Assignment 2

Questions 1 to 4 are based on the following multiple linear regression

$$y = X\beta + \epsilon\,,\ E(\epsilon \mid X) = 0\,,\ Var(\epsilon \mid X) = \sigma^2 I_n\,,$$

where y is an $n \times 1$ vector and X is a $n \times k$ matrix with full column rank, i.e., there are $k-1$ regressors plus an intercept term.

Question 1

(a) Show that $y^{\mathrm{T}}y = y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y + \hat{\epsilon}_{ols}^{\mathrm{T}}\hat{\epsilon}_{ols}$.

(b) If A is a $n \times n$ matrix, and e_i is the $n \times 1$ vector with 1 in the ith position and all other terms zero, show that $e_i^{\mathrm{T}}Ae_i = a_{ii}$, the ith diagonal element of A.

(c) Let h_{ii} is the ith diagonal element of the matrix $X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, i.e.,

$$h_{ii} = e_i^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}e_i = X_{i*}(X^{\mathrm{T}}X)^{-1}X_{i*}^{\mathrm{T}}\,,$$

where X_{i*} is the ith row of the X matrix.

i. Show that $0 \le h_{ii} \le 1$. *Hint: replace y with e_i in part (a).*

ii. Show that $\sum_{i=1}^{n} h_{ii} = k$.

Question 2

(a) Show that $\hat{\epsilon}_{ols} = M\epsilon$ where $M = I_n - P = I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$.

(b) Show that $E(\epsilon_{i,ols}^2 \mid X) = \sigma^2(1 - h_{ii})$. (Hint: $E(\epsilon_{i,ols}^2 \mid X)$ is the i-diagonal element of the variance matrix $Var(\epsilon_{ols} \mid X)$).

Remark: In other words, each squared residual is a downward biased estimate of σ^2.

(c) Show using the result in part (b) of this question and part (b.ii) of Qn. 1, that

$$\widehat{\sigma^2} = \frac{1}{n-k}\sum_{i=1}^{n}\hat{\epsilon}_{ols}^2$$

is an unbiased estimator for σ^2.

Question 3 (Influential Observations)

Suppose the ith observation is omitted, and let $\hat{\beta}_{(-i)}$ is the OLS estimator for β based on the remaining observations. It can be shown (but you're not being asked to do so) that

$$\hat{\beta}_{(-i)}^{ols} = \hat{\beta}^{ols} - \left(\frac{1}{1-h_{ii}}\right)(X^{\mathrm{T}}X)^{-1}X_{i*}^{\mathrm{T}}\hat{\epsilon}_i^{ols}\,.$$

An observation can be considered to be an *influential observation* if omitting it changes its fitted/predicted value substantially. The fitted value for observation i is $X_{i*}\widehat{\beta}^{ols}$ and the predicted value for observation i when it is omitted from the sample is $X_{i*}\widehat{\beta}^{ols}_{(-i)}$. Show that

$$X_{i*}\widehat{\beta}^{ols} - X_{i*}\widehat{\beta}^{ols}_{(-i)} = \left(\frac{h_{ii}}{1-h_{ii}}\right)\widehat{\epsilon}_i^{ols}\,.$$

Remark: One easy way to check for influential observations is to plot $h_{ii}/(1-h_{ii})$ against i. Alternatively, since $0 \le h_{ii} \le 1$ and its average value is k/n, we can simply look at the h_{ii} to see if any are very close to 1.

Question 4 (Leave-One-Out-Cross-Validation)

To check how well the fitted model might perform as a predictive model, we estimate the mean squared prediction error in the following manner:

- for each i, we leave out the i-th observation and and use the remaining $n-1$ observations to fit the model. We use this fitted model to predict the i-th observation and collect the prediction error
$$\widehat{\epsilon}_{(-i)} = Y_i - X_{i*}\widehat{\beta}^{ols}_{(-i)}\,.$$
- After completing the above for all i, we estimated the "Leave-One-Out-Cross-Validated" Mean Square Prediction Error (LOOCV-MSPE) as
$$\frac{1}{n}\sum_{i=1}^{n}\widehat{\epsilon}^{2}_{(-i)}\,.$$

It turns out that we do not need to estimate n regressions in order to calculate the LOOCV-MSPE. Show that

$$\widehat{\epsilon}_{(-i)} = \frac{\widehat{\epsilon}_i^{ols}}{1-h_{ii}}\,.$$

Question 5

The dataset `ceosal1` from the `wooldridge` library contains 209 observations CEO salaries. We will use *lsalary* (log of 1990 salary), *lsales* (log of 1990 sales), *reo* (return on equity 1988-1990 average) and *ros* (return on firm stock 1988-1990).

Create two new series: $rosneg = 1$ if $ros \le 0$, 0 otherwise, and $rosfiltered = ros$ if $ros > 0$, 0 otherwise. You can use the following two commands to do this:

```
dat$rosneg <- (dat$ros<0)
dat$rosfiltered <- ifelse(dat$ros>0, dat$ros, 0)
```

The variable *rosneg* is actually a `TRUE/FALSE` series which will be converted to `1`s and `0`s by the `lm()` function.

(a) Explain how ros affects $lsalary$ in each of the three regressions below:

$$\text{[A]} \qquad lsalary = \beta_0 + \beta_1 lsales + \beta_2 roe + \beta_3 ros + \epsilon$$

$$\text{[B]} \qquad lsalary = \beta_0 + \beta_1 lsales + \beta_2 roe + \beta_3 rosneg + \epsilon$$

$$\text{[C]} \qquad lsalary = \beta_0 + \beta_1 lsales + \beta_2 roe + \beta_3 ros + \beta_4 rosfiltered + \epsilon$$

(b) Estimate the three regressions in part (a), and report your results using heteroskedasticity-robust standard errors (use the `vcovHC()` function from the `sandwich` package, with `type=HC0`). Comment on your results, with a focus on the estimated effect of ros on $lsalary$.

(c) Carry out heteroskedasticity-robust RESET tests on all three models, using the square and cube of the fitted values. Comment on your results.

Question 6

The data set `rdchem` from the `wooldridge` library contains observations on 32 firms from the chemicals industry in 1990. We are interested in the variables $rdintens$ (research and development as a percentage of sales), $sales$ in millions, $salessq$ ($sales^2$), and $profmarg$ (profits as percentage of sales).

(a) First convert $sales$ and $salessq$ to billion dollars and (billion dollars)^2 respectively, then estimate the regression

$$\text{[A]} \qquad rdintens = \beta_0 + \beta_1 sales + \beta_2 salessq + \beta_3 profmarg + \epsilon$$

and report your results (you can assume homoskedastic errors).

(b) Compute h_{ii} as defined in Qn 3. Which observations appear to be influential? (You can construct the X matrix yourself, or use `X <- model.matrix(mdl)` where `mdl` is the name of the regression you estimated with the `lm()` object.)

(c) Drop the most influential observation and re-estimate the regression. Are there any major changes in the estimation results?

(d) Dropping influential observations might not be the best solution. The data observation may contain important information, and dropping one influential observation might turn another observation into an influential one. One alternative is to use Least Absolute Deviation (LAD) estimation, where we choose estimators to minimize the sum of *absolute* residuals rather than the sum of squared residuals. That is, we minimize

$$\sum_{i=1}^{n} \mid Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_k X_{ik} \mid .$$

This can be done in R using the `rq()` function. Install the `quantreg` library and then use the following code to obtain the LAD estimates for equation [A]. Comment on your result.

```
library(quantreg)
mdl <- rq(rdintens ~ sales + salessq + profmarg, data=dat)
summary(mdl, se = "iid")
```