Problem Set 3

If you would like a small amount of feedback on your work, question 1 can be handed in to your class teacher. You can do this through blackboard (under the heading 'Exercise Hand-in'). No marks will be given for this and this will have no impact on your final exam grade. The deadline to hand the work in is Monday 21st 09:00am.

1. Suppose the true model is given by

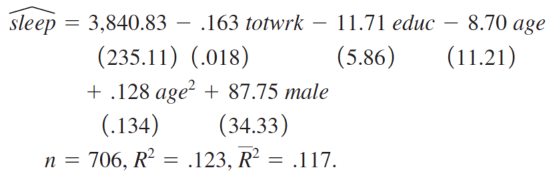
$$y = \beta_1 x + \beta z + \epsilon.$$

(Note that we have not included the constant term to simplify the analysis, it doesn't change anything in a meaningful way. When x, z, and y are zero mean the intercept will always be equal to 0)

(i) However, we estimate a short regression where we omit z. What is the variance of the OLS estimator conditional on x and z? Assume assumptions MLR1-MLR5 hold. (Hint: use scalar notation. Hint 2: this is a bit of a trick question).

(ii) Now suppose you estimate the full model by OLS. Calculate the variance of $\hat{\beta}_1$ conditional on x and z. (Hint: use our variance formula $\sigma^2(X'X)^{-1}$... what is X in this case? Hint 2: $\frac{(\sum x_i z_i)^2}{\sum x_i^2 \sum z_i^2}$ represents the R^2 from a regression of x on z, where both variables have zero mean.)

2. Using the data in 'sleep75', we obtain the estimated equation



$$\begin{aligned}\widehat{sleep} = {} & \underset{(235.11)}{3{,}840.83} - \underset{(.018)}{.163}\, totwrk - \underset{(5.86)}{11.71}\, educ - \underset{(11.21)}{8.70}\, age \\ & + \underset{(.134)}{.128}\, age^2 + \underset{(34.33)}{87.75}\, male \\ n = 706, R^2 = .123, \overline{R}^2 = .117. \end{aligned}$$

The variable *sleep* is total minutes per week spent sleeping at night, *totwrk* is total weekly minutes spent working, *educ* and *age* are measured in years, and *male* is a gender dummy.

(i) All other factors being equal, is there evidence that men sleep more than women? How strong is the evidence?

(ii) Is there a statistically significant tradeoff between working and sleeping? What is the estimated tradeoff?

(iii) What other regression do you need to run to test the null hypothesis that, holding other factors fixed, age has no effect on sleeping? (Note that two variables capture the effect of age)

3. Using the data in 'gpa2', the following equation was estimated:

$$
\begin{aligned}
\widehat{sat} = {} & \underset{(6.29)}{1{,}028.10} + \underset{(3.83)}{19.30}\, hsize - \underset{(.53)}{2.19}\, hsize^2 - \underset{(4.29)}{45.09}\, female \\
& - \underset{(12.71)}{169.81}\, black + \underset{(18.15)}{62.31}\, female{\cdot}black \\
n = {} & 4{,}137,\ R^2 = .0858.
\end{aligned}
$$

The variable *sat* is the combined SAT score; *hsize* is size of the student's high school graduating class, in hundreds; *female* is a gender dummy variable; and *black* is a race dummy variable equal to one for black people, and zero otherwise.

(i) Is there strong evidence that *hsize*2 should be included in the model? From this equation, what is the optimal high school size?

(ii) Holding *hsize* fixed, what is the estimated difference in SAT score between nonblack females and nonblack males? How statistically significant is this estimated difference?

(iii) What is the estimated difference in SAT score between nonblack males and black males? Test the null hypothesis that there is no difference between their scores, against the alternative that there is a difference.

(iv) What is the estimated difference in SAT score between black females and nonblack females? What would you need to do to test whether the difference is statistically significant?

4. Use the data in 'discrim' to answer this question.

(i) Use OLS to estimate the model

$$log(soda) = \beta_0 + \beta_1 prpblck + \beta_2 log(income) + \beta_3 prppov + \epsilon$$

and report the results in the usual form. Is $\hat{\beta}_1$ statistically different from zero at the 5% level against a two-sided alternative? What about at the

1% level?

(ii) What is the correlation between $\log(income)$ and *prppov*? Is each variable statistically significant in any case? Report the two-sided p-values.

(iii) To the regression in part (i), add the variable $\log(hseval)$ (the median house value in the zip code area). Interpret its coefficient and report the two-sided p-value for H_0: $\beta_{log(hseval)}$.

(iv) In the regression in part (iii), what happens to the individual statistical significance of $\log(income)$ and *prppov*? Are these variables jointly significant? (Compute a p-value.) What do you make of your answers?

(v) Given the results of the previous regressions, which one would you report as most reliable in determining whether the racial makeup of a zip code influences local fast-food prices?