Прикладная статистика в R

Лекция 7. Построение модели линейной регрессии для спроса на говядину

Елена Михайловна Парилина

д. ф.-м. н., проф.

2021

Данные "BeefDemand"

Название: BeefDemand.txt
Объем выборки: 36 наблюдений, 10 переменных
Описание: Массив данных состоит из нескольких переменных, которые могут повлиять на спрос на говядину в США.
Описание переменных:
Year: год
ChickPrice: Розничная цена на курицу в центах за фунт
BeefPrice: Розничная цена на говядину в центах за фунт
BeefConsump: Потребление говядины на душу населения в фунтах
CPI: Индекс потребительских цен (CPI) на продукты питания
DPI: Располагаемый личный доход на душу населения в долларах
RealChickPrice: Розничная цена на курицу с поправкой на инфляцию в центах за фунт
RealBeefPrice: Розничная цена на говядину с поправкой на инфляцию в центах за фунт
RealDPI: Располагаемый личный доход на душу населения с поправкой на инфляцию в долларах
(RDPI-Mean): Дисперсия RDPI.

Замечания по массиву данных "BeefDemand"

Первые три из последних четырех переменных выводятся из их аналогов путем деления их значений на соответствующий CPI и умножения на 100. Последняя переменная — дисперсия.

Задача

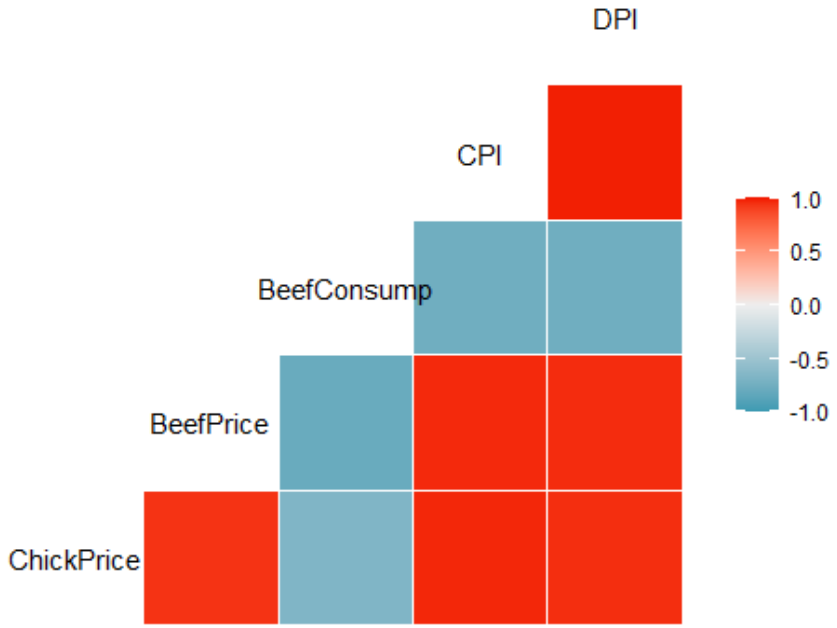
Построить подходящую модель линейной регрессии для спроса на говядину.

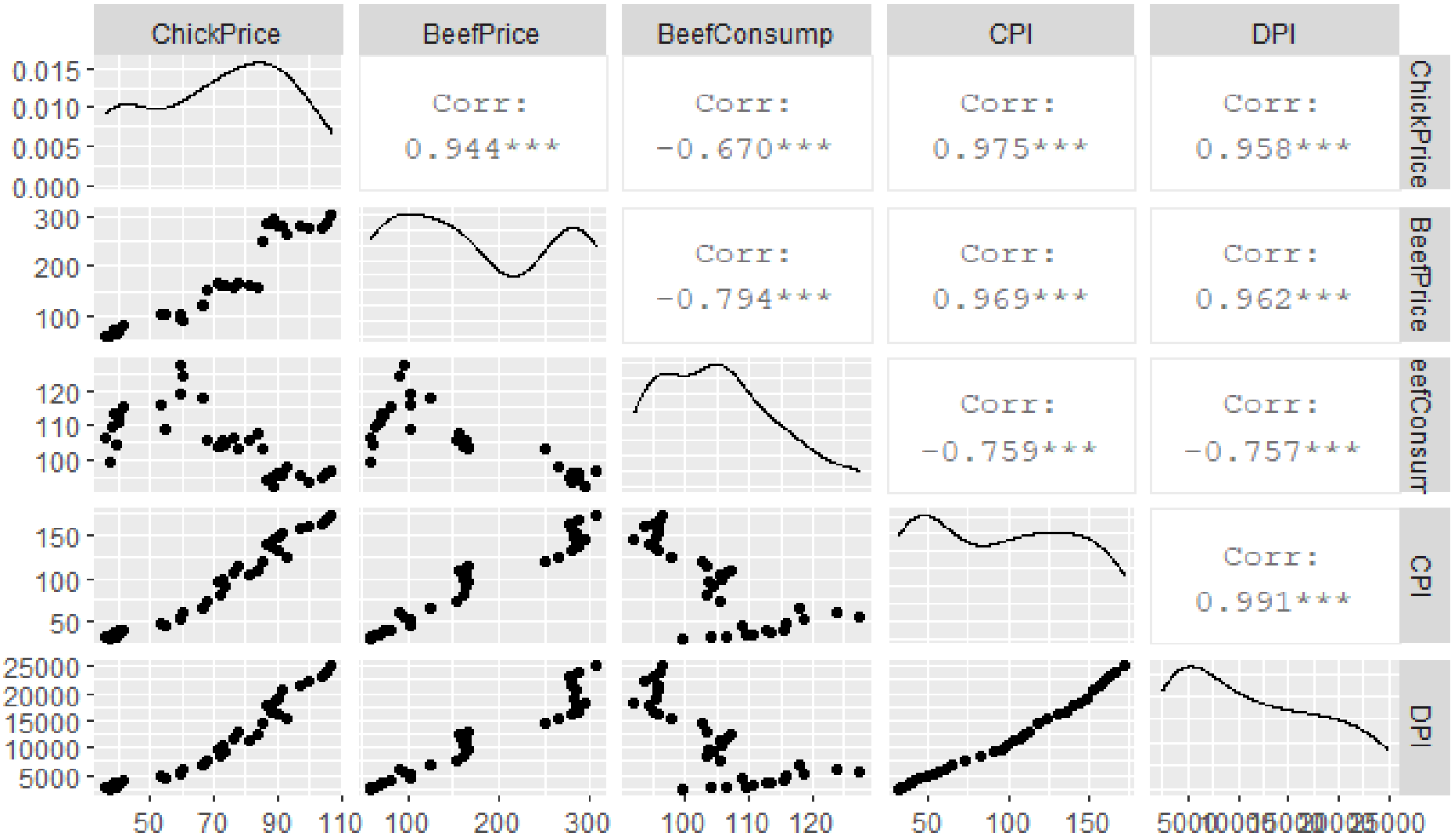
Отчет о построении регрессионной модели в R

Часть 1 (построение и анализ модели)

1. Провести корреляционный анализ имеющихся данных (используем функции `ggpairs`, `ggcorr`).

```
> View(beef)
> beef1<-beef[,-c(1,7,8,9,10)]
> ggcorr(beef1)
> ggpairs(beef1)
```




ChickPrice
BeefPrice
BeefConsump
CPI
DPI
Corr:
0.944***
Corr:
−0.670***
Corr:
0.975***
Corr:
0.958***
Corr:
0.969***
Corr:
0.962***
Corr:
−0.759***
Corr:
−0.757***
Corr:
0.991***
0.015
0.010
0.005
0.000
300
200
100
120
110
100
150
100
50
25000
20000
15000
10000
5000
50
70
90
110
100
200
300
100
110
120
50
100
150

2. Построить базовую модель линейной регрессии (функция `lm`).
3. Вывести результаты анализа базовой модели (функции `lm` и `summary`).

```
> beefc<-lm(beef1$BeefConsump ~., beef1)
> summary(beefc)

Call:
lm(formula = beef1$BeefConsump ~ ., data = beef1)

Residuals:
    Min      1Q  Median      3Q     Max
-11.901  -3.248   0.178   2.971  12.861

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.039e+02  5.280e+00  19.679  < 2e-16 ***
ChickPrice   5.907e-01  1.748e-01   3.380  0.00197 **
BeefPrice   -9.864e-02  3.863e-02  -2.553  0.01582 *
CPI         -3.146e-01  1.842e-01  -1.707  0.09774 .
DPI          5.101e-04  9.025e-04   0.565  0.57599
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.002 on 31 degrees of freedom
Multiple R-squared:  0.7312,    Adjusted R-squared:  0.6965
F-statistic: 21.08 on 4 and 31 DF,  p-value: 1.765e-08
```

4. Записать уравнение линейной регрессии (функция coef).

```
> coef(beefc)
  (Intercept)    ChickPrice     BeefPrice           CPI           DPI
1.038990e+02  5.907387e-01 -9.863606e-02 -3.145701e-01  5.100952e-04
```

$$BeefConsum = 103.9 + 0.59 * ChickPrice - 0.099 * BeefPrice - 0.31 * CPI + 0.00051 * DPI.$$

5. Проверить значимость каждого отдельного коэффициента с помощью T-test. P-values для данного критерия выводятся в Таблице summary в столбце $Pr(>|t|)$.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.039e+02  5.280e+00  19.679  < 2e-16 ***
ChickPrice   5.907e-01  1.748e-01   3.380  0.00197 **
BeefPrice   -9.864e-02  3.863e-02  -2.553  0.01582 *
CPI         -3.146e-01  1.842e-01  -1.707  0.09774 .
DPI          5.101e-04  9.025e-04   0.565  0.57599
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

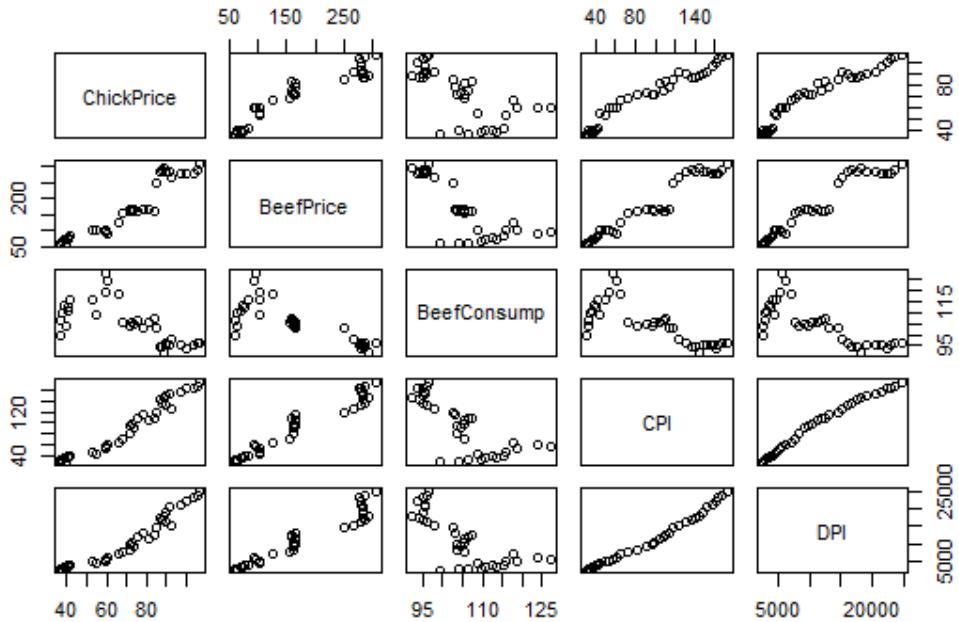
Коэффициенты при ChickPrice, BeefPrice значимы при уровне 0.05, коэффициент при CPI значим при уровне 0.1, коэффициент при DPI незначим.

6. Проверить значимость построенного уравнения регрессии с помощью F-test. Значение статистики теста и соответствующее p-value выводятся в результате работы функции `summary`.

```
Residual standard error: 5.002 on 31 degrees of freedom
Multiple R-squared:  0.7312,    Adjusted R-squared:  0.6965
F-statistic: 21.08 on 4 and 31 DF,  p-value: 1.765e-08
```

Построенное уравнение значимо в целом, т.к. p-value гораздо меньше 0.05. Хотя коэффициент R^2 не очень большой, равен 0.7312.

7. Построить график рассеяния и уравнения регрессии `pairs`.



8. Построить доверительные интервалы для коэффициентов регрессии с помощью функции `confint`.

```
> confint(beefc)
                    2.5 %        97.5 %
(Intercept) 93.131270262 114.666734908
ChickPrice   0.234325382   0.947152084
BeefPrice   -0.177430346  -0.019841765
CPI         -0.690320508   0.061180349
DPI         -0.001330487   0.002350677
```

9. Используя функцию `anova`, построить таблицу моделей линейных регрессии, которая включает статистику F, необходимую для оценки статистической значимости каждой парной линейной регрессионной модели.

```
> anova(beefc)
Analysis of Variance Table

Response: beef1$BeefConsump
           Df  Sum Sq Mean Sq F value    Pr(>F)
ChickPrice  1 1293.67 1293.67 51.7059 4.380e-08 ***
BeefPrice   1  692.38  692.38 27.6731 1.017e-05 ***
CPI         1  116.04  116.04  4.6379   0.03916 *
DPI         1    7.99    7.99  0.3195   0.57599
Residuals  31  775.61   25.02
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

10. В случае подозрения на наличие выбросов, проверить так называемые важные наблюдения, которые значительно влияют на построение модели (функция `influence.measures(m1)`).

    В результате как выбросы отмечаются наблюдения 1, 12, 36.

11. Используя функцию `step` или `stepAIC`, постараться улучшить модель.

```
> stepAIC(beefc)
Start:  AIC=120.52
beef1$BeefConsump ~ ChickPrice + BeefPrice + CPI + DPI

             Df Sum of Sq     RSS    AIC
- DPI         1     7.993  783.61 118.89
<none>                     775.61 120.53
- CPI         1    72.941  848.56 121.76
- BeefPrice   1   163.087  938.70 125.39
- ChickPrice  1   285.904 1061.52 129.82

Step:  AIC=118.89
beef1$BeefConsump ~ ChickPrice + BeefPrice + CPI

             Df Sum of Sq     RSS    AIC
<none>                     783.61 118.89
- CPI         1   116.04  899.65 121.86
- BeefPrice   1   159.37  942.98 123.56
- ChickPrice  1   279.67 1063.27 127.88
```

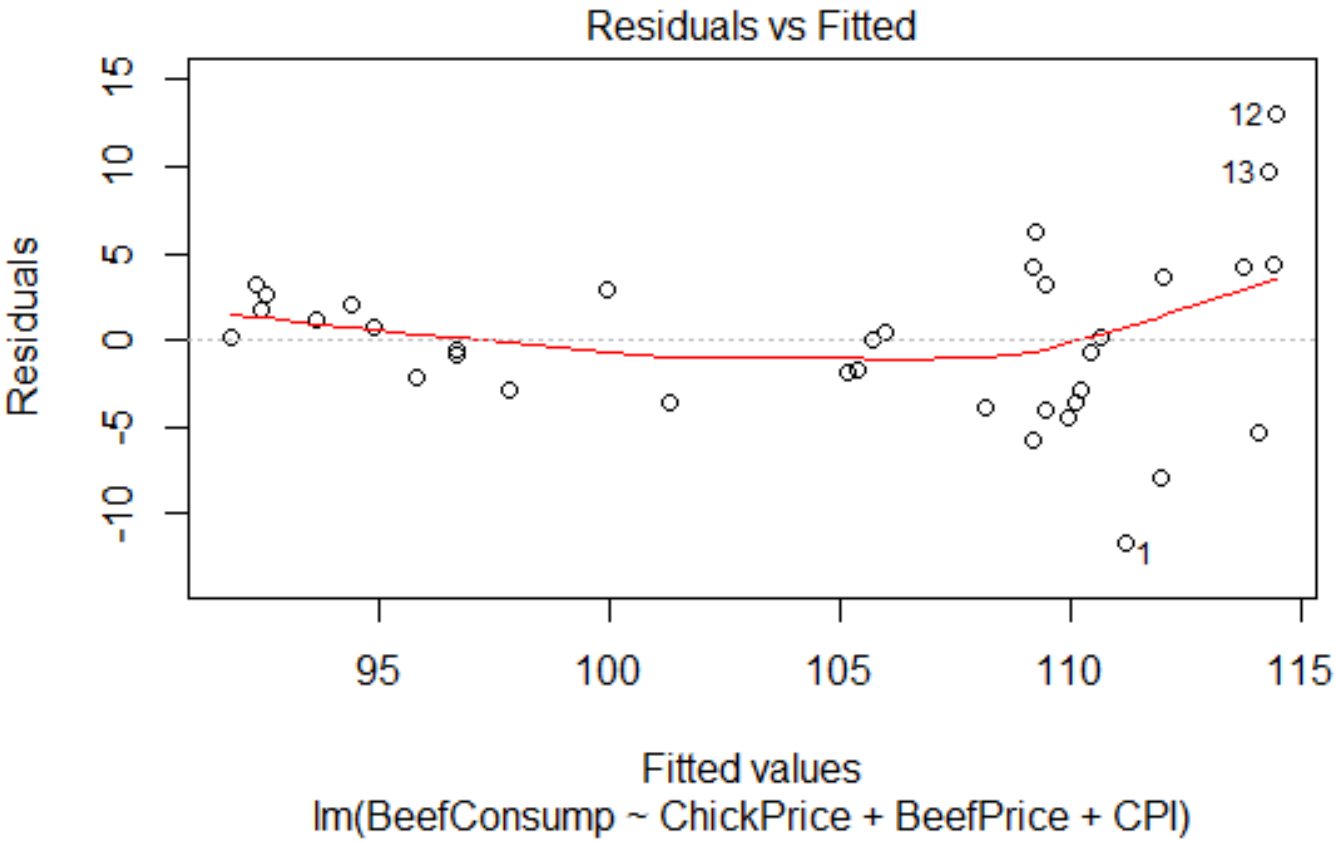
```
Call:
lm(formula = beef1$BeefConsump ~ ChickPrice + BeefPrice + CPI,
    data = beef1)

Coefficients:
(Intercept)   ChickPrice    BeefPrice          CPI
  103.10481      0.56672     -0.09733     -0.22965
```

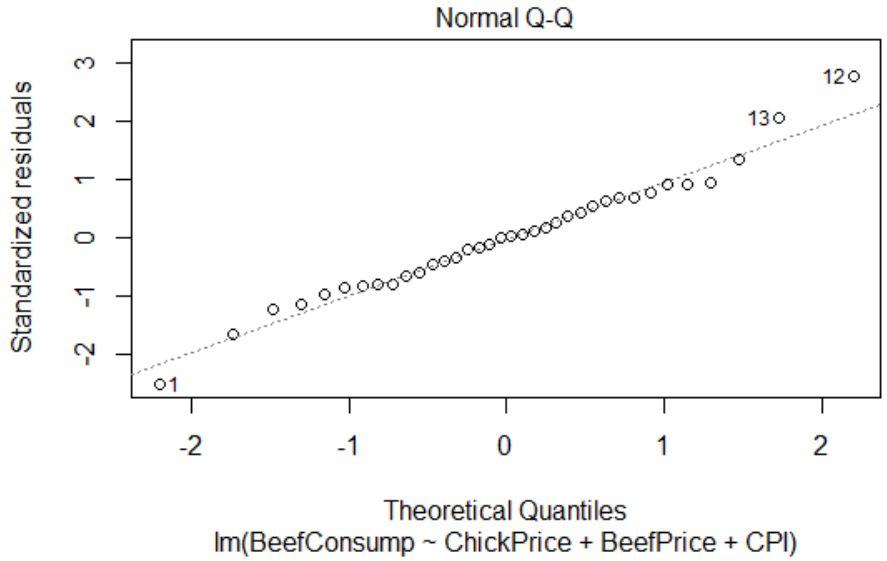
12. В случае получения в предыдущем пункте модели, отличной от базовой, повторить пп. 3–9 для новой модели.

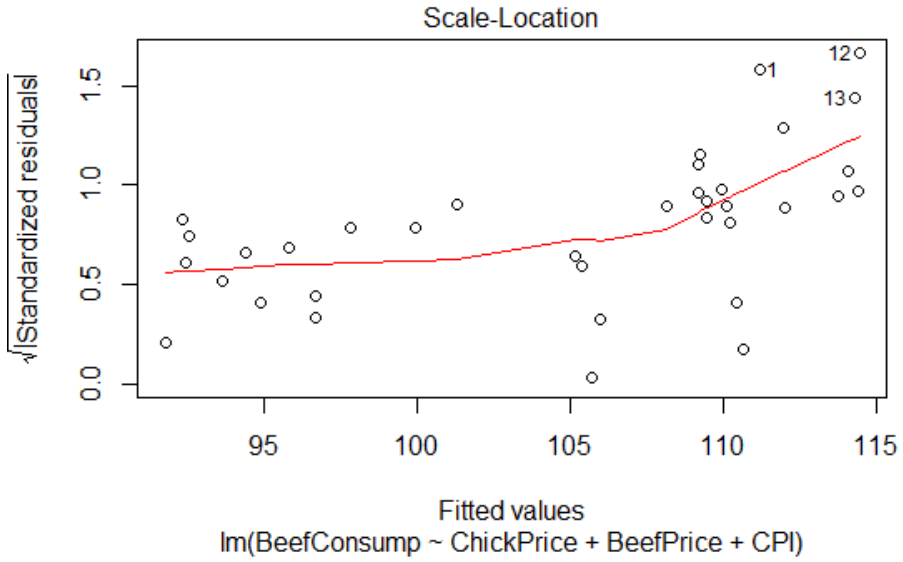
```
> beefc2<-lm(BeefConsump ~ ChickPrice + BeefPrice + CPI, beef1)
> summary(beefc2)

Call:
lm(formula = BeefConsump ~ ChickPrice + BeefPrice + CPI, data = beef1)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6870  -3.0986   0.0664   2.9247  13.0607

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.10481    5.03475  20.479  < 2e-16 ***
ChickPrice    0.56672    0.16770   3.379  0.00193 **
BeefPrice    -0.09733    0.03815  -2.551  0.01572 *
CPI          -0.22965    0.10550  -2.177  0.03698 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.949 on 32 degrees of freedom
Multiple R-squared:  0.7285,    Adjusted R-squared:  0.703
F-statistic: 28.61 on 3 and 32 DF,  p-value: 3.483e-09
```

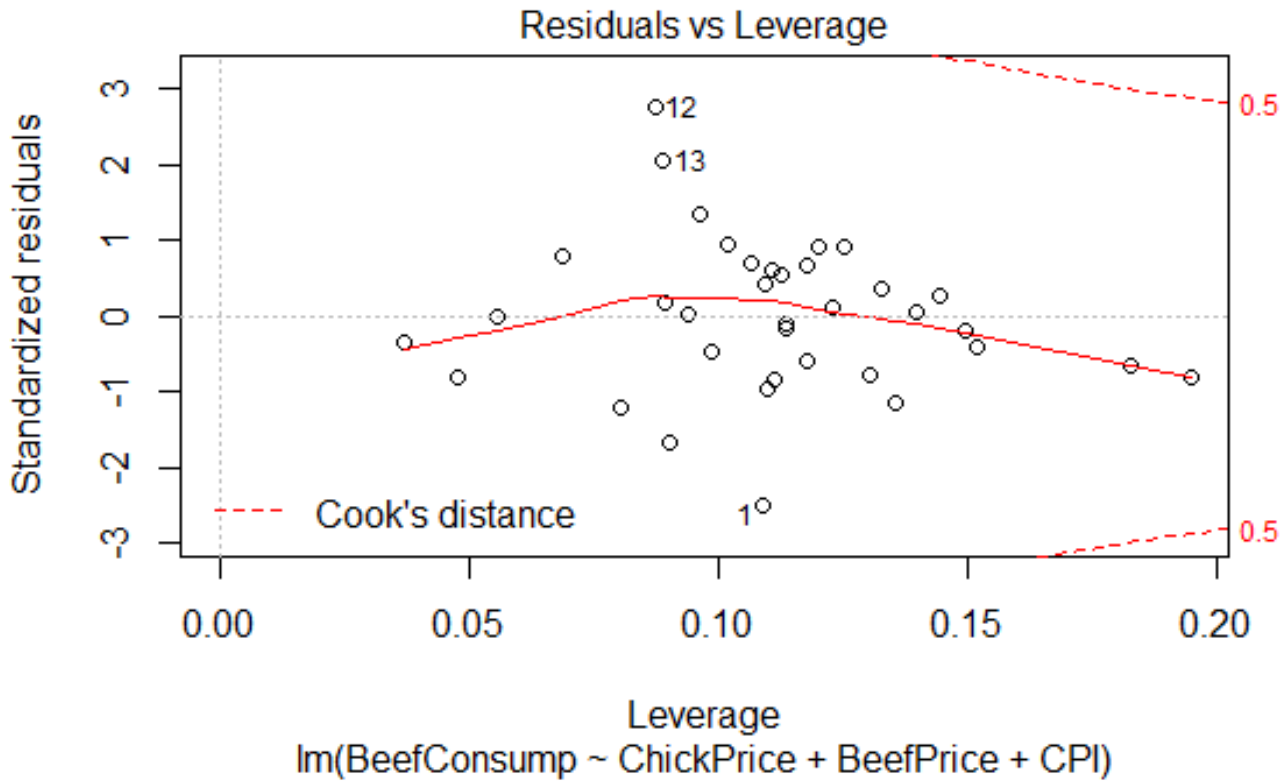
Часть 2 (диагностика модели, проверка предположений)

13. Построить графики: scatterplot, "Residuals vs Fitted" , "Normal Q-Q" , "Residuals vs Leverage" с помощью функции `plot(m1)` и дать интерпретации.




Normal Q-Q
Standardized residuals
-2
-1
0
1
2
3
12
13
1
-2
-1
0
1
2
Theoretical Quantiles
lm(BeefConsump ~ ChickPrice + BeefPrice + CPI)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
12
1
13
95
100
105
110
115
Fitted values
lm(BeefConsump ~ ChickPrice + BeefPrice + CPI)



14. Проверить модель на наличие выбросов с помощью функции `outlierTest(m1)`.

```
> outlierTest(beefc2)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferonni p
12 3.115939          0.0039333       0.1416
```

15. Проверить модель на гетероскедастичность с помощью функции `ncvTest(m1)` или `bptest(m1)`.

```
> ncvTest(beefc2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 11.58252, Df = 1, p = 0.00066575
> bptest(beefc2)

        studentized Breusch-Pagan test

data:  beefc2
BP = 8.113, df = 3, p-value = 0.04373
```

Наблюдаем проблему гетероскедастичности.

16. Проверить остатки модели на автокорреляцию с помощью функции `durbinWatsonTest(m1)` или `dwtest(m1)`.

```
> durbinWatsonTest(beefc2)
 lag Autocorrelation D-W Statistic p-value
   1       0.4975638     0.8253531       0
 Alternative hypothesis: rho != 0
> dwtest(beefc2)

        Durbin-Watson test

data:  beefc2
DW = 0.82535, p-value = 4.418e-06
alternative hypothesis: true autocorrelation is greater than 0
```

Автокорреляция есть.

17. Проверить остатки модели на нормальность распределения с помощью функции `ols_test_normality(m1)`.

```
> ols_test_normality(beefc2)
-----------------------------------------------
       Test             Statistic       pvalue
-----------------------------------------------
Shapiro-Wilk              0.9789         0.7083
Kolmogorov-Smirnov        0.0921         0.8928
Cramer-von Mises          2.5062         0.0000
Anderson-Darling          0.2728         0.6480
-----------------------------------------------
> jarque.bera.test(resbeef)

        Jarque Bera Test

data:  resbeef
X-squared = 1.7112, df = 2, p-value = 0.425
```

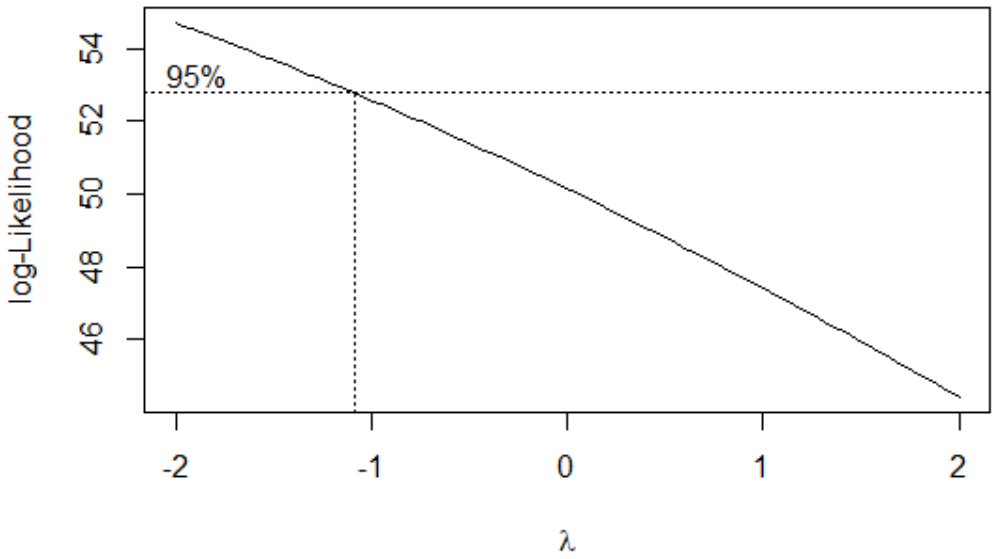
Подтверждаем нормальное распределение остатков по 4 из 5 тестов.

18. Проверить модель на мультиколлинеарность данных с помощью функции `vif(m1)`.

```
> vif(beefc2)
ChickPrice  BeefPrice        CPI
  19.93777   16.33408   35.40881
```

Существует проблема мультиколлинеарности данных.

19. Попробовать применить трансформацию Box-Cox зависимой переменной.




```
> beef_bc <- boxcox(beefc2)
> which.max(beef_bc$y)
[1] 1
> lambda <- beef_bc$x[which.max(beef_bc$y)]
> lambda
[1] -2
> z_beef <- beef1$BeefConsump^lambda
> beef2<-beef1[,-3]
> beef2$BeefConsump <- z_beef
> beefc3 <- lm(beef2$BeefConsump~., data=beef2)
```

```
> summary(beefc3)

Call:
lm(formula = beef2$BeefConsump ~ ., data = beef2)

Residuals:
       Min         1Q     Median         3Q        Max
-1.464e-05 -5.030e-06 -7.680e-08  4.742e-06  1.974e-05

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.379e-05  7.594e-06  12.351 1.65e-13 ***
ChickPrice  -9.479e-07  2.514e-07  -3.771 0.000688 ***
BeefPrice    1.795e-07  5.557e-08   3.231 0.002921 **
CPI          4.030e-07  2.650e-07   1.521 0.138489
DPI         -2.404e-10  1.298e-09  -0.185 0.854257
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.195e-06 on 31 degrees of freedom
Multiple R-squared:  0.8015,    Adjusted R-squared:  0.7759
F-statistic:  31.3 on 4 and 31 DF,  p-value: 1.745e-10
```

20. В случае получения новой модели в предыдущем пункте проанализировать новую модель.
    - Значимыми являются коэффициенты при ChickPrice, BeefPrice.
    - Уравнение регрессии значимо в целом.
    - Коэффициент R^2 значительно увеличился 0.8015 против 0.7285.
    - Анализ графиков `plot(beefc3)` показывает нормальное распределение остатков и отсутствие выбросов.

Итоги

Что мы сделали на Лекции 7?

- Мы разобрали пример построения регрессионной модели на примере базы данных о спросе на говядину.
- Мы проделали все вычисления в R.
- Мы составили отчет о проделанной аналитической работе.

Что мы узнаем на Лекции 8?

Мы узнаем,

- что такое бинарная регрессия.
- почему модель бинарной регрессии используется как классификатор.
- как проверять значимость построенной модели бинарной регрессии.
- как можно улучшить модель бинарной регрессии.

Спасибо за внимание и до встречи на Лекции 8!