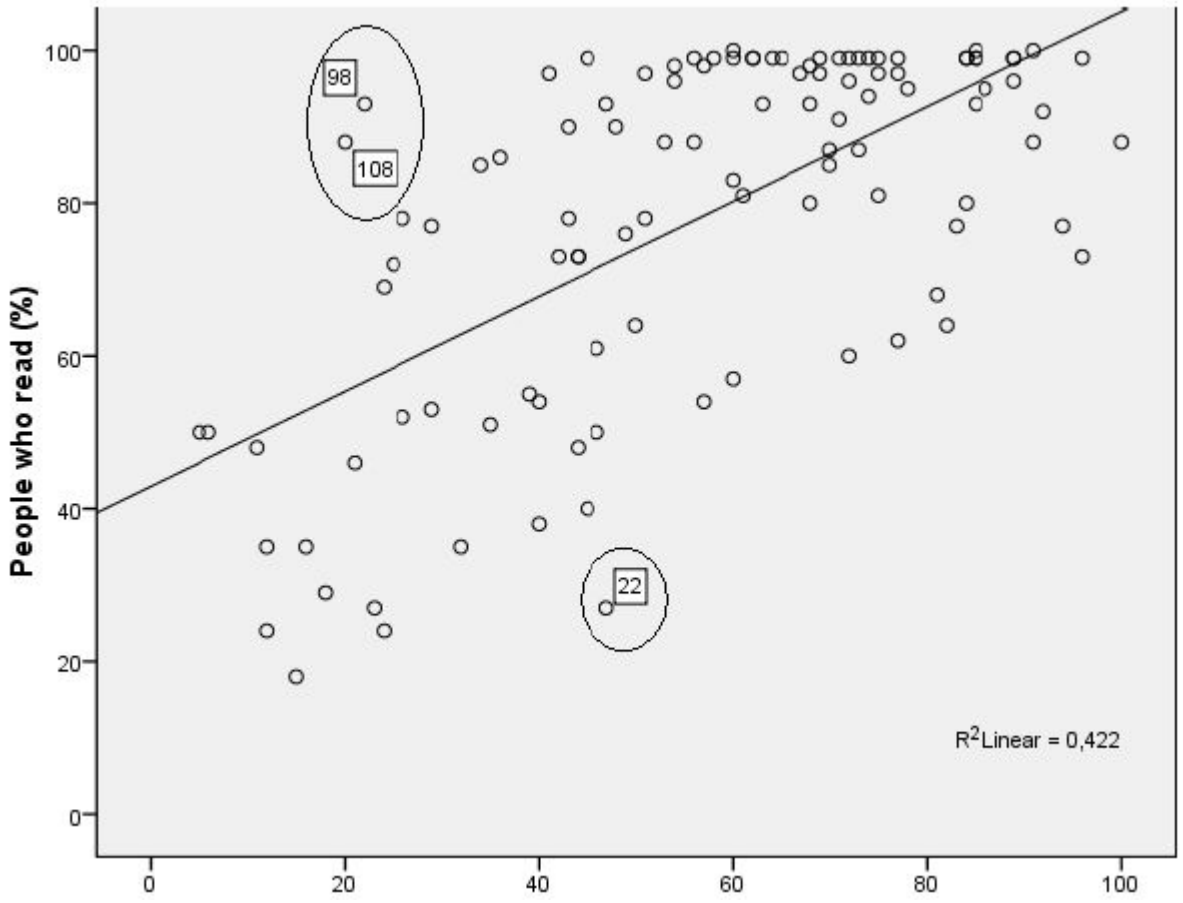
РЕГРЕССИОННЫЙ АНАЛИЗ

Если расчёт коэффициентов корреляции характеризует силу связи между двумя переменными, то регрессионный анализ служит для определения ***вида*** этой ***связи*** и дает возможность для **прогнозирования** значения одной (зависимой) переменной отталкиваясь от значения другой (независимой) переменной. Регрессионные модели: линейная и множественная.

Линейная регрессионная модель

- Прежде чем приступать к построению регрессионной модели обратимся к диаграмме рассеивания, она поможет нам визуально оценить наличие линейной связи и выявить выбросы, которые могут существенно повлиять на результаты построенной модели.
- Для того чтобы построить диаграмму рассеивания выполним команду **Graphs→Legacy Dialogs→Scatter/Dot** и выбираем построение простой диаграммы (**Simple Scatter**). В появившемся окне расставляем наши переменные по осям координат: x - независимая переменная, y - зависимая переменная. В Output появляется наш график, на который нам необходимо нанести прямую. Для этого дважды щелкаем левой кнопкой мыши по графику и в появившемся окне, на графике, щелкаем правой кнопкой мыши и выбираем вторую строчку снизу Add Fit Line at Total.

- На графике мы можем наблюдать линейную зависимость между переменными, а, следовательно, приступить к построению линейной регрессии. Однако мы можем наблюдать три выброса, которые, возможно, могут повлиять на конечный результат (для того, чтобы узнать номера выбросов, в том же окне где мы рисовали прямую, кнопкой в виде мишени нажимаем на интересующие нас случаи). После построения регрессионной модели мы можем попробовать убрать наши выбросы из базы посредством фильтрации и еще раз построить регрессионную модель.


People who read (%)
100
80
60
40
20
0
0
20
40
60
80
100
98
108
22
R2Linear = 0,422

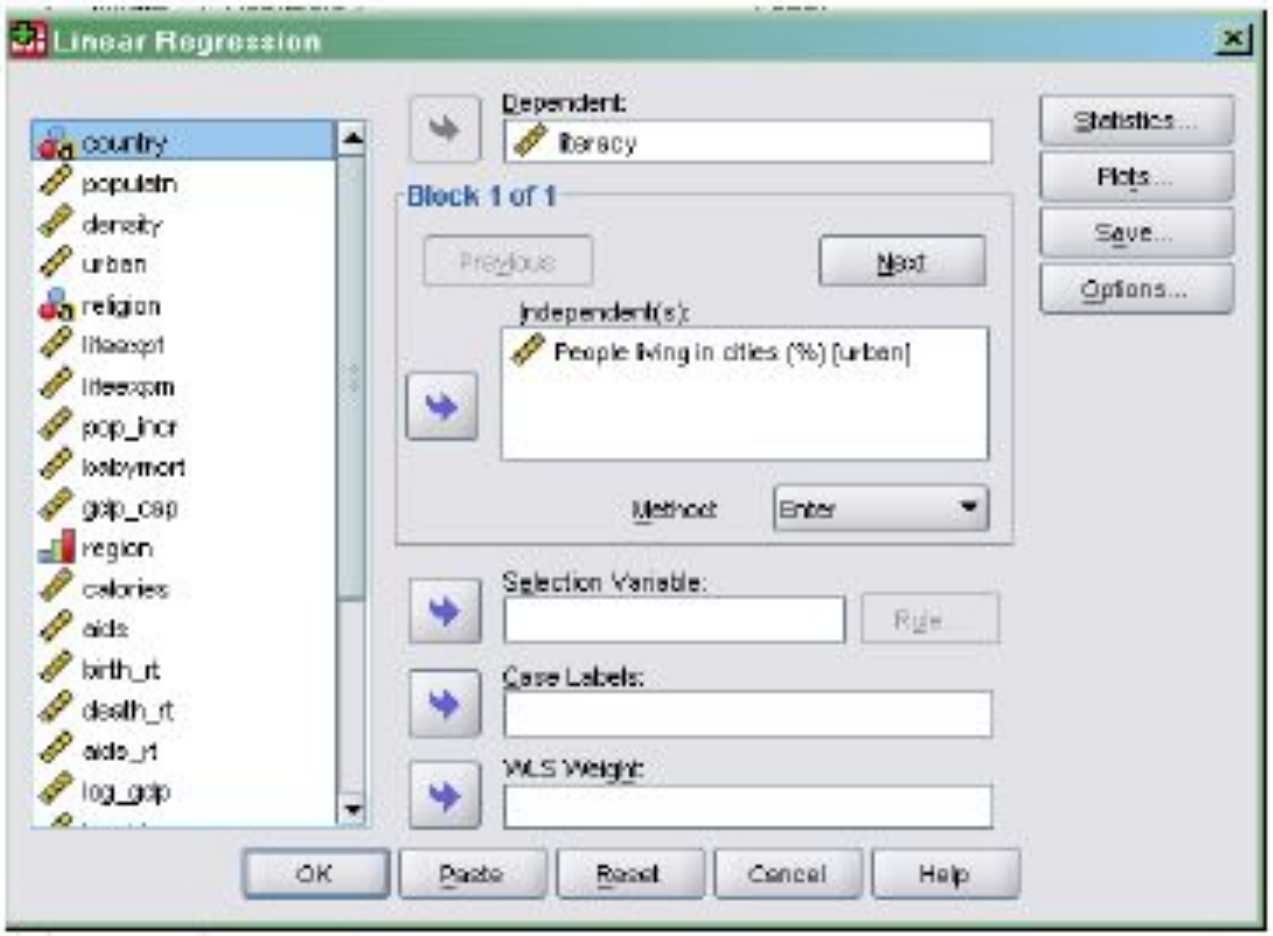
- И последнее что необходимо проверить перед построением регрессионной модели это нормальность распределения наших переменных (это необходимо только для интервальных шкал). Чтобы это сделать нам необходимо провести тест Колмогорова-Смирнова или Шапиро-Уилка для выборки меньше 50. Выполним команду **Analyze→Descriptive Statistics→Explore.** Добавляем наши переменные в Dependend list и не забываем отметить во вкладке Plots галочку напротив Histogram, это позволит визуализировать распределение для переменных и напротив Normality plots with tests, это выведет нам тест Колмогорова-Смирнова.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
People living in cities (%)	,089	107	,036	,969	107	**,013**
People who read (%)	,172	107	,000	,843	107	**,000**

- Для того чтобы интерпретировать полученные результаты сформулируем две гипотезы.
- H0: распределение значений переменной не отличается от нормального распределения
- H1: распределение отличается от нормального
- Для того чтобы подтвердить нормальность распределения нам необходимо чтобы вероятность/значимость (sig.) была больше **0, 05.** В нашем случае значимость меньше, поэтому мы делаем вывод о том, что распределение значимо отличается от нормального.

- Теперь приступим к построению линейной регрессионной модели. Для того чтобы осуществить линейный регрессионный анализ необходимо выполнить команду **Analyze→Regression→Linear Regression.**


Linear Regression
country
populatn
density
urban
religion
lifeexpt
lifeexpm
pop_incr
lexpymort
gdp_cap
region
calories
aids
birth_rt
death_rt
aids_rt
log_gdp
Dependent:
literacy
Block 1 of 1
Previous
Next
Independent(s):
People living in cities (%) [urban]
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
OK
Paste
Reset
Cancel
Help

- После добавления переменных заходим во вкладки. Во вкладке Save сохраняем предсказанные значения (Predicted Values- Unstandardized) и наши остатки (Residuals-Standardized) и нажимаем ОК.
- Теперь пришло время интерпретации результатов. В отчете мы получили несколько таблиц идем по порядку. В первой таблице мы смотрим на второй столбец R Square, который показывает нам качество регрессионной модели, в отличие от R, который представляет собой коэффициент корреляции между переменными. R2, лежит в диапазоне от 0 до 1, соответственно, чем он больше, тем наша регрессионная модель лучше и объясняет большое количество случаев. Интерпретировать регрессионную модель можно в том случае если она объясняет хотя бы 30% случаев и больше, то есть когда R2 больше или равен 0,3. При маленьком R2 дальнейшая интерпретация регрессионной модели является бессмысленной, поскольку предсказанные значения объясняют маленький процент случаев.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,650[a]	**,422**	,416	17,481

a. Predictors: (Constant), People living in cities (%)

b. Dependent Variable: People who read (%)

- Итак, мы с вами видим, что наша модель неплохая, и объясняет 42% случаев. Казалось бы, все хорошо и на этом наша интерпретация может завершиться, однако нам необходимо убедиться в том, что наша модель значима, т.е. корреляционные связи между переменными являются значимыми. Для этого нам необходимо обраться ко второй таблице

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23419,021	1	23419,021	76,636	,000[a]
	Residual	32086,866	105	305,589		
	Total	55505,888	106			

- Для того чтобы правильно интерпретировать результаты этой таблицы нам необходимо сформулировать гипотезы, как мы это делали для теста Колмогорова-Смирнова. Здесь нас также интересует значимость **0,05**, но в этот раз нам важно, чтобы она была ниже этого значения. В нашем случае она приближается к нулю, а, следовательно, мы можем говорить о том, что наша построенная модель статистически значима.
- В таблице коэффициенты нас интересует статистическая значимость нашей переменной, так же как и в предыдущей таблице смотрим на значимость. Делаем вывод о том, что наша переменная является статистически значимой.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	42,928	4,384		9,793	,000
	People living in cities (%)	,622	,071	,650	8,754	**,000**

a. Dependent Variable: People who read (%)

- Таким образом, мы можем сделать вывод о том, что чем больше численность городского населения, тем больше количество читающих людей (R=0, 62). Наша модель получилась статистически значимой и объясняет 40% случаев. Однако следует помнить о том, что распределение наших переменных было ненормальным, поэтому предсказанные нами значения не могут приниматься однозначно.

Множественная регрессия

- Множественная регрессия отличается от простой только количеством независимых переменных, поэтому порядок действий остается прежним. Зависимой переменной у нас будет процент городского населения (Sb), а независимыми переменными будут все оставшиеся (кроме переменной страна).

- Самым важным действием на данном этапе является работа с вкладками. Итак, первая вкладка Statistics, в ней мы отмечаем галочками R2 и вывод корреляционных связей (см. Рис 1). Вкладка Plots построит нам диаграмму рассеивания, позволяющий нам визуально оценить гомо или гетероскедастичность, т.е. однородность или неоднородность наблюдений. По оси Y у нас распологаются остатки (ZRESID), по оси X предсказанные значения (ZPRED). Кроме того, наши остатки должны быть нормальными, поэтому попросим программу построить купол Гаусса (отметим две галочки внизу окна) (см. Рис.2). И наконец последняя вкладка Save, которая позволит нам сохранить наши стандартизированные остатки и предсказанные значения, а также статистики влияния, на основе которых мы сможем исключить сильно влияющие случаи (см. Рис.3)

Рис. 1

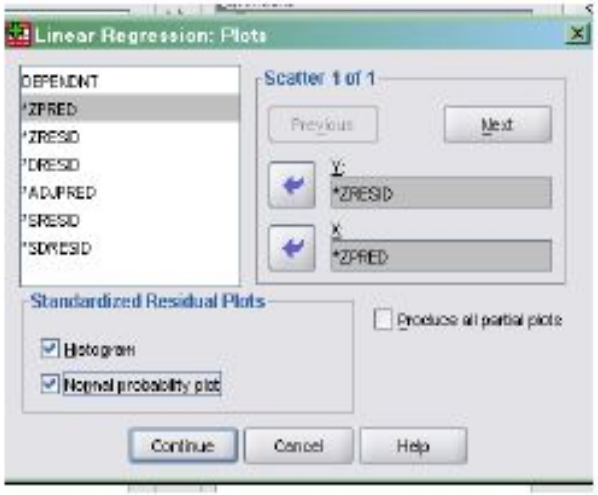


Рис. 2.

Рис.3.

- Интерпретация в целом ничем не отличается от того, что мы делали в простой модели. В первую очередь смотрим на R2, он равен 0,574, это говорит нам о том, что построенная нами модель объясняет 57% случаев. Таблица ANOVA показывает нам, что построенная нами модель является статистически значимой (Sig.=0,003, что меньше 0,05). Теперь смотрим на самую важную таблицу коэффициентов.

Coefficients*

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-11,165	145,016		-,077	,939					
	Средняя продолжительность жизни мужчин	-1,251	1,774	-,222	-,705	,488	,553	-,152	-,101	,204	4,894
	Средняя продолжительность жизни женщин	2,443	2,054	,482	1,189	,248	,681	,251	,169	,124	8,062
	Детская смертность на 1000 новорожденных	-,566	,622	-,299	-,910	,373	-,715	-,195	-,130	,188	5,315
	Количесвто пасмурных дней	,062	,203	,089	,305	,764	,539	,066	,043	,244	4,102
	Средняя температура в январе	,293	,510	,092	,574	,572	-,126	,124	,082	,794	1,260
	Средняя температура в июле	-,933	1,309	-,196	-,837	,531	-,610	-,139	-,091	,215	4,662

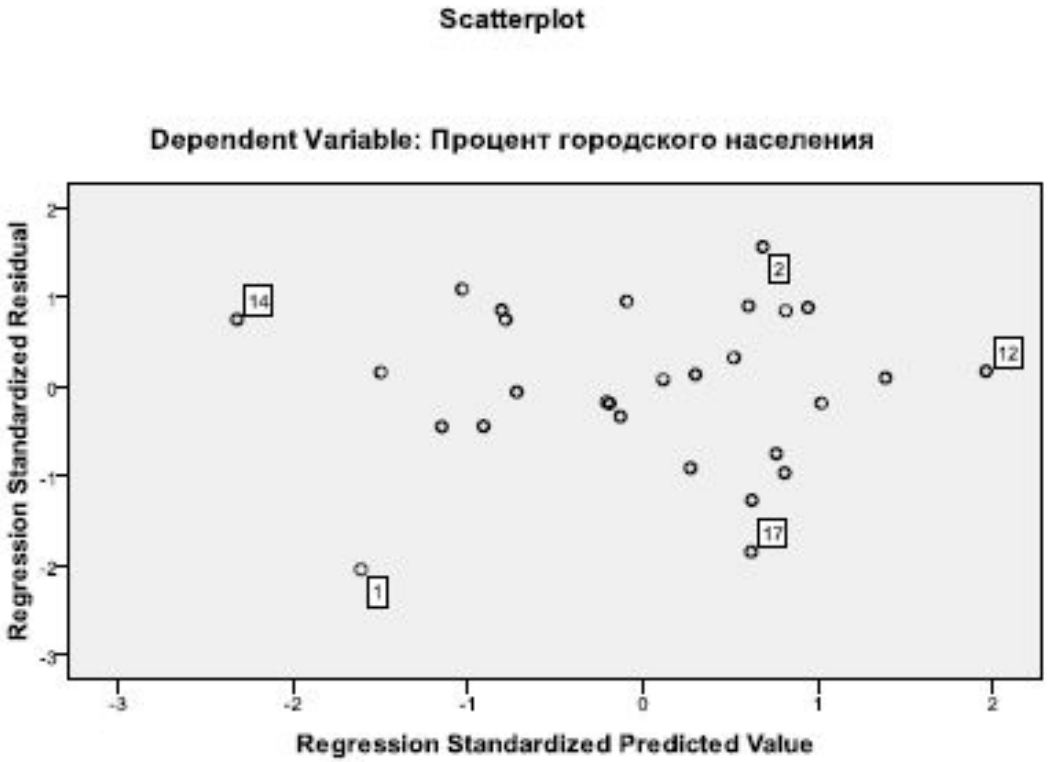
- В первую очередь обращаем внимание на то, какие переменные наиболее важные, т.е. те, чьи значения высокие (столбец B). Как мы видим, есть две переменные, чьи значения практически приближаются к нулю и, следовательно, практически не влияют на нашу модель.

- Поскольку в нашей модели несколько независимых переменных, то мы можем говорить о таком понятии как мультиколлинеарность, т.е. наличии корреляционных связей между независимыми переменными, что, в свою очередь, делает оценку уникальной роли каждой независимой переменной трудной или невозможной. В таблице с коэффициентами мы смотрим на статистики коллинеарности - последний столбец (вывод этого столбца производится посредством вкладки Statistics). Обратимся к толерантности. Коэффициент толерантности показывает уровень мультиколлинеарности данной независимой переменной с другими. Правило большого пальца: на существование мультиколлинеарности указывает значение толерантности <0, 2. Чем ближе толерантность к нулю, тем выше мультиколлинеарность. Второе, что помогает нам выявить наличие мультиколлинеарности это анализ VIF (фактор инфляции дисперсии).

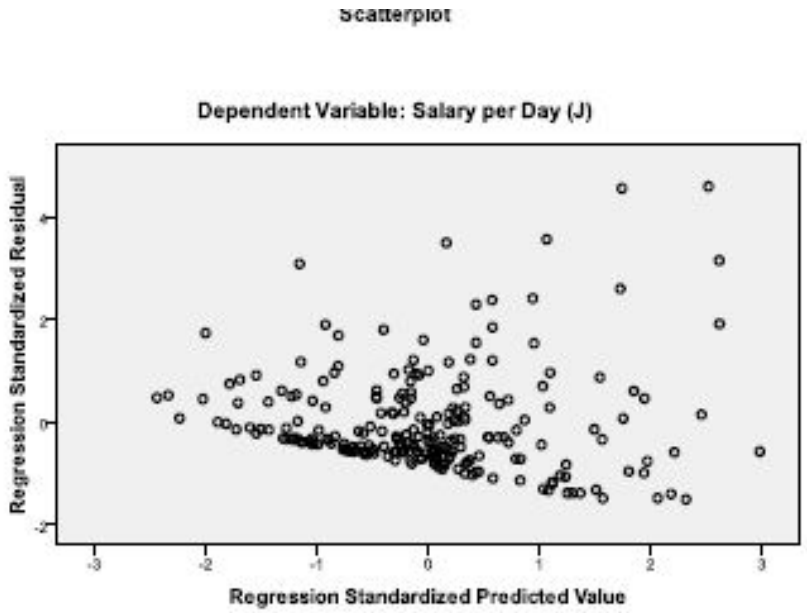
- Данный показатель противоположен по смыслу толерантности, поэтому высокие значения VIF говорят нам о высокой мультиколлинеарности. Принято считать, что значения превышающие 4 указывают на проблемы мультиколлинеарности, однако некоторые ученые используют более мягкий критерий 10 в качестве данного показателя. Вы можете удалить из регрессионной модели переменную с высоким показателем мультиколлинеарности в том случае, если это не противоречит теоретическим ожиданиям. В нашем случае, мы можем удалить вторую и третью переменные, поскольку значение толерантности ниже 0, 2 и показатели VIF довольно высоки. Проблему мультиколлинеарности может решить факторный анализ, который соединяет переменные, измеряющие одинаковые показатели.

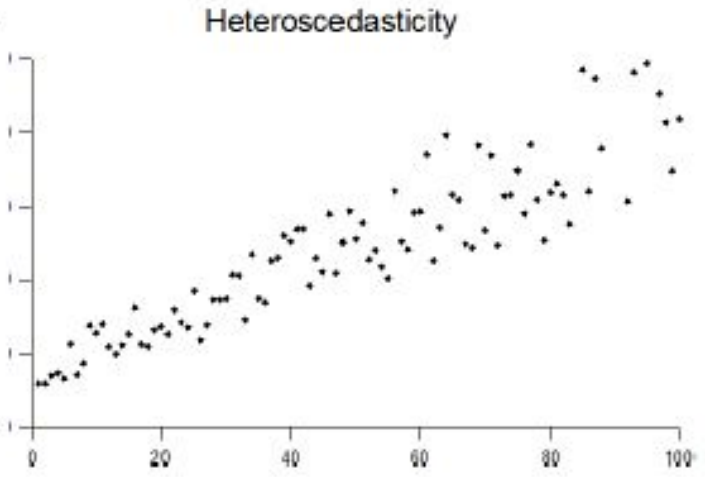
!!! Следует помнить, что если по каким то причинам мы не можем удалить наши переменные из регрессии и сама модель получилась неплохая в интерпретации результатов следует обратить особое внимание на наличие мультиколлинеарности и на то, что это может повлиять на предсказанную модель.

- Теперь нам осталось оценить наши остатки на гомо или гетероскедастичность. Для этого мы построили диаграмму рассеивания. Хотелось бы обратить внимание на выбросы, которые возможно оказывают значительное влияние на конечную регрессионную модель. Эти случаи мы можем на время исключить из конечных выводов, однако помним о том, что удаление не должно противоречить нашим теоретическим ожиданиям. Купол гаусса нам также показывает, что полученные нами остатки отличаются от нормального распределения, однако мы можем заметить, что искажает весь график колонка слева, для того, чтобы посмотреть наличие влияющих наблюдений необходимо обратиться к сохраненным значениям (возвращаемся в основное окно на вкладку DataView).


Scatterplot
Dependent Variable: Процент городского населения
Regression Standardized Residual
Regression Standardized Predicted Value
2
14
12
17
1
-3
-2
-1
0
1
2

- Смотря на диаграмму, мы можем говорить об однородности наблюдений, т.е. о гомоскедастичности (нет общей тенденции, наблюдения располагаются однородно по всей плоскости). Для примера приведем диаграмму рассеивания с гетероскедастичностью.


Scatterplot
Dependent Variable: Salary per Day (J)
Regression Standardized Residual
Regression Standardized Predicted Value
-3
-2
-1
0
1
2
3


Heteroscedasticity
0
20
40
60
80
100

- В обоих представленных случаях можно наблюдать линейную зависимость между значениями переменных. В первом случае остатки уменьшаются с увеличением значения независимой переменной, во втором случае значения остатков возрастают. Оценивание гетероскедастичности является важным пунктом в анализе регрессионной модели, поскольку ее наличие будет указывать нам на неоднородность распределения значения, следовательно, наши статистические выводы будут не совсем адекватны.
- В начале, мы с вами сохраняли остатки, предсказанные значения и статистики влияния, все эти значения у нас появились новыми переменными в нашей базе.
- PRE1 – предсказанные значения
- ZRE1 – стандартизированные остатки SDF1 – статистики влияния
- Для того чтобы выявить влияющие значения отсортируем наши статистики влияния. Самые большие значения по модулю (и положительные, и отрицательные) в большей степени влияют на финальную регрессионную модель. Мы можем сравнить эти значения с выбросами, которые проявились на диаграмме рассеивания. А также сравнить реальные значения с предсказанными. Большое различие говорит о том, что данное наблюдение может являться влияющим.