数据分析与处理技术——线性回归模型

商学院 徐宁

优化与回归模型

简单线性回归

基本回归模型

模型预测方法

线性回归的原理

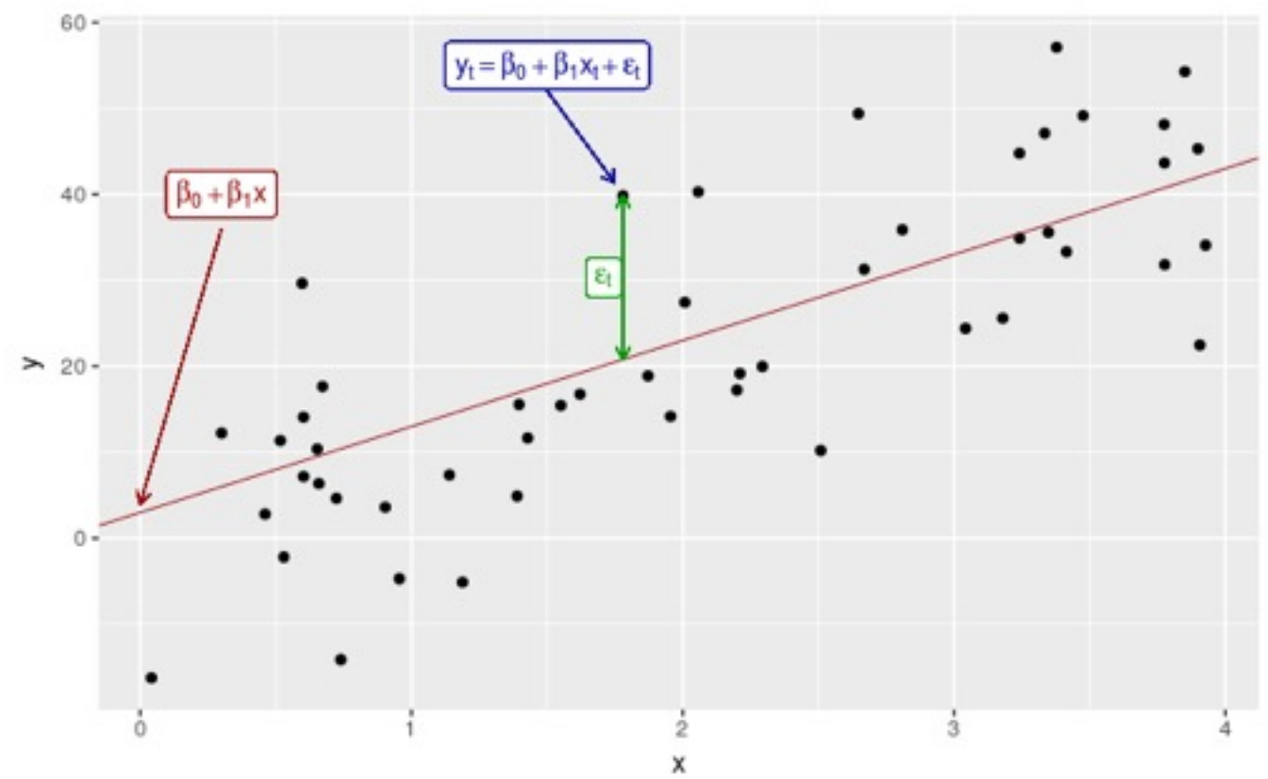
做出直线方程，斜率与截距为待定参数，问题转化为如何根据确定趋势线的两参数?

$$y = \beta_0 + \beta_1 x$$

点到直线L的离差

$$\varepsilon_i = y_t - (\beta_0 + \beta_1 x_t)$$

$$S = \sum_{i=1}^{n} \varepsilon_i^2 \to \min$$

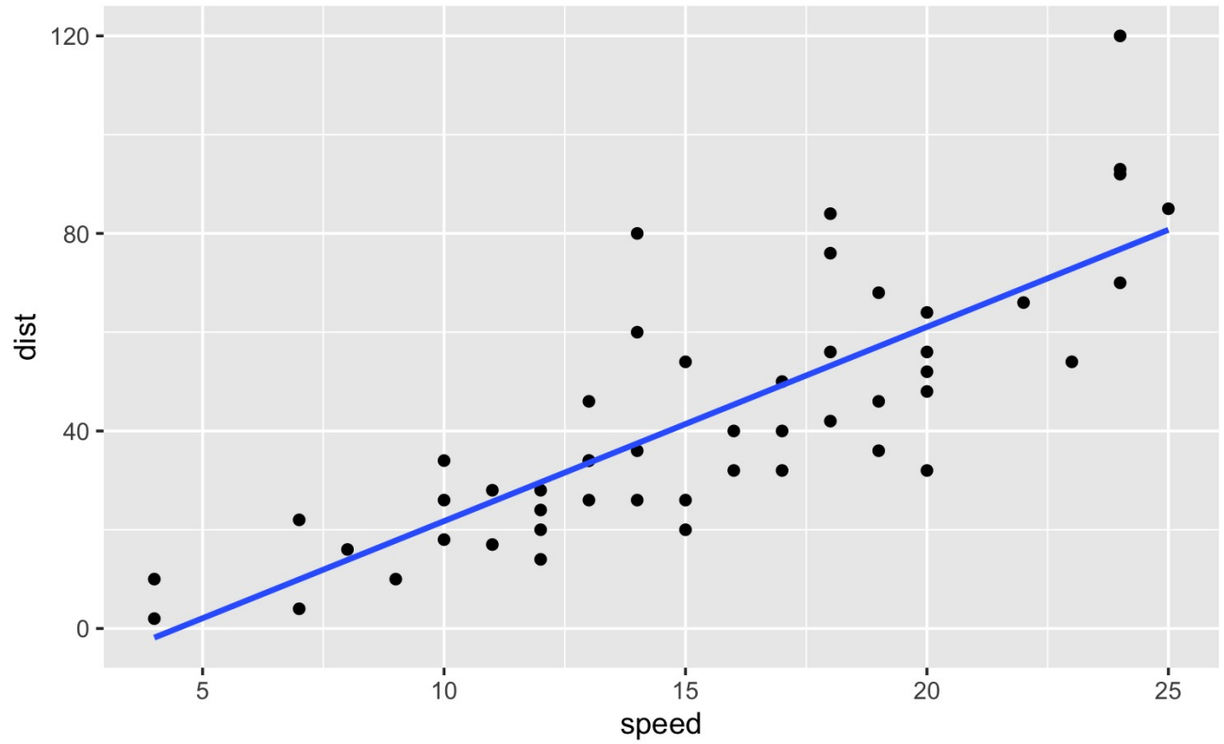


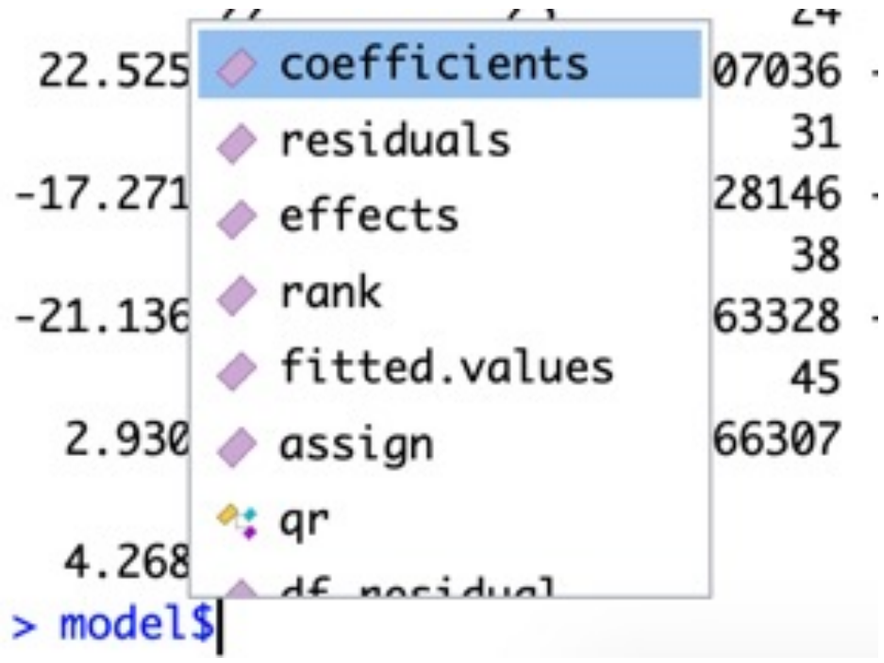
建模代码

lm()函数将线性回归计算结果存入一个变量中，以列表变量方式输出模型

模型实际是一个列表变量，装载了所有计算结果

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





可视化代码

```
ggplot(cars,aes(speed,dist))+
  geom_point()+
  geom_smooth(method="lm",se=F)
```

读取计算结果

通过数据简报方式读取计算结果

```
summary(model)
```

```
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
```

模型可决系数

$$R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

调整后的可决系数

$$\bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)}$$

区间估计

fitted和resid函数常用来提取拟合值和误差。

可预测的前提是数据分布具备正态特征，进而拟合和预测均可以估计置信区间

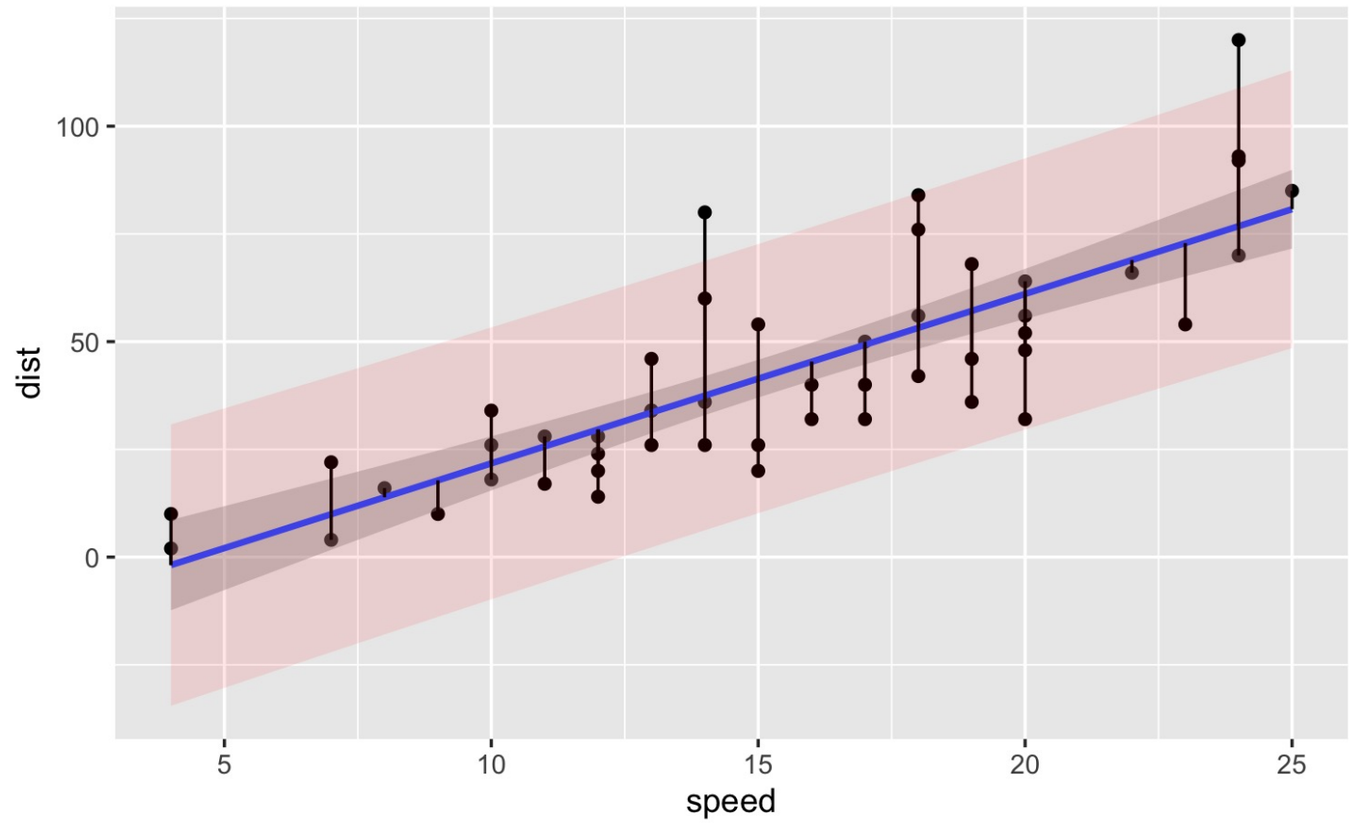
模型变量并没有给出置信区间数值，predict可以利用模型变量完成置信区间的计算，在没有给出新数据的情况下该函数计算拟合区域的情况

```
fit <- fitted(model)
res <- resid(model)
pp <- predict(model,
              interval = "prediction",
              level = 0.95)
```

```
        fit       lwr      upr
1 -1.849460 -34.499842 30.80092
2 -1.849460 -34.499842 30.80092
3  9.947766 -22.061423 41.95696
4  9.947766 -22.061423 41.95696
5 13.880175 -17.956287 45.71664
6 17.812584 -13.872245 49.49741
7 21.744993  -9.809601 53.29959
```

模型可视化

线性回归模型可视化图形（不含预测部分）




```
p <- ggplot(cars,aes(speed,dist))+
  geom_point()+
  geom_smooth(method="lm")+
  geom_linerange(aes(ymin=fit,ymax=cars$dist))+
  geom_ribbon(aes(x=speed,y=NULL,ymin=pp[,'lwr'],ymax=pp[,'upr']),
              alpha=I(1/10),
              fill="red")
p
```

区间预测和可视化

利用predict函数生成区间预测值

```
pre <- predict(model,
               newdata = data.frame(speed=c(26,27,28)),
               interval = "prediction",
               level = 0.95)
df <- as.data.frame(cbind(c(26,27,28),pre))
```

在原拟合图基础上添加图层作出预测点和预测区间

```
p+geom_point(aes(V1,fit),data=df)+
  geom_ribbon(data=df,
              aes(x=V1,y=NULL,ymin=lwr,ymax=upr),
              alpha=I(1/10),
              fill="dark red")
```

回归模型

回归模型的拓展

公式符号的使用

回归模型拓展类型

公式符号

~将左右两边变量构成公式对象，记录自变量和因变量关系。

+ - : ^等运算符表示自变量间关系，并非数值运算含义；

I()函数还原了运算符的数值运算含义

公式代码	对应数学式
y~x	$y=\beta_0+\beta_1 x$
y~x-1	$y=\beta_1 x$
y~x+z	$y=\beta_0+\beta_1 x+\beta_2 z$
y~x+z+x:z	$y=\beta_0+\beta_1 x+\beta_2 z+\beta_3 x\cdot z$
y~(x+z)^2	$y=\beta_0+\beta_1 x+\beta_2 z+\beta_3 x\cdot z$
y~x+I(x^2)+I(x^3)	$y=\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3$

公式符号含义表

符　　号	用　　途
~	分隔符号，左边为响应变量，右边为解释变量。例如，要通过x、z和w预测y，代码为y ~ x + z + w
+	分隔预测变量
:	表示预测变量的交互项。例如，要通过x、z及x与z的交互项预测y，代码为y ~ x + z + x:z
*	表示所有可能交互项的简洁方式。代码y~ x * z * w可展开为y ~ x + z + w + x:z + x:w + z:w + x:z:w
^	表示交互项达到某个次数。代码y ~ (x + z + w)^2可展开为y ~ x + z + w + x:z + x:w + z:w
.	表示包含除因变量外的所有变量。例如，若一个数据框包含变量x、y、z和w，代码y~.可展开为y ~ x + z + w
-	减号，表示从等式中移除某个变量。例如，y ~ (x + z + w)^2 - x:w可展开为y ~ x + z + w + x:z + z:w
-1	删除截距项。例如，表达式y ~ x - 1拟合y在x上的回归，并强制直线通过原点
I()	从算术的角度来解释括号中的元素。例如，y~x+(z+w)^2将展开为y~x+z+w+z:w。相反,代码y ~ x + I((z + w)^2)将展开为y ~ x + h, h是一个由z和w的平方和创建的新变量
function	可以在表达式中用的数学函数。例如，log(y) ~ x + z + w表示通过x、z和w来预测log(y)

多项式回归

数据集women包含身高、体重数据，对比简单回归模型与二次项回归模型的效果。

```
> md1=lm(weight~height,data=women)
> md1

Call:
lm(formula = weight ~ height, data = women)

Coefficients:
(Intercept)       height
     -87.52         3.45
```

对应数学公式：

$$\hat{y} = 3.45x\text{-}87.52$$

```
> md2=lm(weight~height+I(height^2),data=women)
> md2

Call:
lm(formula = weight ~ height + I(height^2), data = women)

Coefficients:
(Intercept)       height  I(height^2)
  261.87818     -7.34832      0.08306
```

$$\hat{y}_i = 261.88 - 7.35x_i + 0.08x_i^2$$

含交互相项的模型

mtcars数据集，以mpg变量为因变量，hp和wt为自变量建立回归模型

```
> md3=lm(mpg~hp+wt+hp:wt,data=mtcars)
> md3

Call:
lm(formula = mpg ~ hp + wt + hp:wt, data = mtcars)

Coefficients:
(Intercept)          hp          wt       hp:wt
   49.80842    -0.12010    -8.21662     0.02785
```

模型含有两自变量的交互项，需用冒号构造交互项。

对应数学公式：

$$mpg_i = 49.81 - 0.12hp_i - 8.22wt_i + 0.03hp_i \cdot wt_i$$

可线性化的模型

根据Cobb-Douglas生产函数建模分析gdp、投资和劳动力。该模型通过取对数可以转化为线性模型，进而按照线性回归模型方式计算。

对应数学公式：

$$GDP = AL^{\alpha}C^{\beta}$$

$$\ln(GDP) = \ln A + \alpha \ln L + \beta \ln C$$

```
> md4=lm(log(gdp)~log(capital)+log(labor),data=nanjinggdp)
> md4

Call:
lm(formula = log(gdp) ~ log(capital) + log(labor), data = nanjinggdp)

Coefficients:
 (Intercept)  log(capital)    log(labor)
      1.5022        0.6781        0.5717
```