Диагностика регрессионных моделей

Построение регрессионных моделей – это многоступенчатый, итерационный процесс. Первая построенная модель в процессе статистического анализа, может оказаться не адекватной данным. Диагностика регрессионных моделей позволяет обнаружить несоответствие модели данным и наметить пути для дальнейшего улучшения построенной модели.

Регрессионная модель

Регрессионная модель – это модель вида $Y = X\beta + \varepsilon$. Параметр β может быть оценен при помощи метода наименьших квадратов. Далее, при условии нормальности $\varepsilon \sim N(0, \sigma^2 I)$, условия модели позволяют строить доверительные интервалы как для предсказаний, так и для параметров модели, а также сравнивать модели по F-критерию. Соответственно, для адекватности регрессионной модели должны быть выполнены следующие условия:

1. *Линейность модели*
2. *Нормальное распределение остатков*
3. *Одинаковое распределение остатков*
4. *Независимость остатков*

Проверка каждого из этих предположение производится при помощи соответствующей диагностической процедуры.

Остатки и плечи

Напомним, что:

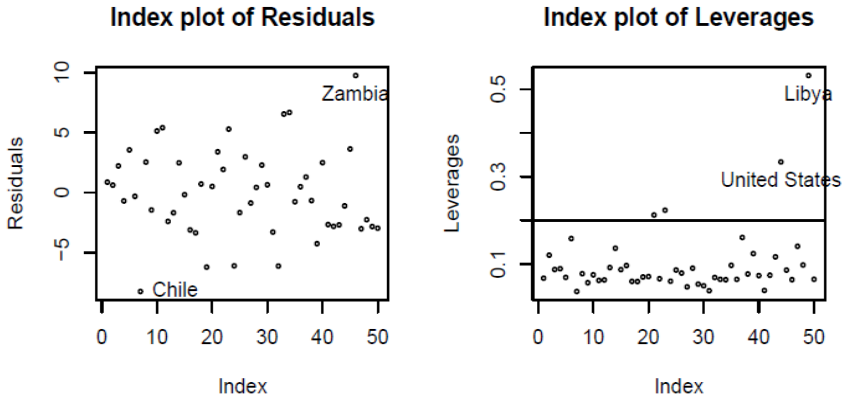
$$\hat{y} = X(X^T X)^{-1} X^T y = Hy$$

$$\varepsilon = y - \hat{y} = (I - H)\sigma^2$$

Здесь $h_i = H_{ii}$ – плечи.

Если плечо h_i большое, воздействие y_i на параметры модели будет велико.

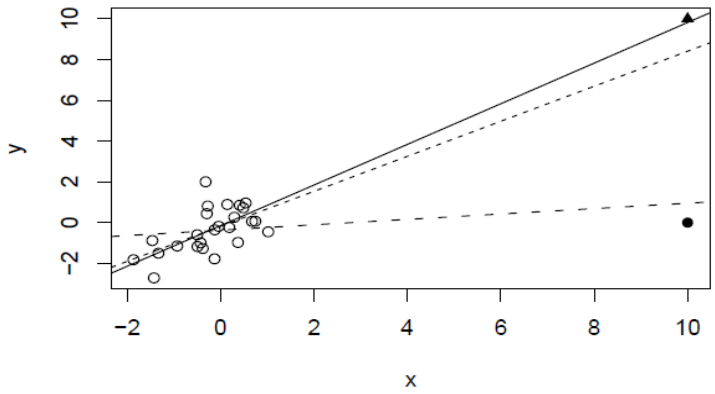
Если $h_i \geq 2p/n$, то необходима проверка значения y_i.



Пример 1. Значения Libya и United States требуют более пристального внимания, поскольку их влияние на параметры модели велико.

Тест на выбросы

Выбросом называется точка, значение которой не подходит к данной модели. Тест на выбросы – полезная процедура, поскольку позволяет отделить точки, являющиеся выбросами, от точек, остатки которых велики, но значение которых подходит к выбранной модели.



Пример 2. Точка, отмеченная кругом, является выбросом. Точка, отмеченная треугольником, не является выбросом. Однако, если мы построим регрессионную модель, включающую все точки, значение остатка для выброса будет невелико, а остаток точки, отмеченной треугольником, будет, наоборот, велико.

Таким образом, тест на выбросы заключается в следующем:

1. Исключаем *i*-ю точку
2. Проводим регрессионный анализ на данных без *i*-й точки.

Критерий: *если скорректированный остаток* $y_{(i)} - y_i$ *велик, значит i – выброс.*

Существует также точный тест на выбросы. Определим статистику Джекнайфа следующим образом:

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\sigma^2{}_{(i)}(1 + x_i{}^T\left(X_{(i)}{}^T X_{(i)}\right)x_i)^{1/2}}$$

Данная статистика распределена $t_i \sim t(n-p-1)$. Поэтому мы можем строить точные критерии на выбросы.

Для проверки модели на содержание хотя бы одного выброса применяется тест Бофферони. Смысл заключается в следующих рассуждениях: вероятность P(все точки не являются выбросами)=1-P(для данной модели в данных присутствует хотя бы один выброс)=1-nα, где α - уровень значимости в тесте Джекнайфа.

Замечания:

1. Если есть два или более выброса, один может скрыть другой
2. Выброс в одной модели может не быть выбросом в другой
3. Распределение остатков может не быть нормальным
4. Если выборка большого размера, выбросы представляют опасность только в том случае, когда они сгруппированы в кластер

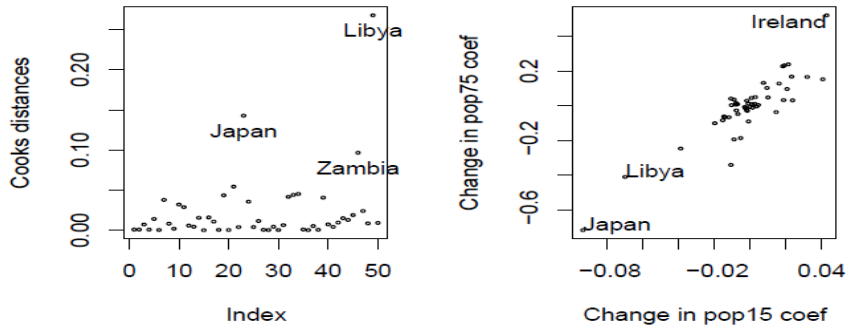
Влияющие наблюдения

Влияющим называется наблюдение, удаление которого из структуры данных влечет за собой существенные изменения в параметрах модели. Изменения бывают двух типов:

1. Изменение в β-коэффициентах модели
2. Изменение в Fit модели.

Критерием того, является ли наблюдение влияющим, служит статистика Кука:

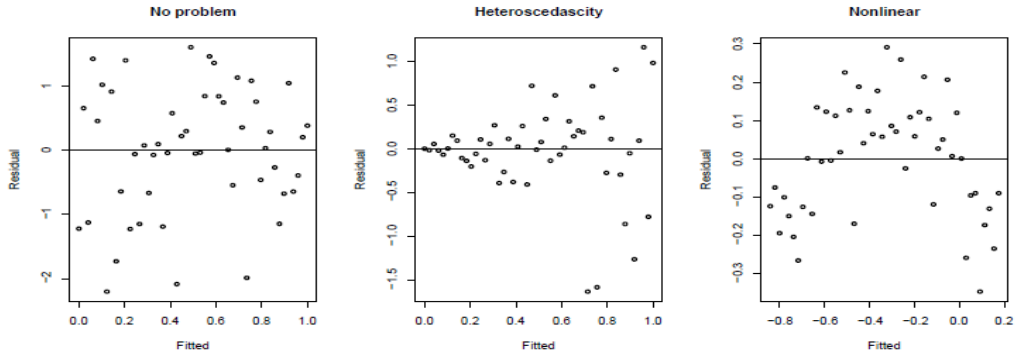
$$D_i = \frac{1}{p} r_i{}^2 \frac{h_i}{1 - h_i}$$



Пример 3. В приведенном примере наблюдения Japan и Libya являются влияющими.

Графики остатков

Наиболее эффективным и вместе с тем простым средством диагностики является построение графика распределения остатков. В приведенных ниже примерах показаны различные виды проблем регрессионных моделей.



Пример 4. Графики для различных видов «проблем» регрессионных моделей. На первом графике нет проблем, на втором – проблема неоднородной дисперсии, на третьем – проблема нелинейности модели.

Оценка нормальности

Для оценки данных модели на нормальность применяется тест Колмогорова-Смирнова:

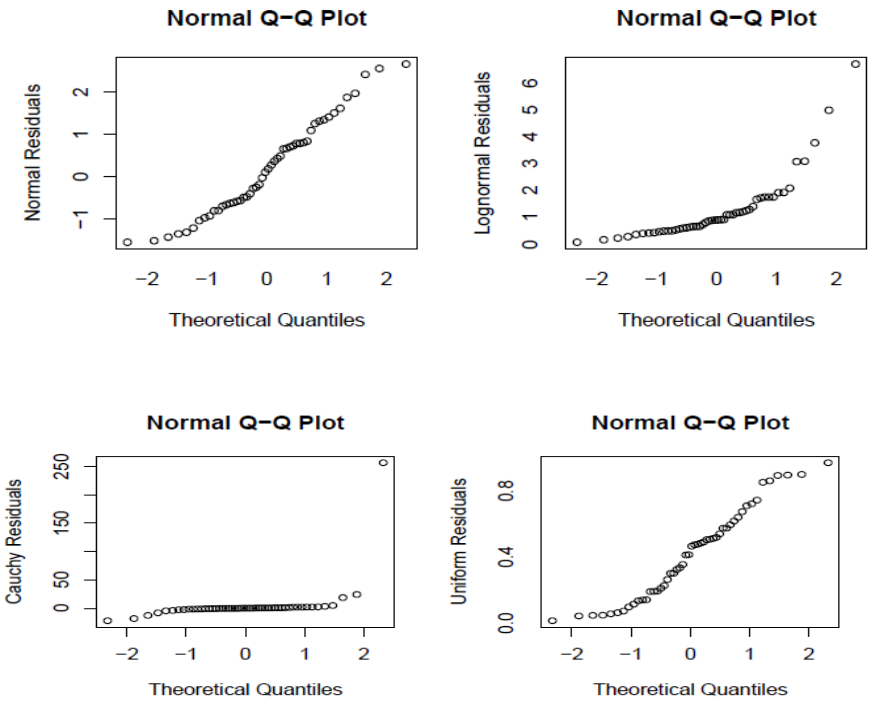
$$D_n = \sup |F_n(x) - F(x)|$$

Проблема теста состоит в том, что p-значение не дает нам представления о причинах ненормальности распределения.

Более содержательным тестом для проверки на нормальность, является график Q-Q plot. Его построение происходит в три этапа:

1. Сортировка остатков $\varepsilon_{[1]} \dots \varepsilon_{[n]}$
2. Вычисление $u_i = \Phi^{-1}(\frac{i}{n+i})$
3. Построение графика зависимости $\varepsilon_{[i]}$ от u_i

Если в итоге мы получим зависимость, близкую к прямой линии, это будет означать выполнение условия нормальности.



Пример 5. На первом рисунке представлено Q-Q plot для нормального распределения остатков, на втором – для логнормального, на третьем – для распределения Коши, на четвертом – для равномерного распределения. Видим существенное отклонение от прямой линии во всех случаях, кроме первого.

В том случае, когда модель не удовлетворяет условию нормальности, существуют следующие проблемы таких моделей:

- Оценка методом наименьших квадратов может быть неоптимальной, она останется лучшей в классе всех несмещенных оценок, но некоторые робастные методы могут оказаться эффективней
- Доверительные интервалы и тесты могут оказаться неверными. Тем не менее, проблему могут вызвать только распределения с «очень» тяжелыми хвостами. При увеличении выборки проблема нивелируется

Для нивелирования проблем, связанных с ненормальностью остатков, применяются следующие меры:

- Трансформация величин

- Использование других методов. Например, в случае с тяжелыми хвостами могут помочь робастные методы, которые придают меньшее значение выбросам
- Для распределений с короткими хвостами проблема с ненормальностью несущественна