НИУ ВШЭ, ОП «Политология»
Курс «Научно-исследовательский семинар: регрессионный анализ», 2025

# Занятие 2

## Множественная линейная регрессия
## Спецификация модели: выбор контрольных переменных
## Оценивание параметров

1. **Решение задач: таблица разложения вариации**

**Задание 1.** Ниже представлены результаты анализа разложения вариации по линейной парной регрессионной модели, построенной по выборке из 15 наблюдений.

```
Analysis of Variance Table

Response: y
          df  sum_sq mean_sq   f      PR(>F)
x         ...  ...     ...    0.6526   ...
Residual  ... 56.116   ...
```

Восстановим пропуски в таблице:

- df для x = k = 1
- df для Residual $= n - k - 1 = 15 - 2 = 13$
- mean_sq для Residual $= \dfrac{56.116}{13} \approx 4.317$
- $f = \dfrac{mean\_sq(x)}{mean\_sq(Residual)} = \dfrac{mean\_sq(x)}{4.317} = 0.6526$

  Следовательно, mean_sq для x $= 4.317 \times 0.6526 \approx 2.817$
- Рассчитаем p-value, помнив о том, что нас интересует односторонняя альтернатива. В Python можно рассчитать следующим образом:

  ```
  from scipy.stats import f
  f.sf(0.6526, 1, 13)
  ```

  В итоге получили 0.434, что говорит о том, что $R^2$ неотличим от 0.

  $R^2 = \dfrac{2.817}{56.116 + 2.817} \approx 0.048$

```
Analysis of Variance Table

Response: y
          df  sum_sq mean_sq   f      PR(>F)
x         1   2.817  2.817   0.6526   0.434
Residual  13 56.116  4.316
```

**Задание 2.**

На данных по 44 городам построена модель, обясняющая динамику уровня преступности за последние 10 лет. `change_in_crime_rate` — прирост преступности в %, `change_in_pop` — прирост численности населения, %; `kids` — процент детей; `free_lunch` — процент бесплатных школьных обедов; `income_change` — прирост доходов домохозяйств.

Восстановим пропуски в таблице:

```
Coefficients:
                 coef      std. err   t       Pr>|t|   [0.025  0.975]
Intercept      -22.3548    12.3097  -1.816  0.0771  -47.253; 2.544
change_in_pop    0.3188     0.2052   1.533  0.1333   -0.096; 0.734
kids             1.1128     0.2869   3.879  0.0004    0.532; 1.693
free_lunch      -0.3681     0.0973  -3.783  0.0005   -0.565;-0.171
income_change   -0.1944     0.3681  -0.528  0.6004   -0.939; 0.551
---

              df   sum_sq  mean_sq  f      PR(>F)
change_in_pop  1    803.2   803.2  6.248  .000
kids           1   1380.1  1380.1
free_lunch     1   3186.6  3186.6
income_change  1     60.6    60.6
Residual      39   8476.0   217.3
---
```

$$R^2 = \frac{803.2 + 1380.1 + 3186.6 + 60.6}{803.2 + 1380.1 + 3186.6 + 60.6 + 8476} = 0.39$$

В силу того, что p-value получился близким к 0, данная модель лучше, чем модель на константу. Те есть, $R^2$ статистически значимым образом отличается от 0.

2. **Запишем спецификацию модели множественной регрессии**

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_k x_{ki}$$

У предикторов появляется 2 субиндекса: первый субиндекс обозначает номер предиктора, второй – номер наблюдения.

Запишем ту же спецификацию в векторно-матричном виде:

$$\begin{pmatrix} y_1 \\ ... \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & ... & x_{k1} \\ ... & ... & ... & ... \\ 1 & x_{1N} & ... & x_{kN} \end{pmatrix} \times \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ ... \\ \hat{\beta}_k \end{pmatrix} + \begin{pmatrix} \hat{\epsilon}_1 \\ ... \\ \hat{\epsilon}_N \end{pmatrix}$$

$$\vec{y} = X\vec{\hat{\beta}} + \vec{\hat{\epsilon}}$$

Мы уже показывали, что в линейной регрессии вектор остатков ортогонален предикторам (столбцам матрицы X).

$$X^T(\vec{y} - X\hat{\vec{\beta}}) = 0$$

$$X^T\vec{y} = X^T X\hat{\vec{\beta}}$$

$$\boxed{\hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{y}}$$

3. **Выбор контрольных переменных. Backdoor criterion**
Формализуем задачу выбора подходящих контрольных переменных для исследования (к примеру, можно непосредственно включить в регрессионную модель или реализовать мэтчинг на основе релевантных «третьих» факторов). Что мы уже знаем о контрольных переменных?

- Мы включаем их в регрессионную модель наряду с ключевыми предикторами для того, чтобы уменьшить смещение в оценках. Так, они играют вспомогательную роль, относительно контрольных переменных мы не формулируем гипотез. В частности, мы предполагаем, что их эффект на зависимую переменную однородный, контрольные характеристики задают некоторый общий контекст.
- Контрольные переменные мы, как правило, находим в литературе – предшествующих исследованиях, изучающих, какие факторы влияют на зависимую переменную, интересующую и нас.
- Контрольные переменные должны быть экзогенны не только по отношению к зависимой переменной, но и по отношению к ключевым предикторам. В противном случае можно «своими руками» привнести post-treatment bias.
- Ошибочно думать, что контрольные характеристики не должны быть никак скоррелированы с другими объясняющими переменными в модели. Для того, чтобы контрольные переменные выполняли свою задачу уменьшения смещения в оценках коэффициентов при ключевых объясняющих переменных, у контролей и ключевых предикторов, разумеется, должна быть совместная изменчивость.
- Даже если эффект каких-то отдельных контрольных переменных оказался незначимым, убирать их из модели не стоит. Они в комбинации с другими контролями могут корректировать эффект ключевых предикторов.

Когда Вы задумываете количественное исследование, прежде чем переходить к реализации эмпирической части, постарайтесь наглядно изобразить, как работает механизм связи между ключевыми переменными, и обозначить, какие вспомогательные переменные можно выделить, так или иначе связанные с ключевыми предикторами и зависимой переменной (примеры таких визуализаций в виде графов – см. ниже в текущем файле). Далее на основе полученного графа можно отобрать подходящие контрольные переменные. Прежде чем сформулировать правило такого отбора, разграничим переменные-confounders и переменные-colliders.

**Переменные-confounders** – переменные, которые влияют и на предиктор, и на зависимую переменную (см. ниже Рис. 1), их также можно называть «вмешивающимися», или «мешающими» переменными. Это явный претендент на включение в модель в качестве контрольной переменной. Дело в том, что выявление истинного каузального эффекта переменной-treatment на зависимую переменную осложняется ложной корреляцией, возникающей между ними из-за наличия «третьей» переменной, влияющей и на X, и на Y. К примеру, вспомним распространенный пример: мы наблюдаем, что при росте количества продаж мороженого растет и количество нападений акул на людей. Из этого, конечно, было бы ошибочно сделать вывод, что рост продаж мороженого положительно влияет на количество нападений акул. Просто есть третий фактор – летний период, в который вместе с повышением температуры увеличиваются продажи мороженого и купания людей, в том числе, и в океане, где есть опасность встретиться с акулами.

**Переменные-colliders** – переменные, которые являются следствием и предиктора, и зависимой переменной (или переменных, непосредственно связанных с ними). Соответствующий граф с переменной L в качестве коллайдера изображен ниже на Рис. 2. Можно представить себе мишень, в которую входят стрелы. Контролировать такую переменную нельзя, так как это приведет к ложной связи между переменными. К примеру, рассматриваем, связаны ли актерский талант и внешняя привлекательность. Изначально связи между этими переменными нет, однако если проконтролируем факт, является ли человек актером (актерский талант и внешняя привлекательность могут выступить факторами, определившими выбор профессии), то можем получить ложную связь между актерским талантом и внешней привлекательностью.

Теперь введем **«критерий черного хода»** в качестве правила, на основании которого можно по графу определить, какие переменные стоит контролировать, а какие - наоборот, было бы опасно. Для возможности интерпретировать результаты в терминах причинно-следственной связи и во избежание ложного вывода (к примеру, изначально связь отсутствует, а мы делаем вывод об отрицательном или положительном значимом эффекте) необходимо «заблокировать» все дополнительные пути (будем называть их «черным ходом») кроме основного, через которые от переменной treatment можно дойти до зависимой переменной. В свою очередь, «заблокировать» можно, проконтролировав любую переменную, появляющуюся как промежуточное звено на дополнительном пути – в «черном ходе», при условии, что данная переменная НЕ является переменной-collider или post-collider (то есть, следствием коллайдера).

Перед нами стоит следующая задача: определить, стоит ли включать ту или иную обозначенную на графе переменную как контрольную в модель. В качестве treatment на последующих графиках используется A / X, в качестве outcome – всегда Y.

Пройдемся последовательно по DAGs (directed acyclic graphs – название говорит о том, что связи однонаправленны, при этом в графе отсутствуют циклы).

Рис. 1: Стоит ли контролировать L?

$$\begin{tikzcd} L \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

*Комментарий*: это классический случай контрольной переменной: L влияет и на Y, и на X - confounder. От A можно дойти к Y через «черный ход»: A – L – Y. При этом L – промежуточное звено на этом дополнительном пути. Вывод: надо контролировать L.

Рис. 2: Стоит ли контролировать L?

$$\begin{tikzcd} A \arrow[rr, bend left] & Y \arrow[r] & \boxed{L} \end{tikzcd}$$

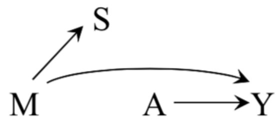
*Комментарий*: это классический случай переменной-коллайдера: изначально между A и Y нет связи, и переменная treatment, и outcome влияют на L. Если бы мы проконтролировали L, то между A и Y могла бы появиться ложная связь. Поэтому L контролировать не надо.

Рис. 3: Стоит ли контролировать M?

$$\begin{tikzcd} M \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

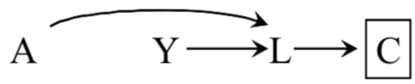
*Комментарий*: «черного хода» от переменной A к Y нет, поэтому M можно не проконтролировать. Между M и A нет совместной изменчивости, поэтому включение M никак не повлияет на оценку коэффициента при A. Можно проконтролировать, если хотим увеличить объяснительную силу модели. Однако на каузальный вывод о связи A и Y это никак не повлияет.

Рис. 4: Стоит ли контролировать S?



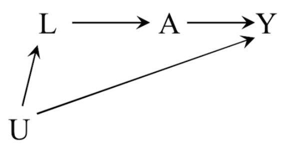

*Комментарий*: схожая ситуация, что и в предыдущем графе. Нейтрально, если нужно улучшить предсказание Y, а S тесно связан с M, то можно S использовать вместо M. Однако на каузальный вывод о связи A и Y это никак не повлияет.

Рис. 5: Стоит ли контролировать C?



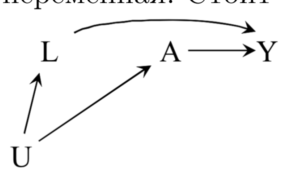

*Комментарий*: здесь C является post-collider, то есть, следствием коллайдера. Контролировать не надо, иначе появится ложная связь между A и Y.

Рис. 6: U - латентная переменная. Стоит ли контролировать L?



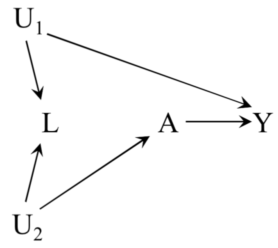

*Комментарий*: «черный ход» – A – L – U – Y. Блокируем его через L – контролируем переменную. Гипотетически можно было бы также заблокировать через U, но она латентная, у нас нет ее в массиве данных.

Рис. 7: U – латентная переменная. Стоит ли контролировать L?



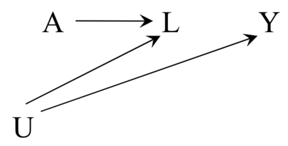

*Комментарий*: схожая ситуация с предыдущим графом, «черный ход» – A – U – L – Y. Блокируем его через L – контролируем переменную. Гипотетически можно было бы также заблокировать через U, но она латентная, у нас нет ее в массиве данных.

Рис. 8: U-переменные – латентные. Стоит ли контролировать L?



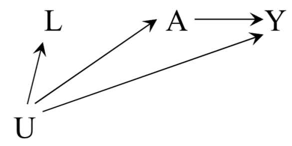

*Комментарий*: «черный ход» – A – $U_2$ – L – $U_1$ – Y. Переменная L в данном случае – коллайдер, на нее влияет и $U_1$, и $U_2$. Если бы ее проконтролировали, то открыли бы ложную связь между $U_1$ и $U_2$ и, соответственно, был бы открыт следующий «черный ход»: A – $U_2$ – $U_1$ – Y. Таким образом, L контролировать не надо.

Рис. 9: U – латентная переменная. Стоит ли контролировать L?



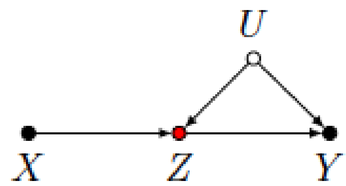

*Комментарий*: схожая ситуация с предыдущим графом, L – коллайдер. Если проконтролируем, то откроем «черный ход»: A – U – Y. Поэтому контролировать L не надо.

Рис. 10: U – латентная переменная. Стоит ли контролировать L?




*Комментарий*: на этом графе явный претендент на контроль – это переменная U. Однако U – латентная, и в массиве у нас ее нет, следовательно, как замещающую переменную можно использовать L, которая имеет совместную изменчивость с U.

Рис. 11: Стоит ли контролировать Z?




*Комментарий*: если проконтролировать Z, то частично учитывается U, а значит откроется backdoor X – U – Y. Это приведет к смещению каузального эффекта, поэтому контролировать не надо.