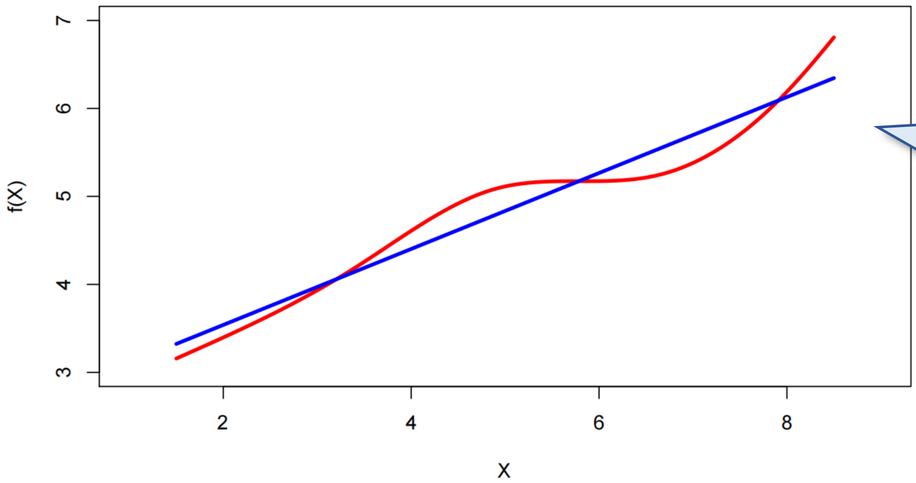
第三章 线性回归

1. 简单线性回归

- 线性回归 (linier regression) 是一种简单的监督学习方法。它假定预测变量$X=\{X_1, X_2, \cdots, X_p\}$和响应变量$Y$之间存在线性关系

01 线性回归定义

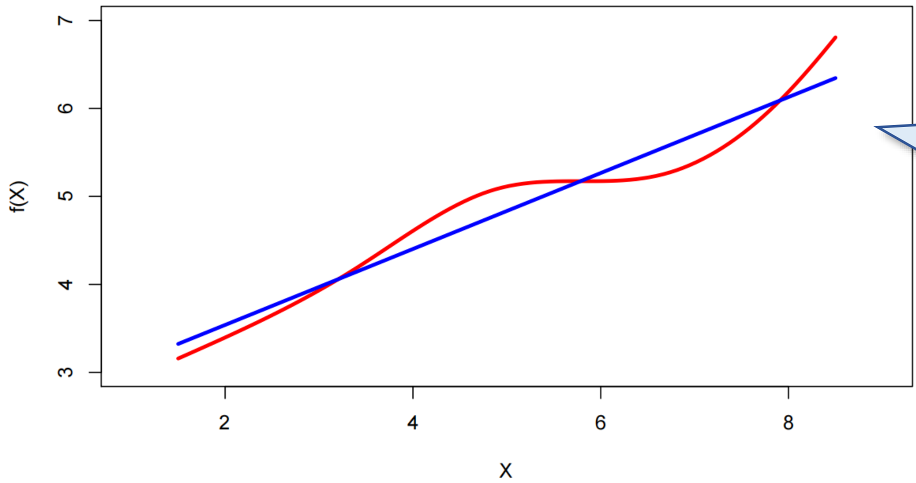
- 线性回归 (linier regression) 是一种简单的监督学习方法。它假定预测变量$X=\{X_1,X_2,\cdots,X_p\}$和响应变量Y之间存在线性关系



X和Y的真实关系（回归函数）往往并非线性

01 线性回归定义

- 线性回归 (linier regression) 是一种简单的监督学习方法。它假定预测变量$X=\{X_1,X_2,\cdots,X_p\}$和响应变量Y之间存在线性关系

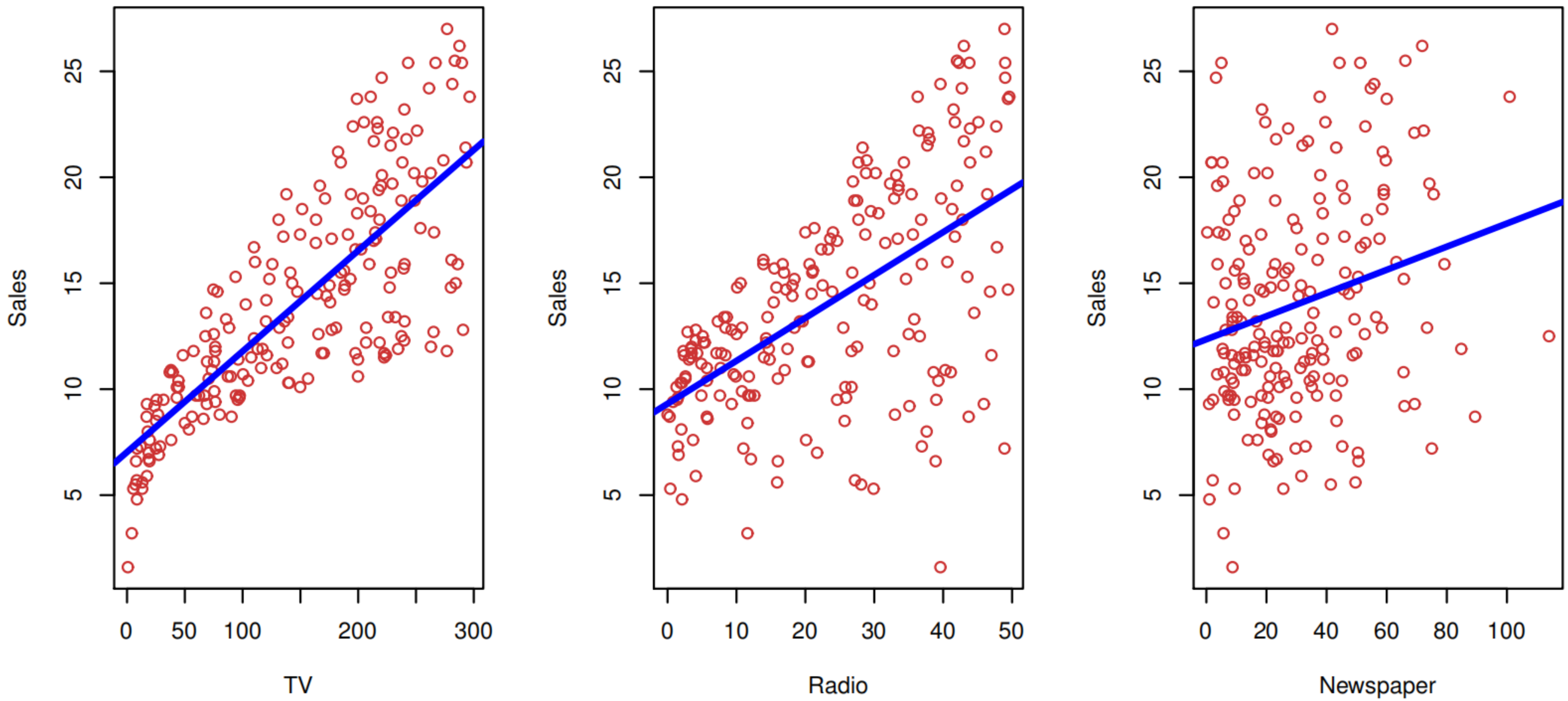


X和Y的真实关系（回归函数）往往并非线性

- 尽管这种假设可能把问题过分简化，但在概念上和实践中，线性回归都非常有用

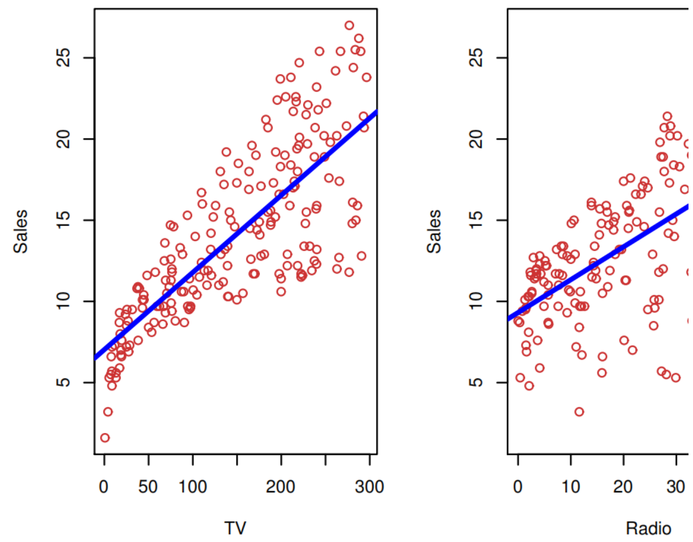
01 线性回归用于Advertising数据分析

- Advertising（广告）数据集记录了某产品在200个不同市场的销量情况及在每个市场中3类广告媒体的预算，分别为TV（电视）、radio（广播）和newspaper（报纸）



线性回归用于Advertising数据分析

- 假设我们的角色是统计咨询师，需要根据这一数据提出一份营销计划，提高明年的产品销量，可能需要考虑：
  - 广告预算和销量有关吗?
  - 广告预算和销量间的关系有多强?
  - 哪种媒体能促进销售?
  - 如何精确地估计每种媒体对销量的影响?
  - 对未来销量的预测精度如何?
  - 这种关系是否是线性的?
  - 广告媒体间是否存在协同效应?



- 用本章学习的线性回归模型，可以回答这些问题。

- 我们假设一种非常简单的，根据单一预测变量X预测定量响应变量Y的方法。假定两者存在线性关系，记为：

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

$$Sale = \beta_0 + \beta_1 TV$$

- 其中， β_0和 β_1是两个未知的常量，分别表示线性模型中的截距和斜率。 β_0和 β_1被称为模型的系数(coefficient) 或参数(parameter)。 ϵ 是误差项

- 我们假设一种非常简单的，根据单一预测变量X预测定量响应变量Y的方法。假定两者存在线性关系，记为：

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

$$Sale = \beta_0 + \beta_1 TV$$

- 其中， β_0和 β_1是两个未知的常量，分别表示线性模型中的截距和斜率。 β_0和 β_1被称为模型的系数(coefficient) 或参数(parameter)。 ϵ 是误差项

- 一旦使用训练数据估计出模型系β_0和 β_1 ，我们就可以根据给定的电视广告费，通过计算

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

- 预测未来的销量。其中$\hat{y}$表示在$X = x$ 的基础上对Y的预测。“^”表示对一个未知的参数或系数的估计值，或表示响应变量的预测值。

- 根据变量X的第i个值，用 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$来估计$Y$。 $e_i = y_i - \hat{y}_i$ 代表第i 个**残差**（观测值与预测值的距离）。

估计系数——最小二乘估计

- 根据变量X的第i个值，用 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$来估计$Y$。 $e_i = y_i - \hat{y}_i$ 代表第i 个**残差**（观测值与预测值的距离）。
- 定义**残差平方和**（residual sum of squares RSS）为:

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2,$$

或等价地定义为:

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

估计系数——最小二乘估计

- 根据变量X的第i个值，用 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$来估计$Y$。 $e_i = y_i - \hat{y}_i$ 代表第i 个**残差**（观测值与预测值的距离）。
- 定义**残差平方和**（residual sum of squares RSS）为：

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2,$$

或等价地定义为：

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

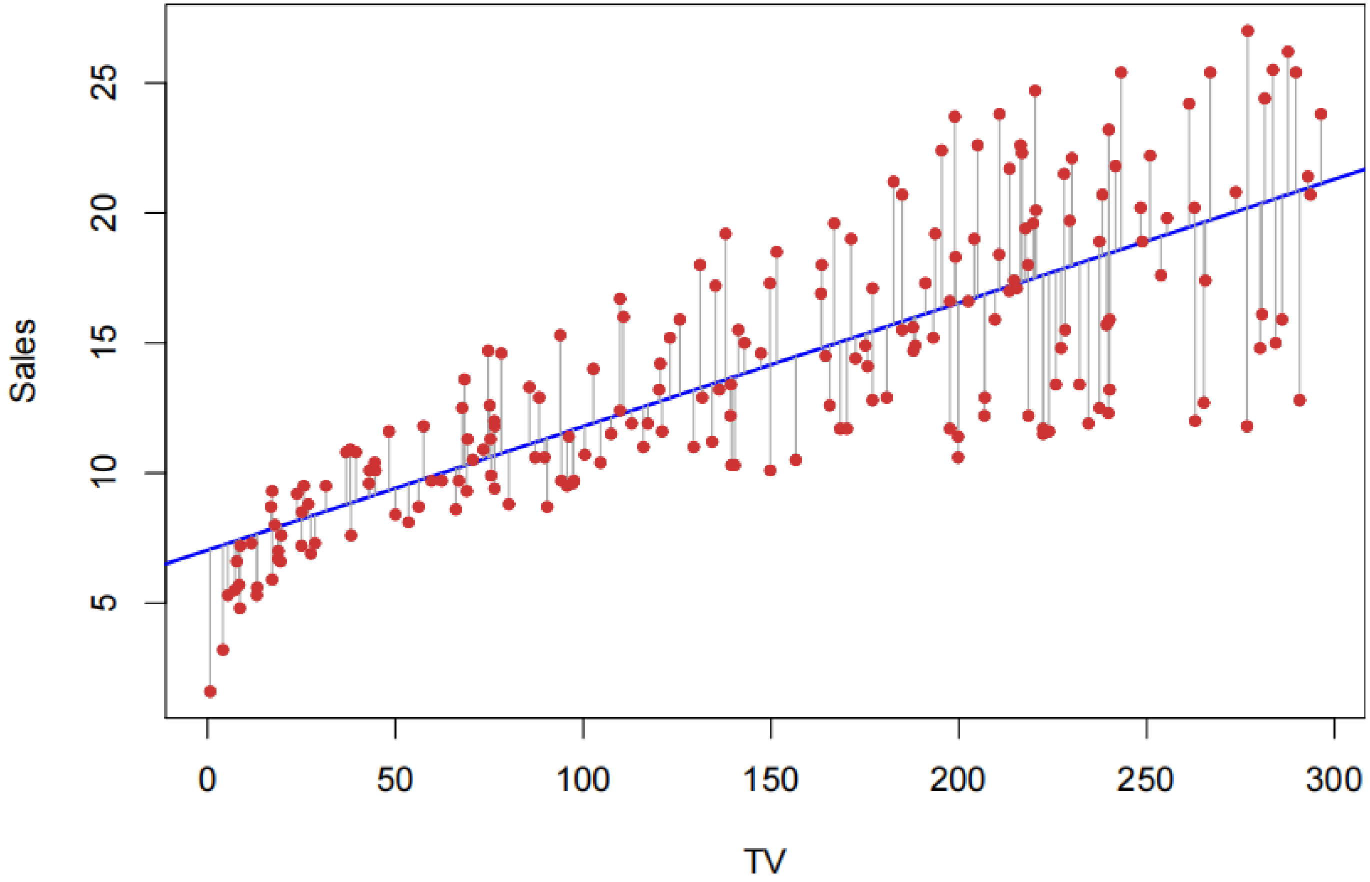
- 最小二乘法选择β_0和β_1 来使RSS达到最小。通过微积分运算可知，使RSS最小的参数估计值为：

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

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

这里$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ 和$\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ 是样本均值。

示例：广告数据



- 对于Advertising数据集，最小二乘法拟合sales关于TV的回归。这种拟合是通过使残差平方和最小化得到的，每条线段代表一个残差。这里的线性拟合抓住了变量间关系的本质，尽管它对图中左侧区域的拟合稍有缺陷。

- **标准误差(*standard error*)**告诉我们一个估计值偏离其实际值的平均量。可以用标准误差探究$\hat{\beta}_0$和$\hat{\beta}_1$与真实值β_0和β_1的接近程度：

$$SE(\hat{\beta}_1)^2=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2},\quad SE(\hat{\beta}_0)^2=\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right),$$

其中$\sigma^2=\mathrm{Var}(\sigma)$：

- **标准误差(standard error)**告诉我们一个估计值偏离其实际值的平均量。可以用标准误差探究$\hat{\beta}_0$和$\hat{\beta}_1$与真实值β_0和β_1的接近程度：

$$SE\left(\hat{\beta}_1\right)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \quad SE\left(\hat{\beta}_0\right)^2 = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right),$$

其中$\sigma^2 = \mathrm{Var}(\sigma)$：

- 标准误差可用于计算置信区间（confidence interval）。95%置信区间被定义为一个取值范围：该范围有95%的概率会包含未知参数的真实值。此范围是根据从样本数据计算出的上下限来定义的。对于线性回归模型， β_1 的95%置信区间约为:

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

- 也就是说，下述区间

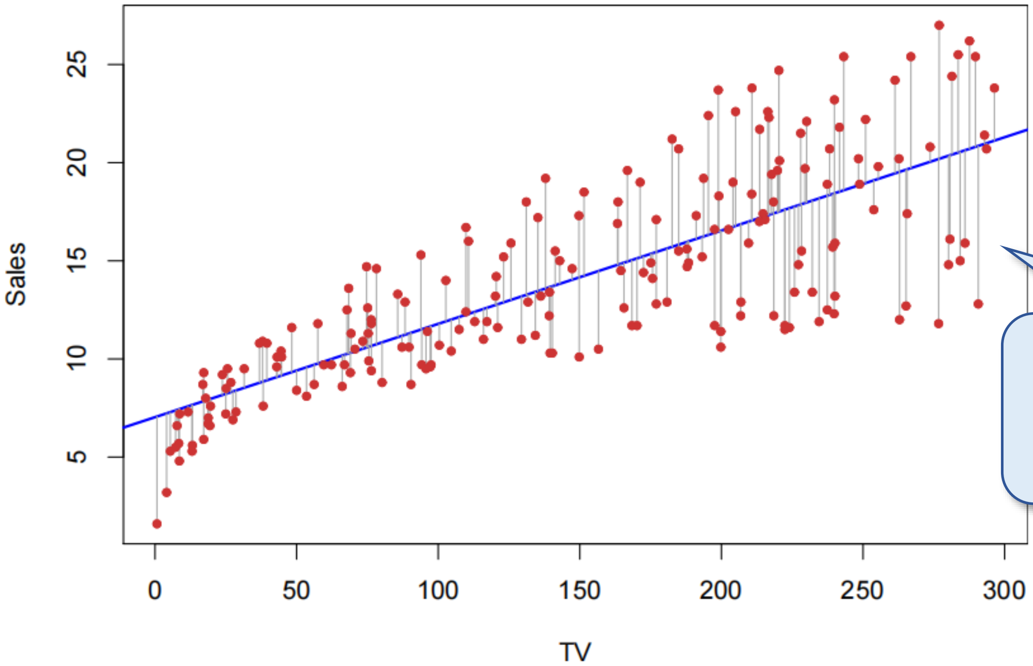
$$[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)]$$

- 有约95%的可能会包含β_1 的真实值。同样，β_0的95%置信区间约为：

$$\hat{\beta}_0 \pm 2 \cdot SE(\hat{\beta}_0)$$

置信区间

- 在Advertising数据的例子中，β_0的95%置信区间为$[6.130, 7.935]$，β_1的95%置信区间为$[0.042, 0.053]$
- 我们可以得出结论，在没有任何广告的情况下，销售平均会下降至6130到7935单位。此外，电视广告每增加一千美元，销售的平均增加值将在 42到53个单位之间。



$$Y = \beta_0 + \beta_1 X$$
$$Sale = \beta_0 + \beta_1 TV$$

- 标准误差也可以用来对系数进行**假设检验(*hypothesis tests*)**。最常用的假设检验包括对**零假设(*null hypothesis*)**：

$$H_0\text{：}X\text{和}Y\text{之间没有关系}$$

- 和**备择假设(*alternative hypothesis*)** 进行检验

$$H_1\text{：}X\text{和}Y\text{之间有一定的关系}$$

- 标准误差也可以用来对系数进行**假设检验(*hypothesis tests*)**。最常用的假设检验包括对**零假设(*null hypothesis*)**：

$$H_0\text{：}X\text{和}Y\text{之间没有关系}$$

- 和**备择假设(*alternative hypothesis*)** 进行检验

$$H_1\text{：}X\text{和}Y\text{之间有一定的关系}$$

- 数学上来说，这就相当于检验

$$H_0{:}\,\beta_1=0\text{和}H_1{:}\,\beta_1\neq 0$$

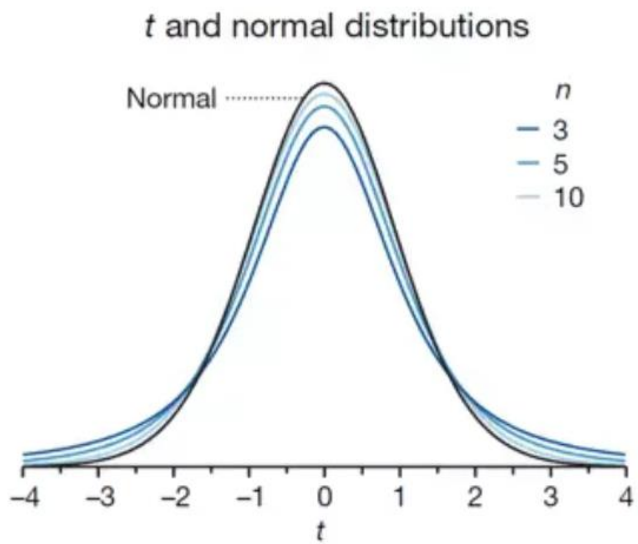
- 因为如果$\beta_1=0$，则模型简化为$Y=\beta_0+\varepsilon$，且X与Y不相关

01 假设检验

- 实践中，为了检验零假设，我们计算**t统计量(*t*-statistic)**

$$t=\frac{\hat{\beta}_1-0}{SE(\hat{\beta}_1)}$$

- 如果$\beta_1=0$，那我们预期它将服从自由度为$n-2$的t分布
- 利用统计软件，假设$\beta_1=0$，很容易计算任意观测值大于等于$|t|$的概率，我们称这个概率为**p值(*p*-value)**



t-分布（*t*-distribution）用于根据小样本来估计呈正态分布且方差未知的总体的均值

Advertising数据集上计算结果

大连理工大学 Dalian University of Technology

- 对于Advertising数据，下表是销量对电视广告预算的最小二乘回归模型的系数。电视广告预算每增加一千美元，销量增加约50个单位。（sales变量是以一千台为单位，而TV变量是以一千美元为单位。）

	系数	标准误差	t统计量	p值
	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

$Sale = \beta_0 + \beta_1 TV$

$\hat{\beta}_0 \pm 2 \cdot SE(\hat{\beta}_0)$
$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$

$H_0: \beta_0 = 0?$
$H_0: \beta_1 = 0?$

- 一旦我们拒绝零假设，就会很自然地想要量化模型拟合数据的程度。判断线性回归的拟合质量通常使用两个相关的量：
  - 残差标准误 (residual standard error , RSE)
  - R^2统计量

- 我们计算**残差标准误 (residual standard error , RSE)**

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

其中，**残差平方和**$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- 我们计算**残差标准误 (residual standard error , RSE)**

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

其中，**残差平方和**$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- ***R^2*统计量**用下列公式计算：

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

其中，$\mathrm{TSS} = \sum(y_i - \bar{y})^2$ 是**总平方和(total sum of squares)**

- 我们计算**残差标准误 (residual standard error , RSE)**

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

其中，**残差平方和**$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- ***R^2*统计量**用下列公式计算：

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

其中，$\mathrm{TSS} = \sum(y_i - \bar{y})^2$ 是**总平方和(total sum of squares)**

- 事实上，在简单线性回归模型中，$R^2 = r^2$，其中r是X和Y之间相关性：

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- 关于Advertising 数据集上销量对于电视广告预算的最小二乘回归模型的更多信息

	Quantity	Value
残差标准误	Residual Standard Error	3.26
R^2统计量	R^2	0.612

代码实现——简单线性回归

- MASS库中包含Boston　(波士顿房价)数据集，它记录了波士顿周围506个街区的medv(房价中位数)。我们将设法用13个预测变量如rm（每栋住宅的平均房间数)，age　(平均房龄)，lstat(社会经济地位低的家庭所占比例)等来预测medv(房价中位数)。

```
> library(MASS)
> library(ISLR)
> fix(Boston)
```

数据编辑器

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
6	0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
7	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
8	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
9	0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
10	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9
11	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15
12	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9

01 代码实现——以Boston数据集为例

大连理工大学 Dalian University of Technology

```
> library(ISLR)
> fix(Boston)
> names(Boston)
 [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"     "rad"     "tax"
[11] "ptratio" "black"   "lstat"   "medv"
> lm.fit=lm(medv~lstat)
Error in eval(predvars, data, env) : 找不到对象'medv'
> lm.fit=lm(medv~lstat,data=Boston)
> attach(Boston)
> lm.fit=lm(medv~lstat)
> lm.fit

Call:
lm(formula = medv ~ lstat)

Coefficients:
(Intercept)        lstat
      34.55        -0.95

> summary(lm.fit)

Call:
lm(formula = medv ~ lstat)

Residuals:
    Min      1Q  Median      3Q     Max
-15.168  -3.990  -1.318   2.034  24.500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,    Adjusted R-squared:  0.5432
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

从用lm()函数拟合一个简单线性回归模型开始，将lstat作为预测变量，medv作为响应变量。基本句法是 lm(y ~ x , data)， 其中y是响应变量，x是预测变量，data是这两个变量所属的数据集。

输入 lm. fit 指令，则会输出模型的一些基本信息。

用 summary(lm. fit)函数 了解更多详细信息。运行这条命令将输出系数的p值和标准误，以及模型的R^2统计量和F统计量。

代码实现——简单线性回归

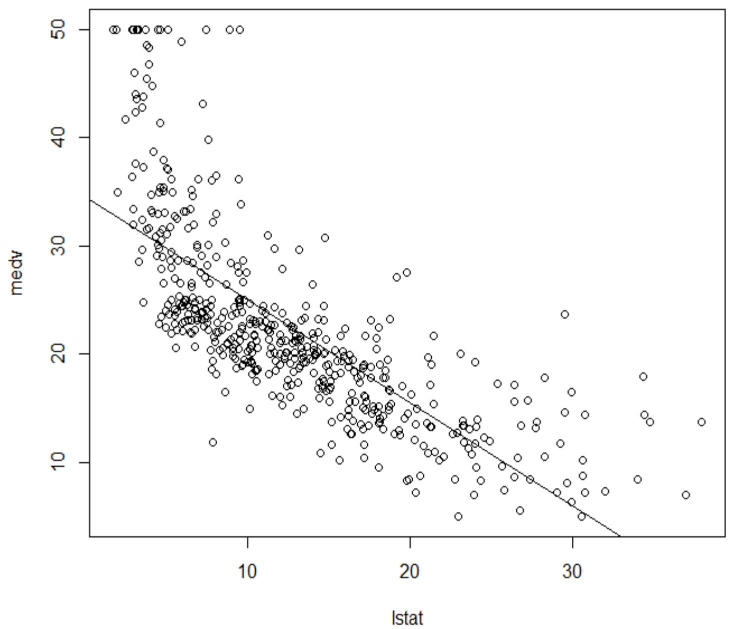
```
> names(lm.fit)
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"   "xlevels"       "call"          "terms"         "model"
> coef(lm.fit)
(Intercept)       lstat
 34.5538409  -0.9500494
> confint(lm.fit)
                2.5 %     97.5 %
(Intercept) 33.448457 35.6592247
lstat       -1.026148 -0.8739505
> predict(lm.fit,data.frame(lstat=(c(5,10,15))), interval="confidence")#计算置信区间
       fit      lwr      upr
1 29.80359 29.00741 30.59978
2 25.05335 24.47413 25.63256
3 20.30310 19.73159 20.87461
> predict(lm.fit,data.frame(lstat=(c(5,10,15))), interval="prediction")#计算预测区间
       fit       lwr      upr
1 29.80359 17.565675 42.04151
2 25.05335 12.827626 37.27907
3 20.30310  8.077742 32.52846
```

可以使用 names ()函数 找出 1m. fit 中存储的其他信息。虽然可以用名称提取这些量，例如：lm. fit $ coefficients . 但用提取功能如 coef ()函数 访问它们会更安全。

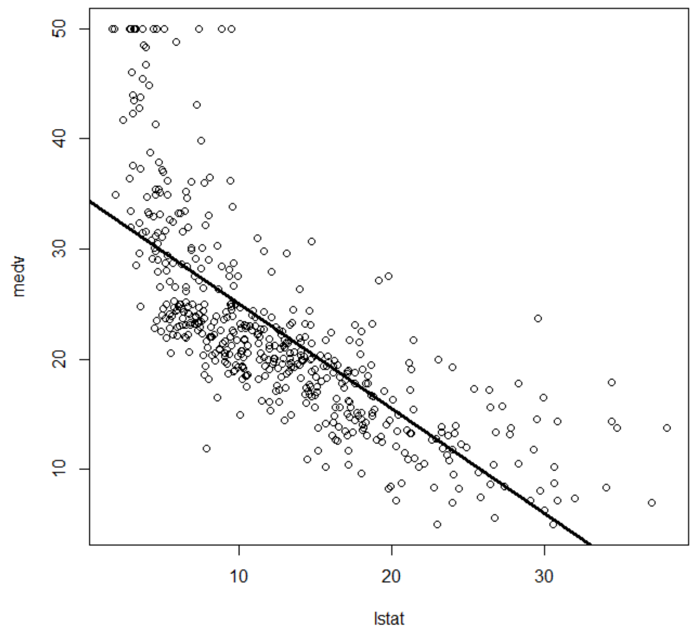
confint()函数 可以得到系数估计值的置信区间。在根据给定lstat的值预测medv时，predict ()函数 可以计算置信区间和预测区间。例如当lstat等于10时，相应的95%置信区间为(24.47, 25. 63)，相应的95%预测区间为(12. 828, 37. 28)。正如预期的那样，置信区间和预测区间有相同的中心点(当lstat等于10时，medv的预测值是25. 05)，但后者要宽得多。

代码实现——简单线性回归

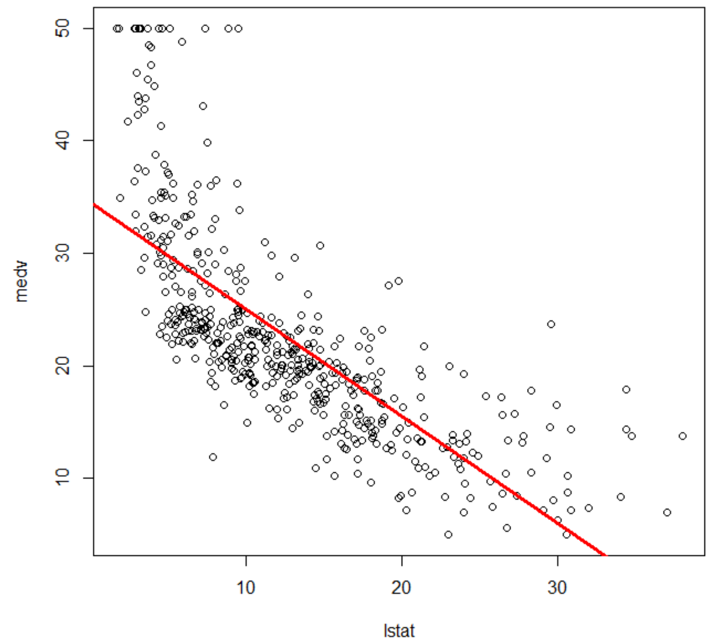
```
> plot(lstat,medv)
> abline(lm.fit)
```




```
> abline(lm.fit,lwd=3)
```




```
> abline(lm.fit,lwd=3,col="red")
```

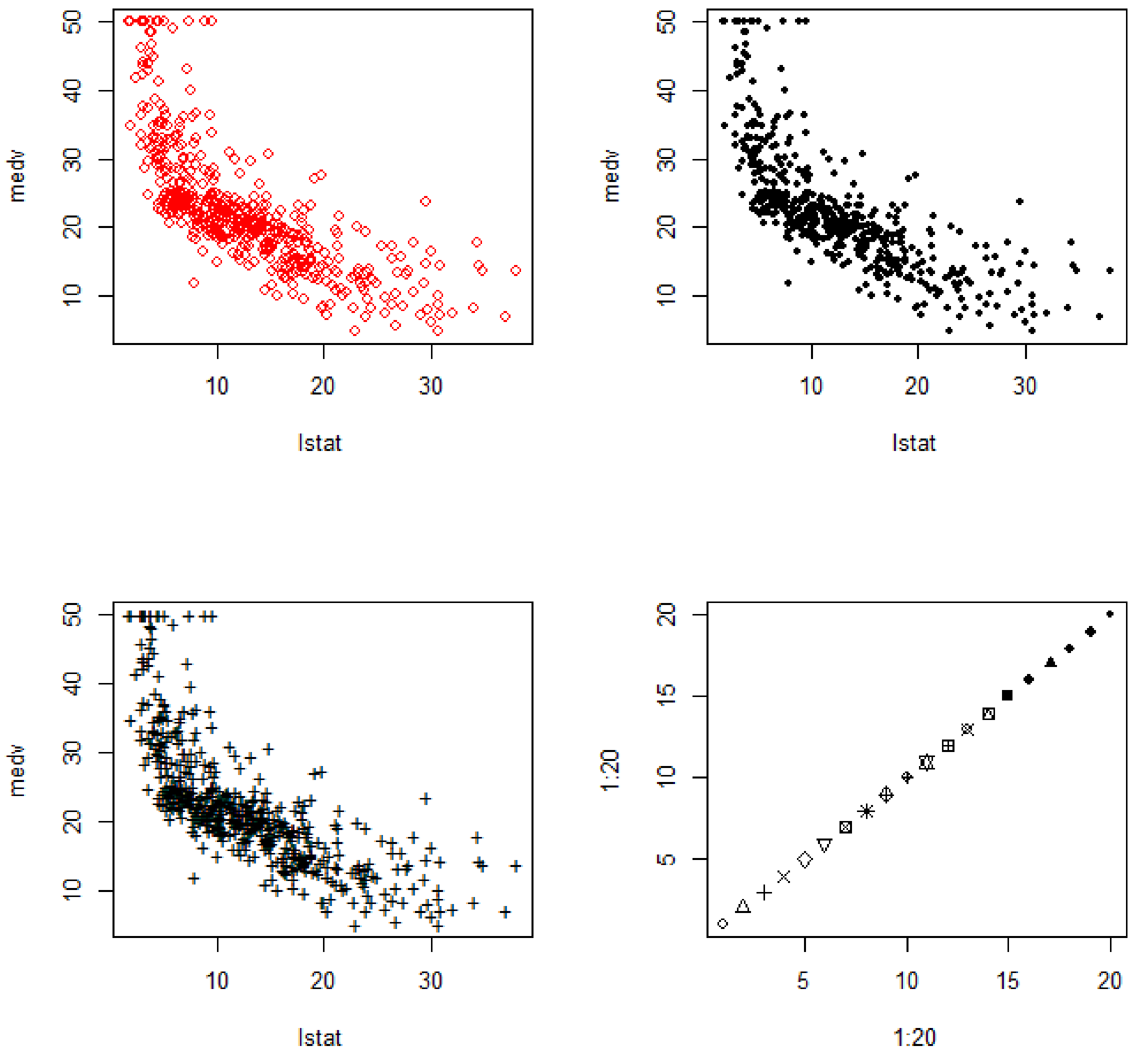


使用 函数plot () 和 abline () 绘制medv和lstat的散点图以及最小二乘回归直线。abline ()函数 可以用来绘制任意直线，而不只是最小二乘回归直线。输入abline(a, b) 可以画一条截距为a，斜率为b的直线。下面尝试一些用于绘制线和点的附加设置。lwd=3命令将使回归直线的宽度增加3倍，这一设置在 plot() 和 lines()函数中也可使用。我们还可以用 pch选项创建不同的图形符号。

01 代码实现——简单线性回归

```
> par(mfrow=c(2,2))
> plot(lstat,medv,col="red")
> plot(lstat,medv,pch=20)
> plot(lstat,medv,pch="+")
> plot(1:20,1:20,pch=1:20)
>
```

可以用 par()函数同时显示多张图表，它指示R将显示屏分割成独立的面板，所以可以同时查看多个图。例如，par(mfrow=c(2,2)) 把绘图区域划分成2x2的网格面板。



本周作业
(9月20日第三周)

教材3.7习题1、3、4、8；2.4习题9、10

上述内容下周二之前交（9月27日第四周）
本周三（9月21日）上机做/检查2.4习题8、10，3.7习题8

2. 多元线性回归

- 多元线性回归模型的形式为：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- β_j 可解释为在**所有其他预测变量保持不变的情况**下，X_j 增加一个单位对Y产生的**平均**效果。以广告数据集为例，即为：

$$sales = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \varepsilon$$

- 这里做的假设是，预测变量X是不相关的
  - 每个系数都可以单独的估计和检验
  - 也就是说，满足“**X_j变化一个单位，相应的Y会发生β_j个单位的变化，而所有其他变量保持不变**”
- 预测变量之间的相关性会导致问题：
  - 会对模型产生错误的解读，因为当 X_j 改变时，其它预测变量也改变了
- 预测变量和响应变量存在线性关系，**要避免解读为两者之间存在因果关系(claims of causality)**

- 来自“Data Analysis and Regression” Mosteller和Tukey 1977
  - 对于回归系数 β_j 的估计是说，**在所有其它预测变量保持不变的情况下**， X_j 一个单位的变化，所导致 Y 变化的程度。但预测变量往往是共同变化的！
  - 例如： Y 表示一个赛季中一个球员铲球的次数； W 和 H 是球员的身高和体重。通过数据分析的回归模型，可以表示为 $\hat{Y}=b_0+0.50W-0.10H$。

- "Essentially, all models are wrong, but some are useful"（本质上，所有模型都是错误的，但有些是有用的）

George Box (1919-2013)

- "The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively"（找出复杂系统受到干扰时会发生什么的唯一方法是干扰系统，而不仅仅是被动地观察它）

Fred Mosteller (1916-2006)和John Tukey (1915-2000)

- 对于给定的$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$，可以用如下公式进行预测：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

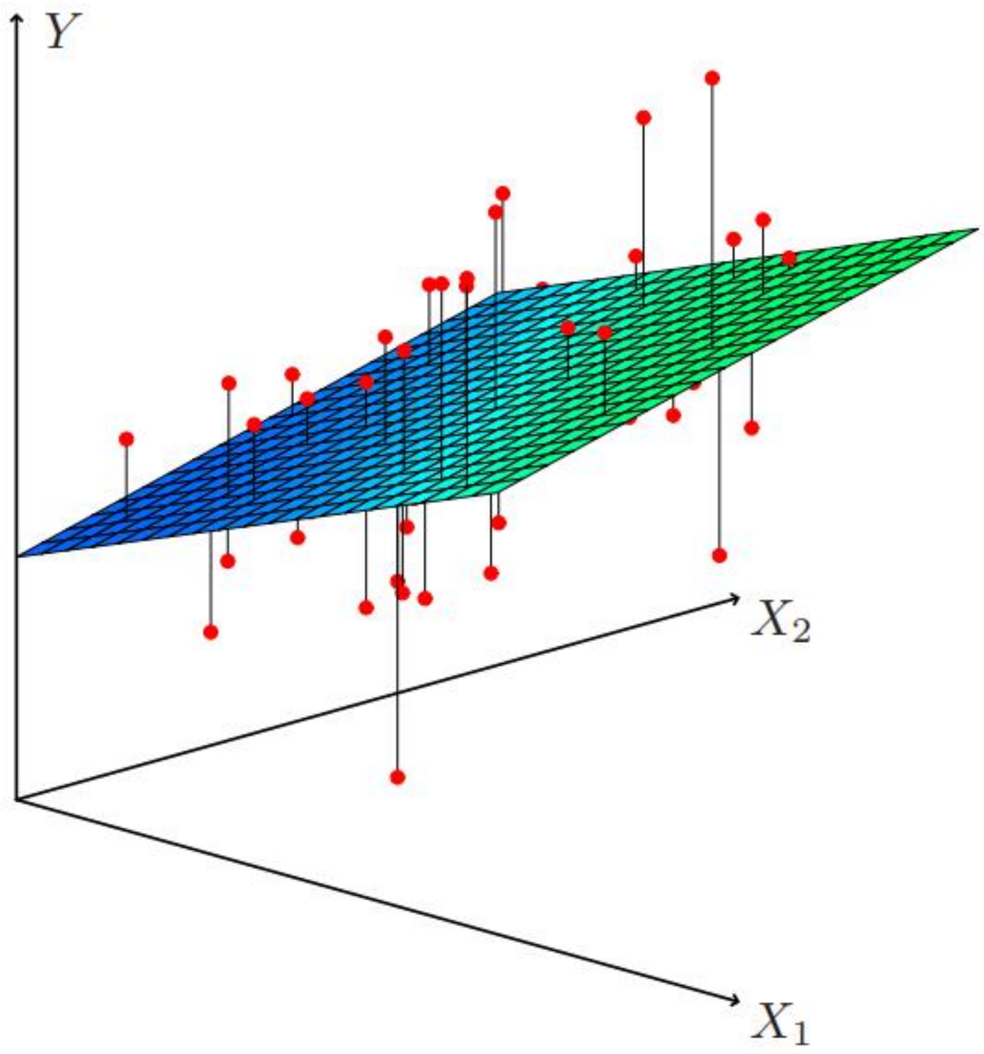
- 与在简单线性回归中相同，这里也是用最小二乘法进行估计。选择$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ 使残差平方和最小：

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

$$= \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}\right)^2$$

- 能最大限度地减小 RSS 的 $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ 取值，即为多元回归系数的最小二乘估计（可通过统计软件计算）

- 右图是用 $p=2$ 个预测变量对某数据集进行最小二乘拟合的一个例子。
- 这个三维图中有两个预测变量和一个响应变量，最小二乘回归直线变成了一个平面。这个平面使得每个观测值（以红色显示）与平面之间的垂直距离的平方和尽量减小。



Advertising数据集上计算结果

- sales关于radio、TV、newspaper的多元线性回归的最小二乘估计系数

	Coefficient	Std. Error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

- Radio、TV、newspaper的相关矩阵

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

- Q1：预测变量 $X_1, X_2, \cdots, X_p$ 中是否至少有一个可以用来预测响应变量?

- Q1：预测变量 $X_1, X_2, \cdots, X_p$ 中是否至少有一个可以用来预测响应变量?
- Q2：所有预测变量都有助于解释Y吗?或仅仅是预测变量的一个子集对预测有用?

- Q1：预测变量 $X_1, X_2, \cdots, X_p$ 中是否至少有一个可以用来预测响应变量?
- Q2：所有预测变量都有助于解释Y吗?或仅仅是预测变量的一个子集对预测有用?
- Q3：模型对数据的拟合程度如何?

- Q1：预测变量 $X_1, X_2, \cdots, X_p$ 中是否至少有一个可以用来预测响应变量?
- Q2：所有预测变量都有助于解释Y吗?或仅仅是预测变量的一个子集对预测有用?
- Q3：模型对数据的拟合程度如何?
- Q4：给定一组预测变量的值，响应值应预测为多少?所作预测的准确程度如何?

- 对于Q1，我们用F统计量回答

$$F=\frac{(\mathrm{TSS}-\mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}\sim F_{p,n-p-1}$$

- 在Advertising数据集中，有关销量sales对电视TV、报纸newspaper、广播radio广告预算的最小二乘回归模型的统计信息

	Quantity	Value
残差标准误	Residual Standard Error	1.69
R^2统计量	R^2	0.897
F统计量	F-statistic	570

Q2：判断重要的变量

- 一个最直接的方法叫做**全量子集(all subsets)**或**最优子集(best subsets)**回归：我们计算所有预测变量子集的最小二乘估计，并根据一些准则，如训练集误差或模型大小，确定最重要的预测变量

- 一个最直接的方法叫做**全量子集(all subsets)**或**最优子集(best subsets)**回归：我们计算所有预测变量子集的最小二乘估计，并根据一些准则，如训练集误差或模型大小，确定最重要的预测变量
- 但这种方法是不可行的！因为包含 p 个变量的模型共有 2^p 个变量子集。例如，若 $p=30$，则共有1 073 741 824（>10亿）种预测变量组合需要分析
- 因此，我们需要一种自动、高效的方法来选出少量待考虑的模型。这里讨论两种方法。

前向选择(Forward selection)

- 从**零模型(null model)**开始——一个只含有截距但不含预测变量的模型;
- 建立简单线性回归模型，并把使RSS最小的变量添加到零模型中;
- 再加入一个新变量，得到新的双变量模型，加入的变量是使新模型的RSS最小的变量;
- 这一过程持续到满足某种停止规则为止。

后向选择(Backward selection)

- 先从包含所有变量的模型开始
- 删除p值最大的变量——统计学上最不显著的变量；
- 拟合完包含$(p-1)$个变量的新模型后，再删值p值最大的变量；
- 此过程持续到满足某种停止规则为止。例如，当所有剩余变量的p值均低于某个阈值时，我们会停止删除变量。

- 注：这里斜体p表示预测变量的个数，p值中的p表示假设检验的显著性

- 第六章将详细讨论，在进行前向选择或后向选择的过程中，确定“最优”模型（即最佳预测变量组合）的方法

- 这些方法包括，**Mallow's C_p 统计量、赤池信息量准则(Akaike information criterion, AIC)、贝叶斯信息准则(Bayesian information criterion, BIC)、调整R^2(adjusted R^2) 和交叉验证(Cross-validation, CV)**。

Q3：模型对数据的拟合程度

- 回顾前面讲的内容，判断线性回归的拟合质量通常使用两个相关的量：
- **残差标准误 (residual standard error , RSE)**

$$RSE = \sqrt{\frac{1}{n-p-1} RSS}$$

其中，**残差平方和**$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- ***R^2*统计量**用下列公式计算：

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

其中，$\mathrm{TSS} = \sum(y_i - \bar{y})^2$ 是**总平方和(total sum of squares)**

Q4：响应值应预测

- 利用预测变量 $X_1, X_2, \ldots, X_p$ 的取值和计算的系数 β 预测响应值：

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

- 它是对真实总体回归平面的一个估计。

- 可以计算置信区间以确定$\hat{y}$与真实值的接近程度

代码实现——多元线性回归

- MASS库中包含Boston　(波士顿房价)数据集，它记录了波士顿周围506个街区的medv(房价中位数)。我们将设法用13个预测变量如rm（每栋住宅的平均房间数)，age　(平均房龄)，lstat(社会经济地位低的家庭所占比例)等来预测medv(房价中位数)。

```
> library(MASS)
> library(ISLR)
> fix(Boston)
```

数据编辑器

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
6	0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
7	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
8	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
9	0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
10	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9
11	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15
12	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9

代码实现——多元线性回归

为了用最小二乘法拟合多元线性回归模型，再次调用 lm ()函数。语句 lm(y ~ x1 + x2 +x3)用于建立有三个预测变量x1，x2和x3的拟合模型。summary()函数输出所有预测变量的回归系数。

```
> lm.fit=lm(medv~lstat+age,data=Boston)
> summary(lm.fit)

Call:
lm(formula = medv ~ lstat + age, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.981  -3.978  -1.283   1.968  23.158

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.22276    0.73085  45.458  < 2e-16 ***
lstat       -1.03207    0.04819 -21.416  < 2e-16 ***
age          0.03454    0.01223   2.826  0.00491 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.173 on 503 degrees of freedom
Multiple R-squared:  0.5513,    Adjusted R-squared:  0.5495
F-statistic:   309 on 2 and 503 DF,  p-value: < 2.2e-16
```

代码实现——多元线性回归

Boston数据集包含13个变量，所以在用所有的预测变量进行回归时，一一输入会很麻烦。可以使用下面的快捷方法:

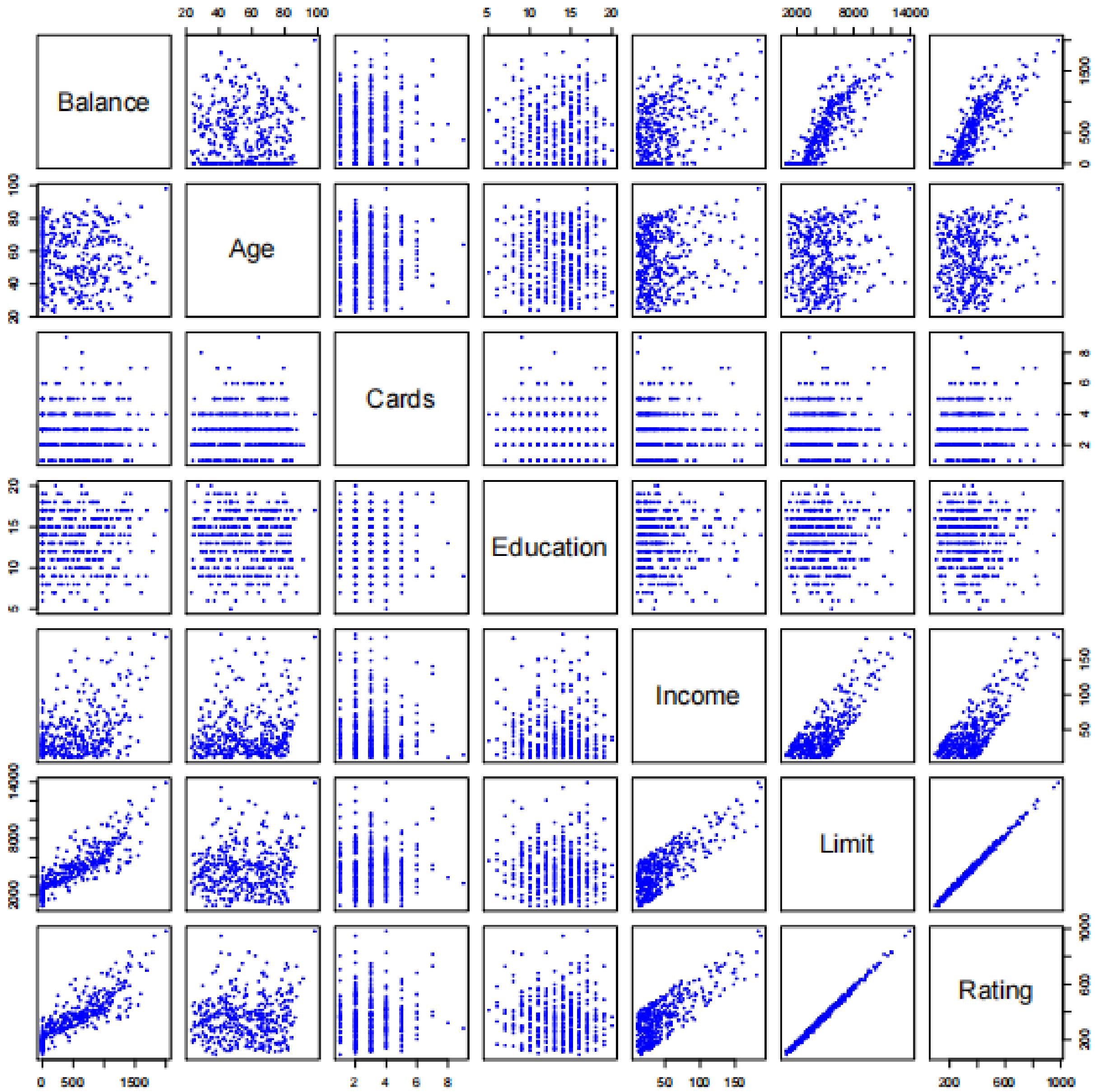
```
> lm.fit=lm(medv~.,data=Boston)
> summary(lm.fit)

Call:
lm(formula = medv ~ ., data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

3. 其他注意事项

- 我们一直假设线性回归模型中的所有变量都是**定量的(quantitative)**，但在实践中，这并不一定，往往有些预测变量是**定性的(qualitative)**，即变量的取值为一组离散数值的集合
- 它们也称作**分类变量(categorical predictors)**或**因子(factor)**
- 例如，性别、学生状态、婚姻状态、种族（亚洲人、白种人、非裔美国人）等

03 回归模型中的其他注意事项

- 右图Credit数据集包含**balance** (个人平均信用卡债务)、**age** (年龄)、**cards (信用卡数量)**、**education** (受教育年限)、**income** (收入，单位:千美元)、**limit** (信用额度)、**rating (信用评级)**等。另外包含**gender** (性别)、**student** (学生状态)、**status** (婚姻状态)、**ethnicity** (种族)四个变量



- 假设我们希望暂时忽略其他变量，调查男性和女性的信用卡债务差异。我们只需给二值变量创建一个指标，或称**哑变量(dummy variable)**。例如，

$$x_i=\begin{cases}1 & 女性\\ 0 & 男性\end{cases}$$

- 并在回归方程中使用这个变量。从而有以下模型：

$$y_i=\beta_0+\beta_1x_i+\varepsilon_i=\begin{cases}\beta_0+\beta_1+\varepsilon_i & 女性\\ \beta_0+\varepsilon_i & 男性\end{cases}$$

- β_0可以解释为男性的平均信用卡债务， $\beta_0+\beta_1$为女性的平均信用卡债务，所以 β_1是男性和女性之间信用卡债务的平均差异。

- Credit数据集中balance和gender回归的最小二乘估计系数。性别gender被编码为一个哑变量。

	Coefficient	Std. Error	t-statistic	p-value
Intercept	509.80	33.13	15.389	< 0.0001
gender[Female]	19.73	46.05	0.429	0.6690

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{女性} \\ \beta_0 + \varepsilon_i & \text{男性} \end{cases}$$

- 当一个定性预测变量有两个以上的水平，一个单一的虚拟变量不能代表所有可能的值。在这种情况下，我们可以创建更多的虚拟变量。例如，我们对于种族(ethnicity)变量（取值亚洲人、白种人、非裔美国人）创建两个哑变量。第一个哑变量是：

$$x_{i1}=\begin{cases}1 & \text{亚洲人}\\ 0 & \text{非亚洲人}\end{cases}$$

- 第二个哑变量是：

$$x_{i2}=\begin{cases}1 & \text{白种人}\\ 0 & \text{非白种人}\end{cases}$$

- 这两个变量都可以用于回归方程中，得如下模型：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{亚洲人} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{白种人} \\ \beta_0 + \varepsilon_i & \text{非裔美国人} \end{cases}$$

- 哑变量个数总是比水平数少1。没有相对应的哑变量的水平——本例中的非裔美国人——被称为**基准水平(baseline)**。

- Credit数据集中balance和ethnicity回归的最小二乘估计系数。种族变量通过前边介绍的虚拟x_{i1} x_{i2}变量来编码。

	Coefficient	Std. Error	t-statistic	p-value
Intercept	531.00	46.32	11.464	< 0.0001
ethnicity[Asian]	-18.69	65.02	-0.287	0.7740
ethnicity[Caucasian]	-12.50	56.68	-0.221	0.8260

- 标准线性回归模型提供了可解释的结果，但它作了一些高度限制性的假设。两个最重要的假设是预测变量和响应变量的关系是**可加(additive)**和**线性(linear)**的。
  - 可加性假设是指，预测变量 x_j 的变化对响应变量 Y 产生的影响与其他预测变量的取值无关；
  - 线性假设是指，无论 x_j 取何值，x_j 变化一个单位引起的响应变量 Y 的变化是恒定的。

- 放宽这两个假设——**交互作用(interaction)**和**非线性(nonlinearity)**
- **交互作用**
  - 在之前对Advertising数据集的分析中，线性模型假设，一种媒体的广告支出增加引起的sales变化与其他媒体的广告支出无关
  - 例如，线性模型

$$\widehat{\text{sales}} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper}$$

解释为， TV 增加一个单位对销售额 sales 的平均影响，始终是 β_1，无论在radio上支出多少

- 然而，这个简单的模型可能并不正确。假设对广播广告的投入事实上增强了电视广告的有效性，这时 TV 的斜率项应随着radio的增加而增加。
- 在这种情况下，给定10万美元的预算，在两种媒体上均分预算，可能比将资金全部投入其中一种媒体，更能增加销售。
- 这种现象在营销中被称为**协同(synergy)效应**，而在统计学中被称为**交互作用(interaction)**。

- 用包含radio和TV以及两者之间的交互项的线性模型来预测sales

$$\begin{aligned}\text{sales} &= \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times (\text{radio} \times \text{TV}) + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \varepsilon\end{aligned}$$

- 该回归模型的最小二乘系数估计如下：

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

- 上表中的结果说明交互作用的重要性
- 交互项 $\text{radio} \times \text{TV}$ 的p值是非常低的，这强有力地证明了 $H_A: \beta_3 \neq 0$。换言之，真正的关系明显是不可加的

- 上表中的结果说明交互作用的重要性
- 交互项 $\text{radio} \times \text{TV}$ 的p值是非常低的，这强有力地证明了 $H_A: \beta_3 \neq 0$。换言之，真正的关系明显是不可加的
- 模型的 R^2 为96.8%，而不含交互项的模型只有89.7%这意味着交互项解释了拟合可加性模型之后sales剩余变异的(96.8-89.7)/(100 -89.7)=69%。

- 上表中的结果说明交互作用的重要性
- 交互项 $\text{radio} \times \text{TV}$ 的p值是非常低的，这强有力地证明了 $H_A\colon \beta_3 \neq 0$。换言之，真正的关系明显是不可加的
- 模型的 R^2 为96.8%，而不含交互项的模型只有89.7%这意味着交互项解释了拟合可加性模型之后sales剩余变异的(96.8-89.7)/(100 -89.7)=69%。
- 系数估计表明，电视TV广告费用每增加一千美元，销量将增加$(\hat{\beta}_1+\hat{\beta}_3 \times \text{radio}) \times 1000 = 19 + 1.1 \times \text{radio}$ 个单位。
- 广播radio广告费用每增加一千美元，销量将增加 $(\hat{\beta}_2+\hat{\beta}_3 \times \text{TV}) \times 1000 = 29 + 1.1 \times \text{TV}$个单位。

- 在少数情况下，交互项的p值很小，而相关的主效应（本例中的TV和radio) 值却不然。
- **实验分层原则(hierarchical principle)**：
  - 如果模型中含有交互项，那么即使主效应的系数的p值不显著，也应包含在模型中；
  - 换句话说，如果X_1和X_2之间的交互作用是重要的，那么即使X_1和X_2的系数估计的p值较大，这两个变量也应该被包含在模型中。

- 这一原则的合理性在于，如果没有主效应项（如TV和radio），那两者之间的交互作用将很难解释
- 具体而言，如果模型中不包含主效应项，但交互项（如$\text{radio} \times \text{TV}$）却包含了主效应项，这是说不通的

- 考虑Credit数据集，假设我们希望用income(定量)和 student(定性)预测balance。
- 在没有交互项的情况下，模型是：

$$\text{balance}_i \approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 & \text{学生} \\ 0 & \text{非学生} \end{cases}$$

$$= \beta_1 \times income_i + \begin{cases} \beta_0 + \beta_2 & \text{学生} \\ \beta_0 & \text{非学生} \end{cases}$$

- 考虑Credit数据集，假设我们希望用income(定量)和 student(定性)预测balance。
- 在没有交互项的情况下，模型是：

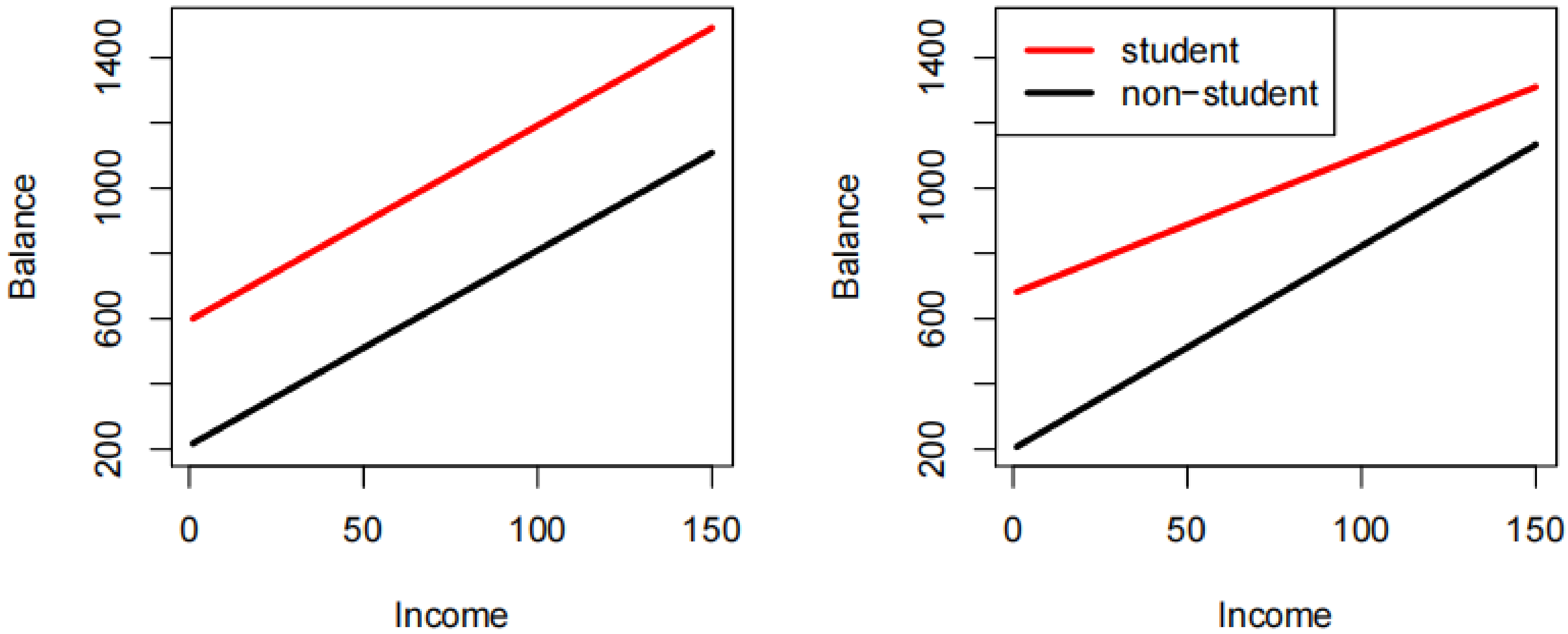
$$\text{balance}_i \approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 & \text{学生} \\ 0 & \text{非学生} \end{cases}$$

$$= \beta_1 \times income_i + \begin{cases} \beta_0 + \beta_2 & \text{学生} \\ \beta_0 & \text{非学生} \end{cases}$$

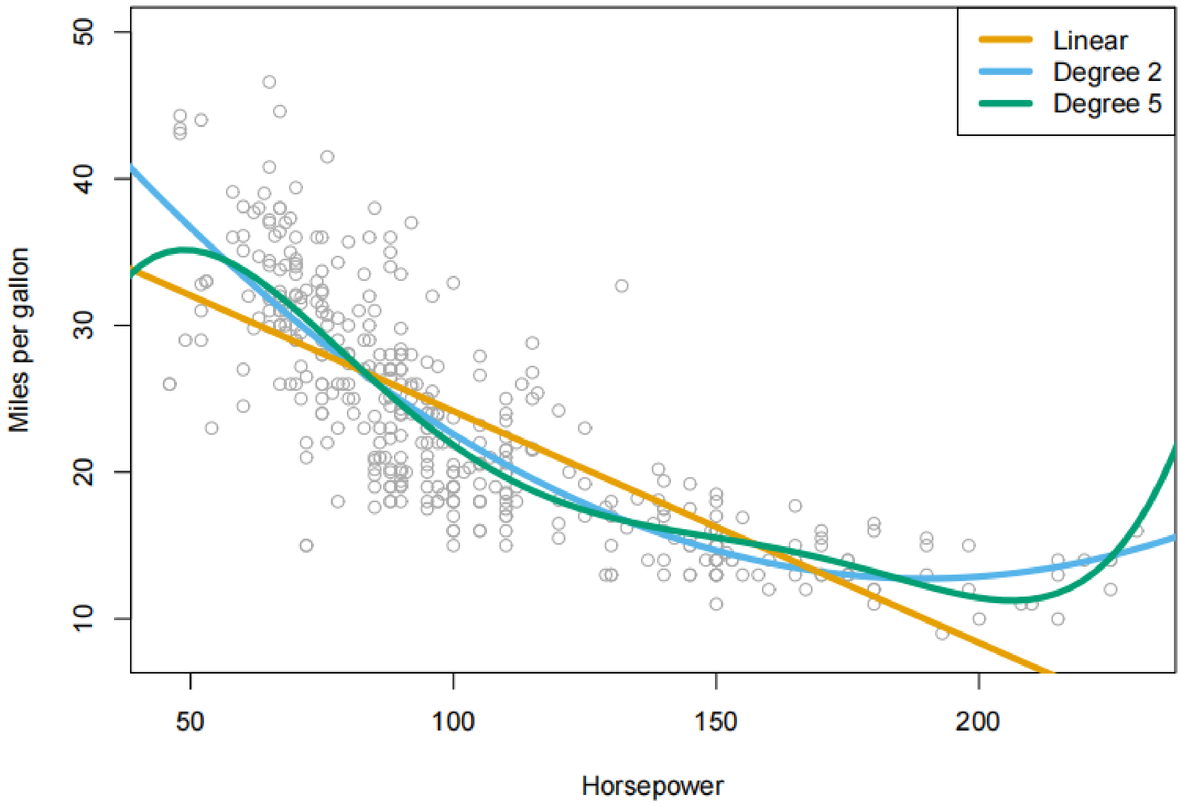
- 如果考虑交互项，模型变为：

$$\text{balance}_i \approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 + \beta_3 \times \text{income}_i & \text{学生} \\ 0 & \text{非学生} \end{cases}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{income}_i & \text{学生} \\ \beta_0 + \beta_1 \times \text{income}_i & \text{非学生} \end{cases}$$



- Credit数据集上，用income预测学生和非学生的balance的最小二乘线。
  - 左：模型不含income和student之间的交互项；
  - 右：模型含有income和student之间的交互项。



- Auto 数据集。汽车的油耗 (mpg) 和马力 (horsepower)。线性回归拟合线是橙色线。包含马力(horsepower)^{2}变量的线性回归拟合线是蓝色线。包含马力(horsepower)的所有五次以内项的线性回归拟合线是绿色线。

- 散点图中显示，油耗（mpg）和马力（horsepower)似乎有二次方的形状特征：

$$\text{mpg} = \beta_0 + \beta_1 \times \textit{horsepower} + \beta_2 \times \text{horsepower}^2 + \varepsilon$$

Auto数据集中，mpg对horsepower和horsepower^2
的回归的最小二乘估计

	Coefficient	Std. Error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

代码实现——交互项

用 lm()函数使线性模型包括交互项是很容易的。语句lstat: black命令R将lstat和black的交互项纳入模型。语句lstat* age将lstat，age和交互项lstat * age同时作为预测变量，它是lstat + age + lstat : age 的简写形式。

```
> summary(lm(medv~lstat*age,data=Boston))

Call:
lm(formula = medv ~ lstat * age, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.806  -4.045  -1.333   2.085  27.552 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 36.0885359  1.4698355  24.553  < 2e-16 ***
lstat       -1.3921168  0.1674555  -8.313 8.78e-16 ***
age         -0.0007209  0.0198792  -0.036   0.9711    
lstat:age    0.0041560  0.0018518   2.244   0.0252 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.149 on 502 degrees of freedom
Multiple R-squared:  0.5557,    Adjusted R-squared:  0.5531 
F-statistic: 209.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

大连理工大学 Dalian University of Technology

lm()函数也可以容纳预测变量的非线性变换。例如，给定预测变量X，可以用 I (X^2) 创建预测变量X^2。函数 I ()是必要的，因为^在公式中有特殊的含义，这是R软件里把X转换成其二次方的标准方法。建立medv对 lstat和 lstat^{2}的回归。

```
> lm.fit2=lm(medv~lstat+I(lstat^2),data=Boston)
> summary(lm.fit2)

Call:
lm(formula = medv ~ lstat + I(lstat^2), data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.2834  -3.8313  -0.5295   2.3095  25.4148 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 42.862007   0.872084   49.15   <2e-16 ***
lstat       -2.332821   0.123803  -18.84   <2e-16 ***
I(lstat^2)   0.043547   0.003745   11.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.524 on 503 degrees of freedom
Multiple R-squared:  0.6407,    Adjusted R-squared:  0.6393 
F-statistic: 448.5 on 2 and 503 DF,  p-value: < 2.2e-16
```

代码实现——预测变量的非线性变换

二次项的p值接近零表明它使模型得到了改进。使用 anova() 函数进一步量化二次拟合在何种程度上优于线性拟合。

```
> lm.fit=lm(medv~lstat,data=Boston)
> anova(lm.fit,lm.fit2)
Analysis of Variance Table

Model 1: medv ~ lstat
Model 2: medv ~ lstat + I(lstat^2)
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)
1    504 19472
2    503 15347  1    4125.1 135.2 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

这里的模型1代表只包含一个预测变量lstat的线性子模型，模型2则对应具有两个预测变量lstat和lstat的二次模型。anova() 函数通过假设检验比较两个模型。零假设是这两个模型对数据的拟合同样出色，备择假设是全模型更优。这里的F统计量是135，相应的p值几乎为零。这提供了非常明确的证据表明包含预测变量lstat和 $lstat^2$的模型远远优于只包含预测变量lstat的模型。

代码实现——预测变量的非线性变换

要创建一个三次拟合，可以向模型中加入一个形如I(x^3)的预测变量。然而，这种方法对于高阶多项式就会变得繁琐。更好的方法是用 poly() 和 lm()函数创建多项式。例如,下面的命令会产生一个5阶多项式拟合，可以看到，模型拟合得到了改善。

```
> lm.fit5=lm(medv~poly(lstat,5),data=Boston)
> summary(lm.fit5)

Call:
lm(formula = medv ~ poly(lstat, 5), data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5433  -3.1039  -0.7052   2.0844  27.1153

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       22.5328     0.2318  97.197  < 2e-16 ***
poly(lstat, 5)1 -152.4595     5.2148 -29.236  < 2e-16 ***
poly(lstat, 5)2   64.2272     5.2148  12.316  < 2e-16 ***
poly(lstat, 5)3  -27.0511     5.2148  -5.187 3.10e-07 ***
poly(lstat, 5)4   25.4517     5.2148   4.881 1.42e-06 ***
poly(lstat, 5)5  -19.2524     5.2148  -3.692 0.000247 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.215 on 500 degrees of freedom
Multiple R-squared:  0.6817,    Adjusted R-squared:  0.6785
F-statistic: 214.2 on 5 and 500 DF,  p-value: < 2.2e-16
```

也可以对预测变量进行对数变换：

```
> summary(lm(medv~log(rm),data=Boston))#对数转换

Call:
lm(formula = medv ~ log(rm), data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.487  -2.875  -0.104   2.837  39.816 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -76.488      5.028  -15.21   <2e-16 ***
log(rm)       54.055      2.739   19.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.915 on 504 degrees of freedom
Multiple R-squared:  0.4358,	Adjusted R-squared:  0.4347 
F-statistic: 389.3 on 1 and 504 DF,  p-value: < 2.2e-16
```

代码实现——定性预测变量

- 现在，我们将研究carseats （汽车座椅）数据，它是ISLR库的一部分。我们试图根据一些预测变量预测400个地区的sales (儿童座椅销量)。

```
> fix(Carseats)
> names(Carseats)
 [1] "Sales"       "CompPrice"   "Income"      "Advertising" "Population"  "Price"
 [7] "ShelveLoc"   "Age"         "Education"   "Urban"       "US"
```

数据编辑器

	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education	Urban	US
1	9.5	138	73	11	276	120	Bad	42	17	Yes	Yes
2	11.22	111	48	16	260	83	Good	65	10	Yes	Yes
3	10.06	113	35	10	269	80	Medium	59	12	Yes	Yes
4	7.4	117	100	4	466	97	Medium	55	14	Yes	Yes
5	4.15	141	64	3	340	128	Bad	38	13	Yes	No
6	10.81	124	113	13	501	72	Bad	78	16	No	Yes
7	6.63	115	105	0	45	108	Medium	71	15	Yes	No
8	11.85	136	81	15	425	120	Good	67	10	Yes	Yes
9	6.54	132	110	0	108	124	Medium	76	10	No	No
10	4.69	132	113	0	131	124	Medium	76	17	No	Yes

07

代码实现——定性预测变量

- Carseats数据含有定性预测变量，如 shelveloc(每个地区搁架位置的质量指标)，即在每个地区汽车座椅在商店内的展示空间。预测变量shelveloc有三个可能的取值:坏(bad)，中等(medium)，好(good)。给出定性变量如shelveloc，R将自动生成虚拟变量。下面构建一个含有交互项的多元回归模型。

```
> lm.fit=lm(Sales~.+Income:Advertising+Price:Age,data=Carseats)#含有交互项
> summary(lm.fit)

Call:
lm(formula = Sales ~ . + Income:Advertising + Price:Age, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9208 -0.7503  0.0177  0.6754  3.3413 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         6.5755654  1.0087470   6.519 2.22e-10 ***
CompPrice           0.0929371  0.0041183  22.567  < 2e-16 ***
Income              0.0108940  0.0026044   4.183 3.57e-05 ***
Advertising         0.0702462  0.0226091   3.107 0.002030 ** 
Population          0.0001592  0.0003679   0.433 0.665330    
Price              -0.1008064  0.0074399 -13.549  < 2e-16 ***
ShelveLocGood       4.8486762  0.1528378  31.724  < 2e-16 ***
ShelveLocMedium     1.9532620  0.1257682  15.531  < 2e-16 ***
Age                -0.0579466  0.0159506  -3.633 0.000318 ***
Education          -0.0208525  0.0196131  -1.063 0.288361    
UrbanYes            0.1401597  0.1124019   1.247 0.213171    
USYes              -0.1575571  0.1489234  -1.058 0.290729    
Income:Advertising  0.0007510  0.0002784   2.698 0.007290 ** 
Price:Age           0.0001068  0.0001333   0.801 0.423812    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.011 on 386 degrees of freedom
Multiple R-squared:  0.8761,    Adjusted R-squared:  0.8719 
F-statistic:   210 on 13 and 386 DF,  p-value: < 2.2e-16
```

contrasts ()函数返回R虚拟变量的编码:

```
> attach(Carseats)
> contrasts(ShelveLoc)#返回虚拟变量的编码，此函数还有其他编码方式，可自行设置
       Good Medium
Bad       0      0
Good      1      0
Medium    0      1
```

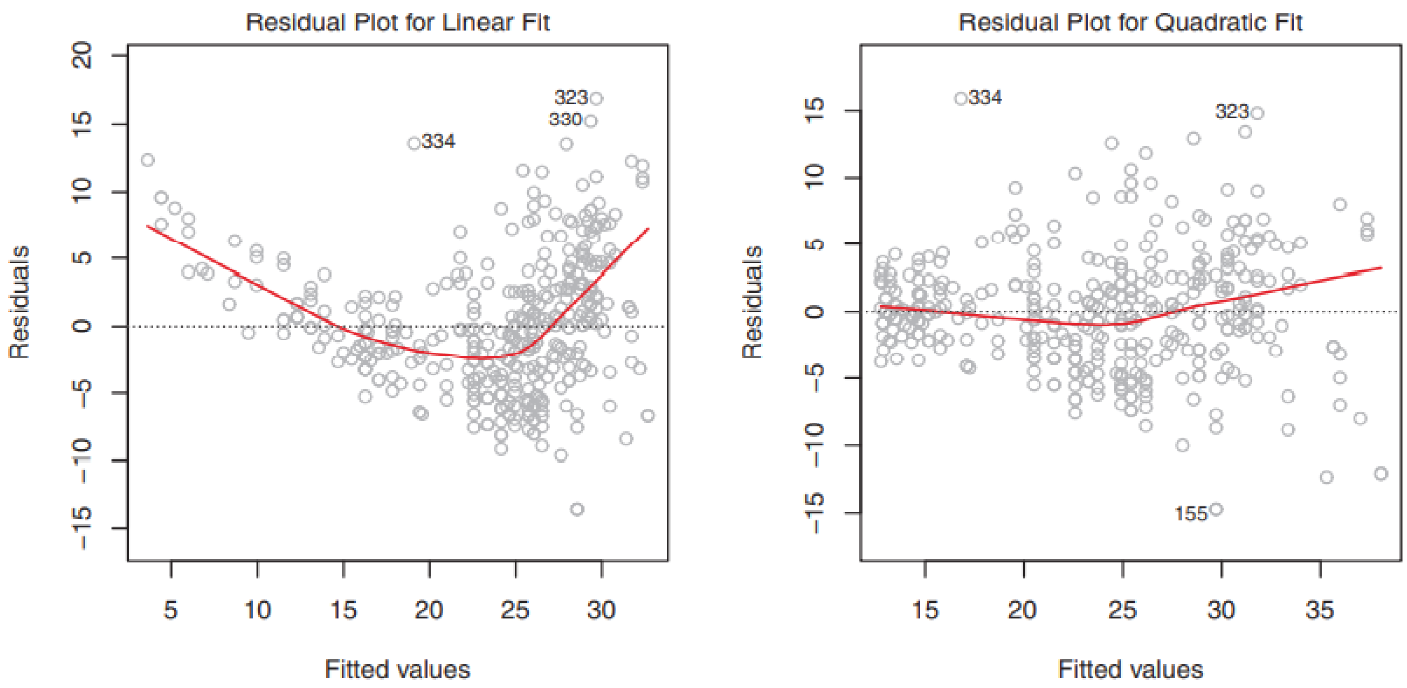
R创建了一个虚拟变量shelveLocGood，如果货架位置好，它的值为1，否则为0。R还创造了一个虚拟变量shelveLocMedium，如果货架位置属于中等水平，它的值为1，否则为0。坏的搁置位置则对应两个虚拟变量均为0。在回归输出中，若变量shelveLocGood的系数为正，则表明好的货架位置与高销售额相关(与坏位置相比)。若变量 shelveLoc-Medium的系数为较小的正值，则表明中等水平的货架位置的销量比坏位置高，但比一个好位置差。

4. 潜在的问题

- 使用线性回归模型进行数据集拟合可能遇到的潜在问题:
  - 数据的非线性（nonlinearity of response-predictor relationship）
  - 误差项自相关(correlation of error term)
  - 误差项方差非恒定(non-constant variance of error term)
  - 离群点(outlier)
  - 高杠杆点(high-leverage point)
  - 共线性(collinearity)

- 线性回归模型假定预测变量和响应变量之间有直线关系。如果真实关系是非线性的，那么得出的几乎所有结论都是不可信的。
- 残差图（residual plot）可用于识别非线性。
- 给定一个简单线性回归模型，我们可以绘制残差 $e_i = y_i - \hat{y}_i$ 和预测变量x_i的散点图。在多元回归中，因为有多个预测变量，我们转而绘制残差与预测值（或拟合值） $\hat{y}_i$的散点图。
- 理想情况，残差图显示不出明显的规律。若存在明显规律，则表示线性模型的某些方面可能有问题。
- 如果残差图表明数据中存在非线性关系，那么一种简单的方法是在模型中使用预测变量的非线性变换，例如$\log X$, $\sqrt{X}$, X^2等

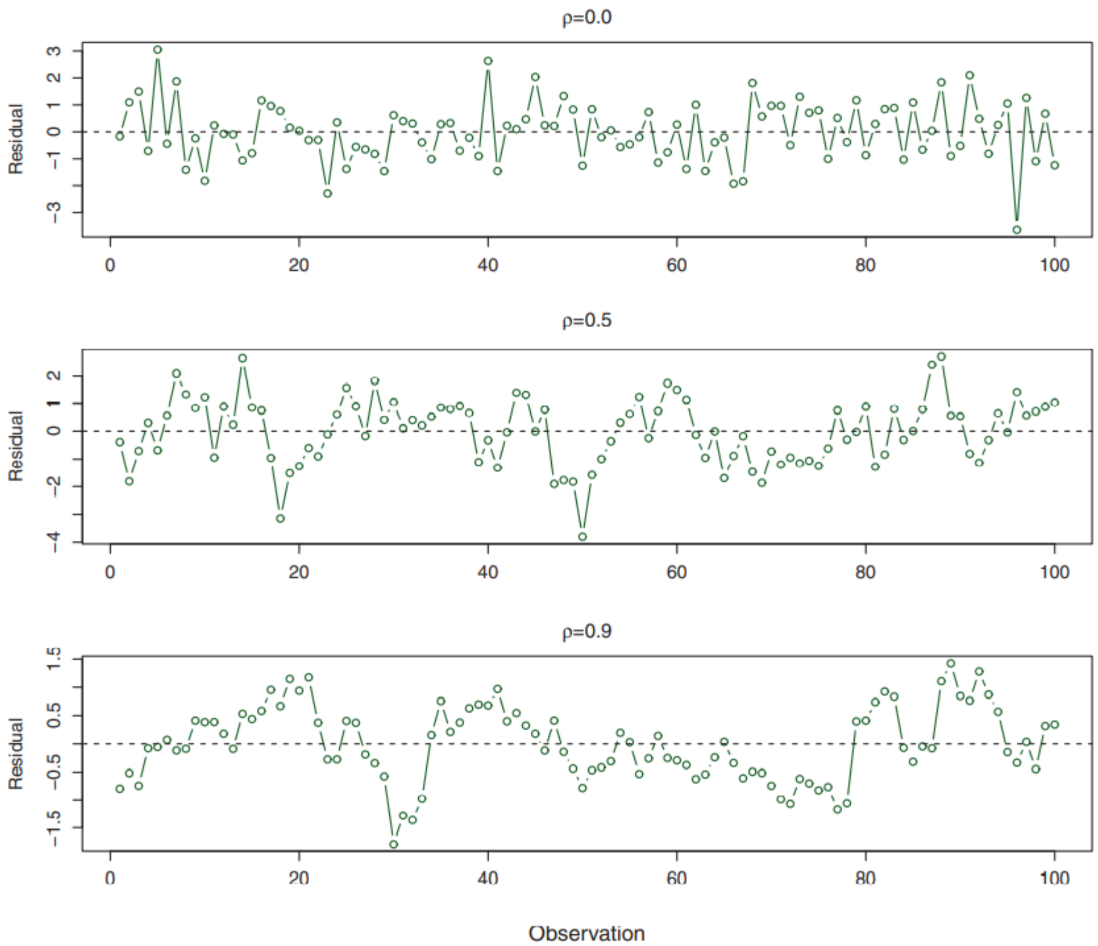
- Auto数据集mpg对horsepower的线性回归的残差图。红线是对残差的一个光滑拟合。左图残差呈现明显的U型，这为数据的非线性提供了强有力的证据。相比之下，右图展示了含有一个二次项的模型的结果，残差似乎没有规律，表明二次项加入提升了模型对数据的拟合度。



- 线性回归模型的一个重要假设是误差项$\varepsilon_1, \varepsilon_2, \ldots\ \varepsilon_n$不相关。如果误差项相关，那么估计标准误往往低估了真实标准误，因此，置信区间和预测区间比真实区间窄。例如95%置信区间包含真实参数的实际概率将远低于0.95。这可能导致得出错误的结论。
- 举个例子，假设我们不小心把数据重复了一遍，导致相同的观测和误差项成对出现。那么我们似乎是在计算一个规模为2n的样本的标准误，但事实上，样本仅为n。我们对2n个样本的参数估计和对n个样本的估计是相同的，但后者置信区间的宽度是前者的$\sqrt{2}$倍

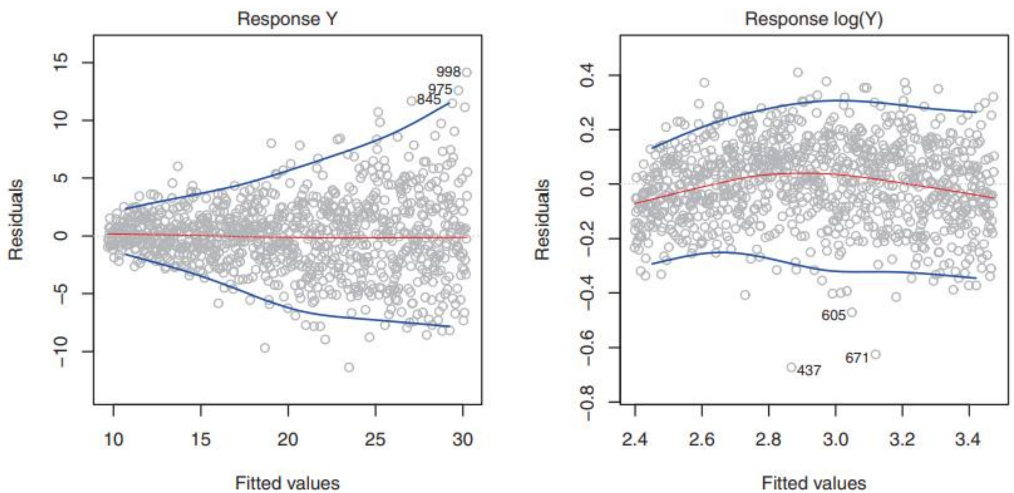
$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$，β_1的95%置信区间约为：$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$

- 误差项自相关经常出现在时序序列数据，很多时候相邻时间点获得的观测的误差有正相关关系。
- 我们可以根据模型绘制作为时间函数的残差，如果误差项不相关，图中没有明显规律；否则，可能在残差中看到跟踪现象。



大连理工大学 Dalian University of Technology

- 线性回归模型的另一个重要假设是误差项的方差是恒定的 $\mathrm{VAR}(\varepsilon_i) = \sigma^2$。假设检验和标准误差、置信区间计算依赖这一假设。
- 但通常，误差项的方差不是恒定的。例如，误差项的方差可能会随响应值的增加而增加。如果残差图呈漏斗形，说明误差项方差非恒定或存在异方差性。
- 下左图，残差随拟合值增加而增加。可以用凹函数对响应值y做变换 $\log Y, \sqrt{Y}$。 进而使较大的响应值有更大的收缩。右图是对数变换后的残差。



- 有时我们可以估计每个响应值的方差。例如，第i个响应值可能是n_i个原始观测值的平均值。如果每个原始观测都与方差 σ^2 无关，那么他们均值的方差是$\sigma_i^2 = \sigma^2 / n_i$。
- 在这种情况下，一个简单的补救办法是用加权最小二乘法拟合模型，即权重与方差的倒数成比例。

离群点

- 离群点是指y_i远离模型预测值的点。
- 预测变量的离群点通常对最小二乘拟合几乎没有影响。但它仍能导致其他问题。例如，下图中含离群点的回归RSE是1.09，但去除离群点后，RSE仅为0.77。
- 单个数据点造成的急剧增加可能影响对拟合的解释。同样，加入离群点导致R^2从89.2%下降到80.5%

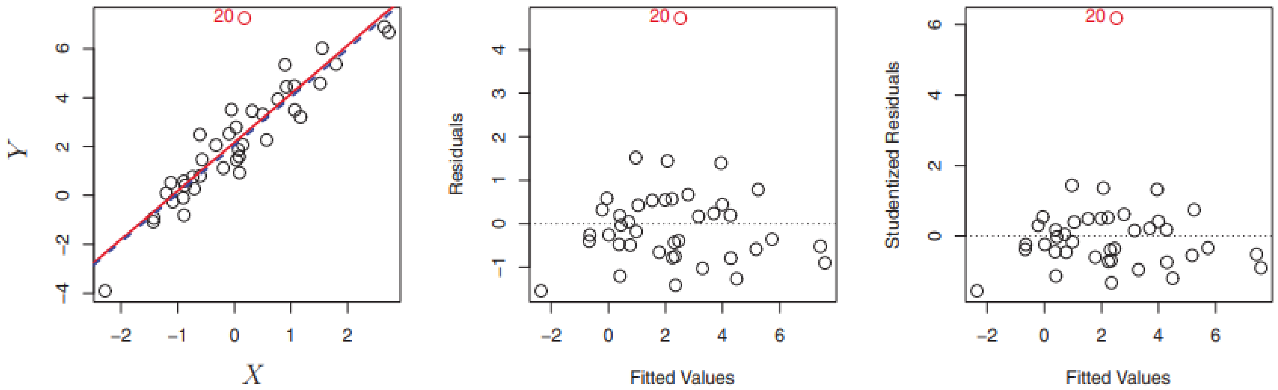
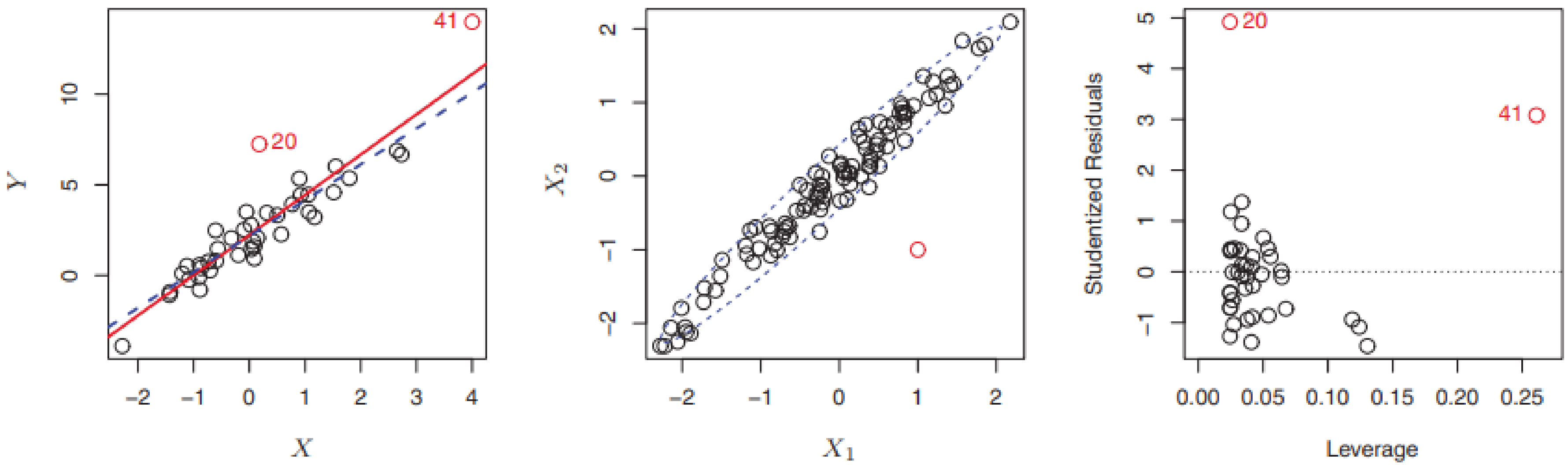


图 3-12　左：最小二乘回归线为红线，而删除离群点后的回归线用蓝色表示。中：残差图清楚地识别出了离群点。右：离群点的学生化残差为 6，通常的学生化残差在 −3 至 3 之间。

- 残差图可以用来识别离群点。但确定残差多大的点可以被认为是一个离群点会十分困难。我们可以绘制学生化残差，即由残差e_i除以它的估计标准误得到。学生化残差绝对值大于3的观测点可能是离群点。
- 如果确信是离群点是由于数据采集或记录中的错误导致的，可以直接删除。但应该小心，有时是因为暗箱模型存在缺陷，比如缺少预测变量。

08 高杠杆点

- 高杠杆表示观测点x_i是异常的（如左图观测点41），高杠杆的观测往往对回归直线的估计有很大的影响
- 在简单线性回归中，我们通过找预测变量的取值超出正常范围的观测点，辨认高杠杆点。多元线性回归中，可能存在观测点、它取值都在正常范围，但从整个预测变量集的角度看，它不寻常（中图）



- 为了量化观测的杠杆作用，可计算杠杆统计量。一个大的杠杆统计量对应一个高杠杆点。
- $$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$
- 方程中h_i随着$x_i - \bar{x}$的增加而增加。
- 杠杆统计量 h_i 的取值总是在1/n和1之间，且所有观测的平均杠杆值总是等于(p+1)/n。因此，如果给定观测的杠杆统计量大大超过(p+1)/n，那么我们可能会怀疑对应点有较高杠杆作用。

- 共线性是指两个或更多的预测变量高度相关。如下图Credit数据集（右图），它会导致难以分离单个变量对响应值的影响。

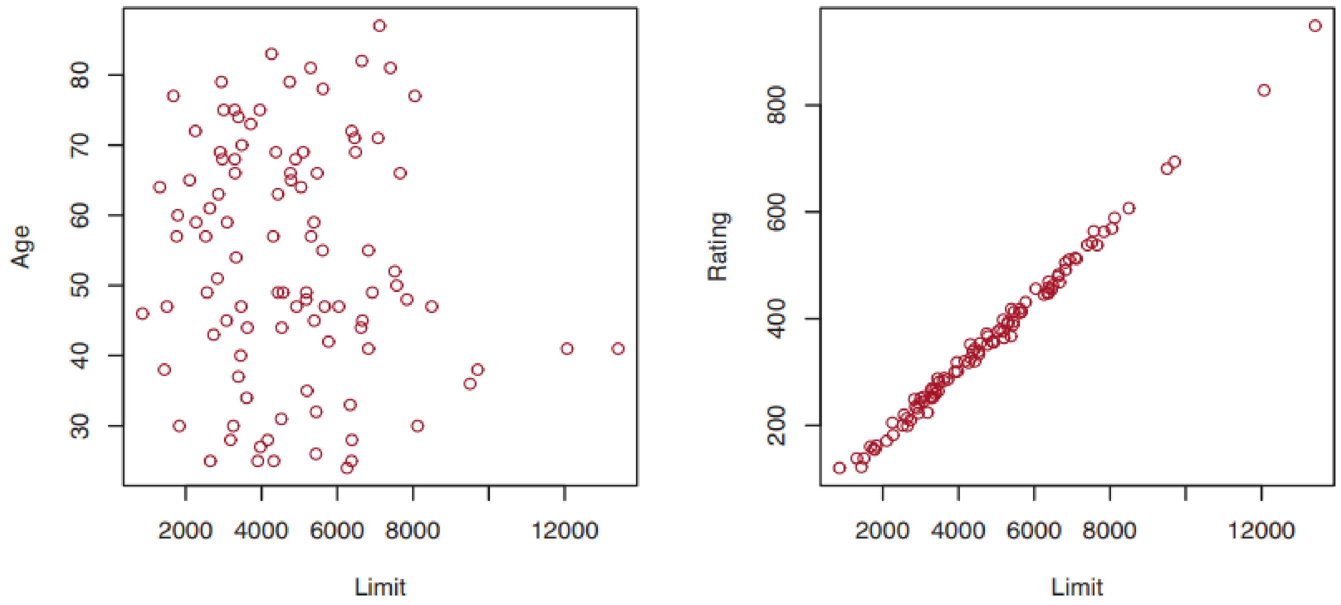


图 3-14　Credit 数据集观测值的散点图。左：age 与 limit 的图。这两个变量没有共线性。右：rating 与 limit 的图。这两个变量有很高的共线性。

共线性

- 下图是共线性可能导致的问题：数据的微小变化可能导致RSS最小的系数估计——即最小二乘估计——沿着这条山谷的任何地方移动。这导致系数估计有很大的不确定性。
- 如果存在共线性，我们可能无法拒绝 $H_0: \beta_j = 0$， 即假设检验的效力——正确地检测出非零系数的概率——被共线性减小了。

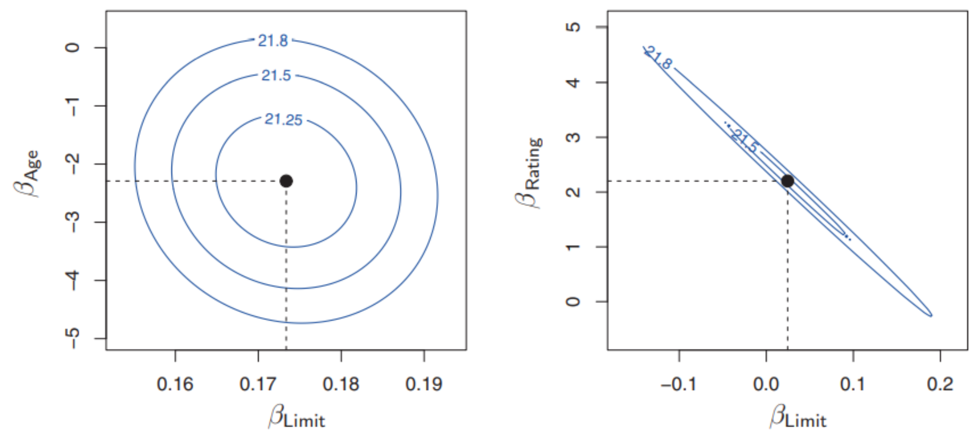


图 3-15　Credit 数据集上多种回归的 RSS 值的等高线图，RSS 是参数 β 的函数。每张图中的黑点代表最小 RSS 对应的系数。左：balance 对 age 和 limit 回归的 RSS 等高线图。最小值被很好地定义。右：balance 对 rating 和 limit 回归的 RSS 等高线图。由于共线性的存在，许多对系数（β_{Limit}，β_{Rating}）都有类似的 RSS。

- 下表比较了两个独立的多元回归模型的系数估计。在第二个回归中，因为共线性存在，系数估计标准误增加了12倍而且p值增加到了0.701。也就是说，limit变量的重要性被掩盖了。

表 3-11 Credit 数据集的两个多元回归模型。模型 1 是 balance 对 age 和 limit 的回归，模型 2 是 balance 对 rating 和 limit 的回归。由于共线性的存在，第二个回归中 $\hat{\beta}_{limit}$ 的标准误是第一个的 12 倍。

		系数	标准误	t 统计量	p 值
Model 1	Intercept	−173.411	43.828	−3.957	<0.000 1
	age	−2.292	0.672	−3.407	0.000 7
	limit	0.173	0.005	34.496	<0.000 1
Model 2	Intercept	−377.537	45.254	−8.343	<0.000 1
	rating	2.202	0.952	2.312	0.021 3
	limit	0.025	0.064	0.384	0.701 2

- 检测共线性的一个简单方法是看预测变量的相关系数矩阵。但即使没有某对变量具有特别高的相关性，有可能三个或更多变量之间存在共线性，成为多重共线性。
- 更好的方法是计算方差膨胀引子（VIF），VIF是拟合全模型时的系数$\hat{\beta}_j$的方差除以单变量回归中$\hat{\beta}_j$的方差所得的比例。VIF最小可能值是1，表示完全不存在共线性。通常情况，VIF值超过5或10就表示有共线性问题。
- $$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1-R^2_{X_j|X_{-j}}}$$
- 其中$R^2_{X_j|X_{-j}}$是X_j对所有预测变量回归的R^2
- 解决方法：（1）从回归中剔除一个问题变量；（2）把共线性变量组合成一个单一的预测变量

- 线性回归是参数方法的一个特例，因为它将f(x)假设为线性函数
  - 优点：需要估计的系数较少、容易拟合、系数有简单的解释、容易进行统计显著性检验
  - 缺点：假设过强、如果指定的函数形式与实际相差太远则表现不佳
- 非参数方法：不明确假设一个参数化的形式
- 最简单而知名的——K最近邻回归（KNN回归）：给定K值和预测点x_0，K最近邻回归首先确定K个最接近x_0的训练观测，记为 $\mathcal{N}_0$ 。然后用$\mathcal{N}_0$中所有训练数据的平均值来估计$f(x_0)$
- $$\hat{f}(x_0) = \frac{1}{K}\sum_{x_i \in \mathcal{N}_0} y_i$$

- 下图为预测变量数p=2的数据集的两个KNN拟合。
- 最优K值的选择依赖于第2章介绍的偏差-方差的权衡：小的K值提供了最灵活的拟合，导致低偏差和高方差；更大的K值提供的拟合更平滑、方差更小，但可能会隐藏f(X)的部分结构而导致偏差

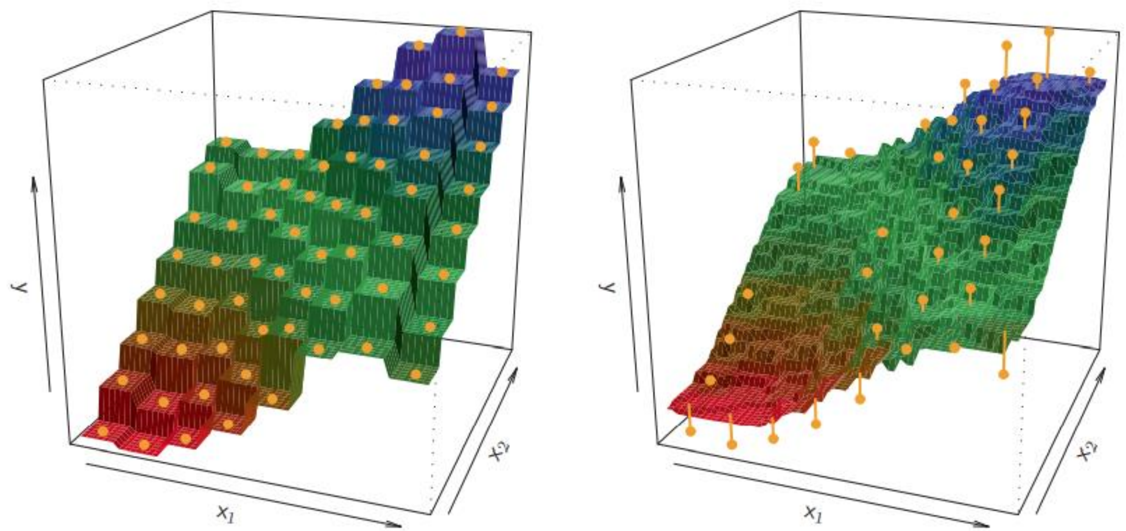


图 3-16　对一个含 64 个观察值（橙色点）的二维数据集进行 KNN 回归得到的 $\hat{f}(X)$ 拟合图。左：取 $K=1$ 可得到一个粗略的阶梯函数拟合。右：取 $K=9$ 产生更平滑的拟合。

线性回归与K最近邻法的比较

- 如何选择：如果选定的参数形式接近f的真实形式，则参数方法更优。如下图，因为真正的关系是线性的，所以非参数方法（K=1或9）很难与线性回归竞争。
- 但实践中X和Y的真实关系很少是完全线性的，这种情况KNN可能会优于线性回归。但即使这种情况，特别是在高维时，KNN也可能比线性回归差。

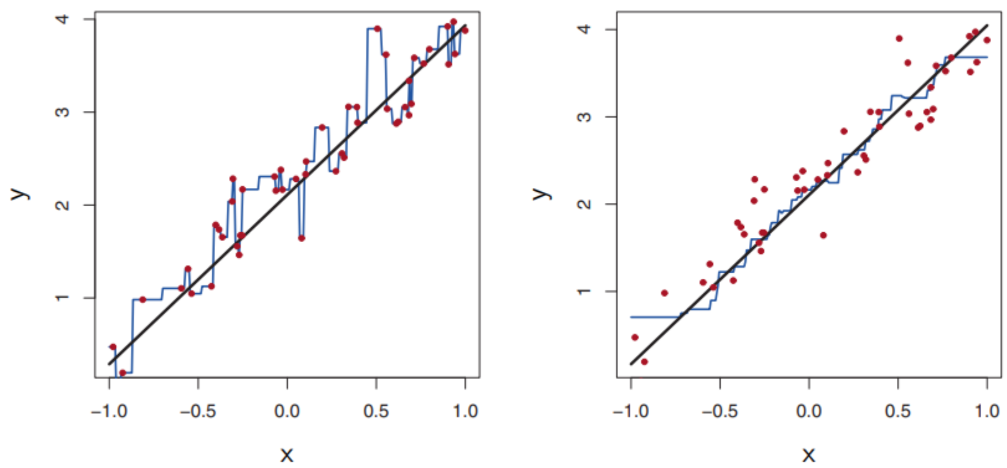


图 3-17　对一个含 100 个观测的一维数据集进行 KNN 回归的 $\hat{f}(X)$ 拟合图。真实关系由较粗直线表示。左：较细曲线对应 $K=1$，它插入（即直接穿过）了训练数据。右：较细曲线对应 $K=9$，代表了更光滑的拟合。

线性回归与K最近邻法的比较

- 预测效果随着维数的增加而恶化是KNN一个普遍问题，因为在高维中样本量大大减少。下图有100个训练观察，，当p=1时，这些点提供了足够的信息来准确估计f(X)。然而，当这100个观测值分布在p=20个维度上时，将使给定的观测附近没有邻点——即维度灾难。
- 若每个预测变量仅有少量观测，参数化方法往往优于非参数方法。

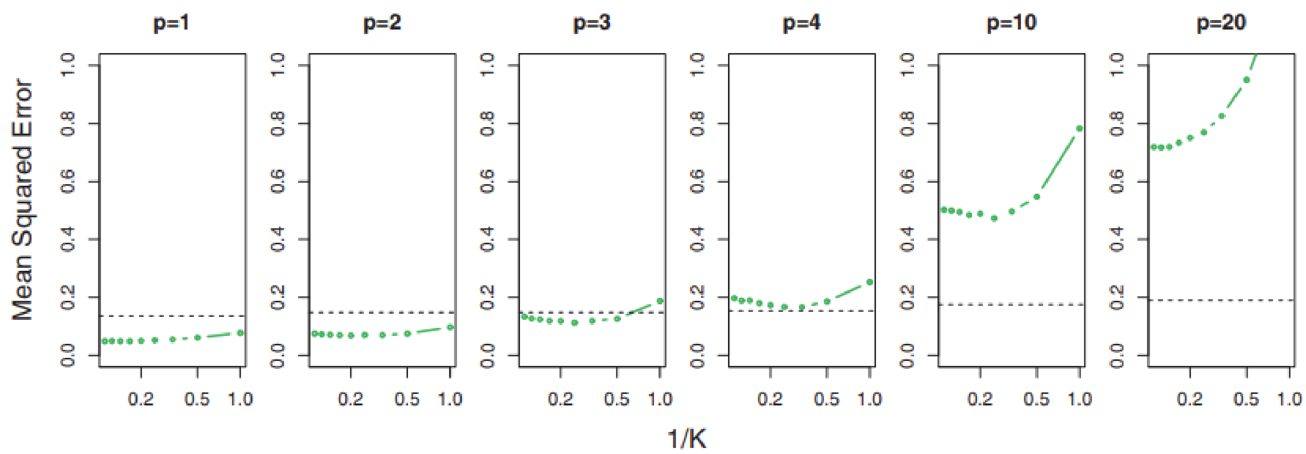


图 3-20　随着变量个数 p 的增加，线性回归（水平虚线）和 KNN（曲线）的 MSE。与图 3-19中的下图相同，第一个变量的真实函数在是非线性的且不依赖于其他变量。随着噪声变量的加入，线性回归的拟合效果逐渐变差，但 KNN 的拟合效果随着 p 的增加恶化得更快。

5. 线性回归应用与扩展

- 假设我们的角色是统计咨询师，需要根据这一数据提出一份营销计划，提高明年的产品销量，可能需要考虑：
- 广告预算和销量有关吗?
  - 可以通过拟合sales对TV、radio、newspaper的多元回归模型，再检验假设 $H_0: \beta_{TV} = \beta_{radio} = \beta_{newspaper} = 0$ ，分析F统计量可用于确定是否应该拒绝零假设。发现F统计量，其对应的p值是非常低的，表明有明确的证据支持广告投入和销量间存在相关性。

表 3-6 有关销量对电视，报纸，广播广告预算的最小二乘回归模型的更多信息，在 Advertising 数据集中。此模型的其他信息显示在表 3-4 内。

量	值
残差标准误	1.69
R^2	0.897
F 统计量	570

线性回归用于Advertising数据分析

- 广告预算和销量间的关系有多强?
  - 用两种测量模型精度的方法
  - 其一，RSE估计响应偏离总体回归直线的标准差。Advertising数据集的RSE为1681单位，而响应变量的平均值为14022，误差百分比约为12%。
  - R^2 统计量记录预测变量解释的响应变量变异的百分比。改预测解释几乎90%的销量方差。

表 3-6　有关销量对电视，报纸，广播广告预算的最小二乘回归模型的更多信息，在 Advertising 数据集中。此模型的其他信息显示在表 3-4 内。

量	值
残差标准误	1.69
R^2	0.897
F 统计量	570

- 哪种媒体能促进销售?
  - 可以检查每个预测变量的t统计量的p值。再多元线性回归中，TV和radio的p值很小，但newspaper的p值泽不然。这表明，只有TV和radio与sales相关。

	Coefficient	Std. Error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

- 如何精确地估计每种媒体对销量的影响?
  - $\hat{\beta}_j$ 的标准误差可以用来构造 β_j 的置信区间。Advertising数据集中，TV的95%置信区间是（0.043,0.049），radio的95%置信区间是（0.172,0.206），newspaper的95%置信区间是（-0.013,0.011）。TV和radio的置信区间都很窄且远离零点，这证明两种媒体都与sales相关。但newspaper的置信区间包括了零，这表明当TV和radio的费用给定时，报纸广告时统计不显著的。

- 标准误差可用于计算置信区间（confidence interval）。95%置信区间被定义为一个取值范围：该范围有95%的概率会包含未知参数的真实值。此范围是根据从样本数据计算出的上下限来定义的。对于线性回归模型， β_1 的95%置信区间约为:

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

线性回归用于Advertising数据分析

- 如何精确地估计每种媒体对销量的影响?
  - 为了评估每个媒体对销量的影响，可以建立三个独立的简单线性回归。有证据表明TV与sales之间、radio和sales之间有非常强的关联性。再忽略TV及radio两个变量的前提下，newspper和sales之间有适度的关联。

	系数	标准误	t统计量	p值
Intercept	7.032 5	0.457 8	15.36	<0.000 1
TV	0.047 5	0.002 7	17.67	<0.000 1

sales 关于 radio 的简单线性回归

	系数	标准误	t统计量	p值
Intercept	9.312	0.563	16.54	<0.000 1
radio	0.203	0.020	9.92	<0.000 1

sales 关于 newspaper 的简单线性回归

	系数	标准误	t统计量	p值
Intercept	12.351	0.621	19.88	<0.000 1
newspaper	0.055	0.017	3.30	<0.000 1

线性回归用于Advertising数据分析

- 对未来销量的预测精度如何?
  - $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$
  - 估计的准确性取决于我们想预测的时单个响应值 $Y = f(X) + \epsilon$ 还是平均响应值 $f(X)$。如果是前者，我们使用预测区间，如果是后者，我们使用置信区间。预测区间永远比置信区间宽，因为预测区间解释了的不确定性，是不可约误差。

```
> predict(lm.fit,data.frame(lstat=(c(5,10,15))), interval="confidence")#计算置信区间
       fit      lwr      upr
1 29.80359 29.00741 30.59978
2 25.05335 24.47413 25.63256
3 20.30310 19.73159 20.87461
> predict(lm.fit,data.frame(lstat=(c(5,10,15))), interval="prediction")#计算预测区间
       fit       lwr      upr
1 29.80359 17.565675 42.04151
2 25.05335 12.827626 37.27907
3 20.30310  8.077742 32.52846
```

confint()函数 可以得到系数估计值的置信区间。在根据给定lstat的值预测medv时，predict ()函数 可以计算置信区间和预测区间。例如当lstat等于10时，相应的95%置信区间为(24.47, 25. 63)，相应的95%预测区间为(12. 828, 37. 28)。正如预期的那样，置信区间和预测区间有相同的中心点(当lstat等于10时，medv的预测值是25. 05)，但后者要宽得多。

线性回归用于Advertising数据分析

- 这种关系是否是线性的?
  - 用残差图识别非线性，如果该关系是线性的，那么残差图应该显示不出规律（教材3.3.3节）

- 广告媒体间是否存在协同效应?
  - 标准线性回归模型假设预测变量和响应变量之间的关系是可加的，每个预测变量对响应变量的影响与其他预测变量无关。我们可以在回归模型中加入交互项以适用于非可加性关系。交互项的p值很小表明存在协同效应。对于Advertsing数据集，把交互项纳入模型将使统计量从90%大幅增加到97%。

$$
\begin{aligned}
\text{sales} &= \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times (\text{radio} \times \text{TV}) + \varepsilon \\
&= \beta_0 + (\beta_1 + \beta_3 \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \varepsilon
\end{aligned}
$$

- 该回归模型的最小二乘系数估计如下:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

- 在本课程的其余部分，许多时候在讨论扩展线性模型的方法，使之可以适应于：
  - **分类问题**：逻辑斯谛回归(logistic regression)、支持向量机(support vector machines)
  - **非线性模型**：核平滑(kernel smoothing)、样条和广义可加模型(splines and generalized additive models)、最近邻法(nearest neighbor methods)
  - **交互作用**：基于树的方法(tree-based methods)、袋装法(bagging)、随机森林(random forests)、自助法(boosting)
  - **正则化**：岭回归(Ridge regression)、lasso

本周作业
(9月27日第四周)

教材3.7习题2、5、9、10、11、13、14

上述内容下周二之前交（10月4日第五周）
本周三（9月28日）上机做/检查3.7习题9、10、11