03305      -2.02    0.050

S = 2.389       R-Sq = 61.2%     R-Sq(adj) = 57.7%

Analysis of Variance

Source          DF          SS         MS        F        P
Regression       4      404.42     101.10    17.72    0.000
Residual Error  45      256.80       5.71
Total           49      661.22
```

Problem 3

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 4

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 5

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 6

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 7

Lets 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 8

In linear regression, does a point with high leverage necessarily cause the fitted line to change?

(a) Yes

(b) No

. . . .

Problem 9

Recall that the formula for Akaike's Information Criterion (AIC) is

$$\text{AIC} = n \log \frac{\text{SSE}}{n} + 2(k+1).$$

where SSE is the sum of squares of residual errors and k is the number of predictors in the model. Which of the following model selection methods attempts to find the optimal tradeoff between bias and variance:

(a) Choosing the model with the smallest value of AIC

(b) Choosing the model with the largest value of AIC

. . . .

Problem 10

Suppose we are trying to predict the total box office gross of a movie, in millions of dollars. We measure the following predictors:
Advertising = amount spent advertising for the movie, in hundreds of thousands of dollars
Genre = Comedy, Action, or Drama
We introduce dummy variables for "Comedy" and "Drama", and use an interaction between Genre and Advertising. The fitted model is

```
Gross = 2.1 + 3.11 Advertising + 10 Comedy - 5.7 Drama
        - 2.84 Advertising*Comedy + 5.92 Advertising*Drama
```

According to the model, the increase in mean Gross (in millions) corresponding to a $100,000 increase in advertising for a Comedy movie is

(a) -2.84

(b) 0.27

(c) 3.11

(d) 10.27

(e) 13.11

. . . .