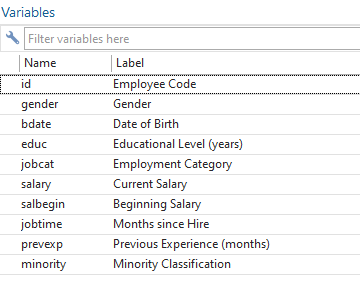
**ФИО студента(ов) Гаврилина А., Глушкова Д.**

**Откройте файл Работники\_предприятия.dta и выполните задания.**

Проведите множественный линейный регрессионный анализ с использованием переменных базы данных. Зависимая переменная - **salary**. Выберите не менее 3-ех предикторов (независимых переменных).

Изучим список переменных и данные:

****

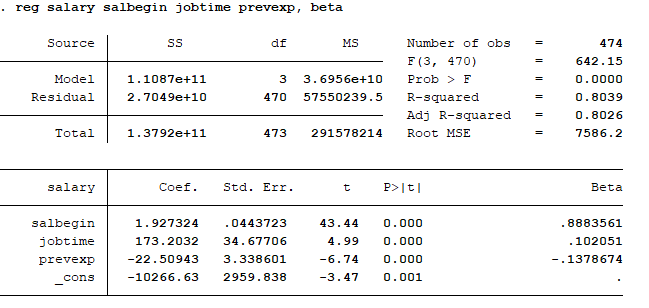
****

Зависимая переменная - salary *(зарплата в настоящий момент)*, в качестве предикторов (независимых переменных) возьмем salbegin *(начальная зарплата)*, jobtime *(время с момента найма, в месяцах)*, prevexp *(прошлый опыт в месяцах)*

**Ø Напишите уравнение регрессии.**

Рассчитаем со стандартизированными коэффициентами

reg salary salbegin jobtime prevexp, beta



Уравнение регрессии:

P>|t| < 0.05 для всех коэффициентов:  
Коэффициенты значимы.

**Ø Оцените значимость регрессионной модели.**

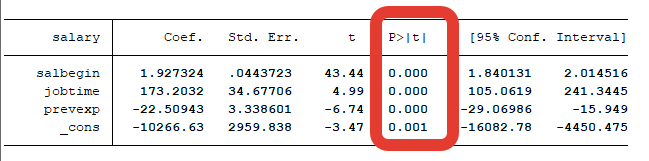
Prob > F = 0.0000: следовательно, **модель значима**

**Ø** **Чему равен процент дисперсии зависимой переменной, объяснённый независимыми переменными?**

**Это можно определить по значению R-squared: коэффициент детерминации** (*R-квадрат*) — это доля [дисперсии](https://ru.wikipedia.org/wiki/%D0%94%D0%B8%D1%81%D0%BF%D0%B5%D1%80%D1%81%D0%B8%D1%8F_%D1%81%D0%BB%D1%83%D1%87%D0%B0%D0%B9%D0%BD%D0%BE%D0%B9_%D0%B2%D0%B5%D0%BB%D0%B8%D1%87%D0%B8%D0%BD%D1%8B) зависимой переменной, объясняемая рассматриваемой [модель](https://ru.wikipedia.org/wiki/%D0%9C%D0%BE%D0%B4%D0%B5%D0%BB%D1%8C)ю зависимости, то есть объясняющими переменными. Чем ближе значение коэффициента к 1, тем сильнее зависимость. При оценке регрессионных моделей это интерпретируется как соответствие модели данным. Для приемлемых моделей предполагается, что коэффициент детерминации должен быть хотя бы не меньше 0.5. Модели с коэффициентом выше 0.8 можно признать достаточно хорошими.

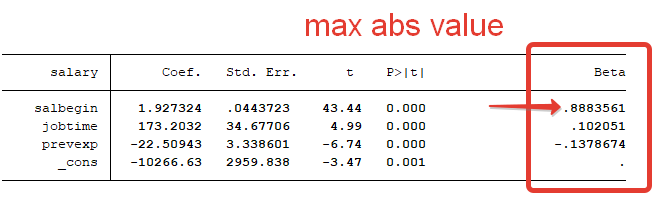
В нашем случае 0.8039 ≈ **80%**, можно сказать, что **модель достаточно хорошая.**

**Ø Какие предикторы статистически значимо взаимосвязаны с зависимой переменной?**



Все < 0.05: следовательно, все значимо взаимосвязаны.

**Ø Какая из независимых переменных наиболее сильно взаимосвязана с зависимой переменной?**

****

Смотрим на максимальное абсолютное значение: 0.8883561 для переменной salbegin. Значит, наиболее сильно взаимосвязана с зависимой переменной salary (зарплата в настоящий момент) независимая переменная salbegin (начальная зарплата).

**Ø Подробно интерпретируйте особенности взаимосвязи между одним из предикторов и зависимой переменной.**

Увеличение времени работы (jobtime) на 1 месяц дает в среднем прирост зарплаты (salary) на 173.2. *(единицы измерения зарплаты не указаны)*

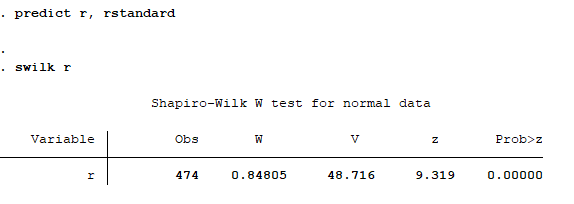
**Ø Проведите диагностику модели.**

**o Проверьте остатки на нормальность;**

Сохранение стандартизированных остатков и их проверка на нормальность с помощью теста Шапиро-Уилка:

predict r, rstandard

swilk r

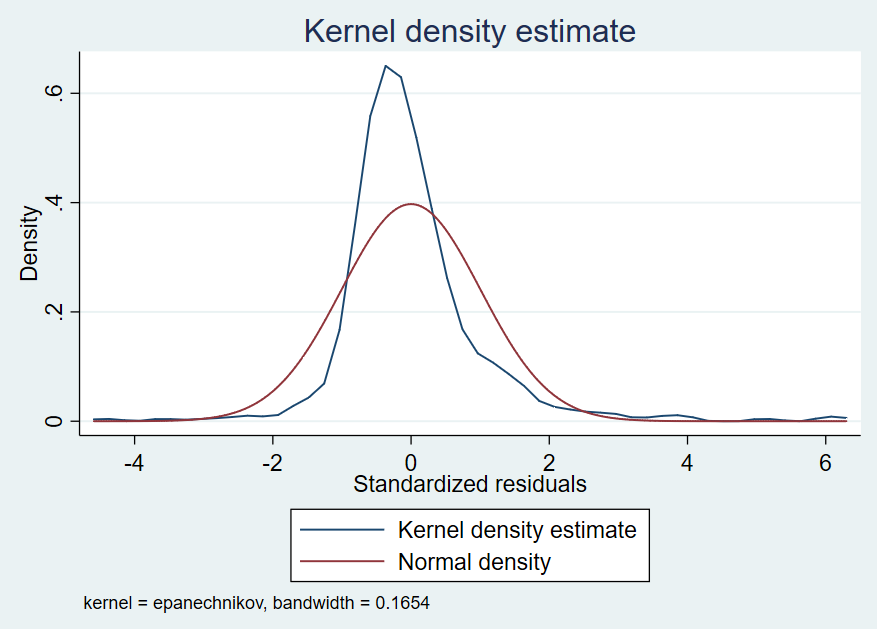


Prob>z: 0.00000

Значимость менее 0.05, следовательно, данное распределение **значимо отличается от нормального**.

График:

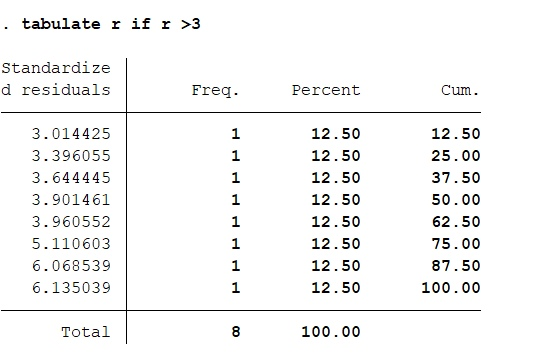
kdensity r, normal



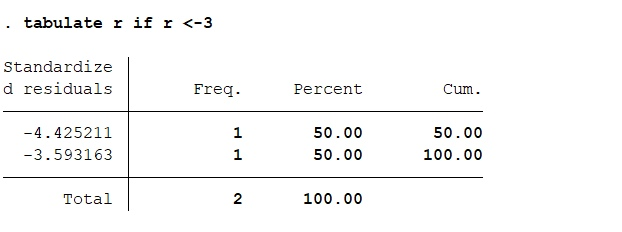
**o Выявите выбросы или влияющие наблюдения.**

Выбросы - результаты измерения, выделяющиеся из общей выборки. Если стандартизированный остаток (r) по модулю больше 3 (*правило 3-х сигм*), то значение является выбросом

tabulate r if r >3



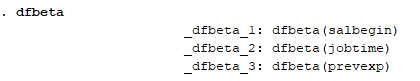
tabulate r if r <-3



В нашей модели 10 выбросов.

Определим, есть ли работники, чьи данные искажают результаты. Для этого произведем расчет статистик влияния:

dfbeta



Граница определения, является переменная влияющей или нет - это единица по модулю. Переменная является влияющей, если значение хотя бы одного из коэффициентов пересекает границу.

Однако, если значения коэффициента примерно равны -1 или 1, их можно не исключать, потому что в таком случае они считаются влияющими, но не сильно.

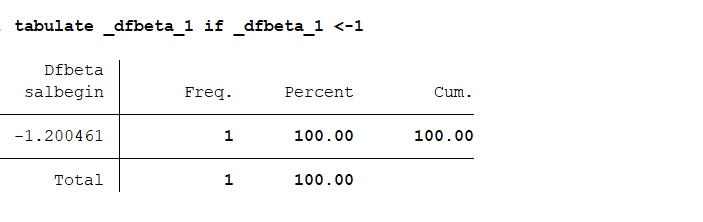
Проверим, есть ли у нас влияющие наблюдения

tabulate \_dfbeta\_1 if \_dfbeta\_1 >1

*таких наблюдений нет*

tabulate \_dfbeta\_1 if \_dfbeta\_1 <-1

*1 наблюдение*



tabulate \_dfbeta\_2 if \_dfbeta\_2 >1

*таких наблюдений нет*

tabulate \_dfbeta\_2 if \_dfbeta\_2 <-1

*таких наблюдений нет*

tabulate \_dfbeta\_2 if \_dfbeta\_2 >1

*таких наблюдений нет*

tabulate \_dfbeta\_3 if \_dfbeta\_3 <-1

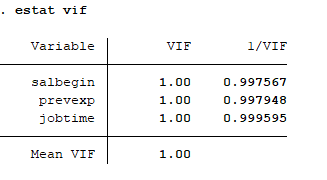
*таких наблюдений нет*

Таким образом, мы получили одно наблюдение, у которого коэффициент влияния больше 1 по модулю. Однако он равен -1,2, поэтому мы можем диагностировать это наблюдение как выброс и не исключать его, так как значение -1,2 почти не отличается от пограничного.

**o Для проверки мультиколлинеарности используйте значения показателя VIF и корреляционной матрицы.**

Вычислим значение variance inflation factor (VIF) – меры мультиколлинеарности.

estat vif

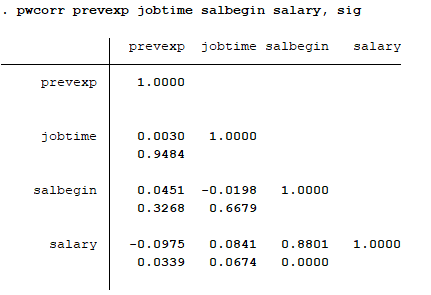
**

Считается, что мультиколлинеарность существует, если общее значение VIF >5. Если оно на уровне 1-5, то мультиколлинеарности нет.

Чтобы убедиться в этом, проведем еще один тест, а именно - оценим корреляцию. Это поможет нам сделать вывод о наличии или отсутствии мультиколлинеарности.

Матрица:

pwcorr prevexp jobtime salbegin salary, sig

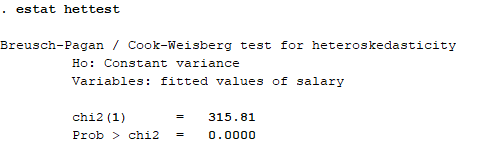


Из корреляционной матрицы видим, что все коэффициенты корреляции очень низкие. Может смело сделать вывод о том, что мультиколлинеарности нет.

**o Оцените гетероскедастичность.**

Проявляется в неоднородности наблюдений, выражающейся в неодинаковой дисперсии случайной ошибки регрессионной модели. Проверим гипотезу H0 о том, что дисперсия остатков при разных значениях зависимой переменной (salary) стабильна

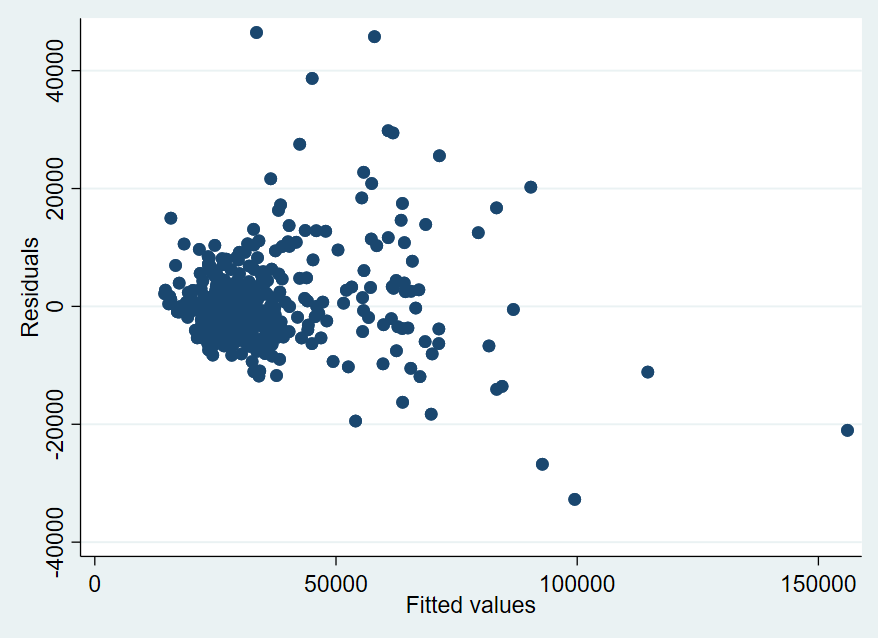
estat hettest



Prob > chi2 = 0.0000:  
Вероятность, что мы ошибемся, отклонив H0 и приняв альтернативную гипотезу, меньше 0.05. Поэтому, у нас **гетероскедастичность**.

Построим rvf plot (residual-versus-fitted plot)

rvfplot



Если бы было случайное облако точек, без скоплений, мы бы могли сказать, что нет взаимосвязи (*гомоскедастичность*). Здесь имеются скопления, выраженные взаимосвязи (*гетероскедастичность*).

Сохраните результаты в этом файле и загрузите в MS Teams. Если задание выполнялось в группе, то укажите фамилии всех членов группы.