**Методы анализа данных**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Название** | **Зависимая переменная** | **Допущения** | **Проверка допущений** | **Интерпретация** |
| 1 | МНК  (линейная регрессия) | **Цель**: узнать, есть ли линейная связь между нашими переменным | 1. Линейная взаимосвязь X и Y 2. Нормальное распределение остатков. 3. Нет выбросов 4. Гомоскедастичность – постоянная изменчивость остатков на всех уровнях независимой переменной. 5. Коллинераность | **Допущение 1**    **Допущение** **2**    Допущение 3    Допущение 4    4. Тест Левина Pr(>F) > 0.05 (все хорошо)  5. VIF (variance inflation factors) – тест (в квадрате) должен показать FALSE (все хорошо) | **R2**показывает, сколько процентов нашей выборки объясняется получившейся моделью (меньше 50% не оч).  **H0**: связи между переменными нет.  **T value** > 1.96 - значимо  **P value** < 0.05 – значимо (H0 неверна)  Каждая последующая единица «независимой переменной» в среднем увеличивает «зависимую переменную» на «значение коэффициента независимой переменной» единиц. |
| **Код:**  lm (*dependent var* ~ *indep. var* + *indep. var*, *data*) | **Код:**  1.2.3.4 - plot(*model*)  4. - leveneTest (*var*, center = median)  4. - sqrt(vif(*model*)) > 2  gvlma(*model, data*) – проверит на все! (пакет gvlma) | **Код:**  tab\_model(*model*)  summary(*model*) |
|  | **Название** | **Зависимая переменная** | **Допущения** | **Проверка допущений** | **Интерпретация** |
| 2 | Логит (пробит) регрессия | Биноминальная – принимает лишь два значения.  Не желательно, чтобы одно из значений переменной встречалось меньше, чем 5. Т.е. количество одного значения чересчур мало, а другое огромно.  **Цель**: узнать, что из предикторов увеличивает или уменьшает вероятность наступления нашего исхода | 1. Линейная взаимосвязь X и Y 2. Нормальное распределение остатков 3. Нет выбросов 4. Гомоскедастичность – постоянная изменчивость остатков на всех уровнях независимой переменной. 5. Коллинераность | См. МНК (линейная регрессия)  3. Смотрим на значение «.std.resid», если оно больше 3, то выбросы оказывают значение на нашу переменную, если нет, то все хорошо. | Модели можно сравнивать по значению AIC (чем оно меньше, тем лучше).  Вероятность наступления того, что «значение зависимой переменной равно 1» равно «значению эксп. коэффициента независимой переменной».  **OR**: 1 – это сто процентов, если показатель меньше 1, то вероятность будет равна 1 – это значение. Например, 0.23 – это снижение шансов на 77%. Если больше 1, то увеличивает шансы на количество свыше 1. Например, 3.87 – это увеличение шансов на 287% или в 3 раза.  Можно интерпретировать по значению предельных эффектов (AME): в среднем, при «значении независимой переменной» вероятность «зависимой переменной» уменьшается на «значение AME» %. |
| **Код:**  glm (*dependent var* ~ *indep. var* + *indep. var*, *data*, family = binomial ()) | **Код:**  3. model.data <- augment(*model*) %>%  mutate(index = 1:n())  model.data %>% top\_n(2, .cooksd) | **Код:**  exp(coef(*model*))  tab\_model(*model*)  summary(margins(mod\_good)) |
|  | **Название** | **Зависимая переменная** | **Допущения** | **Проверка допущений** | **Интерпретация** |
| 3 | Порядковая регрессия | Не равноинтервальная: шкала Ликкерта: согласен, не согласен, нет знаю ответа, образование – уровень между образованием может быть разным – начальная, бакалавр, магистратура.  Если мы можем расположить значения от меньшего к большему, от лучшего к худшему. Если нельзя так сделать, то мультиноминальная.  **Цель**: мы хотим, например, предсказать, ответ респондента. | 1. О параллельности прямых (но оно часто нарушается, поэтому можно не учитывать). Однако если есть сильный перекос, то лучше использовать мультиноминальную модель. | 1. Тест Бранта: если probability больше 0.05, то допущение соблюдается | Вероятность перехода от одной категории к следующей категории.  Принадлежность к / Каждая последующая единица «независимой переменной» увеличивает /снижает вероятность перехода из одной категории «зависимой переменной» в последующую. |
| **Код:**  polr(*dependent var* ~ *indep. var* + *indep. var*, *data*, Hess=TRUE) | **Код:**  brant::brant(*model*) | **Код:**  exp(coef(*model*))  tab\_model(*model*) |
| 4 | Мультиноминальная регрессия | Неупорядоченная переменная, если ее значения нельзя расположить от меньшего к большему, от лучшего к худшему. Т.е. это не равноинтервальные. |  |  | Вероятность наступления нашей категории по отношению к любой категории.  Принадлежность к / Каждая последующая единица «независимой переменной» увеличивает /снижает вероятность перехода из одной категории «зависимой переменной» в другую. |
| **Код:**  multinom(*dependent var* ~ *indep. var* + *indep. var*, *data*) | **Код:**  exp(coef(*model*))  tab\_model(*model*) |
|  | **Название** | **Зависимая переменная** | **Допущения** | **Проверка допущений** | **Интерпретация** |
| 5 | Пуассоновская регрессия (частотная) | Количество чего-то там: количество забитых мячей, количество детей, число правонарушений.  Должна быть не отрицательной и не дробной.  **Цель**: узнать, что из предикторов увеличивает или уменьшает вероятность перехода из одного значения зависимой переменной в другую | 1. Овердесперсия – дисперсия не должна быть очень большой (можно сравнить со средним). Тогда использовать negative binominal model. 2. Избыточные нули – большой количество нулей в значении – плохо. Тогда использовать zero-inflated model. | 1. Разделить дисперсию на среднее – если больше ну 10, то кошмар. 2. Нарисовать гистограмму зависимой переменной и если там столбик нулей, то плохо. | Вероятность возникновения последующего события, например, вероятность того, что следующий мяч будет забит.  **IRR**: Принадлежность к / Каждая последующая единица «независимой переменной» увеличивает/снижает вероятность того, что и «зависимая переменная» увеличится на единицу на «значение IRR» относительно «предыдущего значения зависимой переменной»  Модели можно сравнивать по значению Log-Likelihood (чем больше, тем лучше) |
| **Код:**  2. zeroinfl(*dependent var* ~ *indep. var* + *indep. var*, *data*) | **Код:**  exp(coef(*model*))  tab\_model(*model*) |
| 6 | Многоуровневая регрессия | Количество чего-то там (как у пуассоновской).  Однако мы не можем оценить важность и значимость принадлежности к группе для каждого члена внутри группы.  **Цель**: узнать является ли принадлежность к группе значимым фактором | 1. Надо проверить, если ли у нас какая-то вариация внутри групп 2. Наблюдений в группе больше, чем 30 3. Линейная взаимосвязь X и Y 4. Нормальное распределение остатков 5. Нет выбросов 6. Гомоскедастичность – постоянная изменчивость остатков на всех уровнях независимой переменной. 7. Коллинераность | См. МНК (линейная регрессия) | **ICC**: – какой процент вариации объясняется межгрупповой. Вклад группы в объяснение, если больше 50, значит вклад группы объясняет 50% выборки. |
|  |  |  |  |