1 ДЗ-02

1.1 Задача 1.

1.1.1 Построение модели оценки за экзамен.

Во-первых, по имеющимся данным мы сможем оценить только их влияние на оценку за второй экзамен, так как данные агрегированы за весь семестр.

Во-вторых имеет смысл вместо суммы баллов за домашние задания использовать средний балл, так как он точнее отражает качество выполнения домашнего задания:

$$MHW_i = \frac{SHW_i}{NHW_i} \tag{1.1}$$

Для определения влияния различных факторов на оценки построим модель, включающую в себя все объективные характеристики, а также дамми на пол:

Source	SS	df	MS	Number of obs	=	223
				F(6, 216)	=	79.19
Model	70477.5036	6	11746.2506	Prob > F	=	0.0000
Residual	32041.2498	216	148.339119	R-squared	=	0.6875
				Adj R-squared	=	0.6788
Total	102518.753	222	461.796186	Root MSE	=	12.179

exam2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
NA	-.401993	.2116452	-1.90	0.059	-.8191473	.0151612
NHW	1.216641	.3094121	3.93	0.000	.6067875	1.826495
mean_SHW	.132852	.0594394	2.24	0.026	.0156965	.2500076
exam1	.7557981	.0557075	13.57	0.000	.6459982	.865598
f	5.693563	4.49672	1.27	0.207	-3.169505	14.55663
f_mean_SHW	-.1110807	.066832	-1.66	0.098	-.242807	.0206457
_cons	2.456695	3.705912	0.66	0.508	-4.847685	9.761074

Сразу заметим, что коэффициент при количестве посещенных семинаров значим и меньше 0. Можно подумать, что если не сходить 10 семинаров при прочих равных можно получить оценку на 4 балла выше, что очевидно не так. Причиной этому является тот факт, что в регрессии есть пропущенные переменные, отвечающая за знания эконометрики на момент начала курса и ”талант”, который позволяет студентов быстро готовиться за ночь до экзамена.

Можно предположить, что переменная $Expect$ частично отражает эти факторы, так как высокую оценку будут ожидать только те, кто более-менее уверен в своих знаниях:

Source	SS	df	MS	Number of obs	=	223
				F(7, 215)	=	125.69
Model	82385.9846	7	11769.4264	Prob > F	=	0.0000
Residual	20132.7688	215	93.640785	R-squared	=	0.8036
				Adj R-squared	=	0.7972
Total	102518.753	222	461.796186	Root MSE	=	9.6768

exam2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
NA	-.1452487	.1696905	-0.86	0.393	-.4797187	.1892213
NHW	.8366313	.2481328	3.37	0.001	.3475468	1.325716
mean_SHW	.0817556	.0474427	1.72	0.086	-.0117567	.1752679
exam1	.2489226	.0630816	3.95	0.000	.124585	.3732601
f	2.48884	3.584017	0.69	0.488	-4.575469	9.553148
f_mean_SHW	-.0552568	.0533296	-1.04	0.301	-.1603726	.0498589
Expect	.6905583	.0612357	11.28	0.000	.5698591	.8112575
_cons	-2.343719	2.975031	-0.79	0.432	-8.207682	3.520244

Действительно, добавив $Expect$ в регрессию видим, что эта переменная значима, а количество посещенных семинаров NA – нет, также теперь нет различия во влиянии среднего балла на оценки в зависимости от пола. Все это позволяет сделать вывод, что из-за пропущенной переменной оценки были несостоятельными, а $Expect$ из регрессии убирать не стоит.

Проверим с помощью F-теста гипотезу об одновременном равенстве коэффициентов при NA, f, $f * MHW$:

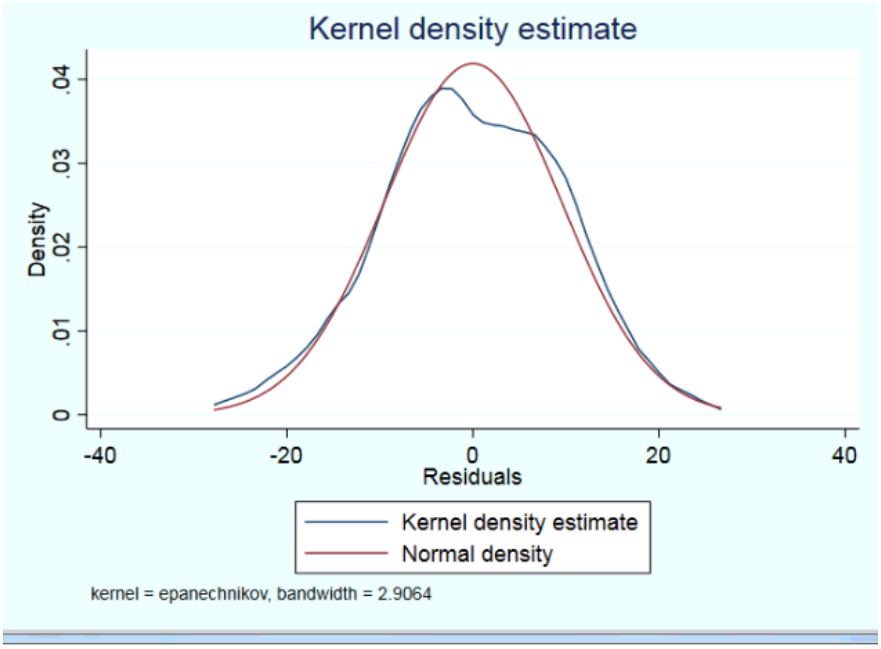
```
( 1)  NA = 0
( 2)  f = 0
( 3)  f_mean_SHW = 0

       F(  3,    215) =    0.77
            Prob > F =    0.5133
```

Согласно результатам теста далее будем использовать короткую модель:

Source	SS	df	MS	Number of obs	=	223
				F(4, 218)	=	220.08
Model	82170.3586	4	20542.5896	Prob > F	=	0.0000
Residual	20348.3948	218	93.3412605	R-squared	=	0.8015
				Adj R-squared	=	0.7979
Total	102518.753	222	461.796186	Root MSE	=	9.6613

exam2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
NHW	.8185909	.2421921	3.38	0.001	.3412531	1.295929
mean_SHW	.060276	.0386379	1.56	0.120	-.0158758	.1364277
exam1	.2490318	.0625426	3.98	0.000	.1257662	.3722973
Expect	.7039217	.0603514	11.66	0.000	.5849747	.8228687
_cons	-2.513992	1.839057	-1.37	0.173	-6.138601	1.110616

Визуально и согласно критерию Шапиро-Уилка нет оснований предполагать что остатки не распределены нормально:



Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
res	223	0.99414	0.961	-0.092	0.53678

Также согласно тесту Бройша-Пагана гипотеза о гомоскедастичности ошибок не отвергается:

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of exam2

         chi2(1)      =     0.08
         Prob > chi2  =   0.7812
```

У модели также достаточно высокий коэффициент детерминации. Поэтому можем считать модель ”хорошей”. Согласно полученным оценкам можно сделать вывод, что количество сданных домашних заданий, оценка за 1-ый экзамен и ожидания положительно влияют на оценку за 2-ой экзамен (менее чем на 1% уровне значимости), что согласуется с логикой.

1.1.2 Построение модели ожидаемой оценки за экзамен.

Далее определим, какие факторы могут влиять на ожидания. Построим модель, включающую в себя все объективные характеристики:

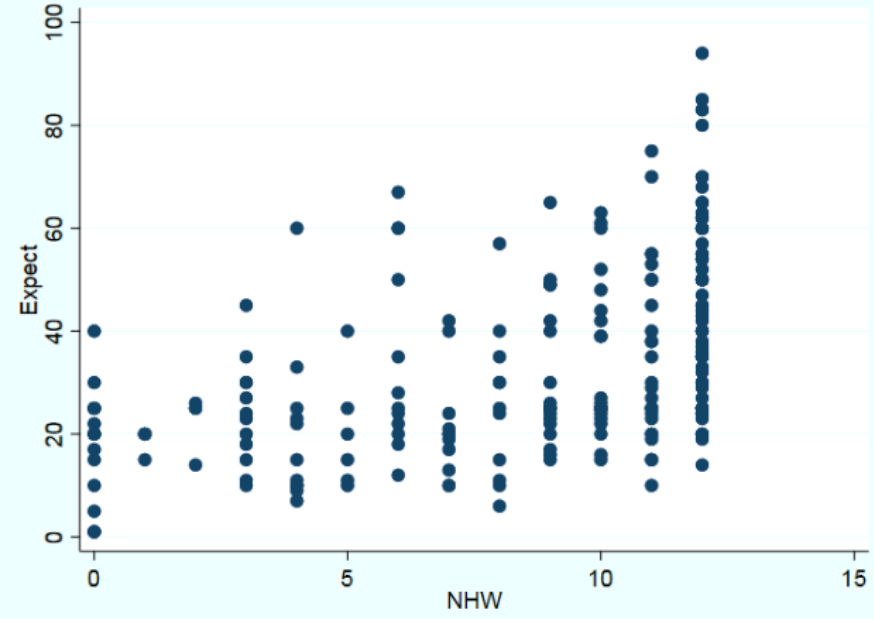
Source	SS	df	MS	Number of obs	=	223
				F(6, 216)	=	71.22
Model	49772.0951	6	8295.34919	Prob > F	=	0.0000
Residual	25157.3309	216	116.469125	R-squared	=	0.6643
				Adj R-squared	=	0.6549
Total	74929.426	222	337.519937	Root MSE	=	10.792

Expect	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
NA	-.2438366	.2776303	-0.88	0.381	-.791048	.3033749
NHW	.5688303	.2750383	2.07	0.040	.0267278	1.110933
mean_SHW	.0324381	.0436314	0.74	0.458	-.0535597	.1184359
exam1	.730806	.0494148	14.79	0.000	.6334091	.8282028
f	.4944433	2.274476	0.22	0.828	-3.988566	4.977452
f_NA	-.1802905	.3448319	-0.52	0.602	-.8599568	.4993758
_cons	8.890677	2.913933	3.05	0.003	3.147293	14.63406

Вероятно, в этой модели есть пропущенные переменные такого же рода, однако их влияние должно быть не так критично, как в предыдущем случае, поскольку ожидания зависят не только от способностей экзаменуемого, но также и от особенностей экзамена.

Согласно тесту Бройша-Пагана ошибки в этой модели гетероскедастичны:

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of Expect

         chi2(1)      =     3.62
         Prob > chi2  =   0.0570
```



Визуально можно определить, что причиной гетероскедастичности является переменная NHW, поскольку большинство студентов сдавали домашние задания, из-за чего у больших значений NHW больше разброс:

Поэтому имеет смысл взвесить регрессию по HNW, а также использовать робастные оценки ковариационной матрицы HC3:

```
. regress Expect NA NHW mean_SHW exam1 f f_NA [aweight = NHW], vce(hc3)
(sum of wgt is 1,877)

Linear regression                               Number of obs     =        208
                                                F(6, 201)         =      61.30
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6524
                                                Root MSE          =     11.144

------------------------------------------------------------------------------
             |             Robust HC3
      Expect |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          NA |  -.2550786   .3915256    -0.65   0.515    -1.027103    .5169458
         NHW |   .6793886   .3477677     1.95   0.052    -.0063526    1.36513
    mean_SHW |   .0497841   .0590341     0.84   0.400    -.0666215    .1661898
       exam1 |   .7074181   .0571808    12.37   0.000     .5946668    .8201693
           f |   .4954608   2.687069     0.18   0.854    -4.803001    5.793922
        f_NA |  -.2613373   .4377063    -0.60   0.551    -1.124423     .601748
       _cons |   7.507877   4.591814     1.64   0.104    -1.546429    16.56218
------------------------------------------------------------------------------
```

Согласно тесту Бройша-Пагана проблема гетероскедастичности частично решена:

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of Expect

         chi2(1)      =     2.15
         Prob > chi2  =   0.1424
```

Поэтому теперь можно пользоваться t-статистиками для определения значимости коэффициентов регрессии.

Проверим с помощью F-теста одновременное равенство нулю коэффициентов при MHW, f, $f * NA$:

```
( 1)  mean_SHW = 0
( 2)  f = 0
( 3)  f_NA = 0

       F(  3,    201) =    0.41
            Prob > F =    0.7469
```

По результатам F-теста оценим ограниченную регрессию:

```
. regress Expect NA NHW exam1 [aweight = NHW], vce(hc3)
(sum of wgt is 1,877)

Linear regression                               Number of obs     =        208
                                                F(3, 204)         =     112.57
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6499
                                                Root MSE          =     11.102

------------------------------------------------------------------------------
             |             Robust HC3
      Expect |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          NA |  -.4611818   .2169487    -2.13   0.035    -.888931   -.0334326
         NHW |   .7729683   .3059366     2.53   0.012     .169765    1.376172
       exam1 |   .7258106   .0521346    13.92   0.000    .6230188    .8286024
       _cons |   10.00136   3.490756     2.87   0.005     3.11877    16.88395
------------------------------------------------------------------------------
```

```
. swilk resexp

                    Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V          z       Prob>z
-------------+-------------------------------------------------------
      resexp |        223    0.99622      0.621     -1.103    0.86502
```

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of Expect

         chi2(1)      =     2.15
         Prob > chi2  =   0.1426
```

Ошибки нормальны и гомоскедастичны. R^2 достаточно большой. Поэтому по данной модели имеем положительну связь между числом выполненных заданий и оценкой за 2-ой экзамен (на 5% уровне значимости), положительну связь между оценкой за 1-ый экзамен и оценкой за 2-ой экзамен (менее, чем на 1% уровне значимости), отрицательную свзяь между

числом посещенных семинаров и оценкой за 2-ой экзамен (на 5% уровне значимости). Последнее, вероятно, можно объяснить тем, что из-за пропущенной переменной отвечающей за начальные эконометрические навыки студента, мы видим это влияние через переменную NA, которая на самом деле отражает не прямую связь между прогулами семинаров и ростом оценки за экзамен, а косвенную через тот факт, что те кто не ходят на семинары предполагают, что знают эконометрику, а соответственно и выше оценивают свои ожидания за экзамен.

1.1.3 Построение модели точности прогноза.

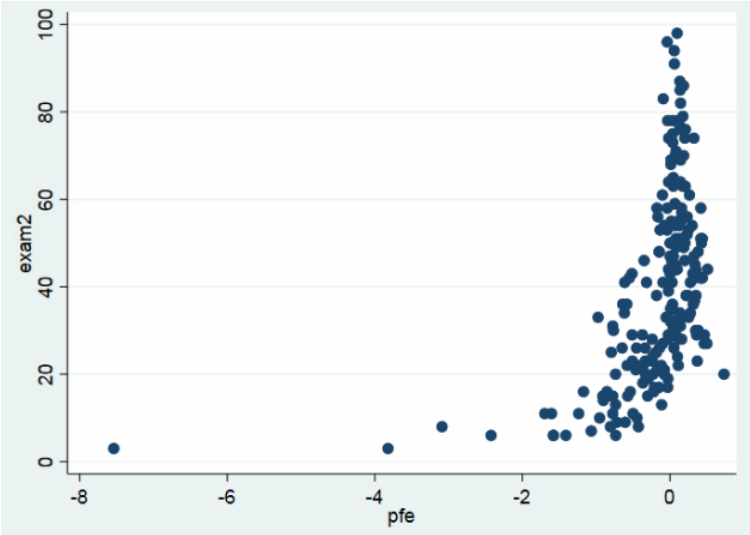
Чтобы определить, насколько точно предсказывают оценку за экзамен студенты, построим регрессию оценок за экзамен на ожидания студентов:

Source	SS	df	MS	Number of obs	=	223
				F(1, 221)	=	666.81
Model	76999.1843	1	76999.1843	Prob > F	=	0.0000
Residual	25519.5691	221	115.473163	R-squared	=	0.7511
				Adj R-squared	=	0.7499
Total	102518.753	222	461.796186	Root MSE	=	10.746

exam2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Expect	1.013717	.0392567	25.82	0.000	.9363518	1.091083
_cons	4.324701	1.500415	2.88	0.004	1.367748	7.281653

Можно заметить, что коэффициент при $Expect$ примерно равен 1 (причем $t_{\{H_0:\beta=1\}} = \frac{\hat{\beta}-1}{se_\beta}$ = 25.82 - 1/0.0393 = 0.3466), а по метрике RMSE можно увидеть, что студенты в среднем ошибаются в своих расчетах на 10 баллов. То есть относительно точно прогнозируют свои оценки.

Для того, чтобы определить, какие факторы влияют на точность прогноза введем переменную ошибка прогноза в долях $PFE_i = \frac{exam_i - \widehat{exam}_i}{exam_i}$:



Так как для ответа на поставленный вопрос не важно, кто завысил, а кто занизил свои ожидания, возьмем этот показатель по модулю, а также возьмем логарифм от этой переменной (из-за взятия логарифма на 2 наблюдения меньше):

```
. regress log_pfe NA NHW mean_SHW exam1 f

      Source |       SS           df       MS      Number of obs   =       221
-------------+----------------------------------   F(5, 215)       =      8.22
       Model |  60.589434         5  12.1178868   Prob > F        =    0.0000
    Residual |  316.932756      215  1.47410584   R-squared       =    0.1605
-------------+----------------------------------   Adj R-squared   =    0.1410
       Total |  377.52219       220  1.71600995   Root MSE        =    1.2141

------------------------------------------------------------------------------
      log_pfe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
           NA |  -.0286377   .021196    -1.35   0.178    -.0704164    .013141
          NHW |  -.0573911   .0309412   -1.85   0.065     -.118378    .0035959
     mean_SHW |  -.0039223   .0049645   -0.79   0.430    -.0137076    .005863
        exam1 |  -.0200552   .0055545   -3.61   0.000    -.0310034   -.0091071
            f |  -.2514773   .1682326   -1.49   0.136    -.5830738    .0801192
        _cons |  -.1955137   .3214631   -0.61   0.544    -.8291364    .438109
------------------------------------------------------------------------------
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of log_pfe

         chi2(1)      =     0.06
         Prob > chi2  =   0.8069
```

Ошибки гомоскедастичны (но на всякий случай будем использовать робастные оценки), поэтому воспользуемся F-тестом, чтобы проверить значимость коэффициентов при NA, MHW, f

```
( 1)  NA = 0
( 2)  mean_SHW = 0
( 3)  f = 0

      F(  3,    215) =    1.60
           Prob > F =    0.1908
```

Согласно F-тесту можем перейти к ограниченной модели:

```
. regress log_pfe NHW exam1, vce(hc3)

Linear regression                               Number of obs     =        221
                                                F(2, 218)         =      24.14
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1418
                                                Root MSE          =     1.2191

------------------------------------------------------------------------------
             |             Robust HC3
      log_pfe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         NHW |  -.0702157   .0250954    -2.80   0.006    -.1196763   -.0207551
       exam1 |  -.0180826   .0044801    -4.04   0.000    -.0269125   -.0092527
       _cons |  -.6578115   .2101586    -3.13   0.002    -1.072014   -.2436087
------------------------------------------------------------------------------
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of log_pfe

         chi2(1)      =     1.72
         Prob > chi2  =   0.1894
```

Ошибки гомоскедастичны (но на всякий случай будем использовать робастные оценки). В результате, чем чаще посещаются семинары, тем точнее прогнозы, чем выше результаты за 1-ый экзамен, тем точнее прогнозы.

1.1.4 Выводы:

- посещение семинаров не влияет на оценку за экзамен, однако влияет на ожидания;
- количество сданных домашних заданий влияет как на оценки, так и на ожидания;
- на ожидания влияют количество посещенных семинаров, количество сданных домашних заданий и оценка за 1-ый экзамен;
- фактор $expect$ является значимым фактором для прогноза оценки $exam_2$;
- студенты ошибаются в среднем в своих прогнозах на 10 баллов; точнее свои оценки предсказывают те, кто сдавал больше домашних заданий и кто лучше написал экзамен.

1.1.5 Программа для STATA:

```
clear
set more off
cd C:\Users\kasyanova\Desktop\stata

log using homeass2_kasianova_shulyak.log, text replace

import excel data_HW_02.xlsx, clear sheet("Sheet1") firstrow

drop if missing(SHW)
drop if missing(Expect)
drop if Expect=="no"
drop if exam1=="n/a"

generate f:f = (man == "")
destring exam1 Expect exam2, replace
recast int NA NHW f exam1 exam2 Expect

generate mean_SHW:mean_SHW = SHW/NHW
replace mean_SHW = 0 if  mean_SHW == .

generate f_mean_SHW:f_mean_SHW = f*mean_SHW

// part I
regress exam2 NA NHW mean_SHW exam1 f f_mean_SHW
regress exam2 NA NHW mean_SHW exam1 f f_mean_SHW Expect
predict res, resid

test NA f f_mean_SHW

regress exam2 NHW mean_SHW exam1 Expect
kdensity  res, normal
```

```
swilk res
estat hettest

// part II
regress Expect NA NHW mean_SHW exam1 f f_NA
regress Expect NHW exam1
estat hettest

scatter Expect NHW

regress Expect NA NHW mean_SHW exam1 f f_NA [aweight = NHW], vce(hc3)
estat hettest

test NA mean_SHW f f_NA
test NA f f_NA
test mean_SHW f f_NA

regress Expect NA NHW exam1 [aweight = NHW], vce(hc3)
predict resexp, resid
kdensity  resexp, normal
swilk resexp
estat hettest

// part III
regress exam2 Expect
predict fe, resid
generate afe:afe = abs(fe)
generate pfe = fe/exam2
generate log_pfe: log_pfe = log(abs(pfe))

scatter exam2 pfe
scatter exam2 log_pfe
```

```
regress afe NA NHW SHW exam1 f f_mean_SHW
estat hettest
regress afe NA NHW SHW exam1 f f_mean_SHW [aweight = NHW], vce(hc3)
estat hettest

regress log_pfe NA NHW mean_SHW exam1 f
test NA mean_SHW f
regress log_pfe NHW exam1
estat hettest

log close
```