



北京航空航天大学

— 经济管理学院 —

BEIHANG UNIVERSITY  
SCHOOL OF ECONOMICS AND MANAGEMENT

## 第 2.2 讲：一元线性回归模型

康雁飞

2020-05-11

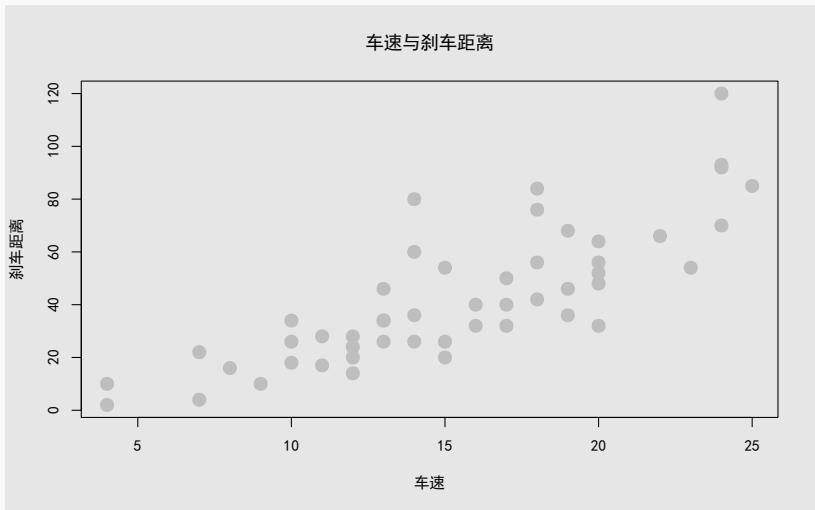
# Outline

- 1 回归模型介绍
- 2 一元线性回归模型
- 3 模型估计
- 4 模型推断
- 5 模型诊断
- 6 预测
- 7 作业

# Outline

- 1 回归模型介绍
- 2 一元线性回归模型
- 3 模型估计
- 4 模型推断
- 5 模型诊断
- 6 预测
- 7 作业

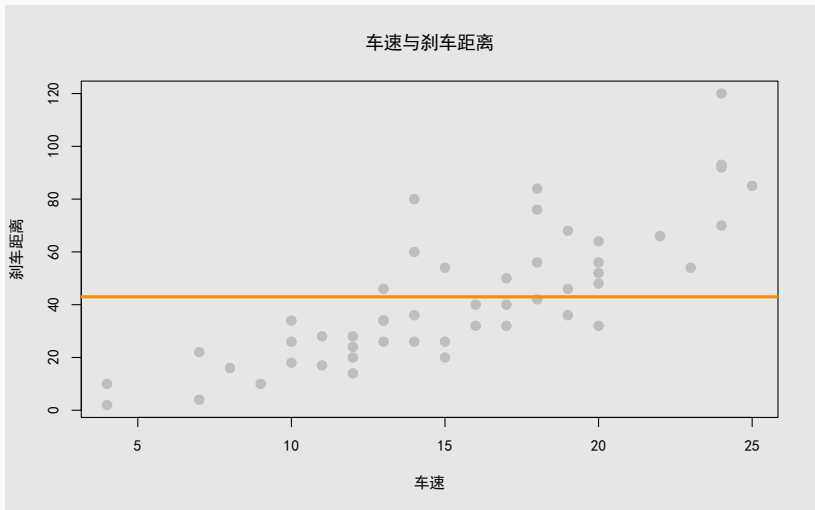
# 一个例子：车速和刹车距离



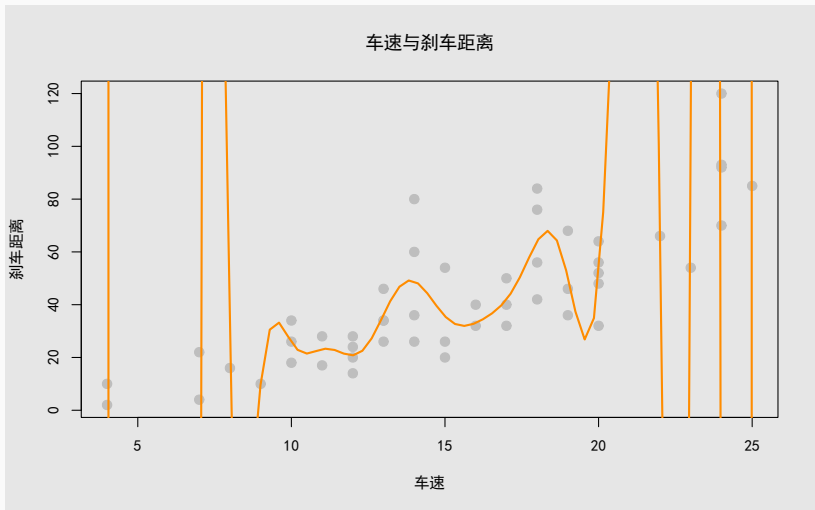
## 问题定义

- 我们有观测值  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ ，其中  $n$  为观测值数量。其中  $x_i, i = 1, \dots, n$  代表自变量， $y_i, i = 1, \dots, n$  代表因变量。
- 回归模型的目的就是找到  $Y = f(X) + \epsilon$ 。
- 解释 + 预测。
- 如何找到  $f()$ ?

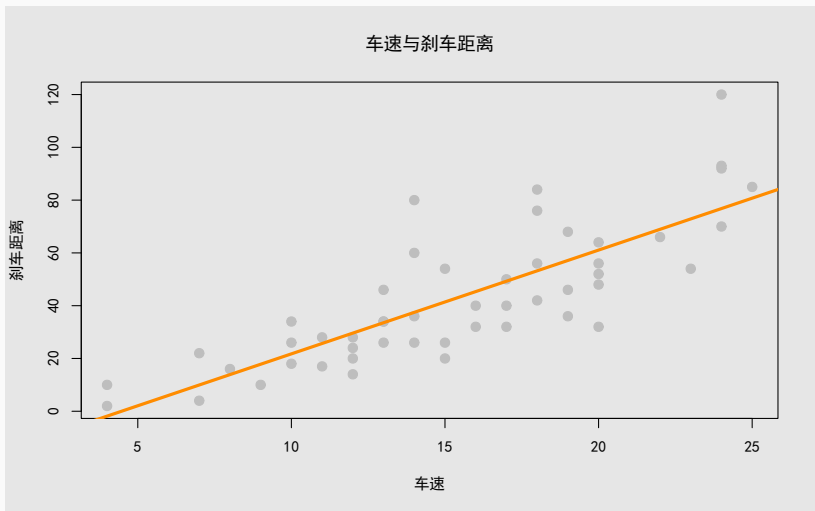
# 尝试 1



## 尝试 2



# 尝试 3





# Outline

- 1 回归模型介绍
- 2 一元线性回归模型
- 3 模型估计
- 4 模型推断
- 5 模型诊断
- 6 预测
- 7 作业

## 一元线性回归模型

一元线性回归模型形式为：

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, n,$$

其中  $\epsilon_i \sim N(0, \sigma^2)$ 。

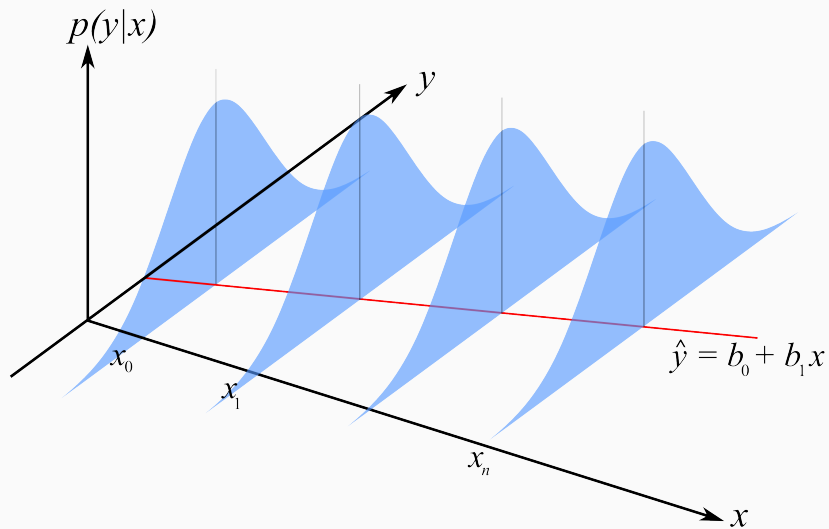
# 一元线性回归模型

- 1  $x_i$  是已知观测值。
- 2  $Y_i$  是随机变量。
- 3  $y_i$  是  $Y_i$  的可能取值。
- 4 模型中需要估计的参数为  $\beta_0, \beta_1, \sigma$ 。
- 5 又叫简单线性回归模型。

## $Y_i$ 的分布?

- $E(Y_i|X_i = x_i) = \beta_0 + \beta_1 x_i.$
- $\text{var}(Y_i|X_i = x_i) = \sigma^2.$
- $Y_i$  的分布?

# $Y_i$ 的分布

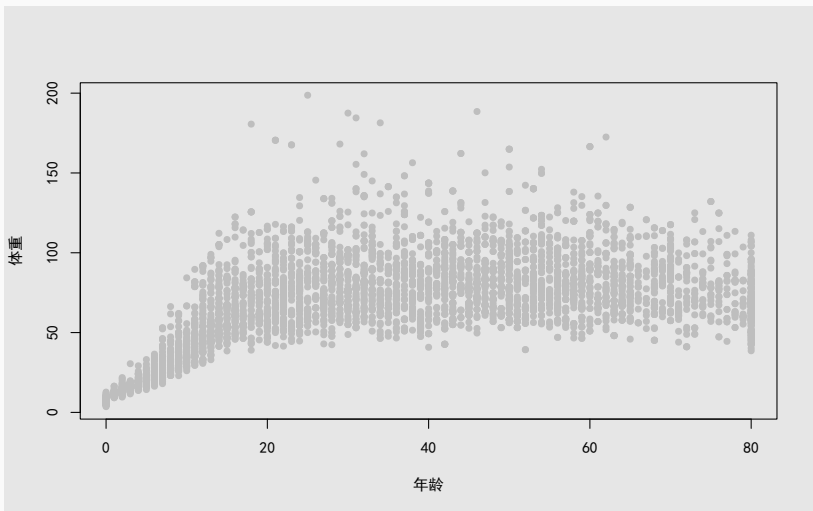


## 一元线性回归模型假设 (LINE)

- 线性关系 (Linear)
- 误差项独立 (Independent)
- 正态 (Normal)
- 同方差 (Equal Variance)

# 年龄和体重

数据下载：[点击这里](#)。



# Outline

- 1 回归模型介绍
- 2 一元线性回归模型
- 3 模型估计**
- 4 模型推断
- 5 模型诊断
- 6 预测
- 7 作业



## 最小二乘法

我们有观测值  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , 那么模型估计的第一个问题是: 如果根据观测数据估计出  $\beta_0$  和  $\beta_1$ ?

- $\hat{y}_i \neq y_i$
- $e_i = y_i - \hat{y}_i$
- 最小二乘法:

- 最小化残差平方和 (Sum of Squared Errors, SSE)

$$f(\beta_0, \beta_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2.$$

- $\arg \min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$

## 最小化残差平方和 → 优化问题

求偏导：

$$\frac{\partial f}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial f}{\partial \beta_1} = -2 \sum_{i=1}^n (x_i) (y_i - \beta_0 - \beta_1 x_i)$$

化简得到：

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\sum_{i=1}^n (x_i) (y_i - \beta_0 - \beta_1 x_i) = 0$$

## 最小二乘估计

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

# 最大似然估计

- 我们的模型为：

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, n,$$

其中  $\epsilon_i \sim N(0, \sigma^2)$ 。

- 我们知道：  $Y_i | X_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ 。

似然函数？

$$\begin{aligned} & L(\beta_0, \beta_1, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2} \left( \frac{y_i - \beta_0 - \beta_1 x_i}{\sigma} \right)^2 \right] \\ &= \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \right] \end{aligned}$$

## 对数似然

上述似然函数取对数：

$$\begin{aligned} & \log L(\beta_0, \beta_1, \sigma^2) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

通过求导可以最大化对数似然：

$$\frac{\partial \log L}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial \log L}{\partial \beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

# 最大似然估计

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

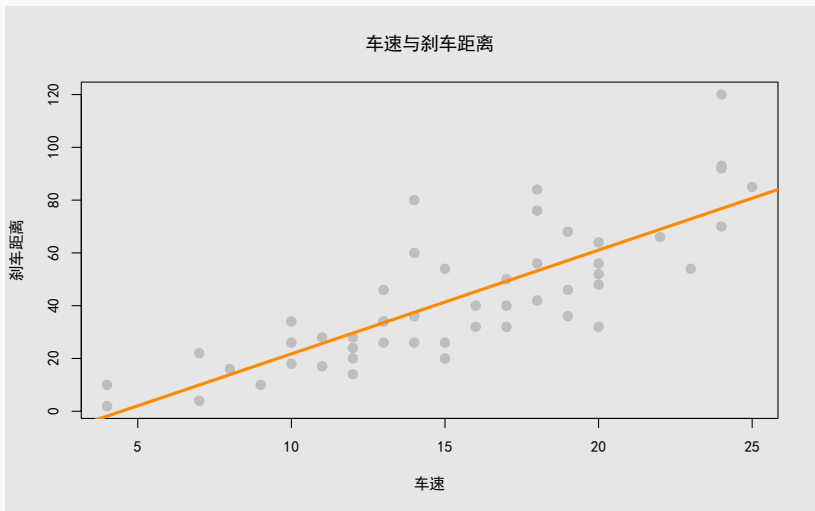
$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



## 在 R 中建立一元线性回归模型

```
stop_dist_model <- lm(dist ~ speed,  
                      data = cars)  
plot(dist ~ speed, data = cars,  
      xlab = " 车速",  
      ylab = " 刹车距离",  
      main = " 车速与刹车距离",  
      pch = 20,  
      cex = 2,  
      col = "grey")  
abline(stop_dist_model, lwd = 3,  
        col = "darkorange")
```

# 在 R 中建立一元线性回归模型



```
coef(stop_dist_model)
```

```
## (Intercept)          speed  
## -17.579095      3.932409
```

# 残差 (Residuals)

残差就是真实值和估计值之间的差：

$$e_i = y_i - \hat{y}_i.$$

```
stop_dist_model$residuals
```

```
##           1           2           3           4           5           6           7
##  3.849460  11.849460 -5.947766  12.052234  2.119825 -7.812584 -3.744993
##           8           9          10          11          12          13          14
##  4.255007  12.255007 -8.677401  2.322599 -15.609810 -9.609810 -5.609810
##          15          16          17          18          19          20          21
## -1.609810 -7.542219  0.457781  0.457781  12.457781 -11.474628 -1.474628
##          22          23          24          25          26          27          28
##  22.525372  42.525372 -21.407036 -15.407036  12.592964 -13.339445 -5.339445
##          29          30          31          32          33          34          35
## -17.271854 -9.271854  0.728146 -11.204263  2.795737  22.795737  30.795737
##          36          37          38          39          40          41          42
## -21.136672 -11.136672  10.863328 -29.069080 -13.069080 -9.069080 -5.069080
##          43          44          45          46          47          48          49
##  2.022820  2.022820  18.866207  6.708715  15.201285  16.201285  42.201285
```

## 方差估计

误差项方差  $\sigma^2$  的无偏估计为:

$$\begin{aligned} s_e^2 &= \frac{1}{n-2} \sum_{i=1}^n \left( y_i - \left( \hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right)^2 \\ &= \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \frac{1}{n-2} \sum_{i=1}^n e_i^2 \end{aligned}$$

注意:  $\bar{e} = 0$ .

## 练习：车速与刹车距离

请估计误差项标准差。

```
stop_dist_model_summary <- summary(stop_dist_model)
stop_dist_model_summary$sigma
```

```
## [1] 15.37959
```

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

- SST = SSE + SSR
- SST: Sum of Squares Total, 总平方和
- SSE: Sum of Squares Error, 残差平方和
- SSR: Sum of Squares Regression, 回归平方和

## 拟合优度 $R^2$

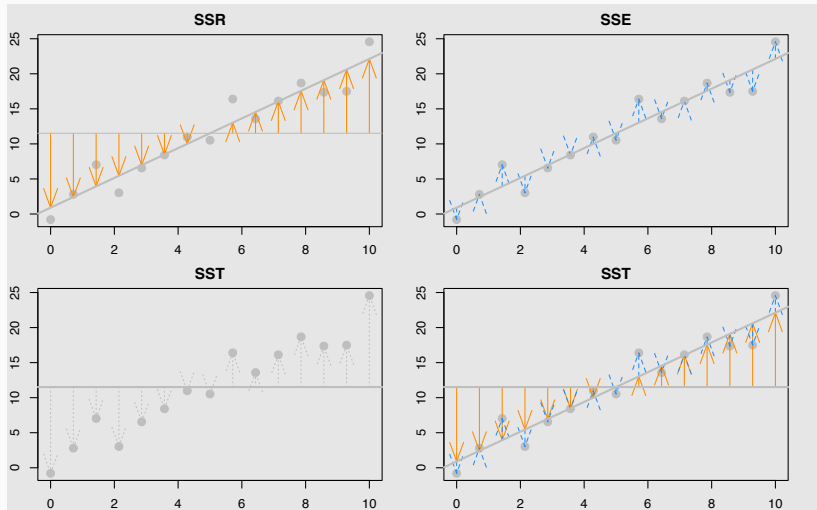
拟合优度 (Goodness of Fit), 又叫测定系数 (Coefficient of Determination) :

$$R^2 = \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

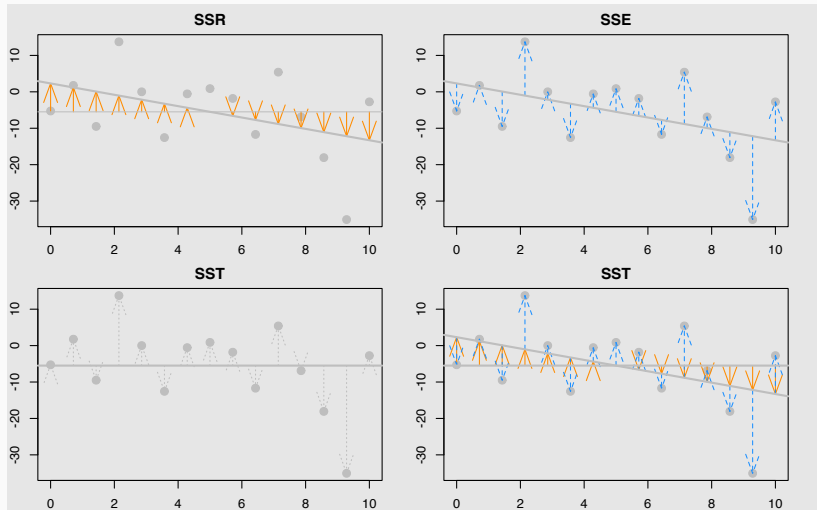
- SSR 越大: 用回归方程解释  $Y$  变异的部分越多。
- SSE 越小: 观测值  $Y$  绕回归线越紧密, 拟合越好。



$$R^2 = 0.92$$



$$R^2 = 0.19$$



# $R^2$ 的性质

1  $0 \leq R^2 \leq 1$

2  $R^2 = \hat{\beta}_1^2 \frac{s_x^2}{s_y^2}$

3  $R^2 = [r(X, Y)]^2$

## R 中求拟合优度

```
stop_dist_model_summary$r.squared
```

```
## [1] 0.6510794
```

# 如何进行模型评价?

样本数据的回归模型总是可以求到的，但是它是否确实是总体回归模型的正确估计呢？

- 1 该模型能否较好地解释  $y_i$  的取值变化规律？
- 2 自变量  $X$  真的可以解释  $Y$  吗？
- 3 关于一元线性回归模型的几个基本假设条件是否得到满足？

# Outline

- 1 回归模型介绍
- 2 一元线性回归模型
- 3 模型估计
- 4 模型推断**
- 5 模型诊断
- 6 预测
- 7 作业

# Gauss-Markov 定理

如果基本假设成立，最小二乘估计量是总体参数  $\beta_0$  和  $\beta_1$  的线性最小方差无偏估计量。

## 标准误差 (Standard Errors, SE)

我们可以推导出:

$$\hat{\beta}_0 \sim N \left( \beta_0, \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \right)$$
$$\hat{\beta}_1 \sim N \left( \beta_1, \frac{\sigma^2}{S_{xx}} \right)$$



## 标准误差 (Standard Errors, SE)

但是  $\sigma$  未知，所以需要用到  $s_e$  代替：

$$\begin{aligned} \text{SE}[\hat{\beta}_0] &= s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \\ \text{SE}[\hat{\beta}_1] &= \frac{s_e}{\sqrt{S_{xx}}} \end{aligned}$$

从而我们有：

$$\begin{aligned} \frac{\hat{\beta}_0 - \beta_0}{\text{SE}[\hat{\beta}_0]} &\sim t_{n-2} \\ \frac{\hat{\beta}_1 - \beta_1}{\text{SE}[\hat{\beta}_1]} &\sim t_{n-2} \end{aligned}$$

## $\beta_0$ 和 $\beta_1$ 的置信区间

$\beta_0$  的置信区间:

$$\hat{\beta}_0 \pm t_{\alpha/2, n-2} \cdot \text{SE} [\hat{\beta}_0],$$

$\beta_1$  的置信区间:

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \cdot \text{SE} [\hat{\beta}_1],$$

其中  $P(t_{n-2} > t_{\alpha/2, n-2}) = \alpha/2$ .

## R 中求置信区间

```
confint(stop_dist_model, level = 0.95)
```

```
##                2.5 %    97.5 %  
## (Intercept) -31.167850 -3.990340  
## speed        3.096964  4.767853
```

## 假设检验

以  $\beta_0$  为例，对于  $\beta_0$  的假设检验为：

$$H_0 : \beta_0 = 0 \quad \text{vs} \quad H_1 : \beta_0 \neq 0$$

检验统计量：

$$t = \frac{\hat{\beta}_0 - 0}{\text{SE}[\hat{\beta}_0]} = \frac{\hat{\beta}_0}{s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}.$$

# R 中的假设检验

```
stop_dist_model_summary$coefficients
```

##	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	-17.579095	6.7584402	-2.601058	1.231882e-02
## speed	3.932409	0.4155128	9.463990	1.489836e-12

## F 检验

F 检验是为了检验回归的显著性。 $H_0$  为  $Y$  不依赖于  $X$ 。

通过误差分解，我们有方差分析表：

来源	平方和	自由度	均方
回归	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	1	SSR/1
误差	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n - 2$	SSE/( $n - 2$ )
总	$\sum_{i=1}^n (y_i - \bar{y})^2$	$n - 1$	

## F 检验

$$F = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / 1}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - 2)} \sim F_{1, n-2}$$

注意：在一元线性回归中， $F$  检验和回归系数  $\beta_1$  的  $t$  检验是等价的。

## R 中的 F 检验

```
anova(stop_dist_model)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: dist
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## speed      1  21186 21185.5   89.567 1.49e-12 ***
## Residuals 48  11354   236.5
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0
```



# Outline

- 1 回归模型介绍
- 2 一元线性回归模型
- 3 模型估计
- 4 模型推断
- 5 模型诊断**
- 6 预测
- 7 作业

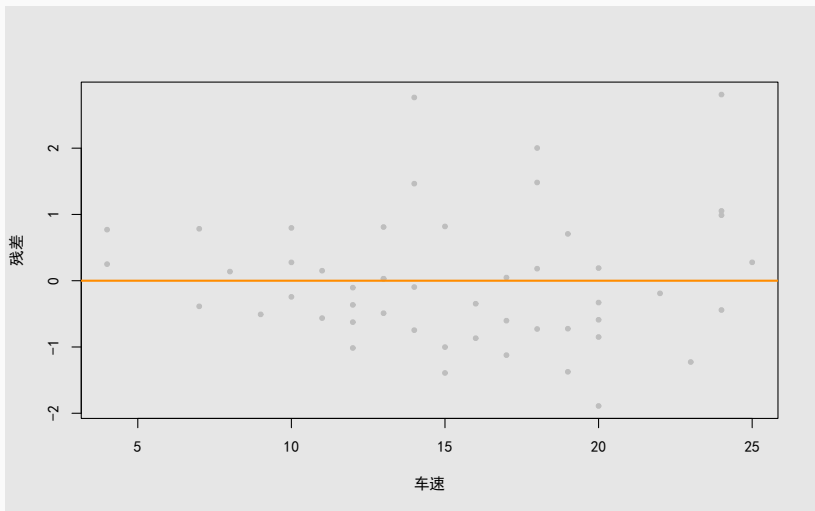
## 还记得一元线性回归模型的假设吗？

- 如果满足假设，完美  $\checkmark$
- 如果不满足假设？ Garbage in, garbage out.

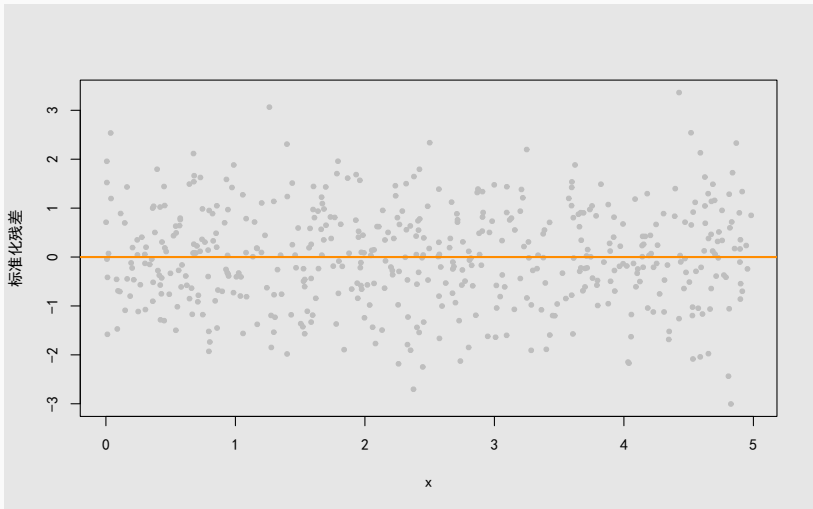
在满足假设的条件下：

- $\bar{e} = 0$ .
- $\text{Var}(e_i) = s_e^2$ .
- 标准化残差  $e_i^* = \frac{e_i - 0}{s_e}$ .
- 当  $n \rightarrow +\infty$  时,  $e_i^* \sim N(0, 1)$ .

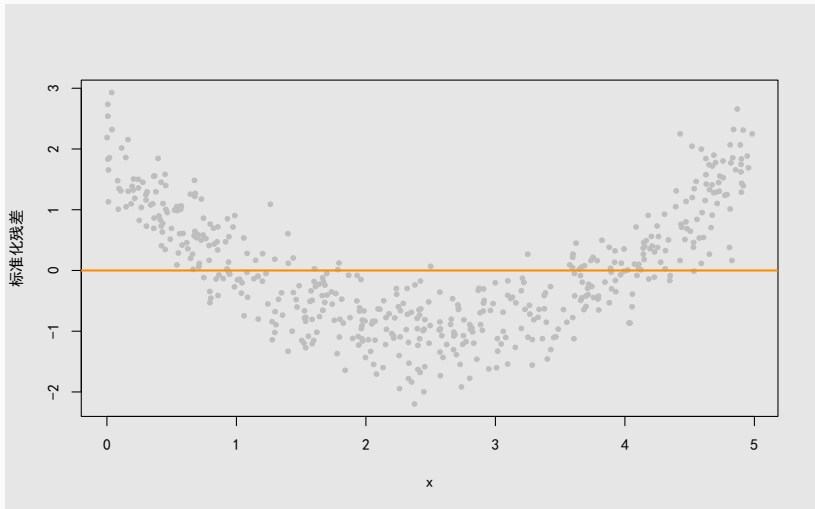
# 标准化残差图



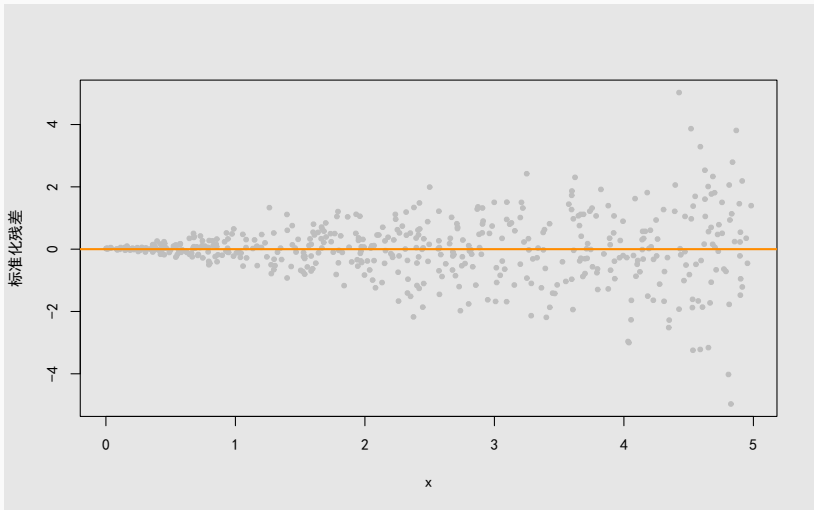
# 拟合良好



# 非线性



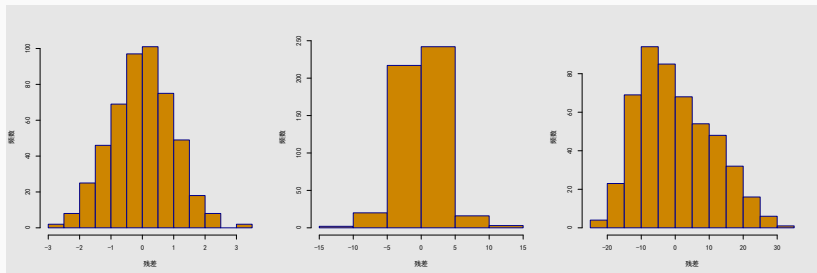
# 异方差



- 1 直方图
- 2 PP 图或者 QQ 图
- 3 正态性假设检验



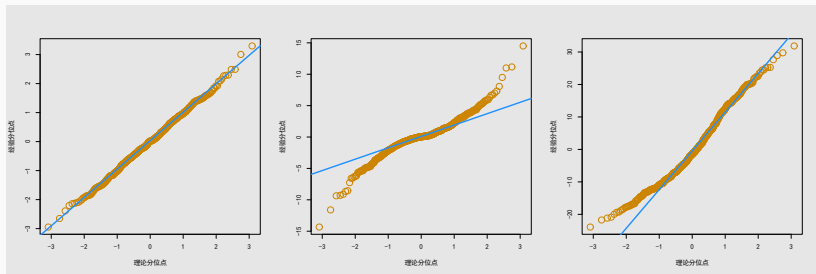
# 正态性-直方图



# 正态性 - PP 图或 QQ 图

1 PP 图 (Percentile-Percentile)

2 QQ 图 (Quantile-Quantile)



Shapiro 检验，请点击[这里](#)查看更多检验理论细节。

```
shapiro.test(stop_dist_model$residuals)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data:  stop_dist_model$residuals  
## W = 0.94509, p-value = 0.02152
```

不满足“残差独立”的假设，可从残差图可以看出。产生的主要原因有：

- 1 重要的解释变量被遗漏
- 2 模型函数形式错误
- 3 时间序列自变量

# Outline

- 1 回归模型介绍
- 2 一元线性回归模型
- 3 模型估计
- 4 模型推断
- 5 模型诊断
- 6 预测**
- 7 作业

- 给定未知的  $x$ ，点预测为  $\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ 。
- $E(Y|X = x) = \beta_0 + \beta_1 x$ 。
- 我们可以用  $\hat{y}(x)$  来估计  $E(Y|X = x)$ ，因为  $E(\hat{y}(x)) = \beta_0 + \beta_1 x$ 。

## 置信区间

- $\text{Var}[\hat{y}(x)] = \sigma^2 \left( \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}} \right)$ .
- $\hat{y}(x) \sim N \left( \beta_0 + \beta_1 x, \sigma^2 \left( \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}} \right) \right)$ .

### 置信区间

$$\hat{y}(x) \pm t_{\alpha/2, n-2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}$$

## R 中求置信区间

```
new_speeds = data.frame(speed = c(5, 21))
predict(stop_dist_model, newdata = new_speeds,
        interval = c("confidence"), level = 0.99)
```

```
##           fit      lwr      upr
## 1  2.082949 -10.89309 15.05898
## 2 65.001489  56.45836 73.54462
```



## 预测区间

给定  $x$ ，我们想预测  $Y$ ，我们知道这个新的观测值和  $\hat{y}$  之间（也就是真实值和回归线之间）差的是一个  $\epsilon$ 。我们对于  $Y$  的预测依然是  $\hat{y}$ ，但是这个预测的方差更大：

$$\begin{aligned}\text{Var}[\hat{y}(x) + \epsilon] &= \text{Var}[\hat{y}(x)] + \text{Var}[\epsilon] \\ &= \sigma^2 \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right) + \sigma^2 \\ &= \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right)\end{aligned}$$

### 预测区间

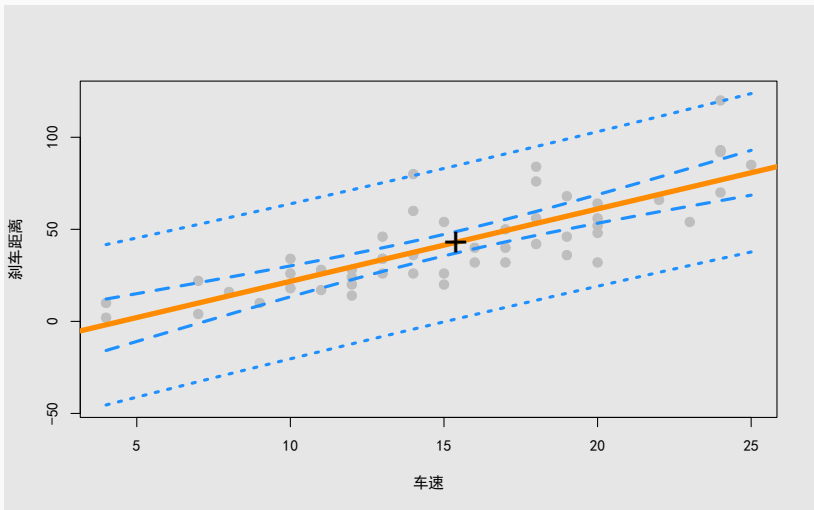
$$\hat{y}(x) \pm t_{\alpha/2, n-2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}.$$

## R 中求预测区间

```
new_speeds = data.frame(speed = c(5, 21))
predict(stop_dist_model, newdata = new_speeds,
        interval = c("prediction"), level = 0.99)
```

```
##           fit      lwr      upr
## 1  2.082949 -41.16099  45.32689
## 2 65.001489  22.87494 107.12803
```

# 置信带和预测带



# Outline

- 1 回归模型介绍
- 2 一元线性回归模型
- 3 模型估计
- 4 模型推断
- 5 模型诊断
- 6 预测
- 7 作业

应用上一讲所采集的班级同学数据，考虑构建一个回归模型，并对研究结果进行分析和解释。