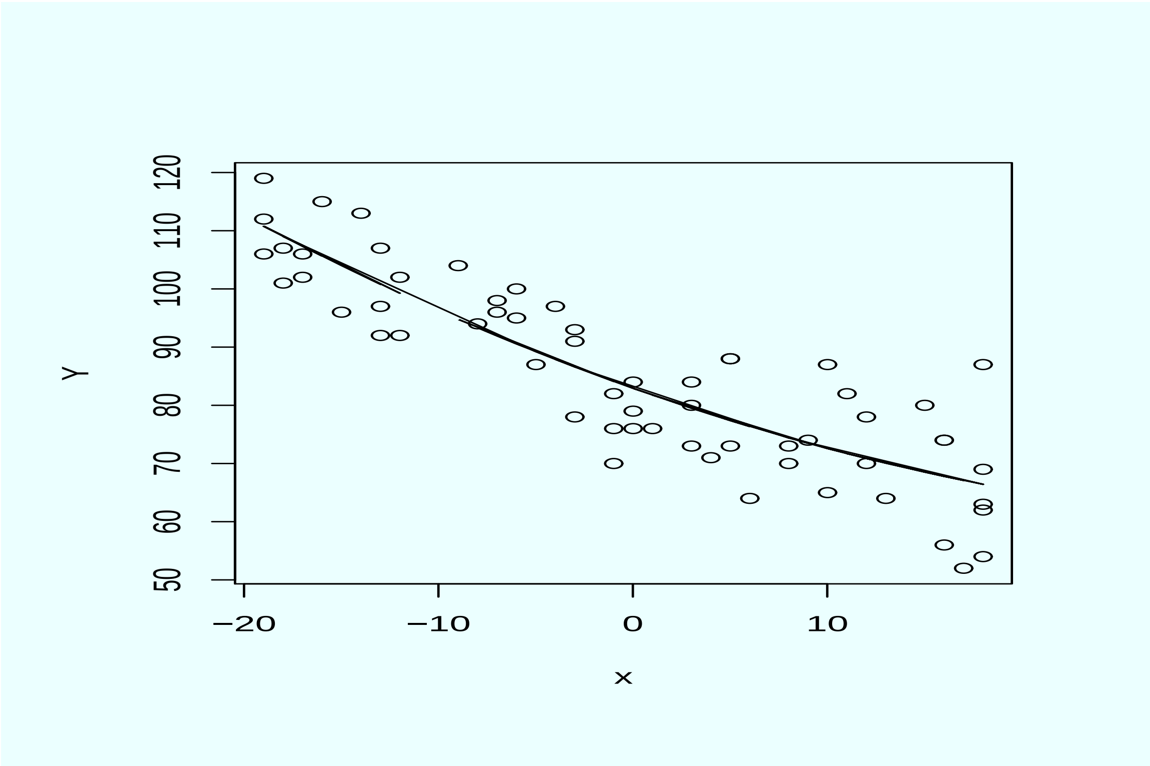
MA 542 REGRESSION ANALYSIS
SPRING 2018
HW - 8 - Solution Key

1. (Chapter 8 question 4)
a) The fitted Regression model is :

$$\hat{Y} = 82.93571.18396x + .0148405x^2,$$

where x is the centralized data of aged (X).



Yes the fitted regression model appears to be a god fit. $R^2 = 0.76317$ is also confirm that.

b)

$$H_0 : \beta_1 = \beta_{11} = 0 \quad \text{vs} \quad H_1 : not \quad H_0$$

$$F^* = \frac{MSR}{MSE} = \frac{5915.31}{64.409} = 91.8398$$

But,

$$F_{0.95,2,57} = 3.15884 \quad \text{OR} \quad P - value = 2.2 \times 10^{-16}$$

Since $F^* = 91.84 > F_{0.99,2,57} = 3.16$ or $P - value = 2.2 \times 10^{-16} < 0.05$, H_0 is rejected. So there is a regression relation between the age and the muscle mass.

c)

```
> Xh=48
> xh=Xh-mean(X)
> xh_2=xh^2
> newpx=data.frame(x=xh,x_2=xh_2)
> predict.lm(QRM,newpx,interval="confidence",level=0.95)
       fit      lwr      upr
1 99.25461 96.28436 102.2249
```

So the 95% confidence interval for the mean muscle mass for women aged 48 years is (96.28436, 102.2249). So we have 95% confidence that the **mean muscle mass** for women aged 48 years is in between 96.28436 and 102.2249.

d)

```
> predict.lm(QRM,newpx,interval="prediction",level=0.95)
       fit     lwr      upr
1 99.25461 82.9116 115.5976
```

So the 95% prediction interval for a women whose age is 48 years, is (82.9116, 115.5976). So we have 95% confidence that a women whose age is 48 years has muscle mass from 82.9116 to 115.5976.

e) i) Using the t-test. The following is a part of the summary output.

```
Call:
lm(formula = Y ~ x + x_2)

Residuals:
    Min      1Q  Median      3Q     Max
-15.086  -6.154  -1.088   6.220  20.578

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 82.935749   1.543146  53.745   <2e-16 ***
x           -1.183958   0.088633 -13.358   <2e-16 ***
x_2          0.014840   0.008357   1.776   0.0811 .
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Residual standard error: 8.026 on 57 degrees of freedom
Multiple R-squared:  0.7632,    Adjusted R-squared:  0.7549
F-statistic: 91.84 on 2 and 57 DF,  p-value: < 2.2e-16
```

$$H_0 : \beta_{11} = 0 \quad \text{vs} \quad H_1 : \beta_{11} \neq 0$$

$$T^* = 1.776$$

But,

$$t_{0.975,57} = 2.00 \quad \text{OR} \quad P - value = 0.0811$$

```
Call:
lm(formula = Y ~ x + x_2)

Residuals:
    Min      1Q  Median      3Q     Max
-15.086  -6.154  -1.088   6.220  20.578

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 82.935749   1.543146  53.745   <2e-16 ***
x           -1.183958   0.088633 -13.358   <2e-16 ***
x_2          0.014840   0.008357   1.776   0.0811 .
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Residual standard error: 8.026 on 57 degrees of freedom
Multiple R-squared:  0.7632,    Adjusted R-squared:  0.7549
F-statistic: 91.84 on 2 and 57 DF,  p-value: < 2.2e-16
```

Since $T^* = 1.776 < t_{0.975,57} = 2.00$ or $P - value = 0.0811 > 0.05$, H_0 is not rejected. So the quadratic term can be dropped from the model under the significance level $\alpha = 0.5$

ii) Using extra sum of squares (F-test)

$$H_0 : \beta_{11} = 0 \quad \text{vs} \quad H_1 : \beta_{11} \neq 0$$

$$F^* = 3.1538$$

But,

$$F_{0.95,1,57} = 4.01 \quad \text{OR} \quad P - value = 0.0811$$

Since $F^* = 3.1538 < F_{0.95,1,57} = 4.01$ or $P - value = 0.0811 > 0.05$, H_0 is not rejected. So the quadratic term can be dropped from the model under the significance level $\alpha = 0.5$

f)

$$\begin{aligned}
\hat{Y} &= 82.9357 - 1.18396x + .0148405x^2 \\
\hat{Y} &= 82.9357 - 1.18396(X - 59.98833) + .0148405(X - 59.98833)^2 \\
\hat{Y} &= 207.3478 - 2.964263X + 0.01484X^2
\end{aligned}$$

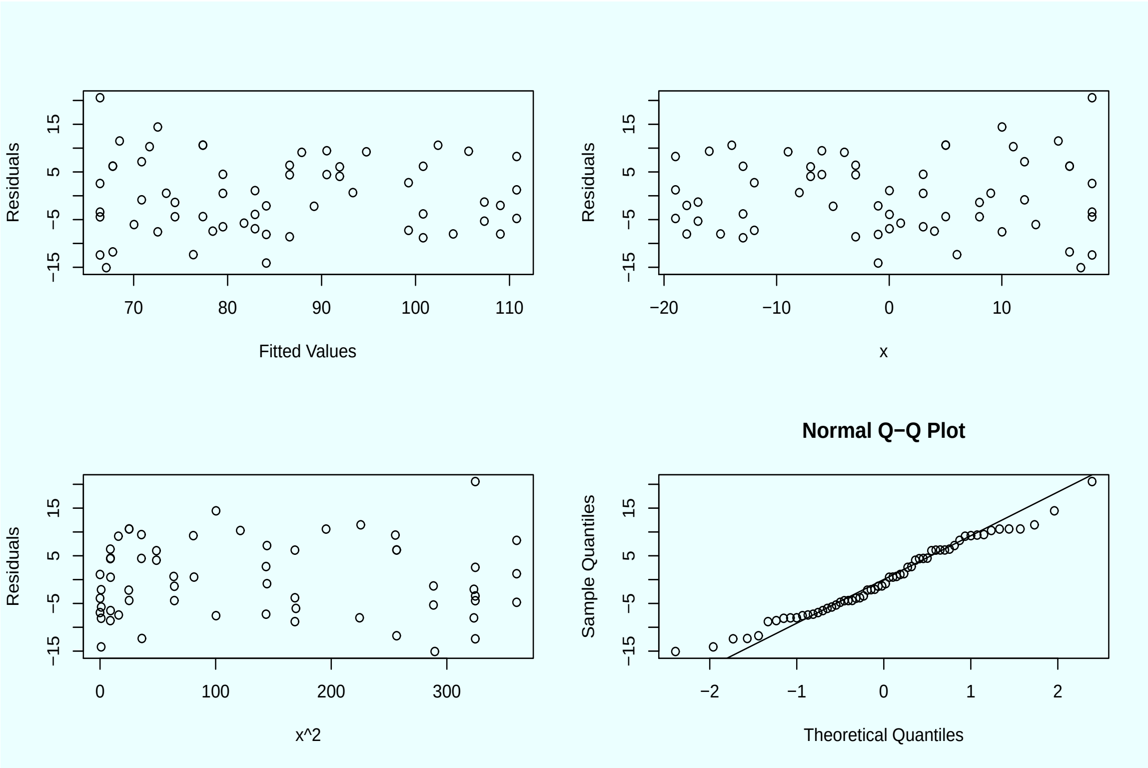
g)

```
> cor(X,X^2)
[1] 0.9960939
> cor(x,x_2)
[1] -0.03835694
```

So the Correlation between X and X^2 is 0.9960939 and correlation between x and x^2 is -0.03835694. So the centered variable is helpful here, it reduces the correlation between the variables.

2. (Chapter 8 question 5)

a) The graphs of residuals against the fitted values and against the X and the normal probability plots are:



Residual plots show the appropriateness of the model (i.e., there is no a considerable violation of the model assumptions.) Since the normal probability plot is approximately linear, there is no violation in the normality assumption too.

b)

```
> Means=aov(Y~factor(x)*factor(x_2))
> anova(QRM,Means)
Analysis of Variance Table

Model 1: Y ~ x + x_2
Model 2: Y ~ factor(x) * factor(x_2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     57 3671.3
2     28 1849.7 29    1821.7 0.9509 0.5539
```

$$H_0 \ : \ E[Y] = \beta_0 + \beta_1 x + \beta_{11} x^2 \qquad H_1 \ : \ not \ \ H_0$$

$$F^* = \frac{MSPE}{SSLF} = 0.9509$$

ut,

$$F_{0.95,29,28} = 1.875 \quad \text{OR} \quad P - value = 0.5539$$

Since $F^* = 0.9509 < F_{0.95,29,28} = 1.875$ or $P - value = 0.5539 > 0.05$, H_0 is not rejected. So the regression model is a good fit for the data.

c) i) t-test.

$$H_0 : \beta_{111} = 0 \quad \text{vs} \quad H_1 : \beta_{111} \neq 0$$

```
> x_3=x^3
> CRM=lm(Y~x+x_2+x_3)
> summary(CRM)

Call:
lm(formula = Y ~ x + x_2 + x_3)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3671  -5.8483  -0.6755   6.1376  20.0637

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 82.9273444  1.5552264  53.322  < 2e-16 ***
x           -1.2678894  0.2489231  -5.093 4.28e-06 ***
x_2          0.0150390  0.0084390   1.782   0.0802 .
x_3          0.0003369  0.0009327   0.361   0.7193
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1        1
```

$$T^* = 0.361$$

But,

$$t_{0.975,57} = 2.003 \quad \text{OR} \quad P - value = 0.7193$$

Since $T^* = 0.361 < t_{0.975,57} = 2.003$ or $P - value = 0.7193 > 0.05$, H_0 is not rejected. So the cubic term can be dropped from the model under the significance level $\alpha = 0.5$

ii) Using extra sum of squares (F-test)

```
> anova(CRM)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq  F value  Pr(>F)
x          1 11627.5 11627.5 177.7720 < 2e-16 ***
x_2        1   203.1   203.1   3.1057 0.08348 .
x_3        1     8.5     8.5   0.1305 0.71928
Residuals 56  3662.8    65.4
```

$$H_0 : \beta_{111} = 0 \quad \text{vs} \quad H_1 : \beta_{111} \neq 0$$

$$F^* = 0.1305$$

But,

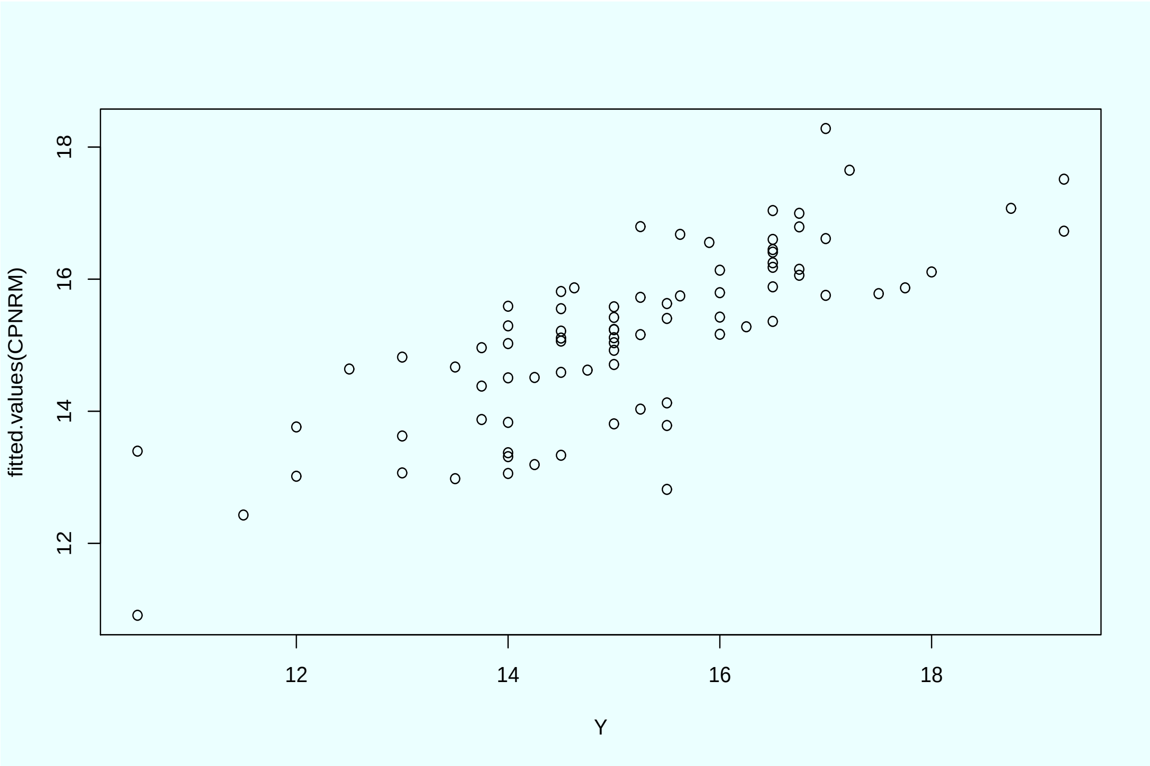
$$F_{0.95,1,57} = 4.012 \quad \text{OR} \quad P - value = 0.71928$$

Since $F^* = 0.1305 < F_{0.95,1,57} = 4.012$ or $P - value = 0.71928 > 0.05$, H_0 is not rejected. So the cubic term can be dropped from the model under the significance level $\alpha = 0.5$.

3. (Chapter 8 question 8)
a)

The fitted Regression Model is:

$$\hat{Y} = 10.19 - 0.1818x_1 + 0.01415x_1^2 + 0.3140X_2 + 0.000008X_4$$



The plot is nearly linear so the model is a good fit.

b)

The adjusted R^2 value is 0.5926885. This value gives proportion of variation explained by the regression model considering the number of predictors in the model.

c)

```
> anova(lm(Y~x1+X2+X4+x1_2))
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x1         1  14.819  14.819 12.3036 0.0007627 ***
X2         1  72.802  72.802 60.4463 2.968e-11 ***
X4         1  50.287  50.287 41.7522 8.907e-09 ***
x1_2       1   7.115   7.115  5.9078 0.0174321 *
Residuals 76  91.535   1.204
```

$$H_0 : \beta_{11} = 0 \quad \text{vs} \quad H_1 : \beta_{11} \neq 0$$

$$F^* = 5.9078$$

But,

$$F_{0.95,1,76} = 3.96676 \quad \text{OR} \quad P - value = 0.0174321$$

Since $F^* = 5.9078 > F_{0.95,1,76} = 3.96676$ or $P - value = 0.0174321 < 0.05$, H_0 is rejected. So x_1^2 term should be in the model under the significance level $\alpha = 0.5$.

d)

```
> x1h=8-mean(X1)
> x1_2h=x1h^2
> X2h=16
> X4h=250000
> newpx=data.frame(x1=x1h,x1_2=x1_2h,X2=X2h,X4=X4h)
> predict.lm(CPNRM,newpx,interval="confidence",level=0.95)
       fit     lwr      upr
1 17.20089 16.4571 17.94468
```

So the 95% confidence interval is (16.4571, 17.94468). So we have 95% confidence to say that **mean rental rate** when the age 8 years, operating expenses and the taxes 16, and the total square footage 25000 is in between 16.4571 and 17.94468.

d)

$$\begin{aligned}
\hat{Y} &= 10.19 - 0.1818x_1 + 0.01415x_1^2 + 0.3140X_2 + 0.000008X_4 \\
\hat{Y} &= 10.19 - 0.1818(X_1 - 7.864198) + 0.01415(X_1 - 7.864198)^2 + 0.3140x_2 + 0.000008X_4 \\
\hat{Y} &= 12.4938 - 2.407368X_1 + 0.01415X_1^2 + 0.3140X_2 + 0.000008X_4
\end{aligned}$$

4. (Chapter 8 question 16)
a)
The Regression Model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

Interpretations:
β_0 : Mean GPA when the entrance test score is 0 and the the major field is not decided.
β_1 : Change in mean GPA, when the entrance test score is increased by one unit.
β_2 : Difference in mean GPA for "major field is decided" and "the major field is not decided"'

b)

```
> RM=lm(Y~X1+X2,data=Data)
> RM

Call:
lm(formula = Y ~ X1 + X2, data = Data)

Coefficients:
(Intercept)           X1           X2
    2.19842      0.03789     -0.09430
```

So the estimated regression d=function is:

$$\hat{Y} = 2.19842 + 0.03789X_1 - 0.09430X_2$$

c)

$$H_0 : \beta_2 = 0 \quad \text{vs} \quad H_1 : \beta_2 \neq 0$$

```
> anova(RM)
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value   Pr(>F)
X1          1  3.588  3.5878  9.2103 0.002966 **
X2          1  0.241  0.2407  0.6179 0.433406
Residuals 117 45.577  0.3895
```

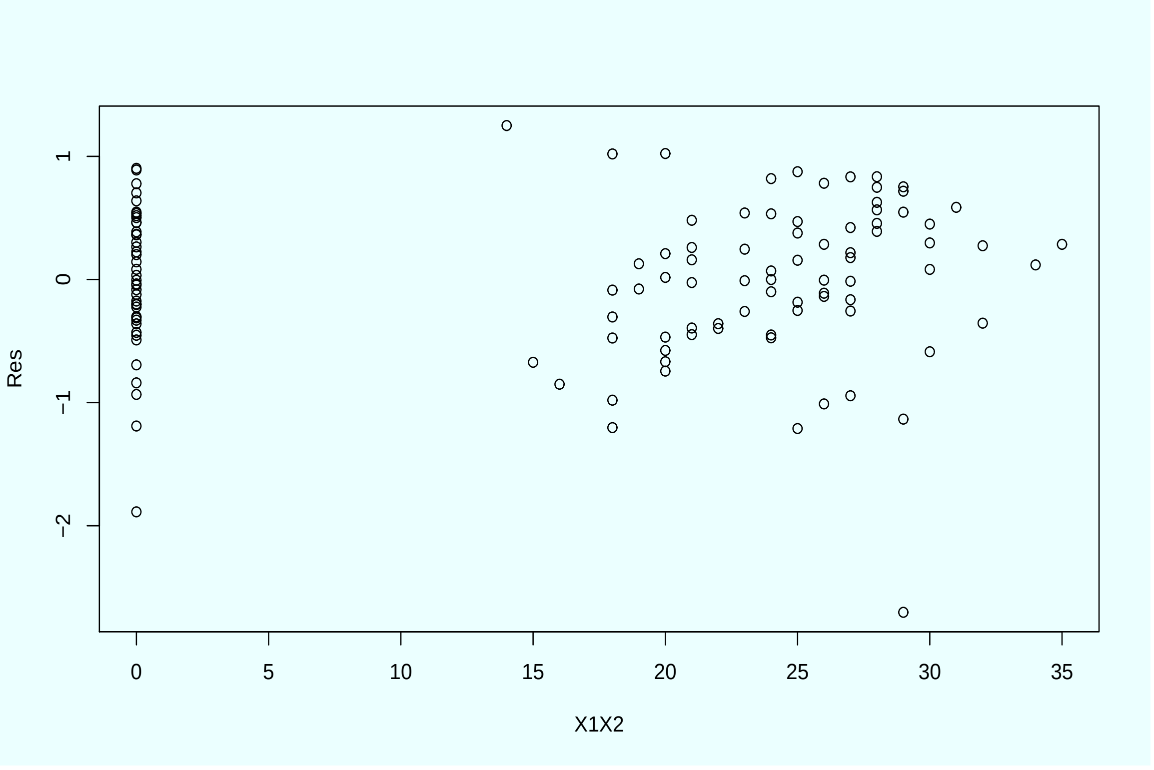
$$F^* = 0.6179$$

But,

$$F_{0.99,1,117} = 6.856564 \quad \text{OR} \quad P - value = 0.433406$$

Since $F^* = 0.6179 < F_{0.99,1,117} = 6.856564$ or $P - value = 0.433406 > 0.05$, H_0 is not rejected. So X_2 term can be dropped from the model under the significance level $\alpha = 0.01$.

d)



Since there is no any specific pattern, in the residual plot, the interaction term is not needed in the model.

5. (Chapter 8 question 29)

a)

So for data -1 :

$Cor(X, X^2) = 0.9902871$ and $cor(x, x^2) = 0.3791661$
$Cor(X, X^3) = 0.9659484$ and $cor(x, x^3) = 0.904355$

So for data -2 :

$Cor(X, X^2) = 0.9699782$ and $cor(x, x^2) = 0.8463526$

```
> cor(X1,X1^2)
[1] 0.9902871
> cor(x1,x1^2)
[1] 0.3791661
>
> cor(X2,X2^2)
[1] 0.9699782
> cor(x2,x2^2)
[1] 0.8463526
>
>
> cor(X1,X1^3)
[1] 0.9659484
> cor(x1,x1^3)
[1] 0.904355
>
>
> cor(X2,X2^3)
[1] 0.9290059
> cor(x2,x2^3)
[1] 0.8955835
```

$Cor(X, X^3) = 0.9290059$ and $cor(x, x^3) = 0.8955835$

Here data-2 has a higher variance while the first data set has comparatively lower variance. So the centralized data has higher covariance between x and x^2 when the variance of the data is high. Further centralize data has lower covariance between x and the lower degree terms than x with higher degree terms.