Домашнее задание 3, Часть 2

Переверзев Виктор, Низоля Валерия, группа БЭК173

Бейзлайн для выбора регрессионной модели для безработицы в России в 2005 году

1. Найти дескриптивные статистики (max, min, mean, sd)

2. Создать переменные

3. Построить гистограммы для переменных

4. Диаграммы рассеяния. Похожа ли на линейную?

5. Значимость стартовой простой линейной модели со всеми созданными переменными

6. Работа с выбросами (использовать дамми для выбросов или исключить их совсем? Хубер?)

7. Проверить тест Чоу

8. Выбрать функциональную модель (логарифмическая, линейная или полулогарифмическая)

9. Бокс-Кокс (линейная и логарифмическая, линейная и полулогарифмическая)/Бера-МакАлера

10. С помощью R2adj сравнить логарифмическую или полулогарифмическую, если в предыдущем пункте выбрана не линейная модель

11. Тест Рамсея на спецификацию

12. Считаем VIF – проверяем мультиколлинеарность

13. Метод исключения в случае мультколлинеарности

14. Тест Бройша-Пагана на гетероскедастичность -> коррекция в случае необходимости

15. Qqnorm для остатков

16. Тест Шапиро-Уилка на нормальность остатков

17. Определить лучшую модель и проанализировать её с точки зрения статистики и экономики.

Чтобы построить правдоподобную экономическую модель, мы добавили ряд новых переменных в базу данных:

1) enter – количество предприятий в регионе
2) pop – рождаемость на 1000 населения
3) repub – национальные республики, дамми-переменная
4) west_mult_urb – произведение доли городского населения на дамми-переменную западных регионов.
5) west_mult_gdp – произведение ВРП региона на дамми-переменную западных регионов.

Также, мы использовали переменные из первоначального датасета:

1) unemp – зависимая переменная, доля безработицы, которую необходимо предсказать
2) gdp – валовый региональный продукт по паритету покупательской способности
3) urb – доля городского населения в регионе
4) educ – доля населения с высшим образованием
5) west – западные регионы (дамми-переменная)

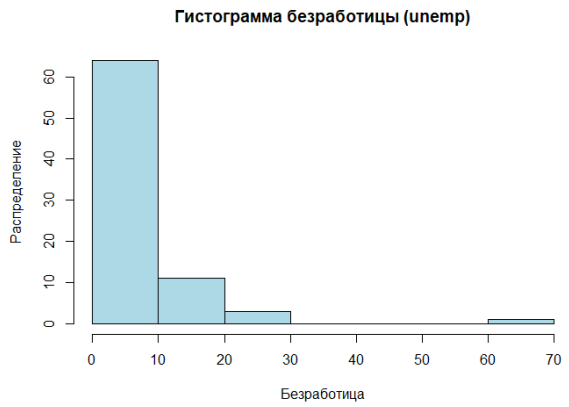
Проверим, правильно ли загрузились данные в R:

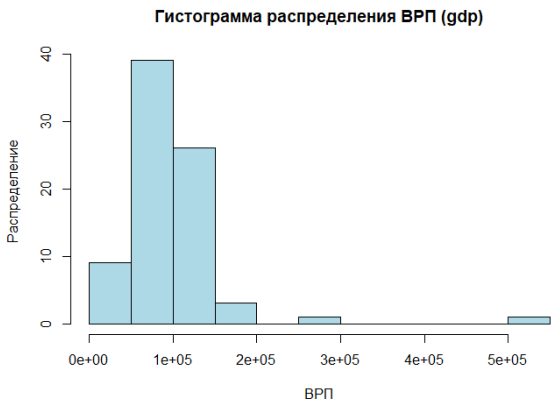
```
   GDPpercapppp Urbanshare Higheduc Enterprises

1      111926.       66.1     19.1     25857
2       58716.       68.1     29.6     20209
3       67449.       77.5     17.8     27868
4       63684.       62.7     25.3     55317
5       46653.       80.6     19.2     29121
6       78708.       75.8     22.1     27880

  PopulationGrowth NationalRepublics WEST Unemployment

1      _8.9                  0        1          6
2       9                    0        1          6.7
3       9.2                  0        1          9
4       8.4                  0        1          7.5
5       8.7                  0        1          6.8
6       8.9                  0        1          5.7
```

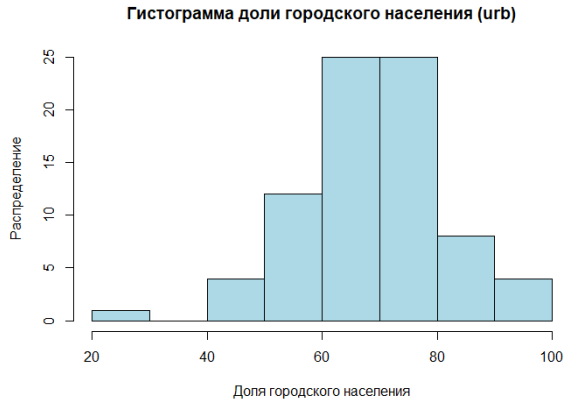
Чтобы лучше ориентироваться в данных, рассмотрим дескриптивные статистики для переменных из датасета:

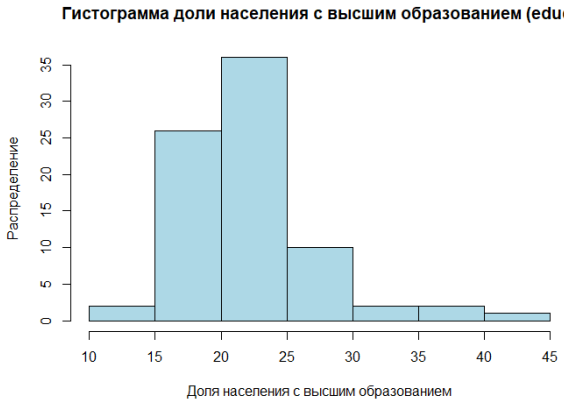
```
                     vars  n     mean        sd    min       max
GDPpercapppp            1 79 96064.68  63566.62 16840.43  545447.0
Urbanshare              2 79    68.86     12.66    26.00     100.0
Higheduc                3 79    22.08      5.00    14.30      43.8
Enterprises             5 79 60125.63 142278.91  1881.00 1221514.0
PopulationGrowth        6 79    10.51      2.10     7.80      19.4
NationalRepublics       7 79     0.25      0.44     0.00       1.0
WEST                    8 79     0.68      0.47     0.00       1.0
Unemployment            9 79     9.08      7.42     0.80      64.9
```

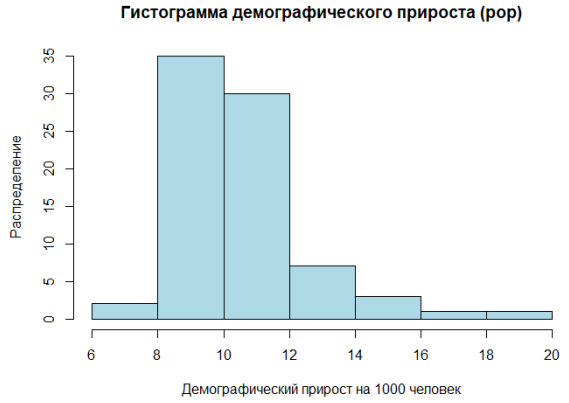
Теперь визуализируем их распределение с помощью гистограмм. Гистограммы для дамми-переменных не принесут новой информации после представленных дескриптивных статистик (выше уже понятно, к примеру, западных или восточных регионов больше), поэтому не будем их изображать):

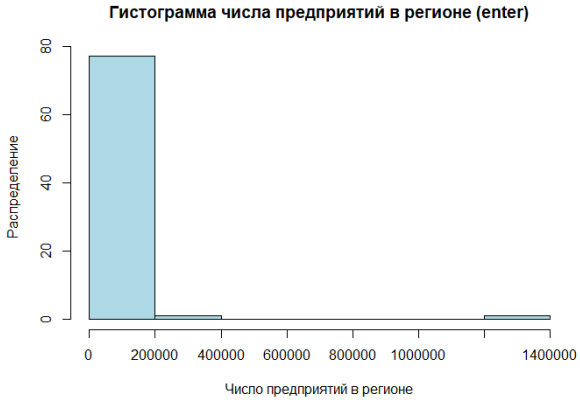






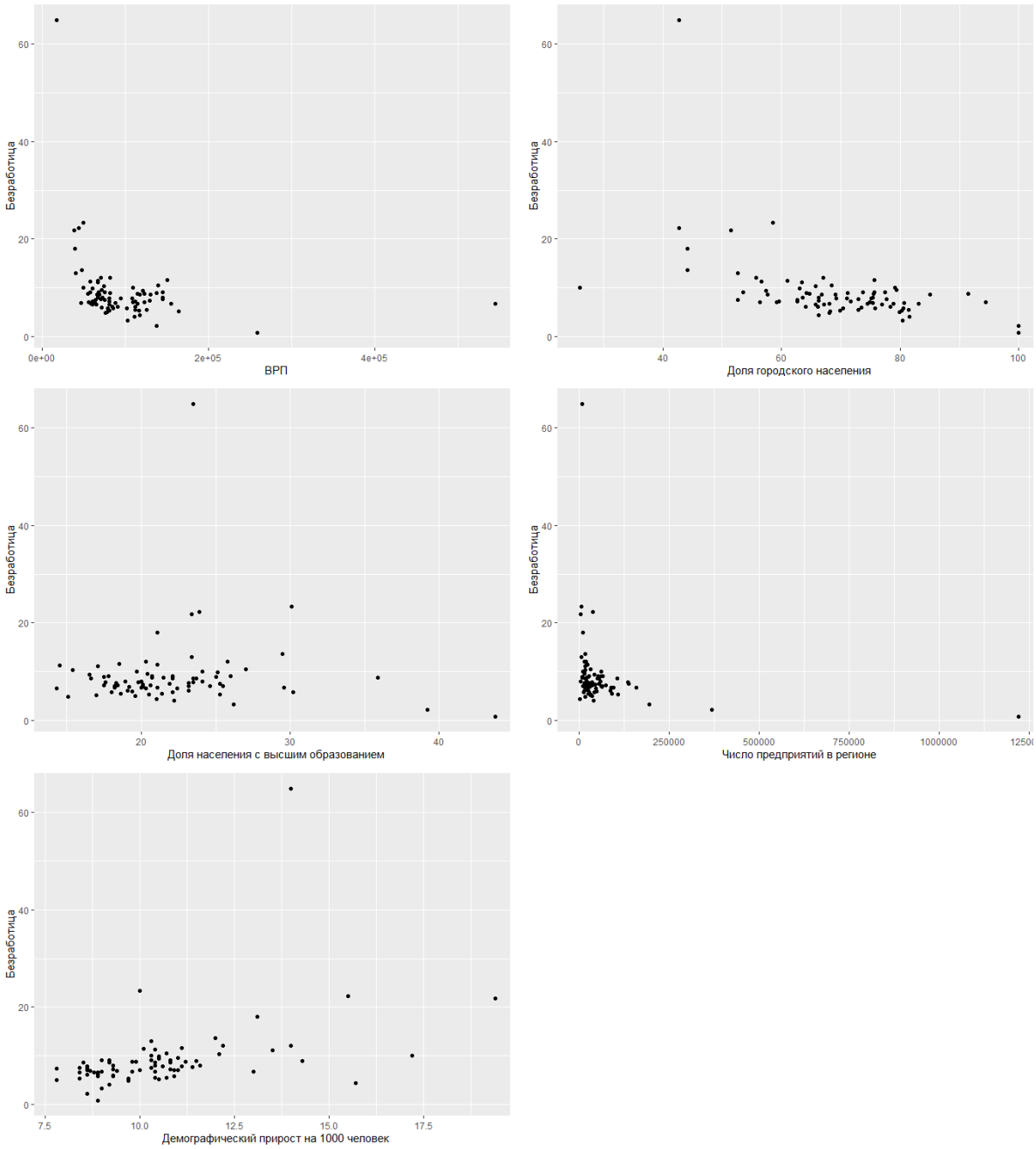






Трудно сказать, что данные по переменным принадлежат нормальному распределению, кроме, может быть, доли городского населения.

Посмотрим на зависимость целевой переменной от регрессоров иначе – с помощью диаграмм рассеяния:



Диаграммы рассеяния явно дают понять, что в данных присутствуют выбросы, которые портят статистику. Если не обращать внимания на выбросы, то доля городского населения и демографический прирост наиболее похожи на нормальное распределение.

Немного познакомившись с данными, перейдём к анализу модели. Составляя бейзлайн работы, мы приняли решение начать с простой модели, включающей в себя все выделенные переменные – *стартовой модели* – и постепенно преобразовывая её, прийти к модели значительно лучшего качества.

Оценим *стартовую модель*:

```
Call:
lm(formula = unemp ~ gdp + urb + educ + west + pop + repub + 
    enter + west_mult_urb + west_mult_gdp, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-9.211 -2.084 -0.294  1.090 38.843 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)   
(Intercept)   -1.678e+01  1.150e+01  -1.459  0.14907   
gdp           -1.307e-05  1.322e-05  -0.988  0.32645   
urb            1.141e-01  1.061e-01   1.076  0.28584   
educ          -3.817e-02  1.835e-01  -0.208  0.83579   
west           2.740e+01  8.307e+00   3.298  0.00154 **
pop            1.601e+00  5.220e-01   3.068  0.00308 **
repub          2.113e+00  2.067e+00   1.023  0.31008   
enter          9.397e-06  7.865e-06   1.195  0.23626   
west_mult_urb -2.803e-01  1.240e-01  -2.261  0.02693 * 
west_mult_gdp -5.670e-05  3.309e-05  -1.714  0.09106 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.788 on 69 degrees of freedom
Multiple R-squared:  0.4615,	Adjusted R-squared:  0.3912 
F-statistic:  6.57 on 9 and 69 DF,  p-value: 1.033e-06
```

R^2_{adj} небольшой, однако нельзя сказать, что модель совсем плоха: ей объясняется почти 40%, что уже достойно. West значим на уровне 1%: если регион западный, то безработица больше на 2.740e+01. Pop значим на уровне 1%, значит, с его увеличением на 1 единицу безработица увеличивается на 1.601e+00. West_mult_urb и west_mult_gdp значимы на уровнях 5% и 10% соответственно, что говорит о том, что на западе доля городского населения будет менее положительно влиять на безработицу, чем на востоке, а gdp будет влиять ещё более отрицательно.

Начнём работать с моделью, а именно – решим вопрос с выбросами. Воспользуемся 3 методами (удаление выбросов, создание дамми-переменной для выбросов и регрессия Хубера), сравним их и определим, какой лучше подходит для нашего случая.

1. Удаление выбросов

```
Call:
lm(formula = unemp1 ~ gdp1 + urb1 + educ1 + west1 + pop1 + repub1 + 
    enter1 + west_mult_urb1 + west_mult_gdp1, data = df[-ind, 
    ])

Residuals:
   Min     1Q Median     3Q    Max 
-4.335 -1.168 -0.034  0.850  4.612 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)   
(Intercept)     1.183e+01  4.129e+00   2.864  0.00573 **
gdp1           -2.425e-06  4.137e-06  -0.586  0.55984   
urb1           -2.937e-02  3.362e-02  -0.874  0.38571   
educ1          -2.707e-02  6.102e-02  -0.444  0.65891   
west1           7.555e-01  3.280e+00   0.230  0.81861   
pop1           -1.731e-02  2.125e-01  -0.081  0.93534   
repub1          1.650e+00  6.352e-01   2.597  0.01176 * 
enter1         -9.622e-06  6.025e-06  -1.597  0.11541   
west_mult_urb1 -2.525e-02  4.468e-02  -0.565  0.57400   
west_mult_gdp1 -4.783e-06  1.030e-05  -0.465  0.64394   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.736 on 61 degrees of freedom
Multiple R-squared:  0.3721,	Adjusted R-squared:  0.2795 
F-statistic: 4.017 on 9 and 61 DF,  p-value: 0.0004506
```

Как можно заметить, R^2_{adj} упал до 0.2795. Новая модель без выбросов стала только хуже.

2. Дамми-переменная для выбросов

Быть может, сформировав дамми-переменную для выбросов, мы улучшим качество модели. Всем регионам с безработицей больше 15% была добавлена дамми-переменная special со значением = 1 (соответственно, 0 при безработице меньше 15%; порог определён на основе диаграмм рассеяния).

Оценим новую модель:

```
Call:
lm(formula = unemp ~ special + gdp + urb + educ + west + west_mult_urb +
    west_mult_gdp + pop + enter + repub, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-13.040  -1.122  -0.253   1.044  32.984

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.899e+00  1.075e+01   0.642    0.523
special1       1.714e+01  3.253e+00   5.269 1.53e-06 ***
gdp           -3.518e-06  1.137e-05  -0.310    0.758
urb            8.670e-03  9.226e-02   0.094    0.925
educ          -7.172e-02  1.559e-01  -0.460    0.647
west           1.026e+01  7.766e+00   1.321    0.191
west_mult_urb -1.166e-01  1.097e-01  -1.063    0.292
west_mult_gdp -2.956e-05  2.856e-05  -1.035    0.304
pop            1.944e-01  5.173e-01   0.376    0.708
enter          2.892e-06  6.789e-06   0.426    0.672
repub          1.722e+00  1.756e+00   0.981    0.330
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.913 on 68 degrees of freedom
Multiple R-squared:  0.6176,   Adjusted R-squared:  0.5613
F-statistic: 10.98 on 10 and 68 DF,  p-value: 7.387e-11
```

Как мы видим, R^2_{adj} повысился до 0.5613. Заметим, что R^2_{adj} увеличился по сравнению с моделью без выбросов и стартовой моделью.

3. Регрессия Хубера

Оценим регрессию Хубера:

```
Call: rlm(formula = unemp ~ gdp + urb + educ + west + west_mult_urb +
    west_mult_gdp + pop + enter + repub, data = df)
Residuals:
     Min       1Q   Median       3Q      Max
-8.52735 -1.00418  0.07762  1.26550 47.14221

Coefficients:
              Value   Std. Error t value
(Intercept)   -2.3000  4.2921    -0.5359
gdp            0.0000  0.0000    -1.7526
urb            0.0004  0.0396     0.0107
educ          -0.0253  0.0685    -0.3695
west           9.4137  3.1010     3.0357
west_mult_urb -0.1021  0.0463    -2.2061
west_mult_gdp  0.0000  0.0000    -1.4862
pop            1.0664  0.1948     5.4731
enter          0.0000  0.0000     0.5350
repub          1.0938  0.7715     1.4178

Residual standard error: 1.7 on 69 degrees of freedom
```

Residual standard error больше в регрессии Хубера, чем когда мы тестируем модель с дамми-переменной. Значит, оставляем дамми-переменную special.

Тест Чоу

Сохранив RSS от каждой из регрессий (отдельно для запада и востока), мы посчитали наблюдаемую F-статистику:

```
Analysis of Variance Table

Response: unemp_west
            Df  Sum Sq Mean Sq F value    Pr(>F)
gdp_west     1  753.32  753.32 18.0424 0.0001013 ***
pop_west     1 1246.62 1246.62 29.8574 1.724e-06 ***
enter_west   1   32.83   32.83  0.7863 0.3797374
repub_west   1   23.12   23.12  0.5538 0.4604599
urb_west     1   19.13   19.13  0.4582 0.5018013
educ_west    1    0.50    0.50  0.0121 0.9129926
Residuals   47 1962.37   41.75
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Analysis of Variance Table

Response: unemp_east
            Df  Sum Sq Mean Sq F value   Pr(>F)
gdp_east     1  23.223  23.223  2.7368 0.115392
pop_east     1  72.751  72.751  8.5736 0.008983 **
enter_east   1   0.072   0.072  0.0085 0.927474
repub_east   1   4.025   4.025  0.4743 0.499787
urb_east     1   1.296   1.296  0.1527 0.700528
educ_east    1   0.048   0.048  0.0056 0.941035
Residuals   18 152.739   8.486
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Analysis of Variance Table

Response: unemp
           Df  Sum Sq Mean Sq F value    Pr(>F)
gdp         1  337.76  337.76  8.6952   0.00430 **
pop         1  835.15  835.15 21.4998 1.541e-05 ***
enter       1    6.59    6.59  0.1696   0.68172
repub       1  199.87  199.87  5.1453   0.02631 *
educ        1    6.12    6.12  0.1576   0.69251
urb         1  109.74  109.74  2.8251   0.09714 .
Residuals  72 2796.82   38.84
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #chow <- {(RSS - (RSS1+RSS2))/(k+1)}/(RSS1+RSS2)/(n-(2(k+1)))
> chow <- ((2796.82-(152.739+1962.37))/(7))/((152.739+1962.37)/65)
> chow

[1] 2.992836
```

Подставив ее в формулу, мы получили P-value = 0.02016. Так как P-value < 0.05, следовательно гипотеза H0 отвергается на уровне 5%. *Для западных и восточных регионов следует применять отдельные модели.*

Далее продолжим оценивать модель для Запада:

Теперь настало время выбрать наиболее подходящую функциональную форму модели. Создав логарифмированные переменные, оценим линейную в логарифмах и полулогарифмическую модели[1].

Линейная в логарифмах модель:

```
Call:
lm(formula = l_unemp_west ~ l_gdp_west + l_urb_west + l_educ_west +
    l_pop_west + l_enter_west + repub_west + special_west, data = df_west)

Residuals:
     Min       1Q   Median       3Q      Max
-0.65731 -0.14833 -0.04883  0.18740  0.49383

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.82541    1.83466   4.810 1.66e-05 ***
l_gdp_west    -0.36748    0.13449  -2.732  0.00889 **
l_urb_west    -0.38869    0.34153  -1.138  0.26097
l_educ_west   -0.42016    0.21370  -1.966  0.05534 .
l_pop_west     0.86139    0.41875   2.057  0.04538 *
l_enter_west  -0.16862    0.05852  -2.882  0.00599 **
repub_west     0.14301    0.11923   1.199  0.23649
special_west1  0.41994    0.20523   2.046  0.04648 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2739 on 46 degrees of freedom
Multiple R-squared:  0.8117,      Adjusted R-squared:  0.783
F-statistic: 28.32 on 7 and 46 DF,  p-value: 1.173e-14
```

Полулогарифмическая модель:

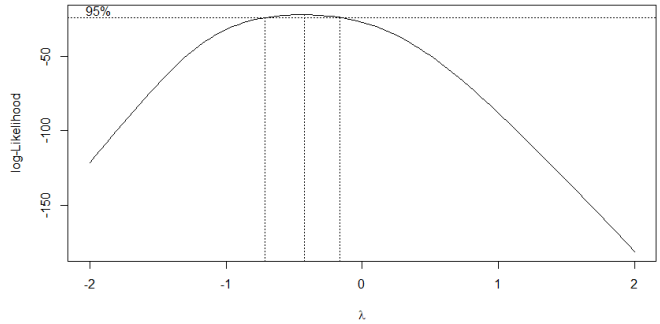
```
Call:
lm(formula = l_unemp_west ~ gdp_west + urb_west + educ_west +
    pop_west + enter_west + repub_west + special_west, data = df_west)

Residuals:
     Min       1Q   Median       3Q      Max
-0.47064 -0.14901  0.00733  0.14928  0.69591

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.030e+00  5.300e-01   3.829 0.000387 ***
gdp_west      -1.928e-06  1.450e-06  -1.330 0.190205
urb_west      -7.313e-03  4.579e-03  -1.597 0.117134
educ_west     -3.032e-03  9.527e-03  -0.318 0.751764
pop_west       7.268e-02  3.644e-02   1.995 0.052026 .
enter_west    -1.352e-06  3.845e-07  -3.517 0.000995 ***
repub_west     1.856e-01  1.040e-01   1.785 0.080911 .
special_west1  6.709e-01  1.880e-01   3.569 0.000852 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2507 on 46 degrees of freedom
Multiple R-squared:  0.8423,      Adjusted R-squared:  0.8183
F-statistic:  35.1 on 7 and 46 DF,  p-value: < 2.2e-16
```

Видно сразу, что полулогарифмическая лучше линейной в логарифмах по R^2_{adj}.

Однако для сравнения моделей с линейной требуется проверить специальные тесты. Остановимся на тестах Бокса-Кокса и Бера-МакАлера:

[1] Выбросы уже учтены как дамми-переменная

Тест Бокса-Кокса:



В интервал не попали ни лямбда = 1, ни лямбда = 0, ни лямбда = 1, следовательно, тест Бокса-Кокса не помогает в этом случае. Обратимся к следующему тесту.

Тест Бера и МакАлера:

Вспомогательные регрессии:

```
Call:
lm(formula = l_unemp_west ~ gdp_west + urb_west + educ_west +
    pop_west + enter_west + v1, data = df_west)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5007 -0.1517  0.0372  0.1386  0.5948

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.426e+00  5.154e-01   2.767 0.008075 **
gdp_west    -2.927e-06  1.399e-06  -2.093 0.041757 *
urb_west    -1.106e-02  4.498e-03  -2.459 0.017686 *
educ_west    4.670e-03  9.119e-03   0.512 0.610944
pop_west     1.622e-01  2.936e-02   5.522 1.41e-06 ***
enter_west  -1.297e-06  3.749e-07  -3.458 0.001165 **
v1           4.649e-02  1.267e-02   3.668 0.000623 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2551 on 47 degrees of freedom
Multiple R-squared:  0.8332,      Adjusted R-squared:  0.8119
F-statistic: 39.12 on 6 and 47 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = unemp_west[-18] ~ gdp_west[-18] + urb_west[-18] +
    educ_west[-18] + pop_west[-18] + enter_west[-18] + v2, data = df_west)[2]

Residuals:
   Min     1Q Median     3Q    Max
-9.437 -2.385 -0.455  1.161 35.602

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.087e+01  1.299e+01  -0.837 0.406776
gdp_west[-18]   -7.043e-05  3.433e-05  -2.051 0.045949 *
urb_west[-18]   -6.813e-02  1.147e-01  -0.594 0.555530
educ_west[-18]   7.111e-02  2.321e-01   0.306 0.760694
pop_west[-18]    3.018e+00  7.211e-01   4.186 0.000127 ***
enter_west[-18] -7.740e-06  2.025e-05  -0.382 0.704071
v2               1.199e+01  6.115e+00   1.961 0.055965 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 6.261 on 46 degrees of freedom
Multiple R-squared:  0.5456,   Adjusted R-squared:  0.4863 F-statistic: 9.205
on 6 and 46 DF,  p-value: 1.266e-06
```

[2] [-18], потому что модель выбрасывает значение с этим индексом, возможно, потому что оно около 0 и нельзя взять логарифм в 1 из вспомогательных регрессий.

Как мы видим, в первом случае нулевая гипотеза о том, что v1=0 отвергается при P-value > 0.05 и, следовательно, коэффициент значим, во втором случае коэффициент при v2 незначим на уровне 5%. Следовательно, линейная модель точно лучше полулогарифмической. Таким образом, выбираем линейную модель для дальнейшего исследования.

Тест Рамсея:

Проведем тест Рамсея:

```
        RESET test

data:  model_west_lin
RESET = 1.4343, df1 = 2, df2 = 44, p-value = 0.2492
```

Нулевая гипотеза не отвергается на уровне значимости в 5%, а значит можно сказать, что нет неучтенных переменных в модели и модель правильно специфицирована.

Проверка на мультиколлинеарность:

Рассчитаем VIF-ы для переменных:

gdp	2.66
repub	3.3
urb	2.58
educ	2.33
pop	1.9
enter	3.49

Ни один из них не получился больше 10, значит, в модели отсутствует мультиколлинеарность, однако, несмотря на адекватность VIFов, в модели только один значимый коэффициент и поэтому попробуем методом исключения сделать так, чтобы стало больше значимых коэффициентов и посмотрим на изменение R^2_{adj}

```
Call:
lm(formula = unemp_west ~ special_west + gdp_west + urb_west + 
    educ_west + pop_west + enter_west + repub_west, data = df_west)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.3414  -1.2511  -0.3365   1.1760  30.7386 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    2.721e+00  1.189e+01   0.229 0.819983    
special_west1  1.691e+01  4.217e+00   4.009 0.000222 ***
gdp_west      -3.803e-05  3.253e-05  -1.169 0.248335    
urb_west      -3.576e-02  1.027e-01  -0.348 0.729326    
educ_west     -6.177e-02  2.137e-01  -0.289 0.773847    
pop_west       1.213e+00  8.174e-01   1.484 0.144503    
enter_west     1.897e-06  8.626e-06   0.220 0.826925    
repub_west     1.151e+00  2.333e+00   0.493 0.624191    
---
```

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

Residual standard error: 5.623 on 46 degrees of freedom
Multiple R-squared:  0.6398,   Adjusted R-squared:  0.585
F-statistic: 11.67 on 7 and 46 DF,  p-value: 2.005e-08
```

Теперь методом пошагового исключения переменных уберем незначимые коэффициенты:

1) уберём enter_west
2) уберём educ_west
3) убираем repub
4) убираем urb

Теперь мы получили модель, где все коэффициенты значимы и при этом R^2_{adj} увеличился.

```
Call:
lm(formula = unemp_west ~ gdp_west + special_west + pop_west,
    data = df_west)

Residuals:
    Min      1Q  Median      3Q     Max
-14.010  -1.254  -0.333   1.226  30.635

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.369e+00  6.518e+00  -0.517   0.6075
gdp_west     -4.089e-05  1.987e-05  -2.058   0.0449 *
special_west1  1.711e+01  3.896e+00   4.391 5.85e-05 ***
pop_west      1.515e+00  6.606e-01   2.294   0.0260 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.424 on 50 degrees of freedom
Multiple R-squared:  0.6358,   Adjusted R-squared:  0.6139
F-statistic: 29.09 on 3 and 50 DF,  p-value: 4.997e-11
```

Гетероскедастичность

Настало время проверить на гетероскедастичность тестом Бройша-Пагана:

```
studentized Breusch-Pagan test

data:  model6_west
BP = 24.913, df = 3, p-value = 1.61e-05
```

Как можно заметить, на уровне 5% мы наблюдаем гетероскедастичность.

Проведем коррекцию, залогарифмировав параметры модели, и оценим снова:

```
Call:
lm(formula = log(unemp_west) ~ log(gdp_west) + log(pop_west) +
    special_west, data = df_west)

Residuals:
     Min       1Q   Median       3Q      Max
-1.30170 -0.13824  0.01098  0.16531  0.66680

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.1150     1.7188   3.558 0.000830 ***
log(gdp_west)  -0.6705     0.1255  -5.343 2.25e-06 ***
log(pop_west)   1.5183     0.4297   3.533 0.000894 ***
special_west1   0.3263     0.2534   1.288 0.203798
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3448 on 50 degrees of freedom
```

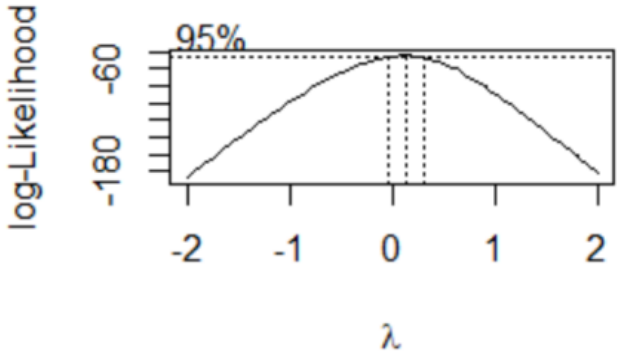
```
Multiple R-squared:  0.6756,   Adjusted R-squared:  0.6562
F-statistic: 34.72 on 3 and 50 DF,  p-value: 2.834e-12
```

Проверим ещё раз на гетероскедастичность:

```
studentized Breusch-Pagan test

data:  model7_west
BP = 17.409, df = 3, p-value = 0.0005822
```

Гетероскедастичность не исчезла, но стала намного меньше.

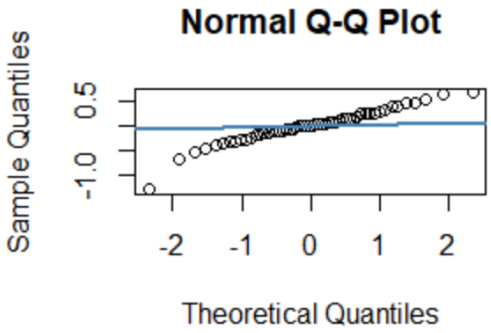
Проведя тест Бокса-Кокса, мы выяснили, ноль что попал в интервал, а значит, логарифмическая модель лучше, чем линейная.



Таким образом, лучшая модель:

```
Call:
lm(formula = log(unemp_west) ~ log(gdp_west) + log(pop_west) +
    special_west, data = df_west)

Residuals:
     Min       1Q   Median       3Q      Max
-1.30170 -0.13824  0.01098  0.16531  0.66680

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.1150     1.7188   3.558 0.000830 ***
log(gdp_west)  -0.6705     0.1255  -5.343 2.25e-06 ***
log(pop_west)   1.5183     0.4297   3.533 0.000894 ***
special_west1   0.3263     0.2534   1.288 0.203798
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3448 on 50 degrees of freedom
Multiple R-squared:  0.6756,   Adjusted R-squared:  0.6562
F-statistic: 34.72 on 3 and 50 DF,  p-value: 2.834e-12
```

Нормальность остатков



Осталось проверить на нормальность остатков:

По графику видно, что наше распределение остатков не совпадает с нормальным распределением, но чтобы окончательно в этом убедиться, проверим тест Шапиро-Уилка (подходит для небольших выборок)

```
Shapiro-Wilk normality test

data:  residuals
W = 0.58194, p-value = 1.175e-13
```

Отвергаем гипотезу о нормальности распределения на уровне 5%. Отсутствие нормальнос ти остатков говорит нам об отсутствии нормальности самого распределения данных (из св ойств нормального распределения). Таким образом, значимость регрессии (см. P-value F-с татистики ниже), может быть под сомнением, потому что нарушена одна из предпосылок F-теста о нормальности распределения, однако на реальных данных так часто бывает и ост аётся либо признать неуверенность в модели, либо исследовать дальше. Мы остановимся на этом шаге, как и в предыдущей домашней работе (3.1). Таким образом, мы получили сл едующую модель для западных регионов:

```
Call:
lm(formula = log(unemp_west) ~ log(gdp_west) + log(pop_west) +
    special_west, data = df_west)

Residuals:
     Min       1Q   Median       3Q      Max
-1.30170 -0.13824  0.01098  0.16531  0.66680

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.1150     1.7188   3.558 0.000830 ***
log(gdp_west)  -0.6705     0.1255  -5.343 2.25e-06 ***
log(pop_west)   1.5183     0.4297   3.533 0.000894 ***
special_west1   0.3263     0.2534   1.288 0.203798
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3448 on 50 degrees of freedom
Multiple R-squared:  0.6756,   Adjusted R-squared:  0.6562
```

```
F-statistic: 34.72 on 3 and 50 DF,  p-value: 2.834e-12
```

R^2_{adj} увеличился почти в 1.5 раза по сравнению со стартовой моделью. Если ВРП увеличит ся на 1%, то безработица упадёт на 0.67%. Если демографический прирост (pop) поднимет ся на 1%, то безработица – на 1.51%.
Отрицательная зависимость безработицы от log ВРП обусловлена действием закона Оукен а. В свою очередь, рост безработицы при увеличении логарифмированного демографическ ого прироста объясняется увеличением количества соискателей работы при ограниченном количестве рабочих мест. Таким образом, мы получили рабочую модель, которая неплохо объясняет сложившиеся зависимости.

Модель для восточных регионов

Теперь подберём лучшую модель зависимости безработицы для восточных регинов России.

Для начала, как и при работе с западом, введём дамми-переменные для регионов-выбросов (special_east). Теперь построим модель с учётом дамми-переменных для выбросов и переменных из *стартовой модели* при условии, что мы рассматриваем выборку с восточными регионами:

```
Call:
lm(formula = unemp_east ~ special_east + gdp_east + urb_east + 
    educ_east + pop_east + enter_east + repub_east, data = df_east)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1980 -0.9064  0.0000  1.0870  2.8274 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    2.426e+01  5.265e+00   4.609  0.00025 ***
special_east1  1.511e+01  2.364e+00   6.391  6.7e-06 ***
gdp_east       1.475e-06  4.213e-06   0.350  0.73063    
urb_east      -8.943e-02  3.655e-02  -2.447  0.02559 *  
educ_east     -5.792e-02  9.574e-02  -0.605  0.55319    
pop_east      -6.807e-01  2.903e-01  -2.345  0.03144 *  
enter_east    -1.339e-05  9.772e-06  -1.371  0.18832    
repub_east     1.589e+00  1.192e+00   1.333  0.20023    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.625 on 17 degrees of freedom
Multiple R-squared:  0.8234,      Adjusted R-squared:  0.7507 
F-statistic: 11.32 on 7 and 17 DF,  p-value: 2.607e-05
```

Согласно результатам модели, R^2_{adj} уже выше, чем в *стартовой модели*. Пояснять значим ые коэффициенты нет смысла (для примера см. модель для западных регионов), потому чт о далее мы будем преобразовывать модель.

Теперь попробуем поэкспериментировать с функциональной формой модели. Введём лога рифмы переменных для НЕ-дамми и оценим логарифмическую, полулогарифмическую и линейную модель. Результаты представлены ниже:

Линейная в логарифмах:

```
Call:
lm(formula = l_unemp_east ~ l_gdp_east + l_urb_east + l_educ_east + 
    l_pop_east + l_enter_east + repub_east + special_east, data = df_east)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4175 -0.1081  0.0000  0.1148  0.3496 

Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.11518    2.31304   3.076  0.00685 **
l_gdp_east    -0.03989    0.14552  -0.274  0.78731
l_urb_east    -0.42988    0.34079  -1.261  0.22420
l_educ_east   -0.27345    0.29416  -0.930  0.36560
l_pop_east    -0.83258    0.66194  -1.258  0.22547
l_enter_east   0.01144    0.06413   0.178  0.86049
repub_east     0.24820    0.16847   1.473  0.15894
special_east1  1.06988    0.35393   3.023  0.00767 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2279 on 17 degrees of freedom
Multiple R-squared:  0.6142,      Adjusted R-squared:  0.4554
F-statistic: 3.867 on 7 and 17 DF,  p-value: 0.01071
```

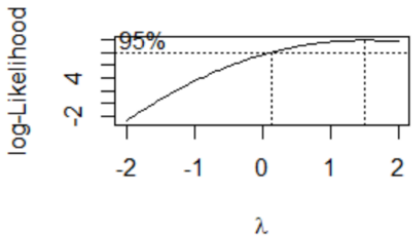
Полулогарифмическая модель:

```
Call:
lm(formula = l_unemp_east ~ gdp_east + urb_east + educ_east +
    pop_east + enter_east + repub_east + special_east, data = df_east)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32681 -0.08758  0.00000  0.14780  0.35398

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.271e+00  6.714e-01   6.362 7.08e-06 ***
gdp_east       2.760e-07  5.372e-07   0.514 0.614000
urb_east      -1.136e-02  4.662e-03  -2.437 0.026101 *
educ_east     -8.383e-03  1.221e-02  -0.687 0.501576
pop_east      -9.957e-02  3.702e-02  -2.689 0.015513 *
enter_east    -1.627e-06  1.246e-06  -1.305 0.209191
repub_east     2.391e-01  1.520e-01   1.573 0.134151
special_east1  1.280e+00  3.015e-01   4.246 0.000545 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2072 on 17 degrees of freedom
Multiple R-squared:  0.6811,      Adjusted R-squared:  0.5498
F-statistic: 5.186 on 7 and 17 DF,  p-value: 0.002636
```

Линейная модель:

```
Call:
lm(formula = unemp_east ~ special_east + gdp_east + urb_east +
    educ_east + pop_east + enter_east + repub_east, data = df_east)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1980 -0.9064  0.0000  1.0870  2.8274

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.426e+01  5.265e+00   4.609  0.00025 ***
special_east1  1.511e+01  2.364e+00   6.391  6.7e-06 ***
gdp_east       1.475e-06  4.213e-06   0.350  0.73063
urb_east      -8.943e-02  3.655e-02  -2.447  0.02559 *
educ_east     -5.792e-02  9.574e-02  -0.605  0.55319
pop_east      -6.807e-01  2.903e-01  -2.345  0.03144 *
enter_east    -1.339e-05  9.772e-06  -1.371  0.18832
repub_east     1.589e+00  1.192e+00   1.333  0.20023
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.625 on 17 degrees of freedom
Multiple R-squared:  0.8234,      Adjusted R-squared:  0.7507
F-statistic: 11.32 on 7 and 17 DF,  p-value: 2.607e-05
```

Для выбора наиболее оптимальной формы модели воспользуемся тестом Бокса-Кокса:



По графику видно, что лямбда = 1 попадает в интервал, а лямбда = 0 – нет. Следовательно, линейная модель является оптимальной.

Проверим специфику выбранной модели (есть ли пропущенные переменные или нет) с по мощью теста Рамсея:

```
        RESET test

data:  model_east_lin
RESET = 2.0339, df1 = 2, df2 = 15, p-value = 0.1654
```

P-value достаточно большое (гораздо больше 0.05), значит, нулевая гипотеза не отвергаетс я на уровне 5% и можно сказать, что модель правильно специфицирована, то есть не включа ет неучтенные переменные.

Что насчёт мультиколлинеарности? Посчитаем VIFы переменных:

```
vif_gdp <- 1/(1-0.2405) = 1.316656
vif_urb <- 1/(1-0.541) = 2.178649
vif_educ <- 1/(1-0.09827) = 1.108979
vif_pop <- 1/(1-0.6192) = 2.62605
vif_enter <- 1/(1-0.3087) = 1.44655
vif_repub <- 1/(1-0.5347) = 2.149151
```

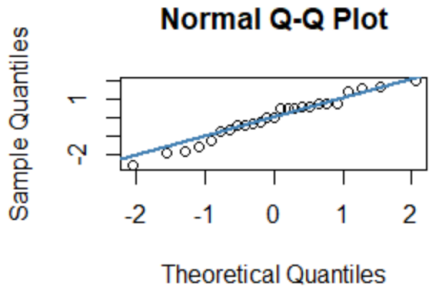
Все VIFы не превышают 10, следовательно, мультиколлинеарность не наблюдается.
Как и при исследовании западных регионов, мы всё-таки применили метод исключения, о днако здесь в результате R^2_{adj} уменьшился, поэтому мы решили не трогать модель – не иск лючать коэффициенты.

Для проверки на гетероскедастичность мы воспользовались тестом Бройша-Пагана и полу чили следующие результаты:

```
        studentized Breusch-Pagan test

data:  model_east_lin
BP = 5.2473, df = 7, p-value = 0.6298
```

P-value > 0.05, следовательно, на уровне 5% нулевая гипотеза о гомоскедастичности не от вергается и нет необходимости делать коррекцию, которая была использована в случае с з ападными регионами.

Наконец, проверим нормальность остатков. Ниже представлен график qqnorm, изображающий наше распределение и нормальное распределение:



По графику нельзя сделать точный вывод, принадлежат ли остатки нормальному распределению или нет, поэтому воспользуемся тестом Шапиро-Уилка (подходит для небольших выборок):

```
Shapiro-Wilk normality test

data:  residuals
W = 0.58194, p-value = 1.175e-13
```

Нулевая гипотеза о нормальности остатков отвергается на уровне 5% ввиду маленького p-value в тесте.

Таким образом, для восточных регионов лучшей моделью является линейная модель с учётом выбросов как дамми-переменных без коррекции на гетероскедастичность и без борьбы с мультиколлинеарностью.

Посмотрим на результаты модели ещё раз и проанализируем их:

```
Call:
lm(formula = unemp_east ~ special_east + gdp_east + urb_east +
    educ_east + pop_east + enter_east + repub_east, data = df_east)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1980 -0.9064  0.0000  1.0870  2.8274

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.426e+01  5.265e+00   4.609  0.00025 ***
special_east1  1.511e+01  2.364e+00   6.391  6.7e-06 ***
gdp_east       1.475e-06  4.213e-06   0.350  0.73063
urb_east      -8.943e-02  3.655e-02  -2.447  0.02559 *
educ_east     -5.792e-02  9.574e-02  -0.605  0.55319
pop_east      -6.807e-01  2.903e-01  -2.345  0.03144 *
enter_east    -1.339e-05  9.772e-06  -1.371  0.18832
repub_east     1.589e+00  1.192e+00   1.333  0.20023
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.625 on 17 degrees of freedom
Multiple R-squared:  0.8234,      Adjusted R-squared:  0.7507
F-statistic: 11.32 on 7 and 17 DF,  p-value: 2.607e-05
```

Таким образом, мы получили модель, которая даёт очень высокий R^2_{adj}, что свидетельствует о хорошем качестве подгонки. Из коэффициентов получились значимые следующие: special_east на уровне 0.001, urb_east на уровне 0.05, pop_east на уровне 0.05. Это значит, что если регион - выброс, то безработица будет больше на 15.11 ед. (%), если же доля городского населения увеличивается на 1 единицу, то безработица будет ниже на 0.08943 ед. (%). А если добавить по одному ребенку на 1000 человек, то безработица сократится на 0.6807 ед. (%).

Это объясяется тем, что восток страны менее развит, чем запад. В городах проживает меньшая доля населения, меньше рабочих мест для желающих найти работу. На каждое рабочее место могут претендовать сразу несколько переселенцев с сельской местности, что усугубляет ситуацию с официальной безработицей. Демографический прирост также положительно влияет на размер безработицы, так как число рабочих мест ограничено, а претендентов становится все больше.

В итоге мы получили для западных и восточных регионов разные модели, что вполне характерно для России. Различное положение в экономике, обусловленное местоположением основных производств и месторождений, долей городского и образованного населения обуславливает разную конъюктуру рынка труда для западных и восточных регионов, что и было отражено в двух эконометрических моделях.