**ФИО слушателя Гаврилина А., Глушкова Д.**

**Откройте файл Burnout.dta и выполните задания.**

Зависимая переменная: **Burnout** *Выгорание на работе.*

Независимые переменные:

Ø **Perceived Control** *Ощущаемый контроль (чем больше значение, тем меньше ощущение контроля над ситуацией)*

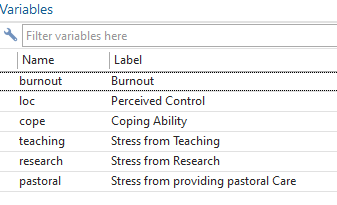
Ø **Coping Ability** *Способность справляться со сложностями (чем больше значение, тем меньше способность)*

Ø **Stress from Teaching** *Стресс от преподавания (чем выше значение, тем выше уровень стресса)*

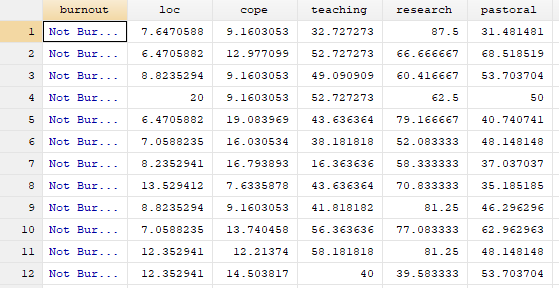
Ø **Stress from Research** *Стресс от проведения исследований (чем выше значение, тем выше уровень стресса)*

Ø **Stress from providing pastoral Care** *Стресс от бескорыстной помощи коллегам и студентам (чем выше значение, тем выше уровень стресса)*

Список переменных:



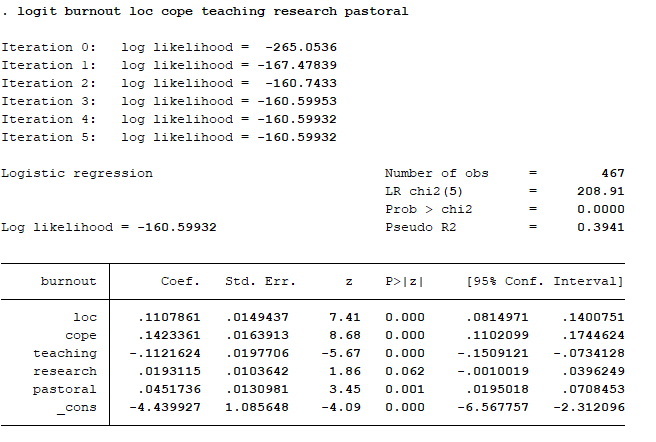
Как выглядят данные:



**Ø Постройте модель бинарной логистической регрессии с использованием переменных базы данных.**

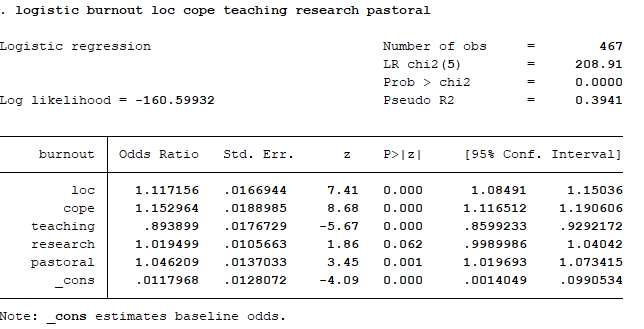
// модель с коэффициентами

logit burnout loc cope teaching research pastoral



// модель, которая позволит нам посмотреть на отношения шансов (odds ratio)

logistic burnout loc cope teaching research pastoral



Значения odds ratio больше 1 при следующих значимых переменных:

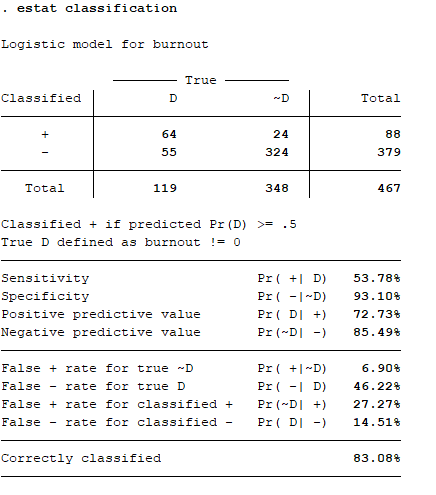
loc, cope, pastoral. Значит, мы можем говорить о том, что

* увеличение значения loc (ощущаемый контроль) на 1 приведет к росту шанса выгореть в 1,117 раз;
* увеличение значения cope (означает степень неспособности справляться со сложностями) на 1 приведет к росту шанса выгореть в 1,15 раз;
* увеличение значения pastoral (стресс от бескорыстной помощи) на 1 приведет к росту шанса выгореть в 1,04 раз.

**Ø Оцените значимость и предсказательную силу регрессионной модели.**

Prob > chi2 = 0.0000

Это говорит о том, что модель **значима** и мы можем принять нашу гипотезу H0 *об отсутствии различий между предсказанными нашей моделью значениями и реально наблюдаемыми.*

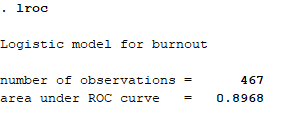


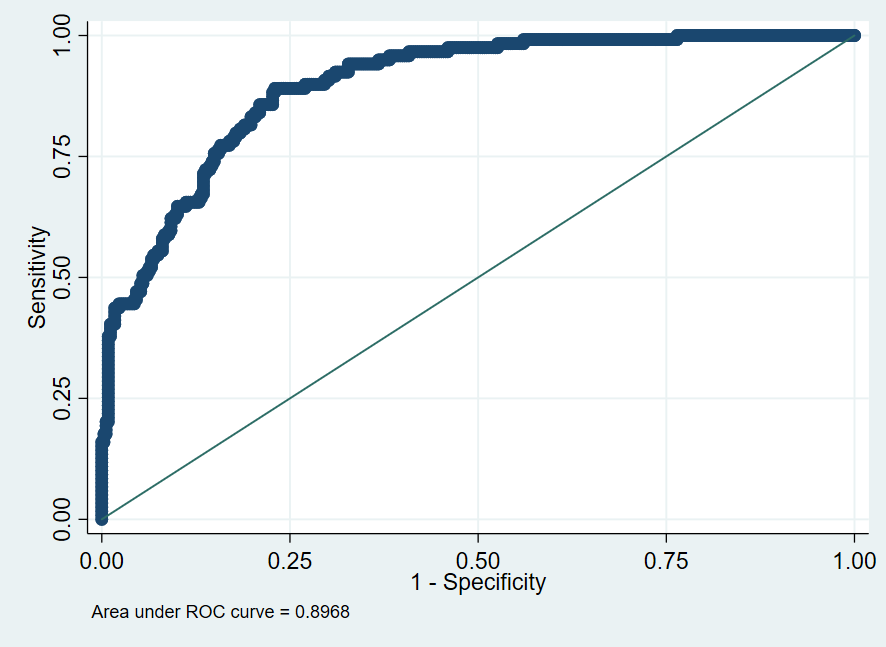
**Верно предсказано 83% случаев.**

График ROC-кривой

**ROC-кривая** (receiver operating characteristic) — график, позволяющий оценить качество [бинарной классификации](https://ru.wikipedia.org/wiki/%D0%91%D0%B8%D0%BD%D0%B0%D1%80%D0%BD%D0%B0%D1%8F_%D0%BA%D0%BB%D0%B0%D1%81%D1%81%D0%B8%D1%84%D0%B8%D0%BA%D0%B0%D1%86%D0%B8%D1%8F). Так же известна как кривая ошибок.

Количественную интерпретацию ROC даёт показатель ***AUC*** (*area under ROC curve*, *площадь под ROC-кривой*) — **площадь**, ограниченная ROC-кривой и осью доли ложных положительных классификаций. Чем выше показатель AUC, тем качественнее классификатор, при этом значение 0,5 демонстрирует непригодность выбранного метода классификации (соответствует случайному гаданию).





Площадь **0.8968**, можно сказать, что качество классификации **хорошее**.

**Ø Напишите уравнение регрессии.**

Вероятность наступления события рассчитывается по формуле

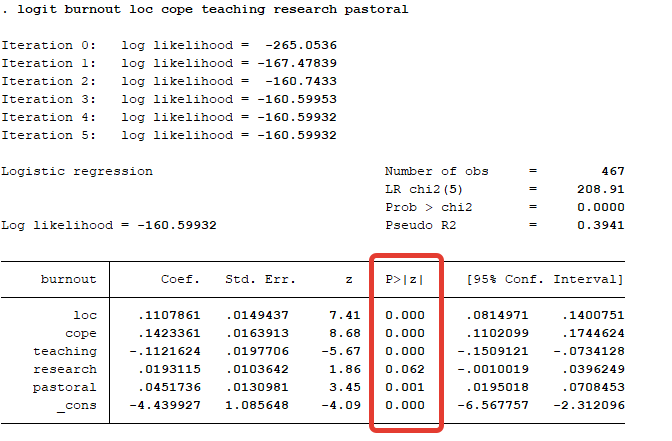
- это значения независимых переменных,

- коэффициенты, расчет которых является задачей бинарной логистической регрессии,

- константа.

В нашем случае **уравнение регрессии** имеет вид:

**Ø Является ли статистически значимым влияние каждой из независимых переменных на зависимую?**

****

Изучим столбец значимости: параметр research имеет (P > |z|) > 0.05, поэтому **не является значимым, остальные параметры значимые.**

**Ø Интерпретируйте влияние одной из независимых переменных на зависимую.**

При уменьшении ощущения контроля над ситуацией увеличивается вероятность выгорания (коэффициент 0.11).

При ослаблении способности справляться со сложностями увеличивается вероятность выгорания (коэффициент 0.14).

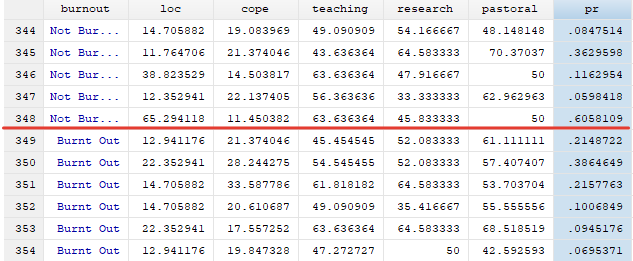
**Ø Проведите диагностику модели**

**o Проверьте остатки на нормальность**

Сохраняем вероятности



Фрагмент таблицы после сохранения вероятностей:



Сохраняем остатки

predict r, rstandard

Оценка нормальности

*Гипотеза H0: распределение значений переменной не значимо отличается от нормального распределения*

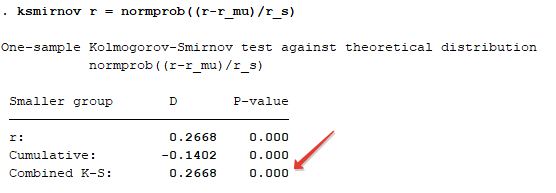
*Гипотеза H1: распределение значений переменной значимо отличается от нормального распределения*

Воспользуемся **тестом Колмогорова-Смирнова**, т.к. выборка больше 100.

egen r\_mu = mean(r)

egen r\_s = sd(r)

ksmirnov r = normprob((r-r\_mu)/r\_s)



В результате получена значимость менее 0.05, следовательно, данное распределение **значимо отличается от нормального**.

**График**

kdensity r, normal

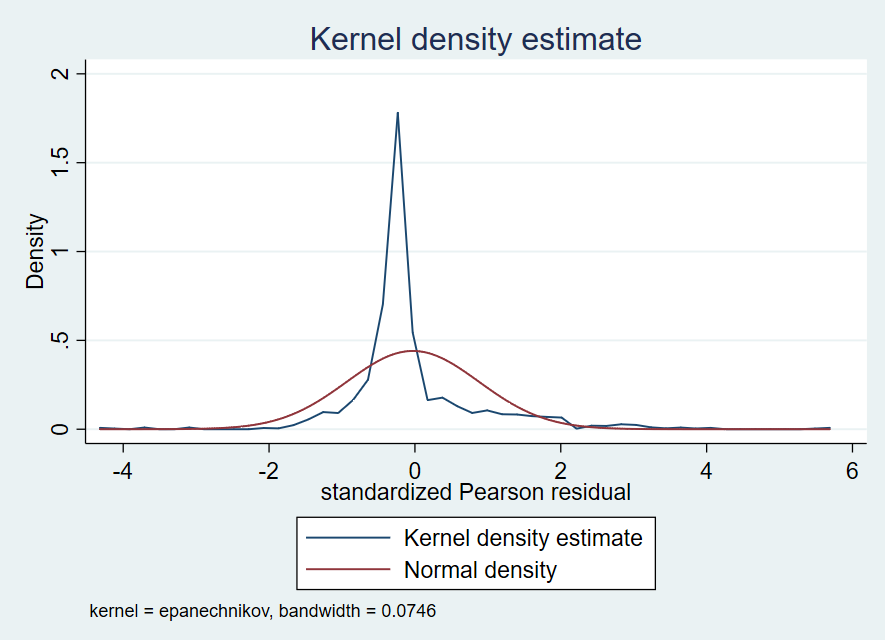
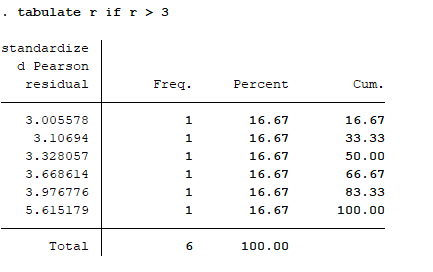


График подтверждает проведенный тест на нормальность

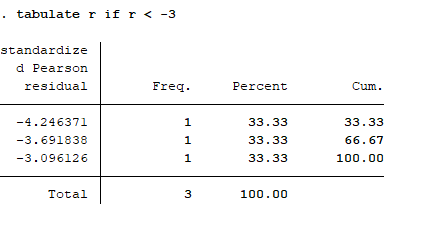
**o Выявите выбросы или влияющие наблюдения.**

Выброс – это наблюдение, которое не вписывается в модель. Выбросы ищем по остаткам. Стандартизированные остатки должны быть от -3 до +3 *(правило 3-х сигм)*. Смотрим на те остатки, которые по модулю превышают 3, это потенциальные выбросы.

tabulate r if r > 3



tabulate r if r < -3

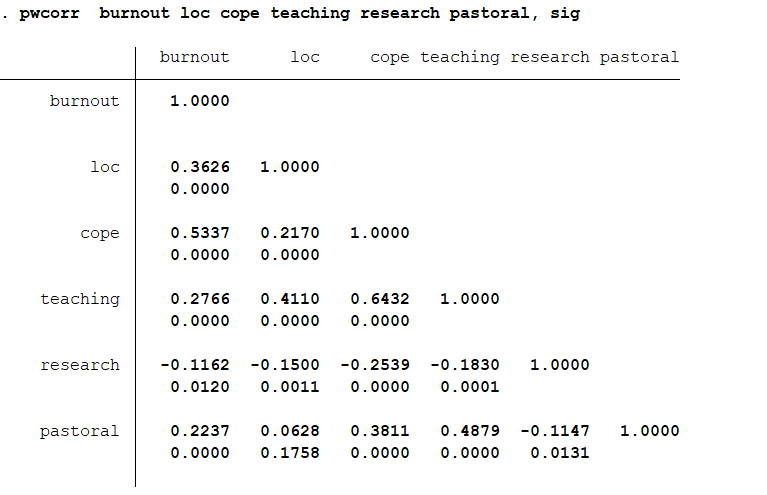


В нашей модели 9 выбросов.

**o Оцените мультиколлинеарность.**

Оценим с помощью корреляционной матрицы

pwcorr burnout loc cope teaching research pastoral, sig



Видим, что коэффициенты корреляции во всех случаях, кроме двух, меньше 0,5, что говорит о том, что корреляция переменных между собой, в основном, слабая.

Тем не менее, мы не можем утверждать, что у нас нет мультиколлинеарности, так как коэффициент корреляции cope и burnout = 0,5337 (говорит о том, что есть средняя корреляция), коэффициент корреляции teaching и cope = 0,6432 (также указывает на среднюю корреляцию).

Загрузите решение в MS Teams. При выполнении задания в группе укажите фамилии всех членов групп

\_\_\_

Когда делаем estat vif, пишет “estat vif not valid” (исправить не удалось)



\_\_\_