НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

ЛЕКЦИЯ 1-5 НОЯБРЯ, 2 МОДУЛЬ

(ЧАСТЬ1, ДЛЯ САМОСТОЯТЕЛЬНОГО ИЗУЧЕНИЯ)

ПРОВЕРКА ГИПОТЕЗ ДЛЯ КОЭФФИЦИЕНТОВ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

ЛЕКТОР: ДЕМИДОВА О.А.

1-5.11.21

ПЛАН ЛЕКЦИИ

- Особенности регрессии без свободного члена
- Проверка гипотез для коэффициентов множественной регрессии

ОСОБЕННОСТИ РЕГРЕССИИ БЕЗ СВОБОДНОГО ЧЛЕНА

Если в модели регрессии нет свободного члена, то не выполняются свойства

1) $\sum_{i=1}^{n} e_i = 0$,

2) TSS = ESS + RSS,

3) $R^2 = 1 - \frac{RSS}{TSS}$,

4) В этом случае R^2 не является показателем качества подгонки регрессии (как и R^2_{adj}).

ПРОВЕРКА ГИПОТЕЗЫ О КОНКРЕТНОМ ЗНАЧЕНИИ КОЭФФИЦИЕНТОВ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

Модель $Y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$,

Основная гипотеза: $H_0: \beta_j = \beta_j^0$

Альтернативная гипотеза: $H_1: \beta_j \neq \beta_j^0$ или $H_1: \beta_j > \beta_j^0$ или $H_1: \beta_j < \beta_j^0$,

Тестовая статистика: $t = \frac{\widehat{\beta}_j - \beta_j^0}{s.e.(\widehat{\beta}_j)} \sim t(n-k-1)$,

Отличие от парной регрессии только в числе степеней свободы.

ПРОВЕРКА ЗНАЧИМОСТИ КОЭФФИЦИЕНТОВ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

Модель $Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i, i = 1, \ldots, n$

Основная гипотеза: $H_0: \beta_j = 0$

Альтернативная гипотеза: $H_1: \beta_j \neq 0$,

Тестовая статистика: $t = \frac{\widehat{\beta}_j}{s.e.(\widehat{\beta}_j)} \sim t(n-k-1)$,

Отличие от парной регрессии только в числе степеней свободы.

ПРОВЕРКА ЗНАЧИМОСТИ КОЭФФИЦИЕНТОВ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

```
. reg EARNINGS S EXP

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =   67.54
       Model |  22513.6473     2  11256.8237            Prob > F      =  0.0000
    Residual |  89496.5838   537  166.660305            R-squared     =  0.2010
-------------+------------------------------           Adj R-squared =  0.1980
       Total |  112010.231   539  207.811189            Root MSE      =   12.91

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   2.678125   .2336497    11.46   0.000     2.219146    3.137105
         EXP |   .5624326   .1285136     4.38   0.000     .3099816    .8148837
       _cons |  -26.48501    4.27251    -6.20   0.000    -34.87789   -18.09213
------------------------------------------------------------------------------
```

Изучается зависимость заработной платы от длительности обучения и опыта работы. t – статистики, p – value, доверительные интервалы рассчитываются аналогично случаю парной регрессии.

СВЯЗЬ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ С ПРОВЕРКОЙ ГИПОТЕЗ

Если β_j^0 попадает в $(1 - \alpha) \cdot 100\%$ доверительный интервал для коэффициента β_j, то на уровне значимости α

гипотеза: $H_0: \beta_j = 0$

при альтернативной гипотезе: $H_1: \beta_j \neq 0$

не отвергается.

ПРОВЕРКА ГИПОТЕЗЫ ОБ АДЕКВАТНОСТИ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

$$H_0: \beta_1 = \ldots = \beta_k = 0$$

$$H_1: \exists\, \beta_i \neq 0$$

Гипотезу H_0 о совместной незначимости коэффициентов при объясняющих факторах можно переформулировать следующим образом: выбранный набор объясняющих переменных не оказывает влияния на зависимую переменную Y.

ПРОВЕРКА ГИПОТЕЗЫ ОБ АДЕКВАТНОСТИ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

$$H_0: \beta_1 = \ldots = \beta_k = 0$$

$$H_1: \exists\, \beta_i \neq 0$$

$$F(k, n-k-1) = \frac{ESS/k}{RSS/(n-k-1)} = \frac{\frac{ESS}{TSS}/k}{\frac{RSS}{TSS}/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

Приведена тестовая статистика для проверки гипотезы о совместной незначимости коэффициентов при объясняющих факторах.

ПРОВЕРКА ГИПОТЕЗЫ ОБ АДЕКВАТНОСТИ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

```
. reg EARNINGS S EXP

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =   67.54
       Model |  22513.6473     2  11256.8237            Prob > F      =  0.0000
    Residual |  89496.5838   537  166.660305           R-squared     =  0.2010
-------------+------------------------------           Adj R-squared =  0.1980
       Total |  112010.231   539  207.811189           Root MSE      =   12.91

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   2.678125   .2336497    11.46   0.000     2.219146    3.137105
         EXP |   .5624326   .1285136     4.38   0.000     .3099816    .8148837
       _cons |  -26.48501   4.27251    -6.20   0.000    -34.87789   -18.09213
------------------------------------------------------------------------------
```

F – статистика для проверки гипотезы о совместной незначимости коэффициентов при выбранных объясняющих факторах выдается любым статистическим пакетом. Если p-value для этой F – статистики меньше выбранного уровня значимости, например, 0.05, то основная гипотеза отвергается, коэффициенты при выбранных факторах значимы в совокупности.

ПРИМЕР ПРОВЕРКИ ОБЩЕЙ ЛИНЕЙНОЙ ГИПОТЕЗЫ

$$S = \beta_0 + \beta_1 ASVABC + \beta_2 SM + \beta_3 SF + \varepsilon$$

Пример зависимости длительности обучения S от способностей индивида, характеризуемых обобщенной переменной ASVAVC, длительности обучения матери индивида SM и отца индивида SF.

ПРИМЕР ПРОВЕРКИ ОБЩЕЙ ЛИНЕЙНОЙ ГИПОТЕЗЫ

```
. reg S ASVABC SM SF

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  3,   536) =  104.30
       Model |  1181.36981     3  393.789935            Prob > F      =  0.0000
    Residual |  2023.61353   536  3.77539837            R-squared     =  0.3686
-------------+------------------------------           Adj R-squared =  0.3651
       Total |  3204.98333   539  5.94616574            Root MSE      =   1.943

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1257087   .0098533    12.76   0.000     .1063528    .1450646
          SM |   .0492424   .0390901     1.26   0.208     -.027546    .1260309
          SF |   .1076825   .0309522     3.48   0.001       .04688    .1684851
       _cons |   5.370631   .4882155    11.00   0.000      4.41158    6.329681
------------------------------------------------------------------------------
```

Коэффициент при переменной SM незначим.

Но это может быть следствием мультиколлинеарности (эта тема будет на одной из следующих лекций).

ПРИМЕР ПРОВЕРКИ ОБЩЕЙ ЛИНЕЙНОЙ ГИПОТЕЗЫ

$$S = \beta_0 + \beta_1 ASVABC + \beta_2 SM + \beta_3 SF + \varepsilon$$

$$H_0: \beta_2 = \beta_3$$

Проверим гипотезу об одинаковом влиянии обоих родителей, равенстве коэффициентов β_2 и β_3 .

ПРИМЕР ПРОВЕРКИ ОБЩЕЙ ЛИНЕЙНОЙ ГИПОТЕЗЫ

$$S = \beta_0 + \beta_1 ASVABC + \beta_2 SM + \beta_3 SF + \varepsilon$$

$$H_0: \beta_2 = \beta_3$$

$$S = \beta_0 + \beta_1 ASVABC + \beta_2 (SM + SF) + \varepsilon =$$
$$= \beta_0 + \beta_1 ASVABC + \beta_2 SP + \varepsilon,$$

$$\mathrm{SP} = \mathrm{SM} + \mathrm{SF}$$

Для этого инкорпорируем ограничение в уравнение регрессии, введя дополнительную переменную.

ПРИМЕР ПРОВЕРКИ ОБЩЕЙ ЛИНЕЙНОЙ ГИПОТЕЗЫ

```
. g SP=SM+SF

. reg S ASVABC SP

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =  156.04
       Model |  1177.98338     2  588.991689            Prob > F      =  0.0000
    Residual |  2026.99996   537  3.77467403            R-squared     =  0.3675
-------------+------------------------------           Adj R-squared =  0.3652
       Total |  3204.98333   539  5.94616574            Root MSE      =  1.9429

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1253106   .0098434    12.73   0.000     .1059743    .1446469
          SP |   .0828368   .0164247     5.04   0.000     .0505722    .1151014
       _cons |    5.29617   .4817972    10.99   0.000     4.349731    6.242608
------------------------------------------------------------------------------
```

Оцениваем вспомогательную регрессию.

ПРИМЕР ПРОВЕРКИ ОБЩЕЙ ЛИНЕЙНОЙ ГИПОТЕЗЫ

```
. reg S ASVABC SM SF

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1257087   .0098533    12.76   0.000     .1063528    .1450646
          SM |   .0492424   .0390901     1.26   0.208     -.027546    .1260309
          SF |   .1076825   .0309522     3.48   0.001       .04688    .1684851
       _cons |   5.370631   .4882155    11.00   0.000      4.41158    6.329681
------------------------------------------------------------------------------

. reg S ASVABC SP

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1253106   .0098434    12.73   0.000     .1059743    .1446469
          SP |   .0828368   .0164247     5.04   0.000     .0505722    .1151014
       _cons |    5.29617   .4817972    10.99   0.000     4.349731    6.242608
------------------------------------------------------------------------------
```

Сравниваем результаты оценивания двух регрессий. Если проверяемое ограничение имеет место, то сумма квадратов RSS должна увеличиться незначительно.

ПРИМЕР ПРОВЕРКИ ОБЩЕЙ ЛИНЕЙНОЙ ГИПОТЕЗЫ

$$S = \beta_0 + \beta_1 ASVABC + \beta_2 SM + \beta_3 SF + \varepsilon \quad (1)$$

$$H_0: \beta_2 = \beta_3, \quad H_1: \beta_2 \neq \beta_3$$

$$S = \beta_0 + \beta_1 ASVABC + \beta_2 (SM + SF) + \varepsilon$$
$$S = \beta_0 + \beta_1 ASVABC + \beta_2 SP + \varepsilon \quad (2)$$

Модель (2) является ограниченной версией модели (1).

ПРИМЕР ПРОВЕРКИ ОБЩЕЙ ЛИНЕЙНОЙ ГИПОТЕЗЫ

$$S = \beta_0 + \beta_1 ASVABC + \beta_2 SM + \beta_3 SF + \varepsilon \quad (1)$$

$$H_0: \beta_2 = \beta_3, \quad H_1: \beta_2 \neq \beta_3$$

$$S = \beta_0 + \beta_1 ASVABC + \beta_2 (SM + SF) + \varepsilon$$
$$S = \beta_0 + \beta_1 ASVABC + \beta_2 SP + \varepsilon \quad (2)$$

Проведем формальную проверку гипотезы, рассчитав значение тестовой статистики.

$$F(1, n-k-1) = \frac{(RSS_R - RSS_U)/1}{RSS_U/(n-k-1)} = \frac{2027.00 - 2023.61}{2023.61/536} = 0.90$$

ПРИМЕР ПРОВЕРКИ ОБЩЕЙ ЛИНЕЙНОЙ ГИПОТЕЗЫ

$$S = \beta_0 + \beta_1 ASVABC + \beta_2 SM + \beta_3 SF + \varepsilon \quad (1)$$

$$H_0: \beta_2 = \beta_3, \quad H_1: \beta_2 \neq \beta_3$$

$$S = \beta_0 + \beta_1 ASVABC + \beta_2 (SM + SF) + \varepsilon$$

$$S = \beta_0 + \beta_1 ASVABC + \beta_2 SP + \varepsilon \quad (2)$$

Проведем формальную проверку гипотезы, рассчитав значение тестовой статистики.

$$F(1, n-k-1) = \frac{(RSS_R - RSS_U)/1}{RSS_U/(n-k-1)} = \frac{2027.00 - 2023.61}{2023.61/536} = 0.90$$

Полученное значение тестовой статистики равно 0.9, что меньше критического значения при уровне значимости 5%, $F^{cr}_{0.05}(1, 536) = 1$. Следовательно, нулевая гипотеза не отвергается.

ПРОВЕРКА ОБЩЕЙ ЛИНЕЙНОЙ ГИПОТЕЗЫ

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

H_0: имеют место q конкретных линейных ограничений на коэффициенты регрессии.
H_1: эти ограничения не имеют места.

Чтобы проверить выполнение ограничений, необходимо:

1) Оценить регрессию без ограничений и найти RSS_{UR} (UR – unrestricted).
2) Оценить регрессию с ограничениями и найти RSS_R (R – unrestricted).
3) Вычислить соответствующую тестовую F – статистику.

ПРОВЕРКА ОБЩЕЙ ЛИНЕЙНОЙ ГИПОТЕЗЫ

$$Y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

H_0: имеют место q конкретных линейных ограничений на коэффициенты регрессии.
H_1: эти ограничения не имеют места.

Тестовая F – статистика:

$$F = \frac{(RSS_R - RSS_U)/q}{RSS_U/(n-k-1)} \sim F(q; n-k-1)$$

Если при выбранном уровне значимости α значение тестовой F - статистики больше, чем F_α^{cr}(q, n-k-1), то гипотеза H_0 отвергается. Если значение тестовой F – статистики меньше, чем F_α^{cr}(q, n-k-1), то H_0 не отвергается.

Спасибо за внимание!