思路：

对于原本的数据集进行清洗，去除了详细的各个家庭成员的资料，因此关于教育程度什么的我没有做。

关于变量的整理，我在进一步地分析中只选取了主要的自变量：家庭人数、家庭结构、家庭所处城市、收入、物质满意度、全面满意度。

初步的分析：首先针对恩格尔系数作为因变量，采用向后逐步选择自变量，得到初步模型，家庭人数和收入有显著影响；之后将恩格尔系数调整成二分变量（<=2为富裕，>2为贫穷），用logistic回归模型得出，家庭人数、是否有宠物有更显著的影响，收入有作用但不显著。

之后的分析采用主要自变量，两种模型得到的效果都比较符合预期，即收入影响最为明显。进一步的分析，将是否有酒水作为因变量，得到对生活总体的满意度是关键的解释变量，即一户人家在吃饭时喝不喝酒和他们对生活的满意度厅有关系的。

R程序代码：

family.whole<-family[1:118,-(3:26)]

family.oamt<-na.omit(family.whole)

library(psych)

cor(family.oamt)

null=lm(engel.coef~1,data=family.oamt)

full=lm(engel.coef~.,data=family.oamt)

step(full, scope=list(lower=null, upper=full), direction="backward")

# backward method to get the final model

summary(lm(formula = engel.coef ~ fam.num + income + money.charged +

other.room + pet + dinner.importance + howdealrestfood +

fixed.stick + fruit, data = family.oamt))

Call:

lm(formula = engel.coef ~ fam.num + income + money.charged +

other.room + pet + dinner.importance + howdealrestfood +

fixed.stick + fruit, data = family.oamt)

Residuals:

Min 1Q Median 3Q Max

-1.19271 -0.36839 -0.03254 0.38132 1.94977

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.12095 0.57244 5.452 4.69e-07 \*\*\*

fam.num -0.21477 0.06245 -3.439 0.000902 \*\*\*

income -0.17305 0.06346 -2.727 0.007746 \*\*

money.charged 0.18709 0.13478 1.388 0.168679

other.room -0.18397 0.08716 -2.111 0.037702 \*

pet -0.29603 0.14230 -2.080 0.040475 \*

dinner.importance -0.17685 0.08850 -1.998 0.048855 \*

howdealrestfood 0.15349 0.10983 1.398 0.165839

fixed.stick 0.21120 0.10064 2.099 0.038775 \*

fruit -0.25658 0.13865 -1.851 0.067669 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6214 on 86 degrees of freedom

Multiple R-squared: 0.2985, Adjusted R-squared: 0.225

F-statistic: 4.065 on 9 and 86 DF, p-value: 0.0002265

# change the type to categorical data, unuseful for lm() function

# family.whole.fact[,1:15] <- data.frame(apply(family.whole.fact[1:15], 2, as.factor))

summary(glm(engel.coef~fam.num + income + money.charged +

other.room + pet + dinner.importance + howdealrestfood +

fixed.stick + fruit,,family=gaussian,data=family.oamt))

Call:

glm(formula = engel.coef ~ fam.num + income + money.charged +

other.room + pet + dinner.importance + howdealrestfood +

fixed.stick + fruit, family = gaussian, data = family.oamt)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.19271 -0.36839 -0.03254 0.38132 1.94977

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.12095 0.57244 5.452 4.69e-07 \*\*\*

fam.num -0.21477 0.06245 -3.439 0.000902 \*\*\*

income -0.17305 0.06346 -2.727 0.007746 \*\*

money.charged 0.18709 0.13478 1.388 0.168679

other.room -0.18397 0.08716 -2.111 0.037702 \*

pet -0.29603 0.14230 -2.080 0.040475 \*

dinner.importance -0.17685 0.08850 -1.998 0.048855 \*

howdealrestfood 0.15349 0.10983 1.398 0.165839

fixed.stick 0.21120 0.10064 2.099 0.038775 \*

fruit -0.25658 0.13865 -1.851 0.067669 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.3861212)

Null deviance: 47.333 on 95 degrees of freedom

Residual deviance: 33.206 on 86 degrees of freedom

AIC: 192.52

Number of Fisher Scoring iterations: 2

# Same solution for two methods in R

summary(glm(cbind(engel.coef<=2,engel.coef>2)~fam.num + income + money.charged +

other.room + pet + dinner.importance + howdealrestfood +

fixed.stick + fruit,,family=binomial,data=family.oamt))

Call:

glm(formula = cbind(engel.coef <= 2, engel.coef > 2) ~ fam.num +

income + money.charged + other.room + pet + dinner.importance +

howdealrestfood + fixed.stick + fruit, family = binomial,

data = family.oamt)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.84703 0.00003 0.19809 0.44628 1.45586

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -4.8026 3.4313 -1.400 0.1616

fam.num 1.0640 0.4512 2.358 0.0184 \*

income 0.9486 0.5014 1.892 0.0585 .

money.charged -1.2001 0.8575 -1.399 0.1617

other.room 18.7174 2558.3865 0.007 0.9942

pet 1.4452 0.6808 2.123 0.0338 \*

dinner.importance 0.4991 0.6230 0.801 0.4230

howdealrestfood 0.4281 0.7048 0.607 0.5436

fixed.stick -1.2915 0.6749 -1.914 0.0557 .

fruit 0.6672 0.8323 0.802 0.4228

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 68.350 on 95 degrees of freedom

Residual deviance: 47.598 on 86 degrees of freedom

AIC: 67.598

Number of Fisher Scoring iterations: 19

# Assign the engel.coef to the binomial variable, and use logistic model.

indep.variables<-colnames(family.oamt)[c(1:4,7:8)]

summary(glm(engel.coef~fam.num+fam.str+fam.city+income+matter.satisfy+whole.satisfy,

family=gaussian,data=family.oamt))

Call:

glm(formula = engel.coef ~ fam.num + fam.str + fam.city + income +

matter.satisfy + whole.satisfy, family = gaussian, data = family.oamt)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.2380 -0.5697 0.0577 0.2698 2.0235

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.22699 0.50600 6.377 7.84e-09 \*\*\*

fam.num -0.14703 0.06387 -2.302 0.0237 \*

fam.str -0.02566 0.06474 -0.396 0.6928

fam.city -0.02658 0.15166 -0.175 0.8613

income -0.22368 0.07125 -3.139 0.0023 \*\*

matter.satisfy -0.10453 0.10938 -0.956 0.3418

whole.satisfy 0.09367 0.13759 0.681 0.4978

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.4639991)

Null deviance: 47.333 on 95 degrees of freedom

Residual deviance: 41.296 on 89 degrees of freedom

AIC: 207.45

Number of Fisher Scoring iterations: 2

summary(glm(cbind(engel.coef<=2,engel.coef>2)

~fam.num+fam.str+fam.city+income+matter.satisfy+whole.satisfy,

family=binomial,data=family.oamt))

Call:

glm(formula = cbind(engel.coef <= 2, engel.coef > 2) ~ fam.num +

fam.str + fam.city + income + matter.satisfy + whole.satisfy,

family = binomial, data = family.oamt)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.6603 0.2673 0.3936 0.5104 1.0381

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.32899 2.38458 -1.396 0.1627

fam.num 0.63184 0.34936 1.809 0.0705 .

fam.str 0.01046 0.29658 0.035 0.9719

fam.city 0.41674 0.76645 0.544 0.5866

income 0.74425 0.37878 1.965 0.0494 \*

matter.satisfy 0.29377 0.59984 0.490 0.6243

whole.satisfy -0.13158 0.74178 -0.177 0.8592

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 68.350 on 95 degrees of freedom

Residual deviance: 61.563 on 89 degrees of freedom

AIC: 75.563

Number of Fisher Scoring iterations: 5

# Use specific independent variables to build the model.

summary(glm(cbind(wine==0,wine>0) ~ fam.num +

fam.str + fam.city + income + matter.satisfy + whole.satisfy,

family=binomial,data=family.oamt))

Call:

glm(formula = cbind(wine == 0, wine > 0) ~ fam.num + fam.str +

fam.city + income + matter.satisfy + whole.satisfy, family = binomial,

data = family.oamt)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.2151 -0.7617 0.5410 0.7636 1.4562

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.5833 1.8442 -0.316 0.75178

fam.num 0.2592 0.2256 1.149 0.25057

fam.str -0.5511 0.2224 -2.478 0.01322 \*

fam.city 1.1543 0.6177 1.869 0.06167 .

income 0.3243 0.2737 1.185 0.23612

matter.satisfy 1.1313 0.5106 2.215 0.02673 \*

whole.satisfy -1.6369 0.5917 -2.767 0.00566 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 112.144 on 95 degrees of freedom

Residual deviance: 94.584 on 89 degrees of freedom

AIC: 108.58

Number of Fisher Scoring iterations: 5

# similarly, we can do the glm() to all the binominal variables:

# garge, pet, fixed.seat, fixed.sticks, wine, dessert, drink, fruit, other.food, midnight.snack