پروژه درس آمار و احتمال مهندسی

تحلیل داده های خودروها با استفاده از R

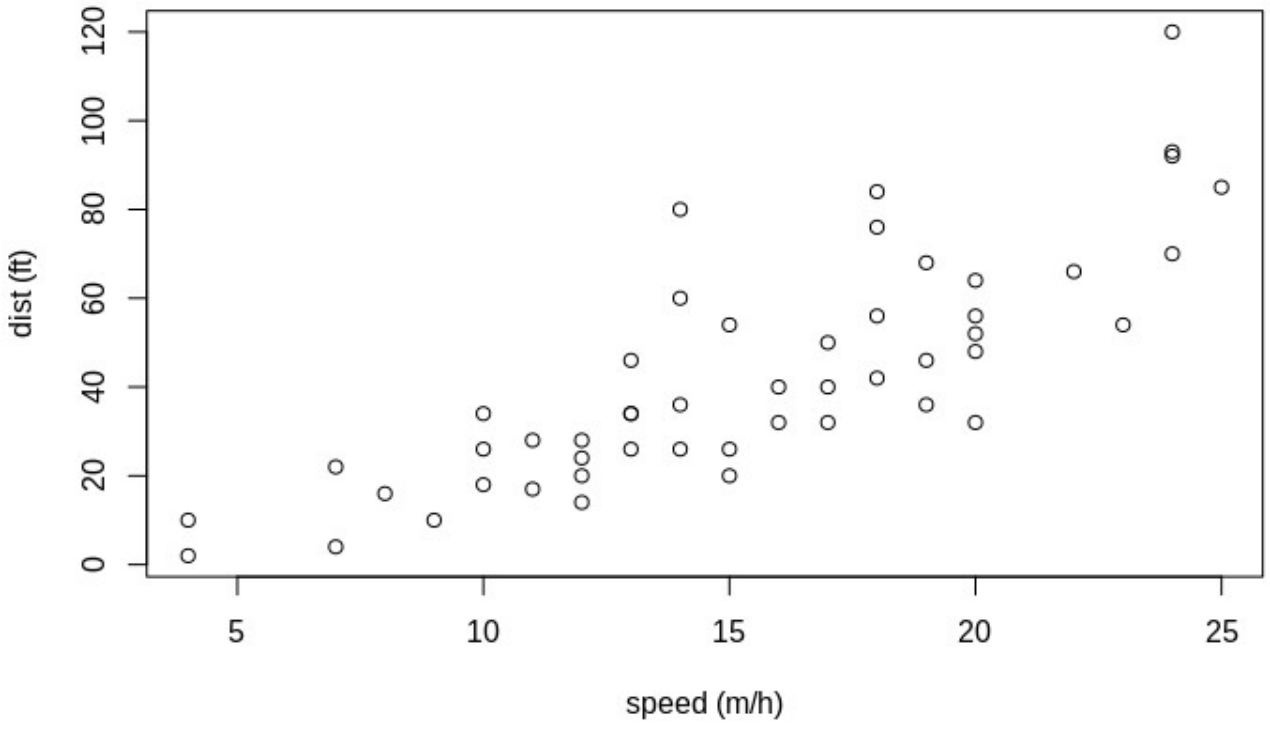
امیرمحمد کریمی

۹۶۳۶۱۳۰۷۷

زمستان ۹۸

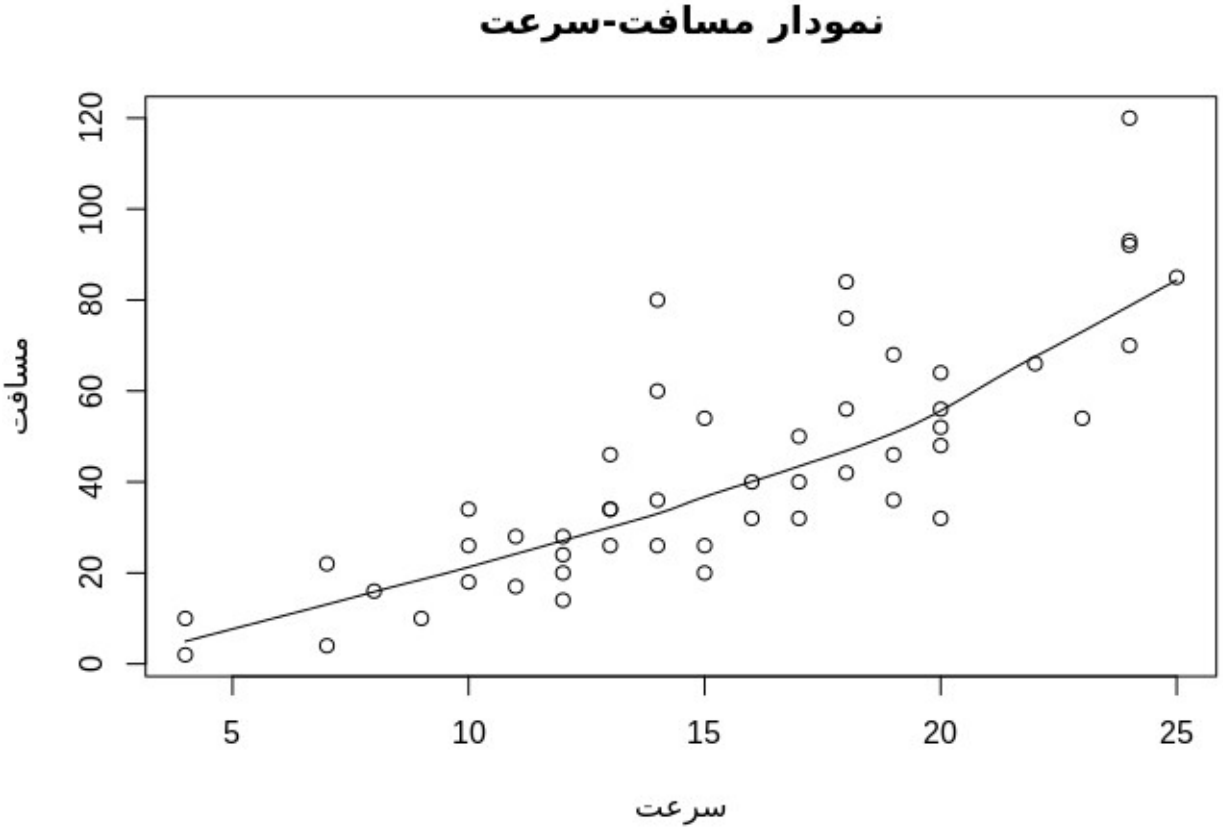
معرفی پروژه:

دراین پروژه از دیتاست cars، که به طور پیش فرض همراه با زبان R نصب می شود استفاده شده و رگرسیون خطی و آماره های رگرسیون مانند جدول ANOVA،آماره F، فواصل اطمینان برای ضرایب رگرسیون، مانده ها با استفاده از زبان R پیاده سازی شده است.سورس کد برنامه که تمام دستورات پروژه را دربر دارد در کنار این فایل ضمیمه شده است.



شرح پروژه:
دیتاست cars، شامل ۵۰ رکورد با دو ستون speed و dist است و می خواهیم رابطه بین سرعت خودروها و مسافتی که تا توقف خودرو موقع ترمز می پیمایند یافته و برای رگرسیون خطی، dist را تابعی از speed تعریف کنیم.پیش از انجام رگرسیون،ابتدا نمودار بین سرعت-مسافت را رسم می کنیم:

```
scatter.smooth(x=cars$speed,xlab = "سرعت", ylab =  "مسافت",  y = cars$dist, main ="نمودار مسافت-سرعت")
```



همانطور که مشاهده می شود به نظر رابطه ای خطی بین این دو متغیر می رسد.
با دستور زیر:

```
cov(cars$speed, cars$dist)
```

به ضریب همبستگی بین دو متغیر می رسیم که برابر ۰.۸۰۶۸۹۴۹ است. این نتیجه نشان می دهد وابستگی مستقیمی بین دو متغیر وجود دارد.
درادامه مدل خود را برای رگرسیون می سازیم.
با استفاده از قطعه کد زیر به ضرایب رگرسیون خطی بین این دو متغیر میرسیم:

linearMod <- lm(dist ~ speed, data=cars)

که شیب خط رگرسیون برابر ۳.۹۳۲ و عرض از مبدا برابر ۱۷.۵۷۹- است.

درباره خلاصه آماره ها:

```
> summary(linearMod)

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

فاصله اطمینان ۹۵ درصد پارامترهای مدل:

```
> confint(linearMod)
                2.5 %    97.5 %
(Intercept) -31.167850 -3.990340
speed         3.096964  4.767853
```

با دستور زیر نیز می بینیم که:

deviance(linearModel)

مجموع مربعات خطا: ۱۱۳۵۳.۵۲

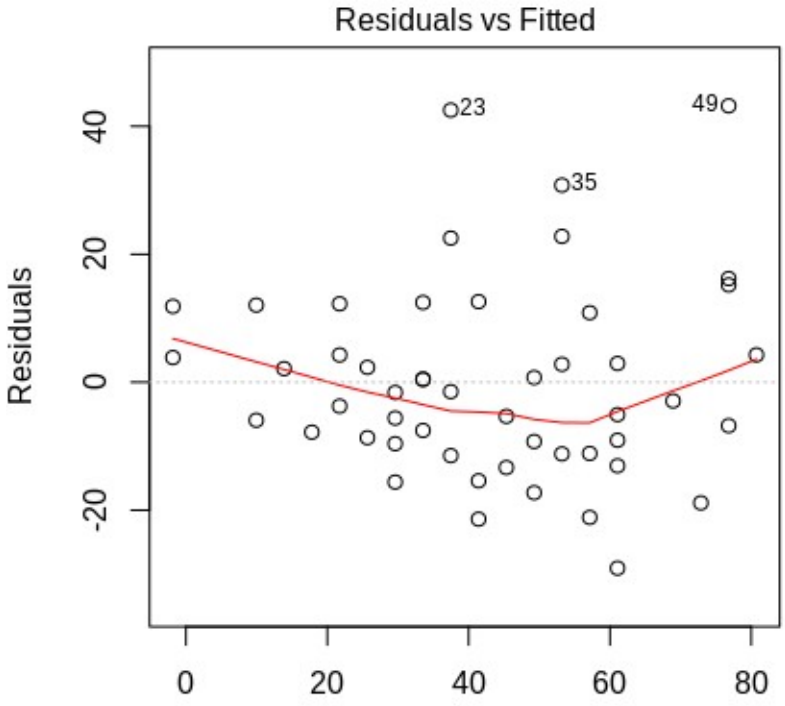
برای حساب مانده ها:

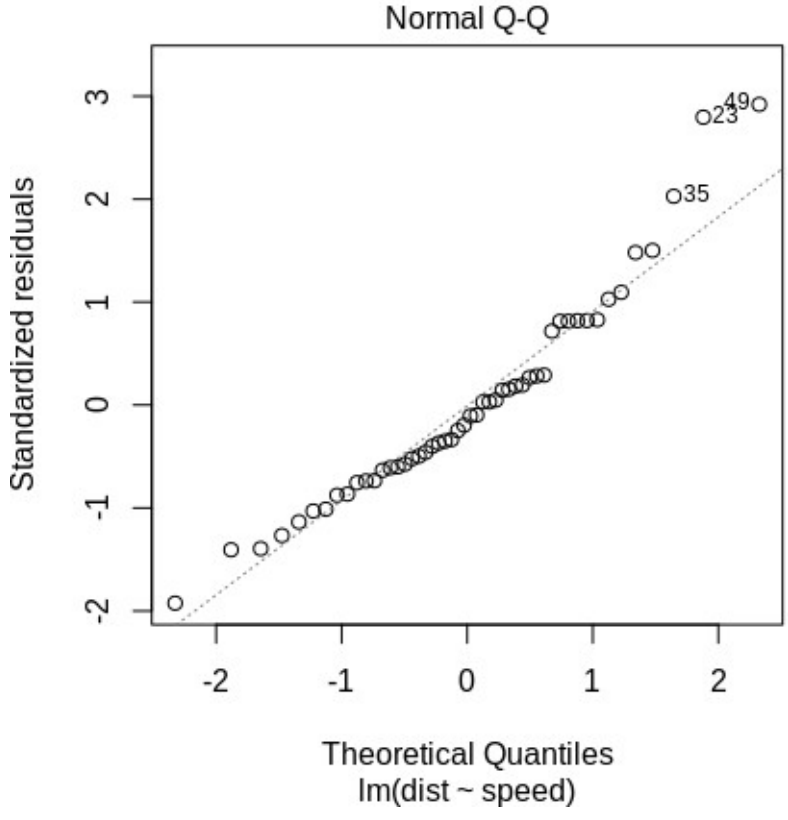
```
> residuals(linearMod)
         1          2          3          4          5
 3.849460  11.849460  -5.947766  12.052234   2.119825
         6          7          8          9         10
-7.812584  -3.744993   4.255007  12.255007  -8.677401
        11         12         13         14         15
 2.322599 -15.609810  -9.609810  -5.609810  -1.609810
        16         17         18         19         20
-7.542219   0.457781   0.457781  12.457781 -11.474628
        21         22         23         24         25
-1.474628  22.525372  42.525372 -21.407036 -15.407036
        26         27         28         29         30
12.592964 -13.339445  -5.339445 -17.271854  -9.271854
        31         32         33         34         35
 0.728146 -11.204263   2.795737  22.795737  30.795737
        36         37         38         39         40
-21.136672 -11.136672  10.863328 -29.069080 -13.069080
        41         42         43         44         45
-9.069080  -5.069080   2.930920  -2.933898 -18.866307
        46         47         48         49         50
-6.798715  15.201285  16.201285  43.201285   4.268876
```

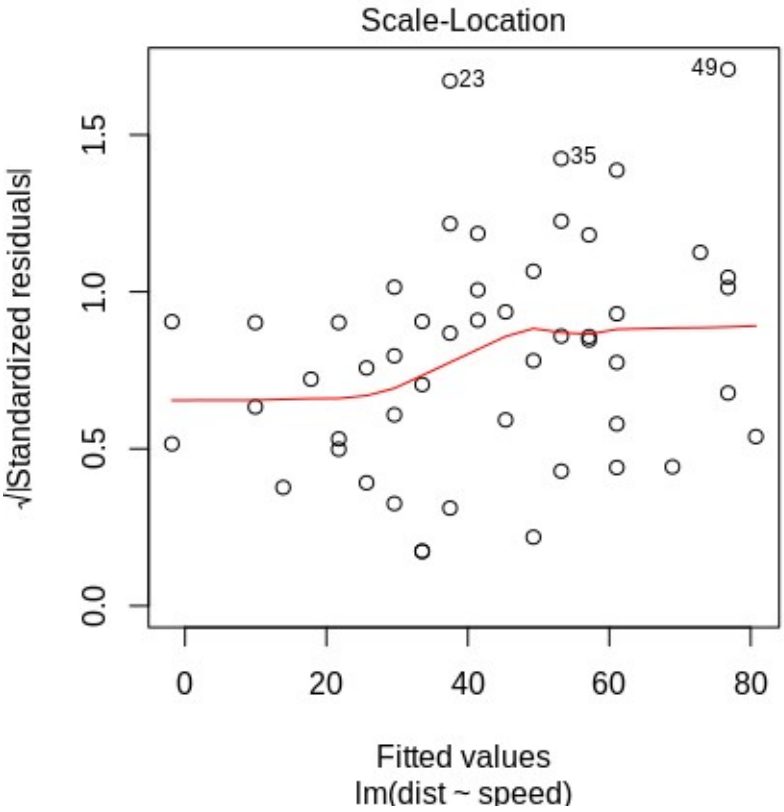
جدول ANOVA :

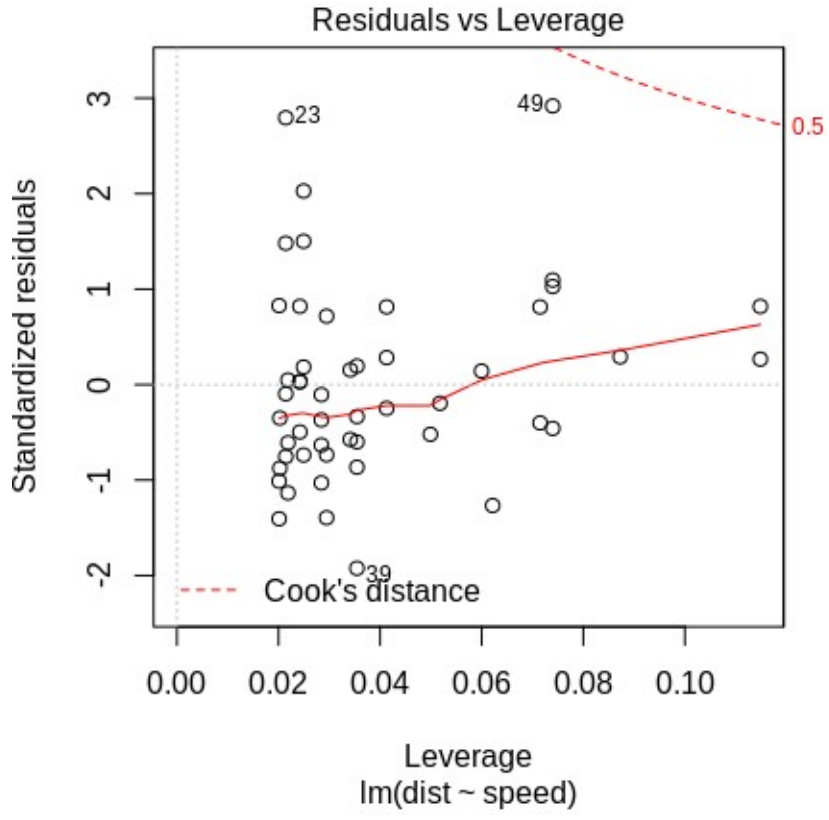
```
> anova(linearMod)
Analysis of Variance Table

Response: dist
          Df Sum Sq Mean Sq F value   Pr(>F)
speed      1  21186 21185.5  89.567 1.49e-12 ***
Residuals 48  11354   236.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

نمودارهای مدل رگرسیون:






Scale-Location
√|Standardized residuals|
23
49
35
0.0
0.5
1.0
1.5
0
20
40
60
80
Fitted values
lm(dist ~ speed)


Residuals vs Leverage
Standardized residuals
23
49
39
0.5
Cook's distance
-2
-1
0
1
2
3
0.00
0.02
0.04
0.06
0.08
0.10
Leverage
lm(dist ~ speed)

درانتها قصد به کاربرد مهم رگرسیون خطی که همان پیش بینی داده هاست پرداخته می شود:

```
> new_datas <- rnorm(20, mean=mean(cars$speed))
> predict(linearMod, data.frame(speed = new_datas))
       1        2        3        4        5        6        7
34.35867 43.48392 49.93557 46.86654 52.71253 36.48363 47.68507
       8        9       10       11       12       13       14
43.75459 41.49359 46.87976 41.50374 34.72181 44.99655 42.83010
      15       16       17       18       19       20
42.86343 35.63866 49.44899 39.24711 39.54567 44.36838
```

همان طور که مشاهده می شود با استفاده از rnorm،
۲۰ داده جدید با میانگینی برابر با میانگین داده های دیتاست ساخته شد و با استفاده از دستور predict و مدل رگرسیونی که ساخته شده بود، مقدار dist برای این ۲۰ داده جدید پیش بینی شد.