Model Diagnostic for regression

Hossein Vatani

April 4, 2016

بنام خدا

عیب شناسی مدل پسرفت(Model Diagnostics for Regression)

پس از برازش نمونه پسرفت(fittung a regression model) ، یکی از نکات مهم بررسی معتبر و معنی دار بودن همه فرض های در نظرگرفته شده در نمونه پسرفت می باشد که باید قبل از تحلیل و استنتاج انجام شود. اگر خطایی در نمونه وجود داشته باشد، ممکن است تحلیل های بعدی دچار خطا باشند و فاقد اعتبار. لذا مشخص کردن نمونه (Model) صحیح بسیار مهم است.

در نمونه ایجاد شده فرض بر آنست که متغیر y با متغیرهای β (یا همان x ها) بصورت خطی رابطه دارد و باقیمانده ها (ی حاصل از نمونه) بصورت مستقل ، با توزیع نرمال $N(0,\sigma^2)$ می باشد.

روش خطایابی نمونه هر دو ابزار نمایش تصویری(graphical method) و آزمون های آماری را بکار می گیرد. این ابزار و روش به ما کمک می کند که بررسی کنیم که آیا پیش فرض های مدل صحیح هستند و آیا ما می توانیم به نتایج حاصله از تحلیل آنها در آینده اعتماد کنیم؟

مثال-(فایل موجود نیست) برای این مثال از داده های مربوط به خودرو که با نام mtcars در نرم افزار وجود دارد استفاده می کنیم. (*توضیح عاملها را در راهنمای نرم افزار R را می توانید بیابید.*) نمونه خطی چندگانه خود را برای عامل mpg (که در مثال قبل در مقاله قبلی عاملهای مناسب مربوط به آنرا پیدا کرده بودیم) می سازیم.

```
Car=mtcars[,c("mpg", "disp", "hp", "wt", "am")]
head(Car)
```

```
##                   mpg disp  hp    wt am
## Mazda RX4         21.0  160 110 2.620  1
## Mazda RX4 Wag     21.0  160 110 2.875  1
## Datsun 710        22.8  108  93 2.320  1
## Hornet 4 Drive    21.4  258 110 3.215  0
## Hornet Sportabout 18.7  360 175 3.440  0
## Valiant           18.1  225 105 3.460  0
```

```
Fit2=lm(mpg ~ hp+ wt, data = Car)
Fit2
```

```
##
## Call:
## lm(formula = mpg ~ hp + wt, data = Car)
##
## Coefficients:
## (Intercept)           hp           wt
##    37.22727     -0.03177     -3.87783
```

الف-بررسی متغیرها

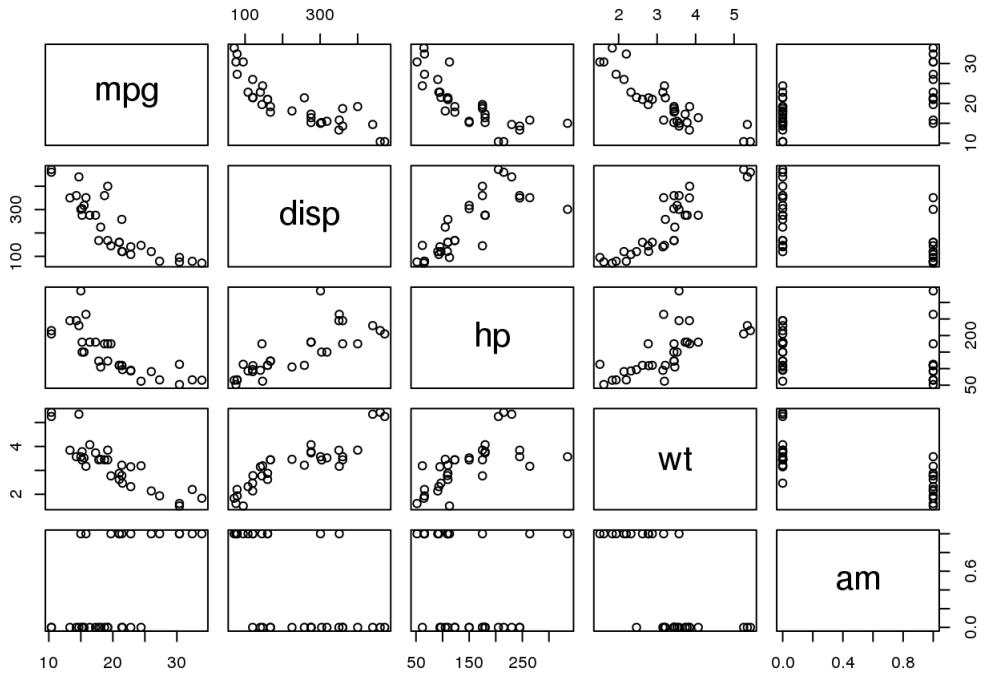
چند روش تصویرسازی مناسب جهت بررسی رابطه متغیر پاسخ با متغیرهای تشریحی وجود دارد. اولین موضوع مورد نظر تعیین محدوده برای متغیرهای تشریحی است و اینکه آیا مقادیری از آنها وجود دارد که خارج از محدوده(Outlier) باشد؟

نمایش تصویری می تواند بصورت ملیه ای(histogram) ویا جعبه ای(boxplot) باشد.البته نوع نمایش نقطه ای تمام عامل ها راکه در گذشته از ان استفاده نموده ایم نیز کاربرد دارد.

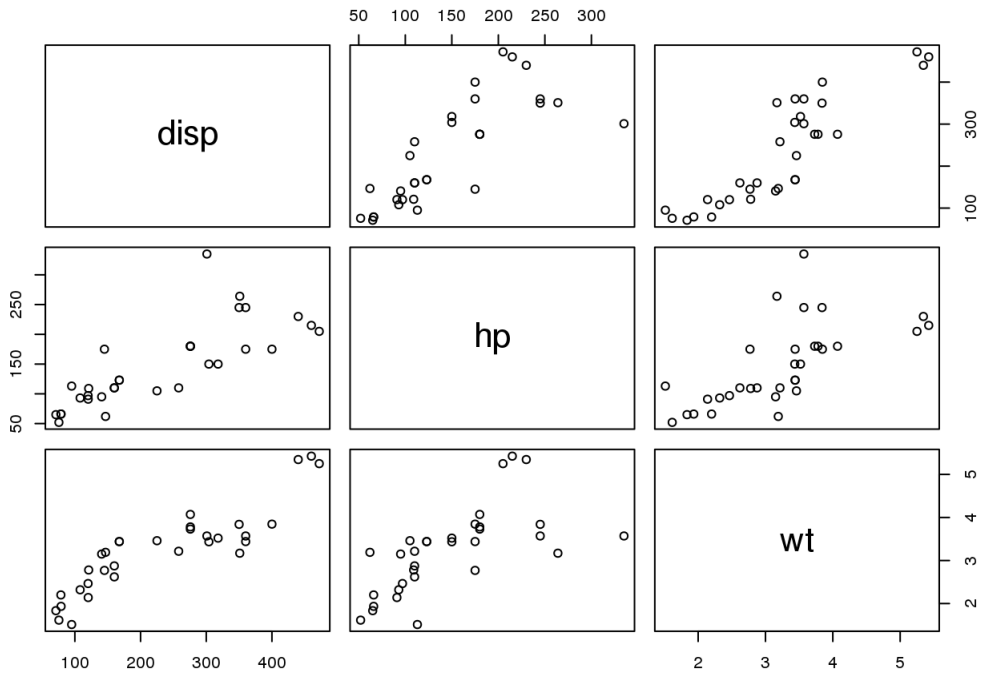
جهت یاد آوری: بدوصورت می توان درواقع نمایش ضربدری متغیرها را داشت ۱-تمام متغیرها و۲-متغیرهای مد نظر، که در زیر آنها را

نشان داده ایم.

```
plot(Car)
```




```
pairs(~disp + hp + wt,data = Car)
```



ب-باقیمانده ها و اهرم(Resiuals and Leverage)

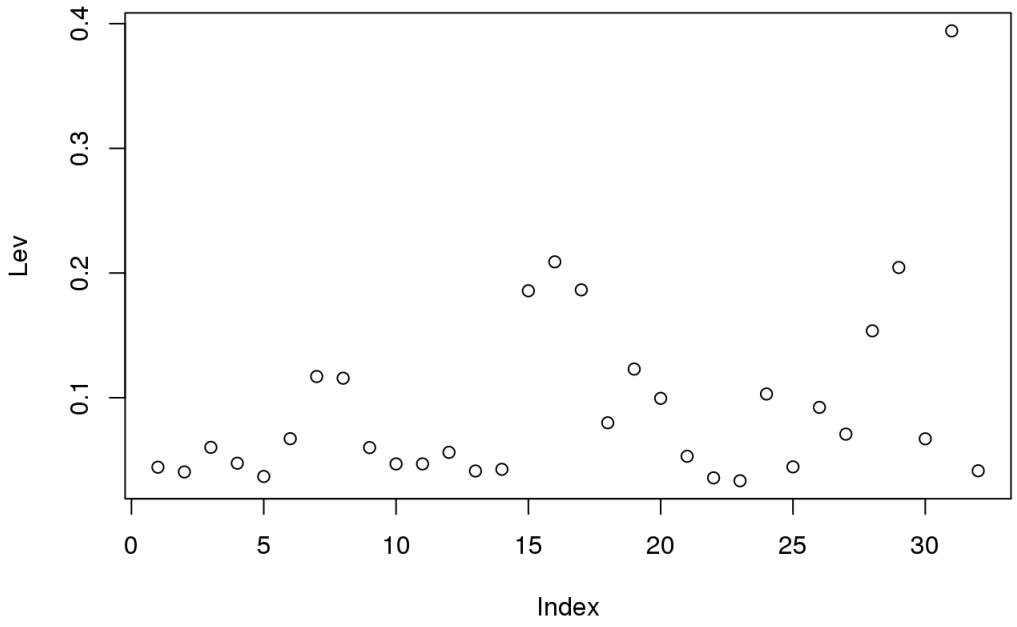
قدرت نفوذ یا همان اهرم، مقداری است که توانایی متغیر تشریح در حرکت و تغییر مقدار نمونه پسرفت(Model Regression) با حرکت

خود در جهت نمودار y(همان نمودار نتغیر پاسخ) دارد. مقدار اهرم یک متغیر، مقدار تغییر نمونه پسرفت به ازای تغییر یک واحد متغیر تشریحی مورد نظر می باشد.

این مقدار، عددی است بین صفر و یک که اگر اندازه آن صفر باشد بدان معناست که تغییر آن بر نمونه بدون تاثیر و اگر یک باشد یعنی نمونه بطور کامل تحت تاثیر این عامل قرار دارد.

برای نمایش تصویری اهرم متغیرها باید از نتیجه حاصله از نمونه ایجاد شده در رابطه زیر استفاده نماییم.

```
Lev=hat(model.matrix(Fit2))
plot(Lev)
```



همانطور که از نمودار مشخص است، در اینجا یک نقطه وجود دارد که قدرت نفوذ(همان تاثیر اهرمی) بیشتری در نمونه ایجاد شده دارد که لازم است بررسی بیشتری برروی آن انجام دهیم. باتوجه به نمودار می توان متوجه شد که مقدار بیش از ۰.۳ اختیار کرده است.

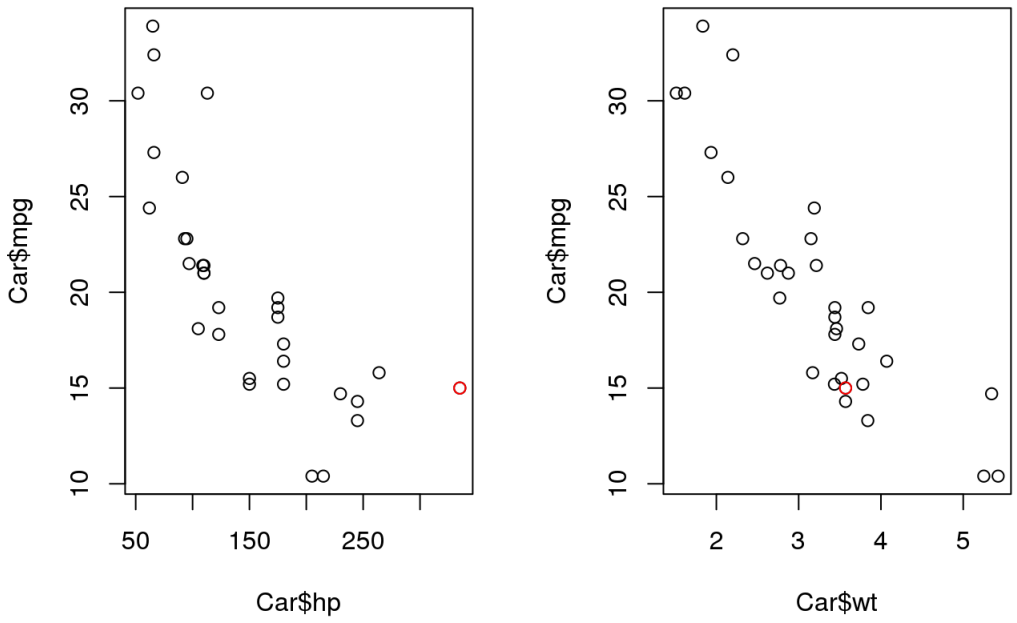
```
Car[Lev>0.3,]
```

```
##               mpg disp  hp   wt am
## Maserati Bora  15  301 335 3.57  1
```

جالب نیست؟ خودروی مازراتی دارای نفوذی زیاد است :-)

حال جهت بررسی دقیق تر می توان نمودارهای قبلی را با مشخص کردن این نقطه خاص دوباره رسم کرد.

```
par(mfrow=c(1,2))
plot(Car$hp,Car$mpg)
points(Car["Maserati Bora",]$hp,Car["Maserati Bora",]$mpg,col="red") #col~color
plot(Car$wt,Car$mpg)
points(Car["Maserati Bora",]$wt,Car["Maserati Bora",]$mpg,col="red") #col~color
```




```
par(mfrow=c(1,1))
```

نقطه مورد نظر را با رنگ قرمز جهت تشخیص راحت، مشخص کرده ایم.

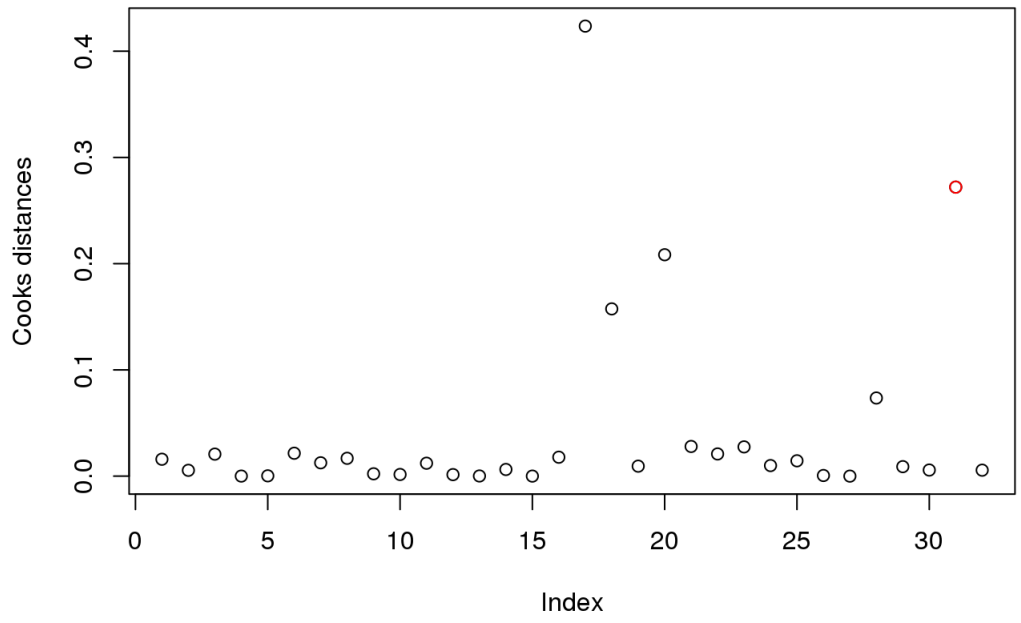
درادامه می خواهیم(ولازم است که جهت بهتر کردن نمونه پسرفت ایجاد شده)این نقطه را از نمونه(در واقع داده اولیه)حذف نماییم ولی ابتدا ضروری است قدرت نفوذ(همان تاثیر) این نقطه در نمونه را بدانیم.

بطور معمول، قبل از مطالعه باقیمانده ها، آنهارا جهت بررسی درست صرفنظر از توان اهرمی(شان) استاندارد می کنند که اینکار با دستور rstudent() انجام می شود که دراصطلاح به آن استیودنتیده کردن می گویند.

```
Res=rstudent(Fit2)
```

زمانی که یک نقطه با نفوذ از نمونه ای حذف شود، ممکن است نمونه را با تغییر معنی داری مواجه کند. یک نقطه با نفوذ ممکن است که یا مقداری پرت داشته باشد و یا مقداراهرمی بالایی ویا هردواولی حتما یکی از ایندو را دارد. عدد Cook Distance حاصلضرب ایندو را مشخص می کند که براحتی در R قابلمحاسبه می باشد.

```
Cook = cooks.distance(Fit2)
plot(Cook,ylab="Cooks distances")
points( which(rownames(Car)=="Maserati Bora"),Cook["Maserati Bora"],col='red')
```



توجه شود که "Maserati Bora" یک مقدار ازستون داده نیست و نام یک ردیف در داده های مورد کاووش می باشد.با دستور which بنحوی که مورد استفاده قرار گرفته است، شماره سطر را پیدا می کنیم

```
Car[Cook>0.5,]
```

```
## [1] mpg  disp hp   wt   am
## <0 rows> (or 0-length row.names)
```

```
Car2=Car[rownames(Car) != ("Maserati Bora"),]
```

حال ما یک قاب-داده(Dataframe) جدید بدون آن نقطه پرت داریم.

ج-نمودار باقیمانده ها

با مطالعه باقیمانده ها، می توانیم به دریافت های زیر نائل شویم که آیا:

۱-تابع پسرفت خیر خطی است؟

۲-جملات خطا پراکنش(regression) غیر ثابت دارند؟

۳-جملات خطا مستقل نیستند؟

۴-آنها مقادیری پرت دارند؟

۵-جملات خطا از توزیع نرمال پیروی نمی کنند.

تخطی باقیمانده ها از هریک از موارد فوق را می توان با رسم نمودارهای مناسب که شامل:

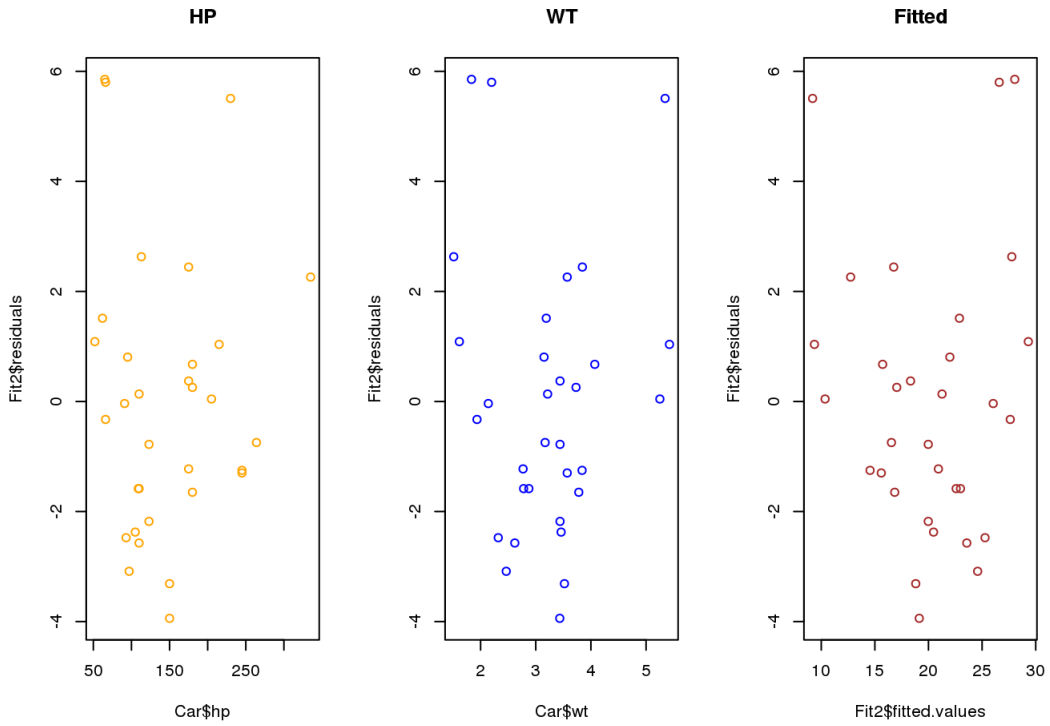
الف-نمودار باقیمانده ها با مقدارهای تشریح یا مقداربرازش شده(fitted values)

ب-نمودار میله ای یا جعبه ای باقیمانده ها

ج-نمایش نقطه ای توزیع نرمال از باقیمانده ها

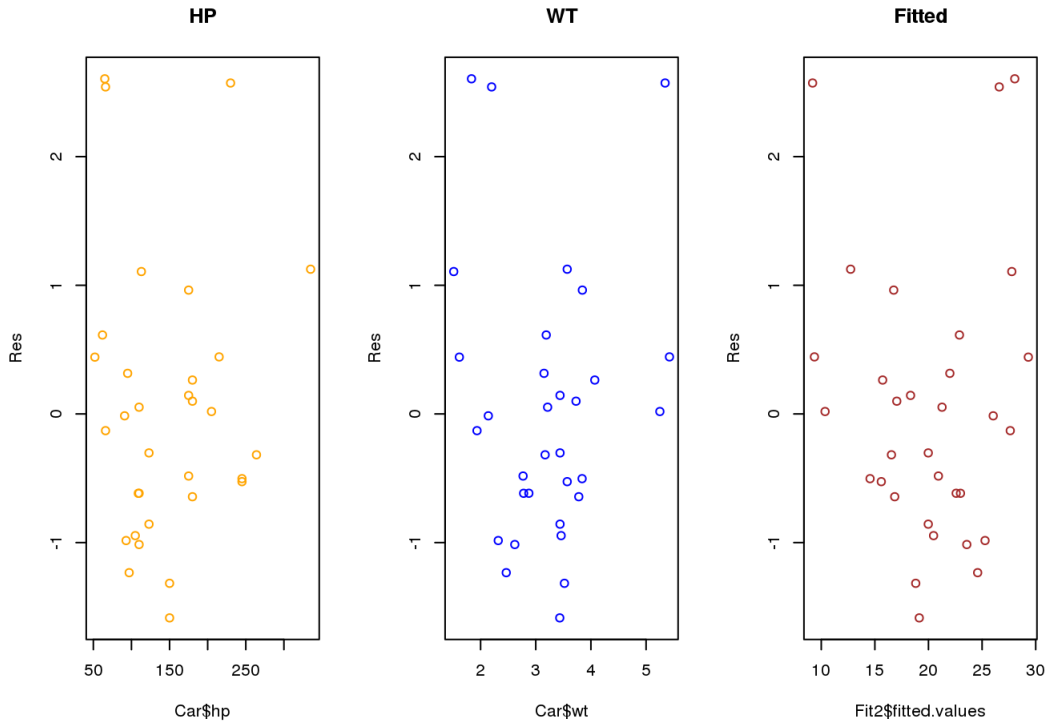
برای نمایش (الف) باتوجه به نمونه مورد نظر که دو متغیر تشریح داریم، سه نمودار مطلوب خواهد بود.

```
par(mfrow=c(1,3))
plot(Car$hp,Fit2$residuals,main = "HP",col="orange")
plot(Car$wt,Fit2$residuals,main = "WT",col="blue")
plot(Fit2$fitted.values,Fit2$residuals,main = "Fitted",col="brown")
```



همچنین جهت ایجاد همان نمودارها بعد از استیدنتیده کردن باقیمانده ها داریم:

```
par(mfrow=c(1,3))
plot(Car$hp,Res,main = "HP",col="orange")
plot(Car$wt,Res,main = "WT",col="blue")
plot(Fit2$fitted.values,Res,main = "Fitted",col="brown")
```



و در نهایت نمودار نرمال احتمال باقیمانده ها جهت بررسی شرط نرمال بودن باقیمانده ها قابل ایجاد است. ما آنرا بدو صورت نقطه ای و میله نمایش خواهیم داد.

```
par(mfrow=c(1,2))
qqnorm(Fit2$residuals)
qqline(Fit2$residuals)
hist(Fit2$residuals)
```