

Федеральное бюджетное государственное образовательное учреждение высшего  
профессионального образования

«РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА

и ГОСУДАРСТВЕННОЙ СЛУЖБЫ

при ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»

**ЭКОНОМИЧЕСКИЙ ФАКУЛЬТЕТ**

**ОТДЕЛЕНИЕ ЭКОНОМИКИ**

Кафедра эконометрики

**Проект по эконометрике:**

**Исследование разрыва заработных плат между мужчинами и женщинами в**

**Челябинской области.**

Выполнила:

студентка группы СП-20

Каменская Елизавета

Преподаватель:

Зямалов В.Е.

**МОСКВА, 2022 г.**

## Введение и использованные данные.

В рамках курса эконометрики нам было предложено задание исследования разрыва между зарплатами у мужчин и женщин. Объектом этого исследования была выбрана Челябинская область. Анализ будет проводиться с помощью построения линейной регрессии с соответствующими переменными. За основу мы взяли уравнение Минцера, которое далее будем преобразовывать и проверять на разных переменных. Мы использовали данные 2020-го года, предложенные национальным исследовательским университетом ВШЭ. Для более точных результатов исключим выбросы следующим образом: нашли медиану заработных плат, вычислим межквартильный диапазон и установим границы. В следующих таблицах приведен небольшой отфильтрованный анализ населения:

	Отрасль	Мужчин (доля населения %)	Женщин (доля населения %)	Средняя ежемесячная з/п мужчин (руб/мес)	Средняя ежемесячная з/п женщин (руб/мес)	Отношение з/п женщин к з/п мужчин, %
1	Тяжелая промышленность	7(6)	3(2.6)	34285.7	25000	27
2	Транспорт, связь	11(9.4)	3(2.6)	36545.45	25000	31.6
3	Торговля, бытовое обслуживание	9(7.7)	18(15.4)	28555.5	22083.33	22.7
4	Образование	2(1.7)	10(8.5)	22000	21600	1.8
5	Легкая, пищевая промышленность	4(3.4)	3(2.6)	40750	32333.3	20.65
6	Юриспруденция	1(0.85)	1(0.85)	70000	70000	0
7	Строительство	7(6)	4(3.4)	49285.71	29500	40.14
8	Жилищно- коммунальное хоз-во	2(6)	2(6)	21500	34000	-58.14
9	Гражданское машиностроение	3(2.6)	1(0.85)	35333	50000	-41.5
10	Финансы	1(0.85)	1(0.85)	25000	35000	-40

11	Энергетическая промышленность	2(6)	2(6)	65000	28500	56
12	Армия, МВД	3(2.6)	2(6)	19333.3	32500	-68.1
13	Операции с недвижимостью	0(0)	1(0.85)		30000	
14	Здравоохранение	0(0)	9(7.7)	31777		
15	Нефть и газ	1(0.85)	0(0)	35000		
16	Военная промышленность	0(0)	1(0.85)		34000	
17	Наука, культура	0(0)	2(6)		17500	
18	IT	1(0.85)	0(0)	30000		

**Всего в опросе участвовало 117 человек. Из них: 51 мужчина и 66 женщин.**

Из этой таблицы видно, что самые высокие зарплаты и мужчины и женщины получают, работая в юриспруденции, самые низкие - в отрасли образования. В строительстве мужчины на 40% получают зп больше, чем женщины, женщины на 68.1% получают зп больше чем мужчины, работая в армии, так же на 58% они получают больше чем мужчины, работая в сфере ЖКХ. Также интересно отметить, что в сфере образования, здравоохранения и бытовых услуг женщин работает гораздо больше чем мужчин, чего не скажешь о сферах тяжелой промышленности и транспорта.

Данных достаточно мало и не для всех сфер деятельности нашлись респонденты, которые участвовали в опросе или их очень мало, это может привести к неточным результатам.

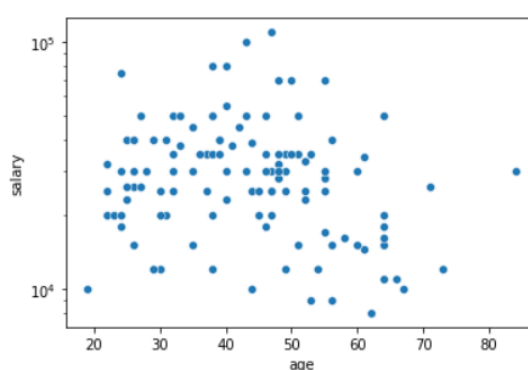
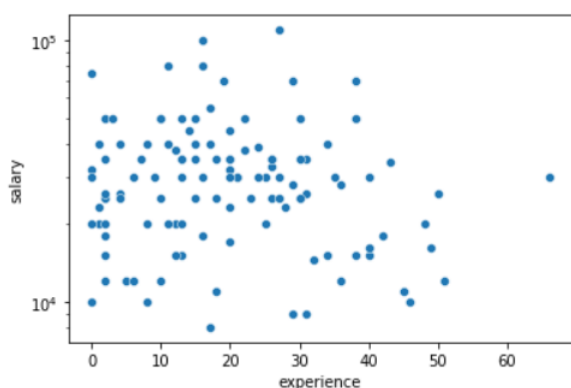
### **Первая построенная модель:**

Как уже было сказано, основой модели стало уравнение Минцера. Оно имеет следующий вид:  $\ln w = a + \alpha_1 edu + \beta_1 exp_i + \beta_2 exp_i^2 + \gamma gend_i + \delta get\_married + u_i$

Здесь

- $\ln w$  – натуральный логарифм среднемесячной заработной платы;
- $age$  – возраст работника;
- $exp$  – стаж работы;
- $gend$  – пол работника (1 – мужчина; 0 – женщина) ;
- $stat$ – тип населенного пункта (1 – город; 0 – село)
- $edu$  – уровень образования:
- $edu$  - (1 – высшее; 0 – остальные);

Протестируем еще две альтернативные модели: модель с учетом возраста и модель с учетом к-ва детей у женщин.



На диаграммах рассеяния видно, что у первой модели экспериментальные точки больше упорядочены как линейная зависимость, поэтому предполагаем, что первая модель окажется лучше.

МНК-оценки:

```

=====
Dep. Variable:      np.log(salary)      R-squared:          0.431
Model:              OLS                 Adj. R-squared:     0.296
Method:             Least Squares       F-statistic:        3.200
Date:               Wed, 21 Dec 2022    Prob (F-statistic): 4.89e-05
Time:               04:10:14           Log-Likelihood:     -60.511
No. Observations:   116                AIC:                167.0
DF Residuals:       93                 BIC:                230.4
DF Model:            22
Covariance Type:    nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
Intercept	9.6984	0.485	19.995	0.000	8.735	10.662
C(work_field)[Т.АРМИЯ, МВД, ОРГАНЫ БЕЗОПАСНОСТИ]	-0.2012	0.519	-0.388	0.699	-1.232	0.830
C(work_field)[Т.ВОЕННО-ПРОМЫШЛЕННЫЙ КОМПЛЕКС]	0.7512	0.685	1.097	0.276	-0.609	2.112
C(work_field)[Т.ГРАЖДАНСКОЕ МАШИНОСТРОЕНИЕ]	0.2112	0.527	0.401	0.689	-0.834	1.257
C(work_field)[Т.ДРУГАЯ ОТРАСЛЬ ТЯЖЕЛОЙ ПРОМЫШЛЕННОСТИ]	0.2736	0.502	0.545	0.587	-0.723	1.270
C(work_field)[Т.ЖИЛИЩНО-КОММУНАЛЬНОЕ ХОЗЯЙСТВО]	0.2028	0.531	0.382	0.704	-0.853	1.258
C(work_field)[Т.ЗДРАВООХРАНЕНИЕ]	0.1984	0.521	0.380	0.704	-0.837	1.234
C(work_field)[Т.ЛЕГКАЯ, ПИЩЕВАЯ ПРОМЫШЛЕННОСТЬ]	0.2914	0.508	0.573	0.568	-0.718	1.301
C(work_field)[Т.НАУКА, КУЛЬТУРА]	-0.0089	0.599	-0.015	0.988	-1.199	1.181
C(work_field)[Т.НЕФТЕГАЗОВАЯ ПРОМЫШЛЕННОСТЬ]	0.4239	0.667	0.635	0.527	-0.901	1.749
C(work_field)[Т.ОБРАЗОВАНИЕ]	-0.2072	0.502	-0.413	0.681	-1.205	0.790
C(work_field)[Т.ОПЕРАЦИИ С НЕДВИЖИМОСТЬЮ]	-0.1328	0.677	-0.196	0.845	-1.477	1.211
C(work_field)[Т.СТРОИТЕЛЬСТВО]	0.2673	0.496	0.539	0.591	-0.718	1.252
C(work_field)[Т.ТОРГОВЛЯ, БЫТОВОЕ ОБСЛУЖИВАНИЕ]	0.0308	0.490	0.063	0.950	-0.942	1.004
C(work_field)[Т.ТРАНСПОРТ, СВЯЗЬ]	0.1508	0.485	0.311	0.757	-0.812	1.114
C(work_field)[Т.ФИНАНСЫ]	0.0566	0.586	0.097	0.923	-1.108	1.221
C(work_field)[Т.ЭНЕРГЕТИЧЕСКАЯ ПРОМЫШЛЕННОСТЬ]	0.2078	0.528	0.394	0.695	-0.840	1.255
C(work_field)[Т.ЮРИСПРУДЕНЦИЯ]	0.6107	0.588	1.039	0.301	-0.556	1.777
gender	0.2725	0.103	2.645	0.010	0.068	0.477
experience	0.0088	0.010	0.895	0.373	-0.011	0.028
I(experience ** 2)	-0.0002	0.000	-1.086	0.280	-0.001	0.000
education	0.3380	0.111	3.034	0.003	0.117	0.559
edu_sub0	0.3141	0.105	2.981	0.004	0.105	0.523

```

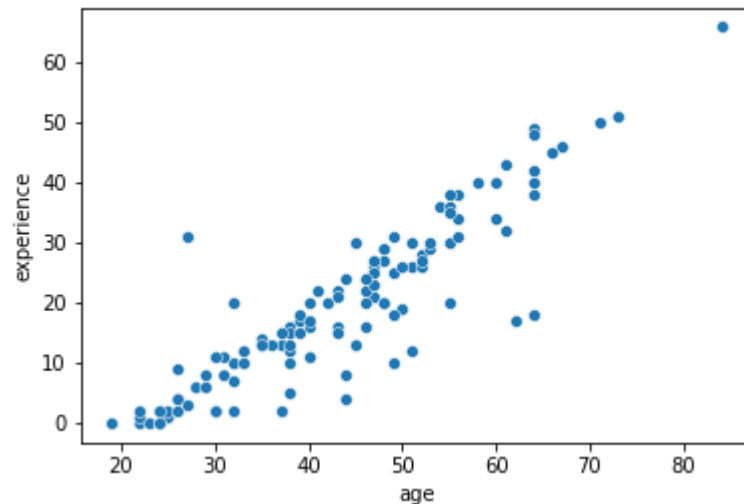
=====
Omnibus:            6.804      Durbin-Watson:          1.726
Prob(Omnibus):      0.033      Jarque-Bera (JB):        9.467
Skew:                0.255      Prob(JB):                0.00879
Kurtosis:            4.304      Cond. No.                 4.48e+04
=====

```

Notes:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 4.48e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Есть все основания полагать, что возраст и опыт нельзя использовать как объясняющие переменные в одном уравнении. Очевидно, они имеют линейную зависимость:



Следующие оценки приведены для модели, где вместо опыта мы включаем возраст агента:

OLS Regression Results						
Dep. Variable:	np.log(salary)	R-squared:	0.455			
Model:	OLS	Adj. R-squared:	0.326			
Method:	Least Squares	F-statistic:	3.525			
Date:	Wed, 21 Dec 2022	Prob (F-statistic):	1.11e-05			
Time:	12:00:37	Log-Likelihood:	-58.031			
No. Observations:	116	AIC:	162.1			
Df Residuals:	93	BIC:	225.4			
Df Model:	22					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	9.3751	0.589	15.928	0.000	8.206	10.544
C(work_field)[Т.АРМИЯ, МВД, ОРГАНЫ БЕЗОПАСНОСТИ]	-0.0875	0.505	-0.173	0.863	-1.091	0.916
C(work_field)[Т.ВОЕННО-ПРОМЫШЛЕННЫЙ КОМПЛЕКС]	0.9168	0.666	1.377	0.172	-0.405	2.239
C(work_field)[Т.ГРАЖДАНСКОЕ МАШИНОСТРОЕНИЕ]	0.2549	0.514	0.496	0.621	-0.766	1.276
C(work_field)[Т.ДРУГАЯ ОТРАСЛЬ ТЯЖЕЛОЙ ПРОМЫШЛЕННОСТИ]	0.3536	0.491	0.720	0.474	-0.622	1.329
C(work_field)[Т.ЖИЛИЩНО-КОММУНАЛЬНОЕ ХОЗЯЙСТВО]	0.2679	0.517	0.518	0.606	-0.760	1.295
C(work_field)[Т.ЗДРАВООХРАНЕНИЕ]	0.3511	0.509	0.689	0.492	-0.660	1.362
C(work_field)[Т.ЛЕГКАЯ, ПИЩЕВАЯ ПРОМЫШЛЕННОСТЬ]	0.3139	0.500	0.628	0.532	-0.679	1.307
C(work_field)[Т.НАУКА, КУЛЬТУРА]	0.1451	0.584	0.249	0.804	-1.014	1.304
C(work_field)[Т.НЕФТЕГАЗОВАЯ ПРОМЫШЛЕННОСТЬ]	0.5088	0.651	0.781	0.437	-0.785	1.802
C(work_field)[Т.ОБРАЗОВАНИЕ]	-0.0250	0.489	-0.051	0.959	-0.996	0.946
C(work_field)[Т.ОПЕРАЦИИ С НЕДВИЖИМОСТЬЮ]	-0.0104	0.661	-0.016	0.987	-1.322	1.302
C(work_field)[Т.СТРОИТЕЛЬСТВО]	0.3664	0.484	0.757	0.451	-0.594	1.327
C(work_field)[Т.ТОРГОВЛЯ, БЫТОВОЕ ОБСЛУЖИВАНИЕ]	0.0908	0.478	0.190	0.850	-0.859	1.041
C(work_field)[Т.ТРАНСПОРТ, СВЯЗЬ]	0.2139	0.475	0.451	0.653	-0.729	1.157
C(work_field)[Т.ФИНАНСЫ]	0.2041	0.571	0.357	0.722	-0.931	1.339
C(work_field)[Т.ЭНЕРГЕТИЧЕСКАЯ ПРОМЫШЛЕННОСТЬ]	0.2830	0.514	0.550	0.583	-0.738	1.304
C(work_field)[Т.ЮРИСПРУДЕНЦИЯ]	0.7551	0.574	1.315	0.192	-0.385	1.895
gender	0.2772	0.100	2.772	0.007	0.079	0.476
age	0.0218	0.020	1.109	0.270	-0.017	0.061
I(age ** 2)	-0.0003	0.000	-1.473	0.144	-0.001	0.000
education	0.3171	0.109	2.918	0.004	0.101	0.533
has_subs	0.3074	0.103	2.980	0.004	0.103	0.512
Omnibus:	7.208	Durbin-Watson:	1.757			
Prob(Omnibus):	0.027	Jarque-Bera (JB):	8.390			
Skew:	0.372	Prob(JB):	0.0151			
Kurtosis:	4.087	Cond. No.	1.16e+05			

Вторая модель учитывает семейное положение у женщин. Таким образом она принимает следующий вид:

$$\ln w = a + \alpha_1 edu + \beta_1 exp_i + \beta_2 exp_i^2 + \gamma gend_i + \delta get\_married + u_i$$

Здесь появилась еще одна переменная – get\_married

Оценки для этой модели:

```
=====
Dep. Variable:      np.log(salary)    R-squared:      0.431
Model:              OLS              Adj. R-squared: 0.289
Method:              Least Squares    F-statistic:    3.029
Date:                Wed, 21 Dec 2022  Prob (F-statistic): 9.03e-05
Time:                05:52:11         Log-Likelihood: -60.503
No. Observations:    116             AIC:            169.0
Df Residuals:        92              BIC:            235.1
Df Model:            23
Covariance Type:     nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
Intercept	9.6968	0.488	19.875	0.000	8.728	10.666
C(work_field)[Т.АРМИЯ, МВД, ОРГАНЫ БЕЗОПАСНОСТИ]	-0.1933	0.527	-0.367	0.715	-1.240	0.853
C(work_field)[Т.ВОЕННО-ПРОМЫШЛЕННЫЙ КОМПЛЕКС]	0.7506	0.689	1.090	0.279	-0.617	2.119
C(work_field)[Т.ГРАЖДАНСКОЕ МАШИНОСТРОЕНИЕ]	0.2193	0.535	0.410	0.683	-0.842	1.281
C(work_field)[Т.ДРУГАЯ ОТРАСЛЬ ТЯЖЕЛОЙ ПРОМЫШЛЕННОСТИ]	0.2801	0.508	0.551	0.583	-0.729	1.289
C(work_field)[Т.ЖИЛИЩНО-КОММУНАЛЬНОЕ ХОЗЯЙСТВО]	0.2140	0.544	0.393	0.695	-0.867	1.295
C(work_field)[Т.ЗДРАВООХРАНЕНИЕ]	0.2083	0.532	0.391	0.696	-0.849	1.265
C(work_field)[Т.ЛЕГКАЯ, ПИЩЕВАЯ ПРОМЫШЛЕННОСТЬ]	0.3001	0.517	0.580	0.563	-0.727	1.327
C(work_field)[Т.НАУКА, КУЛЬТУРА]	-0.0036	0.605	-0.006	0.995	-1.204	1.197
C(work_field)[Т.НЕФТЕГАЗОВАЯ ПРОМЫШЛЕННОСТЬ]	0.4343	0.678	0.641	0.523	-0.912	1.780
C(work_field)[Т.ОБРАЗОВАНИЕ]	-0.2001	0.509	-0.393	0.695	-1.211	0.811
C(work_field)[Т.ОПЕРАЦИИ С НЕДВИЖИМОСТЬЮ]	-0.1219	0.688	-0.177	0.860	-1.488	1.244
C(work_field)[Т.СТРОИТЕЛЬСТВО]	0.2773	0.507	0.547	0.586	-0.730	1.285
C(work_field)[Т.ТОРГОВЛЯ, БЫТОВОЕ ОБСЛУЖИВАНИЕ]	0.0390	0.498	0.078	0.938	-0.951	1.029
C(work_field)[Т.ТРАНСПОРТ, СВЯЗЬ]	0.1604	0.495	0.324	0.747	-0.824	1.144
C(work_field)[Т.ФИНАНСЫ]	0.0671	0.597	0.112	0.911	-1.119	1.253
C(work_field)[Т.ЭНЕРГЕТИЧЕСКАЯ ПРОМЫШЛЕННОСТЬ]	0.2155	0.535	0.403	0.688	-0.847	1.278
C(work_field)[Т.ЮРИСПРУДЕНЦИЯ]	0.6210	0.598	1.038	0.302	-0.567	1.809
gender	0.2732	0.104	2.632	0.010	0.067	0.479
experience	0.0090	0.010	0.897	0.372	-0.011	0.029
I(experience ** 2)	-0.0002	0.000	-1.085	0.281	-0.001	0.000
marriage_status	-0.0118	0.108	-0.109	0.914	-0.227	0.204
education	0.3390	0.112	3.017	0.003	0.116	0.562
has_subs	0.3139	0.106	2.962	0.004	0.103	0.524

```
=====
Omnibus:      6.821    Durbin-Watson:      1.725
Prob(Omnibus): 0.033    Jarque-Bera (JB):      9.561
Skew:         0.252    Prob(JB):      0.00839
Kurtosis:     4.313    Cond. No.      4.53e+04
=====
```

Видим, что к-во детей все же почти не повлияло на результаты, хотя это кажется достаточно весомым аргументом. Также на результаты почти не влияет добавление и остальных переменных(таких как уровень здоровья, семейное положение и проч.), что говорит о достаточности построенной модели. При переменной gender получаем оценку коэффициента на уровне 0.26. Это говорит о том, что средняя зарплата женщин меньше на 26% при прочих равных. Этот коэффициент значим на 1% уровне значимости.

## Проверка на мультиколлинеарность:

Наличие мультиколлинеарности мы проверяли при помощи vif-теста

## Для первой и второй модели:

-----VIF-тест на мультиколлинеарность-----

```
C(work_field)[Т.АРМИЯ, МВД, ОРГАНЫ БЕЗОПАСНОСТИ]: 6.2189
C(work_field)[Т.ВОЕННО-ПРОМЫШЛЕННЫЙ КОМПЛЕКС]: 2.2447
C(work_field)[Т.ГРАЖДАНСКОЕ МАШИНОСТРОЕНИЕ]: 5.1649
C(work_field)[Т.ДРУГАЯ ОТРАСЛЬ ТЯЖЕЛОЙ ПРОМЫШЛЕННОСТИ]: 11.1023
C(work_field)[Т.ЖИЛИЩНО-КОММУНАЛЬНОЕ ХОЗЯЙСТВО]: 5.2619
C(work_field)[Т.ЗДРАВООХРАНЕНИЕ]: 10.8874
C(work_field)[Т.ЛЕГКАЯ, ПИЩЕВАЯ ПРОМЫШЛЕННОСТЬ]: 8.2007
C(work_field)[Т.НАУКА, КУЛЬТУРА]: 3.4063
C(work_field)[Т.НЕФТЕГАЗОВАЯ ПРОМЫШЛЕННОСТЬ]: 2.1304
C(work_field)[Т.ОБРАЗОВАНИЕ] : 13.0892
C(work_field)[Т.ОПЕРАЦИИ С НЕДВИЖИМОСТЬЮ]: 2.1910
C(work_field)[Т.СТРОИТЕЛЬСТВО]: 11.8226
C(work_field)[Т.ТОРГОВЛЯ, БЫТОВОЕ ОБСЛУЖИВАНИЕ]: 23.3676
C(work_field)[Т.ТРАНСПОРТ, СВЯЗЬ]: 13.9652
C(work_field)[Т.ФИНАНСЫ] : 3.2599
C(work_field)[Т.ЭНЕРГЕТИЧЕСКАЯ ПРОМЫШЛЕННОСТЬ]: 5.1841
C(work_field)[Т.ЮРИСПРУДЕНЦИЯ]: 3.2730
gender : 1.4776
experience : 10.6011
I(experience ** 2) : 10.9950
education : 1.6360
has_subs : 1.2879
```

-----VIF-тест на мультиколлинеарность-----

```
C(work_field)[Т.АРМИЯ, МВД, ОРГАНЫ БЕЗОПАСНОСТИ]: 6.1532
C(work_field)[Т.ВОЕННО-ПРОМЫШЛЕННЫЙ КОМПЛЕКС]: 2.2123
C(work_field)[Т.ГРАЖДАНСКОЕ МАШИНОСТРОЕНИЕ]: 5.1401
C(work_field)[Т.ДРУГАЯ ОТРАСЛЬ ТЯЖЕЛОЙ ПРОМЫШЛЕННОСТИ]: 11.1107
C(work_field)[Т.ЖИЛИЩНО-КОММУНАЛЬНОЕ ХОЗЯЙСТВО]: 5.2048
C(work_field)[Т.ЗДРАВООХРАНЕНИЕ]: 10.8393
C(work_field)[Т.ЛЕГКАЯ, ПИЩЕВАЯ ПРОМЫШЛЕННОСТЬ]: 8.2830
C(work_field)[Т.НАУКА, КУЛЬТУРА]: 3.3702
C(work_field)[Т.НЕФТЕГАЗОВАЯ ПРОМЫШЛЕННОСТЬ]: 2.1178
C(work_field)[Т.ОБРАЗОВАНИЕ] : 12.9460
C(work_field)[Т.ОПЕРАЦИИ С НЕДВИЖИМОСТЬЮ]: 2.1784
C(work_field)[Т.СТРОИТЕЛЬСТВО]: 11.7321
C(work_field)[Т.ТОРГОВЛЯ, БЫТОВОЕ ОБСЛУЖИВАНИЕ]: 23.2506
C(work_field)[Т.ТРАНСПОРТ, СВЯЗЬ]: 13.9645
C(work_field)[Т.ФИНАНСЫ] : 3.2306
C(work_field)[Т.ЭНЕРГЕТИЧЕСКАЯ ПРОМЫШЛЕННОСТЬ]: 5.1431
C(work_field)[Т.ЮРИСПРУДЕНЦИЯ]: 3.2614
gender : 1.4521
age : 39.8178
I(age ** 2) : 41.1439
education : 1.6241
has_subs : 1.2876
```

## Для третьей модели и модели со всеми возможными переменными:

-----VIF-тест на мультиколлинеарность-----

```
C(work_field)[Т.АРМИЯ, МВД, ОРГАНЫ БЕЗОПАСНОСТИ]: 6.3890
C(work_field)[Т.ВОЕННО-ПРОМЫШЛЕННЫЙ КОМПЛЕКС]: 2.2555
C(work_field)[Т.ГРАЖДАНСКОЕ МАШИНОСТРОЕНИЕ]: 5.3080
C(work_field)[Т.ДРУГАЯ ОТРАСЛЬ ТЯЖЕЛОЙ ПРОМЫШЛЕННОСТИ]: 11.2832
C(work_field)[Т.ЖИЛИЩНО-КОММУНАЛЬНОЕ ХОЗЯЙСТВО]: 5.4631
C(work_field)[Т.ЗДРАВООХРАНЕНИЕ]: 11.3583
C(work_field)[Т.ЛЕГКАЯ, ПИЩЕВАЯ ПРОМЫШЛЕННОСТЬ]: 8.4465
C(work_field)[Т.НАУКА, КУЛЬТУРА]: 3.4306
C(work_field)[Т.НЕФТЕГАЗОВАЯ ПРОМЫШЛЕННОСТЬ]: 2.1788
C(work_field)[Т.ОБРАЗОВАНИЕ] : 13.3138
C(work_field)[Т.ОПЕРАЦИИ С НЕДВИЖИМОСТЬЮ]: 2.2404
C(work_field)[Т.СТРОИТЕЛЬСТВО]: 12.2299
C(work_field)[Т.ТОРГОВЛЯ, БЫТОВОЕ ОБСЛУЖИВАНИЕ]: 23.9257
C(work_field)[Т.ТРАНСПОРТ, СВЯЗЬ]: 14.4347
C(work_field)[Т.ФИНАНСЫ] : 3.3635
C(work_field)[Т.ЭНЕРГЕТИЧЕСКАЯ ПРОМЫШЛЕННОСТЬ]: 5.2947
C(work_field)[Т.ЮРИСПРУДЕНЦИЯ]: 3.3591
gender : 1.5086
experience : 10.7659
I(experience ** 2) : 11.0809
marriage_status : 1.1924
children : 1.4812
education : 1.7833
has_subs : 1.2895
```

-----VIF-тест на мультиколлинеарность-----

```
C(work_field)[Т.АРМИЯ, МВД, ОРГАНЫ БЕЗОПАСНОСТИ]: 6.6920
C(work_field)[Т.ВОЕННО-ПРОМЫШЛЕННЫЙ КОМПЛЕКС]: 2.3390
C(work_field)[Т.ГРАЖДАНСКОЕ МАШИНОСТРОЕНИЕ]: 5.4385
C(work_field)[Т.ДРУГАЯ ОТРАСЛЬ ТЯЖЕЛОЙ ПРОМЫШЛЕННОСТИ]: 11.4732
C(work_field)[Т.ЖИЛИЩНО-КОММУНАЛЬНОЕ ХОЗЯЙСТВО]: 5.6487
C(work_field)[Т.ЗДРАВООХРАНЕНИЕ]: 11.6941
C(work_field)[Т.ЛЕГКАЯ, ПИЩЕВАЯ ПРОМЫШЛЕННОСТЬ]: 8.6933
C(work_field)[Т.НАУКА, КУЛЬТУРА]: 3.6614
C(work_field)[Т.НЕФТЕГАЗОВАЯ ПРОМЫШЛЕННОСТЬ]: 2.2855
C(work_field)[Т.ОБРАЗОВАНИЕ] : 13.9375
C(work_field)[Т.ОПЕРАЦИИ С НЕДВИЖИМОСТЬЮ]: 2.3436
C(work_field)[Т.СТРОИТЕЛЬСТВО]: 12.7602
C(work_field)[Т.ТОРГОВЛЯ, БЫТОВОЕ ОБСЛУЖИВАНИЕ]: 24.8113
C(work_field)[Т.ТРАНСПОРТ, СВЯЗЬ]: 15.2333
C(work_field)[Т.ФИНАНСЫ] : 3.5488
C(work_field)[Т.ЭНЕРГЕТИЧЕСКАЯ ПРОМЫШЛЕННОСТЬ]: 5.4330
C(work_field)[Т.ЮРИСПРУДЕНЦИЯ]: 3.5582
gender : 1.6194
experience : 11.5503
I(experience ** 2) : 11.6152
marriage_status : 1.2363
children : 1.5547
health_problems : 1.5007
health_lvl : 1.4915
education : 1.9371
workweek : 1.5638
has_subs : 1.3252
count_subs : 1.1690
```

Как видим, каждый тест дает результаты, говорящие о наличии мультиколлинеарности.

Однако, эти результаты связаны с присутствующими функциями от опыта или возраста, что обязательно должно было присутствовать. Поэтому, в целом, будем считать что мультиколлинеарности нет.

## Проверка условий Гаусса-Маркова

Проверка гетероскедастичности:

Ошибки на гомоскедастичность проверены с помощью теста Бройша-Пагана.

**Нулевая гипотеза ( $H_0$ ):** присутствует гомоскедастичность.

**Альтернативная гипотеза: ( $H_a$ ):** гомоскедастичность *отсутствует* (т.е. гетероскедастичность существует)

Результат теста (значения  $f$  p-value):

1) 0.833

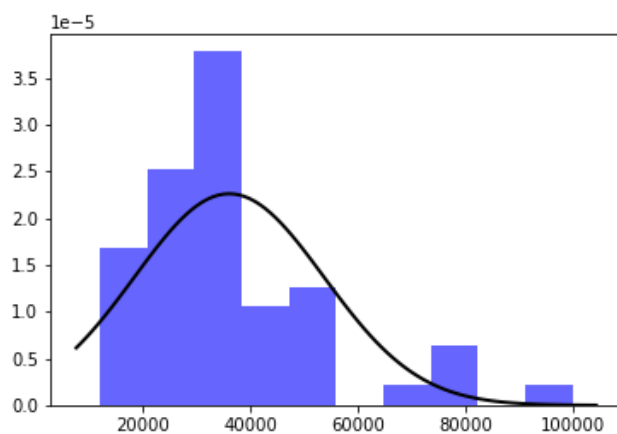
2) 0.822

3) 0.853744

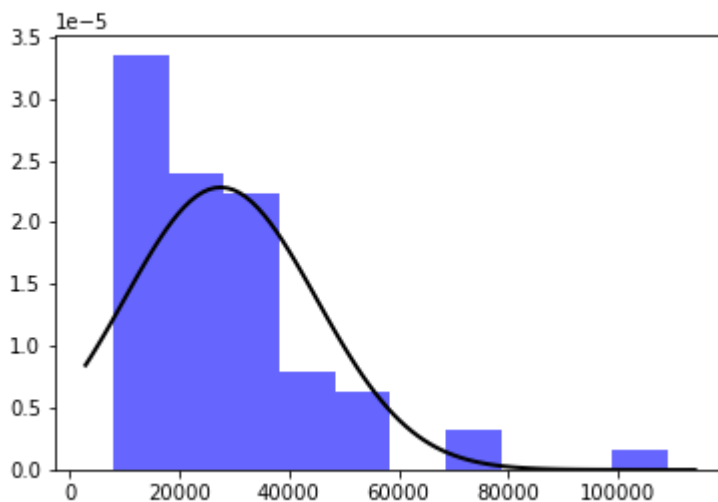
Нулевая гипотеза не отвергнута, т.к. все значения не меньше 0.05

**Проверка на распределение:**

Распределение зарплат по мужчинам:



и распределение зарплат по женщинам:

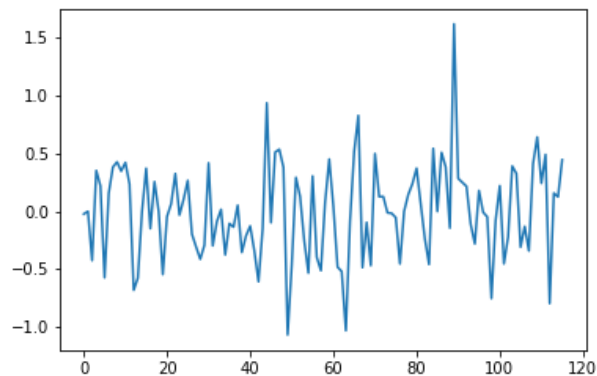




Как видим, распределение зарплат у мужчин больше похоже на нормальное, чем у женщин. Это объясняется тем, что в нашей выборке мужчины затрагивают больше профессий. Большее к-во данных дало бы лучшие результаты.

### Проверка матожидания ошибок

Построим динамику ошибок по имеющимся точкам. Видим, что они колеблются около нуля, поэтому считаем, что матожидание ошибки примерно ноль.



### Выводы

Были построены три модели для изучения разрыва заработных плат между мужчинами и женщинами. Полученные результаты говорят о разрыве в 27% (в среднем по всем моделям)), при этом мужчины получают больше зарплату, что является значимым результатом. Наилучшей моделью оказалась первая, т.к. она дает наилучшие результаты по vif-тесту и оценки для нее могут быть отвергнуты на меньших уровнях значимости.