Лекции 7

КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

1. Понятие о корреляции, регрессии и ковариации
2. Прямолинейна корреляция и регрессия
3. Нелинейная корреляция и регрессия

1. Понятие о корреляции, регрессии и ковариации. В агрономических исследованиях крайне редко функциональные зависимости, когда одному значении. Факториального признака X соответствует строго определенное значение результативного признака Y. Наиболее часто встречаются такие зависимости, при которых одному значению признака X соответствует множество значений признака Y. Такие связи называются ***вероятностными***, или ***корреляционными***, или ***стохастическими***.

Для определения корреляционной зависимости используют специальные статистические методы, которые называются *корреляция* и *регрессия*.

Корреляция определяет форму, направление и тесноту связи

По ***форме*** корреляция может быть линейной или нелинейной:

- при линейной или прямолинейной корреляции одинаковому приращению признака X соответствует одинаковое изменение признака Y;

- при нелинейной корреляции одинаковому приращению признака X соответствует неодинаковое изменение признака Y.

По ***направлению*** корреляция может быть прямой и обратной:

- при прямой или положительной зависимости с увеличением признака X значение признака Y увеличивается;

- при обратной или отрицательной зависимости с увеличением признака X значение признака Y уменьшается.

По ***количеству изучаемых признаков*** корреляция может быть простой и множественной:

- при простой исследуется взаимосвязь между 2 признаками;

- при множественной – между 3 и более признаками.

Под ***регрессией*** понимают изменение результативного признака Y при определенном изменении одного или нескольких факториальных признаков. Связь между признаками выражается *уравнением регрессии.* По уравнению регрессии можно определить вероятное значение результативного признака Y для определенного значения факториального признака X.

Совместное применение дисперсионного и корреляционно-регрессионного анализов для уточнения результатов эксперимента называется ***ковариацией***. Сущность ковариационного анализа заключается в следующем. Если между результативным признаком и сопутствующим эксперименту не изучаемым признаком имеется значимая линейная связь, то метод ковариации позволяет статистически выровнять условия проведения эксперимента в отношении не изучаемого признака и, тем самым, снизить ошибку эксперимента и получить более точные данные.

2. Прямолинейна корреляция и регрессия. Прямолинейная или линейная корреляционная зависимость между 2 признаками является самой простой формой корреляционных связей. Линейная или прямолинейная регрессия – это такая зависимость, когда при любом значении факториального признака X одинаковые приращения его вызывают одинаковые изменения результативного признака Y.

Для анализа линейной корреляции между признаками X и Y проводят n-количество парных наблюдений, в результате которых мы получаем ряд парных чисел (X_1, Y_1; X_2, Y_2;…;X_n, Y_n). На основании полученных значений рассчитываем эмпирические коэффициенты корреляции, детерминации и регрессии.

Для определения тесноты и формы связи рассчитываем ***коэффициент корреляции r*** по формуле:

$$r = \frac{\sum (Y - \bar{y}) \cdot (X - \bar{x})}{\sqrt{\sum (Y - \bar{y})^2 \cdot \sum (X - \bar{x})^2}}$$

Коэффициент корреляции изменяется в пределах от -1 до +1. Знак коэффициента корреляции показывает направление связи, Числовое значение – тесноту связи.

Если значение коэффициента корреляции равно 0, то линейная связь между изучаемыми признаками отсутствует, но может быть нелинейная; если значение коэффициента корреляции равно 1, то связь между признаками функциональная. Считается, что если коэффициент корреляции меньше 0,3, то связь слабая, а если больше 0,7 – сильная.

Коэффициент детерминации $d_{yx}=r^2$ показывает степень сопряженности признаков, т.е. ту часть изменчивости признака X, которая определяется влиянием признака Y.

Коэффициент регрессии рассчитывается для определения уравнения регрессии. В зависимости от направленности взаимовлияния признаков рассчитывается один или два коэффициента регрессии: b_{yx} и b_{xy} .

Коэффициент регрессии b_{yx} показывает в каком направлении и на какую величину изменяется признак Y при увеличении признака X на единицу измерения.

Коэффициент регрессии b_{xy} показывает в каком направлении и на какую величину изменяется признак X при увеличении признака Y на единицу измерения.

Коэффициент регрессии имеет то же знак, что и коэффициент корреляции.

Для оценки существенности корреляционной связи рассчитывают критерий существенности:

$$t_{факт} = \frac{r}{s_r}$$

где r – коэффициент корреляции;

s_r – ошибка коэффициента корреляции.

Ошибка коэффициента корреляции определяется по формуле:

$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

где n – число пар значений признаков.

Знак критерия существенности не несет смысловой нагрузки. Критерий существенности сравнивают с критерием Стьюдента. Значение критерия Стьюдента берется из статистической таблицы с учетом принятого уровня значимости и числа степеней свободы, которое рассчитывается по формуле $v = n - 2$.

Если $t_{факт} \geq t_{теор}$, связь между признаками значима; если $t_{факт} < t_{теор}$, связь незначима.

На основании коэффициента регрессии и средних значений изучаемых признаков выводим уравнение регрессии. Уравнение линейной регрессии Y по X имеет следующий вид

$$Y = \bar{y} + b_{yx}(X - \bar{x})$$

где $\bar{y}$ и $\bar{x}$ - средние арифметические для ряда Y и X,

b_{yx} – коэффициент регрессии Y по X.

По уравнению регрессии находят теоретические значения Y для двух экстремальных значений признака X - Xmin и Xmax. Найденные точки наносят на график и соединяют прямой. Данная прямая является теоретической линией регрессии Y по X.

3. Нелинейная корреляция и регрессия. При нелинейной (криволинейной) регрессии одинаковому изменению факториального признака X соответствует неодинаковое изменение результативного признака Y. Для оценки степени криволинейной зависимости используется показатель, предложенный Пирсоном, который называется *корреляционное отношение*, обозначается буквой греческого алфавита η (эта) и рассчитывается по формуле:

$$\eta = \sqrt{\frac{\sum(Y-\bar{y})^2 - \sum(Y-\bar{y}_x)^2}{\sum(Y-\bar{y})^2}}$$

где $\sum(Y-\bar{y})^2$ - сумма квадратов отклонений индивидуальных значений Y от общей средней арифметической $\bar{y}$, она характеризует общее варьирование признака Y;

$\sum(Y-\bar{y}_x)^2$ - сумма квадратов отклонений вариант от частных средних $\bar{y}_x$, соответствующих определенным, фиксированным значениям независимой переменной X, она характеризует ту часть варьирования признака Y, которая связана с изменчивостью признака X.

Изменяется корреляционное отношение в пределах от 0 до 1. Чем ближе значение корреляционного отношения к 1, тем сильнее зависимость.

Пример. ***Определить влияние семенной инфекции яровой пшеницы (X) на массу зерен (*Y*).*** Были получены следующие данные:

Пораженные растения, %	66.7	38.5	45.4	17.6	37.5	16.0	66.2	23.0
Масса 1000 шт семян, г	14.9	25.8	28.0	36.5	30.1	37.2	10.5	34.7

Для расчета корреляционного отношения необходимо заполнить вспомогательную таблицу. Значения признака X ранжируются в порядке возрастания. Напротив каждого значения X записываем соответствующее ему значение Y. Т.е. *ранжируются только значения признака X*.

X	$\bar{x}_y$	Y	$\bar{y}_x$	$Y-\bar{y}_x$	$Y-\bar{y}$	$(Y-\bar{y}_x)^2$	$(Y-\bar{y})^2$
$\sum$	-		-	-	-		

Значения разбиваем на группы. В каждой группе должно быть не менее 2 значений, причем размер может быть неодинаковым. Одинаковые значения признака X, если они есть, должны быть в одной группе.

Определяем корреляционное отношение

$$\eta = \sqrt{\frac{\sum(Y-\bar{y})^2 - \sum(Y-\bar{y}_x)^2}{\sum(Y-\bar{y})^2}} =$$

Для оценки значимости корреляционного отношения рассчитывают ошибку корреляционного отношения

$$s_\eta = \sqrt{\frac{1-\eta^2}{n-2}}$$

и критерий существенности

$$t_\eta = \frac{\eta}{s_\eta}$$

Критерий существенности сравнивают с критерием Стьюдента, значение которого берут из статистической таблицы с учетом принятого уровня значимости и числе степеней свободы v = n-2.

При v = n-2 =6 t_{05}=

Корреляционное отношение измеряет степень корреляции как при нелинейной так и при прямолинейной корреляционной зависимости. Для определения формы связи используют критерий Фишера, вычисляемый по формуле:

$$F = \frac{(\eta^2 - r^2)\cdot(n-k)}{(1-\eta^2)\cdot(k-2)}$$

Затем фактическое значение критерия Фишера сравнивают с теоретическим, которое берут из статистической таблицы с учетом принятого уровня значимости, объема выборки и числа групп по ряду X.

Если $F_{факт.} < F_{теор.}$, связь носит линейный характер и можно определять показатели для прямолинейной корреляции и регрессии. Если $F_{факт.} \geq F_{теор.}$, корреляция является нелинейной.

Криволинейные зависимости между двумя переменными могут быть выражены в виде кривых и соответствующих им математических уравнений. Кривые могут иметь вид парабол, гипербол, логарифмических кривых.