数理统计大作业

GoodMorning29

修订日期：2024 年 12 月 25 日

§1 问题描述

在本次报告中，我们将研究影响 covid-19 的因素数据，数据来源自 `https://cmu-delphi.github.io/delphi-epidata/api/covidcast-signals/fb-survey.html` 变量包含 COVID-19 病例指数，流感样症状，家庭社区病例指数的加权平均值，口罩有效性信念加权平均值，社交距离有效性信念加权平均值，已接种 COVID 疫苗朋友数量加权平均值，室内大型活动可能性加权平均值，公共场所他人戴口罩加权平均值，公共场所他人保持社交距离加权平均值，室内购物可能性加权平均值，室内用餐可能性加权平均值，担心感染 COVID-19 加权平均值，家庭社区病例指数平均值，没有家庭社区病例指数数量，过去 7 天中戴口罩人数，使用公共交通人数，担心财务状况人数，测试结果阳性人数

§2 描述性统计分析

通过使用 describe 函数，我们可以得到数据的描述性统计信息，如下表所示：

	Min	Max	Mean	Std
COVID-19 病例指数	0.4596288	7.0576803	1.891488271915747	1.1412202859064147
流感样症状	0.4529342	7.1726694	1.9334389623871617	1.168167653033389
家庭社区病例指数的加权平均值	7.2873267	49.4753452	20.30829823640923	9.487699806663626
口罩有效性信念加权平均值	51.7568264	83.3207336	69.09639470065196	6.489895162207121
社交距离有效性信念加权平均值	52.3722481	82.5270216	70.0286744392678	5.2239347850366045
已接种 COVID 疫苗朋友数量加权平均值	47.0049154	84.5433939	64.13917479087262	8.158084484583995
室内大型活动可能性加权平均值	11.2290116	37.8592614	24.394597767652957	4.859814158550932
公共场所他人戴口罩加权平均值	1.6733289	68.4026846	22.80301803761284	18.523129395468253
公共场所他人保持社交距离加权平均值	8.9646801	33.7126527	18.68268443671013	4.286317050823027
室内购物可能性加权平均值	48.4968402	73.7287701	65.3819182100301	3.7606294336258044
室内用餐可能性加权平均值	18.5192973	50.538638	35.859607834954865	5.623884549313024
担心感染 COVID-19 加权平均值	23.4348538	67.4995176	46.109595108174524	7.196239098164835
家庭社区病例指数平均值	10.530873	55.9081779	25.199641660030093	10.407240894016546
没有家庭社区病例指数数量	7.0061728	50.0369904	20.17048090120361	9.671972145423368
过去 7 天中戴口罩人数	21.9239867	88.4838566	61.220362083400204	14.196858793585253
使用公共交通人数	1.5038476	14.8170065	4.604654748645938	2.1652528917039344
担心财务状况人数	27.9148098	48.0493836	36.71750434623872	3.5302399441284193
测试结果阳性人数	2.4418616	49.1204078	17.444779817101303	9.438732032936676

以测试结果阳性人数为例，可以发现其大致遵循正态分布。

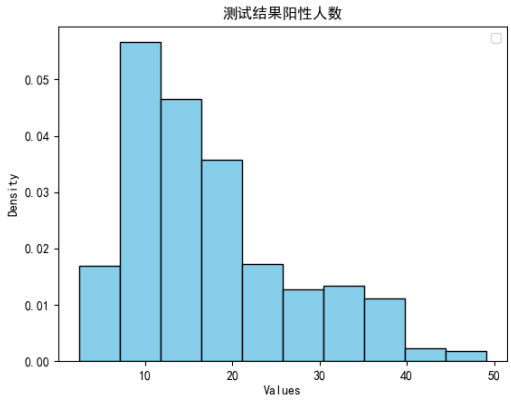

图 1: 测试结果阳性人数分布图

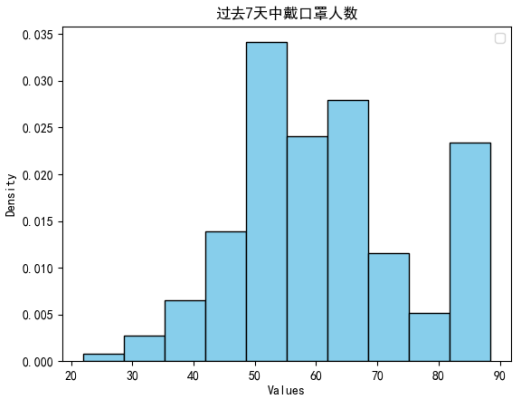

图 2: 七天内戴口罩人数分布图

再查看七天内戴口罩的人数，其分布并没有“中间多，两边少”的特点，这可能是因为过去七天中发生了一些特殊事件，导致人们的行为发生了变化。

§3 统计推断

3.1 分布函数的估计

由于因素过多，我们此处以担心感染 covid19 加权平均值数据为例，绘制数据的核密度估计（Kernel Density Estimation, KDE）曲线，以及正态分布拟合曲线，如下图所示：

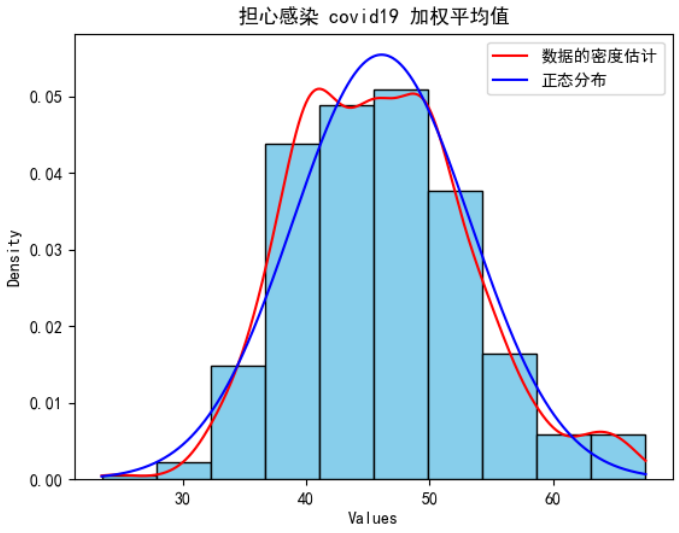

图 3: 担心感染 covid19 加权平均值数据的核密度估计曲线

可以看出，其分布与正态分布较接近，我们可以求出置信度为 95% 的置信区间，如下所示：

$$\text{均值的置信区间}: \left(\bar{\xi} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{\xi} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

$$\text{方差的置信区间}: \left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\right)$$

得到均值置信水平为 0.95 的置信区间为：(45.79362500004988, 46.42556521629917)，置信水平为 0.95 的方差置信区间为：(48.69075245509998, 55.130121666086886)

3.2 检验

在 Python 中，我们可以用 probplot 函数绘制 Q-Q 图，以检验数据是否符合正态分布。

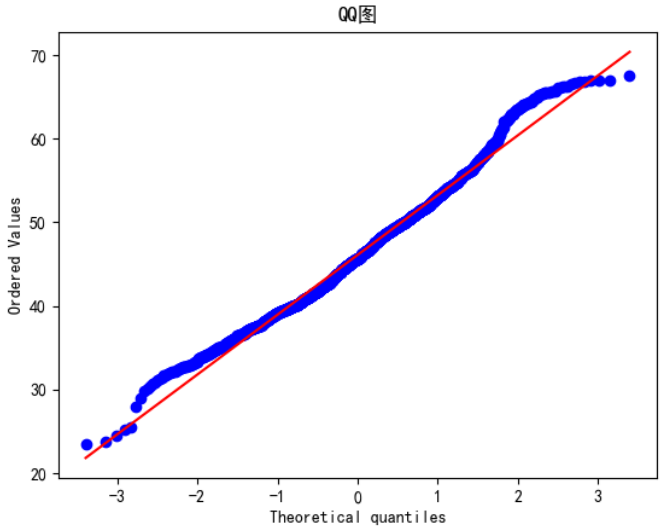

图 4: Q-Q 图

从图中可以看出，除两端数据点有偏差以外，数据点基本在直线上。同时我们可以绘制出其经验曲线和正态分布曲线，可以发现其大体符合正态分布。

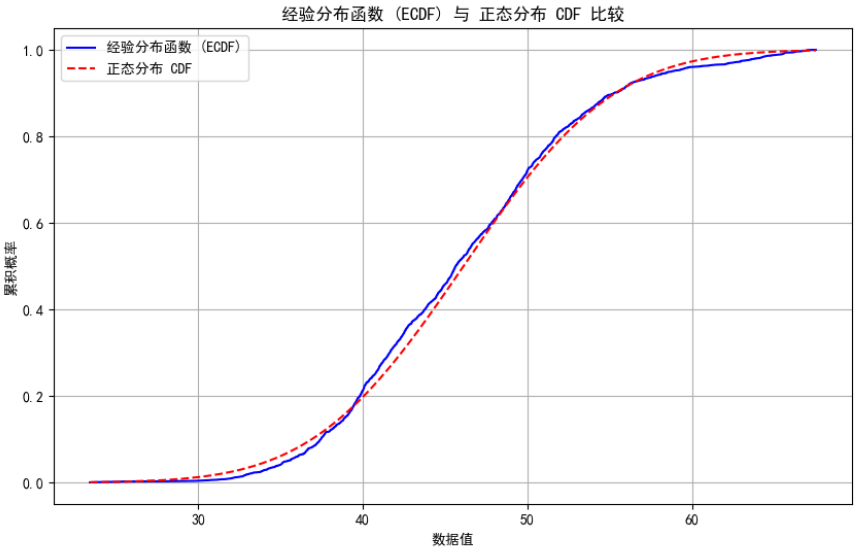

图 5: 经验曲线和正态分布曲线

但是通过 kurtosis 和 normaltest 函数进行峰度值和偏度值的检验，我们发现数据并不符合正态分布。

数据的偏度值：0.3728307323760319

数据的峰度值：0.1013558431733732

偏度显著性检验的统计量：44.586417382085095, p 值：2.080570716267187e-10

故正态性分析结果为：拒绝原假设，数据显著偏离正态分布。数据右偏（正偏）。数据具有较低的峰度，分布较平坦。

§4 方差分析

以七天内佩戴口罩为例，我们想知道佩戴口罩对检测为阳性的人数是否有影响。按照等距分组，我们将数据依据佩戴口罩人数分为 4 组，进行方差分析。通过调用函数 anova_lm 可以得到方差分析表，如下所示：

方差来源	平方和	自由度	样本方差	F 值	p 值
因素 (Factor: C(mask_group))	12475.625198	3	4158.541733	50.130206	3.133201×10^{-31}
误差 (Error)	165080.071944	1990	82.954810	–	–
总和 (Total)	177555.697142	1993	–	–	–

可以看出，F=50.130206>$F_{0.05}(3, 1990) \approx 0.12$，p 值为 3.133201×10^{-31}，拒绝原假设，即佩戴口罩人数对检测为阳性的人数有显著影响。

§5 回归分析

5.1 相关性及其度量

在进行相关分析和回归分析之前，可先通过不同变量之间的散点图直观地了解它们之间的关系和相关程度。若图中数据点分布在一条直线（曲线）附近，表明可用直线（曲线）近似地描述变量间的关系。若有多个变量，常制作多幅两两变量间的散点图来考察变量间的关系。先以家庭社区病例指数平均值和为例，画出散点图。Pyhon 中使用函数 plot( ) 可以方便地画出两个样本的散点图，从而直观地了解对应随机变量之间的相关关系和相关程度。输出结果如下

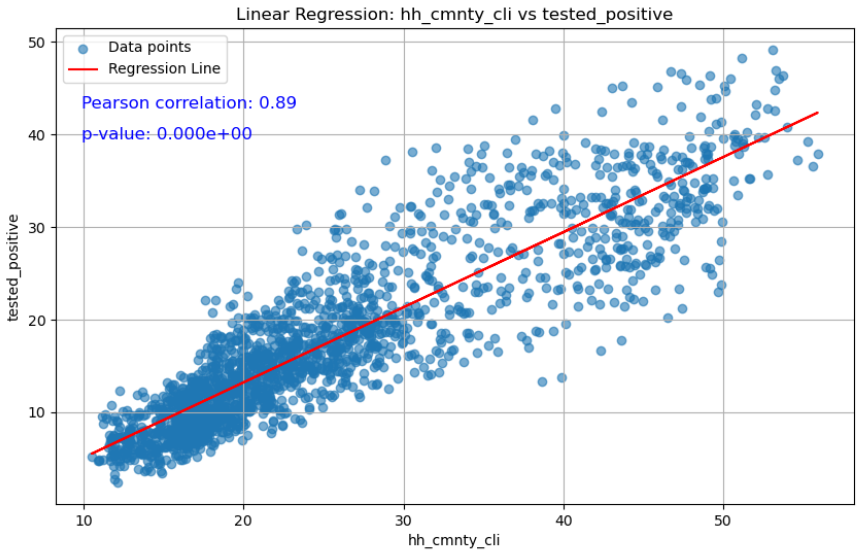

图 6: 家庭社区病例指数平均值与测试结果阳性人数的散点图

从图中可以看出，家庭社区病例指数平均值与测试结果阳性人数之间存在一定的线性关系。

进一步，它们之间的相关性可以用 Pearson 相关系数来度量。Pearson 相关系数是用来度量两个变量之间的线性相关程度的统计量，其值介于 -1 和 1 之间。当相关系数为 1 时，表示两个变量之间存在完全的正线性关系；当相关系数为 -1 时，表示两个变量之间存在完全的负线性关系；当相关系数为 0 时，表示两个变量之间不存在线性关系。在 Python 中，可以使用函数 corrcoef( ) 来计算两个变量之间的相关系数。输出结果如下：

Pearson correlation coefficient : 0.89

p-value : 0.000

由于 p-value 小于 0.05，故可以认为家庭社区病例指数平均值与测试结果阳性人数之间存在显著的线性相关关系。

5.2 一元线性回归分析

可以通过调用 statsmodels 包中的 OLS 函数来进行一元线性回归分析。可以通过调用函数 summary( ) 来查看回归分析的结果。输出结果如下：

系数 (Coef.)	标准误差 (Std. Err.)	t 值 (t)	p 值 (P>|t|)	95% 置信区间 (95% Conf. Interval)
常数项 (const)	-2.9848	0.248	-12.019	[-3.472, -2.498]
hh_cmnty_cli	0.8107	0.009	89.000	[0.793, 0.829]

回归模型摘要：

依赖变量	测试阳性 (tested_positive)	**Omnibus 检验值**	120.957
R 平方	0.799	**Omnibus 的 p 值**	0.000
调整后的 R 平方	0.799	**Jarque-Bera 检验值**	211.775
F 统计量	7921	**Durbin-Watson 统计量**	2.057
F 统计量的 p 值	0.00	**偏度**	0.455
样本量	1994	**峰度**	4.311
残差自由度	1992		
模型自由度	1		
对数似然值	-5705.1		
Akaike 信息准则 (AIC)	11410		
贝叶斯信息准则 (BIC)	11430		
协方差类型	非稳健 (nonrobust)		

检验误差

类似于 R 语言中的 lm.reg，我们可以通过 python 画四个图形

- 残差与拟合值的散点图
- Normal Q-Q 图
- 标准化残差的平方根的分布
- Cook's 距离

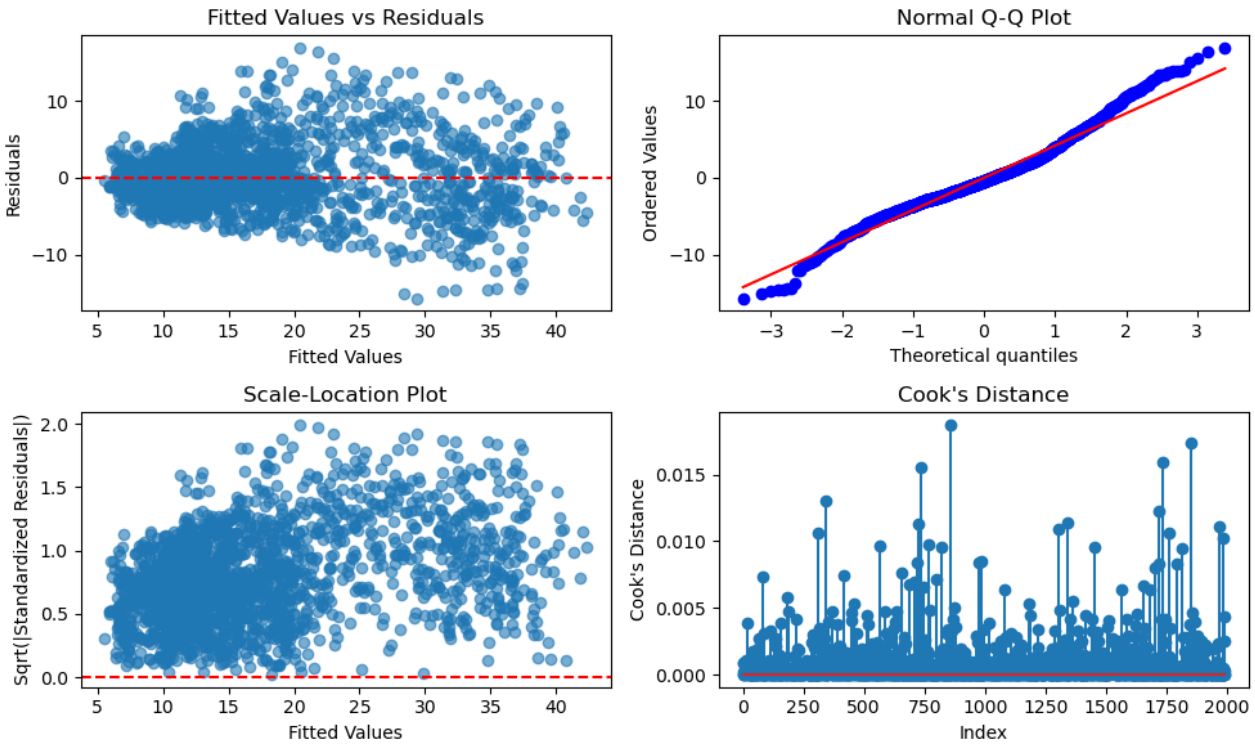

图 7: 检验误差

区间估计

使用 conf_int( ) 函数可以得到回归系数的置信区间，如下所示：

变量	系数	95% 置信区间
常数项 (const)	-2.9848	[-3.4719, -2.4978]
家庭社区病例指数平均值 (hh_cmnty_cli)	0.8107	[0.7928, 0.8286]

表 2: 回归系数的置信区间

5.3 多元线性回归分析

同样地，我们可以通过调用 statsmodels 包中的 OLS 函数来进行多元线性回归分析。可以通过调用函数 summary( ) 来查看回归分析的结果。输出结果如下：

依赖变量	测试阳性 (tested_positive)	**R 平方**	0.886
模型	OLS	**调整后的 R 平方**	0.885
方法	最小二乘法 (Least Squares)	**F 统计量**	906.7
日期	2024 年 12 月 25 日	**F 统计量的 p 值**	0.00
时间	01:23:07	**对数似然值**	-5136.8
样本量	1994	**Akaike 信息准则 (AIC)**	10310
残差自由度	1976	**贝叶斯信息准则 (BIC)**	10410
模型自由度	17	**协方差类型**	非稳健 (nonrobust)

变量	系数 (Coef.)	标准误差	t 值	p 值	95% 置信区间
const	58.9312	3.875	15.210	0.000	[51.333, 66.530]
cli	5.1484	1.510	3.410	0.001	[2.187, 8.110]
ili	-5.1648	1.480	-3.490	0.000	[-8.067, -2.263]
wnohh_cmnty_cli	-0.8288	0.108	-7.663	0.000	[-1.041, -0.617]
wbelief_masking_effective	-0.2443	0.050	-4.918	0.000	[-0.342, -0.147]
wbelief_distancing_effective	-0.3440	0.051	-6.684	0.000	[-0.445, -0.243]
wcovid_vaccinated_friends	0.1536	0.021	7.220	0.000	[0.112, 0.195]
wlarge_event_indoors	-0.1228	0.043	-2.862	0.004	[-0.207, -0.039]
wothers_masked_public	-0.2203	0.015	-14.440	0.000	[-0.250, -0.190]
wothers_distanced_public	0.0529	0.044	1.203	0.229	[-0.033, 0.139]
wshop_indoors	-0.4513	0.040	-11.385	0.000	[-0.529, -0.374]
wrestaurant_indoors	0.0658	0.039	1.704	0.089	[-0.010, 0.142]
wworried_catch_covid	-0.2097	0.032	-6.479	0.000	[-0.273, -0.146]
hh_cmnty_cli	0.6886	0.140	4.933	0.000	[0.415, 0.962]
nohh_cmnty_cli	0.7545	0.169	4.458	0.000	[0.423, 1.086]
wearing_mask_7d	0.3688	0.022	16.890	0.000	[0.326, 0.412]
public_transit	-0.1314	0.048	-2.724	0.006	[-0.226, -0.037]
worried_finances	-0.1086	0.025	-4.340	0.000	[-0.158, -0.060]

表 3: 多元线性回归分析结果

可见回归方程结果较好，大多数变量都通过了显著性检验，说明这些变量对测试阳性人数有显著影响。

变量选择

在代码中，我们实现了逐步回归算法，它是以 Akaike 信息准则（AIC）为准则的逐步回归算法，可以自动选择最优的变量组合。输出结果如下：最后删去了 wrestaurant_indoors 变量，得到了最优的变量组合。结果如下

依赖变量	测试阳性 (tested_positive)	**R 平方**	0.886
模型	OLS	**调整后的 R 平方**	0.885
方法	最小二乘法 (Least Squares)	**F 统计量**	963.0
日期	2024 年 12 月 25 日	**F 统计量的 p 值**	0.00
时间	01:23:07	**对数似然值**	-5137.5
样本量	1994	**Akaike 信息准则 (AIC)**	10310
残差自由度	1977	**贝叶斯信息准则 (BIC)**	10400
模型自由度	16	**协方差类型**	非稳健 (nonrobust)

此处相关回归分析放在附录，最后全部变量的 p 值均小于 0.05，说明这些变量对测试阳性人数有显著影响。

回归诊断

首先画出残差和标准化残差的散点图，如下图所示：

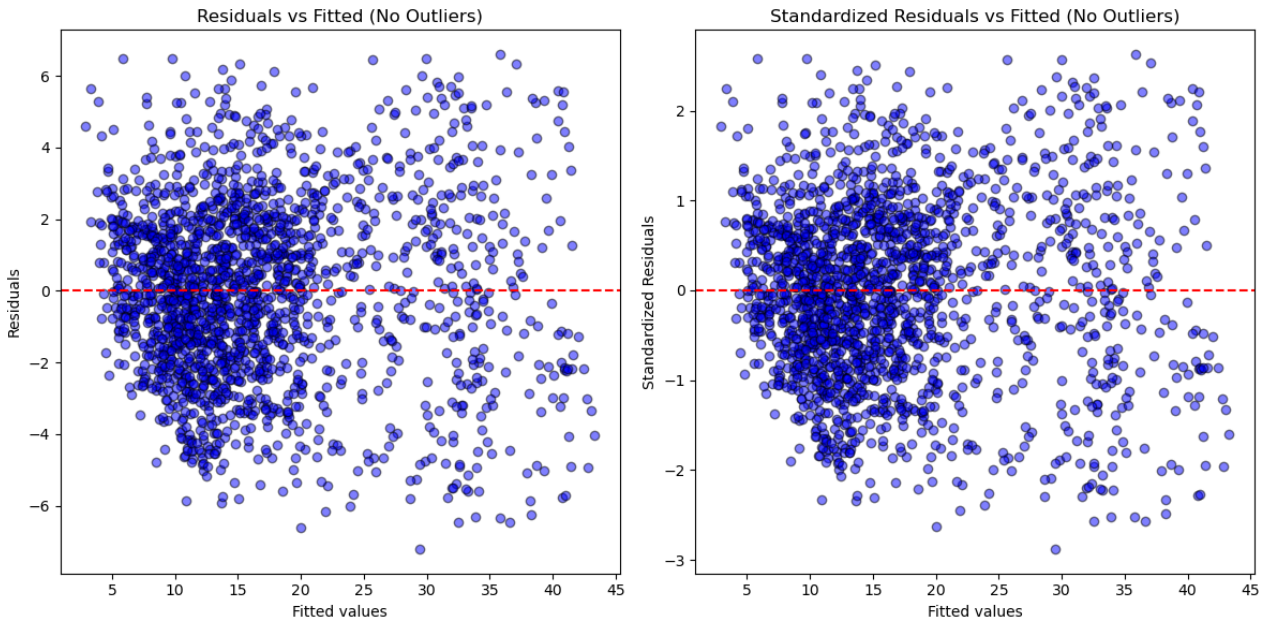

图 8: 残差和标准化残差的散点图

如果多元线性回归模型的假设成立，关于观测值的残差图中散点应该随机分布在 [-2,2] 之间里，此处大部分残差值在此范围内，说明模型的假设成立。

影响分析

从分析观测点对回归结果的影响入手，找出对回归结果影响很大的观测点的分析方法称为影响分析。在 python 中，函数 get_influence( ) 可以做回归诊断中影响分析的概括。输出结果见附录。

§6 附录

6.1 修正后多元线性回归分析

变量	系数 (Coef.)	标准误差	t 值	p 值	95% 置信区间
const	59.5500	3.841	15.505	0.000	[52.018, 67.082]
cli	5.3467	1.501	3.562	0.000	[2.403, 8.291]
ili	-5.3542	1.472	-3.639	0.000	[-8.240, -2.468]
wnohh_cmnty_cli	-0.8331	0.108	-7.705	0.000	[-1.045, -0.621]
wbelief_masking_effective	-0.2454	0.050	-4.941	0.000	[-0.343, -0.148]
wbelief_distancing_effective	-0.3256	0.049	-6.626	0.000	[-0.422, -0.229]
wcovid_vaccinated_friends	0.1455	0.020	7.209	0.000	[0.106, 0.185]
wlarge_event_indoors	-0.1291	0.043	-3.033	0.002	[-0.213, -0.046]
wothers_masked_public	-0.2104	0.013	-16.379	0.000	[-0.236, -0.185]
wshop_indoors	-0.4470	0.039	-11.321	0.000	[-0.524, -0.370]
wrestaurant_indoors	0.0557	0.038	1.477	0.140	[-0.018, 0.130]
wworried_catch_covid	-0.2158	0.032	-6.750	0.000	[-0.279, -0.153]
hh_cmnty_cli	0.6799	0.139	4.876	0.000	[0.406, 0.953]
nohh_cmnty_cli	0.7655	0.169	4.530	0.000	[0.434, 1.097]
wearing_mask_7d	0.3675	0.022	16.849	0.000	[0.325, 0.410]
public_transit	-0.1389	0.048	-2.903	0.004	[-0.233, -0.045]
worried_finances	-0.1042	0.025	-4.210	0.000	[-0.153, -0.056]

6.2 观测点

Strong influence points based on Cook's Distance: [13 41 43 51 53 58 65 84 89 99 112 134 136 154 168 172 175 189 220 239 257 283 291 310 314 328 335 343 355 410 452 457 469 499 516 520 552 566 586 589 594 595 598 613 625 657 664 667 669 688 700 719 723 737 763 765 767 773 777 799 819 820 821 837 858 866 872 954 965 967 973 987 991 1038 1048 1055 1059 1062 1081 1086 1089 1096 1109 1131 1133 1145 1149 1165 1172 1186 1217 1222 1236 1280 1288 1307 1311 1332 1340 1352 1354 1365 1398 1407 1428 1445 1449 1454 1467 1496 1513 1532 1543 1549 1551 1563 1583 1590 1591 1592 1597 1610 1622 1631 1654 1661 1685 1716 1720 1734 1760 1764 1770 1796 1816 1818 1834 1845 1855 1860 1862 1863 1951 1964 1970 1984]

High leverage points based on Leverage: [51 65 89 92 141 151 164 175 210 239 240 247 261 305 319 341 400 405 457 473 486 490 516 517 532 542 571 596 606 634 771 791 799 822 866 873 889 980 990 991 1048 1062 1086 1089 1138 1148 1172 1188 1236 1238 1244 1258 1302 1316 1326 1338 1379 1397 1398 1463 1470 1483 1507 1513 1514 1517 1528 1529 1539 1568 1603 1677 1740 1757 1760 1768 1788 1802 1819 1863 1870 1886 1986 1987]