There are 4 questions. If you get stuck on one part, move on and do the rest. GOOD LUCK!

1. A few years ago, *New York Times* published an article titled "Wine for the Heart: Over All, Risks May Outweigh Benefits."

Motivated by the article, we wish to estimate equations such as

$$heart = \beta_0 + \beta_{alc} alcohol + u.$$

The variables are
alcohol: per capita consumption of liters of wine
heart: deaths due to heart disease per 100,000

a. Say, we first estimate the following simple regression using data on *n* = 150 countries:

$$\widehat{heart} = 239.147 - 19.683 alcohol.$$

Interpret the slope in this equation and explain its sign and magnitude.

Answer: As per capita alcohol consumption increases by 1 liter, deaths due to heart disease decreases by 19.68 per 100,000 people.

b. Next, the following simple regression is also estimated:

$$\log(\widehat{heart}) = 5.361 - 0.353 \log(alcohol).$$

Interpret the slope in this equation and explain its sign and magnitude.

Answer: As per capita alcohol consumption increases by 1%, deaths due to heart disease per 100,000 people decreases by 0.353%.

c. Do the simple regressions above obtain unbiased estimators of the effect of country-level alcohol consumption and deaths due to heart disease? Explain.

Answer: The estimators above based on a simple regression model are unlikely to be unbiased. A number of factors such as a country's average education and income levels are likely correlated with alcohol consumption as well as heart disease.

2. The results below correspond to a regression output. The data set contains data on colleges and the variables are
enroll: total enrollment
police: employed officers
crime: total campus crimes
lcrime: log(crime)
lenroll: log(enroll)
lpolice: log(police).

The equation of interest is given by:

$$\log(crime) = \beta_0 + \beta_{lpolice} \log(police) + \beta_{lenroll} \log(enroll) + u.$$

```
===============================================
                         Dependent variable:
                     ---------------------------
                               lcrime
-----------------------------------------------
lpolice                       0.516***
                               (0.149)

lenroll                       0.923***
                               (0.144)

Constant                      -4.794***
                               (1.112)

-----------------------------------------------
Observations                     97
R2                              0.632
Adjusted R2                     0.624
Residual Std. Error        0.847 (df = 94)
F Statistic            80.720*** (df = 2; 94)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

a. What does the R-squared value of 0.632 imply?

Answer: The R^2 value implies that about 63% of the variation in log(*crime*) is explained by log(*police*) and log(*enroll*).

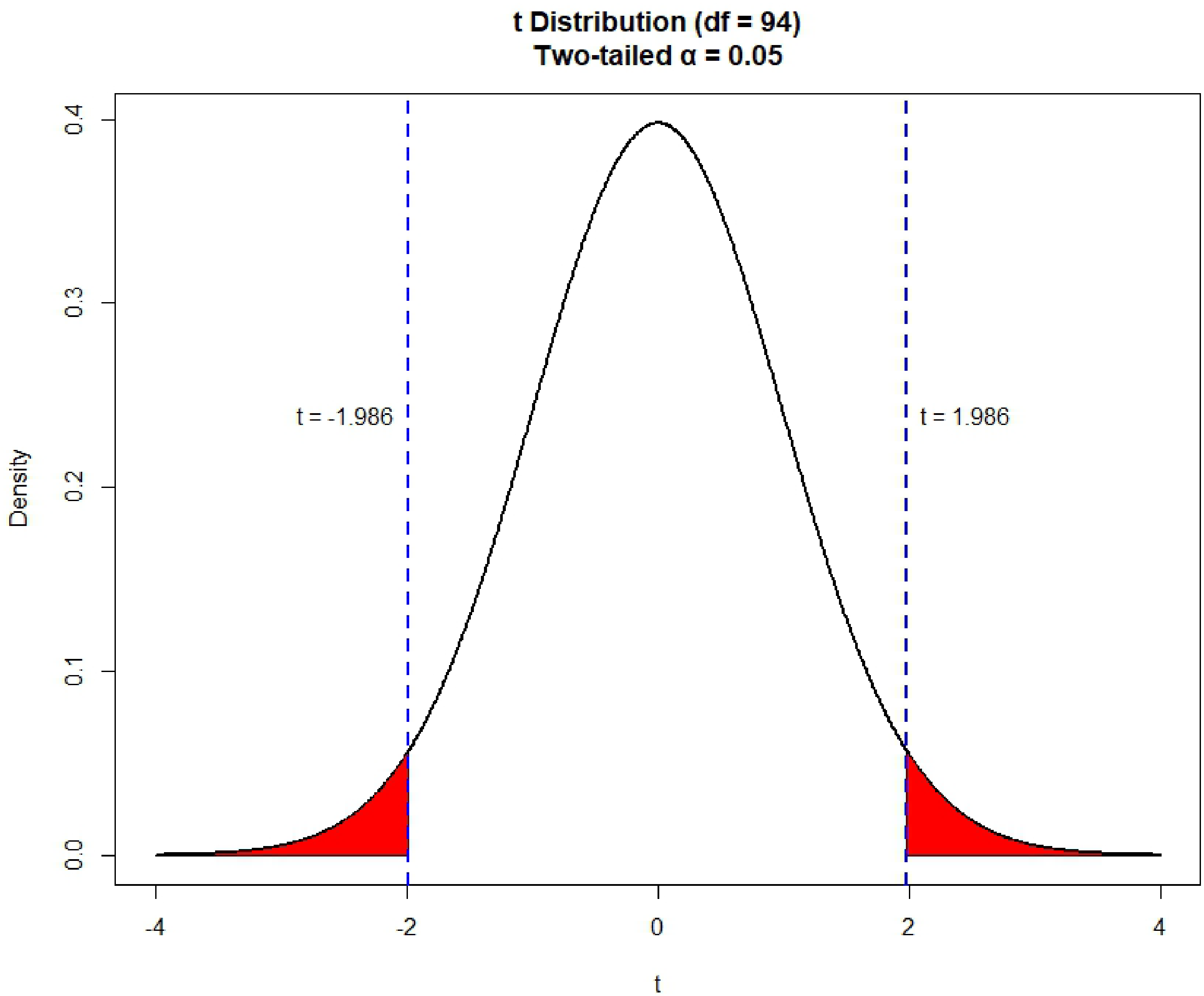
b. Assuming a two-tailed test where H_0: $\beta_{lenroll} = 0$, is the coefficient estimate corresponding to log(*enroll*) statistically significant (i.e., H_0 is rejected) at the 2% level of significance?

Answer: Yes. From the table, the p-value is less than 0.02.

b. For the test described in (a), what is the (numerical) value of the *t* test statistic? What are the critical values at the 5% level of significance? What is the 95% confidence interval? What is the p-value?

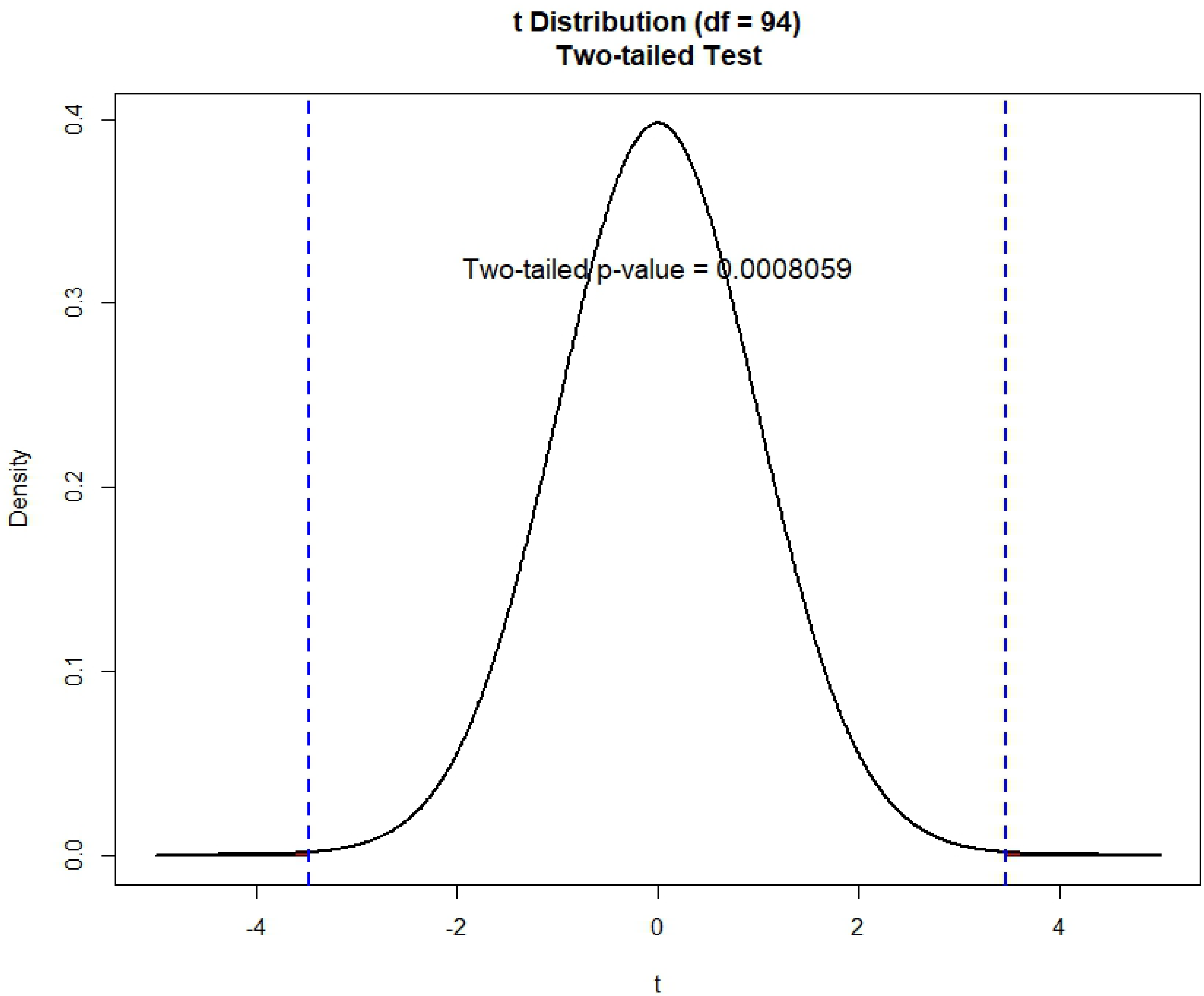
Answer: The test statistic is given by 0.516/0.149 = 3.463.

The critical values are obtained from Table G.2. Here, the degrees of freedom = 97 – 3 = 94; for 90 degrees of freedom, the critical values are -1.987 and 1.987. Using R, the exact critical values are -1.986 and 1.986.



The 95% confidence interval is given by [0.516 – 1.986 × 0.149, 0.516 + 1.986 × 0.149], i.e., [0.220, 0.812].

The p-value is given by 2 × P(|t| > 3.463); using R this is 0.001.

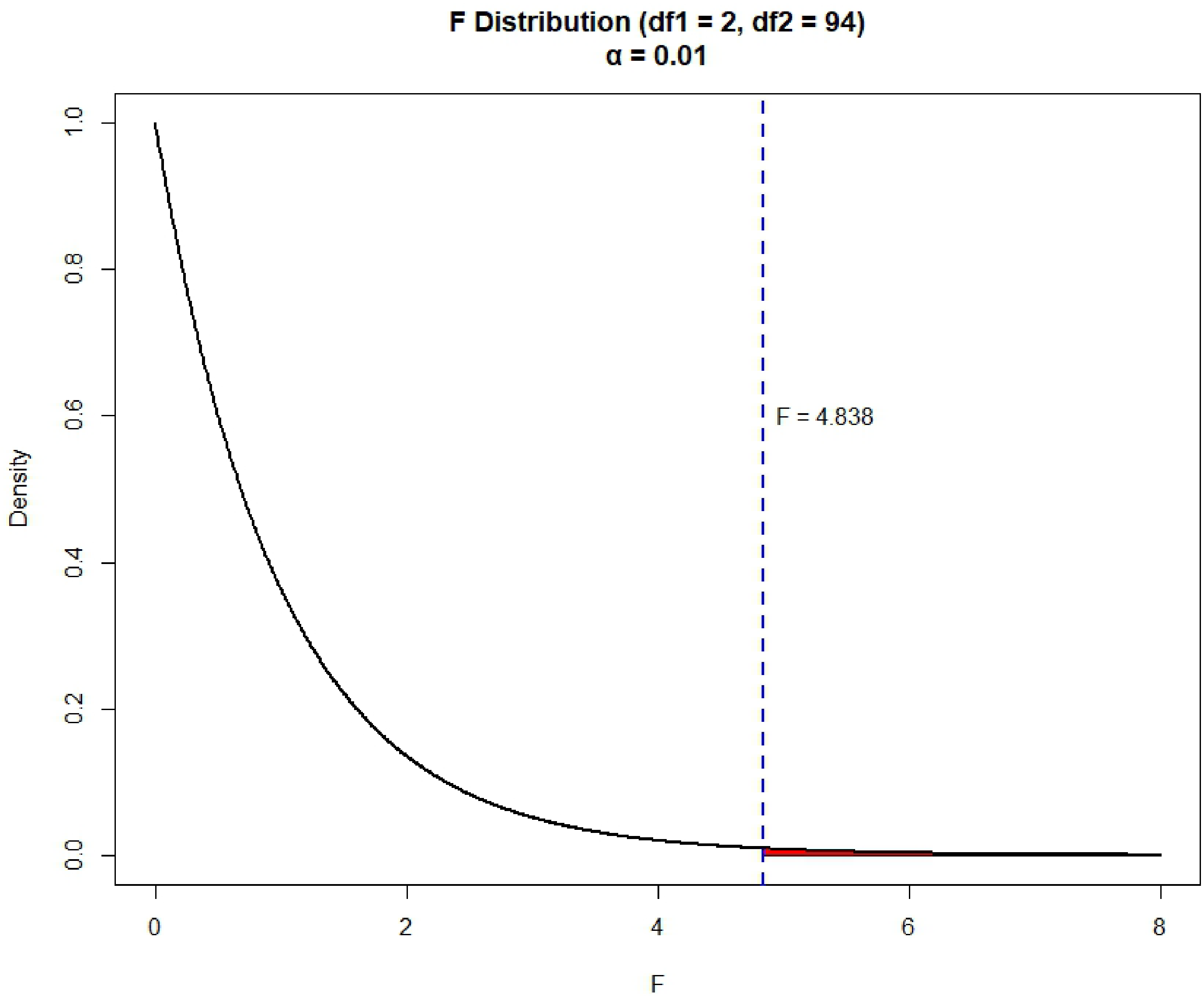


c. What is the (numerical) value of the *t* test statistic to test whether the coefficient estimate corresponding to log(*police*) is significantly different from 0.7 (i.e., H_0: $\beta_{lpolice}$=0.7)? It is fine to use values up to two decimal points.

Answer: The test statistic is given by (0.516 – 0.7)/0.149 = -1.23.

d. Suppose we are jointly testing whether the slope coefficients corresponding to log(*police*) and log(*enroll*) are zero, i.e., H_0: β_{lenroll}=0 and β_{lpolice}=0. Write the R-squared form of *this* F statistic. Do you reject the hypothesis at the 1% level of significance?

Answer: In this case, the unrestricted R^2 is the R^2 from the above regression. The restricted R^2 is zero. So, the F statistic is given by $[R^2/q] / [(1 - R^2)/(n - k - 1)]$, i.e., [0.632/2] / [(1 – 0.632)/(97 – 2 – 1)] = 80.72. This is displayed in the results table. From Table G.3c, the critical value is about 4.85. Hence, we reject H_0.



3. Answer the following briefly:

a. In a regression model, what is the average (numerical) value of the residuals?

Answer: Zero.

b. In a regression model, what is the (numerical) value of correlation between the residuals and each explanatory variable?

Answer: Zero.

c. Is the assumption of homoskedasticity required for unbiasedness of $\hat{\beta}_j$?

Answer: No.

d. Is the assumption of normality required for unbiasedness of $\hat{\beta}_j$?

Answer: No.

4. Consider a data set on prices of homes. The variables are
price: house price, \$1000s
bdrms: number of bdrms
sqrft: size of house in square feet
school: quality (i.e., rating on a 1 to 5 scale) of schools in the neighborhood
crime: crime rate (per 100,000) in the neighborhood.

Our multiple regression of interest is

$$\log(price) = \beta_0 + \beta_1 bdrms + \beta_2 \log(sqrft) + \beta_3 school + \beta_4 crime + u.$$

Here, *u* represents unobserved factors affecting housing prices such as quality, neighboring property values, and proximity to other amenities.

a. Suppose *bdrms* and *sqrft* are endogenous and also correlated with *school* and *crime*. For ordinary least squares (OLS), should we only worry about bias in our estimates of β_1 and β_2?

Answer: No, the estimates of al the β's are potentially biased.

b. Suppose we disregard the information on endogeneity in (a). However, $\log(price)$ affects *crime* just as *crime* influences property values. Should we worry about bias in our estimate of β_4?

Answer: Yes, due to simultaneity.

c. Suppose we disregard the information on endogeneity and simultaneity in (a) and (b). However, $\log(sqrft)$ is measured with error for lower quality houses. Should we worry about bias in our estimate of β_2?

Answer: Yes, due measurement error. Suppose the observed *sqrft* is the sum of true square footage and measurement error. If the error is correlated with *u*, the observed *sqrft* is again endogenous.

d. Suppose *bdrms* and $\log(sqrft)$ are highly correlated. Should we worry about bias in our estimate of β_1 and β_2?

Answer: No, although we have the issue of multicollinearity.

e. If *u* does not follow a normal distribution, can we still assume our test statistics to follow distributions such as *t* or *z*?

Answer: Yes, provided we have a large sample.