Additional topic in LM

Hossein Vatani

March 31, 2016

بنام خدا

بررسی هایی بیشتر در مورد نمونه خطی ساده(Linear Regression Model)

دراین قسمت در مورد موضوعات بیشتری درمورد نمونه ساده خطی بحث خواهیم کرد. ابتدا بررسی نمونه خطی را از منظر تحلیل پراکنش(Analysis Of Variance) آغاز خواهیم کرد. این نوع نگاه و بررسی زمانی رخ نمایان می کند که قصد بررسی و تحلیل پسرفتهای چندگانه(Multiple Regressions) را داشته باشیم. پس از آن در مورد برازشی (fitting) از خط نمونه پسرفت که از مبدا آغاز می شود بحث خواهیم کرد.

الف-پسرفت و تحلیل پراکنش

تحلیل پراکنش، خلاصه ای از اطلاعات در مورد منابع مختلف از داده ها را در اختیار می گذارد. در تحلیل پراکنش مااختلاف بین مقادیر مشخص از متغیر پاسخ با میانگین خودش را بررسی می کنیم.

اما در نمونه پسرفت این تغییرپذیری وابسته است به:

```
۱-فاصله خط پسرفت تا میانگین:اختلافی که توسط نمونه مشخص می شود

۲-فاصله مشاهدات تا خط پسرفت:گونه هایی که توسط نمونه مشخص نمی شوند
```

این تجزیه و تحلیل با کمک دستور anova() که برروی نتیجه دستور lm() اعمال می شود، قابل دستیابی است.

مثال-یک پزشک کودک برای بررسی نظریه خود مبنی بر ارتباط بین قد کودک و اندازه دورسر آنها اطلاعات زیر را جمع اوری کرده است.

```
Height = c(27.75, 24.5, 25.5, 26, 25, 27.75, 26.5, 27, 26.75, 26.75, 27.5)
Circ = c(17.5, 17.1, 17.1,17.3, 16.9, 17.6, 17.3, 17.5, 17.3, 17.5, 17.5)
Dat = data.frame(Height,Circ)
Dat
```

```
##    Height Circ
## 1   27.75 17.5
## 2   24.50 17.1
## 3   25.50 17.1
## 4   26.00 17.3
## 5   25.00 16.9
## 6   27.75 17.6
## 7   26.50 17.3
## 8   27.00 17.5
## 9   26.75 17.3
## 10  26.75 17.5
## 11  27.50 17.5
```

پس از آماده سازی اطلاعات، ابتدا نمونه پسرفت خطی را برای ان ایجاد و سپس تحلیل پراکنش را انجام می دهیم. *توجه شود که در رابطه اول در قسمت عاملهای تشریح ما علامت . را گذاشته ایم که اینکار به تابع lm می گوید تمام داده های موجود(در واقع همان ستون های قاب-داده) بجز آنکه بعنوان متغیر پاسخ است را در ساخت نمونه خطی پسرفت داخل کن.*

```
results = lm(Circ ~ .,Dat)
anova(results)
```

```
## Analysis of Variance Table
##
## Response: Circ
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## Height     1 0.39993 0.39993  43.958 9.59e-05 ***
## Residuals  9 0.08188 0.00910
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

با توجه به خروجی تحلیل پراکنش مقدار عامل F-Statistic برابر با ۴۳.۹۵۸ و همینطور مقدار P-Value برابر با ۹.۵۹e-۵ است که دلیل محکمی بر رد فرض صفر(مبنی بر صفر بودن میانگین) می باشد. مقدارهای فوق را با دستوری که در متن قبلی آموخته ایم مقایسه می کنیم.

```
summary(results)
```

```
##
## Call:
## lm(formula = Circ ~ ., data = Dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16148 -0.05842 -0.01831  0.06442  0.12989
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.49317    0.72968   17.12 3.56e-08 ***
## Height       0.18273    0.02756    6.63 9.59e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09538 on 9 degrees of freedom
## Multiple R-squared:  0.8301, Adjusted R-squared:  0.8112
## F-statistic: 43.96 on 1 and 9 DF,  p-value: 9.59e-05
```

بوضوح می بینیم که یک نتیجه را برای ما دارند.

مربع همبستگی(که با رابطه\(r^2=SSR/SST\) که نسبت مجموع مربعات پسرفت به مجموع مربعات کل محاسبه می شود) اندازه اختلاف مقادیر ارائه شده توسط نمونه پسرفت نسبت به مقدار واقعی را مشخص می کند. با توجه به نتیجه بالا مقدار \(r^2\) (که در خروجی بصورت "Multiple R-Squared"نمایش داده شده است)، مقدار دقت جواب در تعیین متغیر پاسخ با توجه به متغیر(های) تشریحی در نمونه برابر ۸۳٪ می باشد.

ب-نمونه پسرفت بشرط عرض از مبدا!

در برخی از موارد ما نیاز داریم خط نمونه پسرفت را باشرط شروع از مبدا ایجاد کنیم. برای مثال: می خواهیم مسافت پیموده شده در یک مسافرت را بصورت تابعی از زمان بیان کنیم؛مسلما دراین حالت مقدار طی شده در زمان صفر برابر صفر می باشد. در این حالت نمونه پسرفت را نیاز است که با رابطه \(Y_i=\beta_iX_i+\epsilon_i\)بیان کنیم که \(\epsilon_i\)ها مستقل و با توزیع \(N(0,\sigma)\) می باشند. در این نمونه از رابطه، پسرفت لزوما از نقطه صفر شروع خواهد کرد. برای ایجاد نمونه ای که از مبدا آغاز شود، بسادگی کافیست که از یکی از روابط زیر در R استفاده نماییم.

```
lm(response ~ 0 + explanatory)
lm(response ~ explanatory -1)
```

مثال-یک کارخانه تولید لوله های خانگی که دوازده واحد تولید دارد، می خواهد بداند که رابطه بین هر واحد و میزان دستمزد کل کارگران واحد چگونه است.هزینه ها برحسب میلیون تومان می باشد.

```
Units = c(20, 196, 115, 50, 122, 100, 33, 154, 80, 147, 182, 160)
Cost = c(114, 921, 560, 245, 575, 475, 138, 727, 375, 670, 828, 762)
Dat = data.frame(Units,Cost)
Fit = lm(Cost ~ Units -1)
Fit
```

```
##
## Call:
## lm(formula = Cost ~ Units - 1)
##
## Coefficients:
## Units
## 4.685
```

با توجه به خروجی، مقدار \(\hat y=۴.۶۸۵x\) حاصل می شود که بیانگر اینست که برای هر واحد اضافی حدود ۴میلیون و ۶۸۵ هزار تومان نیاز است.

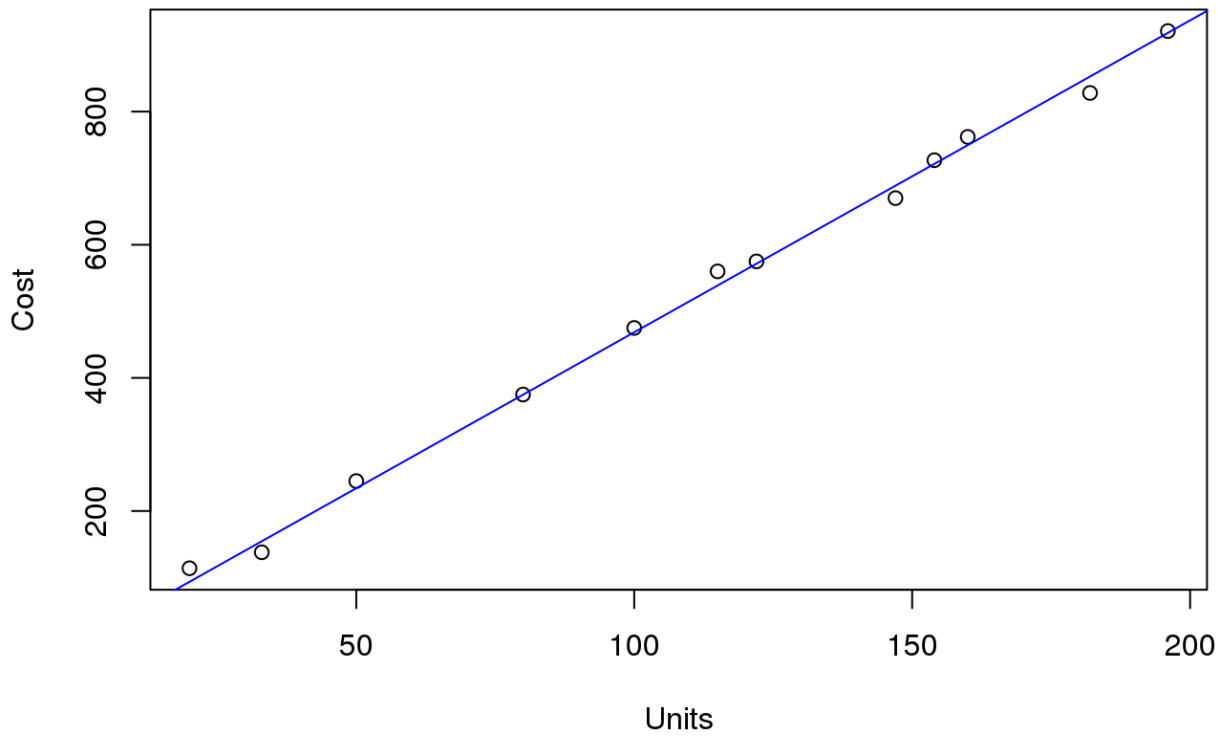
اگر بخواهیم بازه اطمینان را از ۹۵٪ به ۹۰٪ تغییر بدهیم، می توانیم از دستور confit() استفاده کنیم.

```
confint(Fit,level=0.90)
```

```
##            5 %     95 %
## Units 4.623846 4.746702
```

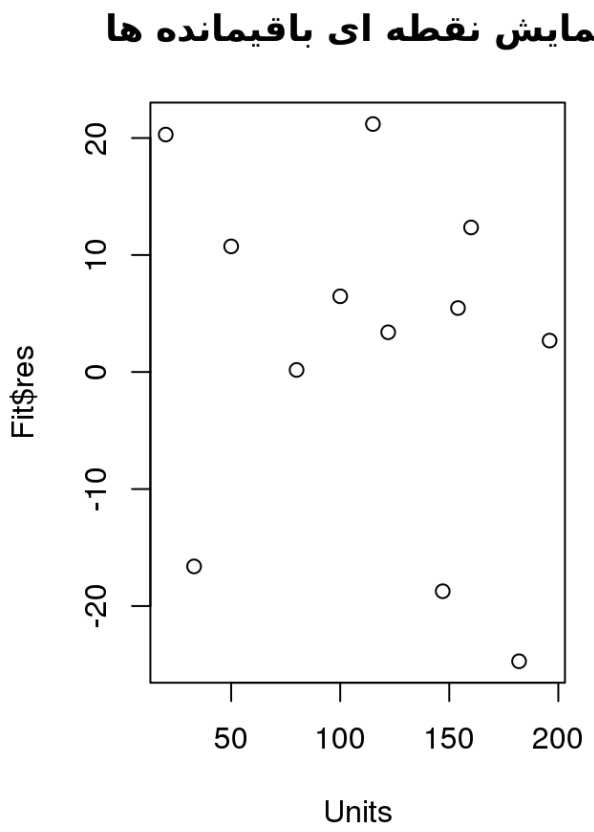
جهت نمایش تصویری داده های فوق با خط نمونه پسرفت (همانطور که در مقاله قبل نیز نشان دادیم) می توان از دستورهای زیر استفاده کرد.

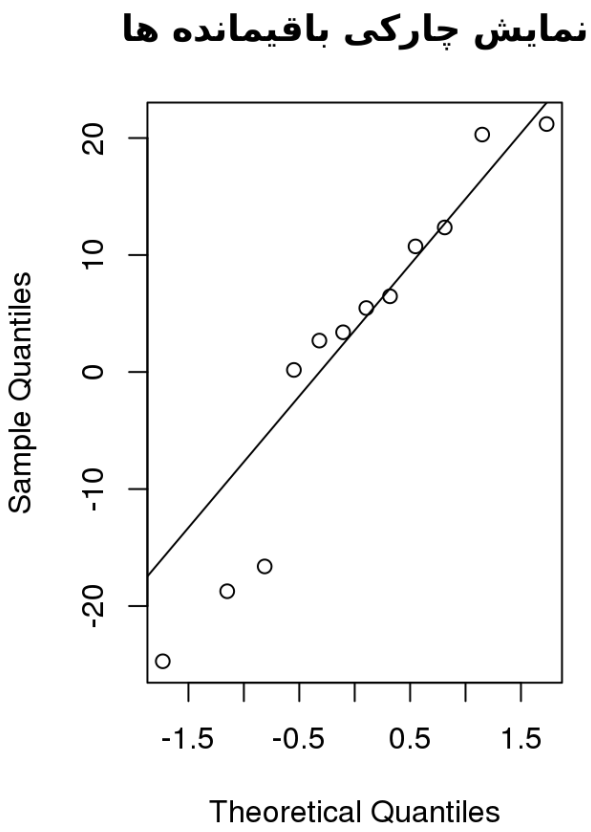
```
plot(Units,Cost)
abline(Fit,col="blue")
```



جهت بررسی قوت نمونه ایجاد شده، نمودار باقیمانده ها را نیز بررسی می کنیم.

```
par(mfrow=c(1,2)) #این دستور صفحه نمایش را به یک سطر و دو ستون جهت نمایش دو نمودار کنار هم تقسیم می کند .
plot(Units,Fit$res,main = "نمایش نقطه ای باقیمانده ها")
qqnorm(Fit$res,main = "نمایش چارکی باقیمانده ها")
qqline(Fit$res)
```






```
par(mfrow=c(1,1)) # حالت خروجی صفحه نمایش را به حالت قبل برمیگردانیم.
```

نمایش نقطه ای باقیمانده ها نسبت به متغیر تشریحی گویای اتفاقی بودن توزیع آنهاست و نمایش چارکی بوضوح نشان از نرمال بودن توزیع آنها دارد.البته در نمودار چارکی، سه باقیمانده وجود دارد که نسبت به سایرین از خط فاصله بیشتری دارد، اما احتمالا نه آنقدر که برروی مدل تاثیر جدی بگذارد. جهت بررسی اهمیت شیب بدست آمده، از دستور summary() استفاده می کنیم.

```
summary(Fit)
```

```
##
## Call:
## lm(formula = Cost ~ Units - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.720  -4.020   4.432  11.141  21.194
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## Units  4.68527    0.03421     137   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.95 on 11 degrees of freedom
## Multiple R-squared:  0.9994, Adjusted R-squared:  0.9994
## F-statistic: 1.876e+04 on 1 and 11 DF,  p-value: < 2.2e-16
```

باتوجه به مقدار P-Value بدست آمده با احتمال بالایی می توان اظهارنظر کرد که بین واحد تولید و هزینه نیروی کار ارتباط مستقیم وجود دارد.

حال فرض کنیم که می خواهیم در بازه اطمینان ۹۰٪ بدانیم که میانگین هزینه برای ۱۰۰ واحد چطور خواهد بود.

```
predict(Fit, data.frame(Units =100), interval="confidence", level=0.90)
```

```
##        fit      lwr      upr
## 1 468.5274 462.3846 474.6702
```

در بازه اطمینان ۹۰٪ مقدار ۴۶۸.۵۲۷ در بازه (۴۶۲.۳۸۴,۴۷۴.۶۷) پیشنهاد می شود.حال اگر بازه اطمینان ۹۰٪ بخواهیم برای دقیقا ۱۰۰ واحد پیش بینی کنیم که چه مقدار هزینه باید در نظر بگیریم.

```
predict(Fit,data.frame(Units =100),interval="prediction",level=0.90)
```

```
##        fit      lwr     upr
## 1 468.5274 440.9898 496.065
```