

ПОНИМАНИЕ И ИНТЕРПРЕТАЦИЯ РЕГРЕССИОННОГО АНАЛИЗА Understanding and interpreting regression analysis

Н. С. Пермяков, магистр 1 курса

Н. И. Шингарева, кандидат сельскохозяйственных наук

Уральский государственный аграрный университет

(Екатеринбург, ул. Карла Либкнехта, 42)

Аннотация

Измеряемые в растениеводстве переменные, такие как количество выпавших осадков, дозы минеральных удобрений, сроки и качество посева обычно являются непрерывными. Существуют статистические методы для подтверждения того, действуют ли эти переменные как независимые переменные, вызывающие результат. Большинство из этих методов являются разновидностями регрессионного анализа. Известно, что увеличение количества внесенных удобрений ведет к повышению урожайности. Если в анализе исследовать достаточно большую площадь посевов, то можно обнаружить прямую корреляционную зависимость между количеством внесенных удобрений и средним уровнем урожайности. Таким образом, можно использовать мощную статистическую процедуру, называемую регрессионным анализом, чтобы определить оптимальные условия для выращивания сельскохозяйственных культур. В этой статье будет рассмотрено, когда и как использовать методы регрессионного анализа.

Ключевые слова: регрессионный анализ, независимые переменные, связь, методы.

Summary

Variables measured in crop production, such as the amount of precipitation, doses of mineral fertilizers, timing and quality of sowing are usually continuous. There are statistical methods to confirm whether these variables act as independent variables causing the result. Most of these methods are variations of regression analysis. It is known that an increase in the amount of fertilizers applied leads to an increase in yields. If a sufficiently large area of crops is examined in the analysis, a direct correlation can be found between the amount of fertilizers applied and the average yield level. In this way, a powerful statistical procedure called regression analysis can be used to determine the optimal conditions for growing crops. This article will discuss when and how to use regression analysis methods.

Keywords: regression analysis, independent variables, relationship, methods.

Регрессионный анализ позволяет исследовать взаимосвязь между переменными. Обычно переменные помечаются как зависимые или независимые. Независимая переменная – это вход, движущая сила или фактор, который оказывает влияние на зависимую переменную (которую также можно назвать результатом). Например, если мы скажем, что дозы удобрений влияют на урожайность пшеницы, какие здесь будут независимые и зависимые переменные? В данном случае удобрения является независимой переменной, и он может влиять на результат/зависимую переменную – в данном случае на урожайность.

Регрессионный анализ преследует четыре основные цели: описание, оценка, прогнозирование и контроль. С помощью описания регрессия может объяснить связь между зависимыми и независимыми переменными. Оценка означает, что с помощью наблюдаемых значений независимых переменных можно оценить значение зависимой переменной. Регрессионный анализ может быть полезен для прогнозирования результатов и изменений зависимых переменных на

основе взаимосвязей зависимых и независимых переменных. Регрессия позволяет контролировать влияние одной или нескольких независимых переменных при исследовании взаимосвязи одной независимой переменной с зависимой переменной.

Самым основным шагом в регрессии является предположение о взаимосвязи между зависимой переменной Y и независимой переменной X . После этого нахождение констант, составляющих реляционное выражение, является ключевой целью регрессионного анализа. Простейшая линейная регрессия предполагает следующую связь между Y и X .

$$y = \beta_0 + \beta_1 x + \varepsilon$$

В этом уравнении β_0 является пересечением y и относится к оценочному значению y , когда x равно 0. Коэффициент β_1 является коэффициентом регрессии и обозначает, что расчетное увеличение зависимой переменной для каждой единицы увеличивается независимой переменной. Символ ε является компонентом случайной ошибки и обозначает неточность регрессии, указывающую на то, что на практике независимые переменные не могут идеально предсказать изменение какой-либо зависимой переменной. Кроме того, для каждого коэффициента предусмотрено каждое P -значение, что означает, что коэффициент не равен нулю при $P < 0,05$. Чем меньше P -значение, тем меньше вероятность совершить ошибку (отклонить истинную нулевую гипотезу), поэтому тем больше вероятность того, что модель регрессии корректна. Множественная линейная регрессия следует той же логике, что и одномерная линейная регрессия, за исключением (а) множественной регрессии, в которой существует более одной независимой переменной и (б) среди независимых переменных должна быть неколлинеарность.

Линейный и множественный регрессионный анализ зависит от таких факторов, как размер выборки, отсутствующие данные и характер выборки. Небольшой размер выборки может демонстрировать связи только между переменными с сильной связью. Большое количество отсутствующих значений в наборе данных может повлиять на размер выборки. Подвыборки в большей выборке могут маскировать фактическое влияние независимых и зависимых переменных.

Несколько независимых переменных могут влиять на зависимую переменную. В регрессионном анализе одновременный вклад нескольких независимых переменных может быть измерен как одна регрессия. В этом случае вклад каждой независимой переменной уменьшается, также может уменьшаться за счет простой регрессии. С этой точки зрения регрессионный анализ с использованием нескольких независимых переменных также можно назвать скорректированной регрессией. Другие независимые переменные, которые не представляют особого интереса, называются ковариатами.

Начиная с установления реляционного выражения, которое является основой регрессии, линейный регрессионный анализ с n независимыми переменными может быть выражен следующим образом.

$$Y = b_0 + b_1 x_1 + \dots + b_n x_n + \varepsilon$$

Такую регрессию можно назвать множественной регрессией в том смысле, что существует множество независимых переменных. Независимость, нормальность и гомоскедастичность членов ошибки также должны удовлетворяться как в множественной регрессии, так и в простой регрессии. Кроме того, в множественных линейных регрессиях все независимые переменные (x_1 до x_n) должны иметь линейную связь с зависимой переменной (=предположение о линейности) [1].

Было бы идеально, если бы каждая переменная была независима друг от друга, когда в нее включено несколько переменных, но в действительности действие переменной может варьироваться в зависимости от изменения других переменных. Это называется взаимодействием, и его можно обнаружить, добавив новую переменную с именем $X_a * X_b$ к модели, которая является произведением двух переменных (X_a и X_b). Когда Р-значение переменной произведения меньше 0,05, между двумя переменными существует взаимодействие, что предполагает необходимость расширенной модели или стратификационного анализа [2].

При включении в модель нескольких независимых переменных следует учитывать, что некоторые из независимых переменных не являются необходимыми. Например, одна независимая переменная может быть сконфигурирована в виде линейной комбинации остальных независимых переменных. При выполнении множественной линейной регрессии добавление этой переменной не приносит пользы модели. Мультиколлинеарность — это такое свойство, при котором одна переменная может быть объяснена линейной комбинацией остальных переменных. Однако, поскольку идеальной мультиколлинеарности не существует, степень мультиколлинеарности должна быть определена. С помощью метода, использующего условный индекс, обычно подозревают мультиколлинеарность, если условный индекс выше 10. Некоторые исследователи выражают это как собственное значение, которое является обратной величиной условного индекса. Мультиколлинеарность подозревается, когда собственное значение меньше 0,1.

При подозрении на мультиколлинеарность чаще всего выполняется регрессионный анализ для каждой независимой переменной с использованием оставшихся независимых переменных для получения фактора дисперсионной инфляции (VIF). Если значение VIF определенной переменной больше 10, модель может быть улучшена путем удаления этой переменной. В дополнение к VIF, можно изучить, вызывает ли каждая переменная проблемы мультиколлинеарности, с помощью таких методов, как число условий и доля дисперсионной декомпозиции (VDP) [3].

Если сложно определить, какие независимые переменные следует удалить с помощью эмпирических знаний, то такие переменные в модели могут быть уменьшены с помощью методологии, основанной на данных. Метод прямого выбора — это метод добавления значимых независимых переменных одну за другой, и независимые переменные, однажды включенные, никогда не удаляются из модели. Метод обратной оценки удаляет незначимые независимые переменные одну за другой, а независимые переменные после удаления никогда не включаются в модель. В дополнение к вышеуказанным двум методам, независимые переменные модели могут быть уменьшены в соответствии со статистическими процедурами с помощью таких методов, как ступенчатый выбор, максимум R^2 улучшение, минимум R^2 улучшение, R^2 выбор, скорректированный R^2 и выбор С Маллоуза.

Регрессионный анализ является мощной и полезной статистической процедурой, имеющей множество последствий для исследований. Это позволяет исследователям описывать, прогнозировать и оценивать взаимосвязи и делать правдоподобные выводы о взаимосвязанных переменных в отношении любых изучаемых явлений. Регрессия также позволяет контролировать одну или несколько переменных, когда исследователи заинтересованы в изучении взаимосвязи между конкретными переменными. Представлены некоторые из ключевых соображений, которые могут быть полезны исследователям, проводящим регрессионный анализ. При планировании и проведении регрессионного анализа исследователи должны учитывать тип и

количество зависимых и независимых переменных, а также характер и размер выборки. Неправильный выбор метода регрессионного анализа при небольшой выборке может привести к ошибочным выводам об изучаемом явлении

Библиографический список

1. Линейная регрессия: понимание линий тренда с линейной регрессией. URL: <https://fastercapital.com/ru/content>.
2. Регрессионный анализ. URL: <https://doc.arcgis.com/ru/insights/latest/analyze/regression-analysis.htm>.
3. Отбор переменных в моделях линейной регрессии URL: <https://loginom.ru/blog/feature-selection>.