Hw8

Dandong Tu

2017/10/24

a

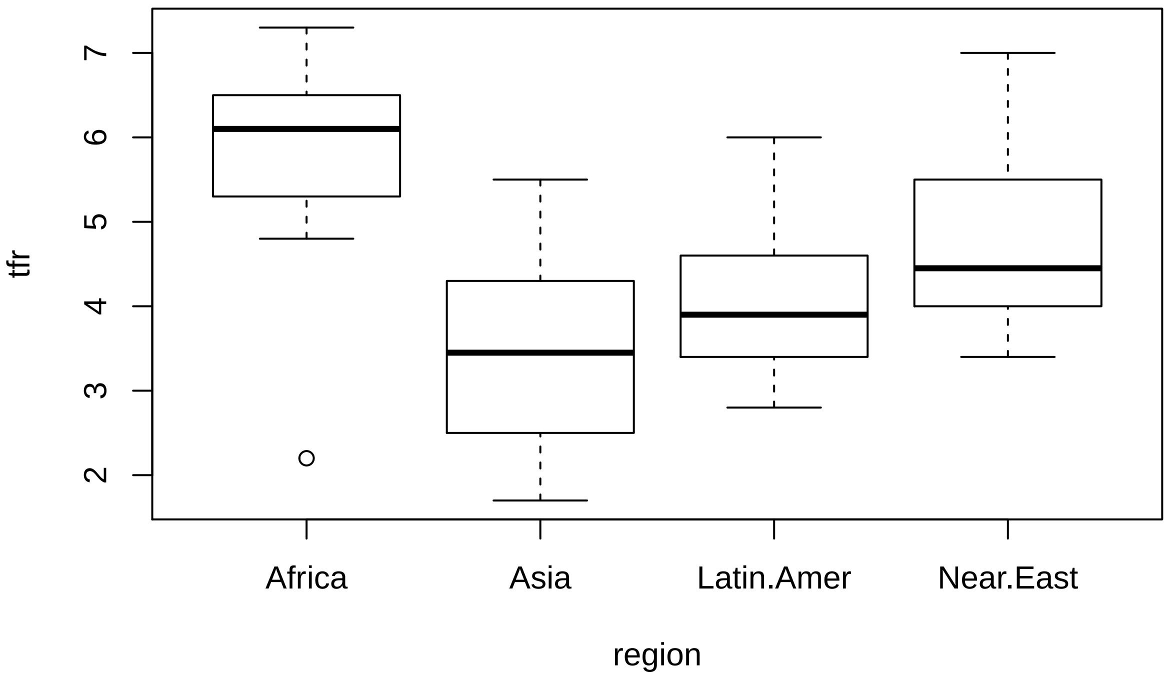
```
library(alr4)
```

```
## Loading required package: car
## Loading required package: effects
##
## Attaching package: 'effects'
## The following object is masked from 'package:car':
##
##     Prestige
```

```
data1=read.table("/Users/dandongtu/Downloads/Robey.txt", header=TRUE)
```

Based on the description of the data, I chose **tfr** as the response.

```
plot(tfr~region,data=data1)
```




```
m1=lm(tfr ~region,data=data1)
summary(m1)$coef
```

```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)       5.855556  0.2674094 21.897344 5.722778e-26
## regionAsia       -2.315556  0.4474615 -5.174871 4.877102e-06
## regionLatin.Amer -1.805556  0.3898128 -4.631853 2.989023e-05
```

```
## regionNear.East  -1.055556  0.5348188 -1.973670 5.444346e-02
```

```r
a=c(0,1,-1,0)
se_b2b3=sqrt(t(a)%*%vcov(m1)%*%a)
b2b3=as.numeric(coef(m1)[2]-coef(m1)[3])
t_val=b2b3/se_b2b3
p_val=2*(1-pt(abs(t_val),m1$df))
c("b2-b3"=b2b3,"SE"=se_b2b3,"t-Value"=t_val,"p-Value"=p_val)
```

```
##      b2-b3         SE    t-Value    p-Value
## -0.5100000  0.4573404 -1.1151431  0.2705813
```

In summary, we have the **p-value=5.444346e-02** when we move from **region=Africa** to **region=Near.East**

After,the calculation, we obtained the **P-value=0.2705813**when we move from **region=Asia** to **region=Latin.Amer**

These two **p-value**s has significant difference on the expected respone, which is **tfr**. The value equal **0.0544** indicate that it is almost statistically significant in the model, while value of **0.277** indicates it is no much significant in our model.

b

```r
summary(m1)$coef
```

```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)       5.855556  0.2674094 21.897344 5.722778e-26
## regionAsia       -2.315556  0.4474615 -5.174871 4.877102e-06
## regionLatin.Amer -1.805556  0.3898128 -4.631853 2.989023e-05
## regionNear.East  -1.055556  0.5348188 -1.973670 5.444346e-02
```

$\hat{\beta}_0 = 5.855556$ means the value of $\mathbf{E(Y)}$ when **region=Africa**

$E(Y|U_1 = 1) = \beta_0$ while $U_2 = 0$ $U_3 = 0$ and $U_4 = 0$

$\hat{\beta}_2 = -2.315556$ means the estimated change in $\mathbf{E(Y)}$ when we move from **region=Africa** to **region=Asia**

$E(Y|U_2 = 1) = \beta_0 + \beta_2$ while $U_3 = 0$ and $U_4 = 0$

$\hat{\beta}_3 = 1.805556$ means the estimated change on $\mathbf{E(Y)}$ when we move from **region=Africa** to **region=Latin.Amer**

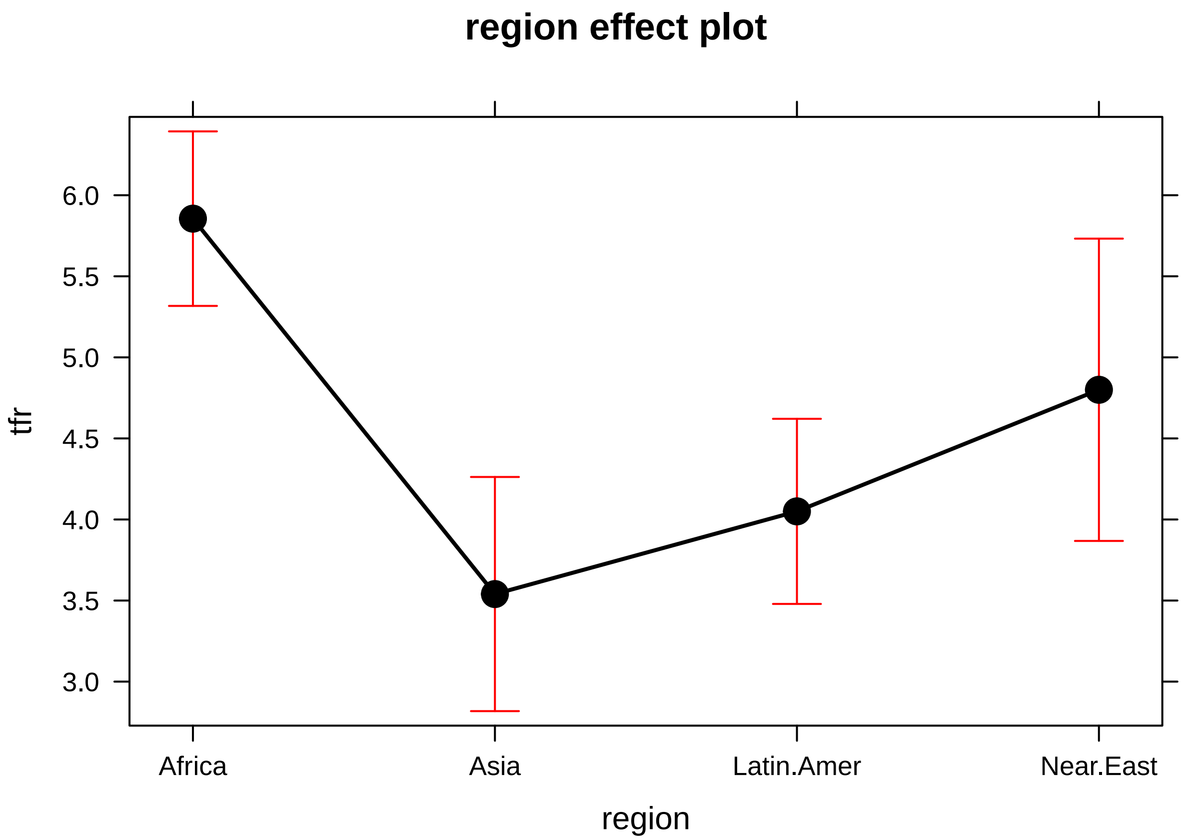
$E(Y|U_3 = 1) = \beta_0 + \beta_3$ while $U_2 = 0$ and $U_4 = 0$

$\hat{\beta}_= - 1.055556$ means the estimated change on $\mathbf{E(Y)}$ when we move from **region=Africa** to **region=Near.East**

$E(Y|U_4 = 1) = \beta_0 + \beta_4$ while $U_2 = 0$ and $U_3 = 0$

c

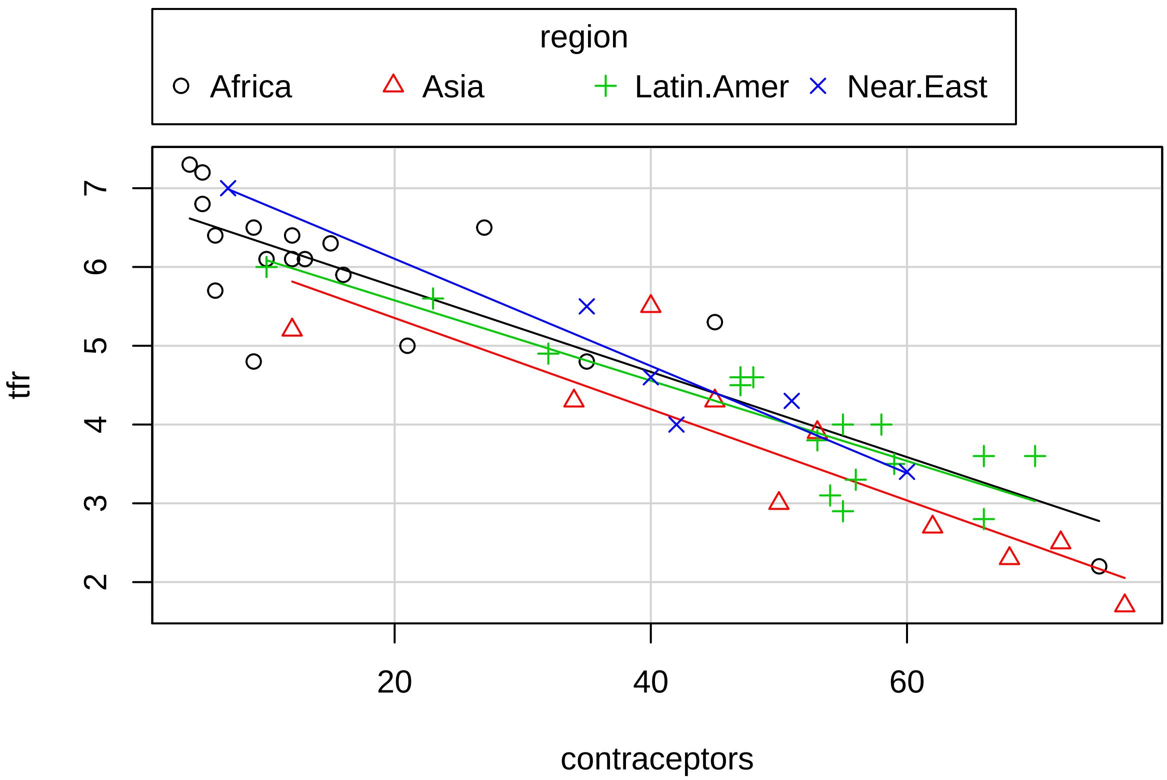
```r
plot(Effect(c("region"), m1))
```



From the effects plot above, we can see that the value for $\mathbf{E(Y)}$ when **region=Africa** is around **5.85** and the value for $\mathbf{E(Y)}$ is about **3.5**when we move from **region=Africa** to **region=Asia**.Similarlly, value for $\mathbf{E(Y)}$ is about **4.0**when we move from **region=Africa** to**region=Latin.Amer**. And value for $\mathbf{E(Y)}$ is about **4.8**when we move from **region=Africa** to **region=Near.East**

d

```
scatterplot(tfr~contraceptors| region, data=data1,
            smooth=FALSE, boxplots=FALSE)
```



From the scatterplot, the black line indicates $\hat{E}(Y|X=x,U_1=1)=\hat{\beta}_0+\hat{\beta}_1 x$

Similarlly, the red line indicates $\hat{E}(Y|X=x,U_2=1)=(\hat{\beta}_0+\hat{\beta}_{02})+(\hat{\beta}_1+\hat{\beta}_{12})x$

The green line indicates $E(Y|X=x,U_3=1)=(\hat{\beta}_0+\hat{\beta}_{02})+(\hat{\beta}_1+\hat{\beta}_{13})x$

The green line indicates $E(Y|X=x,U_4=1)=(\hat{\beta}_0+\hat{\beta}_{02})+(\hat{\beta}_1+\hat{\beta}_{14})x$

Meanwhile, we can easily observe that the slope for these line are less than **0**, so that we could conclude that changes in continuous regressors are associated to change on expected response. Put antoher word, when the **x** is changing, the estimated expected value will change.

Also, it is not difficult to see that the slops for each levels are different and as well as intercepts.

e

```
m2=lm(tfr~region*contraceptors,data=data1)
round(summary(m2)$coef,3)
```

```
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     6.832      0.194  35.202    0.000
## regionAsia                     -0.322      0.564  -0.572    0.570
## regionLatin.Amer               -0.237      0.521  -0.456    0.651
## regionNear.East                 0.632      0.633   0.998    0.324
## contraceptors                  -0.054      0.008  -7.009    0.000
## regionAsia:contraceptors       -0.004      0.012  -0.306    0.761
## regionLatin.Amer:contraceptors  0.003      0.012   0.260    0.796
## regionNear.East:contraceptors  -0.014      0.016  -0.862    0.393
```

The model **m2** is the model we observed with interaction. Similarlly with previous parts, we have the coeffcient of **-0.322,-0.237,-0.633** which indicates the value of **E(Y)**if we move from **region=Africa** to

region=Asia,region=Latin.Amer,region=Near.East,while keep the continues regressor and interaction in the model.

And we have the coeffcient of **-0.064** which indicates the changing in **E(Y)** if we change one unit of **contraceptors**,while keep the factors and the interactions.

Also, we observed the coeffcients for interactions.And now we have the model: $E(Y) = \hat{\beta}_0 + \hat{\beta}_{01}U1 + \hat{\beta}_{02}U2 + \hat{\beta}_{03}U3 + \hat{\beta}_1 x + \hat{\beta}_{12}x + \hat{\beta}_{13}x + \hat{\beta}_{14}x$ When the **region=Africa** we has the $\hat{E}(Y|X = x, U_1 = 1) = \hat{\beta}_0 + \hat{\beta}_1 x$,the term $\hat{\beta}_1$ is the slope and the $\hat{\beta}_0$is the intercept. And when we move the **region=Africa** to **region=Asia** we have $\hat{E}(Y|X = x, U_2 = 1) = (\hat{\beta}_0 + \hat{\beta}_{02}) + (\hat{\beta}_1 + \hat{\beta}_{12})x$ So, we observed the term $(\hat{\beta}_1 + \hat{\beta}_{12})$ as the slope(the slope of red line in part(d)) and the term $(\hat{\beta}_0 + \hat{\beta}_{02})$ as intercept(the intercept for red line in part(d)),and the slope also indicates to the coefficient **regionAsia:contraceptors**. The means for coefficient **regionLatin.Amer:contraceptors** and **regionNear.East:contraceptors** are similar when we move to different **region**.

f

Based on the table in part(e), we observed high **p-value** for interactions(**0.761**,**0.796** and **0.393**). For the hypothesis, $H_0 : \beta_1 2 = 0$ and $H_a : \beta_1 2 \neq 0$, since **p-value** are large, I fail to reject H_0. We do not have sufficient evidence to conclude that the slope of the regression **"tfr~contraceptors"** change while I move from **region=Africa** to **region=Asia**,**region=Latin.Amer** and **region=Near.East**

We conclue that chaning in **E(Y)** for interactions are not statistically significantly differentiate for different **region** levels, therefore, we drop the interaction terms.

The **p-value**for the continues regressor is essencially **0** indicates the statisticly significant in our model, and suggest that we should include **contraceptors** in our model.

```
b=c(0,0,1,-1)
se_b3b4=sqrt(t(b)%*%vcov(m1)%*%b)
b3b4=as.numeric(coef(m1)[3]-coef(m1)[4])
t_val=b3b4/se_b3b4
p_val=2*(1-pt(abs(t_val),m1$df))
c("b3-b4"=b3b4,"SE"=se_b3b4,"t-Value"=t_val,"p-Value"=p_val)
```

```
##      b3-b4         SE    t-Value    p-Value
## -0.7500000  0.5431110 -1.3809331  0.1739721
```

```
c=c(0,1,0,-1)
se_b2b4=sqrt(t(c)%*%vcov(m1)%*%c)
b2b4=as.numeric(coef(m1)[2]-coef(m1)[4])
t_val=b2b4/se_b2b4
p_val=2*(1-pt(abs(t_val),m1$df))
c("b2-b4"=b2b4,"SE"=se_b2b4,"t-Value"=t_val,"p-Value"=p_val)
```

```
##       b2-b4          SE     t-Value     p-Value
## -1.26000000  0.58586463 -2.15066748  0.03679087
```

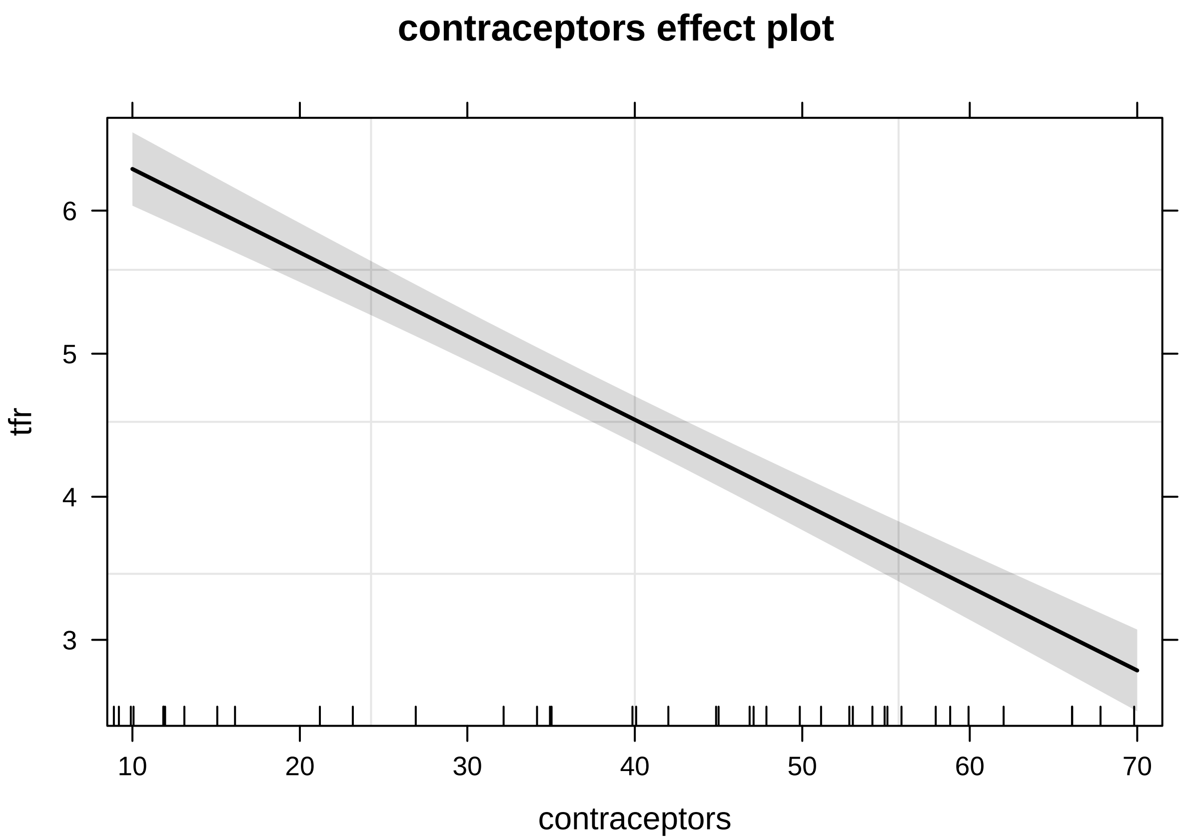
Based on the summary of **m2** and the above test,as well as the test we finished in part(a). The **p-value** for these movements in different factor levels are such high, and it indicates the factors is not statisticly significant. Therefore, we decide to drop the factors as well.

g

```
m3=lm(tfr~contraceptors,data=data1)
round(summary(m3)$coef,3)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.875      0.157  43.829        0
## contraceptors   -0.058      0.004 -16.299        0
```

```
plot(allEffects(m3),
     grid=TRUE, multiline=TRUE)
```



Now, we only have **contraceptors** in our model. The effect plot shows above indicates the changing in **tfr** while we have different value of **contraceptors**. From the observation, we see that when the value of **contraceptors** are in low range or high range, it has more effect rather than it's in the middle.

h

```
aggregate(contraceptors~region,data1,mean)
```

```
##       region contraceptors
## 1     Africa      18.05556
## 2       Asia      51.30000
## 3 Latin.Amer      49.93750
## 4  Near.East      39.16667
```

```r
data_new=data.frame(contraceptors=51.3,region="Asia")
predicted_value=predict(m3,data_new,interval="prediction")
(predicted_value)
```

```
##        fit      lwr      upr
## 1 3.878358 2.707563 5.049153
```

From the result, we observed the fit value of **3.878** and we have the predicted interval in [**2.7**,**5.5**]