Statistical learning assignment 4- chapter 3

孫浩哲　　M072040002

October 11, 2018

4.

(a)

The residual sum of square(RSS) for the polynomial regression is lower than another one, because of the polynomial regression may fit the data better.

(b)

Contrary to (a), polynomial regression have a higher test RSS because the polynomial model may overfit the data.

(c)

The RSS for Polynomial regression is lower than another one.The more flexible model, The lower RSS.

(d)

No enough information, because the true relationship is non-linear, We have to see which model is closer than another one real model.

5.

$$\begin{aligned}
\hat{y}_i = x_i\hat{\beta} = x_i\frac{(\sum_{i'=1}^{n} x_{i'}y_{i'})}{(\sum_{i'=1}^{n} x_{i'}^2)} \\
= \frac{(\sum_{i'=1}^{n} x_i x_{i'}y_{i'})}{(\sum_{i'=1}^{n} x_{i'}^2)} \\
= \left(\frac{\sum_{i'=1}^{n} x_i x_{i'}}{\sum_{i'=1}^{n} x_{i'}^2}\right) y_{i'}
\end{aligned}$$

$\therefore a_{i'} = \left(\frac{\sum_{i'=1}^{n} x_i x_{i'}}{\sum_{i'=1}^{n} x_{i'}^2}\right)$

6.

The simple linear regression ：

$$y = \beta_0 + \beta_1 x$$

We know that $\beta_0 = \bar{y} - \beta_1\bar{x}$, so we can write that：

$$\begin{aligned}
y = \bar{y} - \beta_1\bar{x} + \beta_1 x \\
\Rightarrow (y - \bar{y}) = \beta_1(x - \bar{x})
\end{aligned}$$

No matter what the β_1 is, the simple linear model passes through $(\bar{x}, \bar{y})$

11.

(a)

```
> set.seed(1)
> x=rnorm(100)
> y=2*x+rnorm(100)
> model=lm(y~x)
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-1.8768 -0.6138 -0.1395  0.5394  2.3462

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03769    0.09699  -0.389    0.698
x            1.99894    0.10773  18.556   <2e-16 ***
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.
’  0.1  ‘ ’  1

Residual standard error: 0.9628 on 98 degrees of freedom
Multiple R-squared:  0.7784,Adjusted R-squared:  0.7762
F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

The p-value of β is extremely small and t-statistics is large, so we reject the null hypothesis.

(b)

```
> model2=lm(x~y)
> summary(model2)

Call:
lm(formula = x ~ y)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90848 -0.28101  0.06274  0.24570  0.85736

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03880    0.04266    0.91    0.365
y            0.38942    0.02099   18.56   <2e-16 ***
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.
’  0.1  ‘ ’  1

Residual standard error: 0.4249 on 98 degrees of freedom
```

```
Multiple R-squared:  0.7784,Adjusted R-squared:  0.7762
F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

The p-value is extremely small and t-statistics is large, even larger than the model which has intercept, so we reject the null hypothesis.

(c)

They seem to be the inverse function mutually.

(d)

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$$\begin{aligned}
t = \frac{\hat{\beta}}{SE(\hat{\beta})} &= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \times \sqrt{\frac{(n-1)\sum_{i'=1}^{n} x_{i'}^2}{\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}} \\
&= \frac{\sqrt{(n-1)\sum_{i=1}^{n} x_i^2}\sum_{i=1}^{n} x_i y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2)^2 \sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}} \\
&= \frac{\sqrt{(n-1)}\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n}(y_i^2 - 2x_i y_i\hat{\beta} + (x_i\hat{\beta})^2)}} \\
&= \frac{\sqrt{(n-1)}\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} x_i^2\hat{\beta}(\sum_{i=1}^{n} 2x_i y_i - x_i^2\hat{\beta})}} \\
&= \frac{\sqrt{(n-1)}\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} x_i y_i)^2}}
\end{aligned}$$

(e)

```
> sqrt(99)*sum(x*y)/sqrt(sum(x^2)*sum(y^2)-(sum(x*y))^2)
[1] 18.72593
> sqrt(99)*sum(y*x)/sqrt(sum(y^2)*sum(x^2)-(sum(y*x))^2)
[1] 18.72593
```

(f)

As the program shown in (a) & (b), we can find that the t-statistic of two models are almost same.

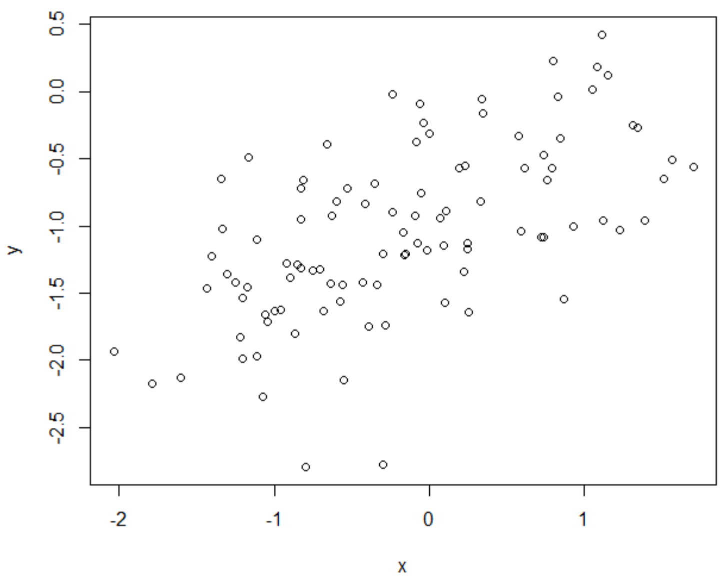
13.

(a)(b)(c)

```
> {
+ x=rnorm(100)
+ eps=rnorm(100,0,0.5)
+ y=-1+0.5*x+eps
+ length(y)
+ }
[1] 100
```

$\beta_0 = -1, \beta_1 = -0.5$

(d)

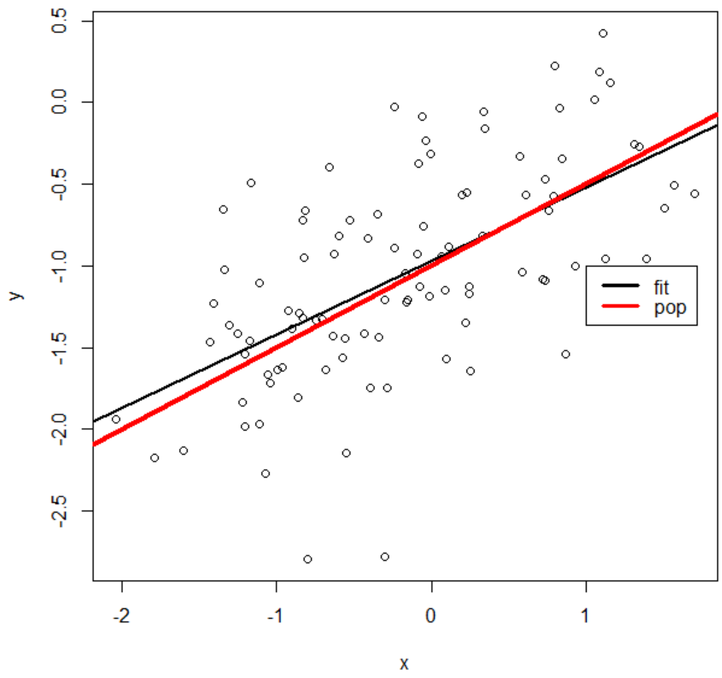


(e)

```
> {
+ fit=lm(y~x)
+ fit
+ summary(fit)
+ }

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.67217 -0.33359 -0.00624  0.36744  1.05246

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.97000    0.05304 -18.289  < 2e-16 ***
x            0.44834    0.06055   7.405 4.62e-11 ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.
'  0.1  ' '  1

Residual standard error: 0.5199 on 98 degrees of freedom
Multiple R-squared:  0.3588,Adjusted R-squared:  0.3522
F-statistic: 54.83 on 1 and 98 DF,  p-value: 4.62e-11
```

They are close to the parameter we construct.The p-value is extremely small, so the null hypothesis can be rejected.

(f)



(g)

```
z=x^2
pol=lm(y~x+z)
> summary(pol)

Call:
lm(formula = y ~ x + z)

Residuals:
     Min       1Q   Median       3Q      Max
-1.69645 -0.32512 -0.01374  0.37001  1.02730

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.94408    0.07213 -13.088  < 2e-16 ***
```

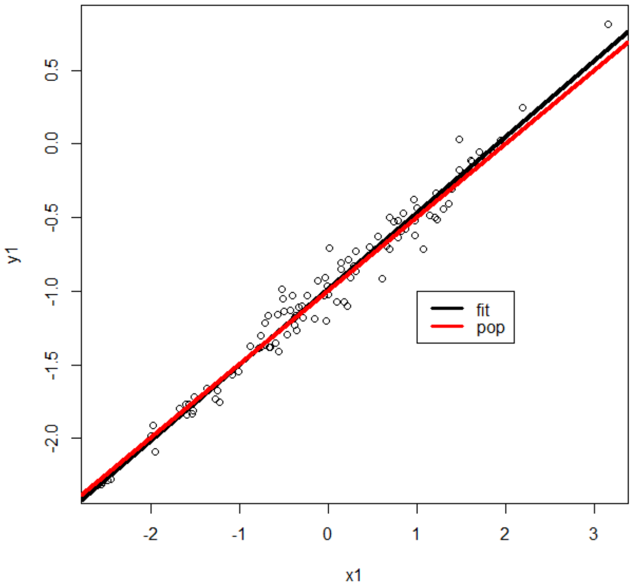
```
x            0.44327    0.06151   7.206 1.25e-10 ***
z           -0.03493    0.06559  -0.533    0.596
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.
'  0.1  ' '  1

Residual standard error: 0.5218 on 97 degrees of freedom
Multiple R-squared:  0.3606,Adjusted R-squared:  0.3474
F-statistic: 27.35 on 2 and 97 DF,  p-value: 3.796e-10
```

The multiple R^2 of this model is slightly larger than simple linear regression, and the p-value is also small.

(h)

```
x1=rnorm(100)
eps1=rnorm(100,0,0.1)
y1=-1+0.5*x1+eps1
fit1=lm(y1~x1)
> summary(fit)

Call:
lm(formula = y1 ~ x1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.283441 -0.052967 -0.001256  0.064842  0.271827

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.98516    0.01051  -93.72   <2e-16 ***
x1           0.51626    0.00894   57.75   <2e-16 ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.
'  0.1  ' '  1

Residual standard error: 0.1048 on 98 degrees of freedom
Multiple R-squared:  0.9715,Adjusted R-squared:  0.9712
```

```
F-statistic:  3335 on 1 and 98 DF,  p-value: < 2.2e-16
```

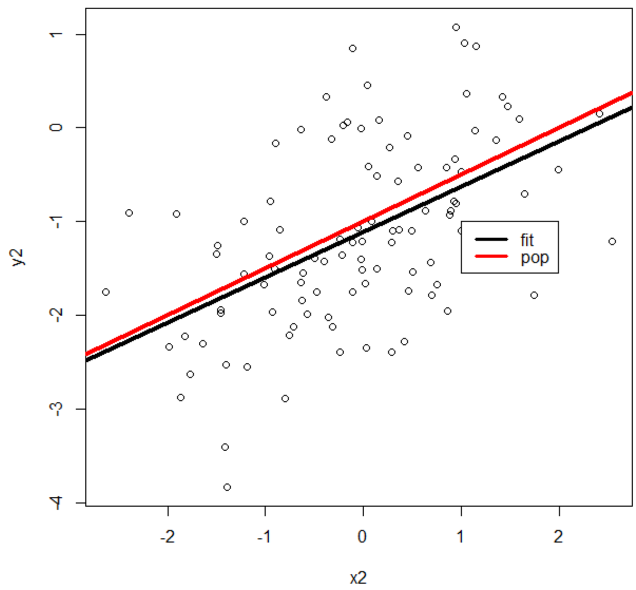


The R^2 is much larger than the model we build before.

(i)

```
> {
+ x2=rnorm(100)
+ eps2=rnorm(100,0,0.8)
+ y2=-1+0.5*x2+eps2
+ fit2=lm(y2~x2)
+ summary(fit)
+ }

Call:
lm(formula = y2 ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.95802 -0.41207 -0.03688  0.60361  1.76914
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.02398    0.07849 -13.045  < 2e-16 ***
x2           0.38069    0.07163   5.315 6.71e-07 ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.
'  0.1  ' '  1

Residual standard error: 0.7839 on 98 degrees of freedom
Multiple R-squared:  0.2237,Adjusted R-squared:  0.2158
F-statistic: 28.25 on 1 and 98 DF,  p-value: 6.707e-07
```



The R^2 is smaller than the original model.

(j)

```
> confint(fit)
                 2.5 %     97.5 %
(Intercept) -1.0332352 -0.9710735
```

```
x             0.4677338  0.5275863
> confint(fit1)
                 2.5 %     97.5 %
(Intercept) -1.0053096 -0.9932385
x1           0.4880421  0.5001488
> confint(fit2)
                  2.5 %     97.5 %
(Intercept) -1.05712151 -0.9677758
x2          -0.03754376  0.0575819
```

The more variance we set, the wider confidence interval we get, vice versa.