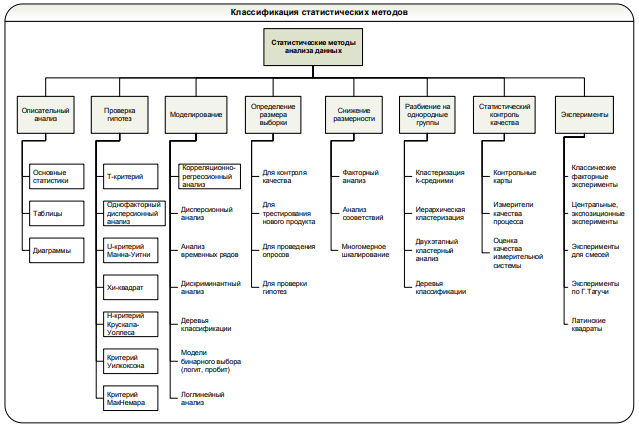
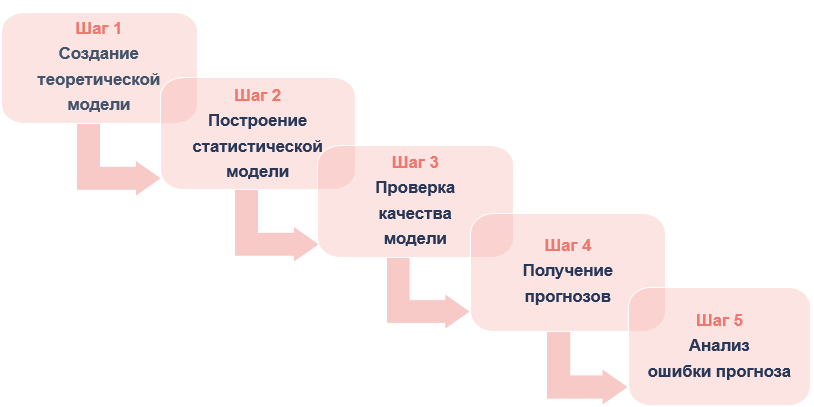
**Взаимосвязь факторов**

**Общая схема методов**



**Схема для регрессионного анализа**



На основании представлений о природе исследуемой проблемы, интуиции специалиста или опыта аналогичных исследований осуществляется выбор отклика и генерация возможных факторов. (тут пообсуждать несколько примеров-задач. В конце - написал список задачек. Если на лекции, то важно, чтобы каждый студент сам поупражнялся. На само деле – не всегда понятна, какая целевая лучше и в каких данных её лучше фикисровать)

Выбор фактора и отклика

Далее – необходимо произвести анализ мультиколлинеарности факторов. Мы могли «сгоряча» нагенерить всякого, наверняка что-то и лишнее попалось)… Факторы называются мультиколлинеарными, если между ними наблюдается достаточно сильная корреляционная связь (коэффициент корреляции 0,9 и более). Эта проблема затрудняет ранжирование факторов по степени влияния на отклик. Мультиколлинеарность определяется либо по корреляционной таблице, либо с помощью коэффициента роста дисперсии (КРД или VIF): если КРД >5, то принято считать, что мультиколлинеарность присутствует. Рекомендуется удалить из модели мультиколлинеарные факторы (один из пары), если это не принципиально для решения поставленной задачи.

Поиск мультиколлинеарных факторов

Из двух мультиколлинеарных факторов из анализа исключается тот, который меньше оказывает влияния на целевую (отклик). Он должен покинуть «наш праздник») (тут важное обсуждение ка это сделать.)

Изучение относительной важности мультиколлинерных факторов

Анализ остатков

Следующий этап регрессионного анализа - очень важный. Это - анализ остатков. Он помогает понять – а хороша ли наша модель (Замечу, речь не идёт ещё о точности модели, она будет разбираться далее). Остатки представляют собой разности фактических значений отклика и значений, предсказанных по уравнению регрессии для одних и тех же факторов. Линейная форма регрессионной модели выбрана верно, если:

- остатки нормально распределены.

- остатки не зависят от предсказанных по уравнению регрессии значений отклика. (тема анализ остатков – рассмотрим отдельно)

(СУПЕР ВАЖНАЯ ТЕМА!!!!!!!!!!!)

Приемлемость модели с точки зрения ее прогнозных качеств определяется с помощью сравнения нашей модели и «наивного» прогноза, т.е. прогноза по среднему (или называют ещё константной моделью).

Оценка приемлемости модели в целом

Анализ регрессионного уравнения и удаление факторов, не влияющих на отклик

Построили уравнение регрессии. Теперь надо понять – факторы которые мы включили в модель – важны? Или на этапе генерации (первый шаг) мы чего-то лишнего «закинули».

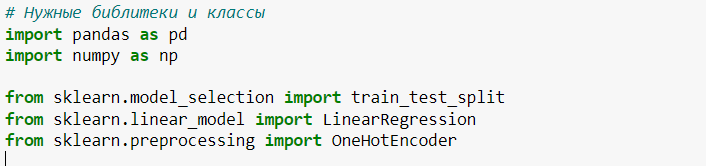
Это делается так: факторы, у которых (р-уровень).> 0,05 , могут быть исключены из анализа, т.е. они несущественно влияют на отклик. После любого исключения весь предшествующий алгоритм анализа нужно повторить.

R2 – коэффициент детерминации, показывает долю изменяемости отклика, происходящую под одновременным воздействием всех включенных в модель факторов. Другими словами, на сколько хорошо факторы объясняют отклик. Чем больше R2 (0≤R2 ≤1), тем выше качество модели. Небольшое значение R2 (менее 0,3) может указывать на неадекватный подбор факторов и говорит о нецелесообразности построения прогнозов по такой модели. Нужно возвращаться на этап генерации факторов модели и искать дополнительные.

Построение прогноза

Анализ R2

И, наконец, «вишенка» всего регрессионного анализа – построение прогноза целевой переменной (отклика). Для построения прогноза необходимо ввести прогнозные значения факторов, влияние которых на отклик установлено, и заново провести регрессионный анализ с сохранением предсказанных по уравнению регрессии значений и 95-процентного интервала прогноза. (тема: анализ остатков и интервал прогноза – рассмотрены в других разделах).



**Дополнительная литература**

1. **Документация по библиотекам**

**По линейной регрессии:**

1. **https://bibl.nngasu.ru/electronicresources/uch-metod/economic\_statistics/859984.pdf**
2. [**https://habr.com/ru/articles/514818/**](https://habr.com/ru/articles/514818/)

**Реализация в питоне:**

1. **1. https://www.kaggle.com/code/muzafferdindar/linear-regression-in-python**
2. **2.** [**https://www.kaggle.com/code/emineyetm/simple-linear-regression-using-python**](https://www.kaggle.com/code/emineyetm/simple-linear-regression-using-python)

**По логистической регрессии: https://www.kaggle.com/code/prashant111/logistic-regression-classifier-tutorial**