Домашняя работа №2

Моделирование статистических зависимостей

Выполнил: Миронюк Даниил Никитич

Группа: ПСА-2

2019 г.

Задание 1

1)Функция спроса

```
> summary(m1<-lm(log(Y)~log(X2),data=df))

Call:
lm(formula = log(Y) ~ log(X2), data = df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.137580 -0.049196  0.008714  0.058689  0.120432

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.57772    0.31051   5.081 9.25e-05 ***
log(X2)      0.55272    0.07989   6.918 2.49e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07812 on 17 degrees of freedom
Multiple R-squared:  0.7379,    Adjusted R-squared:  0.7225
F-statistic: 47.86 on 1 and 17 DF,  p-value: 2.485e-06
```

Из данной модели следует, что с ростом цены цыплят на 1% спрос на них увеличится в среднем на 0,553% Что, конечно, не может быть истинно. Дело в том что данная модель обучалась на данных за 20 лет, и не учитывает рост рынка, рост спроса а так же возможную инфляцию.

2)Функция потребления

```
> summary(m2<-lm(log(Y)~log(X1),data=df))

Call:
lm(formula = log(Y) ~ log(X1), data = df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.072152 -0.020452  0.003223  0.025825  0.059829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.74599    0.12292   14.20 7.33e-11 ***
log(X1)      0.28492    0.01768   16.12 9.84e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03781 on 17 degrees of freedom
Multiple R-squared:  0.9386,    Adjusted R-squared:  0.935
F-statistic: 259.8 on 1 and 17 DF,  p-value: 9.837e-12
```

Из данной модели следует, что с ростом дохода населения на 1% спрос на них увеличится в среднем на 0,285% . Данная модель не учитывает важные факторы цены на товар и цены на товары конкуренты, но имеет хорошее значение скорректированного R квадрата.

3)Функция спроса-потребления

```
> summary(m3<-lm(log(Y)~log(X1)+log(X2),data=df))

Call:
lm(formula = log(Y) ~ log(X1) + log(X2), data = df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.055142 -0.018277 -0.005682  0.022947  0.051872

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.01848    0.12665  15.937 3.07e-11 ***
log(X1)      0.41614    0.04157  10.011 2.70e-08 ***
log(X2)     -0.30482    0.09094  -3.352  0.00405 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02988 on 16 degrees of freedom
Multiple R-squared:  0.9639,    Adjusted R-squared:  0.9594
F-statistic: 213.7 on 2 and 16 DF,  p-value: 2.872e-12
```

По данной модели увеличение среднедушевого дохода на 1% при неизменной цене потребления цыплят вырастет на 0.416% в то же время при фиксированном среднедушевом доходе населения увеличение цены на 1% приведет к снижению спроса на 0.304%

4) Функция спроса с учетом цены на товары-заменители

```
> summary(m4<-lm(log(Y)~log(X2)+log(X3)+log(X4),data=df))

Call:
lm(formula = log(Y) ~ log(X2) + log(X3) + log(X4), data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.06833 -0.02295  0.01050  0.02525  0.04585

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.31519    0.22155  10.450 2.79e-08 ***
log(X2)     -0.48745    0.21100  -2.310   0.0355 *
log(X3)      0.23742    0.15562   1.526   0.1479
log(X4)      0.46005    0.07527   6.112 1.99e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03772 on 15 degrees of freedom
Multiple R-squared:  0.9461,    Adjusted R-squared:  0.9353
F-statistic: 87.74 on 3 and 15 DF,  p-value: 9.73e-10
```

Из модели следует, что при неизменной стоимости двух сопутствующих продуктов увеличение на 1% стоимости цыплят приводит к снижению их потребления в среднем на 0,48%, а увеличение стоимости свинины или говядины на 1% при неизменности цен на остальные входящие в модель продукты приводит к росту потребления цыплят в среднем соответственно на 0,237 и 0,460%.

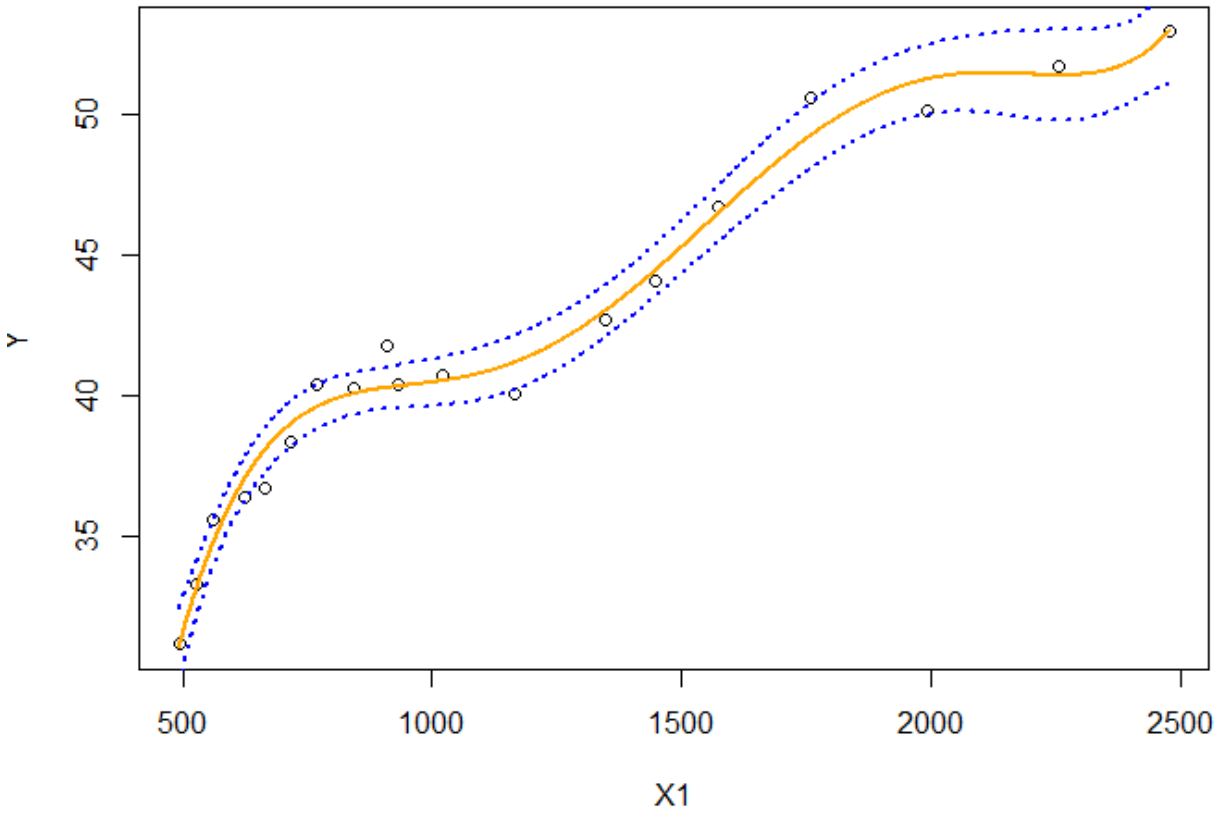
При сравнении моделей мы замечаем, что самое большое значение скорректированного R^2

У 3 модели. Так же все её коэффициенты имеют высокий уровень значимости. Однако все модели с 2-4 приемлемы для использования с определенной точностью. 1 модель является полностью некорректной по причине отсутствия в модели важных факторов.

Задание 2

В ходе данного задания я протестировал модель регрессии с полиномом степени от 2 до 7.С увеличением степени полинома мы получаем все больший прирост скорректированного R^2 данный эффект происходит за счет того, что модель все сильнее подгоняется к точкам наблюдений. Это хорошо лишь до определенного момента, когда эффект выявления закона зависимости сменится переобучением модели и простым подгоном к точкам наблюдений.

Я эмпирически выбрал модель с полиномом 5 степени



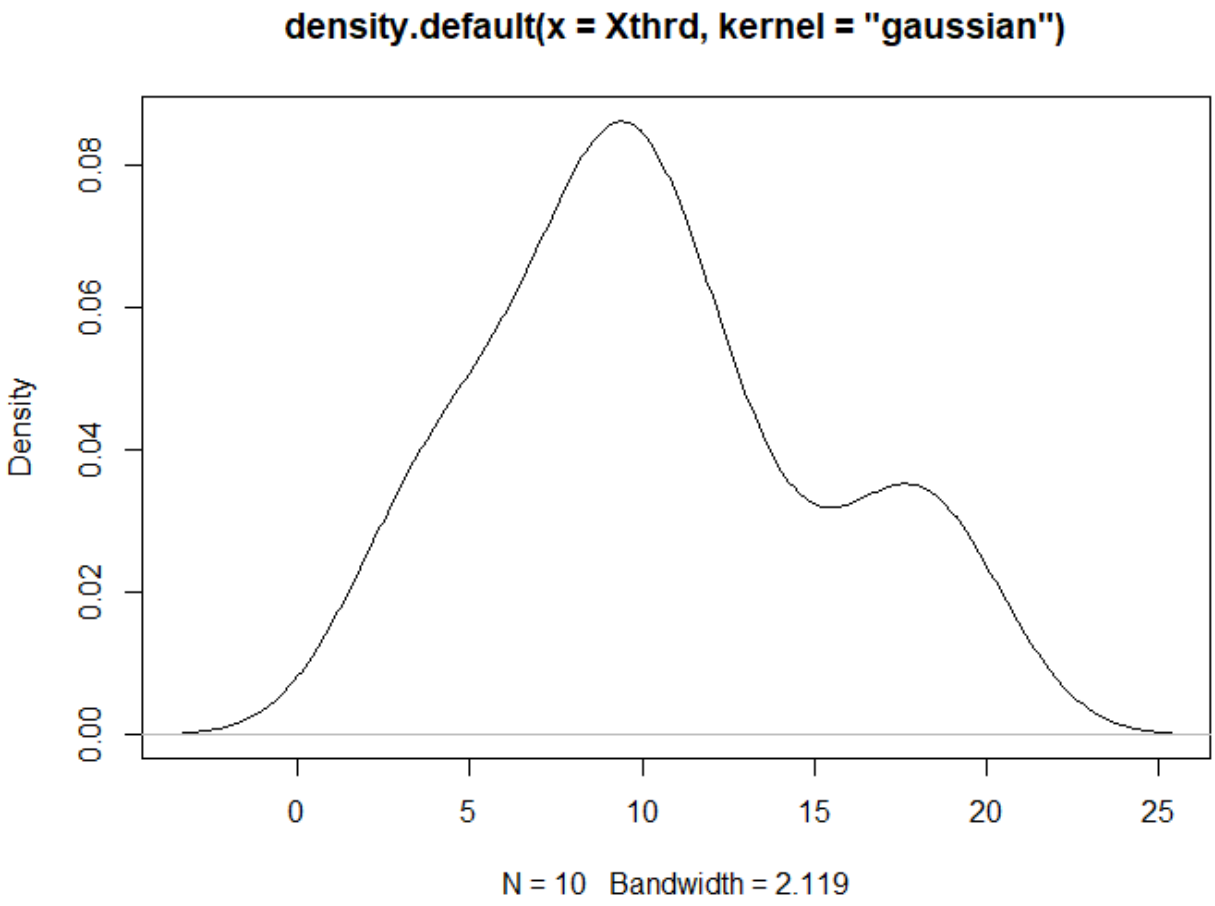
По графику видно, что мы не только выявили основные тенденции, но и отобразили локальные изгибы. Конечно, для проверки нам нужно больше данных и тогда мы сможем проверить верно ли мы выявили закон или же необходимо было слабее подгоняться к точкам.

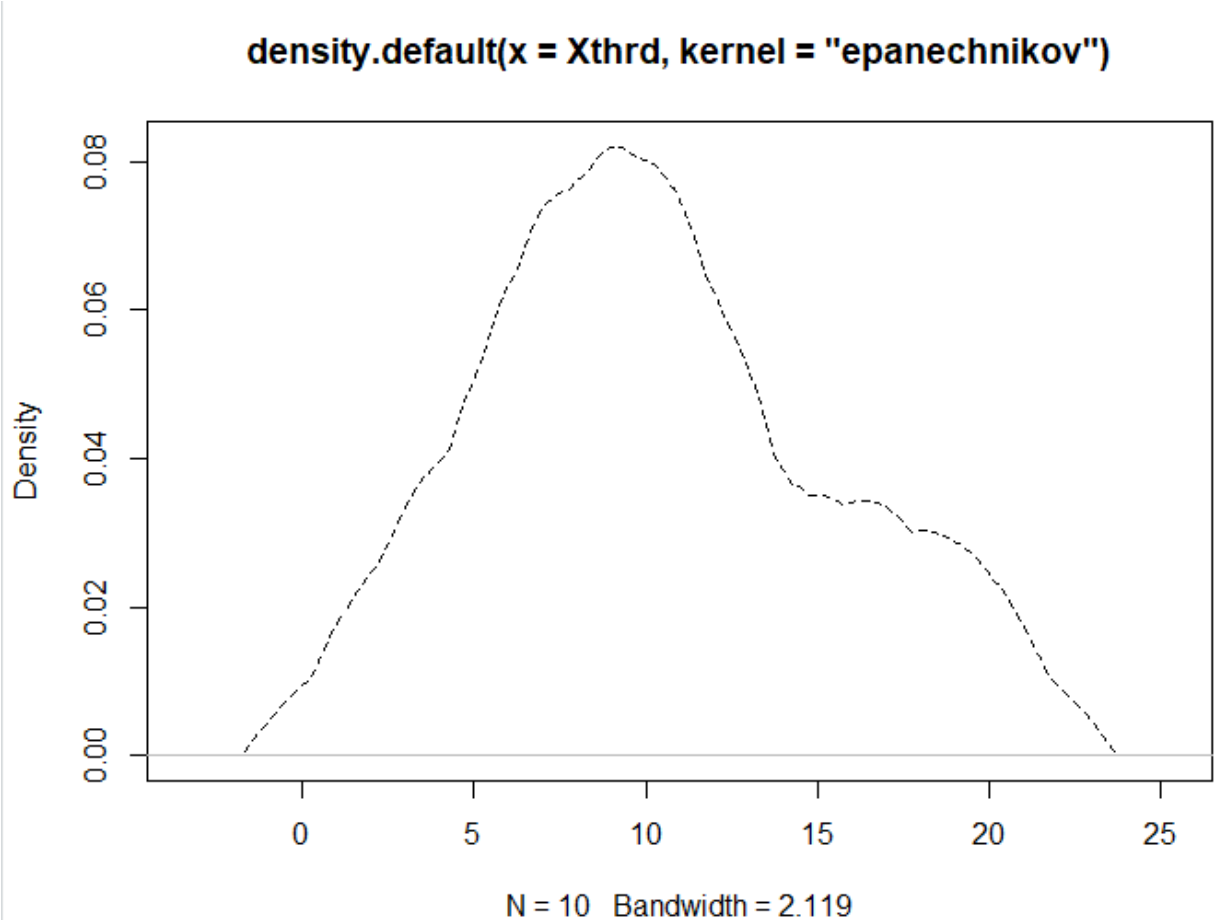
На основе ANOVA теста можно сделать вывод что первые шаги в увеличении степени полинома дают самый большой прирост объяснённой дисперсии.

```
> summary(aov(m5))
                      Df Sum Sq Mean Sq F value   Pr(>F)
poly(X1, 2, raw = TRUE)  2  654.9   327.4   122.2 2.03e-10 ***
Residuals               16   42.9     2.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov(m6))
                      Df Sum Sq Mean Sq F value   Pr(>F)
poly(X1, 3, raw = TRUE)  3  657.1  219.03   80.87 1.73e-09 ***
Residuals               15   40.6    2.71
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov(m7))
                      Df Sum Sq Mean Sq F value   Pr(>F)
poly(X1, 4, raw = TRUE)  4  672.5   168.1   93.25 6.28e-10 ***
Residuals               14   25.2     1.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov(m8))
                      Df Sum Sq Mean Sq F value   Pr(>F)
poly(X1, 5, raw = TRUE)  5  686.4  137.27   157.2 3.76e-11 ***
Residuals               13   11.4    0.87
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov(m9))
                      Df Sum Sq Mean Sq F value   Pr(>F)
poly(X1, 7, raw = TRUE)  7  690.9   98.69   158.5 3.72e-10 ***
Residuals               11    6.8    0.62
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Задание 3

В данном задании я оценил функцию плотности распределения с помощью ядерного сглаживания. В целом это называется ядерной оценкой плотности вероятности.


density.default(x = Xthrd, kernel = "gaussian")
Density
0.00
0.02
0.04
0.06
0.08
0
5
10
15
20
25
N = 10 Bandwidth = 2.119


density.default(x = Xthrd, kernel = "epanechnikov")
Density
0.00
0.02
0.04
0.06
0.08
0
5
10
15
20
25
N = 10 Bandwidth = 2.119

Визуально отметим, что пики функции плотностей соответствуют 9 и 17. Судья по выборке, которую мы использовали так и должно быть. Плотность вероятности в точке 5 по нашим графикам будет равна 0.049 и 0.05, а в точке 10 – 0.08 и 0.083 (для ядра Епанеч. И Гауссова ядра соответственно)

Задание 4

```
> cor(a)
              Lag1         Lag2        Lag3         Lag4         Lag5      volume
Lag1    1.000000000 -0.026294328 -0.01080340 -0.002985911 -0.005674606  0.04090991
Lag2   -0.026294328  1.000000000 -0.02589667 -0.010853533 -0.003557949 -0.04338321
Lag3   -0.010803402 -0.025896670  1.00000000 -0.024051036 -0.018808338 -0.04182369
Lag4   -0.002985911 -0.010853533 -0.02405104  1.000000000 -0.027083641 -0.04841425
Lag5   -0.005674606 -0.003557949 -0.01880834 -0.027083641  1.000000000 -0.02200231
volume  0.040909908 -0.043383215 -0.04182369 -0.048414246 -0.022002315  1.00000000
```

В матрице парных коэффициентов нет больших по модулю значений, это говорит о том, что рынок движется чаще всего непредсказуемо и независимо от прошлых по времени значений курса.

```
> summary(glm.fit)

Call:
glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
    Volume, family = binomial, data = Smarket)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.446  -1.203   1.065   1.145   1.326

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.126000   0.240736  -0.523    0.601
Lag1        -0.073074   0.050167  -1.457    0.145
Lag2        -0.042301   0.050086  -0.845    0.398
Lag3         0.011085   0.049939   0.222    0.824
Lag4         0.009359   0.049974   0.187    0.851
Lag5         0.010313   0.049511   0.208    0.835
Volume       0.135441   0.158360   0.855    0.392

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1731.2  on 1249  degrees of freedom
Residual deviance: 1727.6  on 1243  degrees of freedom
AIC: 1741.6

Number of Fisher Scoring iterations: 3
```

Полученная нами модель бинарного выбора имеет не значимые коэффициенты что может говорить о несостоятельности модели. По модели считается что объем торгов в предыдущий день, а также доходность относительно текущего за 3ий, 4ый и 5ый дни соответственно. Положительным образом влияют на вероятность того, что акция будет расти, а доходность относительно текущего за 1ый и 2ой дни отрицательно влияет на вероятность роста акции.

Возможно, это вызвано тем, что некоторые биржевые стратегии предполагают быстрое закрытие позиции при наличии положительной доходности.

```
> df4$Pred2 = predict(glm.fit, type="response")
> df4$PredCat2 = cut(df4$Pred2, c(0,0.5,1), include.lowest=TRUE, labels=c("Down","U
p"))
> sum(df4$PredCat2 == df4$Direction)/1250
[1] 0.5216
```

Как можно видеть наша модель верно предсказала результат в 52% случаев. Данная точность является неприемлемой. С помощи нашей модели бинарного выбора нельзя предсказать движение рынка.

Задание 5

```
> pcr_model <- pcr(hdi~sub1+sub2+sub3+sub4, data = hdi)
> summary(pcr_model)
Data:   X dimension: 113 4
        Y dimension: 113 1
Fit method: svdpc
Number of components considered: 4
TRAINING: % variance explained
     1 comps  2 comps  3 comps  4 comps
X      80.42    90.01    98.74      100
hdi    98.82    98.84    99.94      100
```

Видно, что модель успешно выделила главные компоненты.

Вынесем ее предсказанные значения отдельным столбиков в датафрейм и сравним с истинными значениями.

	id	hdi	sub1	sub2	sub3	sub4	Preds
1	1	0.9940529	0.998	0.999	0.999	0.990	0.9940529
2	2	0.9890187	0.999	0.980	0.998	0.998	0.9890187
3	3	0.9911061	0.994	0.992	0.993	0.995	0.9911061
4	4	0.9915992	0.991	0.999	0.990	0.990	0.9915992
5	5	0.9920419	0.989	1.000	0.983	0.997	0.9920419
6	6	0.9896459	0.991	0.990	0.991	0.995	0.9896459
7	7	0.9906729	0.977	0.997	0.997	0.993	0.9906729
8	8	0.9899946	0.987	0.997	0.984	0.993	0.9899946
9	9	0.9869671	0.989	0.984	0.996	0.990	0.9869671
10	10	0.9882566	0.982	0.990	0.993	0.993	0.9882566
11	11	0.9879016	0.988	0.991	0.982	0.994	0.9879016
12	12	0.9848095	0.987	0.981	0.991	0.990	0.9848095
13	13	0.9858584	0.995	0.995	0.993	0.965	0.9858584
14	14	0.9874710	0.984	0.997	0.987	0.981	0.9874710
15	15	0.9836428	0.997	0.979	0.990	0.982	0.9836428
16	16	0.9850082	0.985	0.986	0.995	0.981	0.9850082
17	17	0.9850027	0.985	0.986	0.983	0.990	0.9850027
18	18	0.9802060	0.997	0.969	0.992	0.981	0.9802060
19	19	0.9851865	0.996	1.000	0.981	0.962	0.9851865
20	20	0.9866841	0.969	0.999	0.979	0.990	0.9866841
21	21	0.9841944	0.984	0.992	0.991	0.971	0.9841944
22	22	0.9865973	0.967	1.000	0.990	0.981	0.9865973
23	23	0.9857417	0.968	0.999	0.994	0.975	0.9857417
24	24	0.9840671	0.968	0.987	0.990	0.990	0.9840671
25	25	0.9850325	0.978	0.998	0.983	0.975	0.9850325
26	26	0.9828576	0.972	0.997	0.989	0.966	0.9828576
27	27	0.9831145	0.989	0.997	0.957	0.980	0.9831145
28	28	0.9828084	0.965	0.998	0.994	0.965	0.9828084
29	29	0.9844081	0.946	1.000	0.983	0.990	0.9844081
30	30	0.9811343	0.984	0.999	0.986	0.949	0.9811343
31	31	0.9834248	0.943	0.996	0.987	0.991	0.9834248
32	32	0.9792885	0.971	0.979	0.982	0.985	0.9792885
33	33	0.9705713	0.988	0.945	0.981	0.990	0.9705713

Из результатов модели следует что ей удается самостоятельно составлять индекс HDI.