Correlation And Simple Linear regression

Hossein Vatani

March 29, 2016

بنام خدا

همبستگی و پسرفت (رگرسیون) خطی ساده

یکی از مواردی که ما بطور غالب خواهان اطلاع از آن هستیم، نحوه ارتباط بین دو (یا چندین متغیر) می باشد؛اینکه آیا آنها بهم ارتباط دارند و اگر آری، ارتباط آنها چگونه هست.در واقع می خواهیم بدانیم که تغییر در متغیر x (که آنرا متغیر توضیح می نامیم) چه تغییری در متغییر y (که آنرا متغیر وابسته یا جوابگو می نامیم) ایجاد می کند.اگر نمایش نمودار نقطه ای متغیرهای x,y تداعی گر یک رابطه خطی باشد، گوییم یک *مدل خطی* بین آندو وجود دارد که اگر پراکندگی حول آن خط فرضیکم باشد مدل را قوی و در غیر ان ضعیف فرض می کنیم.

در ادامه با مثال ، موضوع روشن تر خواهد شد(انشاءاله).

هم-وردایی(CoVariance) و همبستگی(Correlation) عاملهای اندازه گیری و تشخیص جهت وابستگی دو متغیر کیفی اند که وجود ارتباط خطی بین آنها تشخیص داده شده است.خط رگرسیون یک مدل ریاضی جهت توضیح یک رابطه خطی بین متغیر تشریحی و متغیر پاسخ است که می تواند در پیشبینی متغیر پاسخ در وضعیت های جدید متغیر توضیح(یا همان تشریحی) بکار برد.

همبستگی،هم-وردایی و پسرفت(regression) همگی قابل محاسبه در R هستند.

الف-همبستگی و هم-وردایی

دستور ()cov برای محاسبه هم-وردایی و ()cor برای همبستگی.

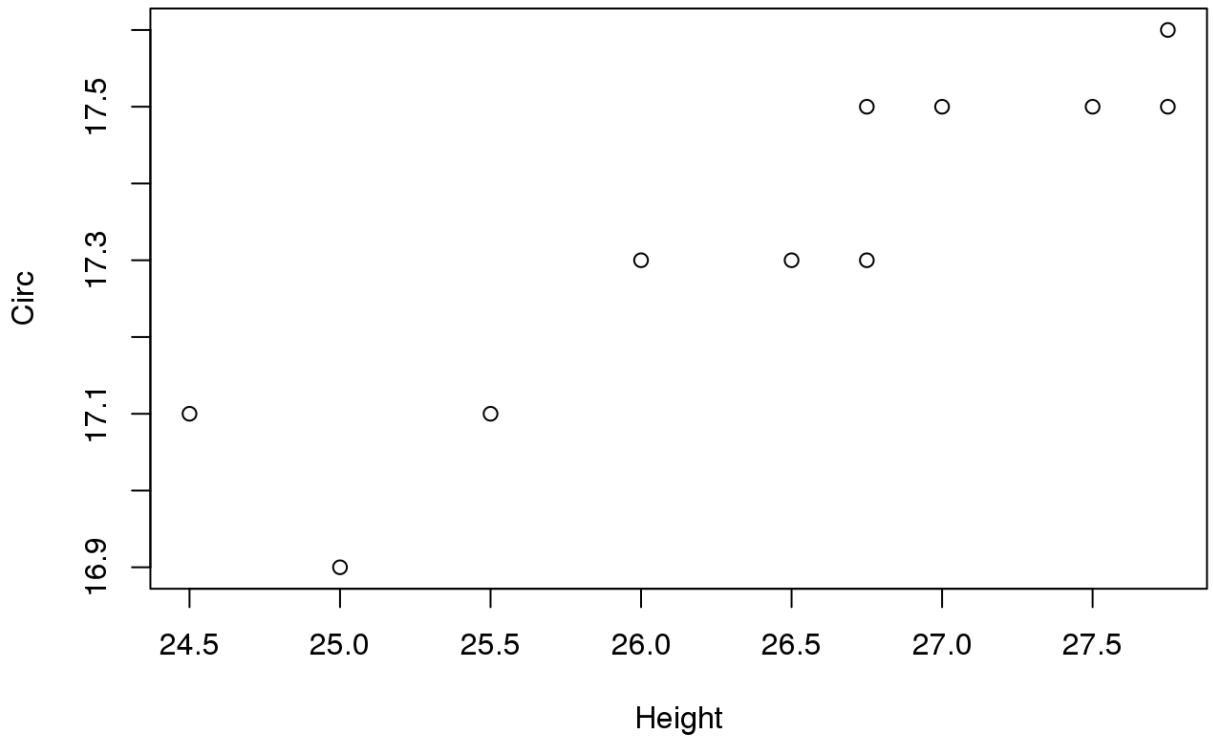
مثال-یک پزشک اطفال می خواهد بررسی کند که آیا بین قد کودک با اندازه دور سر آنها ارتباط دارد یا خیر؟ برای اینکار او آزمایش خود را روی یازده کودک سه ساله انجام داده است که اطلاعات آن بشرح زیر می باشد.

```
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
```

حال نمودار نقطه ای آنرا رسم می کنیم.

```
plot(Circ~Height)
```



ظاهر نمودار مشخص می کند که یک ارتباط خطی بین متغیر تشریح و پاسخ وجود دارد. به کمک دستور abline() می توان خط مورد نظر را جهت بررسی دقیق تر رسم کرد ولیکن چون مورد نظر این مثال نیست در ادامه به آن خواهیم پرداخت.ابتدا هم-وردایی آنها را بررسی می کنیم.

```
cov(Dat) # Covariance matrix
```

```
##           Height       Circ
## Height 1.1977273 0.21886364
## Circ   0.2188636 0.04818182
```

با توجه به خروجی ، مشخص می شود که پراکنش قد و دورسر بترتیب ۱.۱۹۸ و ۰.۰۴۸ می باشد و هم-وردایی آندو ۰.۲۱۹ که به ما می گوید ارتباط مستقیم(مثبت) باهم دارند.

```
cor(Dat) # Correlation matrix
```

```
##           Height      Circ
## Height 1.0000000 0.9110727
## Circ   0.9110727 1.0000000
```

این خروجی به ما می گوید که ارتباط بین دو متغیر بسیار بالاست و نشان بر ارتباط خطی بین آندو دارد.

یادآوری: عامل همبستگی مستقل از واحد و عددی بین ۱- تا ۱+ است

ب-مدل خطی پسرفت(رگرسیون)ساده

زمانیکه مشخص شود یک رابطه خطی قوی بین دو متغیر مورد نظر وجود دارد، می توانیم بسمت ایجاد نمونه(Model) از آن به کمک پسرفت خطی(Linear Regression) حرکت کنیم.دستور lm() پرکاربردترین دستور R دراین زمینه می باشد که عامل های متعدد و زیادی برای استفاده در آن تعریف شده است.

```
lm(formula, data, subset, weights, na.action,method = "qr", model = TRUE, x =
 FALSE, y = FALSE, qr = TRUE,singular.ok = TRUE, contrasts = NULL, offset, ..
.)
```

در ساده ترین حالت می توان بصورت زیر از آن استفاده کرد.

```
lm(response ~ explanatory)
```

مثال-خط توصیف کننده نمونه پسرفت خطی(regression line fitted) را برای داده های بالا محاسبه می کنیم.

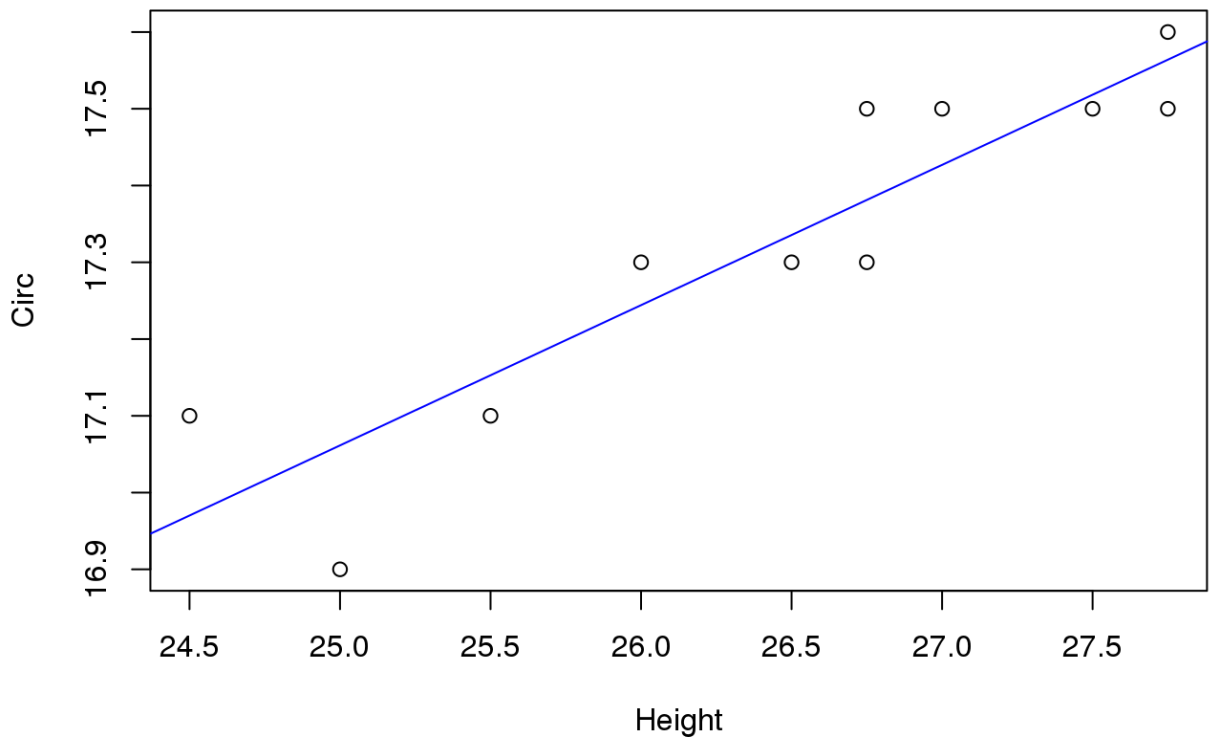
```
Fit= lm(Circ ~ Height)
Fit
```

```
##
## Call:
## lm(formula = Circ ~ Height)
##
## Coefficients:
## (Intercept)       Height
##     12.4932       0.1827
```

این عددها رابطه ۰.۱۸۳Height+۱۲.۴۹۳ را به ما جهت پیش بینی اندازه دور سر با توجه به قد کودک(سه ساله) با خطای ۰.۰۵ درصد پیشنهاد می دهد.

اکنون می توان خط مورد نظر را که در قبل صحبت شده بود در نمودار نقطه ای متغیرها نمایش داد.

```
plot(Height,Circ)
abline(Fit,col="blue")
```



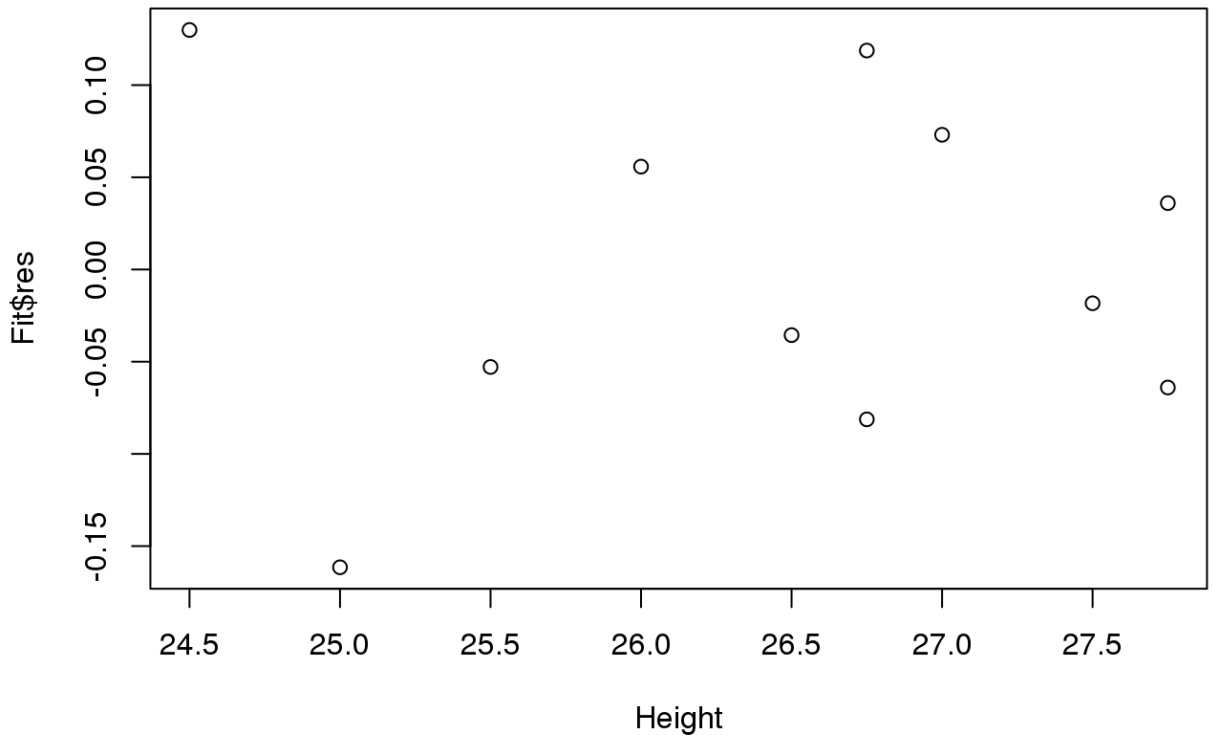
مرحله بعد دراین تحلیل، بررسی درستی تمام فرضیات مربوط به یک نمونه پسرفت (رگرسیون) خطی ساده است:

-باقیمانده (Residual) باید دارای توزیع نرمال با واریانس برابر برای هر مقدار متغیر تشریحی باشد.

توجه: مقدار باقیمانده(یا همان خطا!) با دستور Fit$residual قابل دسترسی است.

-در این قسمت ، ابتدا نمودار باقیمانده ها بر حسب متغیر تشریحی را بررسی می کنیم که آیا توزیع آنها حول محور ۰ نسبت بهم اتفاقی(randomly) می باشد یا خیر؟

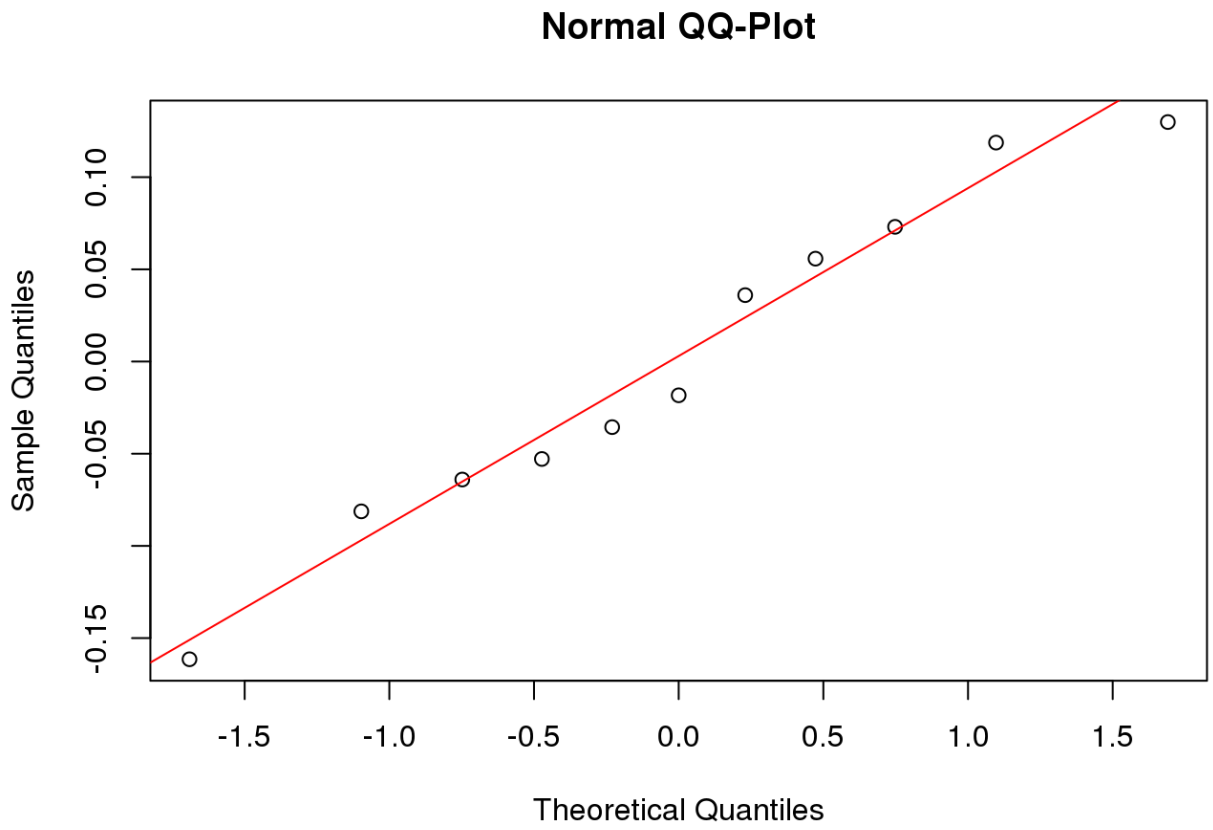
```
plot(Height,Fit$res)
```



همانطور که از تصویر آشکار است، نمودار گویای هیچ توزیع مشخصی نیست و لذا این مورد (فاقد توزیع بودن باقیمانده نسبت به متغیر تشریحی) برقرار می باشد.

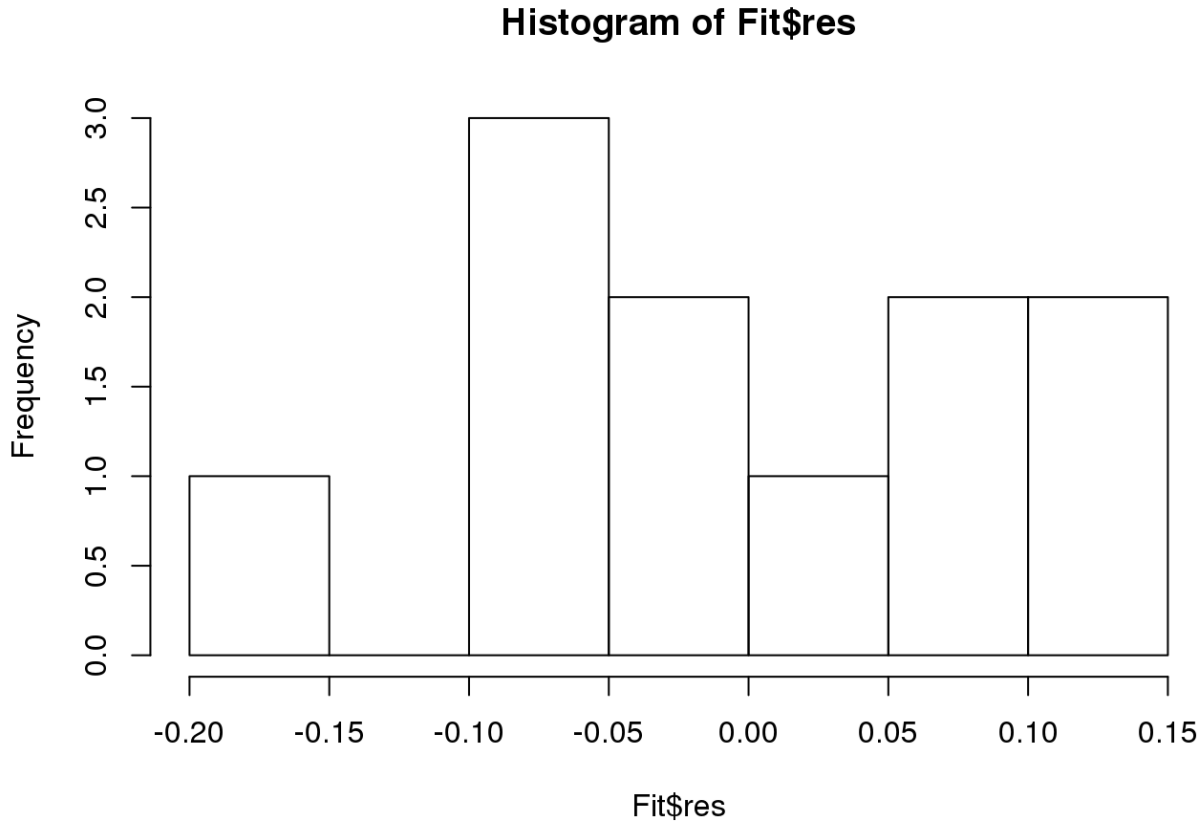
-بررسی نرمال بودن توزیع باقیمانده ها. برای اینکار از رسم چارکی(QQ-Plot) استفاده می کنیم.

```
qqnorm(Fit$res,main = "Normal QQ-Plot")
qqline(Fit$res,col="red")
```



و همچنین جهت نمایش ارتفاع-گونه(histogram) باقیماندهها

```
hist(Fit$res)
```



پس از بررسی درستی فرضیات مورد نیاز، مرحله بعد استنتاج از نتیجه نمونه (Model) ساخته شده است. برای شیب و عرض از مبدا بدست آمده از نمونه(Model) ، جهت بدست آوردن بازه اطمینان برای میانگین متغیر پاسخ و همچنین پیش بینی وضعیت های جدید در اینده باید آزمایش ها و بازه های اطمینان هایی را بوجود آوریم. برای این موضوع از دستور summary استفاده میکنیم.

```
summary(Fit)
```

```
##
## Call:
## lm(formula = Circ ~ Height)
##
## 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
```

مقدار P-Value به ما می گوید که فرض صفر رد می شود و این بدان معتی است که بین قد و اندازه دور سر ارتباط وجود دارد. اندازه عامل "Residual standard error" به ما، میزان انحراف از معیار حول خط نمونه پسرفت خطی(Linear Regression Model) ساخته شده را می دهد که برابر ۰.۰۹۵۳۸ می باشد.

برای ایجاد بازه اطمینان ۹۵٪ برای β_1 می توان از دستور زیر استفاده کرد.

```
confint(Fit)
```

```
##                 2.5 %     97.5 %
## (Intercept) 10.8425070 14.1438307
## Height       0.1203848  0.2450801
```

-درنهایت ما می توانیم از نمونه ایجاد شده در پیش بینی وضعیت ها در آینده استفاده کنیم. می توان میانگین اندازه دور سر کودکانی را که همه آنها قد مشخصی دارند اعلام کرد ویا اینکه برای کودک خاصی، اندازه دورسرش را با توجه به قد او پیش بینی کرد. البته که در حالت دوم پیش بینی بازه خطا بیش از حالت اول است.دستور موجود در R جهت اینکار predic() می باشد. توجه شود که برای حالت اول باید از مورد interval="confidence" استفاده کرد و برای حالت دوم از interval="prediction".

مثال-دریک بازه اطمینان ۹۵٪، میانگین اندازه دورسر کودکان با قد ۲۵ اینچی را محاسبه می کنیم.

```
predict(Fit,data.frame(Height =25),interval="confidence")
```

```
##        fit      lwr      upr
## 1 17.06148 16.94987 17.17309
```

مثال-اندازه دورسر کودکی با قد ۲۵ اینچ را نیز محاسبه می کنیم.

```
predict(Fit,data.frame(Height =25),interval="prediction")
```

```
##        fit      lwr      upr
## 1 17.06148 16.81855 17.30441
```