NATIONAL RESEARCH UNIVERSITY

Лекция 2
2 модуль
Выявление нетипичных наблюдений (выбросов)
Три формы системы нормальных уравнений

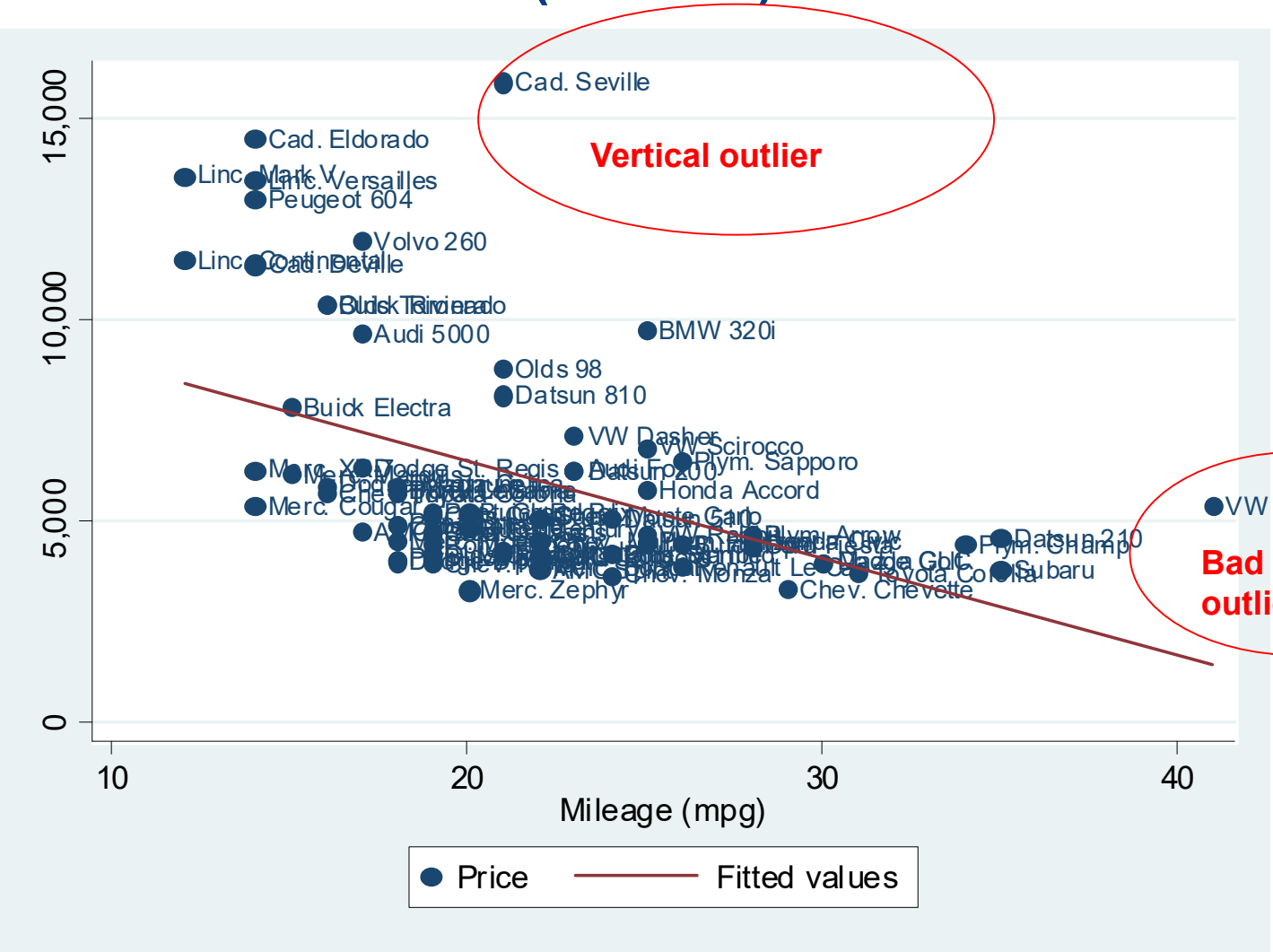
02.11.20

Демидова О.А.
E-mail:demidova@hse.ru

План лекции

1) **Выявление нетипичных наблюдений**
2) **Оценка модели при наличии нетипичных наблюдений**
3) **Влияние изменения масштаба измерения переменных на оценки коэффициентов регрессии и их дисперсий.**
4) **Регрессия в центрированных и нормированных переменных**

Пример нетипичных наблюдений – выбросов (outliers)



Leverage V Residuals

$$Y = X\beta + \varepsilon, e = \hat{Y} - Y$$

$$H = X(X'X)^{-1}X' - hat\ \ matrix$$

$$\hat{Y} = HY, \hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j,$$

$$h_{ii} - leverage \quad , \quad i = 1,..., \ n$$

$$\text{var}[e] = [I - H]\,\text{var}[\varepsilon]$$

$$\text{var}(e_i) = (1 - h_{ii})\sigma_\varepsilon^2$$

$$If\ h_{ii}\ is\ high, then\ \text{var}(e_i)\ is\ small.$$

Vertical Outliers

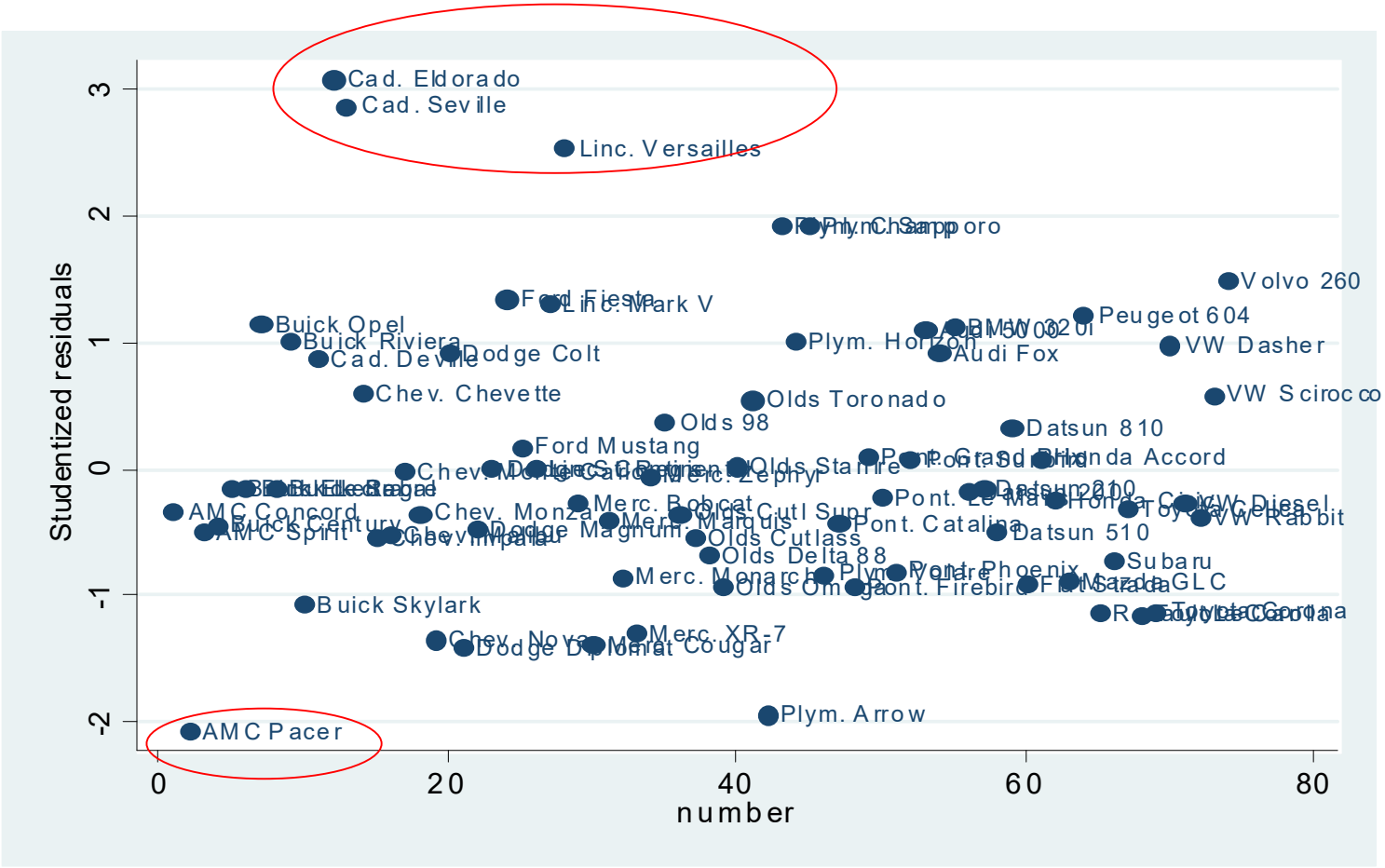
Studentized residues

$$e_i^* = \frac{e_i}{s_{(-i)}\sqrt{1 - h_{ii}}}$$

Rule of thumb: if |e*| > 2 this is outlier

Vertical Outliers

Studentized residues



Results of estimation with outliers and without outliers

```
. reg  price mpg weight  length foreign
```

Source	SS	df	MS
Model	348708940	4	87177235
Residual	286356456	69	4150093.57
Total	635065396	73	8699525.97

Number of obs	74
F(4, 69)	21.01
Prob > F	0.0000
R-squared	0.5491
Adj R-squared	0.5230
Root MSE	2037.2

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-13.40719	72.10761	-0.19	0.853	-157.2579	130.4436
weight	5.716181	1.016095	5.63	0.000	3.689127	7.743235
length	-92.48018	33.5912	-2.75	0.008	-159.4928	-25.46758
foreign	3550.194	655.4564	5.42	0.000	2242.594	4857.793
_cons	5515.58	5241.941	1.05	0.296	-4941.807	15972.97

```
. reg  price mpg weight  length foreign if abs(residst) < 2
```

Source	SS	df	MS
Model	229069621	4	57267405.2
Residual	178137969	65	2740584.14
Total	407207590	69	5901559.27

Number of obs	70
F(4, 65)	20.90
Prob > F	0.0000
R-squared	0.5625
Adj R-squared	0.5356
Root MSE	1655.5

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-33.53987	61.63118	-0.54	0.588	-156.6258	89.54609
weight	3.976065	.9860319	4.03	0.000	2.006823	5.945306
length	-49.88628	32.82274	-1.52	0.133	-115.4378	15.66525
foreign	3403.566	537.8999	6.33	0.000	2329.306	4477.826
_cons	3011.892	4811.133	0.63	0.533	-6596.604	12620.39

Bad leverage points

DF BETA

$$D_{ij} = b_j - b_{j(-i)},$$

$$i = 1, \ldots, n,$$

$$j = 1, \ldots, k$$

Rule of thumb: If

$$\left|D_{ij}\right| \geq 2 / \sqrt{n}\ (in\ practice\ often > 1),$$

$$this\ is\ outlier.$$

But we have many factors!

Bad leverage points

	Rule of thumb
$DFITS_i = e_i^* \sqrt{\frac{h_{ii}}{1-h_{ii}}},$	$DFITS_i > 2\sqrt{\frac{k}{n}},$
$D_i = \frac{1}{k}\frac{s_{(i)}^2}{s^2} DFITS_i^2$ $Cook's\ Distance$	$D_i > \frac{4}{n}$
$W_i = DFITS_i \sqrt{\frac{n-1}{1-h_{ii}}}$ $Welsch's\ Distance$	$W_i > 3\sqrt{k}$

Bad leverage points

```
. predict cooksd, cooksd

. list  make if cooksd > 4/74
```

	make
2.	AMC Pacer
12.	Cad. Eldorado
13.	Cad. Seville
28.	Linc. Versailles
42.	Plym. Arrow
43.	Plym. Champ

Results with and without bad leverage points

```
. reg  price mpg weight  length foreign
```

Source	SS	df	MS
Model	348708940	4	87177235
Residual	286356456	69	4150093.57
Total	635065396	73	8699525.97

Number of obs	74
F(4, 69)	21.01
Prob > F	0.0000
R-squared	0.5491
Adj R-squared	0.5230
Root MSE	2037.2

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-13.40719	72.10761	-0.19	0.853	-157.2579	130.4436
weight	5.716181	1.016095	5.63	0.000	3.689127	7.743235
length	-92.48018	33.5912	-2.75	0.008	-159.4928	-25.46758
foreign	3550.194	655.4564	5.42	0.000	2242.594	4857.793
_cons	5515.58	5241.941	1.05	0.296	-4941.807	15972.97

```
. reg  price mpg weight  length foreign if  cooksd < 4/74
```

Source	SS	df	MS
Model	242230722	4	60557680.6
Residual	161542206	63	2564161.99
Total	403772928	67	6026461.61

Number of obs	68
F(4, 63)	23.62
Prob > F	0.0000
R-squared	0.5999
Adj R-squared	0.5745
Root MSE	1601.3

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-35.5219	62.06735	-0.57	0.569	-159.5536	88.50981
weight	4.953208	1.118628	4.43	0.000	2.717809	7.188606
length	-78.29578	36.2578	-2.16	0.035	-150.7512	-5.840351
foreign	3583.937	532.0723	6.74	0.000	2520.675	4647.199
_cons	5405.938	4834.563	1.12	0.268	-4255.166	15067.04

Оценка модели при наличии нетипичных наблюдений

- Оценка по выборке без выбросов (не очень хорошо)
- Дамми переменные для наблюдений-выбросов
- Робастные оценки параметров
- Критерий $\sum_{i=1}^{n} g(e_i)$, где g(x) медленнее растет при больших x, чем x^2, например, |x| (медианная регрессия).

variable	regols	regmed	mreg	rreg	mmreg70
mpg	-13.407192	6.2978704	-23.500289	-23.873543	-122.79988***
weight	5.7161809***	3.6727916***	4.5636207***	3.2814101***	-3.0145735
length	-92.480183**	-35.375678	-61.1696	-27.096799	93.802086*
foreign	3550.1937***	3222.1316***	3529.0743***	3346.3483***	692.4751*
_cons	5515.5801	-92.609125	3116.0435	480.00227	-1456.2602

legend: * p<0.05; ** p<0.01; *** p<0.001

3 формы системы нормальных уравнений

$$1)\ Y = X\beta + \varepsilon,$$

$$X'X\hat{\beta} = X'Y - 1^{\text{ая}}\ форма\ .$$

$$2)\ Уравнения\quad в\ отклонения\quad x:$$

$$y = Y - \overline{Y}\,\overline{i},$$

$$x_j = X_j - \overline{X}_j\,\overline{i},\ j = 1,\ldots,k$$

$$y = x\alpha + \varepsilon,$$

$$Важный\quad результат\quad .\hat{\alpha} = \hat{\beta}_{-0}$$

3 формы системы нормальных уравнений

Доказательство. $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$

$$\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}_1 + \ldots + \hat{\beta}_k \overline{X}_k$$

$$\hat{y} = \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k \Rightarrow \hat{\alpha} = \hat{\beta}_{-0}.$$

$$x'x\hat{\beta}_{-0} = x'y \Leftrightarrow$$

$$\hat{Var}(X)\hat{\beta}_{-0} = \hat{Cov}(X,Y) - 2^{ая}\ форма.$$

3 формы системы нормальных уравнений

$$3)\ \tilde{x}_j = \frac{x_j}{\hat{\sigma}_j},\ j = 1,\dots,k$$

$$\hat{\sigma}_j = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ji} - \overline{X}_j)^2,\ \tilde{y} = \frac{y}{\hat{\sigma}_y},$$

$$\tilde{x}\tilde{x}\hat{\tilde{\beta}} = \tilde{x}\tilde{y} \Leftrightarrow$$

$$Cor[X]\tilde{\tilde{\beta}} = Cor[X,Y] - 3^{\text{ья}}\ форма\ .$$

Thank you for your attention!

20, Myasnitskaya str., Moscow, Russia, 101000
Tel.: +7 (495) 628-8829, Fax: +7 (495) 628-7931
www.hse.ru