数据分析与处理技术

回归与预测

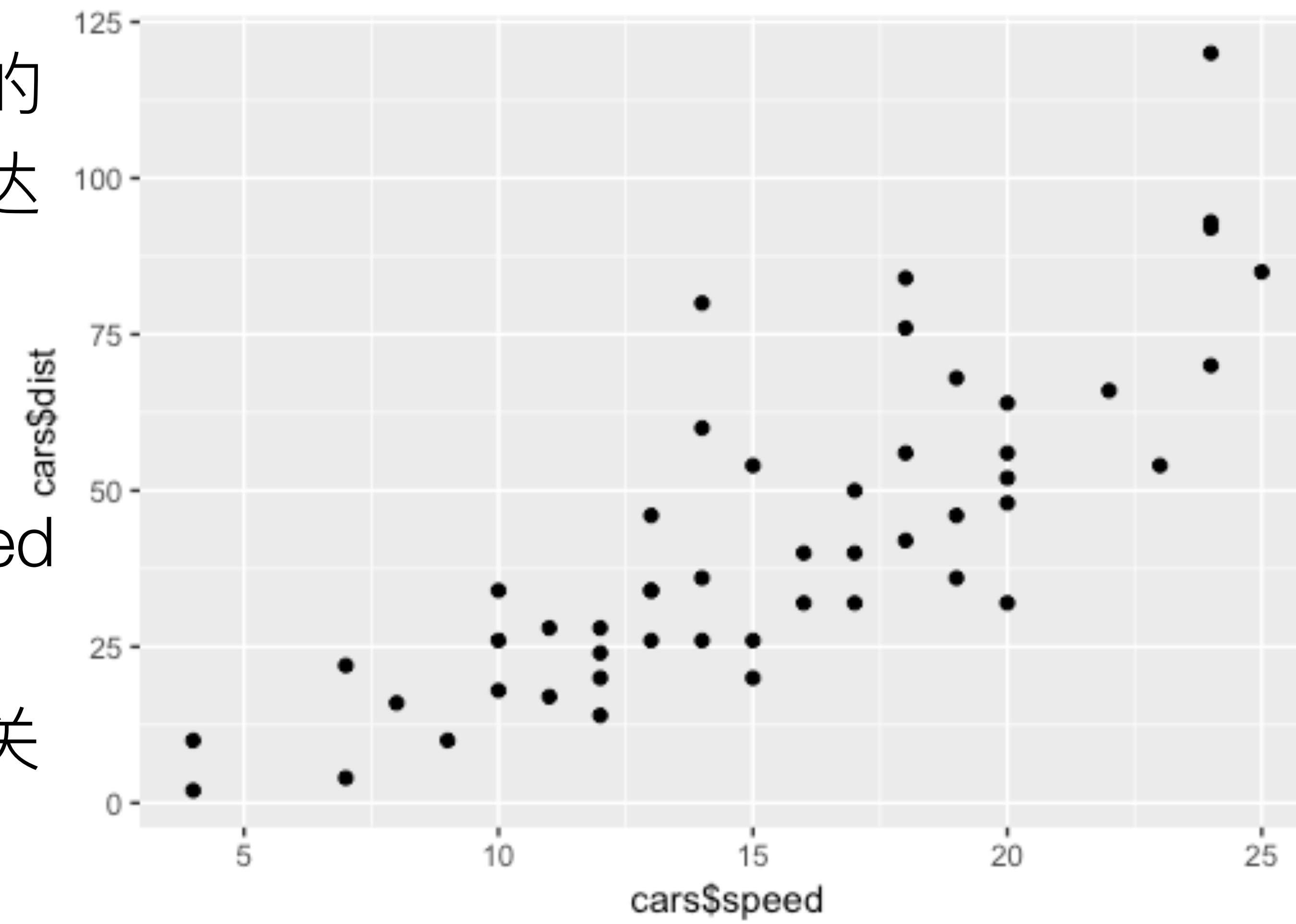
趋势的相关性

cars数据集，dist与speed存在明显的增长关联关系。两变量间相关关系达到超过0.8

```
> cor(cars$speed,cars$dist)
[1] 0.8068949
```

进一步，dist代表的刹车距离与speed代表的车速之间确实存在因果关系，这也不难理解两变量增长间的高相关关系。

问题：dist与speed之间有趋势关联而非确定性函数关系，如何找到最好的描述数据发展趋势的拟合线?




```
> qplot(cars$speed,cars$dist)
```

回归线的误差标准

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?

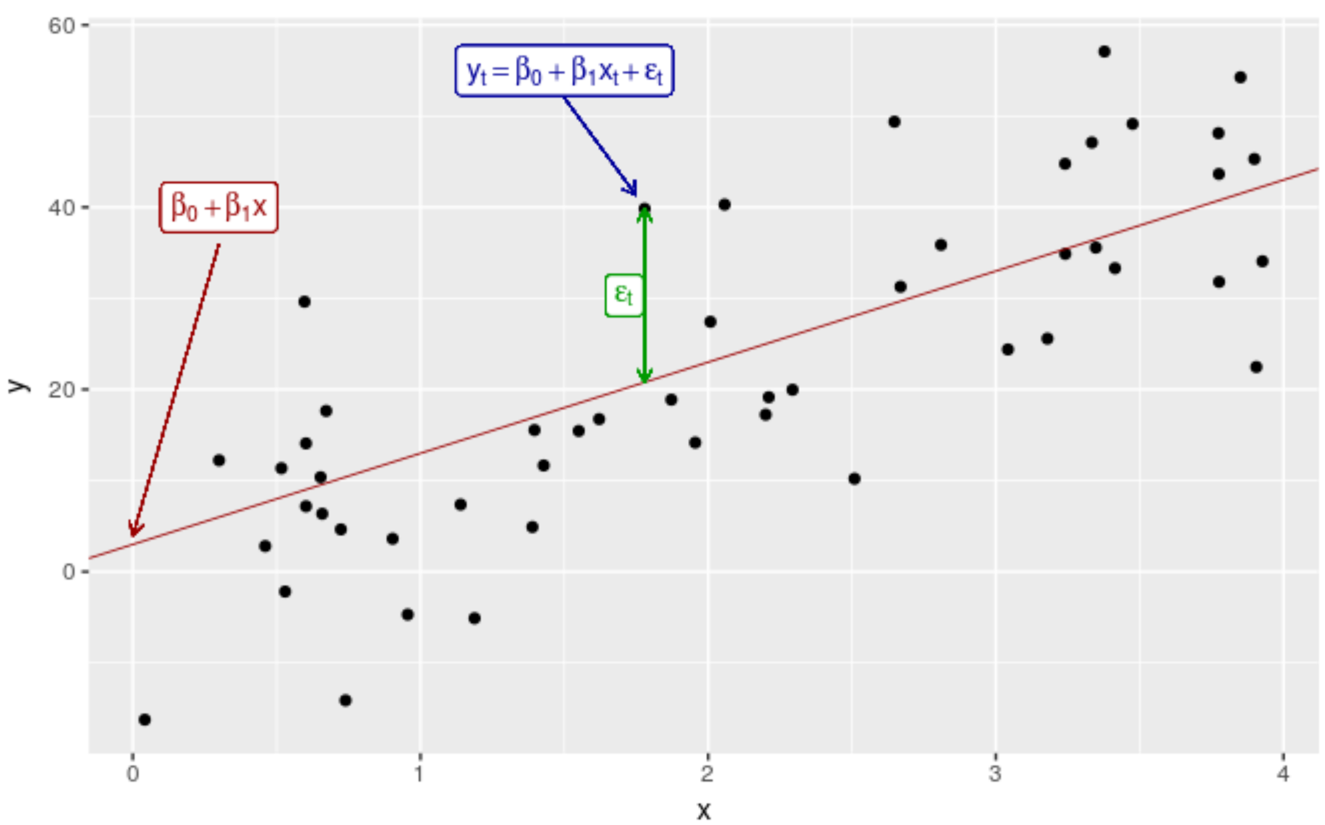
$$y = \beta_0 + \beta_1 x$$

点到直线L的离差 $\varepsilon_i = y_t - (\beta_0 + \beta_1 x_t)$

所有点到直线L的离差之和最小

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

$$\begin{cases} \dfrac{dS}{d\beta_0} = 0 \\ \dfrac{dS}{d\beta_1} = 0 \end{cases}$$



参数估计

最小二乘法的矩阵形式：

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \beta_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$Y = X \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$X^T Y = X^T X \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\left(X^T X \right)^{-1} X^T Y = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

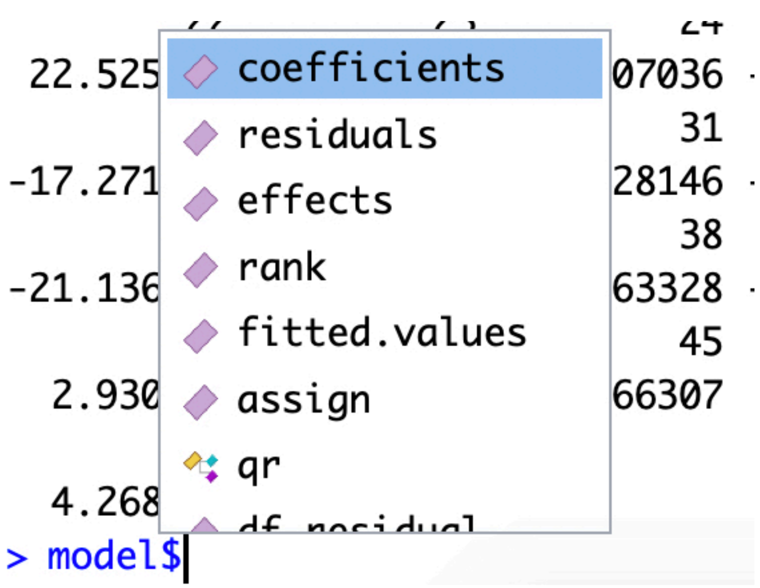
向量y投影到X形成的空间中的投影，投影最短原理

线性回归

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

```
> plot(cars)
> abline(model)
```

观察回归结果




```
> residuals(model)
```

提取残差

```
> coef(model)
```

提取回归参数

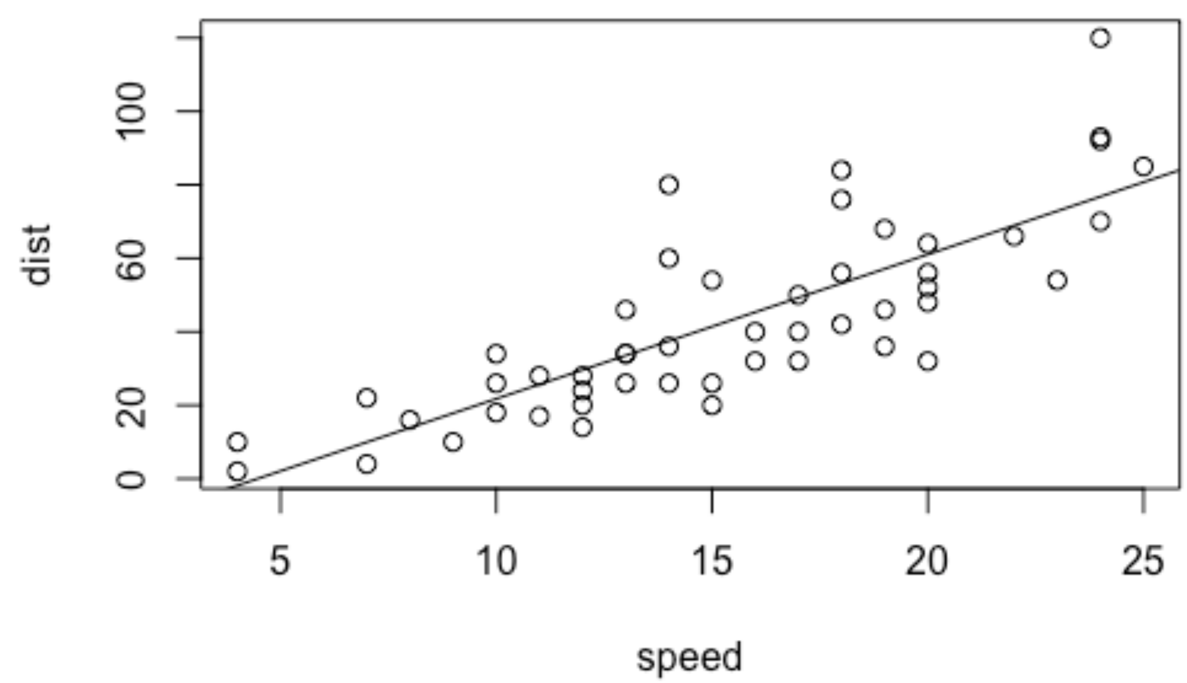
```
> summary(model)

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
```



趋势的正态假设

回归分析可预测的前提假设：

两变量具有明显的相关关系　　两变量具有明确的因果关系

数据与回归线间的误差为噪声

误差近似正态分布

推演回归值，估计预测值的置信区间

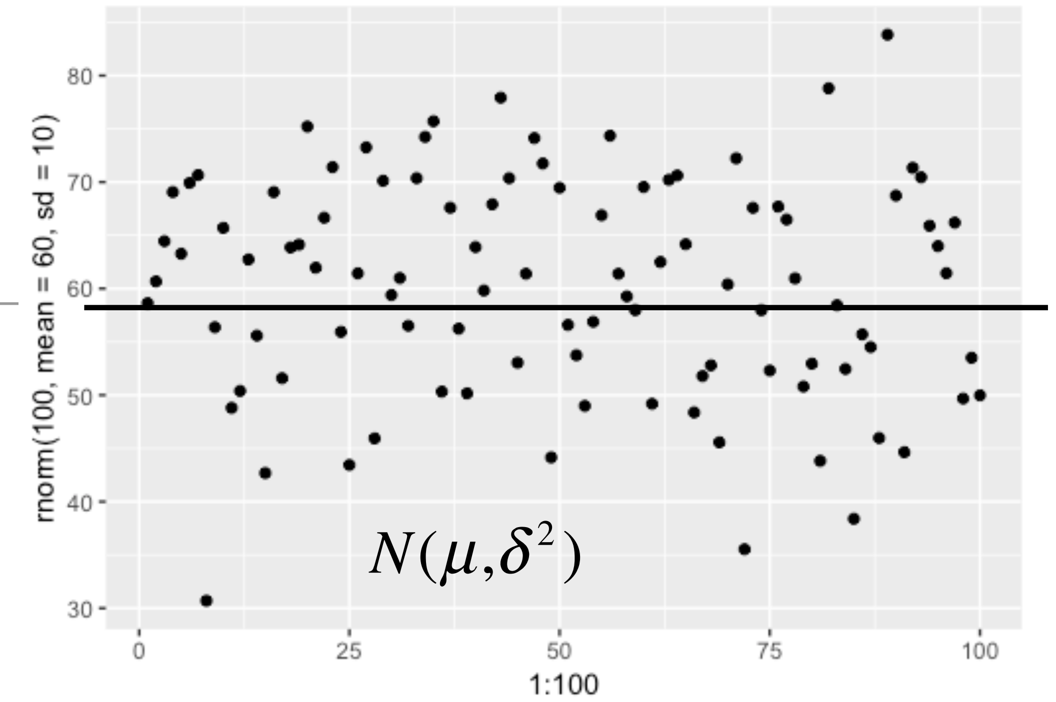
预测

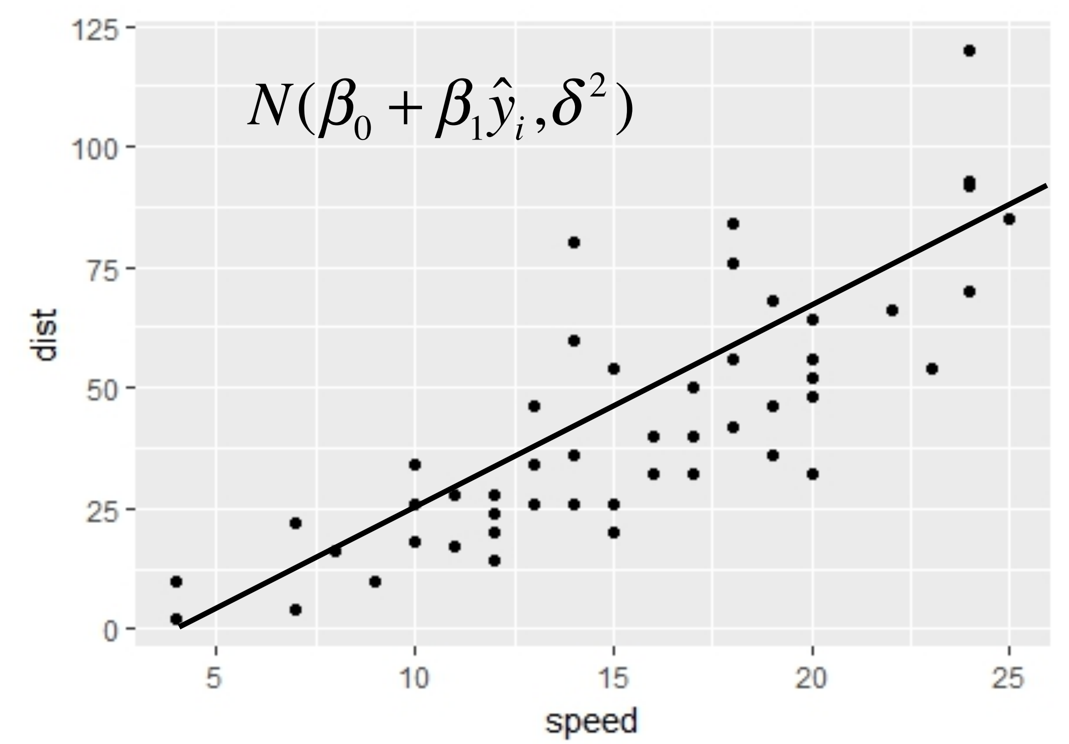
回归拟合的是不确定性关系，无法明确预测未来数据将出现在哪里，但这并不意味着无法预测。如果知道误差的分布特征(正态分布)，那就可以在一定概率基础上预测未来数据将出现的范围。

$$y_0 - \hat{y}_0 \sim N(0,\delta^2(1+\frac{1}{n}+\frac{(x_0-\overline{x})}{\sum x_i^2}))$$

$$\left(\hat{y}_0 - t_{1-\frac{\alpha}{2}} \cdot S_{\hat{y}_0-y_0}, \hat{y}_0 + t_{1-\frac{\alpha}{2}} \cdot S_{\hat{y}_0-y_0}\right)$$

```
> predict(model,newdata = data.frame(speed=c(30,31)))
       1        2 
100.3932 104.3256 
```






```
> predict(model,data.frame(speed=c(30,31)),interval = 'prediction',level=0.9)
       fit      lwr      upr
1 100.3932 72.42500 128.3613
2 104.3256 76.09641 132.5547
```

拟合效果评估

可决系数评估线性回归效果

$$d_i = y_i - \overline{y} = (y_i - \hat{y}) + (\hat{y} - \overline{y})$$

$$TSS = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$$

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 + \sum_{i=1}^{n} \left(\overline{y} - \hat{y}_i\right)^2$$

$$TSS = RSS + ESS$$

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

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

总体平方和TSS:Total sum of squares
回归平方和ESS:Expanded sum of squares
残差平方和RSS:Residual sum of squares

可决系数范围:[0,1]
越接近1说明回归模型的解释能力越强

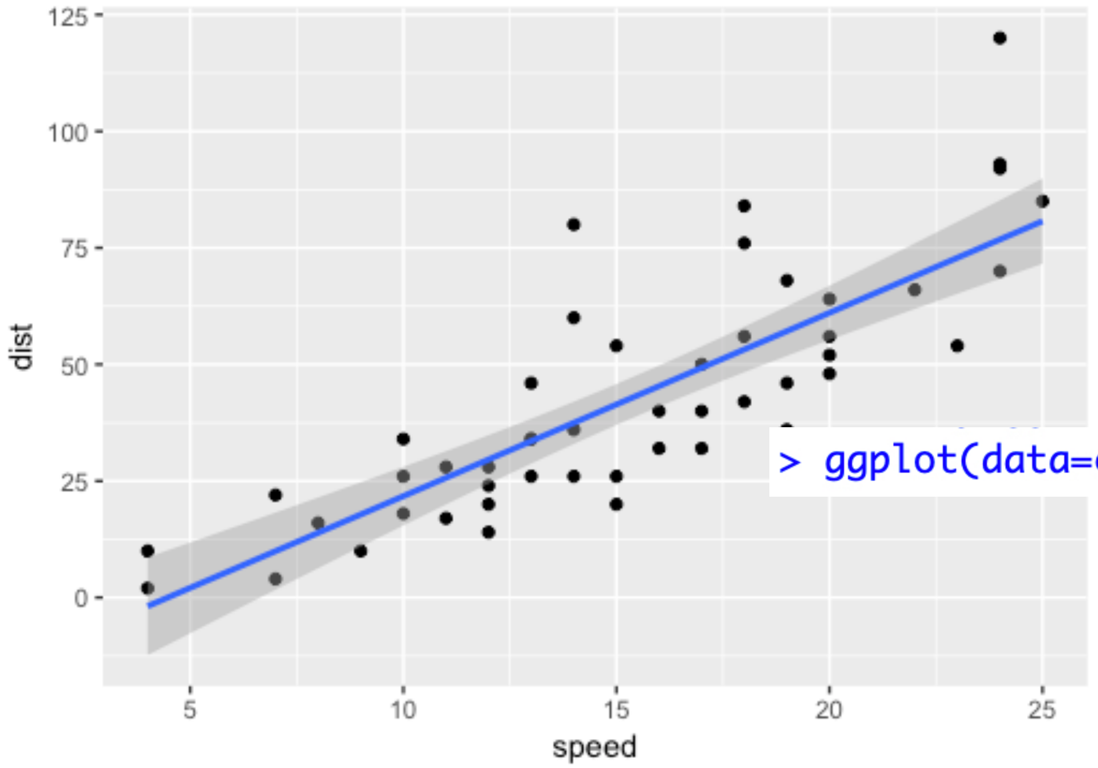
平均绝对误差MAE: Mean absolute errors

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

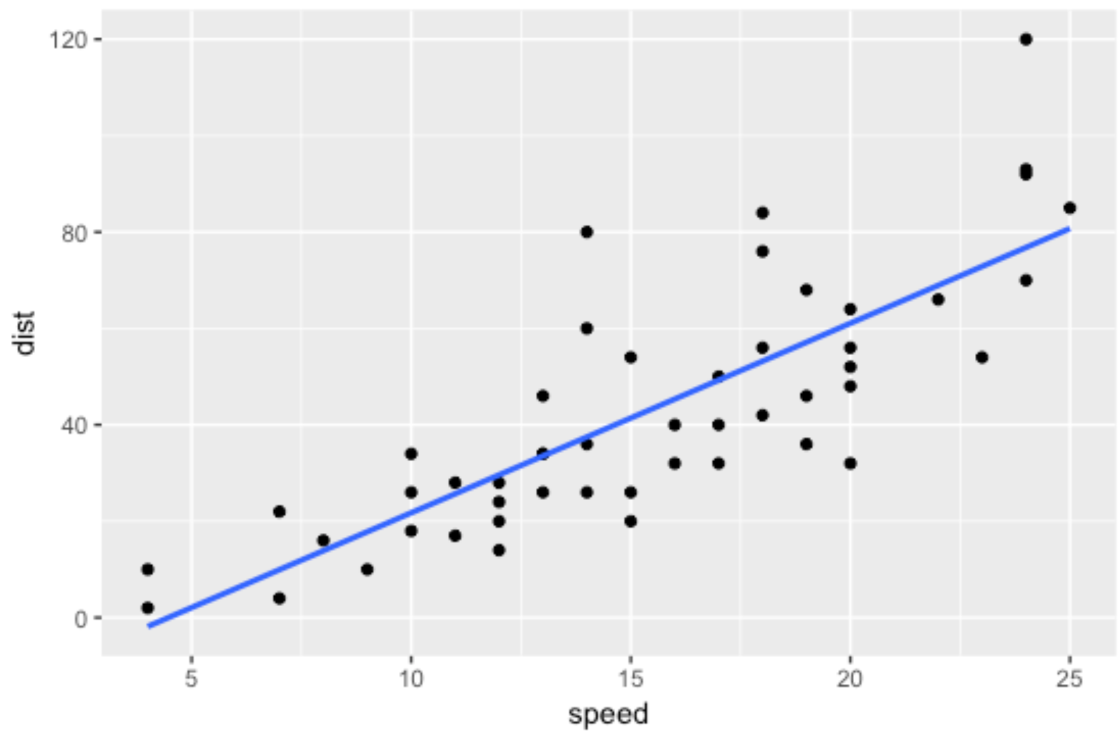
平均相对误差MAPE: Mean absolute percentage errors

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right| = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

拟合图




```
> ggplot(data=cars,aes(x=speed,y=dist))+geom_point()+geom_smooth(method='lm')
```




```
> ggplot(data=cars,aes(x=speed,y=dist))+geom_point()+geom_smooth(method='lm',se=FALSE)
```

公式的变化规则

代码：	回归式：
`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)`	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

公式中符号含义

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)

案例-多项式回归

数据集women，15行身高与体重的数据

对前14个对象建立回归模型，利用第15个数据检验拟合精度

```
> fit<-lm(weight~height,data=women)
> fit

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

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

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

```
> fit2<-lm(weight~height+I(height^2),data=women)

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

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

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

函数检验拟合精度与预测精度

```
testmape<-function(f){
  insample=mean(abs(f$residuals)/f$model$weight)
  pre=predict(f,newdata=data.frame(height=women$height[15]))
  outsample=(pre-women$weight[15])/women$weight[15]
  cat('insample is',insample,'\n')
  cat('outsample is',outsample)
}
```

```
> fit2=lm(weight~height+I(height^2),data=women[-15,])
> testmape(fit2)
insample is 0.001789464
outsample is -0.006801126
```

多元回归

数据集freeny，变量y为因变量，其余为可用的自变量

```
> cov(freeny)
```

```
> plot(freeny)
```

```
> fitmulti=lm(y~lag.quarterly.revenue+price.index,data=freeny)
```

```
> summary(fitmulti)
```

有交互项的多元回归

简单的多元线性回归直接在自变量位置加新的变量即可，如果模型设置有两个自变量的交互项，则需要用到冒号：来表示。

```
> fit3<-lm(mpg~hp+wt+hp:wt,data=mtcars)
> fit3

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

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

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

可线性化的非线性回归

Cobb-Douglas生产函数为例，对gdp、投资和劳动力做回归分析

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

化为线性形式 $\ln(GDP) = \ln A + \alpha \ln L + \beta \ln C$

```
> fit4<-lm(log(gdp)~log(capital)+log(labor),data=nanjinggdp)
> fit4

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

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

课本P255-256

过拟合与多重共线性

多项式回归尝试增加自变量个数回归式在拟合历史数据的时候精度会越来越高，但当变量个数接近对象个数时，预测值的精度会迅速下降。

当自变量可以被其他自变量线性表出时，回归模型的预测能力将大幅下降，但模型对历史数据的拟合能力反而上升。

降低过拟合的方法很多，常用的主要是对自变量进行筛选，降低自变量个数。具体方法见课本P225-229