Дз№1

Иван Мишалкин Олеся Тухватшина

df <- read.dta13("~/priklad/stata2.dta")

# Постановка задачи

В нашей выборке мы рассматриваем спрос только на макароны в 2010-2016 гг. Нашей задачей является предсказать, сколько будет потрачено домохозяйствами на макаронные изделия в неделю.

# Актуальность.

Спрос на продовольственные товары является важным компонентом в экономическом анализе различных продовольственных программ и сельскохозяйственной политике. Спрос помогает прогнозировать ситуацию на рынке, а именно цены, расходы и потребление. Информация о спросе используется во многих маркетинговых решениях.

# Обзор литературы.

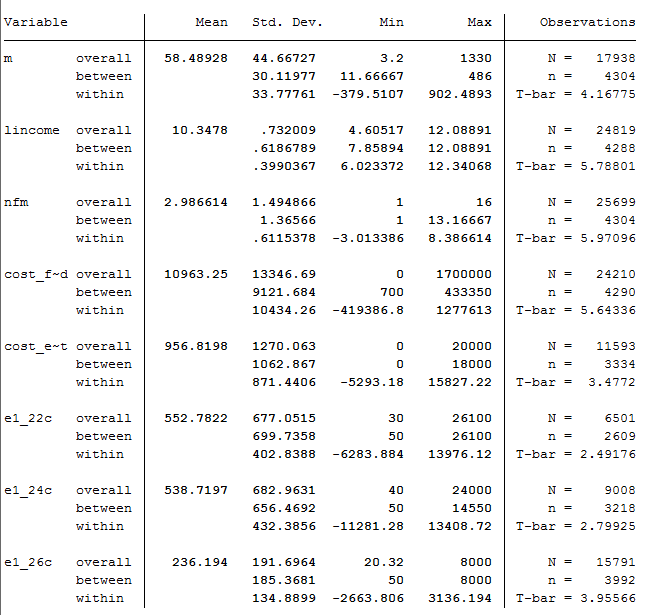
В статье James R. Blaylock и David M. Smallwood (1986) была рассмотрена тобит-модель, в которой в качестве зависимой переменной были взяты средние еженедельные расходы на питание домохозяйств. По выборке исследования Blaylock и Smallwood в рассматриваемый ими период 1881-1882 гг средний размер домохозяйств – 2,59, а процент домохозяйств, которые едят дома равен 67,6. 78,2% опрошенных покупали крупы и бакалейные продукты в неделю. Результатом их работы стало то, что домохозяйства с большим доходом тратят больше денег на продовольственные товары, чем домашние хозяйства с меньшим доходом, что касается макаронных изделий, то домохозяйства с высоким доходом в 1,17 раз тратят больше на бакалейные продукты. Помимо уровня дохода Blaylock и Smallwood рассмотрели так же такие факторы как возраст, регион и раса. Согласно результатам, пожилые американцы тратят меньше на еду, чем молодые люди; домашние хозяйства на северо-востоке и западе тратят больше на продовольствие, чем в странах Южного и Северного регионов; а «белые» люди расходуют больше, чем «черные».

# Описание данных

Были выброшены домохозяйства у которых было меньше 3 значений целевой переменной

**Stata code:**  
**egen dps\_availability = total(!missing(m)), by(id)**  
**drop if dps\_availability < 3**

Далее мы посмотрели на описательную статистику



Всего в выборке 4304 домохозяйств. Из матрицы объектов были удалены наблюдения с пропущенной целевой переменной. Максимальное число пропусков 7 из 7 лет и 9 доступным месяцам. В данны отсутствуют 6, 7 и 8 месяцадля всех лет.

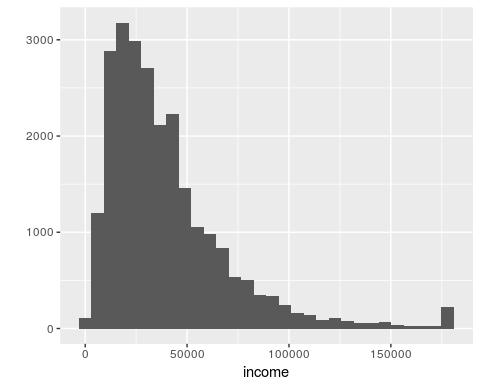
Вашным предиктором в модели может оказаться доход. Этот показатель содержит большое количество выбросов, поэтому все значения 99 перцентили сделаем равными 99 перцентилю.

Но данные имеют тяжелый правый хвост, а распределение похоже на лог-нормальное.

ggplot(df)+geom\_histogram(aes(income))+ylab('')

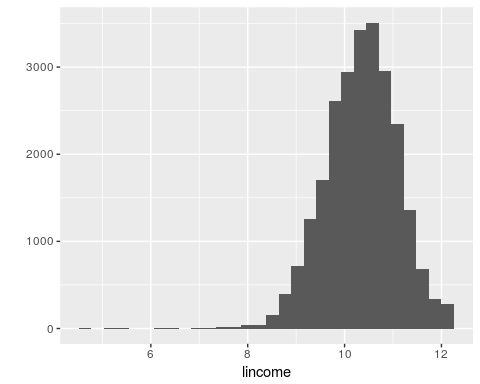
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 855 rows containing non-finite values (stat\_bin).



Прологарифмируем.

ggplot(df)+geom\_histogram(aes(lincome))+ylab('')



Далее соберем данные в панель.

pdata=plm.data(df,index = c("id", "year"))

## series p\_inc\_saved is NA and has been removed

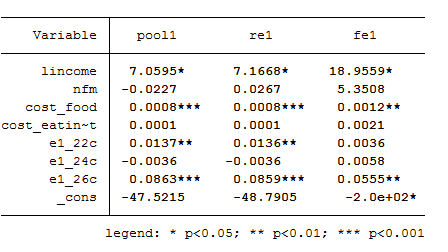
Оценим три модели. Первая Pool модель, затем FE и последняя - RE. Мы будем рассматривать количество потребления макарон от дохода, количества людей в семье, затраты на еду вообще и на еду вне дома. Также кажется разумным посмотреть на потребление мяса и птицы, потомучто макароны являются комплементом к ним. Так как FE регрессия не может учитывать постоянные во времени регрессоры(идея восстановить их с помощью константы для каждой группы), мы не включаем в наши модели тип населенного пункта и регион.

m.pooled <- plm(e1\_5c~lincome+nfm+cost\_food+cost\_eating\_out+e1\_22c+e1\_24c+e1\_26c,data=pdata, model = "pooling")  
m.re <- plm(e1\_5c~lincome+nfm+cost\_food+cost\_eating\_out+e1\_22c+e1\_24c+e1\_26c,data=pdata, model = "random")  
m.fe <- plm(e1\_5c~lincome+nfm+cost\_food+cost\_eating\_out+e1\_22c+e1\_24c+e1\_26c,data=pdata, model = "within")

stargazer(m.pooled,m.re,m.fe,type='text')

=======================================================================================  
 Dependent variable:   
 -----------------------------------------------------------------------  
 e1\_5c   
 (1) (2) (3)   
---------------------------------------------------------------------------------------  
lincome 7.060\*\* 7.611\*\* 18.956\*\*   
 (3.291) (3.329) (8.444)   
   
nfm -0.023 0.188 5.351   
 (1.133) (1.168) (4.031)   
   
cost\_food 0.001\*\*\* 0.001\*\*\* 0.001\*\*\*   
 (0.0002) (0.0002) (0.0004)   
   
cost\_eating\_out 0.0001 0.0001 0.002   
 (0.001) (0.001) (0.001)   
   
e1\_22c 0.014\*\*\* 0.013\*\*\* 0.004   
 (0.005) (0.005) (0.009)   
   
e1\_24c -0.004 -0.003 0.006   
 (0.004) (0.004) (0.009)   
   
e1\_26c 0.086\*\*\* 0.084\*\*\* 0.056\*\*\*   
 (0.012) (0.012) (0.021)   
   
Constant -47.521 -53.848   
 (33.283) (33.654)   
   
---------------------------------------------------------------------------------------  
Observations 960 960 960   
R2 0.176 0.173 0.209   
Adjusted R2 0.170 0.167 -1.562   
F Statistic 29.119\*\*\* (df = 7; 952) 28.375\*\*\* (df = 7; 952) 11.180\*\*\* (df = 7; 296)  
=======================================================================================  
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

То же самое в стате: **reg m lincome nfm cost\_food cost\_eating\_out e1\_22c e1\_24c e1\_26c**  
**est store pool1**   
**xtreg m lincome nfm cost\_food cost\_eating\_out e1\_22c e1\_24c e1\_26c, re**  
**est store re1**  
**xtreg m lincome nfm cost\_food cost\_eating\_out e1\_22c e1\_24c e1\_26c, fe**  
**est store fe1** **est tab pool1 re1 fe1, b(%7.4f) star**



По таблице можно видеть коэффициенты трех моделей. Сразу заметно, что достаточно мал, много незначимых переменных. Значимая константа, и это не очень хорошо, так как это говорит о том, что регрессоры плохо объясняют данные. Также можно заметить, что у FE модели отрицательный . Это так, потому что у нас большое количество групп, больше 10000. Столько же оценивается свободных членов.степеней свободы у FE модели всего 315

Некоторые коэффициенты достаточно сильно разнятся, значит могут быть проблемы со спецификацией. Также есть небольшая разница в коэффициентах между STATA и R, вероятно, это связано с различиями в оптимизациях "под капотом"

Далее необходимо выбрать наилучшую модель из представленных. Сначала сравним Pool с остальными. Для этого используем f test

pFtest(m.fe, m.pooled)

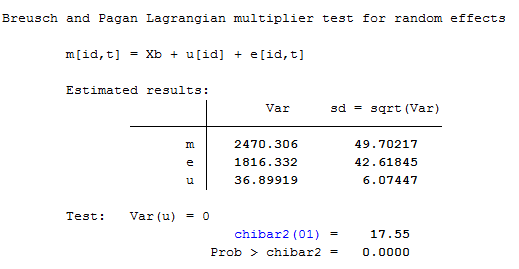
F test for individual effects  
  
data: e1\_5c ~ lincome + nfm + cost\_food + cost\_eating\_out + e1\_22c + ...  
F = 1.1864, df1 = 656, df2 = 296, p-value = 0.04519  
alternative hypothesis: significant effects

Нулевая гипотеза о верной сквозной модели отвергается в пользу модели с фиксированными эффектами.

Следующим шагом проверим Pool про RE. Для этого воспользуемся тестом Бройша-Пагана

plmtest(m.re, type = "bp")

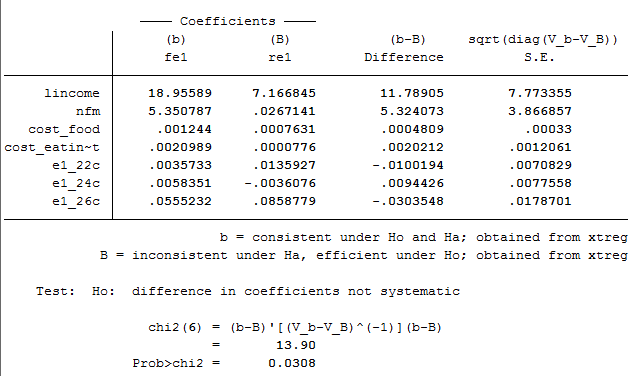
Lagrange Multiplier Test - (Breusch-Pagan) for unbalanced panels  
  
data: e1\_5c ~ lincome + nfm + cost\_food + cost\_eating\_out + e1\_22c + ...  
chisq = 17.547, df = 1, p-value = 0.00002803  
alternative hypothesis: significant effects



Нулевая гипотеза о верности сквозной регрессии отвергается. Остается проверить, какую из двух оставшихся регрессий выбрать. Для этого воспользуемся тестом Хаусмана.

phtest(m.fe, m.re)

Hausman Test  
  
data: e1\_5c ~ lincome + nfm + cost\_food + cost\_eating\_out + e1\_22c + ...  
chisq = 22.759, df = 7, p-value = 0.001877  
alternative hypothesis: one model is inconsistent

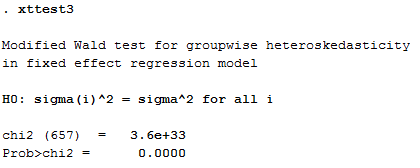


Здесь нулевая гипотеза о состоятельности коэффициентов в обеих моделях (FE и RE) отвергается. Так как этот тест проверяет связаны ли ошибки с регрессорами, то следует в дальнейшем пользоваться моделью FE. Это означает, что ошибки некоррелированны с регрессорами.

Выбор FE модели означает, что мы не можем считать различия в домохозяйствах случайными. Именно от этих особенностей зависит потребление макарон.

# Тесты

## Тест на гетероскедастичность



Тест свидетельствует о том, что присутствует гетероскедастичность и необходимы будут робастные оценки

## Тест на пространственную автокорреляцию

К сожалению не удалось запустить функцию **xttest2**. Любые попытки обрезать датасет не приводили ни к какому результату. Мы попробовали оставить только наблюдения, для которых есть наблюдения целевой переменной за все года, попытались обрезать панель по годам в разных комбинациях. Поэтому мы снова возвращаемся к R.

pcdtest(m.fe, test = 'lm')

## Warning in pcdres(tres = tres, n = n, w = w, form = paste(deparse(x  
## $formula)), : Some pairs of individuals (200 percent) do not have any or  
## just one time period in common and have been omitted from calculation

##   
## Breusch-Pagan LM test for cross-sectional dependence in panels  
##   
## data: e1\_5c ~ lincome + nfm + cost\_food + cost\_eating\_out + e1\_22c + e1\_24c + e1\_26c  
## chisq = 7583.2, df = 3887, p-value < 2.2e-16  
## alternative hypothesis: cross-sectional dependence

Нулевая гипотеза об отсутствии пространственной автокорреляции отвергается

pcdtest(m.fe, test = 'cd')

## Warning in pcdres(tres = tres, n = n, w = w, form = paste(deparse(x  
## $formula)), : Some pairs of individuals (200 percent) do not have any or  
## just one time period in common and have been omitted from calculation

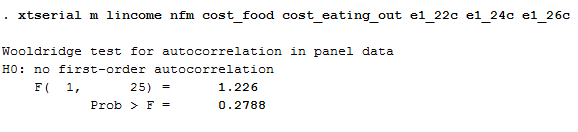
##   
## Pesaran CD test for cross-sectional dependence in panels  
##   
## data: e1\_5c ~ lincome + nfm + cost\_food + cost\_eating\_out + e1\_22c + e1\_24c + e1\_26c  
## z = -0.9624, p-value = 0.3358  
## alternative hypothesis: cross-sectional dependence

CD тест говорит об обратном. Какая же гипотеза верна?

"Monte Carlo experiments show that the standard Breusch–Pagan LM test performs badly for N > T panels, whereas Pesaran’s CD test performs well even for small T and large N" Baltagi (2005, 3rd edition), p. 247

Поэтому доверимся CD тесту и скажем, что нет оснований отвергнуть нулевую гипотезу о том, что пространственной автокорреляции нет.

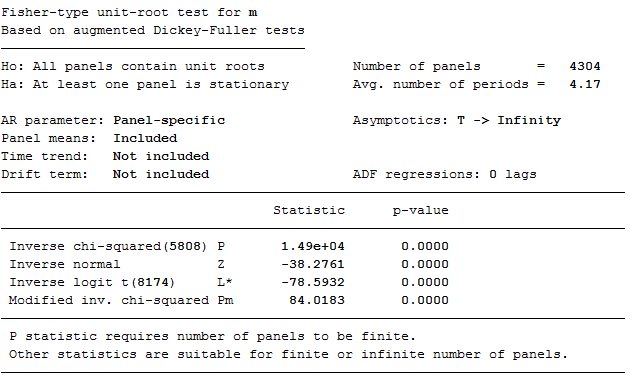
## Тест на временную автокорреляцию



На разумных уровнях значимости мы не можем отвергнуть нулевую гипотезу о том, что в наших отсутсвует автокорреляция первого порядка

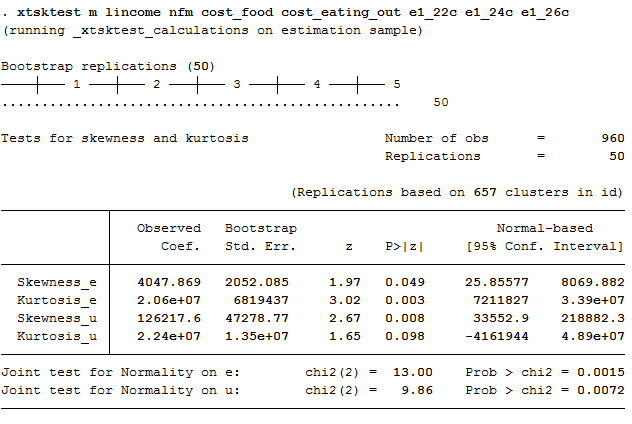
## Тест на стационарность

**xtunitroot fisher m, dfuller lags(0)**



Для многих домохозяйств тест не мог быть посчитан. Однако на тех данных что есть нулевая гипотеза о том, что все панели стационарны отвергается

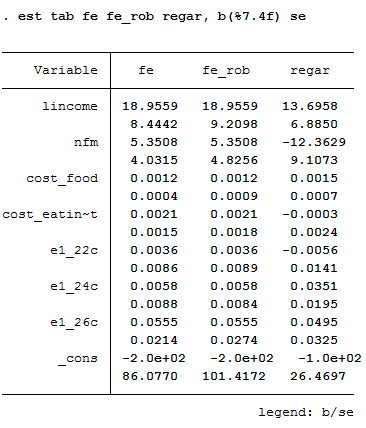
## Тест на нормальность



Нулевая гипотеза о нормальности отвергается.

## Поправка на гетероскедастичность и автокорреляцию

**xtreg m lincome nfm cost\_food cost\_eating\_out e1\_22c e1\_24c e1\_26c, fe**  
**est store fe**  
**xtreg m lincome nfm cost\_food cost\_eating\_out e1\_22c e1\_24c e1\_26c, fe robust**  
**est store fe\_rob**  
**xtregar m lincome nfm cost\_food cost\_eating\_out e1\_22c e1\_24c e1\_26c, fe**  
**est store regar**  
**xtregar m lincome nfm cost\_food cost\_eating\_out e1\_22c e1\_24c e1\_26c, fe** **est store regar**

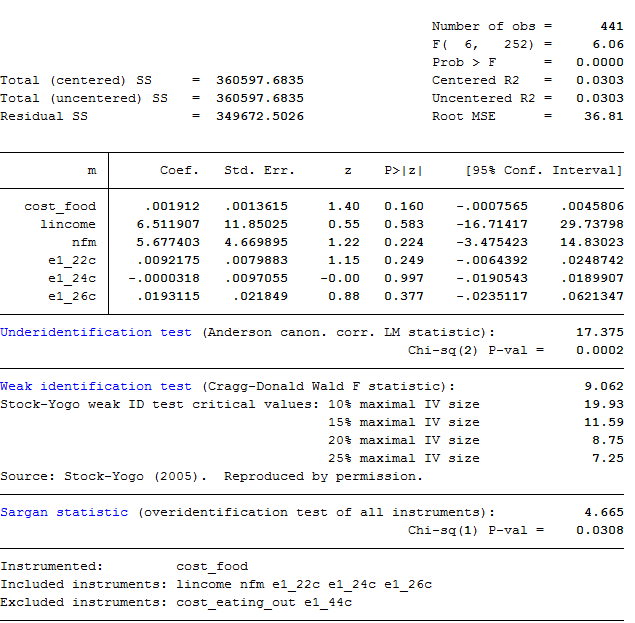


# Инструментальные переменные

Возможно, в нашей модели есть проблема эндогенности. Попробуем оценить модель с инструментальными переменными.

Будем считать, что общие затраты на еду это эндогенная величина. Также будем считать, что еда вне дома и покупка тортов - не связаны с индивидуальными ошибками и коррелированы с количеством денег, потраченных на еду.

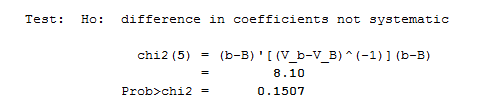
Далее используем наши инструменты и оценим модель с помощью двухшагового МНК. Оценки эффективны и состоятельны только в случае гомоскедастичности. Мы закрываем на это глаза, потому что нам не доступны другие процедуры.



Видим, что F-статистика близка к 10, но все же меньше. Тест Саргана нам говорит, что при 1% уровне значимости мы можем говорить, что инструменты валидны.

Далее мы проводим тест Хаусмана, что бы выбрать между обычной fe моделью и fe с инструментами

**xtivreg m lincome nfm e1\_22c e1\_24c e1\_26c (cost\_food = cost\_eating\_out e1\_44c), fe**  
**est store fe\_iv**  
**xtreg m lincome nfm e1\_22c e1\_24c e1\_26c cost\_food, fe**  
**est store fe**  
**hausman fe\_iv fe**



Результаты говорят о том, что разницы между моделями нет, но, наверное, стоит выбрать ту, которая даст наиболее эффективные оценки. Скорее всего, это будет fe без инструментов.

# Список литературы

James R. Blaylock R., David M. Smallwood. 1986. U.S. Demand for Food: Household Expenditures, Demographics, and Projections. Technical bulletin 1713 (Washington D.C., U.S. Department of Agriculture, Economic Research Service, February)