Проверка гипотез относительно параметров множественной линейной регрессии.

(1) Гипотеза о статистической незначимости коэффициентов уравнения регрессии.

Нулевая гипотеза H_0 формулируется в предположении о том, что теоретический коэффициент регрессионной модели $\beta_i, i=\overline{0,m}$ является статистически незначимым, альтернативная гипотеза H_1 – коэффициент модели β_i является статистически значимым:

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0$

Статистическая значимость параметров линейной регрессии с m факторами проверяется на основе t – статистики (статистика Стьюдента):

$t_{b_i} = \frac{b_i}{S_{b_i}} \sim t_{крит} = t(\alpha/2; n-m-1)$, где

b_i – коэффициент уравнения регрессии;

$S_{b_i} = \sqrt{S^2_{b_i}}$ – стандартная ошибка коэффициента b_i уравнения регрессии;

t_{b_i} – наблюдаемое значение t-статистики гипотезы;

$t_{крит} = t(\alpha/2; n-m-1)$ – значение критической точки распределения Стьюдента при уровне значимости α и значении степеней свободы $\nu = n-m-1$.

При проверке гипотезы H_0 о статистической незначимости коэффициента регрессии b_i, полученной значение наблюдаемой статистики t_{b_i} сравнивается со значением критической точки $t_{\text{крит}}$ распределения Стьюдента. Если $|t_{b_i}| < t_{\text{крит}}$, то нулевая гипотеза принимается и соответствующий параметр считается статистически незначимым, в противном случае, если $|t_{b_i}| \geq t_{\text{крит}}$ – нулевая гипотеза отклоняется в пользу альтернативной и параметр модели b_i статистически значим.

В том случае, когда статистически незначимый параметр соответствует экзогенному фактору, делается вывод, что b_i не отличается значимо от нуля, а значит, фактор x_i линейно не связан с результирующей переменной y. Его наличие среди экзогенных переменных не оправдано со статистической точки зрения. Не оказывая значимого влияния на зависимую переменную, он лишь искажает реальную картину взаимосвязи. Поэтому, после выявления статистической незначимости коэффициента b_i, переменную x_i предлагается исключить из уравнения линейной регрессии, т.к. это не приведет к существенному искажению качества модели, а сделает ее более точной.

Доверительные интервалы коэффициентов b_i, которые с надежностью $(1-\alpha)$ накрывают определяемые параметры β_i, определяются по формуле:

$$\left(b_i - t_{\frac{\alpha}{2};n-m-1} \cdot S_{b_i}; b_i + t_{\frac{\alpha}{2};n-m-1} \cdot S_{b_i}\right) \qquad (1.5)$$

Для того, чтобы определить, коэффициент при какой из переменных x_i не только статистически значим, но и какая переменная оказывает наибольшее влияние на изменение эндогенной переменной y, используют стандартизированные коэффициенты регрессии $\overline{b_i}$, характеризующие насколько изменится стандартное отклонение переменной y при изменении x_i на одно стандартное отклонение.

$$\overline{b_i} = b_i \frac{S_{x_i}}{S_y}.$$

Очевидно, что стандартизированные коэффициенты регрессии $\overline{b_i}$ связаны с понятием эластичности фактора y по фактору x_i в средней точке:

$$Э_i = Э_{y \text{ по } x_i} = b_i \frac{\overline{x}_i}{\overline{y}},$$

где $\overline{b_i}$ показывает на сколько величин отклонений S_y изменится в среднем эндогенная переменная y при увеличении i-ой экзогенной переменной x_i на одно стандартное отклонение S_{x_i}. Коэффициент эластичности в средней точке $Э_i$ показывает на сколько процентов от своей средней величины изменится значение эндогенной переменной y при увеличении экзогенной переменной x_i на один процент относительно своего среднего значения.

<u>Пример 1.</u>

Для показателей, подробное описание которых дано ниже, с помощью МНК оценена регрессионная зависимость и получены соответствующие значения статистических характеристик, которые далее использованы для выводов относительно значимости взаимосвязи и проверки некоторых гипотез.

Food – расходы на питание, тыс. руб.
Wage – размер заработной платы, тыс. руб.
Save – сбережения, тыс. руб.
Pay – обязательные платежи, тыс. руб.
n=100 (перекрестные данные, количество обследованных домохозяйств)

$$Food_t = -97{,}45 + 0{,}724 \cdot Wage_t - 0{,}14 \cdot Save_t - 0{,}358 \cdot Pay_t + e_t$$
$$\begin{array}{lllll} (s) & (45{,}93) & (0{,}04) & (0{,}08) & (0{,}29) \\ (t) & (-2{,}12) & (18{,}1) & (-1{,}75) & (-1{,}23) \\ (p) & (0{,}037) & (0{,}00) & (0{,}083) & (0{,}22) \end{array}$$
$$R^2 = 0{,}818 \quad F(R^2) = 143{,}824$$
$$t(\alpha/2; n-m-1) = t(0{,}025;96) = 1{,}985$$
$$F(\alpha; m; n-m-1) = F(0{,}05;3;96) = 2{,}699$$

Трактовка коэффициентов модели:

При увеличении размера заработной платы на 100 тыс. руб. – расходы на питание возрастут *более, чем на* 7 тыс. руб. (в точности – на 7 тысяч 240 рублей). При уменьшении обязательных платежей на 20000 рублей – расходы на питание увеличатся *примерно на* 7 тыс. руб. (в точности – на 7 тысяч 160 рублей). При росте сбережений на 70 тыс. рублей – расходы на питание снизятся *почти на* 10 тыс. руб. (в точности – на 9 тысяч 800 рублей).

При сравнении семей А и В: если в семье А размер заработной платы выше на 300 тыс. руб, в семье В размер обязательных платежей выше на 50 тыс. руб. и отсутствуют сбережения в отличие от семьи А, где откладывают около 70 тыс. руб. – в семье А тратят на питание примерно на 225 тыс. руб. больше, чем в семье В (в точности – на 225 тысяч 300 рублей).

Проверка гипотезы о статистической значимости коэффициента регрессии:

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$
$$T = \frac{b_0}{S_{b_0}} = \frac{-97,45}{45,93} = -2{,}12 => |T| > t(0,025;96) = 1{,}985 \Rightarrow H_1$$

Общий вывод: коэффициент b_0 статистически значим при уровне значимости α=0,05.

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$$T = \frac{b_1}{S_{b_1}} = \frac{0,724}{0,04} = 18,1 => |T| > t(0,025;96) = 1,985 \Rightarrow H_1$$

Общий вывод: коэффициент при переменной Wage статистически значим при уровне значимости α=0,05.

$H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$

$$T = \frac{b_2}{S_{b_2}} = \frac{-0,14}{0,08} = -1,75 => |T| < t(0,025;96) = 1,985 \Rightarrow H_0$$

Общий вывод: коэффициент при переменной Save статистически незначим при уровне значимости α=0,05.

$H_0 : \beta_3 = 0$

$H_1 : \beta_3 \neq 0$

$$T = \frac{b_3}{S_{b_3}} = -0,358 = -1,23 => |T| < t(0,025;96) = 1,985 \Rightarrow H_0$$

Общий вывод: коэффициент при переменной *Pay* статистически незначим при уровне значимости α=0,05.

Проверка гипотезы о статистической значимости коэффициента регрессии с помощью доверительных вероятностей:

$$\left[\begin{array}{l} H_0 : \alpha < P - value \\ H_0 : \alpha > P - value \end{array}\right.$$

β_0 – статистически значим при уровне значимости, начиная с α=0,04 (т.е. и для больших уровней – α=0,05 и α=0,10), т.к. α=0,04>P=0,037. При уровне значимости α=0,01 коэффициент β_0 статистически незначим, т.к. α=0,01<P=0,037.

β_1 – статистически значим при любом уровне значимости. Например, при α=0,01 (т.е. и для больших уровней – α=0,05 и α=0,10), т.к. α=0,01>P=0,00.

β_2 – статистически значим при уровне значимости, начиная с α=0,09 (т.е. и для большего уровня – α=0,10), т.к. α=0,09>P=0,08. При уровнях значимости α=0,01 и α=0,05 коэффициент β_2 статистически незначим, т.к. α=0,05<P=0,08 и тем более α=0,01<P=0,08.

β_3 – статистически незначим, т.к. может считаться значимым при уровне значимости, начиная с α=0,22, т.е. для уровней значимости α=0,01, α=0,05 и α=0,10 – коэффициент статистически незначим.

Эластичность эндогенного показателя в средней точке по каждому из экзогенных показателей находится по формулам (*при заданных значениях средних, т.е. 560; 70; 240 – по условию*):

$$Э_1 = b_1 \frac{\overline{Wage}}{\overline{Food}} = 0{,}724 \cdot \frac{560}{240} = 1{,}689\%$$

$$Э_2 = b_2 \frac{\overline{Save}}{\overline{Food}} = -0{,}14 \cdot \frac{70}{240} = -0{,}041\%$$

При росте заработной платы на 1% относительно среднего значения, расходы на питание вырастут на 1,689% относительно своего среднего значения. При увеличении заработной платы в два раза относительно среднего значения, расходы на питание возрастут более, чем в 2,5 раза относительно средних расходов на питание. И т.д.

Доверительный интервал для коэффициента угла наклона:

$$\beta_1 \in (0{,}724 - 1{,}985 \cdot 0{,}04; 0{,}724 + 1{,}985 \cdot 0{,}04) = (0{,}645; 0{,}803)$$

Определим, какая переменная x_i оказывает наибольшее влияние на изменение эндогенной переменной y, используя стандартизированные коэффициенты регрессии $\bar{b}_i$.

$$\bar{b}_i = b_i \frac{S_{x_i}}{S_y}.$$

$$\bar{b}_1 = b_1 \frac{S_{x_1}}{S_y}$$

Проверка гипотез об общем статистическом качестве модели множественной линейной регрессии.

(1) Гипотеза о статистической незначимости коэффициента детерминации R^2.

Нулевая гипотеза H_0 формулируется в предположении о том, что коэффициент детерминации R^2 регрессионной модели является статистически незначимым, альтернативная гипотеза H_1 – коэффициент детерминации статистически значим:

$H_0 : R^2 = 0$

$H_1 : R^2 \neq 0$

Статистическая значимость коэффициента детерминации модели линейной регрессии с m факторами проверяется на основе F – статистики (статистика Фишера), поскольку на самом деле речь идет о проверке предположения о равенстве объясненной и необъясненной дисперсий:

$$F = \frac{R^2 / m}{(1 - R^2)/(n - m - 1)} \sim F_{крит} = F(\alpha; m; n - m - 1),$$

где $F_{крит} = F(\alpha; m; n - m - 1)$ – значение критической точки распределения Фишера при уровне значимости α и значениях степеней свободы $\nu_1 == m$, $\nu_2 == n - m - 1$.

Если верна нулевая гипотеза, то статистическая незначимость коэффициента детерминации R^2 свидетельствует о совокупной статистической незначимости коэффициентов при экзогенных переменных, т.е. $\beta_1 = \beta_2 = ... = \beta_m = 0$, модель не может быть признана адекватной, ее дальнейший анализ и применение нецелесообразны. В противном случае, если верна гипотеза H_1,построенная модель статистически адекватна и ее общее качество может быть охарактеризовано непосредственно значением R^2.

Пример 2.

Для показателей, подробное описание которых дано ниже, с помощью МНК оценена регрессионная зависимость и получены соответствующие значения статистических характеристик, которые далее использованы для выводов относительно значимости взаимосвязи и проверки некоторых гипотез.

Food – расходы на питание, тыс. руб.
Wage – размер заработной платы, тыс. руб.
Save – сбережения, тыс. руб.
Pay – обязательные платежи, тыс. руб.
n=100 (перекрестные данные, количество обследованных домохозяйств)

$$Food_t = -97{,}45 + 0{,}724 \cdot Wage_t - 0{,}14 \cdot Save_t - 0{,}358 \cdot Pay_t + e_t$$

$$\begin{array}{lccccc} (s) & (45{,}93) & (0{,}04) & (0{,}08) & (0{,}29) \\ (t) & (-2{,}12) & (18{,}1) & (-1{,}75) & (-1{,}23) \\ (P) & (0{,}037) & (0{,}00) & (0{,}083) & (0{,}22) \end{array}$$

$R^2 = 0{,}818 \quad F(R^2) = 143{,}824$

$t(\alpha/2; n-m-1) = t(0{,}025;96) = 1{,}985$

$F(\alpha; m; n-m-1) = F(0{,}05;3;96) = 2{,}699$

$H_0 : R^2 = 0$

$H_1 : R^2 \neq 0$

$$F = \frac{0{,}818/3}{(1-0{,}818)/(100-3-1)} = 143{,}824 > F_{крит} = F(\alpha; m; n-m-1) = 2{,}699$$

Верна гипотеза H_1,построенная модель статистически адекватна и ее общее качество может быть охарактеризовано непосредственно значением R^2.

Расходы на питание *Food*, тыс. руб. на 81,8 % зависят от экзогенных переменных модели *Wage* – размер заработной платы, тыс. руб., *Save* – сбережения, тыс. руб., *Pay* – обязательные платежи, тыс. руб.
И на 18,2% от других неучтенных факторов.