Занятие 2

На занятии 1 мы выводили оценки коэффициентов в парной линейной регрессии. По построению регрессионной модели справедливо следующее:

1)
$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}\hat{\epsilon}_i = 0$$
Сумма остатков равна 0.

2)
$$\sum_{i=1}^{n}(y_i - \hat{y}_i)x_i = \sum_{i=1}^{n}\hat{\epsilon}_i x_i = 0$$
Скалярное произведение остатков и предиктора $=0$, то есть, предиктор и остатки нескоррелированы. Следовательно, проверить экзогенность посредством выгрузки коэффициента корреляции между объясняющими переменными и остатками не получится! Это справедливо по построению регрессионной модели

Тестирование значимости коэффициентов

На первом шаге, как всегда, формулируем нулевую гипотезу и альтернативу. Обратите внимание, что гипотезы формулируются относительно генеральных параметров, а не об оценках, оценки нам известны по выборочным данным:
$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$
На втором шаге нужно обозначить статистику и ее распределение при верной нулевой гипотезе.
$$\frac{\widehat{\beta}}{st.error(\widehat{\beta})} \overset{H_0}{\sim} t(df = n - k - 1)$$
, где n – количество наблюдений, k – количество предикторов. Так, к примеру, $df = n - 2$ справедливо только для случая, когда в модели один предиктор, так как в парной регрессии оцениваются 2 коэффициента: константа и коэффициент при предикторе. Для проверки гипотезы Вы можете использовать как фиксированный уровень значимости, так и p-value.

Разложение вариации

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$TSS = ESS + RSS$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Общий вид таблицы разложения вариации:

Source	df	Sum of Squares	Mean Square	F
x	k	ESS	$\frac{\text{ESS}}{k}$	$\frac{\text{ESS}/k}{\text{RSS}/(n-k-1)}$
Residual	$n-k-1$	RSS	$\frac{\text{RSS}}{n-k-1}$	
Total	$n-1$	TSS		

Мы проверяем гипотезу о незначимости коэффициента детерминации, или о том, что модель на константу не хуже, чем модель с предикторами.

$$H_0 : R^2 = 0$$

$$H_1 : R^2 > 0$$

При верной нулевой гипотезе $F \sim F(df_1 = k, df_2 = n-k-1)$
Проверяем гипотезу против односторонней альтернативы. Если p-value достаточно мал, делаем вывод в пользу значимого коэффициента детерминации.

Спецификация модели множественной линейной 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).

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

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

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

Условия Гаусса–Маркова

Для того, чтобы получить идентифицируемую модель множественной линейной регрессии с оценками BLUE (то есть, наиболее эффективными среди класса всех линейных несмещенных оценок), должен соблюдаться ряд условий:

1. модель оценивается на случайной выборке наблюдений (то есть, совокупность независимых одинаково распределенных сл.в.)
2. отсутствует строгая мультиколлинеарность (среди предикторов нет линейно зависимых, количество наблюдений превышается количество оцениваемых параметров)
3. модель линейна по параметрам
4. $E(e_i|x) = 0$ – экзогенность
5. $Var(e_i|x) = const$ – гомоскедастичность
6. $Cov(e_i, e_j|x) = 0$ – отсутствие автокорреляции

Важно, что данные условия именно об ошибках, а не остатках. Стоит отметить, что в литературе нет полной согласованности относительно списка данных условий. Более подробно об этом можно прочитать в статье Larocca, 2005