模型三：

模型建立：

我们根据世界卫生组织发布的指南，从受试者体检指标中，得到受试者四种慢性病的情况。

高血压：收缩压>=140mmHg并且舒张压>=90mmHg

糖尿病：空腹血糖>= 7.0 mmol/L

高血脂：总胆固醇 >= 6.2mmol/L 或者 高密度脂蛋白 < 1.0 mmol/L 或者 低密度脂蛋白 >= 24.1 mmol/L 或者 总甘油三酯 >= 2.3mmol/L

高尿酸血症：血尿酸>= 420μmol/L

我们将得某慢性病的居民在对应列记作1，未得病则记作0

我们建立logistic二分类回归模型，来寻找慢性病和与吸烟、饮酒、饮食习惯、生活习惯、工作性质、运动等因素的关系以及相关程度。

logistic回归：

$$\ln (\frac{P(Y=1)}{1-P(Y=1)}) = a\_0 + \sum\_{i=1}^{n}a\_iX\_i$$

利用最小二乘法计算系数大小，从而得到$X\_i$与$Y$的相关性，进一步可以得到相关性显著程度。

模型求解：

高血压（占比5.2%）：

首先考虑所有可能有关变量，得到logistic回归如下：

| 二元Logit回归分析结果汇总 | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 项 | 回归系数 | 标准误 | *z* 值  | Wald χ2  | *p* 值 | OR值 | OR值95% CI |
| 工作主要属于以下何种活动 | 0.026 | 0.148 | 0.174 | 0.030 | 0.862 | 1.026 | 0.768 ~ 1.372 |
| 亲人高血压 | 0.543 | 0.133 | 4.067 | 16.543 | 0.000 | 1.721 | 1.325 ~ 2.235 |
| BMI | -0.001 | 0.007 | -0.137 | 0.019 | 0.891 | 0.999 | 0.986 ~ 1.012 |
| 腰围 | 0.050 | 0.011 | 4.511 | 20.347 | 0.000 | 1.051 | 1.028 ~ 1.074 |
| 臀围 | -0.002 | 0.013 | -0.159 | 0.025 | 0.873 | 0.998 | 0.974 ~ 1.023 |
| 体重 | 0.030 | 0.008 | 3.589 | 12.879 | 0.000 | 1.031 | 1.014 ~ 1.048 |
| 出生年 | -0.074 | 0.007 | -10.489 | 110.020 | 0.000 | 0.929 | 0.916 ~ 0.942 |
| 性别 | 0.249 | 0.199 | 1.251 | 1.566 | 0.211 | 1.283 | 0.868 ~ 1.895 |
| 文化程度 | -0.091 | 0.070 | -1.298 | 1.686 | 0.194 | 0.913 | 0.795 ~ 1.048 |
| 婚姻状况 | -0.161 | 0.143 | -1.124 | 1.262 | 0.261 | 0.852 | 0.644 ~ 1.127 |
| 是否吸烟 | -0.184 | 0.217 | -0.848 | 0.720 | 0.396 | 0.832 | 0.544 ~ 1.272 |
| 烟龄 | -0.011 | 0.010 | -1.177 | 1.385 | 0.239 | 0.989 | 0.970 ~ 1.008 |
| 开始吸烟年龄 | -0.011 | 0.017 | -0.668 | 0.446 | 0.504 | 0.989 | 0.957 ~ 1.022 |
| 平均每天体育锻炼时间 | -0.002 | 0.003 | -0.555 | 0.308 | 0.579 | 0.998 | 0.993 ~ 1.004 |
| 体育锻炼的强度 | -0.239 | 0.216 | -1.108 | 1.227 | 0.268 | 0.787 | 0.516 ~ 1.202 |
| 被动吸烟天数 | 0.013 | 0.025 | 0.527 | 0.278 | 0.598 | 1.013 | 0.965 ~ 1.064 |
| 一共吸烟支数 | -0.000 | 0.000 | -1.166 | 1.359 | 0.244 | 1.000 | 1.000 ~ 1.000 |
| 是否参加体育锻炼 | 0.025 | 0.062 | 0.400 | 0.160 | 0.689 | 1.025 | 0.908 ~ 1.157 |
| 平均每周吸烟天数 | -0.031 | 0.070 | -0.440 | 0.194 | 0.660 | 0.970 | 0.845 ~ 1.113 |
| 您做休闲、家务活动的强度 | -0.067 | 0.138 | -0.488 | 0.238 | 0.626 | 0.935 | 0.714 ~ 1.225 |
| 一天吸烟支数 | 0.025 | 0.021 | 1.152 | 1.328 | 0.249 | 1.025 | 0.983 ~ 1.069 |
| 腌制品 | 0.002 | 0.003 | 0.736 | 0.542 | 0.462 | 1.002 | 0.997 ~ 1.008 |
| 烹调油盐 | 0.001 | 0.002 | 0.649 | 0.421 | 0.516 | 1.001 | 0.998 ~ 1.004 |
| 奶类大豆坚果 | -0.001 | 0.000 | -1.147 | 1.315 | 0.251 | 0.999 | 0.999 ~ 1.000 |
| 是否饮酒 | 0.549 | 0.176 | 3.121 | 9.743 | 0.002 | 1.731 | 1.226 ~ 2.442 |
| 动物性食物 | -0.000 | 0.000 | -0.236 | 0.056 | 0.813 | 1.000 | 0.999 ~ 1.001 |
| 饮酒年数 | 0.023 | 0.011 | 2.115 | 4.474 | 0.034 | 1.023 | 1.002 ~ 1.045 |
| 新鲜蔬果 | -0.000 | 0.000 | -1.317 | 1.736 | 0.188 | 1.000 | 0.999 ~ 1.000 |
| 每周饮用酒精量 | -0.000 | 0.000 | -0.841 | 0.707 | 0.401 | 1.000 | 0.999 ~ 1.001 |
| 度数加权每次饮用量 | 0.000 | 0.001 | 0.415 | 0.172 | 0.678 | 1.000 | 0.999 ~ 1.002 |
| 谷薯类 | 0.001 | 0.000 | 1.366 | 1.867 | 0.172 | 1.001 | 1.000 ~ 1.001 |
| 度数加权饮用频率 | 0.004 | 0.002 | 2.624 | 6.885 | 0.009 | 1.004 | 1.001 ~ 1.008 |
| 不吃早餐 | -0.008 | 0.038 | -0.224 | 0.050 | 0.823 | 0.992 | 0.921 ~ 1.068 |
| 不吃中餐 | -0.372 | 0.210 | -1.770 | 3.132 | 0.077 | 0.689 | 0.457 ~ 1.041 |
| 不吃晚餐 | 0.358 | 0.157 | 2.283 | 5.210 | 0.022 | 1.431 | 1.052 ~ 1.946 |
| 截距 | 136.275 | 13.984 | 9.745 | 94.969 | 0.000 | 1.5260165339126868e+59 | 1.907804175694916e+47 ~ 1.220631808779127e+71 |
| 因变量: 高血压 | | | | | | | |
| McFadden *R* 方: 0.202 | | | | | | | |

观察显著性p值可知，亲人是否有得过高血压, 腰围, 体重, 是否饮酒, 饮酒年数, 酒精度数加权饮用频率, 不吃晚餐会对高血压产生显著的正向影响关系，以及出生年份会对高血压产生显著的负向影响关系。但是其他因素并不会对高血压产生显著影响。

删去影响不显著的变量，再做一次logistic回归，得到

| 二元Logit回归分析结果汇总 | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 项 | 回归系数 | 标准误 | *z* 值  | Wald χ2  | *p* 值 | OR值 | OR值95% CI |
| 亲人高血压 | 0.434 | 0.109 | 3.987 | 15.895 | 0.000 | 1.543 | 1.247 ~ 1.910 |
| 腰围 | 0.051 | 0.008 | 6.702 | 44.918 | 0.000 | 1.052 | 1.037 ~ 1.068 |
| 体重 | 0.019 | 0.006 | 3.248 | 10.550 | 0.001 | 1.019 | 1.007 ~ 1.030 |
| 出生年 | -0.064 | 0.005 | -14.086 | 198.408 | 0.000 | 0.938 | 0.929 ~ 0.946 |
| 是否饮酒 | 0.432 | 0.140 | 3.088 | 9.533 | 0.002 | 1.540 | 1.171 ~ 2.027 |
| 饮酒年数 | 0.016 | 0.008 | 1.900 | 3.608 | 0.057 | 1.016 | 1.000 ~ 1.032 |
| 不吃晚餐 | 0.049 | 0.099 | 0.491 | 0.241 | 0.623 | 1.050 | 0.864 ~ 1.275 |
| 度数加权饮用频率 | 0.002 | 0.001 | 2.315 | 5.359 | 0.021 | 1.002 | 1.000 ~ 1.004 |
| 截距 | 117.263 | 9.098 | 12.888 | 166.105 | 0.000 | 8.445859686979003e+50 | 1.520565658280882e+43 ~ 4.6911848537197707e+58 |
| 因变量: 高血压 | | | | | | | |
| McFadden *R* 方: 0.172 | | | | | | | |

即：

$$\ln(\frac{P(得高血压)}{1-P(得高血压)})=117.263 + 0.434\*亲人是否得过高血压 + 0.051\*腰围 + 0.019\*体重-0.064\*出生年 + 0.432\*是否饮酒 + 0.016\*饮酒年数 + 0.049\*不吃晚餐 + 0.002\*酒精度数加权饮酒频率$$

糖尿病（占比2.9%）：

首先考虑所有可能有关变量，得到logistic回归如下：

| 二元Logit回归分析结果汇总 | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 项 | 回归系数 | 标准误 | *z* 值  | Wald χ2  | *p* 值 | OR值 | OR值95% CI |
| 出生年 | -0.051 | 0.009 | -6.042 | 36.509 | 0.000 | 0.950 | 0.934 ~ 0.966 |
| 性别 | -0.150 | 0.279 | -0.538 | 0.289 | 0.591 | 0.861 | 0.498 ~ 1.487 |
| 是否吸烟 | 0.633 | 0.348 | 1.819 | 3.308 | 0.069 | 1.883 | 0.952 ~ 3.723 |
| 烟龄 | -0.005 | 0.012 | -0.423 | 0.179 | 0.672 | 0.995 | 0.973 ~ 1.018 |
| 平均每周吸烟天数 | 0.174 | 0.102 | 1.702 | 2.897 | 0.089 | 1.190 | 0.974 ~ 1.455 |
| 一天吸烟支数 | 0.035 | 0.025 | 1.430 | 2.044 | 0.153 | 1.036 | 0.987 ~ 1.087 |
| 一共吸烟支数 | -0.000 | 0.000 | -0.602 | 0.362 | 0.547 | 1.000 | 1.000 ~ 1.000 |
| 被动吸烟天数 | 0.053 | 0.031 | 1.701 | 2.894 | 0.089 | 1.055 | 0.992 ~ 1.121 |
| 是否饮酒 | 0.390 | 0.234 | 1.672 | 2.795 | 0.095 | 1.478 | 0.935 ~ 2.335 |
| 饮酒年数 | 0.010 | 0.016 | 0.635 | 0.404 | 0.525 | 1.010 | 0.979 ~ 1.042 |
| 度数加权饮用频率 | -0.000 | 0.002 | -0.060 | 0.004 | 0.952 | 1.000 | 0.996 ~ 1.004 |
| 度数加权每次饮用量 | -0.000 | 0.001 | -0.134 | 0.018 | 0.893 | 1.000 | 0.998 ~ 1.002 |
| 每周饮用酒精量 | -0.000 | 0.001 | -0.131 | 0.017 | 0.896 | 1.000 | 0.998 ~ 1.001 |
| 所有摄入酒精量 | 0.000 | 0.000 | 1.154 | 1.331 | 0.249 | 1.000 | 1.000 ~ 1.000 |
| 不吃早餐 | -0.042 | 0.046 | -0.918 | 0.842 | 0.359 | 0.959 | 0.876 ~ 1.049 |
| 不吃中餐 | 0.097 | 0.159 | 0.611 | 0.373 | 0.542 | 1.102 | 0.807 ~ 1.504 |
| 不吃晚餐 | 0.070 | 0.183 | 0.382 | 0.146 | 0.703 | 1.072 | 0.749 ~ 1.535 |
| 谷薯类 | -0.001 | 0.001 | -1.872 | 3.506 | 0.061 | 0.999 | 0.997 ~ 1.000 |
| 新鲜蔬果 | -0.000 | 0.000 | -1.113 | 1.240 | 0.266 | 1.000 | 0.999 ~ 1.000 |
| 动物性食物 | 0.000 | 0.001 | 0.272 | 0.074 | 0.785 | 1.000 | 0.999 ~ 1.001 |
| 奶类大豆坚果 | -0.001 | 0.001 | -1.210 | 1.465 | 0.226 | 0.999 | 0.998 ~ 1.000 |
| 工作主要属于以下何种活动 | -0.202 | 0.199 | -1.017 | 1.033 | 0.309 | 0.817 | 0.553 ~ 1.206 |
| 您做休闲、家务活动的强度 | -0.040 | 0.179 | -0.222 | 0.049 | 0.824 | 0.961 | 0.676 ~ 1.365 |
| 是否参加体育锻炼 | -0.085 | 0.086 | -0.989 | 0.978 | 0.323 | 0.919 | 0.777 ~ 1.087 |
| 身高 | -0.002 | 0.021 | -0.084 | 0.007 | 0.933 | 0.998 | 0.958 ~ 1.040 |
| 体重 | -0.007 | 0.019 | -0.376 | 0.142 | 0.707 | 0.993 | 0.957 ~ 1.030 |
| BMI | 0.002 | 0.026 | 0.058 | 0.003 | 0.953 | 1.002 | 0.952 ~ 1.053 |
| 腰围 | 0.103 | 0.015 | 6.831 | 46.662 | 0.000 | 1.109 | 1.077 ~ 1.142 |
| 臀围 | -0.040 | 0.012 | -3.246 | 10.534 | 0.001 | 0.961 | 0.937 ~ 0.984 |
| 体育锻炼的强度 | 0.503 | 0.336 | 1.497 | 2.242 | 0.134 | 1.653 | 0.856 ~ 3.192 |
| 平均每天体育锻炼时间 | 0.003 | 0.004 | 0.677 | 0.458 | 0.498 | 1.003 | 0.995 ~ 1.010 |
| 烹调油盐 | 0.004 | 0.002 | 2.176 | 4.737 | 0.030 | 1.004 | 1.000 ~ 1.007 |
| 腌制品 | 0.005 | 0.003 | 1.748 | 3.056 | 0.080 | 1.005 | 0.999 ~ 1.011 |
| 截距 | 90.568 | 17.099 | 5.297 | 28.057 | 0.000 | 2.1547643004419326e+39 | 6.013109184453817e+24 ~ 7.721478270281422e+53 |
| 因变量: 糖尿病 | | | | | | | |
| McFadden *R* 方: 0.176 | | | | | | | |
| Cox & Snell *R* 方：1.000 | | | | | | | |
| Nagelkerke *R* 方：1.000 | | | | | | | |

观察显著性p值可知：腰围, 烹调油盐，腌制品，亲人是否得过糖尿病会对糖尿病产生显著的正向影响关系，以及出生年, 臀围，谷薯类食物会对糖尿病产生显著的负向影响关系。但是其他因素并不会对糖尿病产生影响关系。由于臀围与实际情况过于不符，故删去。

删去影响不显著的变量，再做一次logistic回归，得到

| 二元Logit回归分析结果汇总 | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 项 | 回归系数 | 标准误 | *z* 值  | Wald χ2  | *p* 值 | OR值 | OR值95% CI |
| 出生年 | -0.056 | 0.005 | -10.422 | 108.624 | 0.000 | 0.945 | 0.935 ~ 0.955 |
| 腰围 | 0.066 | 0.007 | 9.751 | 95.073 | 0.000 | 1.068 | 1.054 ~ 1.082 |
| 谷薯类 | -0.002 | 0.001 | -2.689 | 7.230 | 0.007 | 0.998 | 0.997 ~ 1.000 |
| 腌制品 | 0.002 | 0.002 | 0.854 | 0.729 | 0.393 | 1.002 | 0.997 ~ 1.007 |
| 烹调油盐 | 0.003 | 0.002 | 1.670 | 2.787 | 0.095 | 1.003 | 1.000 ~ 1.006 |
| 亲人糖尿病 | 0.896 | 0.164 | 5.477 | 30.003 | 0.000 | 2.450 | 1.778 ~ 3.377 |
| 截距 | 101.760 | 10.670 | 9.537 | 90.958 | 0.000 | 1.5624205262139528e+44 | 1.2931172885651838e+35 ~ 1.8878085710564922e+53 |
| 因变量: 糖尿病 | | | | | | | |
| McFadden *R* 方: 0.143 | | | | | | | |
| Cox & Snell *R* 方：null | | | | | | | |
| Nagelkerke *R* 方：null | | | | | | | |

即：

$$\ln(\frac{P(得糖尿病)}{1-P(得糖尿病)})=101.760-0.056\*出生年 + 0.066\*腰围-0.002\*谷薯类 + 0.002\*腌制品 + 0.003\*烹调油盐 + 0.896\*亲人糖尿病$$

高血脂（占比32.9%）：

由上同理，我们可知：腰围, 体重, 是否饮酒会对高血脂产生显著的正向影响关系，以及出生年, 性别, 身高, 工作主要属于以下何种活动会对高血脂产生显著的负向影响关系。其他因素印象较小。二次处理，得到：

$$\ln(\frac{P(得高血脂)}{1-P(得高血脂)})=32.191-0.017\*出生年 + 0.056\*腰围-0.866\*性别 + 0.012\*体重-0.021\*身高-0.139\*工作主要属于以下何种活动 + 0.105\*是否饮酒$$

高尿酸血症（8.4%）：

出生年, 腰围会对高尿酸血症产生显著的正向影响关系，以及性别, 您做休闲、家务活动的强度, 烹调油盐会对高尿酸血症产生显著的负向影响关系。

$$\ln(\frac{P(得高尿酸血症)}{1-P(得高尿酸血症)})=-26.727 + 0.011\*出生年 + 0.060\*腰围-1.688\*性别-0.307\*您做休闲、家务活动的强度-0.006\*烹调油盐$$

问题四：

模型建立：

借用问题一与问题三的结论，我们可以大致将人群分为几大类：

慢性疾病患者：

高血压：收缩压>=140mmHg并且舒张压>=90mmHg

糖尿病：空腹血糖>= 7.0 mmol/L

高血脂：总胆固醇 >= 6.2mmol/L 或者 高密度脂蛋白 < 1.0 mmol/L 或者 低密度脂蛋白 >= 24.1 mmol/L 或者 总甘油三酯 >= 2.3mmol/L

高尿酸血症：血尿酸>= 420μmol/L

潜在患者：

非上述患者

问题三中四个模型

P(得高血压)>5.2%为潜在高血压患者；

P(得糖尿病)>2.9%为潜在糖尿病（高血糖）患者；

P(得高血脂)>32.9%为潜在高血脂患者；

P(高尿酸血症)>8.4%为潜在高尿酸血症患者

习惯不合理的健康居民：

非以上两种居民

有吸烟的；度数加权饮酒频率>34.5；按照问题一饮食习惯进行分类不合理的；运动低于标准的；BMI不在标准区间的。

习惯合理的健康居民

非以上三种居民

合理建议：

患者与潜在患者：

高血压：饮食要规律（吃晚饭），控制酒精摄入，控制脂肪摄入、多加运动（控制体重和腰围）。

糖尿病：少吃油盐以及腌制品，多加锻炼，多吃粗粮。

高血脂：多加锻炼，日常提高运动量（从事高强度工作，控制体重），戒酒。

高尿酸血症：少吃烹调油盐，日常需要多加活动。

习惯不合理的健康居民：按点随便说说即可