НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

ЛЕКЦИИ 6-7, 1 МОДУЛЬ

МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

ЛЕКТОР: ДЕМИДОВА О.А.

8.10.21, 15.10.21

ПЛАН ЛЕКЦИЙ 6-7

- Множественная линейная регрессия в скалярной и матричной формах
- Метод наименьших квадратов и его геометрическая интерпретация в многомерном случае. Система нормальных уравнений
- Матричное выражение для вектора оценок коэффициентов регрессии
- Ковариационная матрица оценок коэффициентов регрессии
- Теорема Гаусса-Маркова для множественной линейной
- Показатели качества подгонки множественной регрессии
- Особенности регрессии без свободного члена
- Проверка гипотез для коэффициентов множественной регрессии

МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИИ

Общий вид модели множественной регрессии:

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

$X_1,\ldots,X_k$ – факторы (независимые переменные),

Y – зависимая переменная,

ε_i, $i = 1,\ldots,n$ – возмущения,

n – число наблюдений.

МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИИ

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

Обозначим $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{pmatrix},\ X_1 = \begin{pmatrix} X_{11} \\ X_{12} \\ \ldots \\ X_{1n} \end{pmatrix},\ \ldots, X_k = \begin{pmatrix} X_{k1} \\ X_{k2} \\ \ldots \\ X_{kn} \end{pmatrix},\ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{pmatrix}$

Тогда уравнение регрессии можно переписать в векторном виде

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

МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИИ

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

Если ввести матрицу наблюдений X размера (n x (k + 1)) и вектор коэффициентов β размера ((k + 1) x 1),

$$X = \begin{pmatrix} 1 & X_{11} & \ldots & X_{k1} \\ 1 & X_{12} & \ldots & X_{k2} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & X_{1n} & \ldots & X_{kn} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_k \end{pmatrix}$$

то уравнение регрессии можно переписать в матричном виде:

$Y = X\beta + \varepsilon$

МНК ДЛЯ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

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

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki},\ i = 1, \ldots, n$

$e_i = Y_i - \hat{Y}_i,\ i = 1, \ldots, n$

$$RSS = \sum_{i=1}^{n} e_i^2 \to \min$$

$Y = X\beta + \varepsilon$

$\hat{Y} = X\hat{\beta}$

$e = Y - \hat{Y} = Y - X\hat{\beta}$

МНК ДЛЯ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$$RSS = \sum_{i=1}^{n} e_i^2 \to \min$$

$$Y = X\beta + \varepsilon$$

$$\hat{Y} = X\hat{\beta}$$

$$e = Y - \hat{Y} = Y - X\hat{\beta}$$

$$e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}, \sum_{i=1}^{n} e_i^2 = (e_1, \dots, e_n) \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = e'e$$

МНК ДЛЯ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$$RSS(\hat{\beta}) = e'e = (Y - X\hat{\beta})'(Y - X\hat{\beta}) =$$

$$= (Y' - \hat{\beta}'X')(Y - X\beta) = Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$$

$$\underset{(1\times n)}{Y'}\underset{(n\times(k+1))}{X}\underset{((k+1)\times 1)}{\hat{\beta}} \Rightarrow (Y'X\hat{\beta}) = (Y'X\hat{\beta})' = \hat{\beta}'X'Y$$

$$RSS(\hat{\beta}) = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

МНК ДЛЯ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$RSS(\hat{\beta}) = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \to \min$

Лемма 1. Если a, b – n – мерные вектора, то $\frac{\partial(b'a)}{\partial b} = \text{a}$, где $\frac{\partial(\dots)}{\partial b} = \begin{pmatrix} \frac{\partial(\dots)}{\partial b_1} \\ \dots \\ \frac{\partial(\dots)}{\partial b_n} \end{pmatrix}$.

Лемма 2. Если b - n – мерный вектор, A – матрица размера (n x n), то $\frac{\partial(b'Ab)}{\partial b} = 2\text{Ab}$.

Необходимое условие экстремума: $\frac{\partial RSS(\widehat{\beta})}{\partial \widehat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$.

Достаточное условие минимума выполнено, т.к.

Матрица Гессе $H = \frac{\partial^2 RSS(\widehat{\beta})}{\partial \widehat{\beta} \partial \widehat{\beta}'} = 2X'X$ является положительно определенной.

МНК ДЛЯ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$$\frac{\partial RSS(\hat{\beta})}{\partial \hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$$

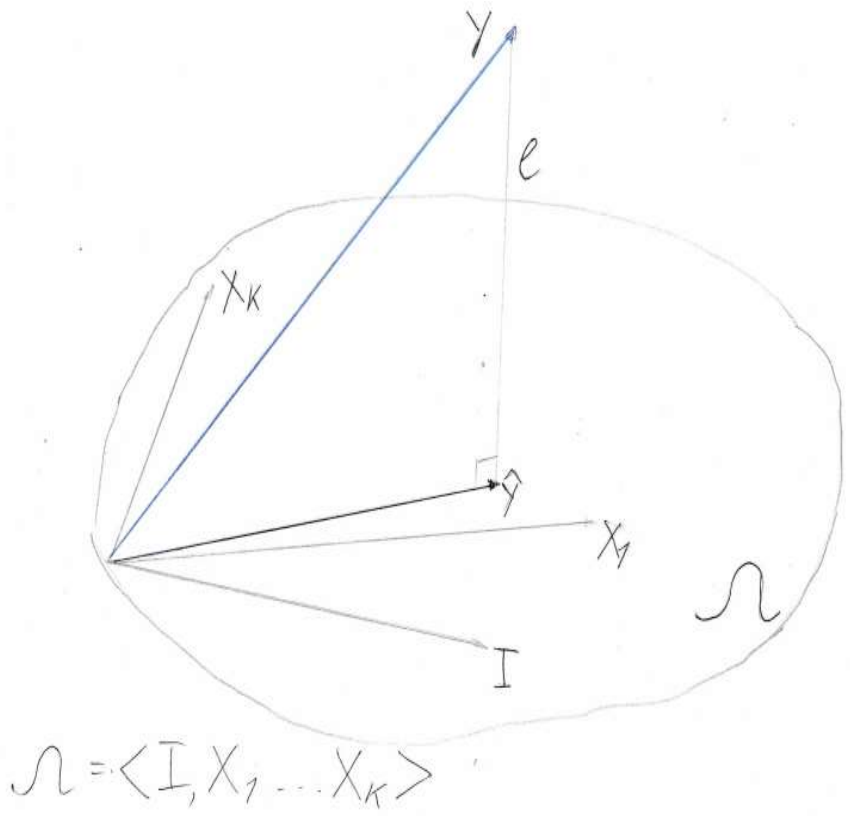
Система нормальных уравнений для множественной линейной регрессии:

$$X'X\hat{\beta} = X'Y$$

Решение системы нормальных уравнений:

$\hat{\beta} = (X'X)^{-1}X'Y$ - это МНК – оценки коэффициентов β.

ГЕОМЕТРИЧЕСКИЙ ВЫВОД ФОРМУЛЫ ОЦЕНОК МНК



$$Y = \hat{Y} + e \Leftrightarrow |e| \to \min \Leftrightarrow$$
$$|e|^2 \to \min$$
$$|e| \to \min \Rightarrow e \perp \Omega$$
$$\Rightarrow e \perp X_j \ \forall\ j = 1, \dots, k \Rightarrow$$
$$X_j' e = 0$$
$$X' e = 0$$
$$X'(Y - \hat{Y}) = 0$$
$$X'(Y - X\hat{\beta}) = 0$$
$$X'Y - X'X\hat{\beta} = 0$$
$$X'X\hat{\beta} = X'Y$$
$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$$

ТЕОРЕМА ГАУССА-МАРКОВА ДЛЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Если модель множественной линейной регрессии

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

1) Правильно специфицирована,

2) Не существует линейной связи между регрессорами,

3) Возмущения имеют нулевое математическое ожидание $E(\varepsilon_i) = 0$,

4) Дисперсии возмущений одинаковы $var(\varepsilon_i) = \sigma_\varepsilon^2$,

5) Возмущения с разными номерами не коррелируют, $cov(\varepsilon_i, \varepsilon_j) = 0$

Тогда оценки МНК являются BLUE (Best Linear Unbiased Estimator).

ТЕОРЕМА ГАУССА-МАРКОВА ДЛЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Estimator – оценка,

Unbiased – несмещенная,

Linear – по Y,

Best – это оценки с наименьшей дисперсией в классе всех линейных несмещенных оценок

ТЕОРЕМА ГАУССА-МАРКОВА ДЛЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Estimator – оценка (т.е. по набору наблюдений мы получаем оценки параметров регрессии)

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

$$\hat{\beta} = (X'X)^{-1} X'Y,$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \ldots \\ \hat{\beta}_k \end{pmatrix}.$$

ТЕОРЕМА ГАУССА-МАРКОВА ДЛЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Unbiased – несмещенная (т.е. математические ожидания оценок коэффициентов регрессии совпадают с истинными значениями параметров).

$$\hat{\beta}_{OLS}(Y) = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y} = (\mathrm{X}'\mathrm{X})^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon$$

$$E(\hat{\beta}_{OLS}) = E(\beta + (X'X)^{-1}X'\varepsilon) = \beta + E[(X'X)^{-1}X'\varepsilon] = \beta + 0 = \beta$$

ТЕОРЕМА ГАУССА-МАРКОВА ДЛЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Linear – по Y.

$$\hat{\beta}_{OLS}(Y) = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}$$

1.1. $\hat{\beta}_{OLS}(Y_1 + Y_2) = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'(\mathrm{Y}_1 + Y_2) = \hat{\beta}_{OLS}(Y_1) + \hat{\beta}_{OLS}(Y_2)$

1.2. $\hat{\beta}_{OLS}(\alpha Y) = \alpha\hat{\beta}_{OLS}(Y), \alpha - const$

ТЕОРЕМА ГАУССА-МАРКОВА ДЛЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Best – это оценки с наименьшей дисперсией в классе всех линейных несмещенных оценок (без доказательства).

$\hat{\beta}_{OLS}(Y) = (\mathrm{X'X})^{-1}\mathrm{X'Y}$

Утверждение. Ковариационная матрица оценок МНК:

$\mathrm{Var}(\hat{\beta}_{OLS}) = \sigma_{\varepsilon}^{2}\,(\mathrm{X'X})^{-1}$.

Вспомогательное утверждение. Если A – постоянная матрица, а Y – случайный вектор, то ковариационная матрица $\mathrm{var(AY)} = \mathrm{Avar(Y)A'}$.

ТЕОРЕМА ГАУССА-МАРКОВА ДЛЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

$$\hat{\beta}_{OLS}(Y) = (X'X)^{-1}X'Y$$

Утверждение. Ковариационная матрица оценок МНК:

$$\mathrm{Var}(\hat{\beta}_{OLS}) = \sigma_\varepsilon^2\,(X'X)^{-1}.$$

Доказательство.

$$\mathrm{Var}(\hat{\beta}_{OLS}) = \mathrm{Var}[(X'X)^{-1}X'Y] = (X'X)^{-1}X'\mathrm{Var}(Y)((X'X)^{-1}X')' =$$

$$= (X'X)^{-1}X'\,\sigma_\varepsilon^2 I\,X\,(X'X)^{-1} = \sigma_\varepsilon^2\,(X'X)^{-1}.$$

ТЕОРЕМА ГАУССА-МАРКОВА ДЛЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

$$\mathrm{Var}(\hat{\beta}_{OLS}) = \sigma_{\varepsilon}^{2}\,(\mathrm{X}'\mathrm{X})^{-1},$$

$$\mathrm{var}(\hat{\beta}_j) = \sigma_{\varepsilon}^{2}(X'X)_{jj}^{-1}, j = 0, \ldots, k.$$

Для множественной регрессионной модели вычисляется ковариационная матрица. Ее диагональными элементами являются дисперсии оценок коэффициентов.

ОЦЕНКИ ДИСПЕРСИЙ ОЦЕНОК КОЭФФИЦИЕНТОВ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

$$\hat{\sigma}_{\varepsilon}^{2} = \frac{RSS}{n-k-1}$$

$$\widehat{var}(\hat{\beta}_j) = \hat{\sigma}_{\varepsilon}^{2}(X'X)_{jj}^{-1}, j = 0, \ldots, k.$$

Приведены формулы для вычисления оценок дисперсий оценок коэффициентов регрессии. Но дисперсия ошибок регрессии неизвестна. Поэтому используется ее оценка (верхняя формула). Подставив вместо дисперсии ошибок регрессии ее оценку, получаем оценку дисперсии коэффициентов.

СТАНДАРТНЫЕ ОШИБКИ ОЦЕНОК КОЭФФИЦИЕНТОВ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

$$s.e.(\hat{\beta}_j) = \sqrt{\hat{\sigma}_\varepsilon^2 (X'X)_{jj}^{-1}}, j = 0, \ldots, k$$

Однако вместо оценок дисперсий коэффициентов обычно используются корни квадратные из них, называемые стандартными ошибками (standard errors). Они выдаются статистическими пакетами.

СТАНДАРТНЫЕ ОШИБКИ ОЦЕНОК КОЭФФИЦИЕНТОВ МНОЖЕСТВЕННОЙ 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
------------------------------------------------------------------------------
```

ДИСПЕРСИОННЫЙ АНАЛИЗ ДЛЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

$$\sum(Y-\bar{Y})^2=\sum(\hat{Y}-\bar{Y})^2+\sum e^2\,,\mathrm{TSS}=\mathrm{ESS}+\mathrm{RSS}$$

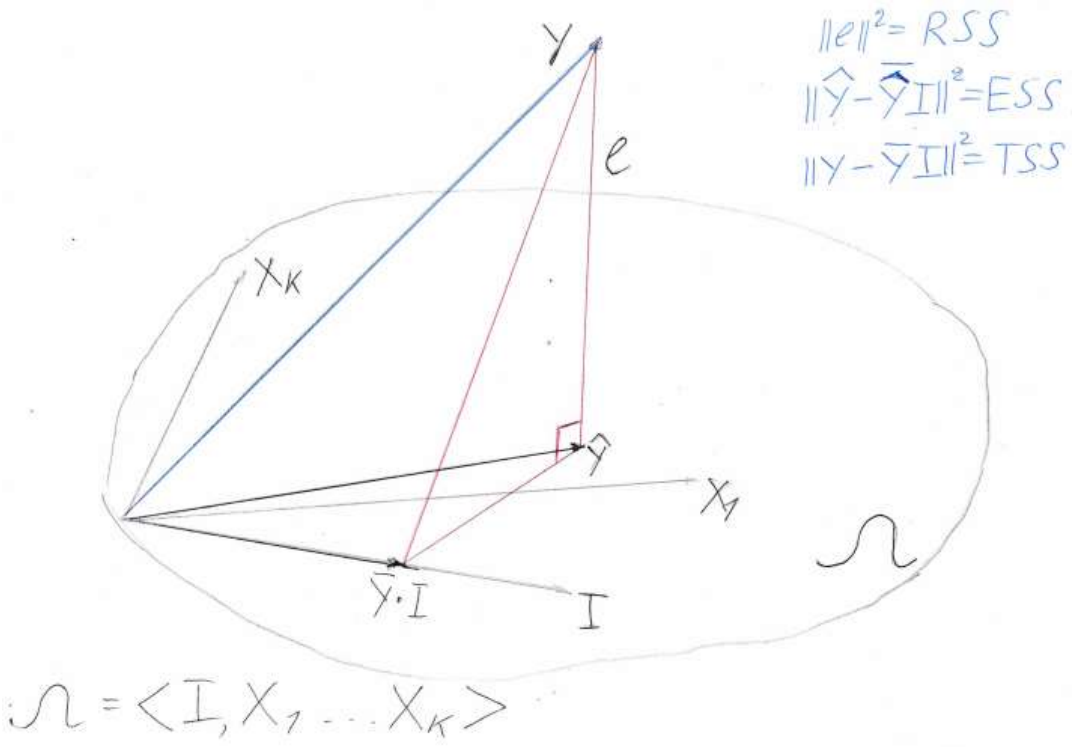
$TSS=\sum(Y-\bar{Y})^2$ - Total sum of squares - общая сумма квадратов отклонений зависимой переменной от среднего значения,

$ESS=\sum(\hat{Y}-\bar{Y})^2$ - Explained sum of squares – объясненная с помощью регрессии сумма квадратов отклонений,

$RSS=\sum e^2$ - Residual sum of squares – сумма квадратов остатков.

ДИСПЕРСИОННЫЙ АНАЛИЗ ДЛЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$$



КОЭФФИЦИЕНТ МНОЖЕСТВЕННОЙ ДЕТЕРМИНАЦИИ

$$R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum(\hat{Y}_i - \bar{Y})^2/(n-1)}{\sum(Y_i - \bar{Y})^2/(n-1)} = \frac{\widehat{\mathrm{var}}(\hat{Y})}{\widehat{\mathrm{var}}(Y)}$$

R^2 - показатель качества подгонки (оценки) регрессии, является отношением ESS к TSS, (или долей дисперсии Y, объясненной с помощью регрессии).

Это неотрицательная и не превышающая 1 величина.

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum {e_i}^2}{\sum(Y_i - \bar{Y})^2}$$

Минимизируя сумму квадратов остатков, мы максимизируем R^2. Таким образом, чем ближе R^2 к 1, тем выше качество подгонки (оценки) регрессии.

КОЭФФИЦИЕНТ МНОЖЕСТВЕННОЙ ДЕТЕРМИНАЦИИ

R^2 действительно является квадратом, а именно, квадратом выборочного коэффициента корреляции Y и Ŷ (без доказательства).

$R^2 = \frac{ESS}{TSS} = \sum_{i=1}^{k} \hat{\beta}_i \frac{\widehat{cov}(X_j,Y)}{\widehat{var}(Y)}$ - факторное разложение R^2.

Доказательство.

$$ESS = TSS - RSS = y'y - e'e,$$

$$RSS = e'e = (y - x\hat{\beta}_{-0})'(y - x\hat{\beta}_{-0}) = y'y - \hat{\beta}'_{-0}x'y - y'x\hat{\beta}_{-0} + \hat{\beta}_{-0}x'x\hat{\beta}_{-0}$$

$$ESS = \hat{\beta}'_{-0}x'y + \hat{\beta}'_{-0}\underbrace{\left(x'y - x'x\hat{\beta}_{-0}\right)}_{=0} = \hat{\beta}'_{-0}x'y.$$

СКОРРЕКТИРОВАННЫЙ КОЭФФИЦИЕНТ МНОЖЕСТВЕННОЙ ДЕТЕРМИНАЦИИ

R^2 имеет существенный недостаток: он возрастает при добавлении любого регрессора X (т.к. при этом уменьшается сумма квадратов остатков).

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

R^2_{adj} - это R^2 , скорректированный с учетом числа степеней свободы. Он «наказывает» за включение лишних регрессоров.

СКОРРЕКТИРОВАННЫЙ КОЭФФИЦИЕНТ МНОЖЕСТВЕННОЙ ДЕТЕРМИНАЦИИ

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

Для сравнения качества подгонки (оценки) двух регрессий с одной и той же зависимой переменной Y используется коэффициент множественной детерминации, скорректированный на число степеней свободы R^2_{adj}.

R^2_{adj} не превышает 1, но может быть меньше 0.

СКОРРЕКТИРОВАННЫЙ КОЭФФИЦИЕНТ МНОЖЕСТВЕННОЙ ДЕТЕРМИНАЦИИ

```
. 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
------------------------------------------------------------------------------
```

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 ЗНАЧИМОСТИ КОЭФФИЦИЕНТОВ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

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

Основная гипотеза: $H_0: \beta_j = 0$

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

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

$$H_0: \beta_2 = \beta_3$$

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%, $F_{0.05}^{cr}(1, 536) = 1$. Следовательно, нулевая гипотеза не отвергается.

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.

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.

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Спасибо за внимание!