回 归 分 析

参考教材:

- 《线性统计模型》, 王松桂、陈敏、陈立萍编著, 高等教育出版社;
- 《回归分析》, 周纪芗, 华东师范大学出版社.

注: 上课内容以这两本教材为基础, 做了一些修改或补充. 预修课程: 《微积分》(或《数学分析》)、《线性代数》(或《高等代数》)、《概率论》和《数理统计》.

课外参考书:

- 《近代回归分析》, 陈希孺、王松桂, 安徽教育出版社.
- 《Linear Models with R》(2nd Edition), Julian J. Faraway, CRC Press.

课程内容:

课程内容:

Chap1: 引论

Chap2: 随机向量

Chap3: 回归参数的估计

Chap4: 假设检验与预测

Chap5: 自变量的选择

Chap6: 含定性变量的回归模型

Chap7: 方差分析模型

课程内容:

Chap1: 引论

Chap2: 随机向量

Chap3: 回归参数的估计

Chap4: 假设检验与预测

Chap5: 自变量的选择

Chap6: 含定性变量的回归模型

Chap7: 方差分析模型

统计软件: R (免费软件).
下载地址：https://www.r-project.org/
注: 由于R的版本的原因, 课件中的某些代码可能会在其它电脑上失效.

课程成绩结构: 期末考试占70%, 平时成绩占10%, 期末大作业占20%.

助教: 毛缘(3140104420@zju.edu.cn).

理 论 作 业

第二章作业:

第二章作业:

1. 设$\boldsymbol{X}$为$n \times p$随机矩阵, $\boldsymbol{A}$为$p \times n$常数矩阵, 证明:

$$\mathsf{E}[\mathsf{tr}(\boldsymbol{AX})] = \mathsf{tr}[\mathsf{E}(\boldsymbol{AX})] = \mathsf{tr}[\boldsymbol{A}\mathsf{E}(\boldsymbol{X})].$$

2. 设$X_1, \cdots, X_n$为随机变量, $Y_1 = X_1, Y_i = X_i - X_{i-1}$, $i = 2, \cdots, n$. 记

$$\boldsymbol{X} = (X_1, \cdots, X_n)', \ \boldsymbol{Y} = (Y_1, \cdots, Y_n)'.$$

(1)若$\mathsf{Cov}(\boldsymbol{X}) = \boldsymbol{I}_n$, 其中$\boldsymbol{I}_n$是$n$阶单位阵, 求$\mathsf{Cov}(\boldsymbol{Y})$;
(2)若$\mathsf{Cov}(\boldsymbol{Y}) = \boldsymbol{I}_n$, 求$\mathsf{Cov}(\boldsymbol{X})$.

3. 设线性模型$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e}$, $\mathsf{E}(\boldsymbol{e})=\boldsymbol{0}, \mathsf{Cov}(\boldsymbol{e})=\sigma^2\boldsymbol{V}$, $\boldsymbol{X}$和$\boldsymbol{\beta}$都是非随机的. 若要$\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}$是$\sigma^2$的无偏估计($\boldsymbol{A}$ 是非随机矩阵), $\boldsymbol{A}$应满足什么条件?

4. 设$\boldsymbol{X}=(X_1,X_2,X_3)'\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, 其中

$$\boldsymbol{\mu}=(2,1,2)',\ \boldsymbol{\Sigma}=\begin{pmatrix}2&1&1\\1&3&0\\1&0&1\end{pmatrix}.$$

求$Y_1=X_1+X_2+X_3$与$Y_2=X_1-X_2$的联合分布.

5. 设$\boldsymbol{X}\sim N_n(\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}>0$, $\boldsymbol{A}$是$n\times n$实对称矩阵, 证明当$\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}=\boldsymbol{A}$时, $(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu})\sim\chi^2(r)$, 其中$r$是矩阵$\boldsymbol{A}$的秩.

6. 若$\boldsymbol{X}\sim N_n(\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}>0$, $\boldsymbol{A}$和$\boldsymbol{B}$是两个$n\times n$的实对称矩阵. 试给出二次型$(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu})$ 与$(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{B}(\boldsymbol{X}-\boldsymbol{\mu})$相互独立的条件.

第三章作业:

第三章作业:

1. 设

$$\begin{cases} y_i = \theta + e_i, & i = 1, \cdots, m, \\ y_{m+i} = \theta + \phi + e_{m+i}, & i = 1, \cdots, m, \\ y_{2m+i} = \theta - 2\phi + e_{2m+i}, & i = 1, \cdots, n, \end{cases}$$

其中θ, ϕ是未知参数, 各e_i相互独立, 且服从$N(0, \sigma^2)$.
(1) 写出设计矩阵$\boldsymbol{X}$;
(2) 求θ, ϕ的最小二乘估计$\hat{\theta}$和$\hat{\phi}$;
(3) 证明当$m = 2n$时, $\hat{\theta}$和$\hat{\phi}$不相关.

2. 对于下列的线性回归模型

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \ \ \boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n),$$

证明$\boldsymbol{\beta}$的最小二乘估计与极大似然估计是一致的.

3. 设

$$y_i = \beta_0 + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + e_i, i = 1, \cdots, n,$$

其中$e_i, i = 1, \cdots, n$独立同分布, 服从$N(0, \sigma^2)$. 令$\hat{\beta}_1$为β_1的最小二乘估计, 试证明

$$\mathsf{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 (1 - r_{12}^2)},$$

其中r_{12}是$(x_{i1}, x_{i2}), i = 1, \cdots, n$间的样本相关系数.

4. 设$y_1, \cdots, y_n$是来自$N(\theta, \sigma^2)$的独立同分布样本, 求θ的最小方差线性无偏估计$\hat{\theta}$, 并求$\mathsf{Var}(\hat{\theta})$.

5. 设$y_i \sim N(i\theta, i^2\sigma^2), i = 1, \cdots, n$且相互独立, 求$\theta$的最小方差线性无偏估计$\hat{\theta}$, 并求$\mathsf{Var}(\hat{\theta})$.

6. 设$\boldsymbol{A}$是$n \times n$的可逆矩阵, $\boldsymbol{u}$、$\boldsymbol{v}$为n维列向量, 试证:

$$(\boldsymbol{A}-\boldsymbol{u}\boldsymbol{v}')^{-1}=\boldsymbol{A}^{-1}+\frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}'\boldsymbol{A}^{-1}}{1-\boldsymbol{v}'\boldsymbol{A}^{-1}\boldsymbol{u}}.$$

并利用此结论证明以下结论: 设$\boldsymbol{X}_n$, $\boldsymbol{Y}_n$, $\hat{\boldsymbol{\beta}}_n$是$p$元(即有$p$个自变量)线性回归模型中基于$n$组观测的设计矩阵、观测向量及$\boldsymbol{\beta}$的最小二乘估计, 现获得了第$n+1$组观测

$$(x_{n+1,1},\cdots,x_{n+1,p},y_{n+1})=(\boldsymbol{x}_{n+1}',y_{n+1}),$$

又记

$$\boldsymbol{X}_{n+1}=\begin{pmatrix}\boldsymbol{X}_n\\ \boldsymbol{x}_{n+1}'\end{pmatrix},\boldsymbol{Y}_{n+1}=\begin{pmatrix}\boldsymbol{Y}_n\\ y_{n+1}\end{pmatrix},$$

$\hat{\boldsymbol{\beta}}_{n+1}$表示基于$n+1$组观测所得到的$\boldsymbol{\beta}$的最小二乘估计, 则有

$$\hat{\boldsymbol{\beta}}_{n+1}=\hat{\boldsymbol{\beta}}_n+\frac{(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{x}_{n+1}(y_{n+1}-\boldsymbol{x}_{n+1}'\hat{\boldsymbol{\beta}}_n)}{1+\boldsymbol{x}_{n+1}'(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{x}_{n+1}}.$$

7. 设$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e}, \mathsf{E}(\boldsymbol{e})=\boldsymbol{0}, \mathsf{Cov}(\boldsymbol{e})=\sigma^2\boldsymbol{I}_n$, $\boldsymbol{X}$是$n\times(p+1)$列满秩设计矩阵. 将$\boldsymbol{X}, \boldsymbol{\beta}$分块成

$$\boldsymbol{X}\boldsymbol{\beta}=(\boldsymbol{X}_1\ \ \boldsymbol{X}_2)\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}.$$

(1) 证明$\boldsymbol{\beta}_2$的最小二乘估计$\hat{\boldsymbol{\beta}}_2$由下式给出:

$$\hat{\boldsymbol{\beta}}_2=[\boldsymbol{X}_2'\boldsymbol{X}_2-\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2]^{-1}[\boldsymbol{X}_2'\boldsymbol{Y}-\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{Y}];$$

提示: 设$\boldsymbol{A}$为非奇异的对称矩阵, 将其分块为

$$\boldsymbol{A}=\begin{pmatrix}\boldsymbol{B} & \boldsymbol{C}\\ \boldsymbol{C}' & \boldsymbol{D}\end{pmatrix},$$

则当$\boldsymbol{B}^{-1}, \boldsymbol{D}^{-1}$都存在时有

$$
\begin{aligned}
\boldsymbol{A}^{-1} &= \begin{pmatrix} \boldsymbol{B}_1 & \boldsymbol{C}_1 \\ \boldsymbol{C}_1' & \boldsymbol{D}_1 \end{pmatrix} \\
&= \begin{pmatrix} (\boldsymbol{B}-\boldsymbol{C}\boldsymbol{D}^{-1}\boldsymbol{C}')^{-1} & -\boldsymbol{B}_1\boldsymbol{C}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}'\boldsymbol{B}_1 & \boldsymbol{D}^{-1}+\boldsymbol{D}^{-1}\boldsymbol{C}'\boldsymbol{B}_1\boldsymbol{C}\boldsymbol{D}^{-1} \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{B}^{-1}+\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{D}_1\boldsymbol{C}'\boldsymbol{B}^{-1} & -\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{D}_1 \\ -\boldsymbol{D}_1\boldsymbol{C}'\boldsymbol{B}^{-1} & (\boldsymbol{D}-\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{C})^{-1} \end{pmatrix}.
\end{aligned}
$$

(2) 求$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_2)$.

8. 对于线性回归模型$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, 假设$\boldsymbol{X}$的第一列元素全为1, 证明:
(1) $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$;
(2) $\sum_{i=1}^{n}\hat{y}_i(y_i - \hat{y}_i) = 0$.
其中$\hat{y}_i$是拟合值向量$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$的第$i$个分量, $\hat{\boldsymbol{\beta}}$是$\boldsymbol{\beta}$的最小二乘估计.

9. 设$\boldsymbol{Y} = \boldsymbol{\beta} + \boldsymbol{e}, \mathrm{E}(\boldsymbol{e}) = \boldsymbol{0}, \mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n$. 试用Lagrange乘子法证明: 在约束条件$\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$下, 使$\|\boldsymbol{Y} - \boldsymbol{\beta}\|^2$达到最小的$\boldsymbol{\beta}$值为

$$\hat{\boldsymbol{\beta}}_c = (\boldsymbol{I}_n - \boldsymbol{A}'(\boldsymbol{A}\boldsymbol{A}')^{-1}\boldsymbol{A})\boldsymbol{Y},$$

其中$\boldsymbol{A}$是已知的$q \times n$矩阵, 其秩为q.

10. 对于线性回归模型

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \mathsf{E}(\boldsymbol{e}) = \boldsymbol{0}, \mathsf{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{V},$$

其中$\boldsymbol{V}$为正定矩阵, $\boldsymbol{X}$为$n \times (p+1)$矩阵.
(1) 证明$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$仍然是$\boldsymbol{\beta}$的一个无偏估计.
(2) 证明$\mathsf{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$.
(3) 记$\hat{\sigma}^2 = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y}/(n-p-1)$, 证明

$$\mathsf{E}(\hat{\sigma}^2) = \frac{\sigma^2}{n-p-1}\mathsf{tr}[\boldsymbol{V}(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')].$$

第四章作业:

第四章作业:

1. 设

$$\begin{cases} y_1 = \beta_1 + e_1, \\ y_2 = 2\beta_1 - \beta_2 + e_2, \\ y_3 = \beta_1 + 2\beta_2 + e_3, \end{cases}$$

这里, $\boldsymbol{e} = (e_1, e_2, e_3)' \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_3)$. 试导出检验$H_0 : \beta_1 = 2\beta_2$的统计量.

2. 设

$$\begin{aligned} y_{1i} &\sim N(\beta_{10} + \beta_{11} x_{1i}, \sigma^2), \quad i = 1, \cdots, n, \\ y_{2i} &\sim N(\beta_{20} + \beta_{21} x_{2i}, \sigma^2), \quad i = 1, \cdots, m \end{aligned}$$

且相互独立, 试导出检验$H_0 : \beta_{11} = \beta_{21}$的统计量.

3. 对于一元线性回归模型

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \cdots, n,$$

$\mathsf{E}(e_i) = 0, \mathsf{Var}(e_i) = \sigma^2$, 各$e_i$互不相关. 记$\hat{\beta}_0$和$\hat{\beta}_1$分别为$\beta_0$和$\beta_1$的最小二乘估计, 记$\mu_0 = \beta_0 + \beta_1 x_0$为因变量$y$在$x_0$处取值$y_0$的均值, 即$\mathsf{E}(y_0) = \mu_0$. 用$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$估计$\mu_0$, 证明

$$\mathsf{Var}(\hat{\mu}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right].$$

4. 假设回归直线通过原点, 即一元线性回归模型为

$$y_i = \beta x_i + e_i, \quad , i = 1, \cdots, n,$$

$\mathsf{E}(e_0) = 0, \mathsf{Var}(e_i) = \sigma^2$, 各$e_i$互不相关.

(1) 写出β和σ^2的最小二乘估计;

(2) 记因变量y在x_0处的值y_0的预测值为$\hat{y}_0 = \hat{\beta} x_0$, 求$\mathsf{Var}(\hat{y}_0 - y_0)$;

(3) 记$\mu_0 = \beta x_0$, $\hat{\mu}_0 = \hat{\beta} x_0$, 求$\mathsf{Var}(\hat{\mu}_0)$.

上 机 作 业

秋第2周上机内容:

学习R语言(《统计建模与R软件》第二章).

秋第4周上机内容:

放假

秋第6周上机内容:

1. 学习R语言(《统计建模与 R 软件》).

2. 在动物学研究中, 有时需要找出某种动物的体积与重量的关系, 因为重量相对容易测量, 而测量体积比较困难. 我们可以利用重量预测体积的值. 下面是某种动物的18个随机样本的体重x(单位:kg) 与体积y(单位:$10^{-3}\mathrm{m}^3$)的数据:

x	y	x	y
17.1	16.7	15.8	15.2
10.5	10.4	15.1	14.8
13.8	13.5	12.1	11.9
15.7	15.7	18.4	18.3
11.9	11.6	17.1	16.7
10.4	10.2	16.7	16.6
15.0	14.5	16.5	15.9
16.0	15.8	15.1	15.1
17.8	17.6	15.1	14.5

(1) 画出散点图.
(2) 求回归直线$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$; 并画出回归直线的图像.
(3) 对体重$x_0 = 15.3$的这种动物, 预测它的体积y_0.

秋第8周上机内容:

为了研究水的耗氧量与周围环境的关系, 在实验室条件下, 对连续放置220天的水进行不断的测试, 共作了20次观测. 选取如下变量进行观察: 水的日耗氧量取对数(y), 生物耗氧量(x_1), 总的耗氧量(x_2), 固定物质含量(x_3), 挥发性固定物质含量(x_4), 化学物质耗氧量(x_5). 其中x_1到x_5的单位都是mg/L, y的单位是mg/min. 数据如下表所示. 试给出回归分析, 并进行回归诊断(包括模型的诊断和数据的诊断).

No	x_1	x_2	x_3	x_4	x_5	y
1	1125	232	7160	85.9	8905	1.5563
2	920	268	8804	86.5	7388	0.8976
3	835	271	8108	85.2	5348	0.7482
4	1000	237	6370	83.8	8056	0.716
5	1150	192	6441	82.1	6960	0.313
6	990	202	5154	79.2	5690	0.3617
7	840	184	5896	81.2	6932	0.1139
8	650	200	5336	80.6	5400	0.1139
9	640	180	5041	78.4	3177	-0.2218
10	583	165	5012	79.3	4461	-0.1549
11	570	151	4825	78.7	3901	0.0000
12	570	171	4391	78.0	5002	0.0000
13	510	243	4320	72.3	4665	-0.0969
14	555	147	3709	74.9	4642	-0.2218
15	460	286	3969	74.4	4840	-0.3979
16	275	198	3558	72.5	4479	-0.1549
17	510	196	4361	57.7	4200	-0.2218
18	165	210	3301	71.8	3410	-0.3919
19	244	327	2964	72.5	3360	-0.5229
20	79	334	2777	71.9	2599	-0.0458

冬第2周上机内容:

题1: 10次试验得观测数据如下:

y	16.3	16.8	19.2	18.0	19.5	20.9	21.1	20.9	20.3	22.0
x_1	1.1	1.4	1.7	1.7	1.8	1.8	1.9	2.0	2.3	2.4
x_2	1.1	1.5	1.8	1.7	1.9	1.8	1.8	2.1	2.4	2.5

若以x_1, x_2为回归自变量, 问它们之间是否存在多重共线性关系?

题2: 10次试验得观测数据如下:

y	16.3	16.8	19.2	18.0	19.5	20.9	21.1	20.9	20.3	22.0
x_1	1.1	1.4	1.7	1.7	1.8	1.8	1.9	2.0	2.3	2.4
x_2	1.1	1.5	1.8	1.7	1.9	1.8	1.8	2.1	2.4	2.5

试用岭迹法求y关于x_1, x_2的岭回归方程, 并画出岭迹图.

题3: 对某种商品的销量y进行调查, 并考虑有关的四个因素: x_1表示居民可支配收入, x_2表示该商品的平均价格指数, x_3表示该商品的社会保有量, x_4表示其它消费品平均价格指数. 下面是调查数据:

序号	x_1	x_2	x_3	x_4	y
1	82.9	92.0	17.1	94.0	8.4
2	88.0	93.0	21.3	96.0	9.6
3	99.9	96.0	25.1	97.0	10.4
4	105.3	94.0	29.0	97.0	10.4
5	117.7	100.0	34.0	100.0	12.2
6	131.0	101.0	40.0	101.0	14.2
7	148.2	105.0	44.0	104.0	15.8
8	161.8	112.0	49.0	109.0	17.9
9	174.2	112.0	51.0	111.0	19.6
10	184.7	112.0	53.0	111.0	20.8

利用主成分方法建立y与x_1,x_2,x_3,x_4的回归方程.

冬第4周上机内容:

河流的一个断面的年径流量y, 该断面的上游流域的年平均降水量x_1, 年平均饱和差x_2, 现共有14年的记录:

序号	x_1	x_2	y
1	720	1.80	290
2	553	2.67	135
3	575	1.75	234
4	548	2.07	182
5	572	2.49	145
6	453	3.59	69
7	540	1.88	205
8	579	2.22	151
9	515	2.41	131
10	576	3.03	106
11	547	1.83	200
12	568	1.90	224
13	720	1.98	271
14	700	2.90	130

(1)检验有无异常点,
(2)对回归方程的显著性作检验(显著性水平$\alpha = 0.05$),
(3)对每一个回归系数的显著性作检验(显著性水平$\alpha = 0.05$),
(4)设某年$x_1 = 600, x_2 = 2.50$, 求y的概率为0.95的预测区间.

冬第6周上机内容:

题1: 中国旅游业的现状分析: 国内旅游市场收入y(亿元)收到许多因素的影响, 我们选取如下的5个因素进行研究.
x_1: 国内旅游人数(万人次);
x_2: 城镇居民平均旅游支出(元);
x_3: 农村居民人均旅游支出(元);
x_4: 公路里程(万公里);
x_5: 铁路里程(万公里).
根据《中国统计年鉴》, 我们收集了1994-2010年度数据, 如下表. 试做自变量选择的分析.

年份	y	x_1	x_2	x_3	x_4	x_5
1994	1023.5	52400	414.7	54.9	111.78	5.9
1995	1375.7	62900	464	61.5	115.7	5.97
1996	1638.4	63900	534.1	70.5	118.58	6.49
1997	2112.7	64400	599.8	145.7	122.64	6.6
1998	2391.2	69450	607	197	127.85	6.64
1999	2831.9	71900	614.8	249.5	135.17	6.74
2000	3175.5	74400	678.6	226.6	140.27	6.87
2001	3522.4	78400	708.3	212.7	169.8	7.01
2002	3878.4	87800	739.7	209.1	176.52	7.19
2003	3442.3	87000	684.9	200	180.98	7.3
2004	4710.7	110200	731.8	210.2	187.07	7.44
2005	5285.86	121200	737.1	227.6	334.52	7.54
2006	6229.74	139400	766.4	221.9	345.7	7.71
2007	7770.62	161000	906.9	222.5	358.37	7.8
2008	8749.3	171200	849.4	275.3	373.02	7.97
2009	10183.69	190200	801.1	295.3	386.08	8.55
2010	12579.77	210300	883	306	400.83	9.12

题2: 对下列数据使用前进法、后退法和逐步回归法选择自变量.

数据: 下表给出了我国1991-2006年猪肉价格及其影响因素的数据. 在这个数据集中, y表示猪肉价格(元/公斤), x_1表示CPI, x_2表示人口数(亿), x_3表示年末存栏量(万头), x_4表示城镇居民可支配收入(元), x_5表示玉米价格(元/吨), x_6表示猪肉生成量(万吨).

年份	y	x_1	x_2	x_3	x_4	x_5	x_6
1990	9.84	103.1	5.28	36241	1510.2	686.7	2281
1991	10.32	103.4	5.89	36965	1700.6	590	2452
1992	10.65	106.4	5.87	38421	2026.6	625	2635
1993	10.49	114.7	6.01	39300	2577.4	726.7	2854
1994	9.16	124.1	6.45	41462	3496.2	1004.2	3205
1995	10.18	117.1	6.95	44169	4283	1576.7	3648
1996	14.96	107.9	7.58	36284	4838.9	1481.7	3158
1997	11.81	102.8	8.18	40035	5160.3	1150.8	3596
1998	10.77	99.2	9.14	42256	5425.1	1269.2	3884
1999	8.38	98.6	10.06	43020	5854	1092.5	3891
2000	8.74	100.4	10.42	44682	6280	887.5	4031
2001	10.18	100.7	10.55	45743	6859.6	1060	4184
2002	9.85	99.2	11.21	46292	7702.8	1033.3	4327
2003	10.7	101.2	11.45	46602	8472.2	1087.5	4519
2004	13.97	103.9	11.60	48189	9421.6	1288.3	4702
2005	13.39	101.8	12.98	50335	10493	1229.2	5011
2006	14.03	101.5	14.39	49441	13172	1280	5197

冬第8周上机内容:

题1: 某经济学家想调查文化程度对家庭储蓄的影响, 在一个中等收入的样本框中, 随机调查了13户高学历家庭和14户中低收入家庭. 因变量y表示上一年家庭储蓄增加额, 自变量x_1 为上一年度家庭总收入, 自变量x_2表示家庭学历, 其中$x_2 = 1$表示高学历家庭, 而$x_2 = 0$表示低学历家庭, 其调查数据如下表所示. 请分析学历对家庭储蓄增加额有无显著影响.

序号	y(元)	x_1(万元)	x_2	序号	y(元)	x_1(万元)	x_2
1	235	2.3	0	15	3265	3.8	1
2	346	3.2	1	16	3265	4.6	1
3	365	2.8	0	17	3567	4.2	1
4	468	3.5	1	18	3658	3.7	1
5	658	2.6	0	19	4588	3.5	0
6	867	3.2	1	20	6436	4.8	1
7	1085	2.6	0	21	9047	5.0	1
8	1236	3.4	1	22	7985	4.2	0
9	1238	2.2	0	23	8950	3.9	0
10	1345	2.8	1	24	9685	4.8	0
11	2365	2.3	0	25	9866	4.6	0
12	2365	3.7	1	26	10235	4.8	0
13	3256	4.0	1	27	10140	4.2	0
14	3256	2.9	0				

题2: 在一次关于公共交通的社会调查中, 一个调查项目是"是乘坐公共汽车上下班还是骑自行车上下班". 因变量$y = 1$表示主要乘公共汽车上下班, $y = 0$表示主要骑自行车上下班. 自变量x_1是年龄, x_2是月收入, x_3是性别(1表示男性, 0表示女性). 调查对象为工薪族群体, 数据见下表. 请建立Logistic回归模型.

序号	x_3	x_1	x_2(元)	y	序号	x_3	x_1	x_2(元)	y
1	0	18	850	0	15	1	20	1000	0
2	0	21	1200	0	16	1	25	1200	0
3	0	23	950	1	17	1	27	1300	0
4	0	23	950	1	18	1	28	1500	0
5	0	28	1200	1	19	1	30	950	1
6	0	31	850	0	20	1	32	1000	0
7	0	36	1500	1	21	1	33	1800	0
8	0	42	1000	1	22	1	33	1000	0
9	0	46	950	1	23	1	38	1200	0
10	0	48	1200	0	24	1	41	1500	0
11	0	55	1800	1	25	1	45	1800	1
12	0	56	2100	1	26	1	48	1000	0
13	0	58	1800	1	27	1	52	1500	1
14	1	18	850	0	28	1	56	1800	1

大 作 业

对大作业的要求: (1)能利用本课程的知识做一个案例分析(如进行数据的诊断、模型的诊断、多重共线性的诊断、线性回归、岭回归、主成分回归、变量选择、Logistic回归、方差分析等); 或(2)对某一主题进行扩展阅读, 完成一篇文献综述; 或(3)理论上的研究. 期末考试前发送文档(在文档中请写上学号、姓名、专业等信息)到txpang@zju.edu.cn 邮箱即可(邮件主题请注明"回归分析大作业"), 收到邮件后我将及时回复确认.

若是做案例分析, 可以自己寻找数据(通过校图书馆的"中国经济与社会发展统计数据"、"中经专网"等数据库寻找), 或者使用以下数据集: Advertising, Auto, Boston, Carseats, Smarket. 数据或数据来源见
https://pan.zju.edu.cn/share/e9ce9031b548bb36c1ca3ef7b5

数据6: 为研究某一单位员工业务能力测试成绩与员工智商之间的相关关系, 研究者对该单位20名员工进行了智商测试, 并收集了本年度这批员工的业务能力测试成绩, 测试数据如下表所示.

能力测试成绩与员工智商测试数据

编号	a	b	c	d	e	f	g	h	i	j
智商	89	98	126	87	119	101	130	115	108	105
成绩	55	74	87	60	71	54	90	73	67	70
编号	k	l	m	n	o	p	q	r	s	t
智商	84	121	97	101	92	110	128	111	99	120
成绩	53	82	58	60	67	80	85	73	71	90

数据7: 美国海军试图建立一些方程来估计操作某些海军设备所需的人力. 下表给出了关于BOQ设备的数据, 这些数据取自25个使用BOQ设备的点, 其中x_1表示每天平均占有率, x_2表示每月平均执勤人数, x_3表示每周服务台工作时间(小时数), x_4表示一般使用面积(平方英尺), x_5表示建筑物侧厅的数目, x_6表示操作余地, x_7表示房间数, y表示每月所需人力(人数每小时).

BOQ数据

序号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
1	2.00	4.00	4.0	1.26	1.0	6.0	6.0	180.23
2	3.00	1.58	40.0	1.25	1.0	5.0	5.0	182.61
3	16.60	23.78	40.4	1.0	1.0	13.0	13.0	164.38
4	7.00	2.37	168.0	1.0	1.0	7.0	8.0	284.55
5	5.30	1.67	42.5	7.79	3.0	25.0	25.0	199.92
6	16.50	8.25	168.0	1.12	2.0	19.0	19.0	267.38
7	25.89	3.00	40.0	0.0	3.0	36.0	36.0	990.09
8	44.42	159.75	168.0	0.60	18.0	48.0	48.0	1103.24

未完, 请转下页:

序号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
9	39.63	50.86	40.0	27.37	10.0	77.0	77.0	944.21
10	31.92	40.08	168.0	5.52	6.0	47.0	47.0	931.84
11	97.33	255.08	168.0	19.0	6.0	165.0	130.0	2268.06
12	56.63	373.42	168.0	6.03	4.0	36.0	37.0	1489.50
13	96.67	206.67	168.0	17.86	14.0	120.0	120.0	1891.70
14	54.38	207.08	168.0	7.77	6.0	66.0	66.0	1387.82
15	113.88	981.00	168.0	24.48	6.0	166.0	179.0	3559.92
16	149.58	233.83	168.0	31.07	14.0	185.0	202.0	3115.29
17	134.32	145.82	168.0	25.99	12.0	192.0	192.0	2227.76
18	188.74	937.00	168.0	45.44	26.0	237.0	237.0	4804.24
19	110.24	410.00	168.0	20.05	12.0	115.0	115.0	2628.32
20	96.83	677.33	168.0	20.31	10.0	302.0	210.0	1880.84
21	102.33	288.83	168.0	21.01	14.0	131.0	131.0	3036.63
22	274.92	695.25	168.0	46.63	58.0	363.0	363.0	5539.98
23	811.08	714.33	168.0	22.76	17.0	242.0	242.0	3534.49
24	384.50	1473.66	168.0	7.36	24.0	540.0	453.0	8266.77
25	95.00	368.00	168.0	30.26	9.0	292.0	196.0	1845.89

数据8: 下表给出了我国1991-2006年猪肉价格及其影响因素的数据. 在这个数据集中, y表示猪肉价格(元/公斤), x_1表示CPI, x_2表示人口数(亿), x_3表示年末存栏量(万头), x_4表示城镇居民可支配收入(元), x_5表示玉米价格(元/吨), x_6表示猪肉生成量(万吨).

年份	y	x_1	x_2	x_3	x_4	x_5	x_6
1990	9.84	103.1	5.28	36241	1510.2	686.7	2281
1991	10.32	103.4	5.89	36965	1700.6	590	2452
1992	10.65	106.4	5.87	38421	2026.6	625	2635
1993	10.49	114.7	6.01	39300	2577.4	726.7	2854
1994	9.16	124.1	6.45	41462	3496.2	1004.2	3205
1995	10.18	117.1	6.95	44169	4283	1576.7	3648
1996	14.96	107.9	7.58	36284	4838.9	1481.7	3158
1997	11.81	102.8	8.18	40035	5160.3	1150.8	3596
1998	10.77	99.2	9.14	42256	5425.1	1269.2	3884
1999	8.38	98.6	10.06	43020	5854	1092.5	3891
2000	8.74	100.4	10.42	44682	6280	887.5	4031
2001	10.18	100.7	10.55	45743	6859.6	1060	4184
2002	9.85	99.2	11.21	46292	7702.8	1033.3	4327
2003	10.7	101.2	11.45	46602	8472.2	1087.5	4519
2004	13.97	103.9	11.60	48189	9421.6	1288.3	4702
2005	13.39	101.8	12.98	50335	10493	1229.2	5011
2006	14.03	101.5	14.39	49441	13172	1280	5197

数据9: 研究者想研究采取某项保险革新措施的速度y与保险公司的规模x_1和保险公司类型之间的关系. 因变量y是某一公司采纳这项革新和给定公司采纳这项革新在时间上先后间隔的月数. 保险公司的规模变量x_1是数量型的, 它用公司的总资产额(百万美元)来计量. 而保险公司的类型变量x_2是定性变量, 它包括两种类型: 股份公司和互助公司. 资料数据如下表所示.

序号	y	x_1	公司类型	序号	y	x_1	公司类型
1	17	151	互助	11	28	164	股份
2	26	92	互助	12	15	272	股份
3	21	175	互助	13	11	295	股份
4	30	31	互助	14	38	68	股份
5	22	104	互助	15	31	85	股份
6	0	277	互助	16	21	224	股份
7	12	210	互助	17	20	166	股份
8	19	120	互助	18	13	305	股份
9	4	290	互助	19	30	124	股份
10	16	238	互助	20	14	246	股份