STAT 3008: Applied Regression Analysis
2019-20 Term 2
Assignment #1 Solutions

Problem 1:

(a) Let $g(\beta)=\sum_{i=1}^{n}(y_i-\beta x_i^2)^2$. Differentiate g wrt β, $\frac{dg}{d\beta}=-2\sum_{i=1}^{n}x_i^3(y_i-\beta x_i^2)$.

Put $\left.\frac{dg}{d\beta}\right|_{\hat\beta=0}=0\Rightarrow\hat\beta=\frac{\sum_{i=1}^{n}x_i^2y_i}{\sum_{i=1}^{n}x_i^4}$. Since $g(\beta)$ is a convex function in β, the turning point $\hat\beta$

is an absolute minimum point. Hence the least squares estimate is given by $\hat\beta=\frac{\sum_{i=1}^{n}x_i^2y_i}{\sum_{i=1}^{n}x_i^4}$

Since $RSS/\sigma^2\sim\chi^2_{n-1}\Rightarrow E(RSS/\sigma^2)=n-1\Rightarrow E(RSS/(n-1))=\sigma^2$, the OLS estimate for σ^2 is the unbiased estimator for σ^2, given by

$$\hat\sigma^2=\frac{RSS}{n-1}=\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-x_i^2\hat\beta\right)^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-x_i^2\frac{\sum_{k=1}^{n}x_k^2y_k}{\sum_{j=1}^{n}x_j^4}\right)^2$$

(b) Since $E(\hat\beta\,|\,X)=\frac{\sum_{i=1}^{n}x_i^2E(y_i)}{\sum_{i=1}^{n}x_i^4}=\frac{\sum_{i=1}^{n}x_i^2(\beta x_i^2)}{\sum_{i=1}^{n}x_i^4}=\beta$, $\hat\beta$ is an unbiased estimator for β.

(c) $(\bar x,\bar y)$ is not on the fitted regression line as $\hat\beta\bar x=\frac{\sum_{i=1}^{n}x_i^2y_i}{\sum_{i=1}^{n}x_i^4}(\bar x)^2\neq\bar y$. However, $\left(\left[\overline{x^4}\right]^{1/2},\overline{x^2y}\right)$

is as $\hat\beta\left[\left(\overline{x^4}\right)^{1/2}\right]^2=\frac{\sum_{i=1}^{n}x_i^2y_i}{\sum_{i=1}^{n}x_i^4}\overline{x^4}=\frac{1}{n}\sum_{i=1}^{n}x_i^2y_i=\overline{x^2y}$

(d) $L(\beta,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}\left(y_i-\beta x_i^2\right)^2\right]$, $l(\beta,\sigma^2)=-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\beta x_i^2\right)^2$

$\frac{\partial l(\beta,\sigma^2)}{\partial\beta}=\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i^2\left(y_i-\beta x_i^2\right)$, $\frac{\partial l(\beta,\sigma^2)}{\partial\sigma^2}=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(y_i-\beta x_i^2\right)^2$. Put

$$\left.\frac{\partial l(\beta,\sigma^2)}{\partial\beta}\right|_{(\tilde\beta,\tilde\sigma^2)}=\left.\frac{\partial l(\beta,\sigma^2)}{\partial\sigma^2}\right|_{(\tilde\beta,\tilde\sigma^2)}=0\Rightarrow\tilde\beta=\frac{\sum_{i=1}^{n}x_i^2y_i}{\sum_{i=1}^{n}x_i^4},\quad\tilde\sigma^2=\frac{RSS}{n}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-x_i^2\frac{\sum_{k=1}^{n}x_k^2y_k}{\sum_{j=1}^{n}x_j^4}\right)^2$$

(e) Based on the data, $\hat\beta=2.1176,\hat\sigma^2=0.8824$.

No, as $\sum_{i=1}^{n}\hat e_i=\sum_{i=1}^{n}\left(y_i-\hat\beta x_i^2\right)=0.8235\neq0$ (Theoretical answer also acceptable).

Note on Problem 1: You can simply view the cubic regression as simple linear regression of y on u = x^2. Therefore the result in part (a) $\hat{\beta} = SUY / SUU$ (with $\bar{u} = 0$) should be similar to the OLS estimates from the lecture notes.
For part (e), the sum of residuals is non-zero because the intercept term β_0 is missing in the regression.

Problem 2: First, note that $\overline{\hat{e}_i} = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \bar{y} - (\bar{y} - \hat{\beta}_1\bar{x}) - \hat{\beta}_1\bar{x} = 0$. Now

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})(\hat{e}_i - \overline{\hat{e}_i}) = \sum_{i=1}^{n}(x_i - \bar{x})\hat{e}_i &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - [\bar{y} - \hat{\beta}_1\bar{x}] - \hat{\beta}_1 x_i) \\ &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}_1\sum_{i=1}^{n}(x_i - \bar{x})^2 = SXY - (SXY / SXX)SXX = 0\end{aligned}$$

Therefore, $\{\hat{e}_i$, i= 1,2,…$n\}$ are uncorrelated with the explanatory variable $\{x_i$, i= 1,2,…$n\}$.

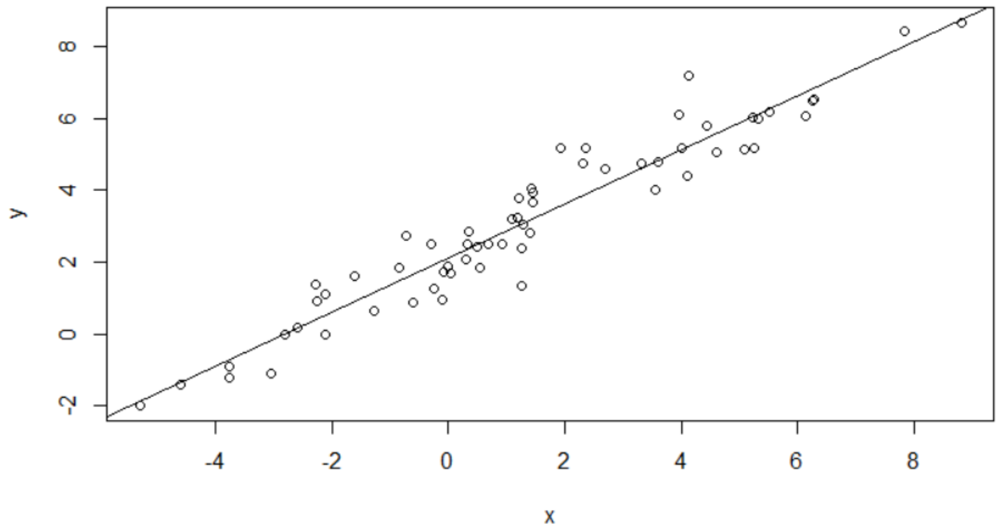
Note on Problem 2: The result is simply the consequence of the geometrical property described on Ch3 page21: $\mathbf{X}'\hat{\mathbf{e}} = 0 \Rightarrow \sum\hat{e}_i = 0$ and $\sum x_i\hat{e}_i = 0$. Hence,

$$\sum_{i=1}^{n}(x_i - \bar{x})(\hat{e}_i - \overline{\hat{e}_i}) = \sum_{i=1}^{n}(x_i - \bar{x})\hat{e}_i - 0 = \sum_{i=1}^{n}x_i\hat{e}_i - \bar{x}\sum_{i=1}^{n}\hat{e}_i = 0$$

Problem 3: (a) From the R codes below, $\hat{\beta}_1 = 0.75169, \hat{\beta}_0 = 2.13479$ and $\hat{\sigma}^2 = 0.6943^2 = 0.4820$

```
library(car); library(alr3) # Initiate the Dataset of the textbook alr3
x<-log(brains$BodyWt); y<-log(brains$BrainWt)
n<-length(x); n # Obtain the number of data points n
fit<-lm(y~x) # fit is an object of regression y by x
summary(fit) # OLS estimates and Test for OLS estimates
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.13479    0.09604   22.23   <2e-16 ***
x            0.75169    0.02846   26.41   <2e-16 ***
Residual standard error: 0.6943 on 60 degrees of freedom
Multiple R-squared:  0.9208,   Adjusted R-squared:  0.9195
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
c(fit$coef, summary(fit)$sigma^2) # Display (beta0.hat, beta1.hat, sigma.hat^2)
  (Intercept)        x
  2.1347883    0.7516861    0.4820435
```

(b) plot(x,y); abline(fit)



(c) The R codes below compute the ratios $\hat{e}_i / \hat{\sigma}$, i=1, 2, … 33, suggesting that observations 31 (ratio 2.81), observation 34 (ratio = -2.47) and observation 35 (ratio = 2.32) are outliers.

```
----------------------------------------
round(fit$residuals/summary(fit)$sigma,2)    # round up the values in 2 decimal places
    1     2     3     4     5     6     7     8     9    10    11    12
 1.07  1.67 -0.39 -1.01 -0.07  0.16  0.61 -0.66  1.22  0.53  1.40 -0.48
   13    14    15    16    17    18    19    20    21    22    23    24
-0.36 -0.17 -0.03 -1.00 -0.21 -1.00  0.58 -0.72 -0.04 -0.51 -1.01  1.27
   25    26    27    28    29    30    31    32    33    34    35    36
 0.30  0.99  0.68 -0.47 -0.20  0.45  0.85  2.81 -0.15 -2.47  2.32 -1.13
   37    38    39    40    41    42    43    44    45    46    47    48
-0.51 -0.78 -0.31 -0.09  0.20 -0.13 -0.48  0.02 -0.78  1.41  1.85 -1.17
   49    50    51    52    53    54    55    56    57    58    59    60
-1.18 -0.08  0.63 -0.77 -0.01  0.79 -1.38 -1.29  0.37 -1.18 -1.58 -0.09
   61    62
 0.70  1.01
### Sum of Residuals ####
sum(fit$residuals)
[1] -1.543904e-16
```

Problem 4:

(a) $SXY = \sum_{i=1}^{n} x_i y_i - n\bar{x}(\bar{y}) = 3373.75 - 11(73.14545)(3.954545) = 191.9227$

$SXX = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 60961.94 - 11(73.14545)^2 = 2109.107$

$SYY = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 202.25 - 11(3.954545)^2 = 30.22727$

(b) $\hat{\beta}_1 = SXY / SXX = 0.09100$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3.954545 - 0.09100(73.14545) = -2.70148$

$RSS = SYY - SXY^2 / SXX = 12.76285, \hat{\sigma}^2 = RSS/(n-2) = 1.418095$

(c) $\hat{V}ar(\hat{\beta}_0 \mid X) = \hat{\sigma}^2(1/n + \bar{x}^2 / SXX) = 3.726, \quad \hat{V}ar(\hat{\beta}_1 \mid X) = \hat{\sigma}^2 / SXX = 0.0006724$

(d) Consider $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -2.70148 + 0.09100(74.2) = 4.0505$.

$\hat{e} = y - \hat{y} = 2 - 4.0505 = -2.0505 = -1.722\hat{\sigma}$. Hence (74.2,2.000) is NOT an outlier.

(e) $\bar{x} = (11(73.14545) + 50.3)/12 = 71.24167$, $\bar{y} = (11(3.95455) + 3)/12 = 3.875$

$$SXY = \sum_{i=1}^{m} x_i y_i - m\bar{x}(\bar{y}) = (3373.75 + 50.3(3)) - 12(71.24167)(3.875) = 211.9125$$
$$SXX = \sum_{i=1}^{m} x_i^2 - m\bar{x}^2 = (60961.94 + 3^2) - 12(3)^2 = 2587.529$$
$$\hat{\beta}_1^* = SXY / SXX = 0.08190$$

(f) $\hat{\beta}_0^* = \bar{y} - \hat{\beta}_0^* \bar{x} = 3.875 - 0.08190(71.24167) = -1.9595$

$$SYY = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = (202.25 + 3^2) - 12(3.875)^2 = 31.0625$$
$$RSS = SYY - SXY^2 / SXX = 13.707, \hat{\sigma}^2 = RSS/(n-2) = 1.3707$$