Энгельсский технологический институт (филиал) федерального
государственного бюджетного образовательного учреждения
высшего образования
«Саратовский государственный технический университет
имени Гагарина Ю.А.»

Кафедра «Оборудование и технологии обработки материалов»

РЕГРЕССИОННЫЕ МОДЕЛИ С НЕСКОЛЬКИМИ ВХОДНЫМИ ПЕРЕМЕННЫМИ

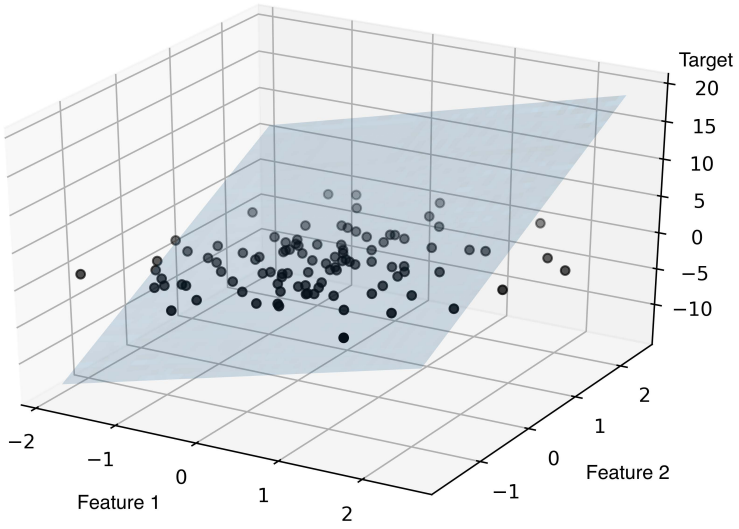
Электронное издание локального распространения

Методические указания к практической работе
по дисциплине «Математическое моделирование технологических процессов»
для студентов направлений подготовки:

15.03.05 «Конструкторско-технологическое обеспечение машиностроительных производств»

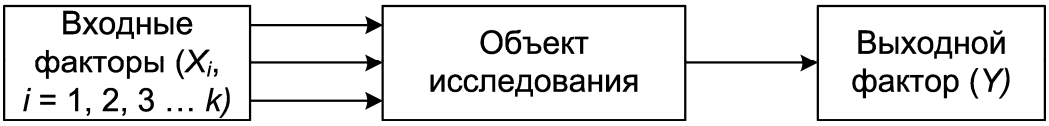
очной и заочной форм обучения

Энгельс 2024



1. Многофакторная (множественная) линейная регрессия

Рассмотрим основы построения регрессионных моделей для объекта с несколькими входными переменными (рисунок).



Объект исследования с несколькими входными факторами

Для построения модели необходимо иметь данные экспериментальных исследований объекта, представленные в виде таблицы, где каждой комбинации значений входных факторов соответствует значение выходного фактора.

Данные экспериментальных исследований объекта

Номер эксперимента	X_1	X_2	…	X_k	Y
1	x_{11}	x_{21}	…	x_{k1}	y_1
2	x_{12}	x_{22}	…	x_{k2}	y_2
3	x_{13}	x_{23}	…	x_{k3}	y_3
…					
m	x_{1m}	x_{2m}	…	x_{km}	y_m

Моделирование объекта со сложным внешним воздействием в виде нескольких входных факторов, так же как и для объекта с одним входным фактором, начинается с линейной модели.

Если иметь неограниченно большое количество экспериментальных точек, то линейная регрессионная модель с несколькими входными переменными имеет вид

$$\bar{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k , \qquad (1)$$

где x_1, x_2, x_3 и т. д. – значения входной переменной;

β_0, β_1, β_2, …, β_k – коэффициенты регрессии.

Имеющиеся экспериментальные данные в виде комбинаций (x_{1i}, x_{2i}, x_{3i}, … x_{ki}, y_i) являются лишь ограниченной выборкой из общего числа состояний исследуемого объекта. Поэтому можно определить только оценки коэффициентов β_0, β_1, β_2, …, которые обозначают, соответственно, b_0, b_1, b_2, b_3, …, b_k.

$$\bar{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_k x_k . \qquad (2)$$

Такие модели в литературе часто называют *многофакторными моделями*. Однако определить коэффициенты регрессии b_i так, как это делается для однофакторной модели – по методу наименьших квадратов, в данном случае не представляется возможным. Необходимо использовать основы алгебры матриц и матричного исчисления.

2. Матричный подход к определению коэффициентов регрессии

Запишем для нашего случая матрицы X, Y, B:

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} & \ldots & x_{k1} \\ 1 & x_{12} & x_{22} & x_{32} & \ldots & x_{k2} \\ 1 & x_{13} & x_{23} & x_{33} & \ldots & x_{k3} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1m} & x_{2m} & x_{3m} & \ldots & x_{km} \end{bmatrix}. \qquad (3)$$

В матрице X все элементы первого столбика равны единице. Будем считать это фиктивной входной переменной X_0 с постоянным значением.

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_m \end{bmatrix}, \qquad B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \dots \\ b_k \end{bmatrix}. \tag{4}$$

Определим размерность этих матриц:

- Y – вектор наблюдений ($m \cdot 1$);
- X – матрица независимых переменных ($m(k + 1)$);
- B – вектор коэффициентов регрессии ($(k + 1) \cdot 1$).

Правила умножения матриц и векторов требуют, чтобы они были *согласованными* и имели соответствующую размерность [9, 12].

Пусть A – матрица размерностью ($n \cdot p$):

- на матрицу A только слева может быть умножена матрица B размерностью ($m \cdot n$):

$$B \cdot A = (m \cdot n \cdot n \cdot p) = C(m \cdot p); \tag{5}$$

- на матрицу A только справа может быть умножена матрица D размерностью ($p \cdot q$):

$$A \cdot C = (n \cdot p \cdot p \cdot q) = F(n \cdot q). \tag{6}$$

Следовательно, произведение $B \cdot X$ не существует и множественная линейная регрессия может быть записана в виде

$$Y = X \cdot B. \tag{7}$$

Использование аппарата линейной алгебры позволяет получить общую формулу для определения вектора, содержащего коэффициенты регрессии [9]:

$$B = (X' \cdot X)^{-1} = X' \cdot Y, \tag{8}$$

где $(X' \cdot X)^{-1}$ – обратная матрица;

X' – транспонированная матрица.

Определением коэффициентов регрессионной модели построение модели не заканчивается. Необходимо также определить адекватность и точность предлагаемой многофакторной модели.

3. Оценка адекватности и точности многофакторной линейной модели

Адекватность модели характеризует соответствие модели экспериментальным данным и статистическую значимость уравнения регрессии. Адекватность регрессионной модели оценивается коэффициентом Фишера

$$F_{\text{расч}} = \frac{\sum_{i=1}^{m}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}. \qquad (9)$$

Расчетное значение коэффициента ($F_{\text{расч}}$) необходимо сравнить с табличным значением ($F_{\text{табл}}(m, \alpha)$), где m – общее количество экспериментальных наблюдений, α – уровень значимости.

α – *уровень значимости* – вероятность, с которой правильная гипотеза о модели может быть отвергнута как неправильная. Обычно в моделировании (и мы об этом уже говорили) используют значения α = 0,05; 0,01. Однако для многофакторных моделей табличное значение F-критерия зависит еще и от числа входных переменных [9].

Если $F_{\text{расч}} > F_{\text{табл}}$, то модель считается адекватной, а регрессия статистически значимой. Если $F_{\text{расч}} < F_{\text{табл}}$, то регрессионная модель неадекватна и регрессия статистически незначима.

Для оценки точности регрессионных моделей с несколькими входными переменными используется множественный коэффициент корреляции (R^2) [4], который определяется по формуле

$$R^2 = \frac{\sum_{i=1}^{m}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{m}(y_i - \overline{y}_i)^2}. \qquad (10)$$

Отношение R^2 характеризует тесноту связи между выходной переменной и входными переменными. Область определения отношения R^2 лежит в пределах от 0 до 1. При $R^2 = 0$ выходной фактор y линейно не за-

висит от входных факторов x_1, x_2, …, x_k – можно сказать, что корреляционная связь между выходным фактором и входными факторами отсутствует. При $R^2 = 1$ выходной фактор y линейно зависит от входных факторов x_1, x_2, …, x_k – имеется в наличии сильная корреляционная связь. Чем выше значение R^2, тем теснее связь в модели между выходной переменной (фактором) и входными переменными (факторами), тем точнее, а следовательно, лучше математическая модель. Если модель имеет низкое значение R^2, то она имеет низкую точность оценки и предсказания поведения или свойств объекта. Использовать такую модель для исследования, описания и предсказания объекта не рекомендуется. Из нескольких моделей для исследования выбирается та, у которой отношение R^2 имеет наибольшее значение.

4. Линейные регрессионные модели с несколькими входными переменными

Если в результате расчета отношения R^2 множественная линейная регрессия признана недостаточно точной, переходят к исследованию более сложных моделей:

- *полинома с одной независимой переменной* $y = b_0 + b_1x + b_2x^2 + b_3x^3$;
- *полинома с несколькими независимыми переменными* $y = b_0 + b_1x_1 + b_2x_2 + b_3x_1^2 + + b_4x_2^2 + b_5x_1x_2$;
- *обратной модели* $y = b_0 + \frac{b_1}{x_1} + \frac{b_2}{x_2} + \frac{b_3}{x_3} + \frac{b_4}{x_4}$ или $y = b_0 + \frac{b_1}{x_1} + \frac{b_2}{x_2} + \frac{b_3}{x_1^2} + \frac{b_4}{x_2^2} + \frac{b_4}{x_1\,x_2}$;
- *комбинированной модели* $y = b_0 + b_1x_1 + b_2\sqrt{x} + b_3\ln x + b_4\exp(x)$.

Исследователь сам может предложить вид многофакторной модели на основе анализа априорной информации. Оценка коэффициентов регрессии, критерия адекватности модели и множественного коэффициента корреляции осуществляется по формулам (6.8), (6.9), (6.10), рассмотренным ранее для «классической» множественной регрессии.

5. Нелинейные регрессионные модели с несколькими входными переменными

Если в результате построения линейных моделей ни одна из них не была признана достаточно точной, переходят к исследованию более слож-

ных моделей. Любую модель, в результате преобразования записываемую в виде (6.1), можно анализировать методами линейного регрессионного анализа.

Виды преобразований для нелинейных моделей [12]:

- обратное преобразование;
- логарифмическое преобразование;
- преобразование типа квадратного корня.

Если при помощи каких-либо преобразований нелинейная модель может быть приведена к виду множественной линейной регрессии, то она называется *нелинейной моделью с «внутренней линейностью»* [11]. К таким моделям относятся:

- *степенная (мультипликативная) модель* $y = b_0 \cdot x_1^{b_1} \cdot x_2^{b_2} \cdot x_3^{b_3} \cdot \ldots \cdot x_k^{b_k}$. Выполним преобразования: $\ln y = \ln b_0 + b_1 \ln x_1 + b_2 \ln x_2 + \ldots + b_k \ln x_k$. Введем переобозначения: $\ln x_i = z_i$, $w = b_0^* + b_1 z_1 + b_2 z_2 + \ldots + b_k z_k$. Данная модель имеет вид многофакторной линейной регрессии (см. формулу (6.1)), следовательно, *степенная* модель может быть преобразована логарифмированием к виду многофакторной линейной регрессии. Это позволяет использовать для расчета коэффициентов регрессии, критерия адекватности и множественного коэффициента корреляции формулы, рассмотренные выше. После вычислений необходимо выполнить потенцирование [11] и вернуться к исходному степенному виду модели;
- *экспоненциальная модель* $y = \exp(b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_k x_k)$. Логарифмирование также позволяет преобразовать экспоненциальную модель к виду многофакторной линейной модели (формула 6.1) и использовать для ее исследования аппарат линейного регрессионного анализа;
- *обратная модель* $y = \dfrac{1}{b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k}$. Выполним преобразования: $\dfrac{1}{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$. Введем переобозначения: $\dfrac{1}{y} = w$, $w = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$. Следовательно, обратная модель также может быть преобразована к виду множественной линейной регрессии, что позволяет использовать для ее исследования аппарат линейного регрессионного анализа.

Если при помощи любых преобразований нелинейная модель не может быть приведена к виду множественной линейной регрессии, то она называется *нелинейной моделью с «внутренней нелинейностью»*. Для иссле-

дования этих моделей в математике разработан аппарат нелинейного регрессионного анализа [12].

6. Шаговые методы построения регрессионных моделей

На практике при исследовании объектов и построении регрессионных моделей с несколькими входными переменными используют шаговые (итерационные) методы, позволяющие обоснованно включать в модель только наиболее влиятельные и значимые входные факторы. Чаще всего применяют *метод исключения переменных* и *метод включения переменных*.

Метод исключения переменных состоит из нескольких этапов [12]:

• Этап 1: предлагается регрессионная модель, включающая все исследуемые входные переменные.

• Этап 2: рассчитывается частный F-критерий Фишера для каждой входной переменной F_{X_i} и оценивается статистическая значимость каждой входной переменной аналогично тому, как это выполняется для модели в целом.

• Этап 3: устанавливается входная переменная с минимальным значением F_{X_i}.

• Этап 4: минимальное значение F-критерия сравнивается с граничным значением.

Если $F_{\min} > F_{\text{табл}}(m, \alpha)$, то соответствующая входная переменная считается статистически значимой и остается в модели. Следовательно, предложенная на этапе 1 модель статистически значима, адекватна и может быть использована для исследования объекта. Анализ оставшихся входных переменных уже не проводится. Далее рассчитываются коэффициенты регрессии и множественный коэффициент корреляции.

Если $F_{\min} < F_{\text{табл}}$, то соответствующая входная переменная статистически незначима и должна быть удалена из модели как неадекватная. Следует помнить, что для многофакторных моделей табличное значение F-критерия зависит еще и от числа входных переменных [9, 12].

• Этап 5: после удаления статистически незначимой входной переменной для оставшихся входных переменных снова пересчитываются их частные F-критерии.

• Этап 6: снова устанавливается входная переменная (из оставшихся) с минимальным значением F_{Xi}.

• Этап 7: этапы 4, 5, 6 повторяются до тех пор, пока в модели не останутся только статистически значимые переменные.

Если после выполнения указанных процедур в модели не останется ни одной статистически значимой переменной, то модель построить нельзя. Скорее всего, причина в просчетах, допущенных при планировании эксперимента на основе анализа априорной информации. В этом случае следует снова внимательно проанализировать априорную информацию, заново спланировать и провести эксперимент и обработать его результаты.

Если в модели остаются входные переменные, то переходят к расчету коэффициентов регрессии и оценке точности модели.

Метод включения переменных также состоит из этапов:

• Этап 0: в «модели» нет ни одной входной переменной.

• Этап 1: рассчитывается корреляционная матрица (из частных коэффициентов корреляции входных переменных друг с другом и с выходной переменной)

$$\begin{bmatrix} R_{1y} & R_{11} & R_{21} & R_{31} & \dots & R_{k1} \\ R_{2y} & R_{12} & R_{22} & R_{32} & \dots & R_{k2} \\ R_{3y} & R_{13} & R_{23} & R_{33} & \dots & R_{k3} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ R_{ky} & R_{1k} & R_{2k} & R_{3k} & \dots & R_{kk} \end{bmatrix},$$

где R_{ky} – частный коэффициент корреляции k-й входной переменной и выходной переменной.

R_{ij} – частный коэффициент корреляции i-*й* и j-*й* входных переменных ($i = \overline{1, k}$; $j = \overline{1, k}$; $i \neq j$).

• Этап 2: выбирается входная переменная с максимальным коэффициентом корреляции R_{iy}. Эта переменная первой вводится в модель.

• Этап 3: определяется частный F-критерий введенной входной переменной, который одновременно является критерием адекватности всей модели.

Если $F_{\text{расч}} < F_{\text{табл}}$, то соответствующая входная переменная статистически незначима, т. е. должна быть удалена из модели, а сама модель неадекватна. Оставшиеся входные переменные имеют с выходной переменной еще менее тесную корреляционную связь. Следовательно, в данной ситуации модель построить не представляется возможным.

Если $F_{расч} > F_{табл}(m, \alpha)$, то соответствующая входная переменная считается статистически значимой и остается в модели.

• Этап 4: корреляционная матрица пересчитывается без учета влияния выбранной входной переменной.

• Этап 5: из оставшихся входных переменных выбирается переменная с максимальным коэффициентом корреляции R_{iy}. Эта переменная вводится в модель следующей.

• Этап 6: определяется F-критерий новой модели.

Если $F_{расч} < F_{табл}$, то вновь введенная в модель входная переменная статистически незначима и должна быть удалена из модели как неадекватная. В модели остается одна входная переменная.

• Этап 7: процедуры 4, 5, 6 повторяются до тех пор, пока не сформируется окончательный вид модели. Дальше переходят к расчету коэффициентов регрессии и оценке точности модели.

Сравнивая данные методы, можно сказать следующее:

• Метод исключения входных переменных дает вполне удовлетворительные результаты при моделировании.

• Метод включения входных переменных более экономичен в вычислительном аспекте.

Вопросы для самоконтроля

1. Что такое многофакторная линейная регрессия?

2. Как оценивается точность многофакторной линейной регрессионной модели?

3. Как оценивается адекватность многофакторной линейной регрессионной модели?

4. Какие значения может принимать множественный коэффициент корреляции?

5. Что такое нелинейные модели с «внутренней линейностью»?

6. Какие бывают нелинейные модели с «внутренней линейностью»?

7. Что такое нелинейные модели с «внутренней нелинейностью»?

8. Обозначьте основные этапы метода включения переменных.

9. Что такое корреляционная матрица?

10. Что такое частный критерий Фишера для входной переменной? Что он характеризует?

11. Обозначьте основные этапы метода исключения переменных.

Список литературы

1. *Кузьмин В. В.* Математическое моделирование технологических процессов сборки и механической обработки изделий машиностроения: учебник для вузов / В. В. Кузьмин [и др.]. Москва: Высшая школа, 2008. 279 с.

2. *Ашихмин В. Н.* Введение в математическое моделирование: учебное пособие / В. Н. Ашихмин [и др.]; под ред. П. В. Трусова. Москва: ЛОГОС, 2005. 440 с.

3. *Советов Б. Я.* Моделирование систем: учебник для вузов / Б. Я. Советов, С. А. Яковлев. 3-е изд., перераб и доп. Москва: Высшая школа, 2001. 343 с.

4. *Зобнин Б. Б.* Моделирование систем: конспект лекций / Б. Б. Зобнин. Екатеринбург: Изд-во УГГГА, 2001. 129 с.

5. *Дьяконов В. П.* Новые информационные технологии: учебное пособие / В. П. Дьяконов [и др.]; под ред. В. П. Дьяконова. Москва: СОЛОН-Пресс, 2005. 640 с.

6. *Саблина Н. Г.* Информационные технологии: конспект лекций: в 2 частях / Н. Г. Саблина, Г. М. Черногородова. Екатеринбург: Изд-во УГТУ – УПИ, 2001. Ч. 2. 119 с.

7. *Дулов В. Г.* Математическое моделирование в современном естествознании: учебное пособие / В. Г. Дулов, В. А. Цибаров; под ред. В. Г. Дулова. Санкт-Петербург: Изд-во С.-Петерб. ун-та, 2001. 244 с.

8. *Зарубин В. С.* Математическое моделирование в технике: учебник для вузов / В. С. Зарубин [и др.]; под ред. В. С. Зарубина. Москва: Изд-во МГТУ им. Н. Э. Баумана, 2001. 496 с.

9. *Корн Г.* Справочник по математике для научных работников и инженеров / Г. Корн, Т. Корн. Москва: Наука, 1972. 830 с.

10. *Спирин Н. А.* Методы планирования и обработки результатов инженерного эксперимента: учебное пособие / Н. А. Спирин [и др.]; под ред. Н. А. Спирина; ГОУ ВПО УГТУ – УПИ. Екатеринбург, 2003. 260 с.

11. *Рогов В. А.* Методика и практика технических экспериментов: учебное пособие / В. А. Рогов. Москва: Академия, 2005. 288 с.

12. *Дрейпер Н.* Прикладной регрессионный анализ: перевод с английского / Н. Дрейпер, Г. Смит. 3-е изд. Москва: Вильямс, 2007. 912 с.