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, 做了一些修改和补充. 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:

Ch.1: 引论

Ch.2: 随机向量

Ch.3: 模型的估计

Ch.4: 模型的推断与预测

Ch.5: 自变量的选择

Ch.6: 含定性变量的回归模型

Ch.7: 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版本上失效

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: 曾子悦(12135031@zju.edu.cn)

上 机 作 业

有 待 更 新

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:

学习R语言(《统计建模与 R 软件》第二章).

秋第6周上机内容:

题1: 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.

题2: 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

秋第8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.

冬第2周上机内容:

题1: 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.

题2: 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.

题3: 数据集punting收集了13名志愿者关于10次美式足球的平均投球距离、悬挂时间与各种腿部力量测量的数据. 数据集中的7个变量及其含义:

Distance: average distance over 10 punts,
Hang: hang time,
RStr: right leg strength in pounds,
LStr: left leg strength in pounds,
RFlex: right hamstring muscle flexibility in degrees,
LFlex: left hamstring muscle flexibility in degrees,
OStr: overall leg strength in foot pounds.

问题:

(1) Fit a regression model with Distance as the response and the right and left leg strengths and flexibilities as predictors. Which predictors are significant at the 5% level?
(2) Use an F-test to determine whether collectively these four predictors have a relationship to the response.
(3) Relative to the model in (1), test whether the right and left leg strengths have the same effect.
(4) Plot a 95% confidence region for (bRStr,bLStr). Explain how the test in (3) relates to this region.
(5) Fit a model to test the hypothesis that it is total leg strength defined by adding the right and left leg strengths that is sufficient to predict the response in comparison to using individual left and right leg strengths.
(6) Relative to the model in (1), test whether the right and left leg flexibilities have the same effect.
(7) Test for left – right symmetry by performing the tests in (3) and (6) simultaneously.

冬第4周上机内容:

中国旅游业的现状分析: 国内旅游市场收入y(亿元)受到许多因素的影响, 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

冬第6周上机内容:

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: 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); 或对某一主题进行扩展阅读, 完成一篇文献综述; 或进行理论上的探索研究. 期末考试前上传到“学在浙大”.

若是做案例分析, 需自己寻找数据(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). 一些公开的数据集可从UCI Repository of machine learning databases找到(网址: https://archive.ics.uci.edu/ml).