



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

数量经济与商务统计系

Outline

- 1 回归模型介绍
- 2 一元线性回归模型
- 3 模型估计
- 4 回归参数的统计推断
- 5 模型诊断
- 6 预测
- 7 Key points

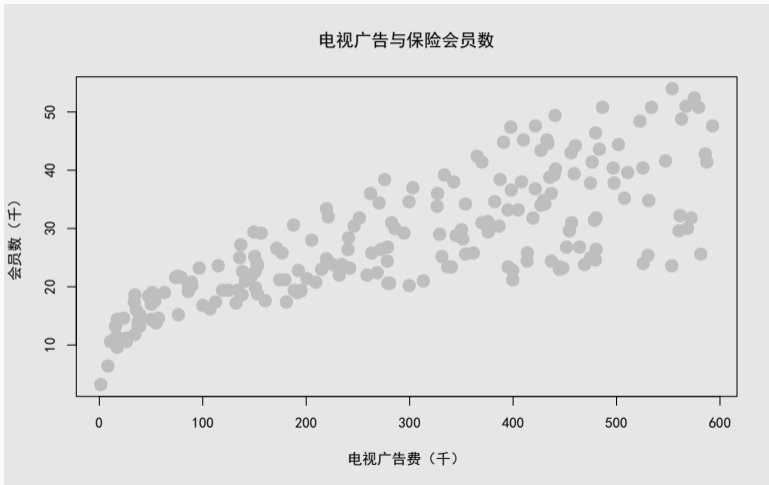
“All models are wrong, but some are useful.”

— George E. P. Box

Outline

- 1 回归模型介绍
- 2 一元线性回归模型
- 3 模型估计
- 4 回归参数的统计推断
- 5 模型诊断
- 6 预测
- 7 Key points

一个例子：广告投入



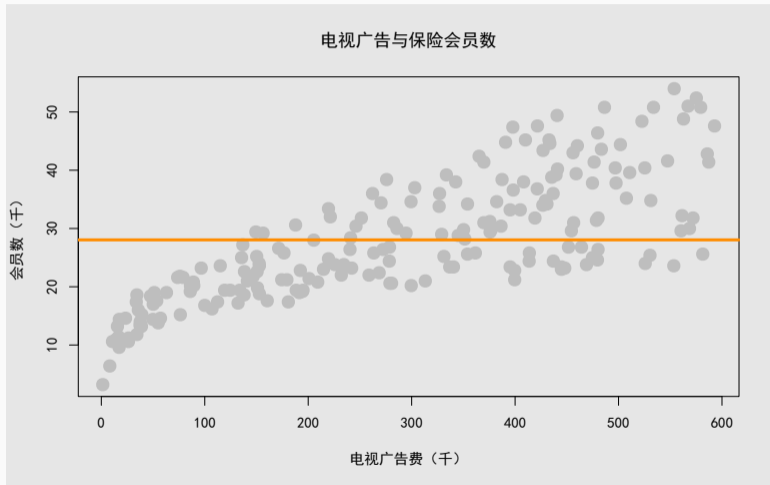
问题定义

- 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 $f()$?

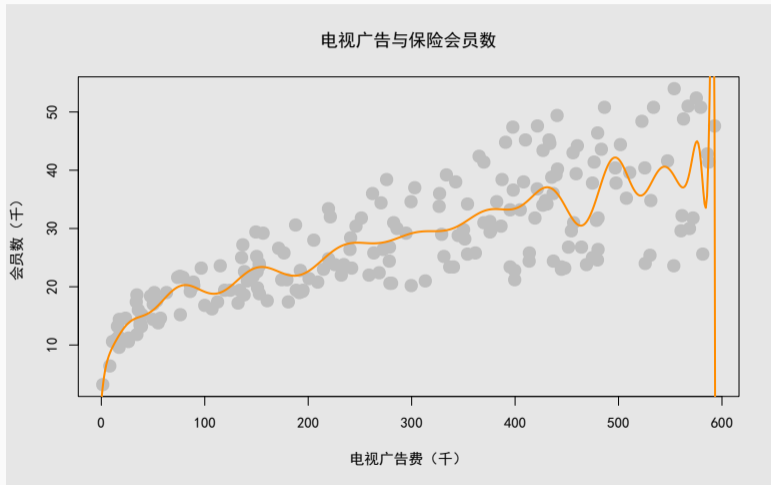
问题定义

- $\text{Response} = \text{Prediction} + \text{Error}$
- $\text{Response} = \text{Signal} + \text{Noise}$
- $\text{Response} = \text{Model} + \text{Unexplained}$
- $\text{Response} = \text{Deterministic} + \text{Random}$
- $\text{Response} = \text{Explainable} + \text{Unexplainable}$

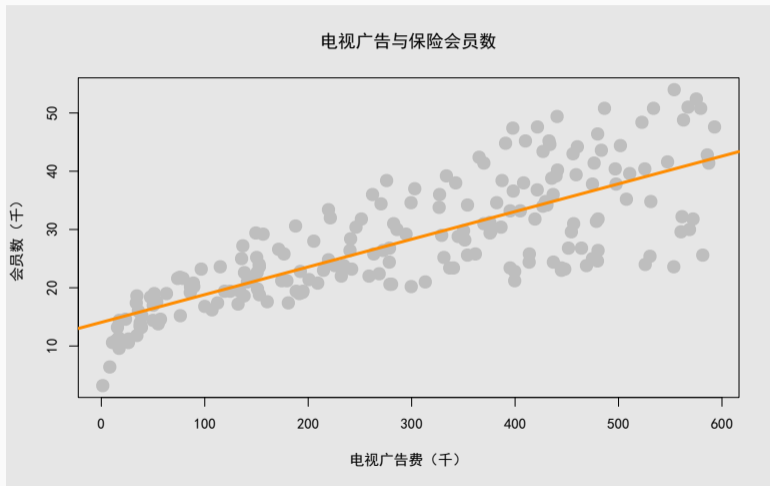
尝试 1



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 回归参数的统计推断
- 5 模型诊断
- 6 预测
- 7 Key points

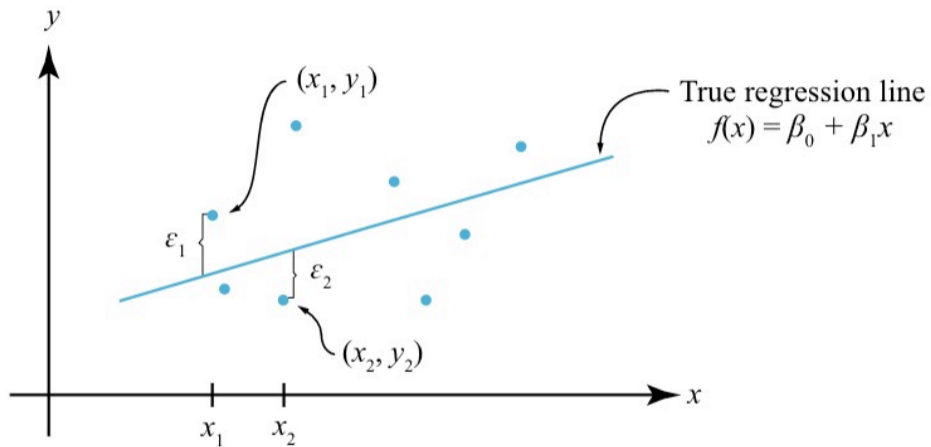
一元线性回归模型

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 $\epsilon_i \sim N(0, \sigma^2)$ 。

- 1 intercept coefficient β_0 : 当 $x = 0$ 时 Y 的期望, 即 $\beta_0 = E(Y|x = 0)$
- 2 slope coefficient β_1 : x 变动一个单位时对 Y 的平均影响
- 3 error term $\epsilon_i \sim N(0, \sigma^2)$, σ 未知
 - 上述回归也称为均值回归, 也就是说 $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$



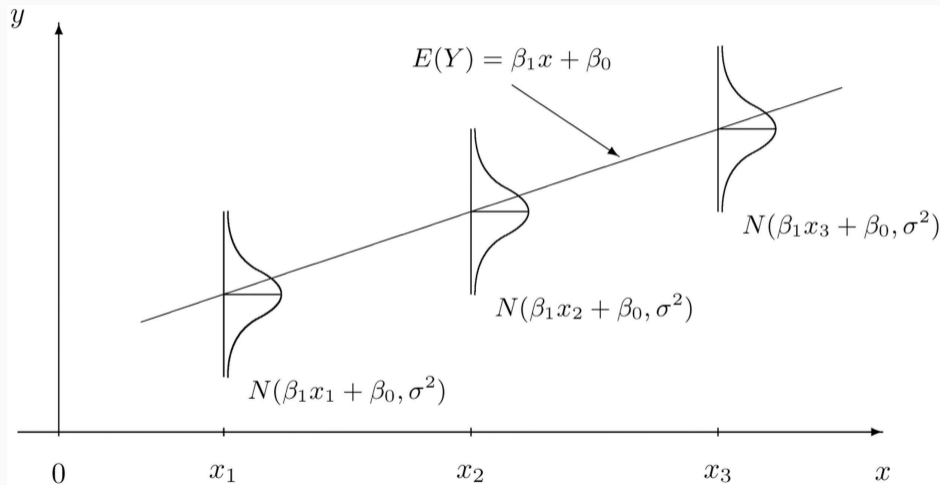
一元线性回归模型

- 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 β_0, β_1, σ 。
- 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

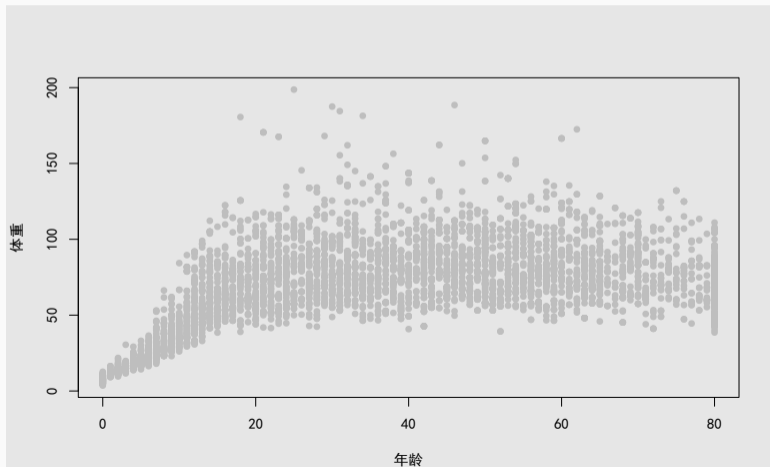


一元线性回归模型假设 (LINE)

- 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](#)。



随堂练习

假设应力 x (单位: kg/mm^2) 和断裂时间 y (单位小时) 之间的关系由简单线性回归模型描述, 其中 $\beta_0 = 65, \beta_1 = -1.2, \sigma = 8$, 则随着应力增加 $1\text{kg}/\text{mm}^2$, 平均 (或期望的) 断裂时间减少了 1.2 小时。请计算

- 1 应力 $x = 20$ 时, 断裂时间 $Y > 50$ 的概率
- 2 应力 $x = 25$ 时, 断裂时间 $Y > 50$ 的概率

注意到给定 x 时，断裂时间 $Y \sim N(65 - 1.2x, 8^2)$ ，那么

- 1 应力 $x = 20$ 时， $Y|x = 20 \sim N(65 - 1.2 * 20, 8^2) = N(41, 8^2)$ ，
那么

$$P(Y > 50|x = 20) = P\left(Z > \frac{50 - 41}{8}\right) = 1 - \Phi(1.13) = 0.1292$$

- 2 类似地，可以得到应力 $x = 25$ 时，断裂时间 $Y > 50$ 的概率

$$P(Y > 50|x = 25) = P\left(Z > \frac{50 - 35}{8}\right) = 1 - \Phi(1.88) = 0.0301$$

Outline

- 1 回归模型介绍
- 2 一元线性回归模型
- 3 模型估计**
- 4 回归参数的统计推断
- 5 模型诊断
- 6 预测
- 7 Key points

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 β_0 和 β_1 ?

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

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 \rightarrow 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

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

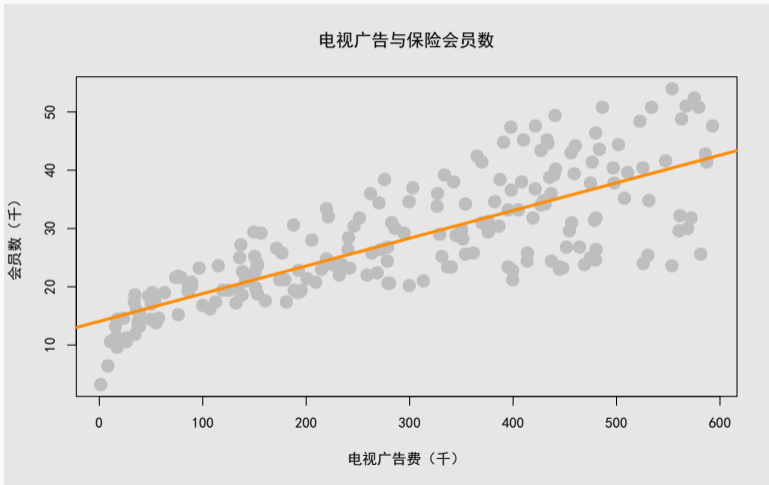
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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

在 R 中建立一元线性回归模型

```
marketing_model <- lm(Members ~ TV,  
                      data = marketing)  
plot(Members~TV, data = marketing,  
     xlab = " 电视广告费 (千) ",  
     ylab = " 会员数 (千) ",  
     main = " 电视广告与保险会员数",  
     pch  = 16,  
     cex  = 2,  
     col  = "grey")  
abline(marketing_model, lwd = 3,  
       col = "darkorange")
```

在 R 中建立一元线性回归模型



模型系数

```
coef(marketing_model)
```

```
## (Intercept)          TV  
## 14.06518710  0.04753664
```

回归方程:

$$\widehat{\text{会员数}} = 14.065 + 0.0475 \cdot \text{电视广告费}$$

残差 (Residuals) 和 σ^2 的估计

- σ^2 表示回归模型的波动性大小

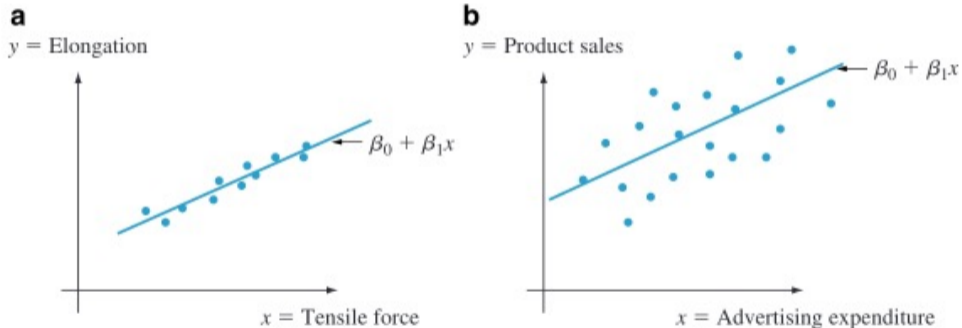


Figure 12.9 Typical sample for σ : (a) small; (b) large

■ 拟合 (fitted) 值: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x})$

■ 残差 (residual) 就是真实值和拟合值之间的差:

$$e_i = y_i - \hat{y}_i = y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}).$$

1 $\sum_{i=1}^n e_i = 0$, i.e. $\bar{e} = 0$

2 $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$

3 e_i 可以作为真实误差 ϵ_i 的“替代”，这时可称为估计误差

■ 残差平方和 (residuals sum of squares, SSE)

$$\text{SSE} = \sum_{i=1}^n (e_i - \bar{e})^2 = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

σ^2 的最小二乘估计

■ σ^2 的最小二乘估计

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

- ▶ $s_e = \sqrt{\text{SSE}/(n-2)}$ 为 σ 的估计, 称为残差标准误差

关于自由度为 $n - 2$ 的几点说明

1 计算 s_e^2 需要估计两个参数 β_0, β_1 (自由度损失 2)

2 s_e^2 是 σ^2 的无偏估计, 即

$$Es_e^2 = \sigma^2$$

3 事实上, $\frac{SSE}{\sigma^2} \sim \chi^2(n - 2)$, 且与 $\hat{\beta}_0, \hat{\beta}_1$ 独立

marketing_model\$residuals

##	1	2	3	4	5	6
##	8.25845097	2.50405190	2.89955247	8.53121085	-5.45443628	-0.49232464
##	7	8	9	10	11	12
##	4.06809925	0.90700454	-5.28281731	-11.86082862	-3.14953096	0.32257950
##	13	14	15	16	17	18
##	2.07206882	-3.93483198	4.53035628	12.15749382	4.48884446	7.98119167
##	19	20	21	22	23	24
##	1.95574187	1.13051863	1.17080836	-11.63558398	-4.12015441	-4.77041712
##	25	26	27	28	29	30
##	-0.58825250	-15.05995264	2.34884107	-5.09228183	0.08058062	0.22263927
##	31	32	33	34	35	36
##	0.88784894	-0.99896051	-4.10631000	-4.51665050	-4.16370008	-16.10298985
##	37	38	39	40	41	42
##	11.35975424	8.23283882	2.03715450	7.25810486	-0.11752647	3.30684219
##	43	44	45	46	47	48
##	-0.57870236	-7.93584891	0.54847355	-0.91251858	-1.39326039	9.52673282
##	49	50	51	52	53	54
##	-6.06583651	-1.02558959	-10.26082862	-2.21054450	10.56095492	10.97443182
##	55	56	57	58	59	60
##	1.35906202	14.42473734	-3.75922205	-0.61416795	13.49336530	2.70287262
##	61	62	63	64	65	66

估计误差项标准差

```
marketing_model_summary <- summary(marketing_model)
marketing_model_summary$sigma
```

```
## [1] 6.517313
```

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

- 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

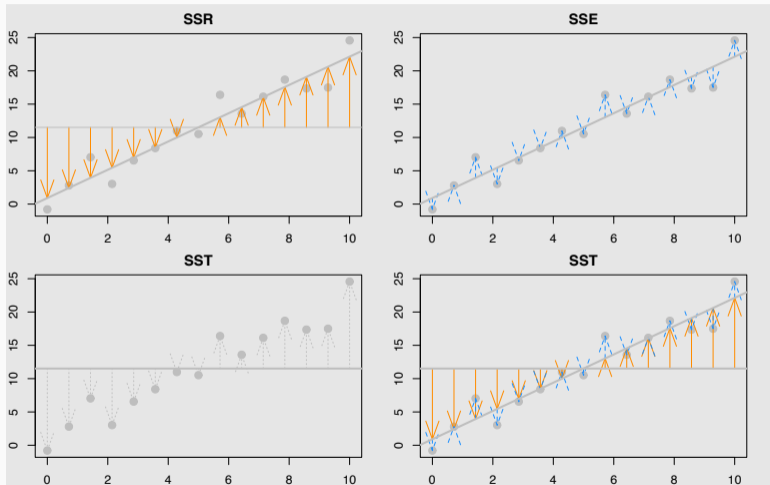
拟合优度 R^2

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

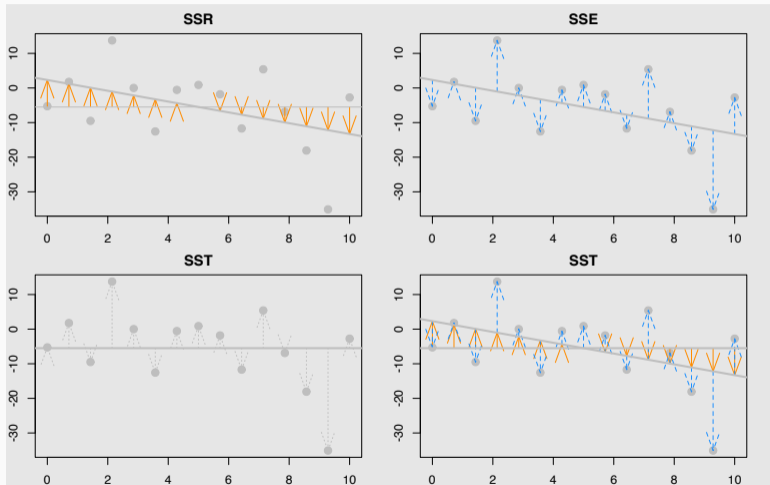
$$R^2 = \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

- 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 $R^2 = \hat{\beta}_1^2 \frac{S_{xx}}{S_{yy}}$

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

R 中求拟合优度

```
marketing_model_summary$r.squared
```

```
## [1] 0.6118751
```

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？
 - 1 和 2 属于模型统计推断的内容：假设检验或置信区间
 - 3 属于模型诊断的内容

Outline

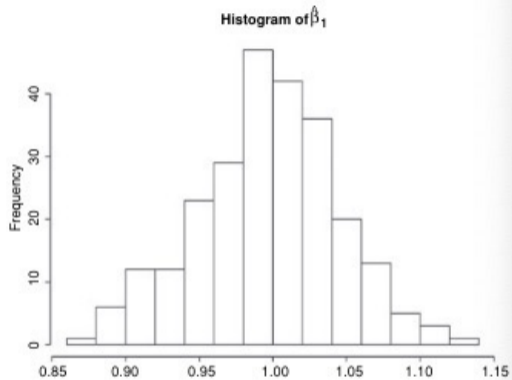
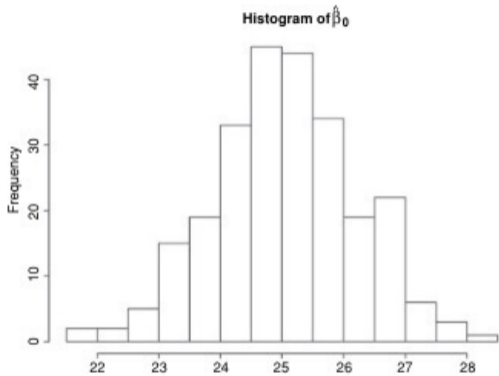
- 1 回归模型介绍
- 2 一元线性回归模型
- 3 模型估计
- 4 回归参数的统计推断**
- 5 模型诊断
- 6 预测
- 7 Key points

R 数值模拟： $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的抽样分布

假设回归模型 $Y_i = 25 + x_i + \epsilon_i$, $\epsilon_i \sim N(0, 1.4^2)$, $i = 1, 2, \dots, 19$, x 的取值为

12, 14, 14, 15, 15, 16, 18, 22, 22, 24, 24, 26, 26, 27, 28, 30, 30, 33, 36

- 1 生成随机误差项 $\epsilon_i \sim N(0, 1.4^2)$
- 2 根据回归方程，用 x_i 生成 y_i
- 3 对于模拟的数据 $(x_i, y_i)_{i=1}^{19}$ 估计回归系数
- 4 重复 1-3 步 250 次，分别画出 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的直方图



Gauss-Markov 定理

如果基本假设成立，最小二乘估计量是总体参数 β_0 和 β_1 的线性最小方差无偏估计量。

- $\hat{\beta}_1$ 可以写为:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n k_i Y_i,$$

其中 $k_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

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

$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 都是线性估计量。

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 σ 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：

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

标准误差 (Standard Errors, SE)

$$\begin{aligned}\frac{\hat{\beta}_1 - \beta_1}{\text{SE}[\hat{\beta}_1]} &= \frac{\hat{\beta}_1 - \beta_1}{s_e / \sqrt{S_{xx}}} \\ &= \frac{\hat{\beta}_1 - \beta_1}{s_e / \sqrt{S_{xx}}} \cdot \frac{\sigma / \sqrt{S_{xx}}}{\sigma / \sqrt{S_{xx}}} \\ &= \frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \cdot \frac{\sigma / \sqrt{S_{xx}}}{s_e / \sqrt{S_{xx}}} \\ &= \frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} / \sqrt{\frac{s_e^2}{\sigma^2}} \\ &= \frac{\hat{\beta}_1 - \beta_1}{\text{SD}[\hat{\beta}_1]} / \sqrt{\frac{(n-2)s_e^2}{\sigma^2}} \sim \frac{Z}{\sqrt{\frac{\chi_{n-2}^2}{n-2}}} \sim t_{n-2}\end{aligned}$$

β_0 和 β_1 的置信区间

β_0 的置信区间:

$$\hat{\beta}_0 \pm t_{1-\alpha/2}(n-2) \cdot \text{SE} [\hat{\beta}_0],$$

β_1 的置信区间:

$$\hat{\beta}_1 \pm t_{1-\alpha/2}(n-2) \cdot \text{SE} [\hat{\beta}_1],$$

其中 $P(t(n-2) > t_{1-\alpha/2}(n-2)) = \alpha/2$.

R 中求置信区间

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

```
##                2.5 %        97.5 %  
## (Intercept) 12.25943854 15.87093566  
## TV          0.04223072  0.05284256
```

假设检验

以 β_0 为例，对于 β_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

```
marketing_model_summary$coefficients
```

```
##              Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 14.06518710 0.915685881 15.36028 1.40630e-35
## TV          0.04753664 0.002690607 17.66763 1.46739e-42
```

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

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

注意：在一元线性回归中，F 检验和回归系数 β_1 的 t 检验是等价的。

R 中的 F 检验

```
anova(marketing_model)
```

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

```
##
```

```
## Response: Members
```

```
##           Df  Sum Sq Mean Sq F value    Pr(>F)
```

```
## TV           1 13258.5 13258.5   312.14 < 2.2e-16 ***
```

```
## Residuals 198  8410.1    42.5
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Outline

- 1 回归模型介绍
- 2 一元线性回归模型
- 3 模型估计
- 4 回归参数的统计推断
- 5 模型诊断**
- 6 预测
- 7 Key points

还记得一元线性回归模型的假设吗？

Assumption	In terms of Y	In terms of ε
Linearity	$E(Y x)$ is a linear function of x .	For any fixed x , $E(\varepsilon) = 0$.
Normality	For any fixed x , the Y distribution is normal.	For any fixed x , the rv ε is normally distributed.
Constant variance	The variance of Y at any fixed x value is independent of x .	$V(\varepsilon) = \sigma^2$, independent of x .
Independence	Y_i 's for different observations are independent.	ε_i 's for different observations are independent.

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

残差分析

残差 $e_i = Y_i - \hat{y}_i$

1 $E(e_i) = 0$

2 $\text{Var}(e_i) = \sigma^2 \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}} \right]$

3 $\widehat{\text{Var}}(e_i) = s_e^2 \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}} \right]$

4 标准化残差

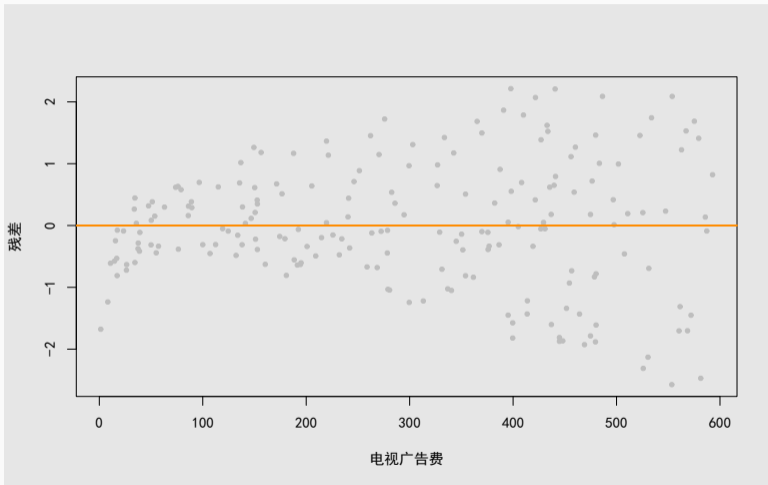
$$e_i^* = \frac{e_i - 0}{s_{e_i}} = \frac{y_i - \hat{y}_i}{s_e \sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}}$$

当 n 充分大时, $e_i^* \approx \frac{e_i}{s_e}$

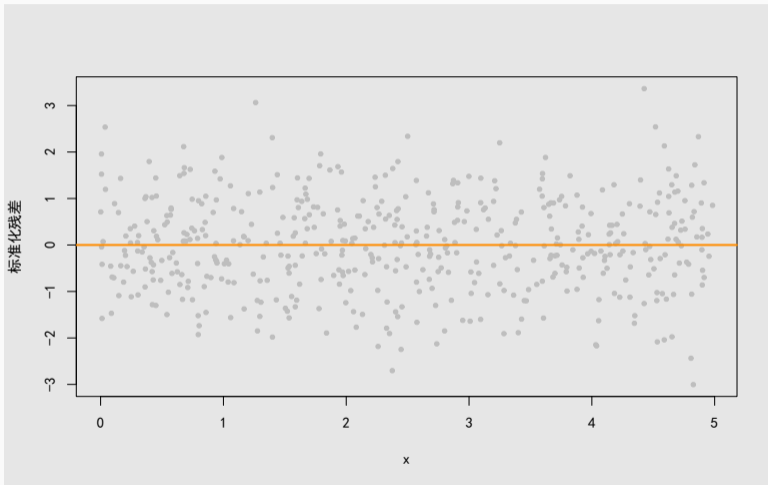
标准化残差图

- Plots of (x_i, e_i^*) or (\hat{y}_i, e_i^*) : 拟合程度、线性性和方差齐性
- 残差应该大致随机分布在通过 0 的水平线周围, 不含任何趋势

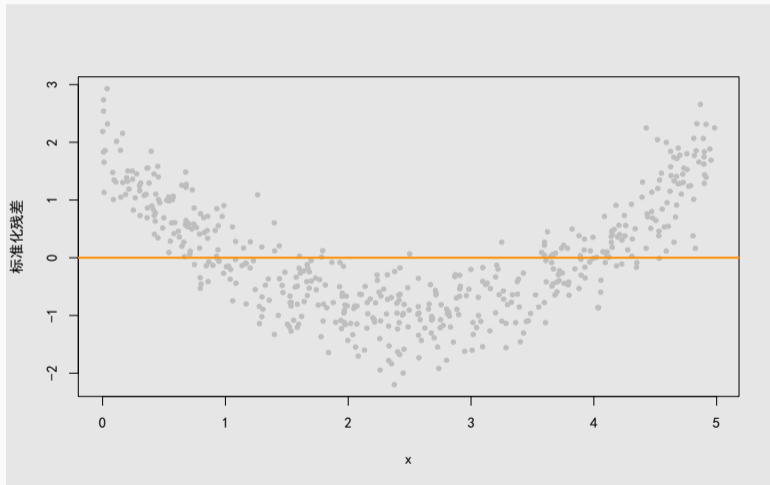
标准化残差图



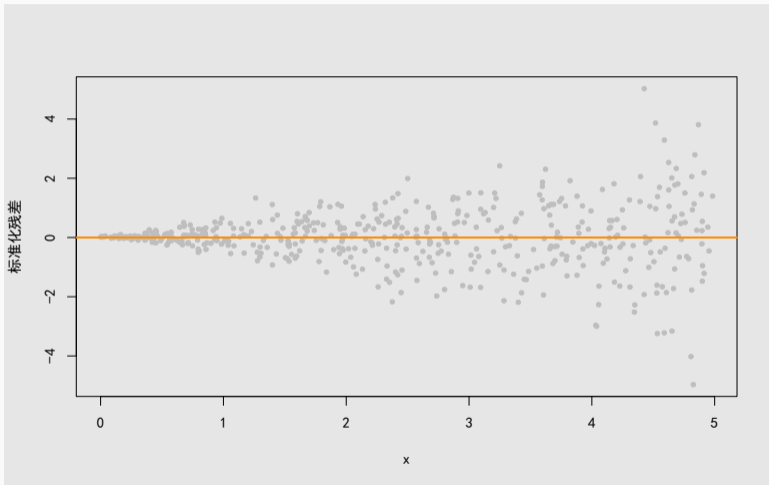
拟合良好



非线性

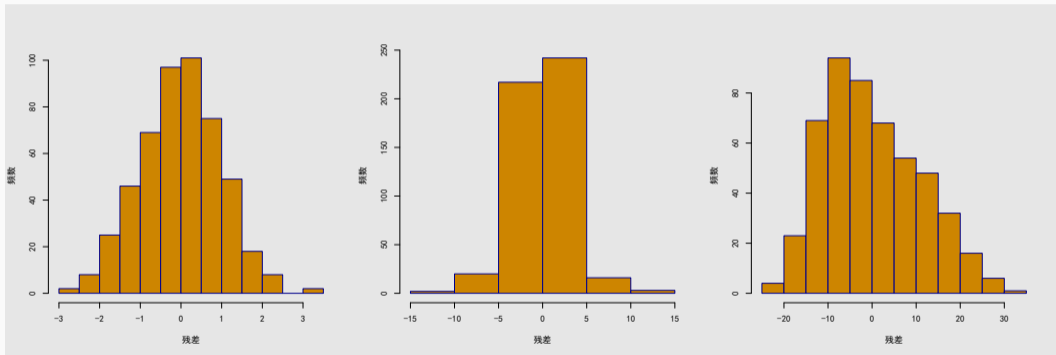


异方差



- 1 直方图
- 2 PP 图或者 QQ 图
- 3 正态性假设检验

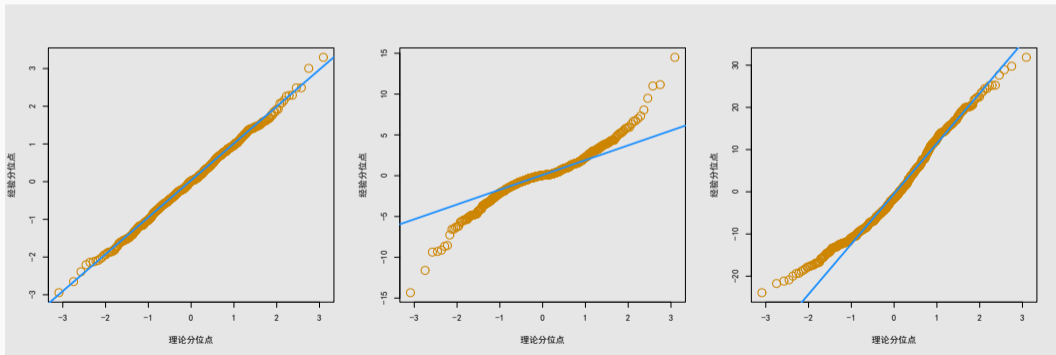
正态性-直方图



正态性 - PP 图或 QQ 图

1 PP 图 (Percentile-Percentile)

2 QQ 图 (Quantile-Quantile)



正态性检验

Shapiro 检验，请点击[这里](#)查看更多检验理论细节。

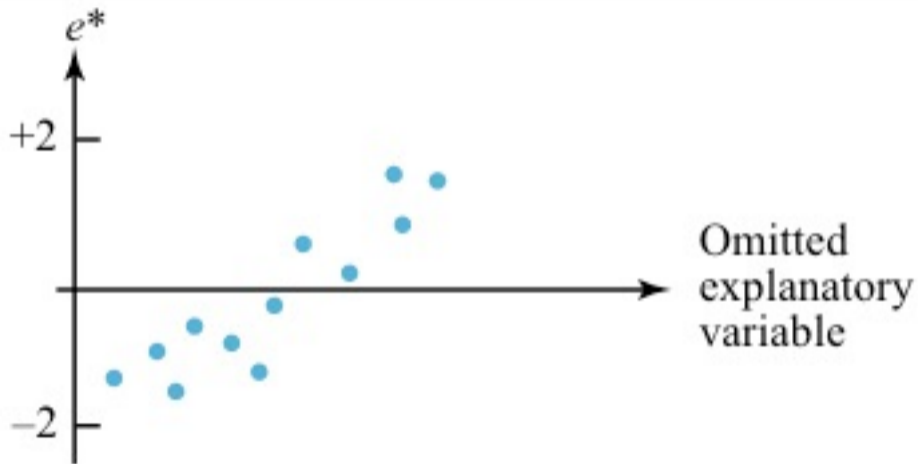
```
shapiro.test(marketing_model$residuals)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data:  marketing_model$residuals  
## W = 0.99053, p-value = 0.2133
```

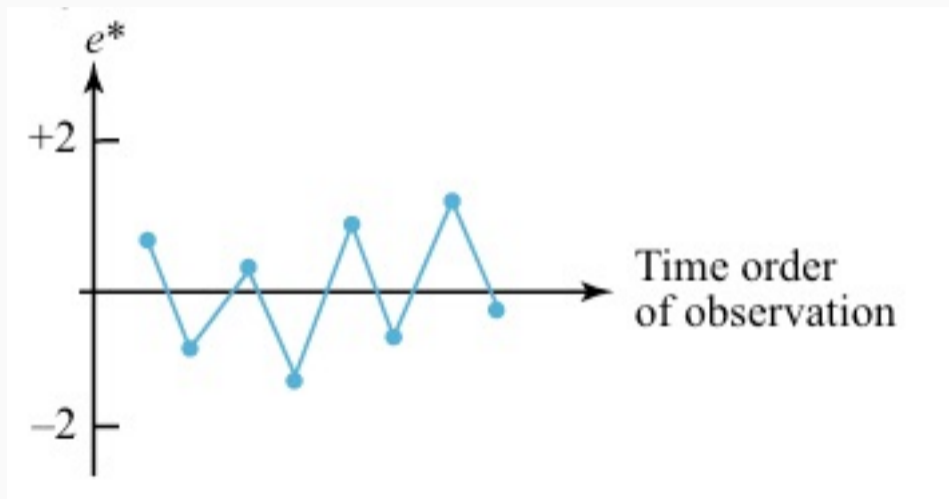
不满足“残差独立”的假设，可从残差图可以看出。产生的主要原因有：

- 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

重要解释变量被遗漏



时间序列数据：误差自相关（交替规律）



如何解决？

- 线性假设不成立：非线性模型（比如，多项式回归模型）或者非参数模型
- 异方差：方差稳定性变换；加权最小二乘估计；位置-刻度回归；分位数回归等等
- 正态假设不成立：大样本时不重要；小样本时，可以采用 Bootstrap 方法进行推荐推断

■ 异常点或者强影响点：

1 去掉后重新回归

2 最小绝对离差估计 (least absolute deviation estimates, LAD)

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min \sum_{i=1}^n |Y_i - \beta_0 - \beta_1 x_i|$$

3 Robust estimation, eg., Huber regression.

Outline

- 1 回归模型介绍
- 2 一元线性回归模型
- 3 模型估计
- 4 回归参数的统计推断
- 5 模型诊断
- 6 预测
- 7 Key points

- 给定 $x = x^*$ ，点估计或点预测为 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ 。
- $\mu_{Y|x^*} = E(Y|X = x^*) = \beta_0 + \beta_1 x^*$ 。
- \hat{y} 作为 $E(Y|X = x^*)$ 的点估计（无偏估计），因为 $E(\hat{y}) = \beta_0 + \beta_1 x^*$ 。
- 下面给出 $\mu_{Y|x^*}$ 的置信区间（confidence interval）和 Y 的预测区间（prediction interval）

线性预测 (估计) 量

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x^* = \bar{Y} + \hat{\beta}_1 (x^* - \bar{x}) \\ &= \bar{Y} + \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})(x^* - \bar{x})}{S_{xx}} \\ &= \sum_{i=1}^n \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x^* - \bar{x})}{S_{xx}} \right] Y_i \\ &= \sum_{i=1}^n d_i Y_i,\end{aligned}$$

其中 $d_i = \frac{1}{n} + \frac{(x_i - \bar{x})(x^* - \bar{x})}{S_{xx}}, i = 1, 2, \dots, n$

置信区间

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

置信区间

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

R 中求置信区间

```
new_TV = data.frame(TV = c(500, 11))  
predict(marketing_model, newdata = new_TV,  
        interval = c("confidence"), level = 0.99)
```

```
##           fit      lwr      upr  
## 1 37.83351 35.95918 39.70783  
## 2 14.58809 12.27269 16.90349
```

预测区间

给定一个新的 x^* , 我们想预测 Y , $Y = \beta_0 + \beta_1 x^* + \epsilon$

- 预测误差 $\hat{y} - Y = \hat{y} - (\beta_0 + \beta_1 x^* + \epsilon)$

Because the future value Y is independent of the observed Y_i 's that determine \hat{y} ,

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

预测区间

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

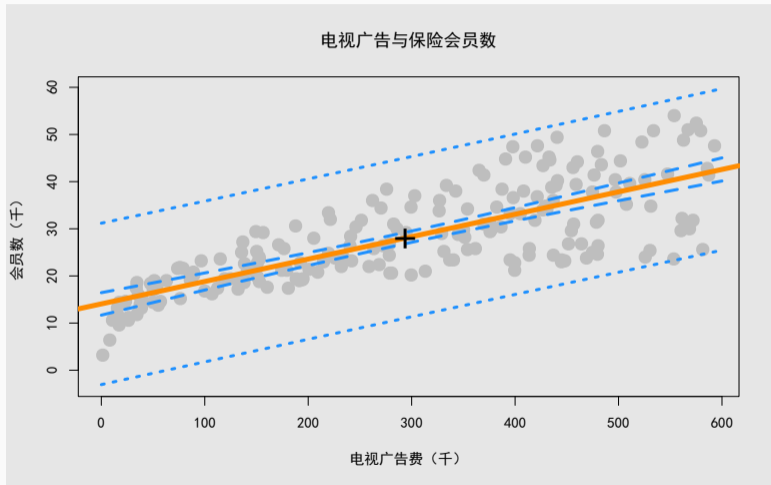
- 预测区间比置信区间的方差大，表现在区间长度稍宽

R 中求预测区间

```
new_TV = data.frame(TV = c(500, 11))  
predict(marketing_model, newdata = new_TV,  
        interval = c("prediction"), level = 0.99)
```

```
##           fit           lwr           upr  
## 1 37.83351 20.779400 54.88761  
## 2 14.58809 -2.520111 31.69629
```

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 回归参数的统计推断
- 5 模型诊断
- 6 预测
- 7 Key points

Key points

- 1 理解并掌握一元线性回归模型、模型假设、最小二乘方法等；
- 2 理解并掌握最小二乘估计的统计性质
- 3 熟悉掌握对回归系数、回归方程和响应变量的统计推断：假设检验和置信区间
- 4 理解并掌握模型诊断方法
- 5 掌握如何利用 R 进行一元线性回归模型

上机实验（一）

- 数据 `poverty.txt`: 美国的 50 个州和哥伦比亚特区搜集到的 15-17 岁女性在 2002 年出生率及其该州的贫困率

- 1 y = 2002 年每 1000 名 15 至 17 岁女性的出生率

- 2 x = 贫困率，即州人口中生活在低于联邦规定贫困水平的家庭的百分比

- 要求：

- 1 给出散点图说明两者之间的相关性关系，并计算相关系数及其显著性检验

- 2 给出一元回归模型结果，并针对结果给出解释

- 3 进行模型诊断

上机实验（二）

- 数据 `skincancer.txt`: 响应变量 y 是皮肤癌死亡率（每 1000 万人口死亡人数），预测变量 x 是美国 48 个州中心的纬度（北纬度）
- 要求：
 - 1 给出散点图说明两者之间的相关性关系，并计算相关系数及其显著性检验
 - 2 给出一元回归模型结果，并针对结果给出解释
 - 3 进行模型诊断