STOR 664 CWE Summer 2017

(1) **Read the following example taken from Faraway's book; answer the inserted questions and make your comments.**

The original data (R output-1) is collected in an experiment to determine the effects of column temperature X_1, gas/liquid ratio X_2 and packing height X_3 in reducing the unpleasant odor Y of a chemical product that was sold for household use. The three predictors have been transformed from their original scale of measurement, e.g. temp = (Fahrenheit - 80)/40, etc.

R output-1

```
> data(odor, package="faraway")
> odor
   odor temp gas pack
1    66   -1  -1    0
2    39    1  -1    0
3    43   -1   1    0
4    49    1   1    0
5    58   -1   0   -1
6    17    1   0   -1
7    -5   -1   0    1
8   -40    1   0    1
9    65    0  -1   -1
10    7    0   1   -1
11   43    0  -1    1
12  -22    0   1    1
13  -31    0   0    0
14  -35    0   0    0
15  -26    0   0    0
```

Now we fit the model:

R output-2

```
> lmod <- lm(odor ~ temp + gas + pack, odor)
> summary(lmod, cor=T)

Call:
lm(formula = odor ~ temp + gas + pack, data = odor)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-50.200 -17.138   1.175  20.300  62.925

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.200      9.298   1.635    0.130
temp         -12.125     12.732  -0.952    0.361
gas          -17.000     12.732  -1.335    0.209
pack         -21.375     12.732  -1.679    0.121

Residual standard error: 36.01 on 11 degrees of freedom
Multiple R-squared:  0.3337,    Adjusted R-squared:  0.1519
F-statistic: 1.836 on 3 and 11 DF,  p-value: 0.1989
```

(1a) What is the expression of the fitted model based on the data? Write down the estimated covariance matrix for the least square estimator (LSE) $\widehat{\beta}$? Why in R output-2 all the components of $\widehat{\beta}$ (except for the intercept) have the same standard errors?

Now we drop the predictor X_1 (temperature) and refit the model:

R output-3

```
> lmod <- lm(odor ~ gas + pack, odor)
> summary(lmod)

Call:
lm(formula = odor ~ gas + pack, data = odor)

Residuals:
    Min      1Q  Median      3Q     Max
-50.200 -26.700   1.175  26.800  50.800

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.200      9.262   1.641    0.127
gas          -17.000     12.683  -1.340    0.205
pack         -21.375     12.683  -1.685    0.118

Residual standard error: 35.87 on 12 degrees of freedom
```

```
Multiple R-squared:  0.2787,    Adjusted R-squared:  0.1585
F-statistic: 2.319 on 2 and 12 DF,  p-value: 0.1408
```

(1b) Note that the coefficients for X_2 and X_3 remain the same as those in R output-2, but their corresponding standard errors get slightly smaller. Are such results coincidental? If not, explain why.

(1c) Are the results of model fitting shown in R output-2 and R output-3 satisfactory? If not, comment on possible reasons for the pitfalls and any steps you may take in a further study.

(2) Suppose a (deterministic) linear equation $y_0 = \beta_0 + \beta_1 x_0$ holds at a known point (x_0, y_0) with unknown parameters β_0 and β_1. A further statistical study is carried out for the regression model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i,\ i = 1, ..., n$ where $\epsilon_i,\ i = 1, ..., n$ are iid $N(0, \sigma^2)$ random variables, $x_1, ..., x_n$ are known constants, and β_2 and σ^2 are unknown parameters.

(2a) Find the least squares estimators for β_0, β_1 and β_2. (**Hint:** Different methods may be available. You can treat the problem as a reduced model under the constraint $y_0 = \beta_0 + \beta_1 x_0$ in a null hypothesis, or transform the responses y_i to $y_i - y_0 = ...$, whichever is more convenient to you.)

(2b) Find an unbiased estimator for $\sigma^2 > 0$.

(2c) Assume the overall setting before (2a), i.e. suppose for each $i = 1, ..., n$, the observed response follows the model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ with the constraint $y_0 = \beta_0 + \beta_1 x_0$. However, we are mislead to believe the simpler model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ is the true one and get the LSE $\widehat{\beta}_0^{(0)}$ and $\widehat{\beta}_1^{(0)}$, and also use $\widehat{y}_i^{(0)} = \widehat{\beta}_0^{(0)} + \widehat{\beta}_1^{(0)} x_i$ to predict $y_i^{(1)} \triangleq \beta_0 + \beta_1 x_i + \beta_2 x_i^2$, (note: no additional error ϵ_i^* involved). Find the mean squared error (MSE) $E\left[\sum_{i=1}^n (\widehat{y}_i^{(0)} - y_i^{(1)})^2\right]$.