# Untitled

12111603

2023-10-02

## ISLR 3.14

## (a)

```
set.seed(1)
x1=runif (100)
x2=0.5*x1+rnorm (100)/10
y=2+2*x1+0.3*x2+rnorm (100)
```

The linear model above is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, where $\beta_0 = 2, \beta_1 = 2, \beta_2 = 0.3$.

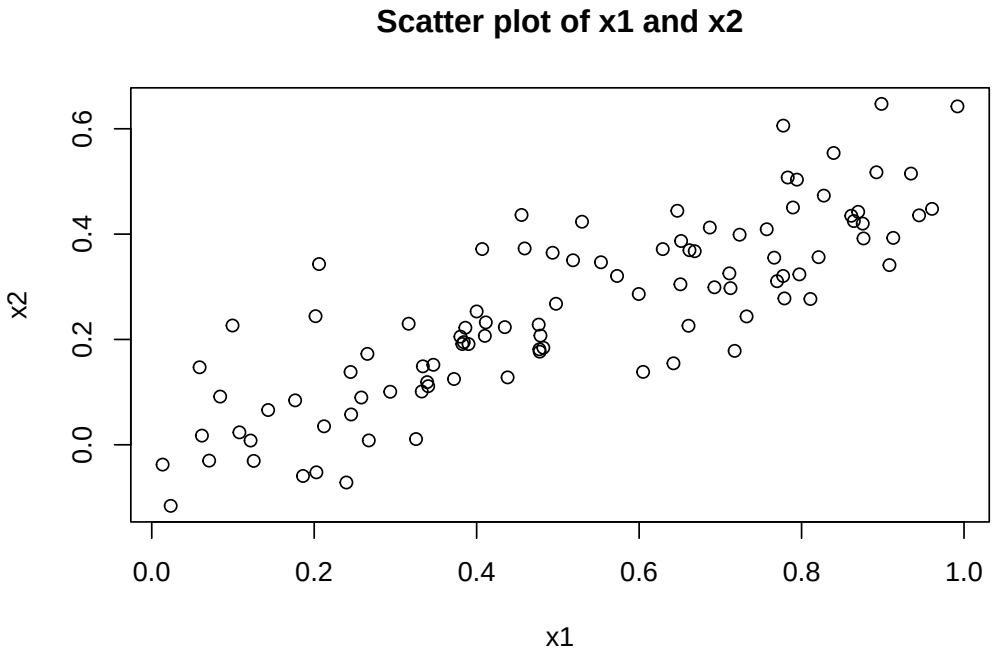
## (b)

```
correlation <- cor(x1,x2)
correlation
```

```
## [1] 0.8351212
```

The correlation is 0.8351212.

```
plot(x1,x2,xlab = "x1",ylab = "x2",main = "Scatter plot of x1 and x2")
```




## (c)

```
model <- lm(y~x1+x2)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

$\hat{\beta_0} = 2.13, \hat{\beta_1} = 1.44, \hat{\beta_2} = 1.01$. The estimated result is close to the true value, however, there is still some distance.

As for $\beta_1$, at 0.05 significance level, we figure out p-value = 0.0487 $<$ 0.05, so we SHOULD reject null hypothesis $\beta_1 = 0$.

As for $\beta_2$, at 0.05 significance level, we figure out p-value = 0.3754 $>$ 0.05, so we CANNOT reject null hypothesis $\beta_2 = 0$.

# (d)

```
model_alter <- lm(y~x1)
summary(model_alter)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

$\hat{\beta_0} = 2.11, \hat{\beta_1} = 1.98$. The estimated result is better than the previous one, which means the distance between $(\hat{\beta_0}, \hat{\beta_1})$ and $(\beta_0, \beta_1)$ is quite small.

As for $\beta_1$, at 0.001 significance level, we figure out p-value = 2.66e-06 $<$ 0.001, so we SHOULD reject null hypothesis $H_0 : \beta_1 = 0$.

# (e)

```
model_alter <- lm(y~x2)
summary(model_alter)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62687 -0.75156 -0.03598  0.72383  2.44890
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
## x2            2.8996     0.6330    4.58 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.072 on 98 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1679
## F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

$\hat{\beta_0} = 2.39, \hat{\beta_2} = 2.90$. The estimated result is NOT close to the true value, thus the accuracy is relatively bad.

As for $\beta_2$, at 0.001 significance level, we figure out p-value = 1.37e-05 $<$ 0.001, so we SHOULD reject null hypothesis $H_0 : \beta_2 = 0$. Furthermore, F-statistics is 20.98 less than that in (d), which implies some information of $x_2$ has been covered by $x_1$.

# (f)

Not contradict.

In (c), because $\beta_2 = 0.3$ is smaller than$\beta_1 = 2$ on earth, so after estimation $\beta_2$ tends to be believed is zero by p-value test. Alternatively, when both $x_1$ and $x_2$ exist, $x_2$ is less significant.

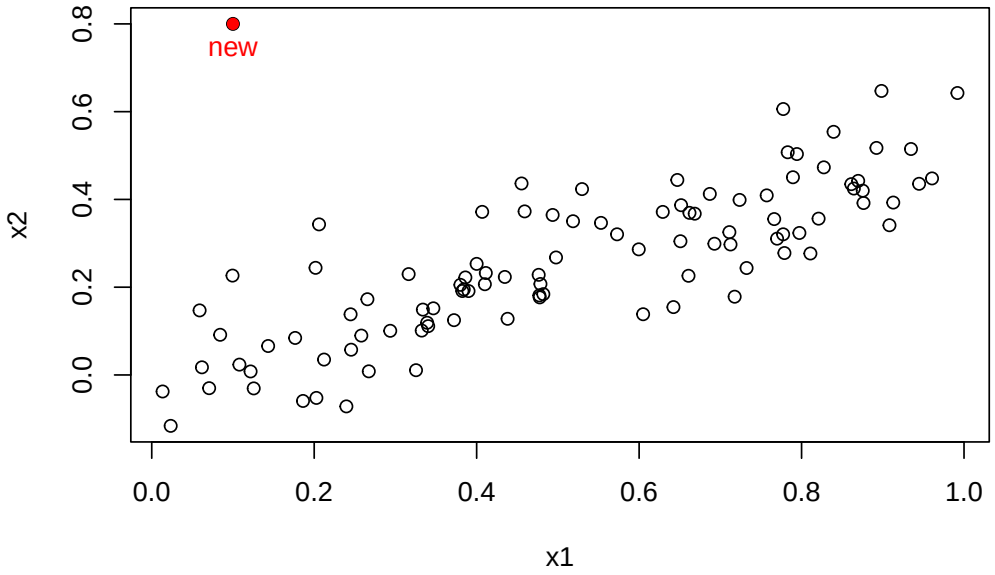
In (d), since there is only one label $x_1$, noticed that $\beta_1 = 0.2 > \beta_2 = 0.3$ which means $x_1$ makes more contributions to $y$, thus in the p-value test, $\beta_1$ shall NOT be zero. Furthermore, since not only $y$, but $x_2$ is also yielded by $x_1$, so the estimated result with only one label $x_1$ is highly consistent with true value.

In (e), because $x_1$ and $x_2$ are correlated, which is so called "**collinearity**", then some information of $x_1$ can be covered by $x_2$. Therefore, $\beta_2 = 0$ should be rejected by p-value test.

# (g)

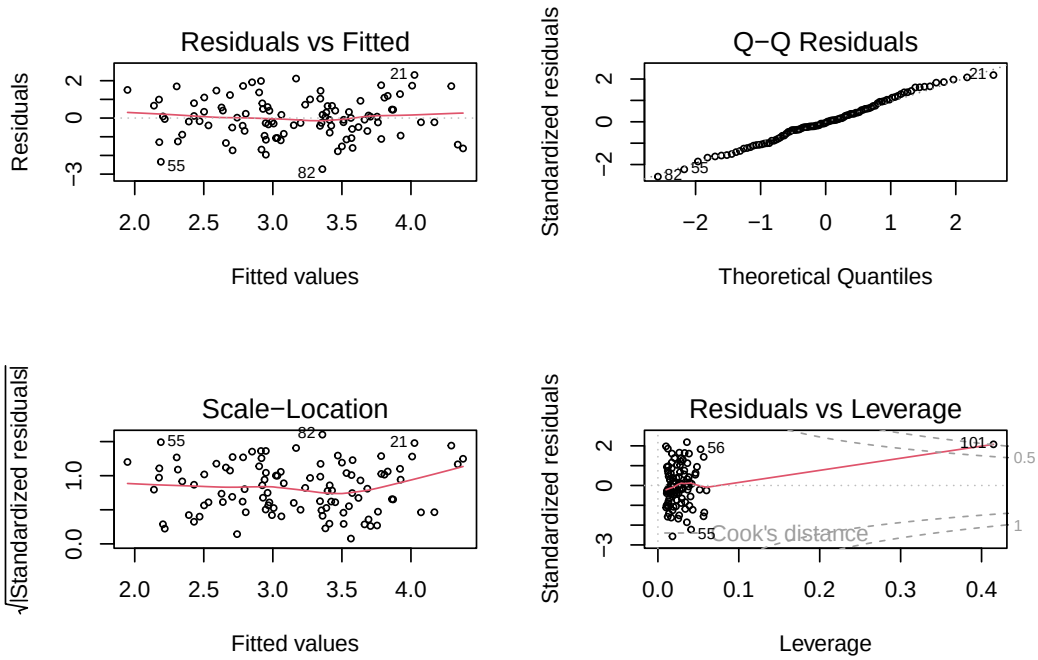
```r
x1=c(x1, 0.1)
x2=c(x2, 0.8)
y=c(y,6)

plot(x1,x2)
points(x1[101], x2[101], pch = 16, col = "red")
text(x=0.1, y=0.8, labels="new",pos=1,col="red")
```



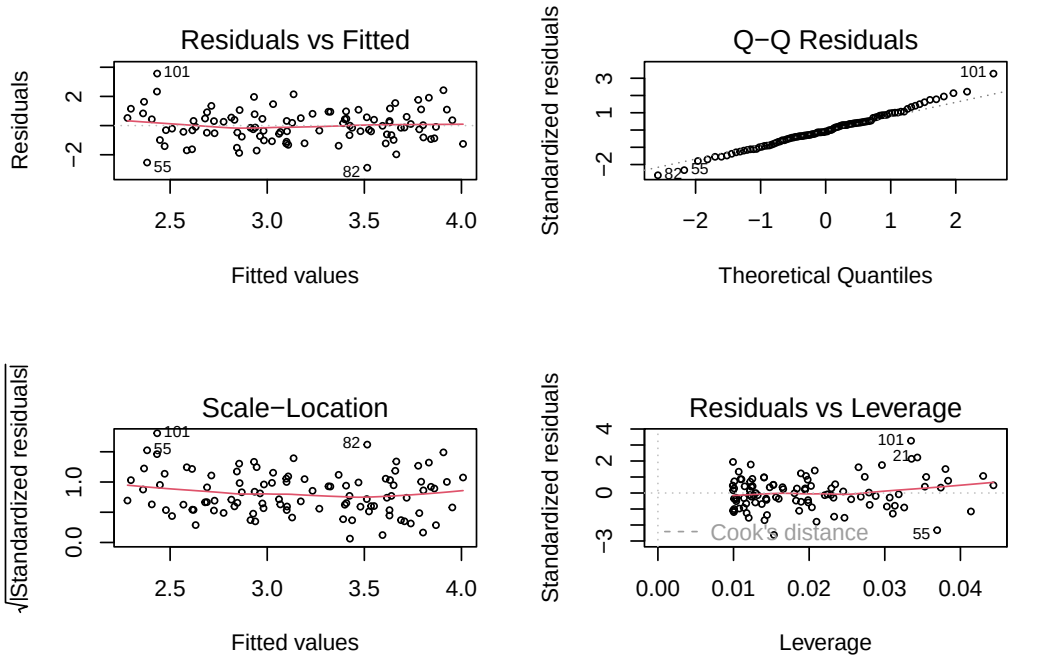

```
model_c <- lm(y~x1+x2)
model_d <- lm(y~x1)
model_e <- lm(y~x2)


par(mfrow = c(2, 2))
plot(model_c,cex = 0.6)
```



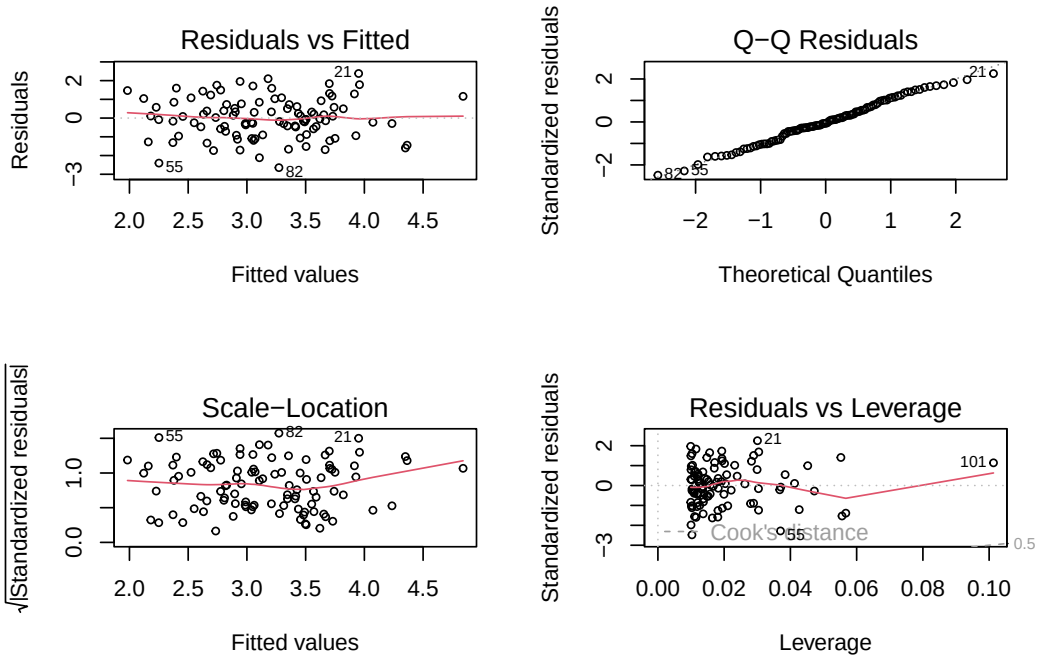

In the first model, from the last figure we can find out that the new point 's leverage is larger than 0.4 and residual is also large enough, so it tends to be a high-leverage point and also an outlier, thus it can a treated as a high influential point, which is consistent with its large cook's distance.

```
par(mfrow = c(2, 2))
plot(model_d,cex = 0.6)
```



In the second model, from the last figure we can find out that the new point is not a high-leverage point but an outlier, because its leverage is relatively normal but its residual is high.

```
par(mfrow = c(2, 2))
plot(model_e,cex = 0.7)
```




In the third model, from the last figure we can find out that the new point is a high-leverage point but not an outlier, because its leverage is high but its residual is small.