S631 HW5

Shibi He

2. Using real data to check question 1a.

a. Use lifeExpF, fertility, and ppgdp in UN11 to check $SYY = SSreg + RSS$.

```
library(alr4)
library(car)
library(pracma)

Y = UN11$lifeExpF
X = cbind(1, UN11$fertility, UN11$ppgdp)
n = length(Y)

H = X%*%qr.solve(t(X)%*%X)%*%t(X)
J = ones(n, n)
I = diag(1, n, n)
```

```
SYY = t(Y)%*%(I-(1/n)*J)%*%Y
RSS = t(Y)%*%(I-H)%*%Y
SSreg = t(Y)%*%(H-(1/n)*J)%*%Y

SYY
```

```
##          [,1]
## [1,] 20293.22
```

```
RSS
```

```
##          [,1]
## [1,] 5652.248
```

```
SSreg
```

```
##          [,1]
## [1,] 14640.97
```

$SYY = 20293.22$, $Rss = 5652.248$, and $SSreg = 14640.97$. Therefore, we have $SYY = SSreg + RSS$.

b. Check you will get the same value for $h_{3,4}$ using two methods:

```
h34 <- t(X[3,])%*%qr.solve(t(X)%*%X)%*%X[4,]
h34
```

```
##               [,1]
## [1,] -0.0003981503
```

```
H[3,4]
```

```
## [1] -0.0003981503
```

Constructing the equation in part 1b for $h_{3,4}$ gives the same value as extracting the value in the third row and fourth column in H. Specifically, $h_{3,4} = -0.000398$

3. ALR 4.2

```
data(Transact)
Transact$a <- (Transact$t1+Transact$t2)/2
Transact$d <- Transact$t1- Transact$t2
```

Fit four mean functions:

```
M1 <- lm(time ~ t1+t2, data=Transact)
M2 <- lm(time ~ a+d, data=Transact)
M3 <- lm(time ~ t2+d, data=Transact)
M4 <- lm(time ~ t1+t2+a+d, data=Transact)

compareCoefs(M1, M2, M3, M4, se=FALSE)
```

```
## Calls:
## 1: lm(formula = time ~ t1 + t2, data = Transact)
## 2: lm(formula = time ~ a + d, data = Transact)
## 3: lm(formula = time ~ t2 + d, data = Transact)
## 4: lm(formula = time ~ t1 + t2 + a + d, data = Transact)
##
##             Model 1 Model 2 Model 3 Model 4
## (Intercept)     144     144     144     144
## t1             5.46                    5.46
## t2             2.03            7.50    2.03
## a                       7.5         aliased
## d                      1.71    5.46 aliased
```

4.2.1

In the fit of M4, the coeffiicients of a and d are labeled as "aliased" because these two variables are simply linear combinations of the other two variables in the model: t1 and t2. That is, a and d can be determined exactly from the others. Therefore, R will omitt them when estimating the model.

4.2.2

The estimated intercepts are the same among these four models ($\hat{\beta_0} = 144$). The estimated coefficients of t1 and t2 are the same in M1 and M4. Other estimated coefficients are different.

4.2.3

The coefficient of a predictor not only depends on the predictor itself, but also depends on the other predictors in the model. In M1, the coefficient of t2 is the effect on time of increasing t2 by 1 unit, holding t1 constant. In M3, the coefficient of t2 is the effect on time of increasing t2 by 1 unit, holding d=t1-t2 constant. Therefore, the coefficients of t2 is different in M1 and M3.

4. ALR 4.6 and 4.7

4.6

$$log(fertility) = 1.501 - 0.01pctUrban$$

The intepretation of the estimated coefficient for pctUrban is: for a one-unit increase in pctUrban, we expect to see a 1% decrease in fertitlity, since $\exp(-0.01) \approx 99\%$.

4.7

```
data(UN11)
Reg1 <- lm(log(fertility) ~ log(ppgdp)+ lifeExpF, data=UN11)
summary(Reg1)
```

```
##
## Call:
## lm(formula = log(fertility) ~ log(ppgdp) + lifeExpF, data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61778 -0.16891  0.03731  0.17591  0.61072
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.50736    0.12707  27.601  < 2e-16 ***
## log(ppgdp)  -0.06544    0.01781  -3.675 0.000307 ***
## lifeExpF    -0.02824    0.00274 -10.306  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.248 on 196 degrees of freedom
## Multiple R-squared:  0.6926, Adjusted R-squared:  0.6894
## F-statistic: 220.8 on 2 and 196 DF,  p-value: < 2.2e-16
```

The estimated coefficient of $\log(ppgdp)$ is -0.06544. A 25% increase in ppgdp is associated with a 1.4% decrease in expected fertility, since $(1 + 0.25)^{-0.06544} - 1 = -0.014$.

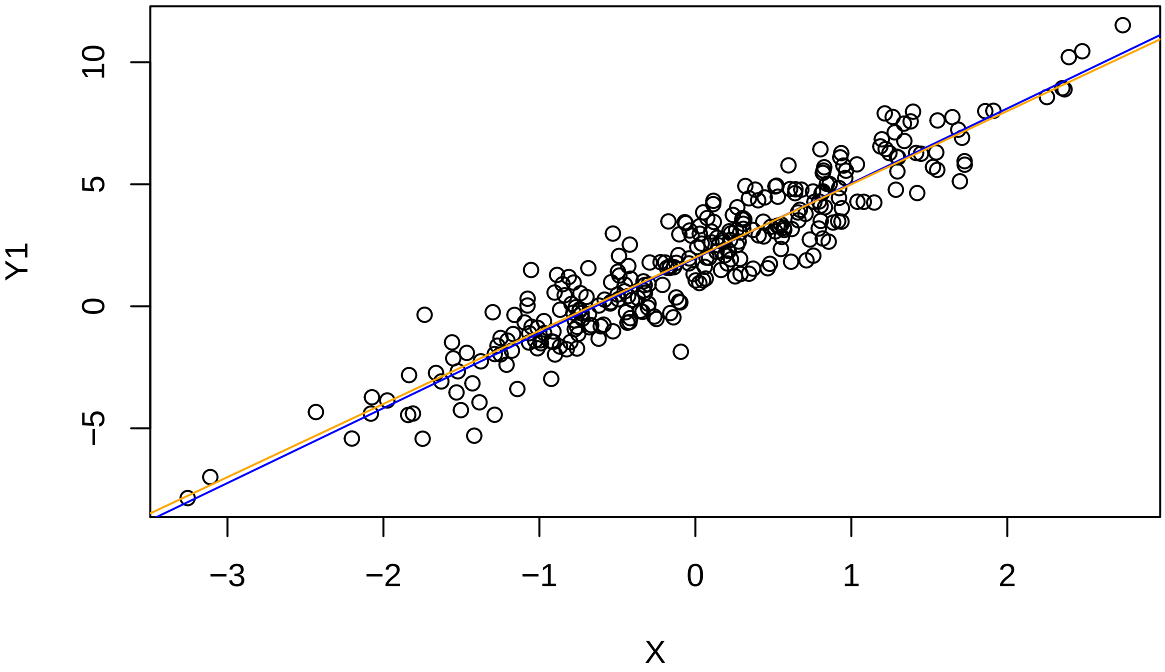
5. ALR 4.12

4.12.1

```
set.seed(520)
n=300
x = rnorm(n, mean = 0, sd = 1)
e = rnorm(n, mean = 0, sd = 1)
y1 = 2+ 3*x + e
df <- data.frame(Y1=y1, X=x, e=e)
```

Create a scatter plot:

```
plot(Y1 ~ X, df)
model1 <- lm(Y1 ~ X, data=df)
#summary(model1)
abline(model1, col= "blue")
abline(2, 3, col="orange")
```

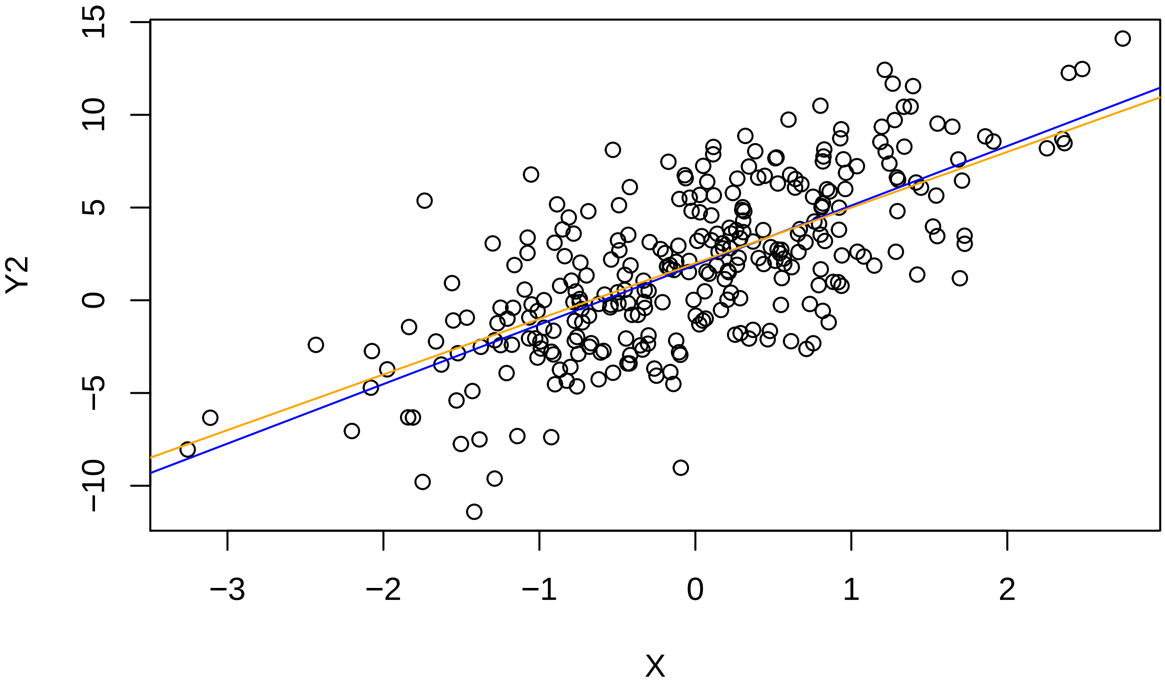


The scatter of points is approximately elliptical. The blue line indicates the OLS regression line and the orange line indicates the true regression line. These two lines are very similar to each other, but not exactly the same.

4.12.2

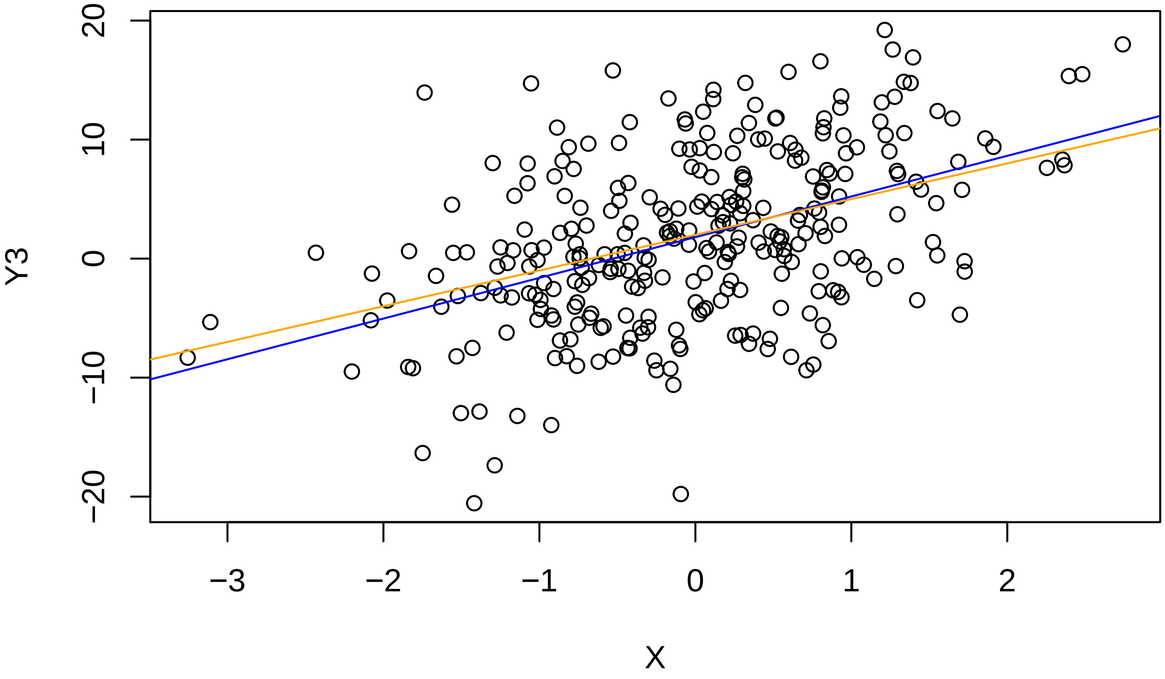
Repeat the exercise with $\sigma = 3$:

```
y2 = 2+ 3*x + 3*e
df$Y2 <- y2

plot(Y2 ~ X, df)
model2 <- lm(Y2 ~ X, data=df)
#summary(model2)
abline(model2, col= "blue")
abline(2, 3, col="orange")
```



Repeat the exercise with $\sigma = 6$:

```
y3 = 2+ 3*x + 6*e
df$Y3 <- y3

plot(Y3 ~ X, df)
model3 <- lm(Y3 ~ X, data=df)
#summary(model3)
abline(model3, col= "blue")
abline(2, 3, col="orange")
```

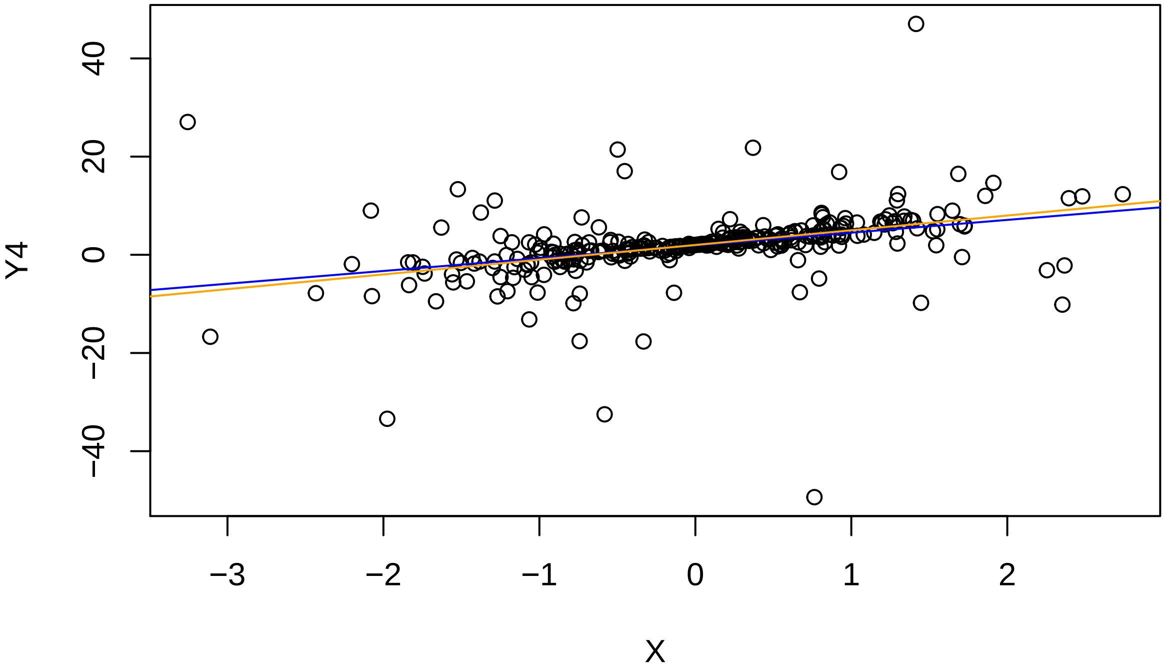


As σ increases, the scatter of points becomes more and more dispersed. The linear trend in the scatter plots becomes less evident.

4.12.3

Repeat the exercise with e following a Cauchy distribution:

```r
set.seed(520)
n=300
V1 <- rnorm(n, mean = 0, sd = 1)
V2 <- rnorm(n, mean = 0, sd = 1)
e2 <- V1/V2

y4 = 2+ 3*x + e2
df$Y4 <- y4

plot(Y4 ~ X, df)
model4 <- lm(Y4 ~ X, data=df)
#summary(model4)
abline(model4, col= "blue")
abline(2, 3, col="orange")
```



With e following a Cauchy distribution, the scatter of points becomes even more dispersed. It's difficult to see the linear trend in the relationship between Y and X.