پروژه درس رگرسیون 1– دانشگاه اراک – نیم‌سال 991

عنوان پروژه

بررسی میزان درآمد راننده آژانس

پدیدآورندگان :
محراب عتیقی
پریسا قائدرحمتی

با توجه به اطلاعات به دست آمده از آژانس معلم ، میخواهیم اثر تعداد سرویس درون و برون شهری راننده و متوسط مدت زمانی که راننده در طول یک ماه درآژانس حضور داشته و تاثر آن بر میزان درآمدش را بررسی کنیم

y: میانگین درآمد راننده در طول یک ماه(به ریال)
x1:میانگین ساعت حضور راننده در آژانس
x2: تعداد سرویس های درون شهری راننده در طول یک ماه
x3: تعداد سرویس های برون شهری راننده در طول یک ماه

	y	X1	X2	X3
1	45582000	12	22	9
2	35400000	10	24	7
3	25507800	11	26	6
4	28542800	9	24	7
5	27686000	10	26	5
6	25921300	9	30	1

ابتدا مدل رگرسیون چندگانه را اجرو سپس ضرایب را تفسیر می کنیم:

```
> fit<-lm(y~x1+x2+x3)
> summary(fit)

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
         1          2          3          4          5          6
 8.207e+05  1.888e+06 -9.837e-09 -8.226e+04 -4.350e+06  1.724e+06

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 356222236  130285959   2.734    0.112
x1            4886831    1995370   2.449    0.134
x2          -12153029    4388433  -2.769    0.109
x3          -11415137    4670299  -2.444    0.134

Residual standard error: 3615000 on 2 degrees of freedom
Adjusted R-squared:  0.7849        Multiple R-squared:  0.914,
F-statistic: 7.082 on 3 and 2 DF,  p-value: 0.1262
```

به ازای صفر بودن مقادیر x1,x2,x3، متوسط y 356222236 واحد خواهد بود.
به ازای افزایش یک واحدی x1(میانگین ساعت حضور)، با شرط ثابت بودن x2,x3 متوسط y، 4886831 واحد افزایش می‌یابد.
به ازای افزایش یک واحدی x2(سرویس های درون شهری)، باشرط ثابت بودن x1,x3 متوسط y، 12153029 واحد کاهش می‌یابد.
به ازای افزایش یک واحدی x3(سرویس های بیرون شهری)، با شرط ثابت بودن x1,x2 متوسط y، 11415137 واحد کاهش می‌یابد.

خطای معیار ضرایب رگرسیون ستون قرمز رنگ به ترتیب از بالا به پایین برای $\beta 0, \beta 1, \beta 2, \beta 3$ است.
همچنین مشاهده می کنیم جذر واریانس با دو درجه آزادی برابر با 3615000 می باشد.

برای به دست آوردن ضرایب رگرسیونی داریم:

```
> confint(fit)
                 2.5 %    97.5 %
(Intercept) -204353000 916797473
x1            -3698552  13472213
x2           -31034930   6728873
x3           -31509812   8679538
```

در بازه های فوق، مقادیری که ضرایب x1,x2,x3 و عرض از مبداء می‌توانند بگیرند،قرار دارند و از بالا به پایین مشخص شده اند.
یعنی با احتمال 95٪ عرض از مبدا و ضرایب هر متغییر ما در این بازه ها قرار دارند.

```
(pred.w.clim<-predict(fit , interval="confidence"))
       fit      lwr      upr
1 44761334 30273070 59249599
2 33511890 25406524 41617256
3 25507800  9952830 41062770
4 28625060 14932847 42317273
5 32036107 24547263 39524950
6 24197709  9820468 38574951
```

برای میانگین پاسخ فاصله اطمینان مقادیر زیر را داریم که مقادیر دقیق در ستون قرمز رنگ و حد بالا در ستون آبی رنگ و حد پایین در ستون سبز رنگ قرار دارند که به ترتیب برای رانندگان اول تا ششم درج شده است.

```
> newd=data.frame(x1=11,x2=20,x3=3)
> (pred.w.clim<-predict(fit ,newdata = newd, interval="confidence"))
       fit       lwr       upr
1 132671384 -28031185 293373953
```

برای پیش‌بینی یک فاصله اطمینان برای متوسط میانگین پاسخ خود زمانی که x1=11,x2=20,x3=3 هستند، با توجه به خروجی مقدار دقیق و حدود بالا و پایین نیز مشخص شده است.

```
> (pred.w.clim<-predict(fit , interval="prediction"))
       fit      lwr      upr
1 44761334 23504173 66018496
2 33511890 15971820 51051960
3 25507800  3509750 47505850
4 28625060  7902275 49347844
5 32036107 14772275 49299938
6 24197709  3016062 45379357
```

برای پیش بینی متوسط مقادیر و نشان دادن یک فاصله اطمینان برای آن به صورت بالا دیدیم که حدود هم نیز مشخص شده است. که هریک از سطر های یکی ازمقادیر وردی و رانده های ما را نشان میدهند.

```
> newd=data.frame(x1=8,x2=12,x3=7)
> (pred.w.clim<-predict(fit ,newdata = newd,interval="prediction"))
       fit       lwr       upr
1 169574575 -52565990 391715140
```

برای پیش بینی مقدار درآمد راننده زمانی که x1=8,x2=12,x3=7 به صورت بالا مشاهده کردیم و حدود بالا و پایین آن نیز مشخص بود.

برای آزمون معنی داری رگرسیون داریم:

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

Response: y
          Df     Sum Sq    Mean Sq F value  Pr(>F)
x1         1 1.3853e+14 1.3853e+14 10.5997 0.08279 .
x2         1 6.1057e+13 6.1057e+13  4.6717 0.16321
x3         1 7.8080e+13 7.8080e+13  5.9741 0.13444
Residuals  2 2.6139e+13 1.3070e+13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ستون قرمز رنگ درجه آزادی ها را نشان داده و ستون سبز رنگ نیز مقادیر SSE را برای متغییر های پیشگو ما نیز نشان میدهد و همچنین ستون آبی رنگ نیز مقادیر MS E را برای متغییر های پیشگو ما نشان میدهد و دو ستون اخر نیز میتواینم نتیجه گیری درمورد هریک از ضرایب رگرسیون خود داشته باشیم که آیا آن ضریب برای ما مفید ه ست یا نه که، هریک از سطر ها که در ستون اخر مقدارش کمتر از 0.05 یا همان آلفا ما هستش،نشان دهنده یک ضریب خوب هستش.

در قسمت بعدی آزمون فرض مربوط به هر ضریب را انجام داده ایم:

```
> fit1<-lm(y~x1)
> anova(fit1)
Analysis of Variance Table

Response: y
          Df     Sum Sq    Mean Sq F value Pr(>F)
x1         1 1.3853e+14 1.3853e+14  3.3528 0.1411
Residuals  4 1.6528e+14 4.1319e+13
> summary(fit1)
Call:
lm(formula = y ~ x1)

Residuals:
       1        2        3        4        5        6
 5887263  4710449 -9684344  2355841 -3003551  -265659

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14336376   25137236  -0.570    0.599
x1            4502593    2459005   1.831    0.141

Residual standard error: 6428000 on 4 degrees of freedom
Adjusted R-squared:   0.32        Multiple R-squared:  0.456,
F-statistic: 3.353 on 1 and 4 DF,  p-value: 0.1411



> fit2<-lm(y~x2)
> summary(fit2)
Call:
lm(formula = y ~ x2)

Residuals:
       1        2        3        4        5        6
 6930546  1075429 -4489889 -5781771 -2311689  4577375

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 86247157   23666132   3.644   0.0219 *
x2          -2163441     929694  -2.327   0.0805 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5681000 on 4 degrees of freedom
Adjusted R-squared:  0.4689       Multiple R-squared:  0.5752,
F-statistic: 5.415 on 1 and 4 DF,  p-value: 0.08051
```

```
> anova(fit2)
Analysis of Variance Table

Response: y
          Df     Sum Sq    Mean Sq F value  Pr(>F)
x2         1 1.7474e+14 1.7474e+14  5.4152 0.08051 .
Residuals  4 1.2907e+14 3.2268e+13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> fit3<-lm(y~x3)
> summary(fit3)

Call:
lm(formula = y ~ x3)

Residuals:
       1        2        3        4        5        6
 7708128  1589637 -6270809 -5267563 -2060855  4301462

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19588083    6432268   3.045   0.0382 *
x3           2031754    1014919   2.002   0.1159
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6160000 on 4 degrees of freedom
Adjusted R-squared:  0.3756       Multiple R-squared:  0.5005,
F-statistic: 4.008 on 1 and 4 DF,  p-value: 0.1159
> anova(fit3)
Analysis of Variance Table
Response: y
          Df     Sum Sq    Mean Sq F value Pr(>F)
x3         1 1.5205e+14 1.5205e+14  4.0076 0.1159
Residuals  4 1.5176e+14 3.7941e+13
```

همانگونه که میبینیم هیچ یک از ضرایب رگرسیونی ما مناسب نبوده ولی بطور کلی در بین ضرایب ما x2بهتر از بقیه بوده زیرا p value کوچک تری دارد.

و برای هریک از متغییر ها بطور جداگانه جدول آنوا را نیز تشکیل داده‌ایم.

حال برای بدست آوردن تاثیر اضافه کردن متغییر پیشگو x1 به مدلی که در آن متغییر پیشگو x2 وجود دارد داریم:

```
> fit4<-update(fit1,.~.+x2)
> anova(fit1,fit4,test="F")
Analysis of Variance Table

Model 1: y ~ x1
Model 2: y ~ x1 + x2
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1      4 1.6528e+14
2      3 1.0422e+14  1 6.1057e+13 1.7576 0.2768
```

حال برای بدست آوردن مدلی که شامل z1,z2 که هریک از آنها در پایین تعریف شده‌انددداریم:

```
> z1=x1
> z2=x2*x3
> Reduce.model<-lm(y~z1+z2)
> anova(fit,Reduce.model)
Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3
Model 2: y ~ z1 + z2
  Res.Df        RSS Df   Sum of Sq      F  Pr(>F)
1      2 2.6139e+13
2      3 1.4701e+14 -1 -1.2087e+14 9.2481 0.09325 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary(fit)

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
         1          2          3          4          5          6
 8.207e+05  1.888e+06 -9.837e-09 -8.226e+04 -4.350e+06  1.724e+06

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 356222236  130285959   2.734    0.112
x1            4886831    1995370   2.449    0.134
x2          -12153029    4388433  -2.769    0.109
x3          -11415137    4670299  -2.444    0.134

Residual standard error: 3615000 on 2 degrees of freedom
Adjusted R-squared:  0.7849        Multiple R-squared:  0.914,
F-statistic: 7.082 on 3 and 2 DF,  p-value: 0.1262
```

ضریب تعیین برابر 0.914 است و ضریب تعدیل یافته برابر 0.7849 می باشد

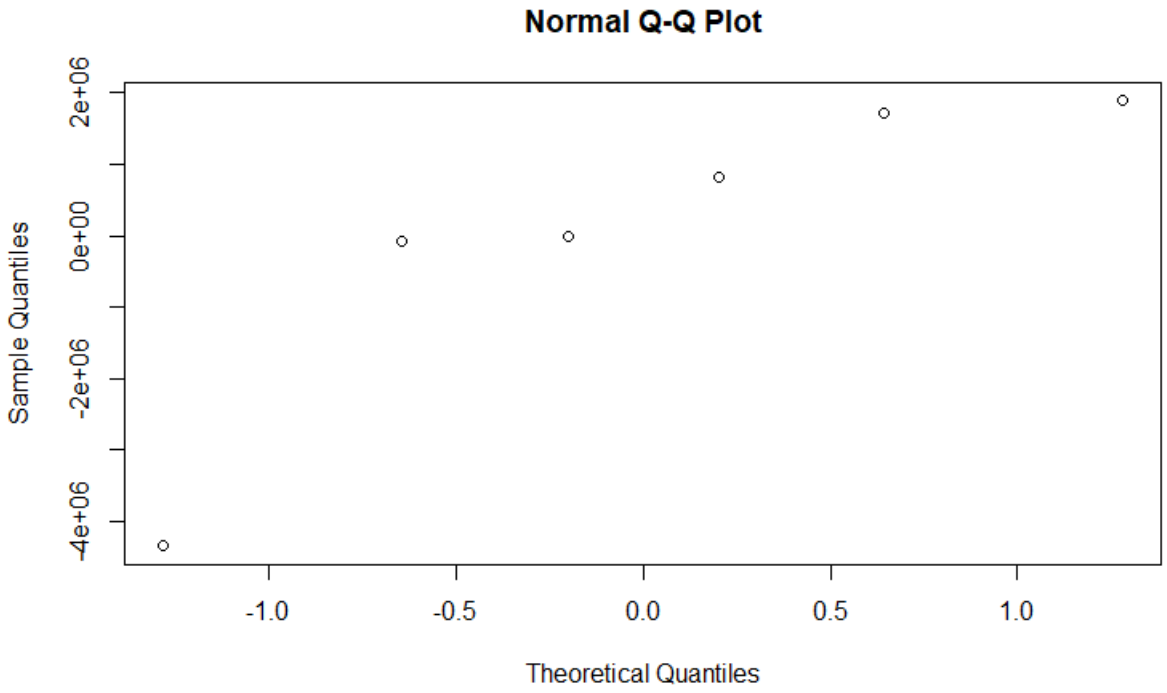
برای بدست آوردن خطاها داریم:

```
> e<-fit$residuals
> print(e)
           1            2            3            4            5
 8.206656e+05  1.888110e+06 -9.837095e-09 -8.225960e+04 -4.350107e+06
           6
 1.723591e+06
```

برای بدست آوردن مقادیر خطاهای استاندارد و استیودنت شده داریم:

```
> s<-rstandard(fit)
> print(s)
          1           2           3           4           5           6
 0.62374459  0.61190854         NaN -0.04795156 -1.37286074  1.24904587
> t<-rstudent(fit)
> print(t)
          1           2           3           4           5           6
 0.49143575  0.47993669         NaN -0.03392638 -4.04389275  1.88325486
```

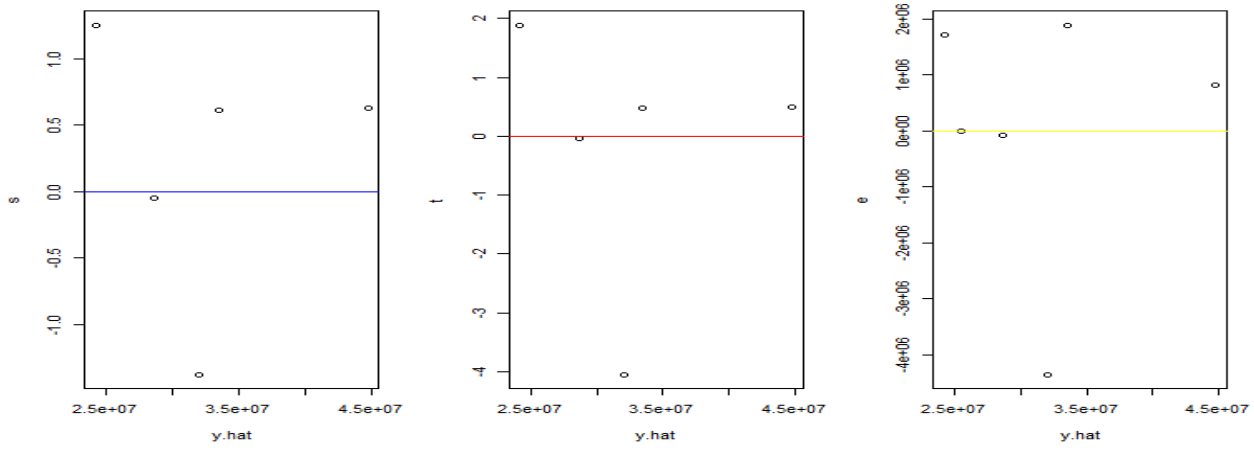
برای بررسی فرض نرمال بودن باقی‌مانده‌ها داریم:

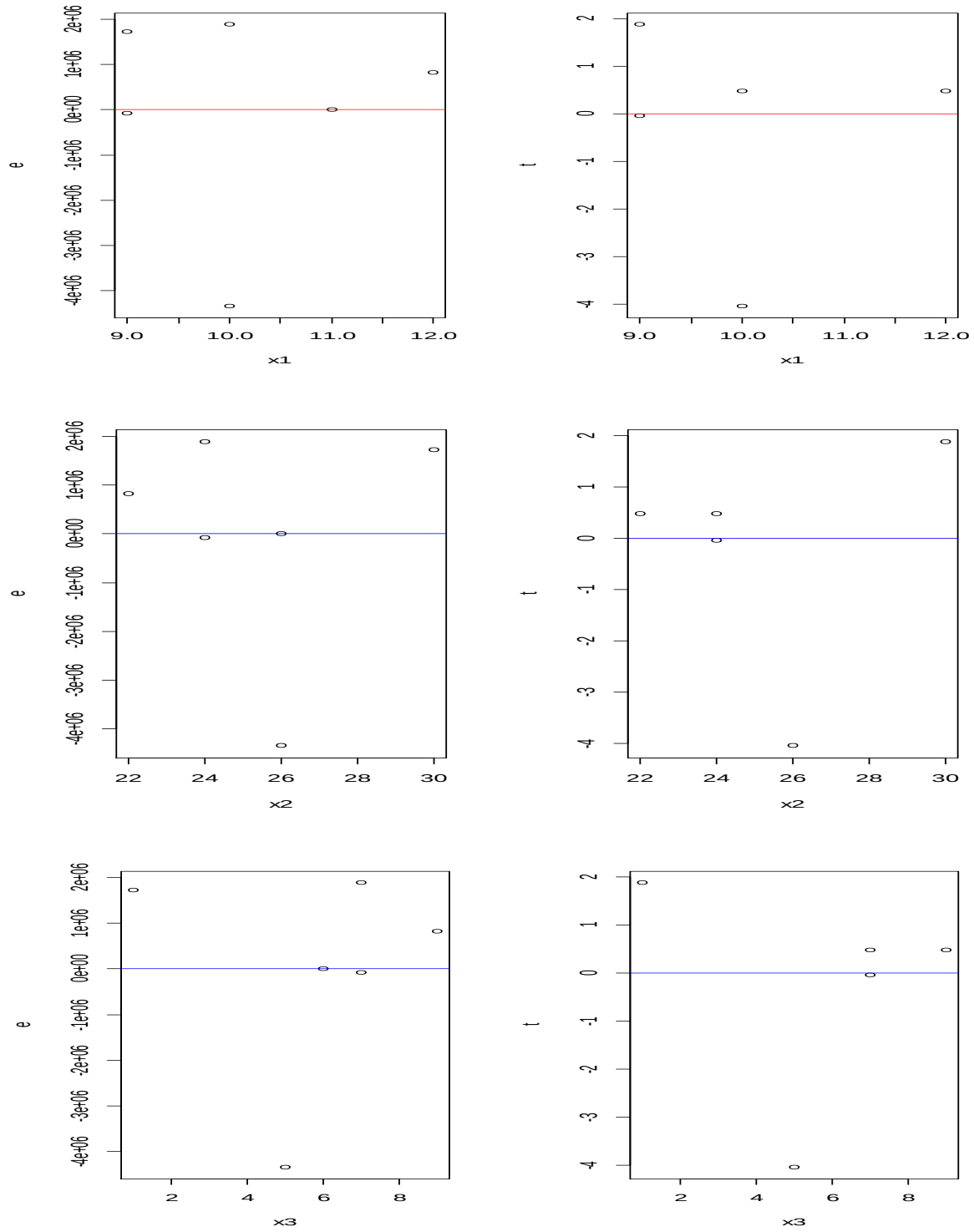


با توجه به نمودار بالا فرض نرمال بودن خطا ها رد می شود.

برای بدست رسم نمودار انواع باقی مانده هایمان در مقابل متغییر های پیشگو و متغییر پاسخ خود،داریم:

```
> y.hat<-fit$fitted.values
> par(mfrow=c(1,3))
> plot(y.hat,s)
> abline(h=0 , col="blue")
> plot(y.hat,t)
> abline(h=0 , col="red")
> plot(y.hat,e)
> abline(h=0,col="yellow")
> par(mfrow=c(1,2))
```





همانگونه که مشخص هست سومین داده ما یا باقی مانده مربوط به سومین داده از بازه منفی 2 تا 2 خارج و در نتیجه میتوانیم آنرا داده پرت معرفی کرده و از مدل حذفش کنیم.