ПРОВЕРКА КАЧЕСТВА РЕГРЕССИИ

Лекции 9-10

СИСТЕМА ПОКАЗАТЕЛЕЙ КАЧЕСТВА ЛИНЕЙНОЙ РЕГРЕССИИ

1. Показатели качества параметров регрессии

2. Показатели качества уравнения регрессии в целом

ПОКАЗАТЕЛИ КАЧЕСТВА ПАРАМЕТРОВ РЕГРЕССИИ

1. Стандартные ошибки оценок (анализ точности определения оценок).
2. Значения t-статистик (проверка гипотез относительно коэффициентов регрессии).
3. Интервальные оценки коэффициентов линейного уравнения регрессии.
4. Доверительные области для зависимой переменной.

РАСПРЕДЛЕНИЕ ОСТАТКОВ РЕГРЕССИИ

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \in N\left(\mathbf{0}; \sigma^2 \mathbf{I}\right)$$

$$\mathbf{e} \in \mathrm{N}\left(\mathbf{0}; \sigma^2 \mathbf{M}\right) \sim \mathrm{N}\left(\mathbf{0}; s^2 \mathbf{M}\right)$$

$$\mathbf{M} = \mathbf{I} - \mathbf{P} = \mathbf{I} - \frac{\mathbf{1}}{N}\mathbf{X}\left(\frac{\mathbf{X}'\mathbf{X}}{N}\right)^{-1}\mathbf{X}' \xrightarrow[N\to\infty]{} \mathbf{I}$$

$$e_t \sim N(0, \sigma^2) \Rightarrow \frac{e_t}{\sigma} \sim N(0,1) \Rightarrow \xi_t = \frac{e_t^{\,2}}{\sigma^2} \sim \chi_1^2$$

РАСПРЕДЛЕНИЕ ОЦЕНКИ ДИСПЕРСИИ ОШИБОК

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \in N\left(\mathbf{0}; \sigma^2 \mathbf{I}\right)$$

$$e_t \sim N(0, \sigma^2) \Rightarrow \frac{e_t}{\sigma} \sim N(0,1) \Rightarrow \xi_t = \frac{e_t^2}{\sigma^2} \sim \chi_1^2$$

$$\eta = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \frac{\sum_t e_t^2}{\sigma^2} = \frac{RSS}{\sigma^2} = \frac{(N-k)s^2}{\sigma^2} \in \chi^2_{N-k}, \quad s^2 = \frac{RSS}{N-k}$$

$$\mathrm{E}\left(\frac{(N-k)s^2}{\sigma^2}\right) = \mathrm{E}\,\eta = N-k$$

$$\mathrm{E}\left(\frac{s^2}{\sigma^2}\right) = 1 \Leftrightarrow \mathrm{E}\,s^2 = \sigma^2$$

т.е. оценка s^2 дисперсии σ^2 несмещенная

РАСПРЕДЛЕНИЕ ОЦЕНКИ ДИСПЕРСИИ ОШИБОК

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \in N\left(\mathbf{0}; \sigma^2 \mathbf{I}\right)$$

$$\eta = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \frac{\sum_t e_t^2}{\sigma^2} = \frac{RSS}{\sigma^2} = \frac{(N-k)s^2}{\sigma^2} \in \chi^2_{N-k}$$

$$\operatorname{plim}_{N\to\infty} \frac{s^2}{\sigma^2} - 1 = \operatorname{plim}_{N\to\infty} \frac{\eta}{N-k} - 1 = \operatorname{plim}_{N\to\infty} \frac{\eta - \left(N-k\right)}{N-k} = 0$$

оценка дисперсии ошибок- состоятельна

ТЕСТИРОВАНИЕ ГИПОТЕЗ О ДИСПЕРСИИ ОШИБОК С ПОМОЩЬЮ СТАТИСТИКИ ПИРСОНА

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \in N\left(\mathbf{0}; \sigma^2 \mathbf{I}\right)$$

$$\eta = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \frac{\sum_t e_t^2}{\sigma^2} = \frac{RSS}{\sigma^2} = \frac{(N-k)s^2}{\sigma^2} \in \chi^2_{N-k}$$

ПОСТРОЕНИЕ ДОВЕРИТЕЛЬНОГО ИНТЕРВАЛА ДЛЯ ДИСПЕРСИИ ОШИБОК

$$\frac{RSS}{\sigma^2} \in \left[K^-_{N-k}\left(\frac{1-\alpha}{2}\right); K^+_{N-k}\left(\frac{\alpha}{2}\right) \right]$$

$$\sigma^2 \in \left[\frac{RSS}{K^+_{N-k}\left(\frac{\alpha}{2}\right)}; \frac{RSS}{K^-_{N-k}\left(\frac{1-\alpha}{2}\right)} \right]$$

ПОСТРОЕНИЕ ДОВЕРИТЕЛЬНОГО ИНТЕРВАЛА ДЛЯ СКО ОШИБОК

$$\sigma \in \left[\sqrt{\frac{RSS}{K^{+}_{N-k}(0.05)}}; \sqrt{\frac{RSS}{K^{-}_{N-k}(0.95)}} \right]_{0.10} =$$

$$= \left[\sqrt{\frac{46362}{42.557}}; \sqrt{\frac{46362}{17.708}} \right] = [33.01; 51.17]$$

- Dependent Variable: RENT/NO
- Method: Least Squares
- Date: 10/11/08 Time: 16:06
- Sample: 1 32
- Included observations: 32

Variable	Coefficient	Std. Error	t-Statistic	Prob.
RM/NO	132.0648	37.64941	3.507751	0.0015
DIST	-1.032338	0.550585	-1.874985	0.0709
C	32.69061	34.24517	0.954605	0.3477

R-squared	0.326268	Mean dependent var	138.1693
Adjusted R-squared	0.279804	S.D. dependent var	47.11474
S.E. of regression	39.98362	Akaike info criterion	10.30388
Sum squared resid	46362.00	Schwarz criterion	10.44129
Log likelihood	-161.8620	F-statistic	7.021925
Durbin-Watson stat	2.245847	Prob(F-statistic)	0.003259

РАСПРЕДЕЛЕНИЕ ОЦЕНОК КОЭФФИЦИЕНТОВ

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \in N\left(\mathbf{0}; \sigma^2 \mathbf{I}\right)$$

$$\hat{\boldsymbol{\beta}} \in \mathrm{N}\left(\boldsymbol{\beta}; \frac{\sigma^2}{N}\mathbf{Q}_N^{-1}\right) \Rightarrow \mathbf{z} = \mathbf{Q}_N^{1/2}\frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{\sigma}\sqrt{N} \in \mathrm{N}\left(\mathbf{0}; \mathbf{I}\right), \quad \mathbf{Q}_N = \left(\frac{\mathbf{X}'\mathbf{X}}{N}\right)$$

это – многомерная **z**-статистика,

с ее помощью строятся доверительные области

и тестируются гипотезы- H_0: $\boldsymbol{\beta} = \boldsymbol{\theta}$

$$\hat{\beta}_i \in \mathrm{N}\left(\beta_i; \frac{\sigma^2}{N}\left(\frac{\mathbf{X}'\mathbf{X}}{N}\right)^{-1}_{ii}\right) \Rightarrow z_i = \frac{\hat{\beta}_i - \beta_i}{\mathrm{s.d.}\hat{\beta}_i} = \frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{\left[\mathbf{Q}_N^{-1}\right]_{ii}}}\sqrt{N} \in \mathrm{N}(0;1)$$

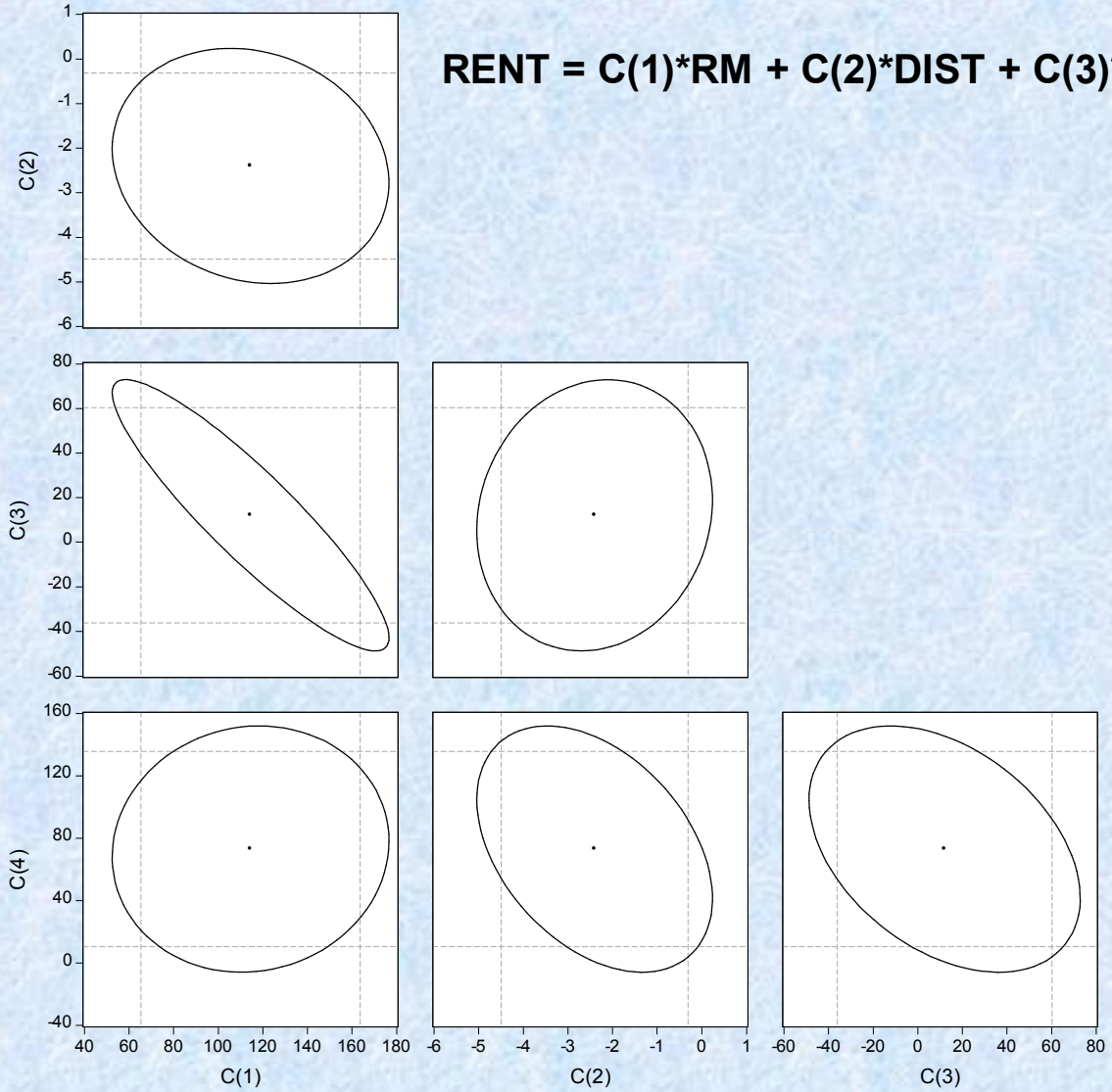
это – индивидуальная z_i -статистика

для построения доверительного интервала для коэффициента β_i

и тестирования гипотез-H_0: $\beta_i = \theta$

ПОСТРОЕНИЕ ДОВЕРИТЕЛЬНЫХ ОБЛАСТЕЙ

RENT = C(1)*RM + C(2)*DIST + C(3)*NO + C(4)



РАСПРЕДЛЕНИЕ ОЦЕНОК КОЭФФИЦИЕНТОВ

в асимптотике

Если $\boldsymbol{\varepsilon} \notin \mathrm{N}(0;\sigma^2\mathbf{I})$,

то можно получить такие же результаты в асимптотике основываясь на ЦПТ

$$\hat{\boldsymbol{\beta}} \in P\left(\boldsymbol{\beta};\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right) \sim P\left(\boldsymbol{\beta};s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right);$$

$$\mathbf{e} \in P\left(\mathbf{0};\sigma^2\mathbf{M}\right) \sim P\left(\mathbf{0};s^2\mathbf{M}\right)$$

$$\hat{\boldsymbol{\beta}} \sim P\left(\boldsymbol{\beta};\frac{\sigma^2}{N}\mathbf{Q}_N^{-1}\right) \Rightarrow \left(\mathbf{ЦПТ}\right):\mathbf{z}=\left(\mathbf{Q}_N^{-1}\right)^{-1/2}\frac{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}}{\sigma}\sqrt{N} \sim \mathrm{N}\left(\mathbf{0};\mathbf{I}\right)$$

$$\hat{\beta}_i \sim P\left(\beta_i;\frac{\sigma^2}{N}\left[\mathbf{Q}_N^{-1}\right]_{ii}\right) \Rightarrow \left(\mathbf{ЦПТ}\right):z_i=\frac{\hat{\beta}_i-\beta_i}{\mathrm{s.d.}\hat{\beta}_i}=\frac{\hat{\beta}_i-\beta_i}{\sigma\sqrt{\left[\mathbf{Q}_N^{-1}\right]_{ii}}}\sqrt{N} \sim \mathrm{N}(0;1)$$

СТАНДАРТНЫЕ ОТКЛОНЕНИЯ И СТАНДАРТНЫЕ ОШИБКИ ОЦЕНОК КОЭФФИЦИЕНТОВ

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\hat{\boldsymbol{\beta}} \in P\left(\boldsymbol{\beta}; \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right) \sim \mathrm{N}\left(\boldsymbol{\beta}; s^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right), \quad \mathbf{Q}_N = \left(\frac{\mathbf{X}'\mathbf{X}}{N}\right)$$

$$\mathbf{e} \in P\left(\mathbf{0}; \sigma^2\mathbf{M}\right) \sim \mathrm{N}\left(\mathbf{0}; s^2\mathbf{M}\right), \quad s^2 = \frac{RSS}{N-k}$$

$$\mathrm{Var}\,\hat{\boldsymbol{\beta}} = \frac{\sigma^2}{N}\mathbf{Q}_N^{-1}, \quad \widehat{\mathrm{Var}}\hat{\boldsymbol{\beta}} = \frac{s^2}{N}\mathbf{Q}_N^{-1}$$

$$\mathrm{Var}\,\hat{\beta}_i = \frac{\sigma^2}{N}\left[\mathbf{Q}_N^{-1}\right]_{ii}, \quad \widehat{\mathrm{Var}}\hat{\beta}_i = \frac{s^2}{N}\left[\mathbf{Q}_N^{-1}\right]_{ii}$$

$$\mathrm{s.d.}\hat{\beta}_i = \sqrt{\mathrm{Var}\,\hat{\beta}_i} = \frac{\sigma\sqrt{\left[\mathbf{Q}_N^{-1}\right]_{ii}}}{\sqrt{N}}, \quad \mathrm{s.e.}\hat{\beta}_i = \sqrt{\widehat{\mathrm{Var}}\hat{\beta}_i} = \frac{s\sqrt{\left[\mathbf{Q}_N^{-1}\right]_{ii}}}{\sqrt{N}}$$

СТАНДАРТНЫЕ ОШИБКИ КОЭФФИЦИЕНТОВ

формула для расчета

выборочных дисперсий

эмпирических коэффициентов регрессии:

$$s^2_{\hat{\beta}_i} = s^2 q'_{ii}, \quad i = \overline{1,k}$$

Здесь $q'_{ii}, \quad i = \overline{1,k}$ – диагональные элементы матрицы

$$(\mathbf{X}'\mathbf{X})^{-1}$$

СТАНДАРТНЫЕ ОШИБКИ КОЭФФИЦИЕНТОВ

стандартные ошибки коэффициентов:

$$s.e._{\hat{\beta}_i} = \sqrt{s^2_{\hat{\beta}_i}}, \quad j = \overline{0,k}$$

стандартная ошибка регрессии:

$$RMSE = \sqrt{s^2} = \sqrt{\frac{RSS}{N-k}}$$

ИСПОЛЬЗОВАНИЕ СТАНДАРТНЫХ ОШИБОК

Сравнивая значение коэффициента с его стандартной ошибкой, можно судить о значимости коэффициента

Коэффициент называется **значимым**, если есть достаточно высокая вероятность того, что его истинное значение отлично от нуля

НЕЗНАЧИМОСТЬ КОЭФФИЦИЕНТА НЕ ЯВЛЯЕТСЯ ДОСТАТОЧНЫМ АРГУМЕНТОМ ДЛЯ ИСКЛЮЧЕНИЯ СООТВЕТСВУЮЩЕГО РЕГРЕССОРА ИЗ МОДЕЛИ

Для стандартных ошибок оценок нет таблиц критических уровней – для точного суждения используются *t*-статистики

ТЕСТИРОВАНИЕ ГИПОТЕЗ О КОЭФФИЦИЕНТАХ С ПОМОЩЬЮ t-СТАТИСТИК

s.d. известно

Используем z-статистику $$z = \frac{\hat{\beta}_i - \beta_i^0}{\text{s.d.}\hat{\beta}_i}$$

s.d. не известно

Используем t-статистику $$t = \frac{\hat{\beta}_i - \beta_i^0}{\text{s.e.}\hat{\beta}_i}$$

$$t_i = \frac{z_i}{\sqrt{\frac{RSS/\sigma^2}{N-k}}} = \frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{\left[\mathbf{Q}_N^{-1}\right]_{ii}}} \frac{\sigma}{\sqrt{\frac{RSS}{N-k}}}\sqrt{N} =$$

$$= \frac{\hat{\beta}_i - \beta_i}{s\sqrt{\left[\mathbf{Q}_N^{-1}\right]_{ii}}}\sqrt{N} = \frac{\hat{\beta}_i - \beta_i}{\text{s.e.}\hat{\beta}_i} \sim \text{St}_{N-k}$$

ЗНАЧИМОСТЬ КОЭФФИЦИЕНТОВ РЕГРЕССИИ

Значимость коэффициентов множественной регрессии проверяется по *t*-критерию Стьюдента:

$$|t_i| = \frac{|\hat{\beta}_i|}{s.e._{\hat{\beta}_i}} > t_{\frac{\alpha}{2};N-k}$$

$$t_i = \frac{\hat{\beta}_i}{s.e._{\hat{\beta}_i}}$$

– расчетное значение *t*-статистики коэффициента b_j

t-тесты обеспечивают проверку значимости предельного вклада каждой переменной при допущении, что все остальные переменные уже включены в модель

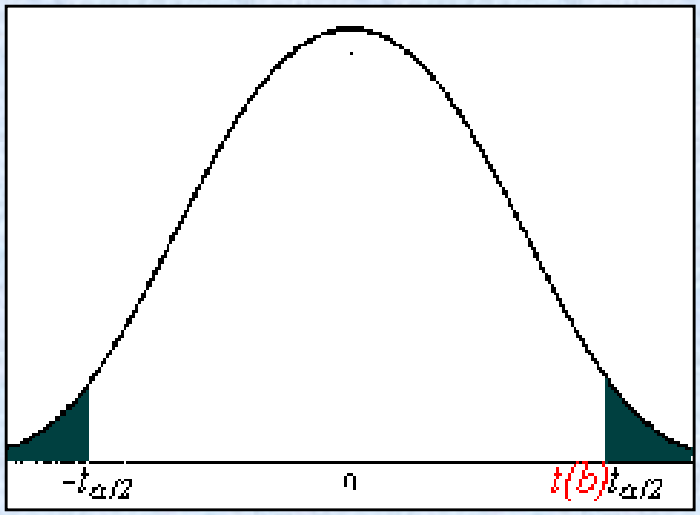
Незначимость коэффициента регрессии не всегда может служить основанием для исключения соответствующей переменной из модели

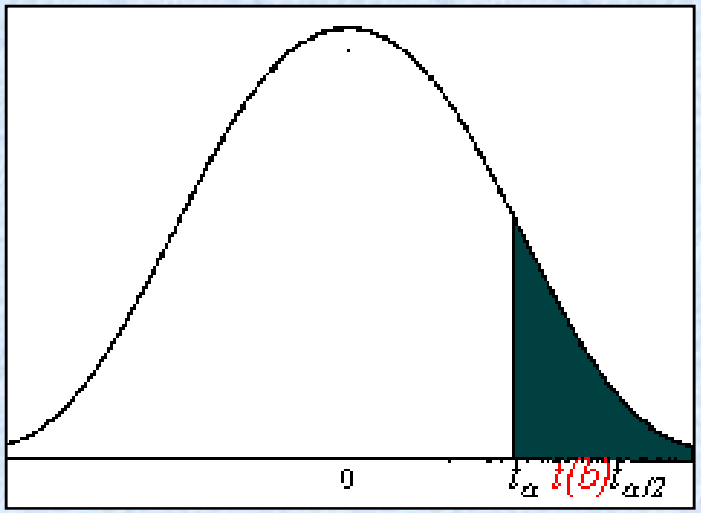
Порядок работы при проверке значимости коэффициента по *t*-статистике

1. Выбираем уровень значимости α (обычно 1% , 5% или 10%).
2. Вычисляем число степеней свободы (N–k).
3. По таблицам распределения Стьюдента определяем критическое значение $t_{\alpha/2;\, N\text{-}k}$ (двухсторонний критерий) или $t_{\alpha;\, N\text{-}k}$ (односторонний критерий).
4. Если модуль *t*-статистики больше критического значения, то коэффициент является значимым на уровне значимости α.
5. В противном случае коэффициент не значим (на данном уровне α).

Использование односторонних гипотез для проверки значимости коэффициентов

Использование односторонних гипотез иногда позволяет «спасти» значимость коэффициентов регрессии при том же уровне значимости





Это требует обязательного экономического обоснования

ПРОВЕРКА ЗНАЧИМОСТИ

Dependent Variable: RENT/NO
Method: Least Squares
Date: 10/11/08 Time: 16:06
Sample: 1 32
Included observations: 32

Variable	Coefficient	Std. Error	t-Statistic	Prob.
RM/NO	132.0648	37.64941	3.507751	0.0015
DIST	-1.032338	0.550585	-1.874985	0.0709
C	32.69061	34.24517	0.954605	0.3477

Стандартная ошибка коэффициента при ***RM/NO*** мала по сравнению с самим коэффициентом

Стандартная ошибка коэффициента при ***константе*** велика по сравнению с самим коэффициентом

Стандартная ошибка коэффициента при ***DIST*** по сравнению с самим коэффициентом может быть признана как малой, так и большой

ПРОВЕРКА ЗНАЧИМОСТИ

Dependent Variable: RENT/NO
Method: Least Squares
Date: 10/11/08 Time: 16:06
Sample: 1 32
Included observations: 32

Variable	Coefficient	Std. Error	t-Statistic	Prob.
RM/NO	132.0648	37.64941	3.507751	0.0015
DIST	-1.032338	0.550585	-1.874985	0.0709
C	32.69061	34.24517	0.954605	0.3477

Коэффициент при ***RM/NO*** **высоко значим** (как минимум на уровне 1.6%)

Коэффициент при ***константе*** **существенно незначим** (по крайней мере на 34% уровне значимости)

Коэффициент при ***DIST*** **незначим на 5%** уровне значимости, но **значим на 10%** уровне значимости (или 5% при односторонней гипотезе)

Правило оценки значимости коэффициентов регрессии без использования таблиц

1. Если $|t_{b_i}| \le 1$, то коэффициент b_i не м.б. признан значимым, т.к. доверительная вероятность менее 0,7.
2. Если $1 < |t_{b_i}| \le 2$, то найденная оценка может рассматриваться как относительно (слабо) значимая. При этом доверительная вероятность лежит между 0,7 и 0,95.
3. Если $2 < |t_{b_i}| \le 3$, то коэффициент значим. Доверительная вероятность лежит между значениями 0,95 и 0,99.
4. Если $|t_{b_i}| > 3$, то это почти полная гарантия значимости коэффициента.

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ДЛЯ КОЭФФИЦИЕНТОВ РЕГРЕССИИ

$$\hat{\beta}_i - t_{\frac{\alpha}{2};\,N-k}\, s.e.\hat{\beta}_i < \beta_i < \hat{\beta}_i + t_{\frac{\alpha}{2};\,N-k}\, s.e.\hat{\beta}_i$$

Данный доверительный интервал накрывает с надежностью $(1-\alpha)$ истинное значение коэффициента регрессии

ПРОВЕРКА ЗНАЧИМОСТИ

Dependent Variable: RENT/NO
Method: Least Squares
Date: 10/11/08 Time: 16:06
Sample: 1 32
Included observations: 32

Variable	Coefficient	Std. Error	t-Statistic	Prob.
RM/NO	132.0648	37.64941	3.507751	0.0015
DIST	-1.032338	0.550585	-1.874985	0.0709
C	32.69061	34.24517	0.954605	0.3477

$$\left[\hat{\beta}^{-}_{DIST};\hat{\beta}^{+}_{DIST}\right]_{0.95}=\left[\hat{\beta}_{DIST}-t_{32-3}(0.05)\cdot s.e.\hat{\beta}_{DIST};\hat{\beta}_{DIST}+t_{32-3}(0.05)\cdot s.e.\hat{\beta}_{DIST}\right]=$$

$$=\left[-1.03-2.04\cdot 0.55;-1.03+2.04\cdot 0.55\right]=\left[-2.152;0.092\right]$$

$$\left[\hat{\beta}^{-}_{DIST};\hat{\beta}^{+}_{DIST}\right]_{0.90}=\left[\hat{\beta}_{DIST}-t_{32-3}(0.1)\cdot s.e.\hat{\beta}_{DIST};\hat{\beta}_{DIST}+t_{32-3}(0.1)\cdot s.e.\hat{\beta}_{DIST}\right]=$$

$$=\left[-1.03-1.70\cdot 0.55;-1.03+1.70\cdot 0.55\right]=\left[-1.965;-0.095\right]$$

ПРОВЕРКА ГИПОТЕЗЫ О КОЭФФИЦИЕНТЕ

Estimation Equation:

==========

RENT/NO = C(1)*RM/NO + C(2)*DIST + C(3)

Substituted Coefficients:

==========

RENT/NO = 132.0647502*RM/NO - 1.032338386*DIST + 32.69060899

Wald Test:
Equation: EQ02

Test Statistic	Value	df	Probability
F-statistic	0.003450	(1, 29)	0.9536
Chi-square	0.003450	1	0.9532

Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
1 + C(2)	-0.032338	0.550585

$$\mathrm{H}_0 : C(2) = -1$$

$$\mathrm{H}_1 : C(2) < -1$$

$$t_{DIST} = \left| \frac{-1.03 - (-1)}{0.55} \right| = 0.0545$$

$$t_{32-3}(0.05) = 1.699 \ (\text{одностороннее})$$

$$t_{DIST} < t_{32-3}(0.05)$$

порядок работы при проверке гипотезы о коэффициенте по доверительному интервалу

1. Выбираем уровень значимости α (обычно 1% , 5% или 10%).
2. Вычисляем число степеней свободы (*N*–k).
3. По таблицам распределения Стьюдента определяем критическое значение $t_{\alpha/2;\, N-k}$ (двухсторонний критерий).
4. Вычисляем границы доверительного интервала.
5. Если проверяемая точка не лежит внутри доверительного интервала, то коэффициент является значимым на уровне значимости α.
6. В противном случае коэффициент не значим (на данном уровне α).

ДОВЕРИТЕЛЬНЫЕ ОБЛАСТИ ДЛЯ ЗАВИСИМОЙ ПЕРЕМЕННОЙ

Одной из центральных задач эконометрики является прогнозирование значений зависимой переменной при определенных значениях объясняющих переменных.

Здесь возможны два варианта:

1. Предсказать условное математическое ожидание зависимой переменной при определенных значениях объясняющих переменных (***предсказание среднего значения***).
2. Предсказать некоторое конкретное значение зависимой переменной (***предсказание конкретного значения***).

Предсказание среднего значения зависимой переменной

Пусть построено уравнение регрессии $\overline{Y}(X_i) = \sum_{m=1}^{k} X_{im} \hat{\beta}_m$

На его основе необходимо предсказать условное м. о.

$$\mathrm{E}(\mathrm{Y} \mid \mathrm{X} = X_i) = \sum_{m=1}^{k} X_{im} \beta_m$$

переменной *Y* при $X = x_i$.

Вопрос: Как сильно может уклониться значение $\overline{y}(x_p)$ от $\mathrm{E}(\mathrm{Y} \mid \mathrm{X} = X_i)$

Предсказание среднего значения зависимой переменной

Доверительная область для условного м. о. E[$Y|X = x_i$]:

$$\hat{\mathrm{E}}\left(Y \mid \mathrm{X} = X_i\right) = \hat{Y}\left(X_i\right)$$

$$\widehat{\mathrm{var}}\left(\mathrm{E}\left(Y \mid \mathrm{X} = X_i\right)\right) = \widehat{\mathrm{var}}\left(\hat{\mathrm{E}}\left(Y \mid \mathrm{X} = X_i\right)\right) = \widehat{\mathrm{V\,ar}}\left(\hat{Y}\left(X_i\right)\right)$$

Предсказание среднего значения зависимой переменной

Доверительная область для условного м. о. $\mathrm{E}[Y|X = x_i]$:

$$\hat{\mathrm{E}}\left(Y \mid \mathrm{X} = X_i\right) = \hat{Y}\left(X_i\right)$$

$$\widehat{\mathrm{var}}\left(\mathrm{E}\left(Y \mid \mathrm{X} = X_i\right)\right) = \widehat{\mathrm{var}}\left(\hat{\mathrm{E}}\left(Y \mid \mathrm{X} = X_i\right)\right) = \widehat{\mathrm{Var}}\left(\hat{Y}\left(X_i\right)\right) =$$

$$= \widehat{\mathrm{var}}\left(\mathbf{X}_\mathbf{i}\hat{\boldsymbol{\beta}}\right) = \widehat{\mathrm{var}}\left(\sum_{m=1}^{k} X_{im}\hat{\beta}_m\right) =$$

Предсказание среднего значения зависимой переменной

Доверительная область для условного м. о. E[$Y|X = x_i$]:

$$\hat{\mathrm{E}}\left(Y \mid \mathrm{X} = X_i\right) = \hat{Y}\left(X_i\right)$$

$$\widehat{\mathrm{var}}\left(\mathrm{E}\left(Y \mid \mathrm{X} = X_i\right)\right) = \widehat{\mathrm{var}}\left(\hat{\mathrm{E}}\left(Y \mid \mathrm{X} = X_i\right)\right) = \widehat{\mathrm{V\,ar}}\left(\hat{Y}\left(X_i\right)\right) =$$

$$= \widehat{\mathrm{var}}\left(\mathbf{X_i}\hat{\boldsymbol{\beta}}\right) = \widehat{\mathrm{var}}\left(\sum_{m=1}^{k} X_{im}\hat{\beta}_m\right) =$$

$$= \sum_{m=1}^{k} X_{im}\,\widehat{\mathrm{v\,ar}}\left(\hat{\beta}_m\right) + 2\sum_{m_1<m_2} X_{im_1}\,\mathrm{cov}\left(\hat{\beta}_{m_1};\hat{\beta}_{m_2}\right) =$$

Предсказание среднего значения зависимой переменной

Доверительная область для условного м. о. $\mathrm{E}[Y|X = x_i]$:

$$\hat{\mathrm{E}}\left(Y \mid \mathrm{X} = X_i\right) = \hat{Y}\left(X_i\right)$$

$$\widehat{\mathrm{var}}\left(\mathrm{E}\left(Y \mid \mathrm{X} = X_i\right)\right) = \widehat{\mathrm{var}}\left(\hat{\mathrm{E}}\left(Y \mid \mathrm{X} = X_i\right)\right) = \widehat{\mathrm{V\,ar}}\left(\hat{Y}\left(X_i\right)\right) =$$

$$= \widehat{\mathrm{var}}\left(\mathbf{X_i}\hat{\boldsymbol{\beta}}\right) = \widehat{\mathrm{var}}\left(\sum_{m=1}^{k} X_{im}\hat{\beta}_m\right) =$$

$$= \sum_{m=1}^{k} X_{im}\,\widehat{\mathrm{v\,ar}}\left(\hat{\beta}_m\right) + 2\sum_{m_1<m_2} X_{im_1}\,\mathrm{cov}\left(\hat{\beta}_{m_1};\hat{\beta}_{m_2}\right) =$$

$$= \mathbf{X_i}\,\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right)\mathbf{X_i'}$$

Предсказание среднего значения зависимой переменной

Доверительная область для условного м. о. $\mathrm{E}[Y|X = x_i]$:

$$\hat{\mathrm{E}}\left(Y \mid \mathrm{X} = X_i\right) = \hat{Y}\left(X_i\right)$$

$$\widehat{\mathrm{var}}\left(\mathrm{E}\left(Y \mid \mathrm{X} = X_i\right)\right) = \widehat{\mathrm{var}}\left(\hat{\mathrm{E}}\left(Y \mid \mathrm{X} = X_i\right)\right) = \widehat{\mathrm{Var}}\left(\hat{Y}\left(X_i\right)\right) =$$

$$= \widehat{\mathrm{var}}\left(\mathbf{X_i}\hat{\boldsymbol{\beta}}\right) = \widehat{\mathrm{var}}\left(\sum_{m=1}^{k} X_{im}\hat{\beta}_m\right) =$$

$$= \sum_{m=1}^{k} X_{im}\,\widehat{\mathrm{var}}\left(\hat{\beta}_m\right) + 2\sum_{m_1<m_2} X_{im_1}\,\mathrm{cov}\left(\hat{\beta}_{m_1};\hat{\beta}_{m_2}\right) =$$

$$= \mathbf{X_i}\,\widehat{\mathrm{Var}}\left(\hat{\boldsymbol{\beta}}\right)\mathbf{X_i'} = s^2\,\frac{\mathbf{X_i}}{\sqrt{N}}\left(\frac{\mathbf{X'X}}{N}\right)^{-1}\frac{\mathbf{X_i'}}{\sqrt{N}} = s^2\,\frac{\mathbf{X_i}}{\sqrt{N}}\mathbf{Q}_N{}^{-1}\,\frac{\mathbf{X_i'}}{\sqrt{N}} \xrightarrow[N\to\infty]{} 0$$

Предсказание среднего значения зависимой переменной

ОЦЕНКА МАТРИЦЫ КОВАРИАЦИЙ ПАРНОЙ РЕГРЕССИИ

$$\mathrm{COV}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} \Rightarrow \widehat{\mathrm{COV}}\left(\hat{\boldsymbol{\beta}}\right) = s^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \frac{1}{N^2\,\mathrm{VAR}\left(\mathbf{X}_1\right)} \begin{bmatrix} \sum X_{1i}^2 & -\sum X_{1i} \\ -\sum X_{1i} & N \end{bmatrix}$$

$$\widehat{\mathrm{var}}\left(\hat{\beta}_0\right) = \frac{s^2 \sum X_{1i}^2}{N^2\,\mathrm{VAR}\left(\mathbf{X}_1\right)} = \frac{s^2}{N}\left(1 + \frac{\bar{\mathbf{X}}_1^2}{\mathrm{VAR}\left(\mathbf{X}_1\right)}\right)$$

$$\widehat{\mathrm{var}}\left(\hat{\beta}_1\right) = \frac{s^2}{N\,\mathrm{VAR}\left(\mathbf{X}_1\right)};$$

$$\widehat{\mathrm{cov}}\left(\hat{\beta}_0;\hat{\beta}_1\right) = \frac{-s^2 \sum X_{1i}}{N^2\,\mathrm{VAR}\left(\mathbf{X}_1\right)} = -\frac{s^2 \bar{\mathbf{X}}_1}{N\,\mathrm{VAR}\left(\mathbf{X}_1\right)}$$

Предсказание среднего значения зависимой переменной

$$\widehat{\operatorname{var}}\left(\hat{\beta}_0\right) = \frac{s^2 \sum X_{1i}^2}{N^2 \operatorname{VAR}\left(\mathbf{X}_1\right)} = \frac{s^2}{N}\left(1 + \frac{\bar{\mathbf{X}}_1^2}{\operatorname{VAR}\left(\mathbf{X}_1\right)}\right)$$

$$\widehat{\operatorname{var}}\left(\hat{\beta}_1\right) = \frac{s^2}{N \operatorname{VAR}\left(\mathbf{X}_1\right)};$$

$$\widehat{\operatorname{cov}}\left(\hat{\beta}_0; \hat{\beta}_1\right) = -\frac{s^2 \bar{\mathbf{X}}_1}{N \operatorname{VAR}\left(\mathbf{X}_1\right)}$$

$$\widehat{\operatorname{var}}\left(\hat{Y}_i\right) = \widehat{\operatorname{var}}\left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right) =$$

$$= \widehat{\operatorname{var}}\left(\hat{\beta}_0\right) + \widehat{\operatorname{var}}\left(\hat{\beta}_1\right) X_i^2 + 2\widehat{\operatorname{cov}}\left(\hat{\beta}_0; \hat{\beta}_1\right) X_i =$$

Предсказание среднего значения зависимой переменной

$$\widehat{\operatorname{var}}\left(\hat{Y}_i\right) = \widehat{\operatorname{var}}\left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right) = \widehat{\operatorname{var}}\left(\hat{\beta}_0\right) + \widehat{\operatorname{var}}\left(\hat{\beta}_1\right) X_i^2 + 2\widehat{\operatorname{cov}}\left(\hat{\beta}_0; \hat{\beta}_1\right) X_i =$$

$$= \frac{s^2}{N}\left(1 + \frac{\bar{\mathbf{X}}_1^2}{\operatorname{VAR}(\mathbf{X}_1)} + \frac{X_i^2}{\operatorname{VAR}(\mathbf{X}_1)} - \frac{2\bar{\mathbf{X}}_1 X_i}{\operatorname{VAR}(\mathbf{X}_1)}\right) = \frac{s^2}{N}\left(1 + \frac{\left(\bar{\mathbf{X}}_1 - X_i\right)^2}{\operatorname{VAR}(\mathbf{X}_1)}\right)$$

$$\widehat{\operatorname{var}}\left(Y_i\right) = \widehat{\operatorname{var}}\left(\operatorname{E} Y_i + \varepsilon_i\right) = \widehat{\operatorname{var}}\left(\hat{\operatorname{E}} Y_i\right) + \widehat{\operatorname{var}}\left(\varepsilon_i\right) =$$

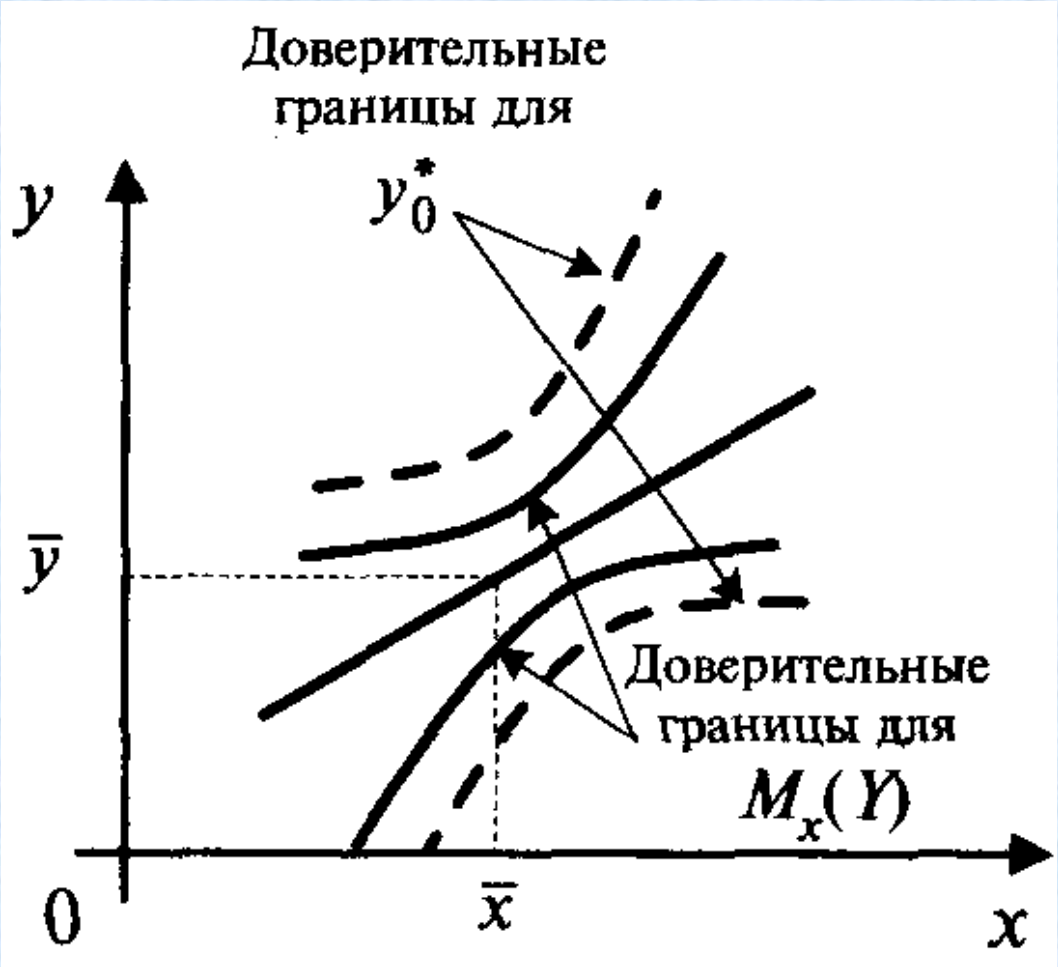
$$= \widehat{\operatorname{var}}\left(\hat{Y}_i\right) + \widehat{\operatorname{var}}\left(\varepsilon_i\right) = \widehat{\operatorname{var}}\left(\hat{Y}_i\right) + s^2 =$$

$$= s^2\left(\frac{1}{N} + \frac{\left(\bar{\mathbf{X}}_1 - X\right)^2}{N \operatorname{VAR}(\mathbf{X}_1)} + 1\right)$$

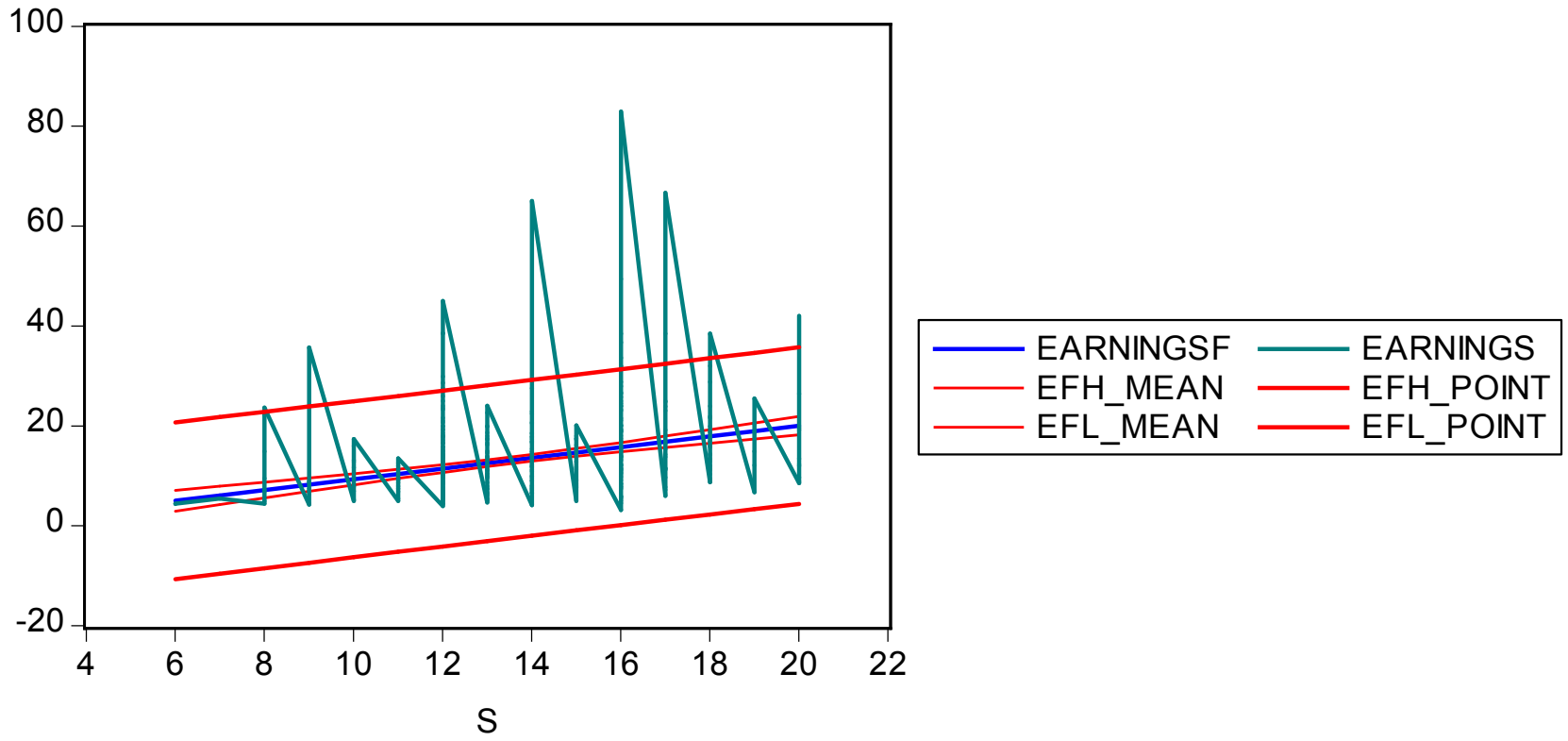
Самое точное прогнозирование в районе центра данных

Точность прогнозного значения выше, чем «наблюдаемого»

Графики доверительных областей для зависимой переменной



ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ПРЕДСКАЗАННЫХ ЗНАЧЕНИЙ



Выводы по доверительным областям для зависимой переменной

1. ***Прогноз значений*** зависимой переменной Y по уравнению регрессии ***оправдан***, если значение x объясняющей переменной X ***не выходит за диапазон ее значений по выборке***. Причем, чем ближе x_p к $\bar{x}$, тем точнее прогноз (уже доверительный интервал).

2. ***Использование линии регрессии вне обследованного диапазона значений объясняющей переменной*** (даже если оно оправдано, исходя из смысла решаемой задачи) ***может привести к значительным погрешностям***.

ПОКАЗАТЕЛИ КАЧЕСТВА УРАВНЕНИЯ РЕГРЕССИИ В ЦЕЛОМ

Суть проверки общего качества уравнения регрессии – оценить насколько хорошо эмпирическое уравнение регрессии согласуется со статистическими данными.

Показатели качества уравнения регрессии в целом

Основные показатели качества:

1. Коэффициент детерминации R^2
2. Скорректированный коэффициент детерминации $\overline{R}^2$
3. Значение *F*-статистики
4. Сумма квадратов остатков (*RSS*)
5. Стандартная ошибка регрессии S_e
6. Прочие показатели: средняя ошибка аппроксимации, индекс множественной корреляции и т.д.

РАЗЛОЖЕНИЕ СУММЫ КВАДРАТОВ

$$\mathrm{Var}(Y) = \mathrm{Var}(\hat{Y} + e) = \mathrm{Var}(\hat{Y}) + \mathrm{Var}(e) + 2\,\mathrm{COVAR}\left(\hat{Y}, e\right)$$

$$\mathrm{Var}(Y) = \mathrm{Var}(\hat{Y}) + \mathrm{Var}(e)$$

$$\sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum e^2$$

$$TSS = ESS + RSS$$

TSS – полная сумма квадратов (total sum of squares)

ESS – оцененная сумма квадратов (explained sum of squares)

RSS – остаточная сумма квадратов (residual sum of squares)

РАЗЛОЖЕНИЕ СУММЫ КВАДРАТОВ, ВЫЧИСЛЕНИЕ R^2 И ВСЕ НИЖЕСЛЕДУЮЩИЕ РЕЗУЛЬТАТЫ ОПРАВДАНЫ ТОЛЬКО ЕСЛИ СРЕДИ РЕГРЕССОРОВ ЕСТЬ КОНСТАНТА

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R^2

$$Y = \alpha + \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

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

$$H_0 : \boldsymbol{\beta} = \mathbf{0} \Rightarrow \forall i \quad \hat{Y}_i \equiv \bar{Y} \Leftrightarrow R^2 = 0$$

$$H_1 : \boldsymbol{\beta} \neq \mathbf{0} \Rightarrow \exists i \quad \hat{Y}_i \neq \bar{Y} \Leftrightarrow R^2 > 0$$

Коэффициент детерминации достигает максимума из возможных значений, когда RSS принимает наименьшее из достижимых значений

КРИТЕРИЙ МАКСИМУМА R^2 ЭКВИВАЛЕНТЕН ПРИНЦИПУ МНК

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R^2

Коэффициент R^2 показывает долю объясненной вариации зависимой переменной:

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

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R^2

Используется для предварительной оценки качества модели (по степени корреляции между объясняемой переменной и ее предсказанным значением)

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R^2

Используется для предварительной оценки качества модели (по степени корреляции между объясняемой переменной и ее предсказанным значением)

$$\operatorname{cov}\left(Y,\hat{Y}\right)=\operatorname{cov}\left(\hat{Y}+e,\hat{Y}\right)=\operatorname{cov}\left(\hat{Y},\hat{Y}\right)+\operatorname{cov}\left(e,\hat{Y}\right)=\operatorname{var}\left(\hat{Y}\right)$$

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R^2

Используется для предварительной оценки качества модели (по степени корреляции между объясняемой переменной и ее предсказанным значением)

$$\operatorname{cov}\left(Y,\hat{Y}\right)=\operatorname{cov}\left(\hat{Y}+e,\hat{Y}\right)=\operatorname{cov}\left(\hat{Y},\hat{Y}\right)+\operatorname{cov}\left(e,\hat{Y}\right)=\operatorname{var}\left(\hat{Y}\right)$$

$$\operatorname{corr}\left(Y,\hat{Y}\right)=\frac{\operatorname{cov}\left(Y,\hat{Y}\right)}{\sqrt{\operatorname{var}\left(Y\right)\operatorname{var}\left(\hat{Y}\right)}}=\frac{\operatorname{var}\left(\hat{Y}\right)}{\sqrt{\operatorname{var}\left(Y\right)\operatorname{var}\left(\hat{Y}\right)}}=\sqrt{\frac{\operatorname{var}\left(\hat{Y}\right)}{\operatorname{var}\left(Y\right)}}$$

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R^2

Используется для предварительной оценки качества модели (по степени корреляции между объясняемой переменной и ее предсказанным значением)

$$\operatorname{cov}\left(Y,\hat{Y}\right)=\operatorname{cov}\left(\hat{Y}+e,\hat{Y}\right)=\operatorname{cov}\left(\hat{Y},\hat{Y}\right)+\operatorname{cov}\left(e,\hat{Y}\right)=\operatorname{var}\left(\hat{Y}\right)$$

$$\operatorname{corr}\left(Y,\hat{Y}\right)=\frac{\operatorname{cov}\left(Y,\hat{Y}\right)}{\sqrt{\operatorname{var}\left(Y\right)\operatorname{var}\left(\hat{Y}\right)}}=\frac{\operatorname{var}\left(\hat{Y}\right)}{\sqrt{\operatorname{var}\left(Y\right)\operatorname{var}\left(\hat{Y}\right)}}=\sqrt{\frac{\operatorname{var}\left(\hat{Y}\right)}{\operatorname{var}\left(Y\right)}}$$

$$R^2=\operatorname{corr}^2\left(Y,\hat{Y}\right)$$

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R^2

В парной регрессии коэффициент детерминации равен квадрату коэффициента между объясняемой и зависимой переменной:

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R^2

В парной регрессии коэффициент детерминации равен квадрату коэффициента между объясняемой и зависимой переменной:

$$R^2 = \frac{\operatorname{var}\left(\hat{Y}\right)}{\operatorname{var}\left(Y\right)} = \frac{\operatorname{var}\left(\hat{\alpha} + X\hat{\beta}\right)}{\operatorname{var}\left(Y\right)}$$

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R^2

В парной регрессии коэффициент детерминации равен квадрату коэффициента между объясняемой и зависимой переменной:

$$R^2 = \frac{\operatorname{var}\left(\hat{Y}\right)}{\operatorname{var}\left(Y\right)} = \frac{\operatorname{var}\left(\hat{\alpha} + X\hat{\beta}\right)}{\operatorname{var}\left(Y\right)}$$

$$R^2 = \hat{\beta}^2 \cdot \frac{\operatorname{var}\left(X\right)}{\operatorname{var}\left(Y\right)} = \frac{\operatorname{cov}^2\left(Y, X\right)}{\operatorname{var}^2\left(X\right)} \cdot \frac{\operatorname{var}\left(X\right)}{\operatorname{var}\left(Y\right)}$$

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R^2

В парной регрессии коэффициент детерминации равен квадрату коэффициента между объясняемой и зависимой переменной:

$$R^2 = \frac{\operatorname{var}\left(\hat{Y}\right)}{\operatorname{var}\left(Y\right)} = \frac{\operatorname{var}\left(\hat{\alpha} + X\hat{\beta}\right)}{\operatorname{var}\left(Y\right)}$$

$$R^2 = \hat{\beta}^2 \cdot \frac{\operatorname{var}\left(X\right)}{\operatorname{var}\left(Y\right)} = \frac{\operatorname{cov}^2\left(Y,X\right)}{\operatorname{var}^2\left(X\right)} \cdot \frac{\operatorname{var}\left(X\right)}{\operatorname{var}\left(Y\right)}$$

$$R^2 = \frac{\operatorname{cov}^2\left(Y,X\right)}{\operatorname{var}\left(X\right)\operatorname{var}\left(Y\right)}$$

$$R^2 = \operatorname{corr}^2\left(Y,X\right)$$

ГЕОМЕТРИЧЕСКИЙ СМЫСЛ R^2

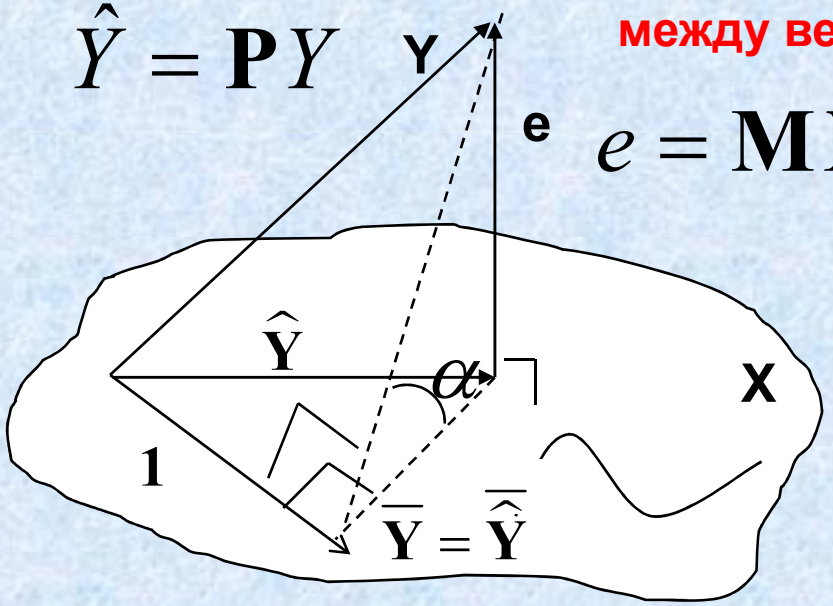
Если среди регрессоров есть константа, то

А) коэффициент детерминации определен

Б) геометрически представляет косинус угла между векторами $\left(\mathbf{Y}-\overline{\mathbf{Y}};\hat{\mathbf{Y}}-\overline{\mathbf{Y}}\right)$

$\hat{Y}=\mathbf{P}Y$

$e=\mathbf{M}Y$



$$\angle\alpha=\angle\left(\mathbf{Y}-\overline{\mathbf{Y}};\hat{\mathbf{Y}}-\overline{\mathbf{Y}}\right)$$

$$R^2=\cos\alpha$$

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R^2

Величину 1-R^2 можно трактовать как вероятность необъяснения разброса значений зависимой переменной выбранными регрессорами.

Тогда:

$$1 - R^2 = \frac{\mathrm{var}(e)}{\mathrm{var}(Y)} = \frac{\mathrm{var}\left(e_{X_1,\ldots,X_k}\right)}{\mathrm{var}(Y)}$$

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R^2

Величину $1-R^2$ можно трактовать как вероятность необъяснения разброса значений зависимой переменной выбранными регрессорами.

Тогда:

$$1-R^2=\frac{\operatorname{var}(e)}{\operatorname{var}(Y)}=\frac{\operatorname{var}\left(e_{X_1,\ldots,X_k}\right)}{\operatorname{var}(Y)}$$

$$1-R^2=\frac{\operatorname{var}\left(e_{X_1}\right)}{\operatorname{var}(Y)}\cdot\frac{\operatorname{var}\left(e_{X_1,X_2}\right)}{\operatorname{var}\left(e_{X_1}\right)}\cdot\frac{\operatorname{var}\left(e_{X_1,X_2,X_3}\right)}{\operatorname{var}\left(e_{X_1,X_2}\right)}\cdot\ldots\cdot\frac{\operatorname{var}\left(e_{X_1,\ldots,X_k}\right)}{\operatorname{var}\left(e_{X_1,\ldots X_{k-1}}\right)}$$

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R^2

Величину 1-R^2 можно трактовать как вероятность необъяснения разброса значений зависимой переменной выбранными регрессорами.

Тогда:

$$1-R^2=\frac{\operatorname{var}(e)}{\operatorname{var}(Y)}=\frac{\operatorname{var}\left(e_{X_1,\ldots,X_k}\right)}{\operatorname{var}(Y)}$$

$$1-R^2=\frac{\operatorname{var}\left(e_{X_1}\right)}{\operatorname{var}(Y)}\cdot\frac{\operatorname{var}\left(e_{X_1,X_2}\right)}{\operatorname{var}\left(e_{X_1}\right)}\cdot\frac{\operatorname{var}\left(e_{X_1,X_2,X_3}\right)}{\operatorname{var}\left(e_{X_1,X_2}\right)}\cdot\ldots\cdot\frac{\operatorname{var}\left(e_{X_1,\ldots,X_k}\right)}{\operatorname{var}\left(e_{X_1,\ldots X_{k-1}}\right)}$$

$$1-R^2=\left[1-\mathrm{r}^2\left(Y,X_1\right)\right]\left[1-\mathrm{r}^2\left(Y,X_2\mid X_1\right)\right]\ldots\left[1-\mathrm{r}^2\left(Y,X_k\mid X_1\ldots X_{k-1}\right)\right]$$

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R^2

Величину 1-R^2 можно трактовать как вероятность необъяснения разброса значений зависимой переменной выбранными регрессорами.

Тогда:

$$1-R^2=\left[1-\mathrm{r}^2\left(Y,X_1\right)\right]\left[1-\mathrm{r}^2\left(Y,X_2\mid X_1\right)\right]\ldots\left[1-\mathrm{r}^2\left(Y,X_k\mid X_1\ldots X_{k-1}\right)\right]$$

$$R_{yx_1x_2\ldots x_m}\geq\max_i r_{yx_i}$$

При правильном включении факторов в модель индекс множественной корреляции будет существенно превосходить наибольшее из значений коэффициента парной корреляции

ОСНОВНЫЕ СВОЙСТВА КОЭФФИЦИЕНТА ДЕТЕРМИНАЦИИ

1. $0 \le R^2 \le 1$.
2. Чем ближе R^2 к 1, тем лучше регрессия аппроксимирует статистические данные, тем теснее линейная связь между зависимой и объясняющими переменными.
3. Если $R^2 = 1$, то статистические данные лежат на линии регрессии, т.е. между зависимой и объясняющими переменными имеется функциональная зависимость. Если $R^2 = 0$, то вариация зависимой переменной полностью обусловлена воздействием неучтенных в модели переменных.
4. Низкое значение R^2 не свидетельствует о плохом качестве модели, и может объясняться наличием существенных факторов, не включенных в модель

НЕДОСТАТКИ КОЭФФИЦИЕНТА ДЕТЕРМИНАЦИИ R^2

1. R^2 всегда увеличивается с включением новой переменной
2. Не позволяет дать окончательного заключения о качестве модели без учета других факторов
3. Подвержен влиянию посторонних факторов и может привести к ошибочным выводам
4. Отсутствуют таблицы распределения R^2
5. R^2 является смещенной оценкой истинного коэффициента детерминации
6. Коэффициенты R^2 в разных моделях (разной спецификации) с разным числом наблюдений (и переменных) несравнимы

СКОРРЕКТИРОВАННЫЙ КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

Корректировка на несмещенность

$$\mathrm{E}\left(1-R^2\right)=\frac{N-k}{N-1}\cdot\frac{\mathrm{var}\left(\varepsilon\right)}{\mathrm{var}\left(Y\right)}=\frac{N-k}{N-1}\cdot\left(1-\overline{R}^2\right)$$

$$1-R^2_{adj}=\left(1-R^2\right)\frac{N-1}{N-k}=\frac{RSS}{TSS}\cdot\frac{N-1}{N-k}$$

$$\mathrm{E}\left(1-R^2_{adj}\right)=1-\overline{R}^2$$

СКОРРЕКТИРОВАННЫЙ КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

$$R^2_{adj} = 1 - \frac{RSS/(N-k)}{TSS/(N-1)} = 1 - \left(1 - \frac{ESS}{TSS}\right)\frac{N-1}{N-k} = 1 - \left(1 - R^2\right)\frac{N-1}{N-k}$$

показывает долю объясненной вариации зависимой переменной с учетом числа объясняющих переменных уравнения регрессии

Скорректированные коэффициенты детерминации в разных моделях с разным числом наблюдений (и переменных) ограниченно сравнимы

F –ТЕСТ КАЧЕСТВА ПОДГОНКИ РЕГРЕССИИ

$$Y = \alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$H_0 : \boldsymbol{\beta} = \mathbf{0} \Rightarrow R^2 = 0$$

$$H_1 : \boldsymbol{\beta} \neq \mathbf{0}$$

Хотя для коэффициента детерминации не существует таблиц распределения, можно попытаться составить статистику, основанную на его значении для тестирования незначимости модели, т.е. незначимости <u>всех</u> (кроме может быть константы) коэффициентов.

F-статистика
для проверки качества уравнения регрессии

F-статистика представляет собой отношение объясненной суммы квадратов (в расчете на одну независимую переменную) к остаточной сумме квадратов (в расчете на одну степень свободы)

$$F(k-1, n-k) = \frac{ESS/(k-1)}{RSS/(N-k)}$$

Поскольку и *ESS*, и *RSS*, по сути – дисперсии, умноженные на соответствующие числа степеней свободы, то их распределение будет связано со статистикой Пирсона, а потому эта дробь окажется распределенной по закону Фишера.

F-статистика
для проверки качества уравнения регрессии

$$F(k-1, N-k) = \frac{ESS/(k-1)}{RSS/(N-k)} =$$

$$= \frac{\left.\frac{ESS}{\sigma^2}\right/(k-1)}{\left.\frac{RSS}{\sigma^2}\right/(N-k)} \sim \frac{\chi^2_{k-1}/(k-1)}{\chi^2_{n-k}/(N-k)} = F_{(k-1,N-k)}$$

***F*-статистики в разных моделях с разным числом наблюдений и (или) переменных несравнимы**

F-статистика
для проверки качества уравнения регрессии

$$F(k-1, N-k) \sim \frac{\chi^2_{k-1}/(k-1)}{\chi^2_{n-k}/(N-k)} = F_{(k-1,N-k)}$$

В случае парной регресии:

$$Y = \beta_1 + \beta_2 X + u$$

$$H_0 : \beta_2 = 0, \quad H_1 : \beta_2 \neq 0$$

$$F_{(1,n-1)} = \frac{\chi^2_1}{\chi^2_{n-1}/(N-2)} = \left(\frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(N-2)}} \right)^2 = t^2_{N-2}$$

Таким образом, в случае парной регрессии,

F-тесту соответствует двусторонний t-тест для коэффициента наклона

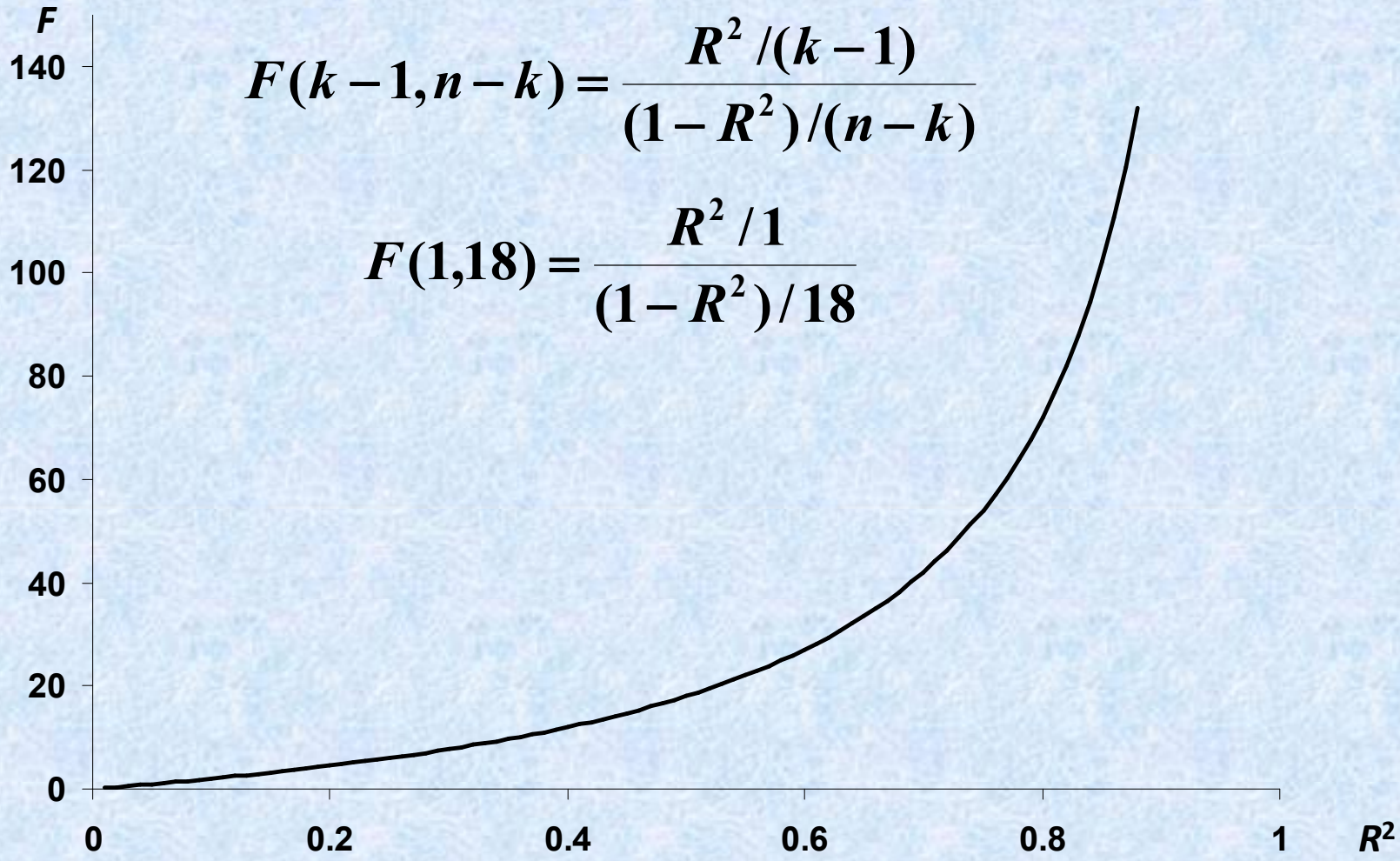
F-статистика
для проверки качества уравнения регрессии

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

F –ТЕСТ КАЧЕСТВА ПОДГОНКИ РЕГРЕССИИ

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

$$F(1,18)=\frac{R^2/1}{(1-R^2)/18}$$



F **статистика монотонно возрастает при росте коэффициента детерминации.**

Значит если нулевая гипотеза верна, то ее значение не будет велико

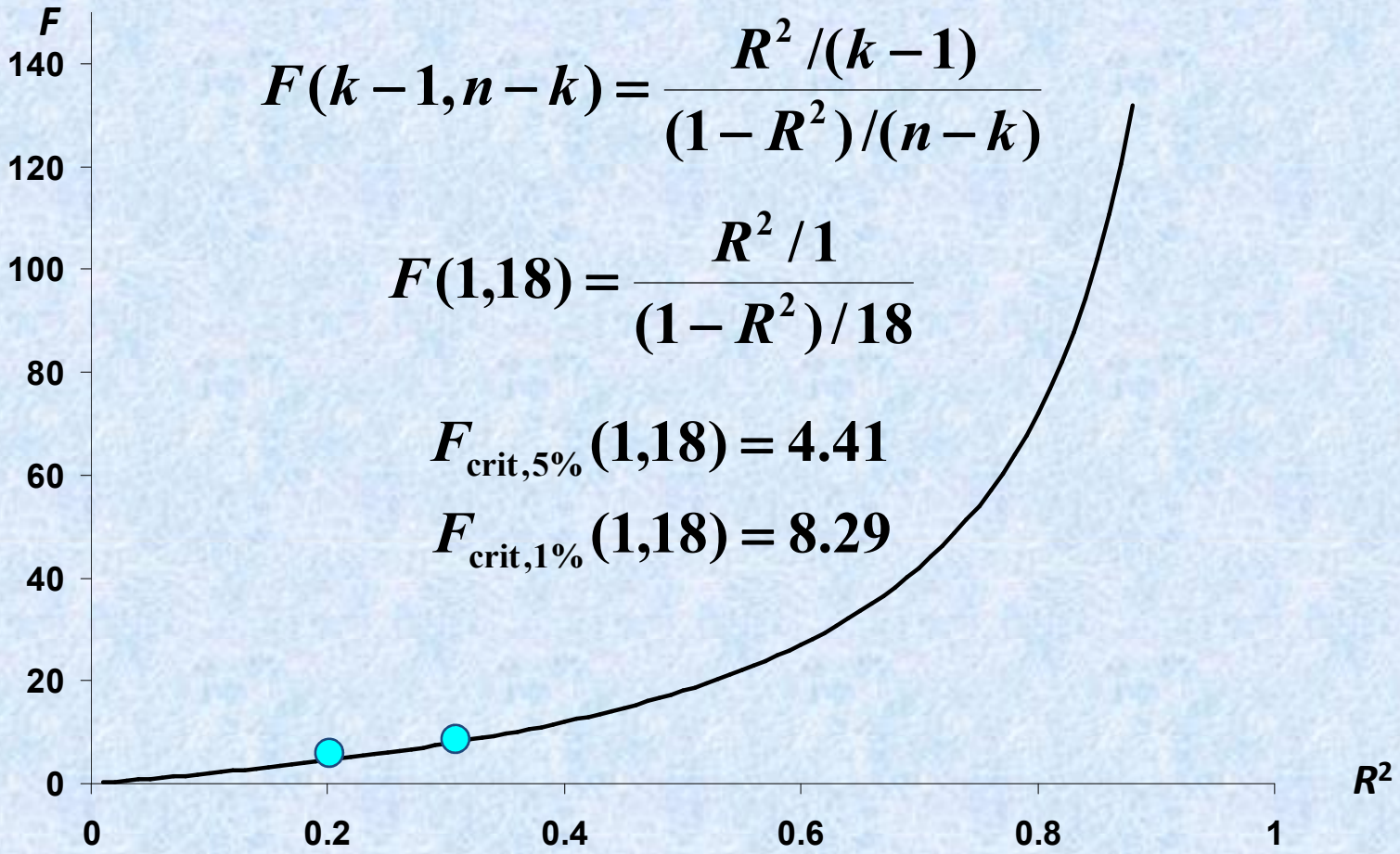
F –ТЕСТ КАЧЕСТВА ПОДГОНКИ РЕГРЕССИИ

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

$$F(1,18)=\frac{R^2/1}{(1-R^2)/18}$$

$$F_{crit,5\%}(1,18)=4.41$$

$$F_{crit,1\%}(1,18)=8.29$$



При фиксированном количестве степеней свободы, в частности когда модель определена хотя бы с точностью до количества регрессоров и наблюдений, между критическими значениями *F*- статистики и R^2 есть взаимнооднозначное соответствие.

В данном случае 5% критической точке соответствует значение R^2 = 0.20, а 1% - R^2 = 0.32

F-статистика
для проверки качества уравнения парной регрессии

$$R^2 = \frac{\mathrm{Var}(\hat{Y})}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}(\hat{\beta}_1 + \hat{\beta}_2 X)}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}(\hat{\beta}_2 X)}{\mathrm{Var}(Y)} = \frac{\hat{\beta}_2^2 \mathrm{Var}(X)}{\mathrm{Var}(Y)}$$

$$R^2 = 1 - \frac{\mathrm{Var}(e)}{\mathrm{Var}(Y)} \quad \Rightarrow \quad 1 - R^2 = \frac{\mathrm{Var}(e)}{\mathrm{Var}(Y)}$$

$$s^2 = \frac{N}{N-2}\mathrm{Var}(e)$$

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

$$= \frac{\dfrac{\hat{\beta}_2^2 \mathrm{Var}(X)}{\mathrm{Var}(Y)}}{\dfrac{\mathrm{Var}(e)}{\mathrm{Var}(Y)} \Big/ (N-2)} = \frac{\hat{\beta}_2^2 \mathrm{Var}(X)}{\dfrac{1}{N} \cdot \dfrac{N}{N-2}\mathrm{Var}(e)} = \frac{\hat{\beta}_2^2 \mathrm{Var}(X)}{\dfrac{s^2}{N}} = \frac{\hat{\beta}_2^2}{\dfrac{s^2}{\mathrm{NVar}(x)}} = t^2$$

F –ТЕСТ КАЧЕСТВА ПОДГОНКИ РЕГРЕССИИ

$$F_{\text{crit},5\%}(1,18) = 4.41$$

$$F_{\text{crit},1\%}(1,18) = 8.29$$

$$t_{\text{crit},5\%}(18) = 2.10$$

$$t_{\text{crit},1\%}(18) = 2.88$$

$$4.41 = 2.10^2$$

$$8.29 = 2.88^2$$

Пример соответствия между критическими значениями *F*-статистики и *t*- статистики

ОСНОВНУЮ ЖЕ РОЛЬ *F*-СТАТИСТИКА ИГРАЕТ В ТЕСТАХ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Порядок работы при проверке значимости парного уравнения по *F*-статистике

1. Выбираем уровень значимости α (1% ,5%, 10% и т.п.).
2. Вычисляем число степеней свободы k и (*N*–k).
3. По таблицам *F*-распределения определяем критическое значение $F_{\alpha;\, k;\, N-k}$ (всегда одностороннее).
4. Если *F*-статистика больше $F_{\alpha;\, k;\, N-k}$, то уравнение в целом является значимым на уровне значимости α.
5. В противном случае уравнение в целом не значимо (на данном уровне α).

Связь между значимостью коэффициента регрессии и уравнения парной регрессии в целом

В парной регрессии *F*-статистика равна квадрату *t*-статистики; то же верно и для их критических уровней (односторонний для *t*-статистики)

$$t^2 = F$$

$$\left(t_{\alpha;n-2}\right)^2 = F_{\alpha;1;n-2}$$

В парной регрессии значимость коэффициента регрессии и значимость уравнения в целом эквивалентны

F-статистики в разных моделях с разным числом наблюдений и (или) переменных несравнимы

Взаимосвязь критериев в парном регрессионном анализе

Коэффициент корреляции по абсолютной величине совпадает с квадратным корнем из коэффициента детерминации

$$\left|r_{xy}\right| = \sqrt{R^2}$$

t-статистики для коэффициента корреляции и коэффициента регрессии b_1 совпадают

Проверка значимости коэффициента регрессии эквивалентна проверке наличия линейной связи

Коэффициент корреляции r_{xy}

Коэффициент корреляции указывает на наличие (или отсутствие) линейной связи между зависимой и объясняющей переменными

Для проверки гипотезы об отсутствии линейной связи используется тот факт, что величина

$$t = \sqrt{F} = \sqrt{\frac{r^2}{(\ )/(N-2)}} = \frac{r}{\sqrt{1-r^2}}\sqrt{N-2}$$

имеет распределение Стьюдента с (*n*–2) степенями свободы

F –ТЕСТ КАЧЕСТВА ПОДГОНКИ РЕГРЕССИИ

```
. reg EARNINGS S

      Source |       SS       df       MS              Number of obs =     570
-------------+------------------------------           F(  1,   568) =   65.64
       Model |  3977.38016     1  3977.38016            Prob > F      =  0.0000
    Residual |  34419.6569   568  60.5979875            R-squared     =  0.1036
-------------+------------------------------           Adj R-squared =  0.1020
       Total |  38397.0371   569  67.4816117           Root MSE      =  7.7845

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |    1.073055   .1324501     8.102   0.000      .8129028    1.333206
       _cons |   -1.391004   1.820305    -0.764   0.445     -4.966354    2.184347
------------------------------------------------------------------------------
```

$$F = \frac{ESS/(k-1)}{RSS/(N-k)} = \frac{EMS}{RMS} = \frac{3977.38}{60.60} = 65.64$$

F –ТЕСТ КАЧЕСТВА ПОДГОНКИ РЕГРЕССИИ

```
. reg EARNINGS S

  Source |       SS       df       MS              Number of obs =     570
---------+------------------------------           F(  1,   568) =   65.64
   Model |  3977.38016     1  3977.38016           Prob > F      =  0.0000
Residual |  34419.6569   568  60.5979875           R-squared     =  0.1036
---------+------------------------------           Adj R-squared =  0.1020
   Total |  38397.0371   569  67.4816117           Root MSE      =  7.7845

------------------------------------------------------------------------------
EARNINGS |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       S |    1.073055   .1324501     8.102   0.000        .8129028    1.333206
   _cons |   -1.391004   1.820305    -0.764   0.445       -4.966354    2.184347
------------------------------------------------------------------------------
```

$$F(1, N-2) = \frac{R^2}{(1-R^2)/(N-2)} = \frac{0.1036}{(1-0.1036)/(570-2)} = 65.65$$

F –ТЕСТ КАЧЕСТВА ПОДГОНКИ РЕГРЕССИИ

```
. reg EARNINGS S

  Source |       SS       df       MS              Number of obs =     570
---------+------------------------------           F(  1,   568) =   65.64
   Model |  3977.38016     1  3977.38016            Prob > F      =  0.0000
Residual |  34419.6569   568  60.5979875            R-squared     =  0.1036
---------+------------------------------           Adj R-squared =  0.1020
   Total |  38397.0371   569  67.4816117            Root MSE      =  7.7845

------------------------------------------------------------------------------
EARNINGS |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       S |   1.073055   .1324501      8.102   0.000        .8129028    1.333206
   _cons |  -1.391004   1.820305     -0.764   0.445       -4.966354    2.184347
------------------------------------------------------------------------------
```

$$65.64 = 8.102^2$$

Сумма квадратов остатков *RSS*

Является оценкой необъясненной части вариации зависимой переменной

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

Используется как основная минимизируемая величина в МНК, а также для расчета других показателей

Значения *RSS* в разных моделях с разным числом наблюдений и (или) переменных несравнимы

Стандартная ошибка регрессии RM*SE*

Является оценкой величины квадрата ошибки, приходящейся на одну степень свободы модели

$$RMSE = \sqrt{\frac{RSS}{N-k}}$$

Используется как основная величина для измерения качества модели (чем она меньше, тем лучше)

Значения S_e в однотипных моделях с разным числом наблюдений и (или) переменных сравнимы

Средняя ошибка аппроксимации *A*

Оценку качества модели дает также средняя ошибка аппроксимации – среднее отклонение расчетных значений $\hat{y}_i$ зависимой переменной от фактических значений y_i

$$A = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%.$$

Допустимый предел значений *A* – не более 10%. Чем меньше значение A, тем лучше

Значения *A* в моделях с разным числом наблюдений и одинаковым количеством переменных сравнимы

Типичные ошибки в использовании показателей качества регрессии

- Величина коэффициентов регрессии не указывает на силу связи или силу влияния на зависимую переменную
- Значимость коэффициентов по *t*-тестам не позволяет сделать вывод о справедливости тех или иных теорий
- *t*-статистики не указывают на относительную важность коэффициентов регрессии
- *t*-статистики предназначены для использования исключительно для выборки и бесполезны для анализа всей совокупности
- Нельзя сравнивать *t*-статистики, *F*-статистики, коэффициенты детерминации и др. у разных уравнений

Ограниченность простой регрессии

1. Никакая единственная переменная за редкими исключениями не в состоянии хорошо «объяснить» изменения зависимой переменной.
2. Могут существовать несколько одинаково хороших и взаимно противоречивых регрессий.
3. Наконец, линейная форма примитивна.

И тем не менее: Нет ничего лучше по простоте и ясности объяснения парной линейной связи. При равной объясняющей способности из двух моделей мы всегда выбираем более простую.

Конец лекции