**HW 6 (Uploaded on Apr. 19, 2017)** **Name:**

**HW 6.1.** To investigate the relationship between hydrocarbon and oxygen purity, the following data are collected.

| Observation Number | Hydrocarbon Level $x(\%)$ | Purity $y(\%)$ | Observation Number | Hydrocarbon Level $x(\%)$ | Purity $y(\%)$ |
|---|---|---|---|---|---|
| 1 | 0.99 | 90.01 | 11 | 1.19 | 93.54 |
| 2 | 1.02 | 89.05 | 12 | 1.15 | 92.52 |
| 3 | 1.15 | 91.43 | 13 | 0.98 | 90.56 |
| 4 | 1.29 | 93.74 | 14 | 1.01 | 89.54 |
| 5 | 1.46 | 96.73 | 15 | 1.11 | 89.85 |
| 6 | 1.36 | 94.45 | 16 | 1.20 | 90.39 |
| 7 | 0.87 | 87.59 | 17 | 1.26 | 93.25 |
| 8 | 1.23 | 91.77 | 18 | 1.32 | 93.41 |
| 9 | 1.55 | 99.42 | 19 | 1.43 | 94.98 |
| 10 | 1.40 | 93.65 | 20 | 0.95 | 87.33 |

The purity is the response $Y$, and hydrocarbon level is the regression $x$. A simple linear regression model is assumed as

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$. Use $R$ to run the following codes (for performing linear regression analysis).

```
Hydro=c(0.99, 1.02, 1.15, 1.29, 1.46, 1.36, 0.87, 1.23, 1.55, 1.4, 1.19,
1.15, 0.98, 1.01, 1.11, 1.2, 1.26, 1.32, 1.43, 0.95)
Purity=c(90.01, 89.05, 91.43, 93.74, 96.73, 94.45, 87.59, 91.77, 99.42, 93.65, 93.54,
 92.52, 90.56, 89.54, 89.85, 90.39, 93.25, 93.41, 94.98, 87.33)
fit=lm(Purity~Hydro)
summary(fit)
```

Then answer the following questions.

(a) Testing $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$. Use significant level $\alpha = 0.05$.

(b) Find the estimate of $\sigma$

(c) Using the codes to find the 95% confidence and prediction intervals at $x = 1\%$. Which one is larger? Why?

```
predict(fit,data.frame(Hydro=1),level=0.95,interval="confidence")
predict(fit,data.frame(Hydro=1),level=0.95,interval="prediction")
```

**Sol.** The followings are `R` outputs:

```
Call:
lm(formula = Purity ~ Hydro)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.83029 -0.73334  0.04497  0.69969  1.96809

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   74.283      1.593   46.62  < 2e-16 ***
Hydro         14.947      1.317   11.35 1.23e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.087 on 18 degrees of freedom
Multiple R-squared:  0.8774,Adjusted R-squared:  0.8706
F-statistic: 128.9 on 1 and 18 DF,  p-value: 1.227e-09
```

(a) Since we have P-value$= 1.23 \times 10^{-9} < 0.05$, we reject the null hypothesis.

Conclusion: At significant level $\alpha = 0.05$, data do provide sufficient evidence that the slope $\beta_1$ is not zero.

(b) $\widehat{\sigma} = 1.087$. (`Residual standard error`)

(c) The followings are `R` outputs:

```
> predict(fit,data.frame(Hydro=1),level=0.95,interval="confidence")
       fit      lwr      upr
1 89.23079 88.48612 89.97547
> predict(fit,data.frame(Hydro=1),level=0.95,interval="prediction")
       fit      lwr     upr
1 89.23079 86.82969 91.6319
```

When hydrocarbon level $x = 1\%$, we are 95% confident that the mean oxygen purity percentage is between 88.486 and 89.975.

When hydrocarbon level $x = 1\%$, we are 95% confident that the oxygen purity percentage for a single run of the exp will be between 86.830 and 91.632.

**HW 6.2.** The electric power consumed each month by a chemical plant is thought to be related to the average ambient temperature ($x_1$), the number of days in the month ($x_2$), the average product purity ($x_3$), and the tons of product produced ($x_4$).

**TABLE • E12-2 Power Consumption Data**

| $y$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|
| 240 | 25 | 24 | 91 | 100 |
| 236 | 31 | 21 | 90 | 95 |
| 270 | 45 | 24 | 88 | 110 |
| 274 | 60 | 25 | 87 | 88 |
| 301 | 65 | 25 | 91 | 94 |
| 316 | 72 | 26 | 94 | 99 |
| 300 | 80 | 25 | 87 | 97 |
| 296 | 84 | 25 | 86 | 96 |
| 267 | 75 | 24 | 88 | 110 |
| 276 | 60 | 25 | 91 | 105 |
| 288 | 50 | 25 | 90 | 100 |
| 261 | 38 | 23 | 89 | 98 |

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
    Min      1Q  Median      3Q     Max
-14.098  -9.778   1.767   6.798  13.016

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -123.1312   157.2561  -0.783    0.459
x1             0.7573     0.2791   2.713    0.030 *
x2             7.5188     4.0101   1.875    0.103
x3             2.4831     1.8094   1.372    0.212
x4            -0.4811     0.5552  -0.867    0.415
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.79 on 7 degrees of freedom
Multiple R-squared:  0.852,     Adjusted R-squared:  0.7675
F-statistic: 10.08 on 4 and 7 DF,  p-value: 0.00496
```

Based on the above R output, answer the following questions.

(a) Write down the fitted least squares regression model (see Page 151 of the lecture note)

(b) What is the value of the estimate of $\sigma$ (residual standard error, see Page 152)

(c) Find the value of $R^2$, and how to interpret it? (see Page 154)

(d) Test the hypothesis at $\alpha = 0.05$: (see Page 155, do not forget write down the interpretation)

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \text{ vs } H_a : \text{ at least one of them is nonzero.}$$

(e) Based on the output, test the following hypotheses at $\alpha = 0.05$ (see page 156-159, do not forget write down the interpretation)

(1) $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$.

(2) $H_0 : \beta_2 = 0$ vs $H_a : \beta_2 \neq 0$.

(3) $H_0 : \beta_3 = 0$ vs $H_a : \beta_3 \neq 0$.

(4) $H_0 : \beta_4 = 0$ vs $H_a : \beta_4 \neq 0$.

**Sol.**

(a)

$$\hat{Y} = -123.1312 + 0.7573x_1 + 7.5188x_2 + 2.4831x_3 - 0.4811x_4.$$

(b) $\widehat{\sigma} = 11.79$. (`Residual standard error`)

(c) $R^2 = 0.852$. The coefficient of determination, $R^2$, is the the proportion of the total variation in the data explained by the linear regression model.

(d) Since P-value$= 0.00496 < 0.05 = \alpha$, we do reject the null hypothesis.

Conclusion: At significant level $\alpha = 0.05$, at least one independent variable $x_1, \ldots, x_4$ is important in describing the response $Y$.

(e) (1) Since P-value$= 0.030 < 0.05 = \alpha$, we do reject the null hypothesis.

Conclusion: At significant level $\alpha = 0.05$, the data do provide sufficient evidence that average ambient temperature ($x_1$) does significantly add to a model that include number of days in the month ($x_2$), the average product purity ($x_3$), and the tons of produced.

(2) Since P-value$= 0.103 > 0.05 = \alpha$, we do not reject the null hypothesis.

Conclusion: At significant level $\alpha = 0.05$, the data do not provide sufficient evidence that $x_2$ does significantly add to a model that include $x_1$, $x_3$, and $x_4$.

(3) Since P-value$= 0.212 > 0.05 = \alpha$, we do not reject the null hypothesis.

Conclusion: At significant level $\alpha = 0.05$, the data do not provide sufficient evidence that $x_3$ does significantly add to a model that include $x_1$, $x_2$, and $x_4$.

(4) Since P-value$= 0.415 > 0.05 = \alpha$, we do reject the null hypothesis.

Conclusion: At significant level $\alpha = 0.05$, the data do not provide sufficient evidence that $x_4$ does significantly add to a model that include $x_1$, $x_2$, and $x_3$.