Регрессионный анализ

Лекция 20

Звоновский, к.с.н.

Регрессионный анализ

Выше мы использовали виды взаимосвязи, которые указывали бы нам на тесноту взаимосвязи между двумя переменными. В то время как нам хотелось бы вычислить силу этой взаимосвязи.

Это позволит рассчитывать значения зависимой переменной у объектов как выборочной, так и генеральной совокупности на основании информации о независимой переменной, а также прогнозировать значение первой в другие моменты времени – в прошлом и будущем.

Регрессионный анализ

Предположим, что нам нужно выяснить насколько будет меняться успеваемость студентов в случае, если мы будем отбирать абитуриентов с высокой предварительной подготовкой.
При этом мы знаем, что другие факторы также влияют на успеваемость, но мы сознательно отказываемся анализировать силу влияния другой величины.

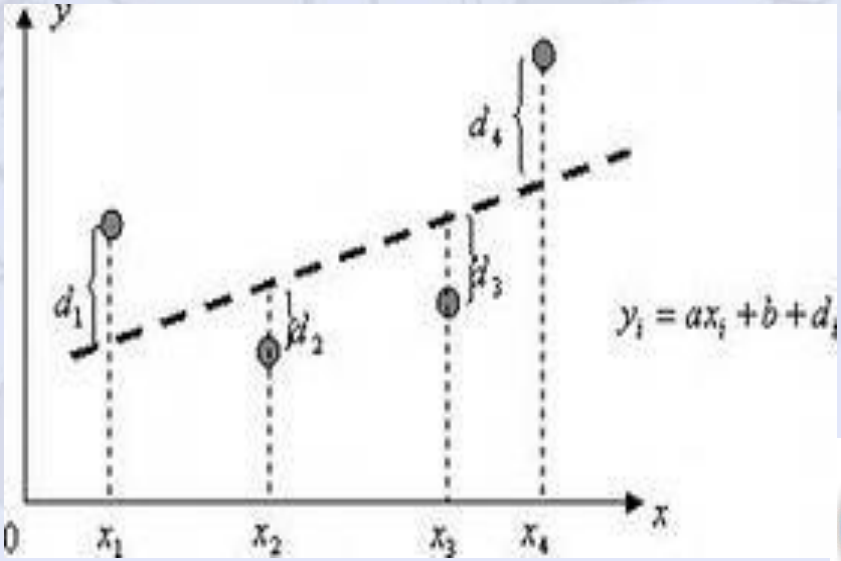
Регрессионный анализ

Принимается, что увеличение успеваемости студента на Y значений возникает если уровень предварительной подготовки возрастает на X значений.

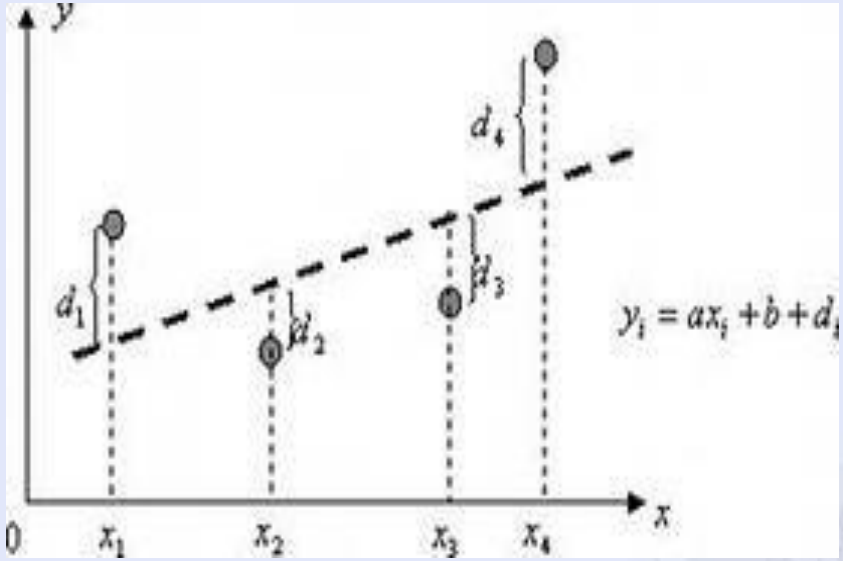
$$y_i = M(Y|X = x_i) + e_i = \beta_0 + \beta_1 x_i + e_i$$

Здесь β_0 – значение зависимой переменной в случае, когда независимая равна нулю, а β_1 - угол наклона прямой регрессии к оси, где расположены значения независимой переменной.

Остатки e – это ошибка между расчетным значением y в точке I и выборочным значением y_i.



Регрессионный анализ



В геометрическом смысле регрессионный анализ состоит в построении прямой, при котором сумма ошибок e_i минимальна. Сумма ошибок, как видно из рисунка, представляет собой расстояния от выборочного значения переменной до расчетного.

Существует несколько способов расчета расстояний. Самым распространенным является метод наименьших квадратов. Наименьшее значение получается в случае

$$\beta_1 = S_{x,y} / D_x$$

$$\beta_0 = y_i - \beta_1 x_i$$

Регрессионный анализ

Нулевая гипотеза в данном случае состоит в том, что между X и Y не существует линейной зависимости. **Альтернативная** предполагает, что между двумя переменными есть положительная или отрицательная линейная связь.

Обычно проводится проверка на основе двустороннего теста.

Также оценивается сила связи между двумя переменными. Для этого используется **коэффициент детерминации**, изменяющийся от 0 до 1 и представляющий собой долю дисперсии независимой переменной в дисперсии зависимой.

Данный коэффициент также должен оцениваться на значимость.

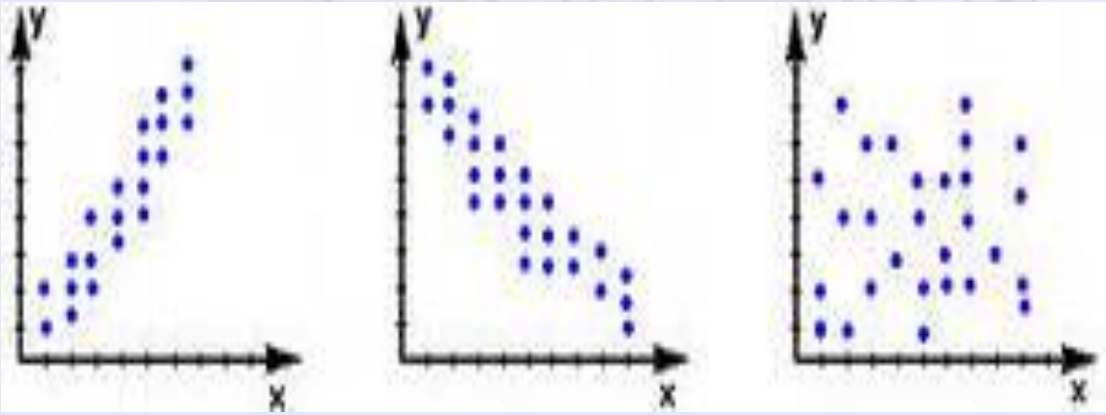
Регрессионный анализ

- Построение корреляционной диаграммы
- Выбор модели (двумерная или многомерная)
- Оценка параметров
- Расчет стандартизированных коэффициентов
- Проверка значимости
- Расчет силы и значимости зависимости
- Расчет точности прогнозирования (СОО)
- Изучение остатков

Регрессионный анализ

Корреляционная диаграмма статистическая взаимосвязь двух или нескольких случайных величин, где изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин.

Выбор модели подразумевает сведение всего разнообразия факторов, влияющих на зависимую переменную, до одной или нескольких независимых переменных. В зависимости от этого различают двумерный или многомерный регрессионный анализ.



Регрессионный анализ

Оценка параметров представляет собой расчет коэффициентов β_0 и β_1.

$$\beta_1 = S_{x,y} / D_x$$

$$S_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Стандартизация - преобразование переменных, имеющих размерность и различный диапазон значений к безразмерной переменной, с диапазоном значений от 0 до 1. Собственно, они и называются бета-коэффициентами.

Регрессионный анализ

Проверка значимости состоит в проверке нулевой гипотезы об отсутствии зависимости (или – о независимости) X и Y, что равнозначно равенству нулю β_1. Значимость проверяют на основании (чаще всего) двустороннего теста Стьюдента, где t=b/SE

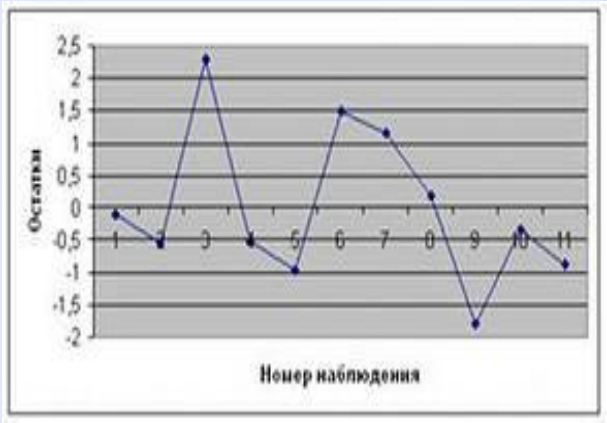
Сила и значимость зависимости. В регрессионном анализе не только фиксируют наличие зависимости между переменными X и Y, но измеряют ее силу и значимость.

Сила выражена через **коэффициент детерминированности,** представляющий собой квадрат совместного коэффициента корреляции. Он же является долей дисперсии зависимой переменной, объясняемой влиянием на нее независимой. Так, если в нашем случае R^2=0,298, это значит, что 29,8% дисперсии текущей успеваемости студента объясняется его предварительной подготовкой.

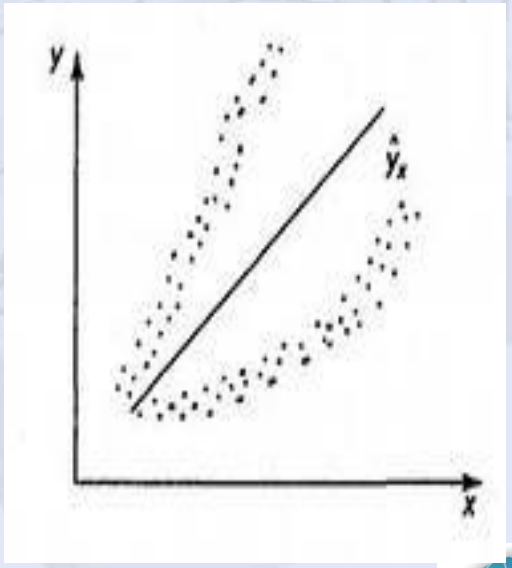
Регрессионный анализ

Точность прогнозирования. Очевидно, что если мы можем прогнозировать значения Y, мы можем оценить точность такого прогноза. Ошибка стандартизована и безразмерна и чем она больше, тем ниже пригодность регрессии.

Изучение остатков. того, как мы получили значения коэффициентов, необходимо убедиться в нормальности распределения остатков.



Если остатки не имеют нормального распределения есть вероятность, что рассчитанная линия регрессии не имеет физического значения



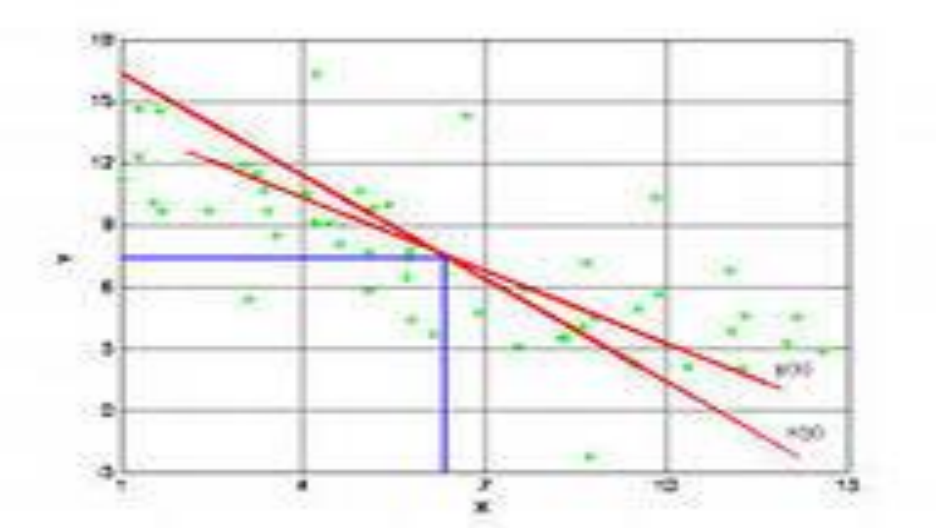
Множественный регрессионный анализ

Если мы будем предполагать, что на успеваемость студента кроме предварительной подготовки влияет и посещение занятий, то анализ влияния этих двух независимых переменных на зависимую будет множественным регрессионным анализом

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i + e_i$$

Регрессионный анализ

Регрессионный анализ может быть крайне полезным при поиске различий между различными социальными группами, например, между мужчинами и женщинами.



В этом случае используют **фиктивные (dummy)** переменные. Они позволяют использовать регрессионный анализ для случая, когда независимые переменные имеют порядковую или номинальную.

В этом случае вместо нескольких парных уравнений используют одно уравнение множественной регрессии.

Регрессионный анализ

Пример преобразования номинальной переменной «семейное положение» в фиктивную переменную.

1. Холост (не замужем) V1. 1 – Холост (не замужем)
   0 – Иное семейное положение
2. Женат (замужем) V2. 1 – Женат (замужем)
   0 – Иное семейное положение
3. Разведен (а) V3. 1 – Разведен (а)
   0 – Иное семейное положение
4. Вдовец (вдова) V4. 1 – Вдовец (вдова)
   0 – Иное семейное положение

Теперь в N – 1 дихотомических переменных содержится информация, находившаяся в номинальной переменной с N градациями.

Логистическая регрессия

Лекция 21

Звоновский, к.с.н.

Логистическая регрессия

Регрессионный анализ может использоваться лишь в случае когда зависимая переменная – метрическая или интервальная.

В случае когда зависимая переменная – **дихотомическая**, используют **логистическую регрессию**.

Очевидно, что число случаев, когда необходимо вычислить силу влияния на факт **события** или его **отсутствия**, например, на выход замуж в текущем году или голосования за определенную партию.

При этом, если в случае метрической зависимой переменной определяется сила воздействия на нее, то в случае дихотомической измеряется **вероятность наступления** события. Вероятность измеряется от 0 до 1.

Таким образом, **логистическая регрессия** решает задачу построения **модели прогноза вероятности** события Y в зависимости от переменных X1, X2,…, XN

Логистическая регрессия

Непосредственно использовать вероятность наступления события в формуле регрессии нельзя. Используют так называемый логит.

Шанс (отношение шансов) – отношение вероятности наступления события к вероятности его ненаступления – P / (1 – P)

Логит – это натуральный логарифм шанса Z = ln (P / (1 – P)).

Тогда $Z = \mathbf{B_0 + B_1x_1 + B_2x_2 +...+ B_nx_n}$

Предположим, что вероятность **голосования за определенную партию** зависит от того, за какую партию человек голосовал на предыдущих выборах ($\mathbf{B_1}$), его социального статуса ($\mathbf{B_2}$), возраста ($\mathbf{B_3}$) и дохода ($\mathbf{B_4}$).

Логистическая регрессия

Логит $Z = B_0 + B_1x_1 + B_2x_2 + ... + B_nx_n$

Предположим, что вероятность **голосования за определенную партию** зависит от того, за какую партию человек голосовал на предыдущих выборах (B_1), его социального статуса (B_2), возраста (B_3) и дохода (B_4).

Результатами логистической регрессии будут: собственно коэффициенты регрессии и классификационная таблица.

Классификационная таблица показывает долю верных предсказаний зависимой переменной с помощью полученных коэффициентов. Например, для значения переменной D1=1 доля верных предсказаний – 67%, а для D1=2 аналогичный показатель 54%. Для обоих значений – 63%.

Значимость рассчитанных коэффициентов рассчитывается либо по статистике Вальда, либо с помощью пошагового расчета коэффициентов.