Регрессионный анализ

Определение

- В основе регрессионного анализа лежит предположение, что зависимая переменная является функций одной или нескольких независимых переменных. Тогда, зная значения независимых переменных, мы можем сделать прогноз об изменении зависимой переменной.
- Регрессионный анализ предполагает построение регрессионного уравнения, его оценку и анализ.

Уравнение парной линейной регрессии

- Простейшей регрессионной моделью является парная линейная регрессия.
- Уравнение парной линейной регрессии в общем виде следующее:

  $y=b_0+b_1x$, где
- b_0 – свободный член уравнения регрессии (Константа);
- b_1 –коэффициент уравнения регрессии.

Требования к исходным данным регрессионного анализа

1. Зависимая (результирующая) переменная должна быть непрерывной количественной переменной. Независимая переменная должна быть непрерывной или дихотомической. Категориальные независимые переменные с более чем двумя значениями перекодируются в набор дихотомических переменных.

2. Изучаемая совокупность должна быть достаточно большой, чтобы показатели связей были статистически надежными (число единиц совокупности должно превосходить число коррелируемых переменных не менее чем в 6-8 раз).

Требования к исходным данным регрессионного анализа

3. Каждое значение зависимой переменной должно быть независимо от других значений. Такие зависимости возникают если опрашивать одного и того же респондента в разные периоды времени или опрашивать респондентов, объединенных в группы (семья, бригада и т. д.).
4. Распределение зависимой переменной должно быть близким к нормальному и не иметь явных выбросов.
5. Должно выполняться требование гомоскедактичности, что означает, что ошибки не становятся меньше, если уменьшается значение **y** и не растут с увеличением значений **y**. Это предположение проверяется при построении диаграммы рассеяния между стандартизованными остатками и стандартизованными предсказанными значениями. Если облако рассеяния овальное – данные гомоскедактичные. Если облако рассеяния принимает форму конуса, требование гомоскедактичности нарушается и данные являются гетероскедактичными.

Требования к исходным данным регрессионного анализа

6. Ошибка предсказания для каждого значения не должна зависеть от ошибки предсказания других значений (тест Дарбина-Уотсона), остатки должны быть нормально распределены (график остатков).

7. Для случая множественной регрессии должно отсутствовать явление мультиколлинеарности, которое возникает, когда независимые переменные сильно коррелируют между собой. Такого рода корреляция может оказать сильное воздействие на зависимый признак и это уже будет иное воздействие, чем независимых переменных по отдельности.

Пример:

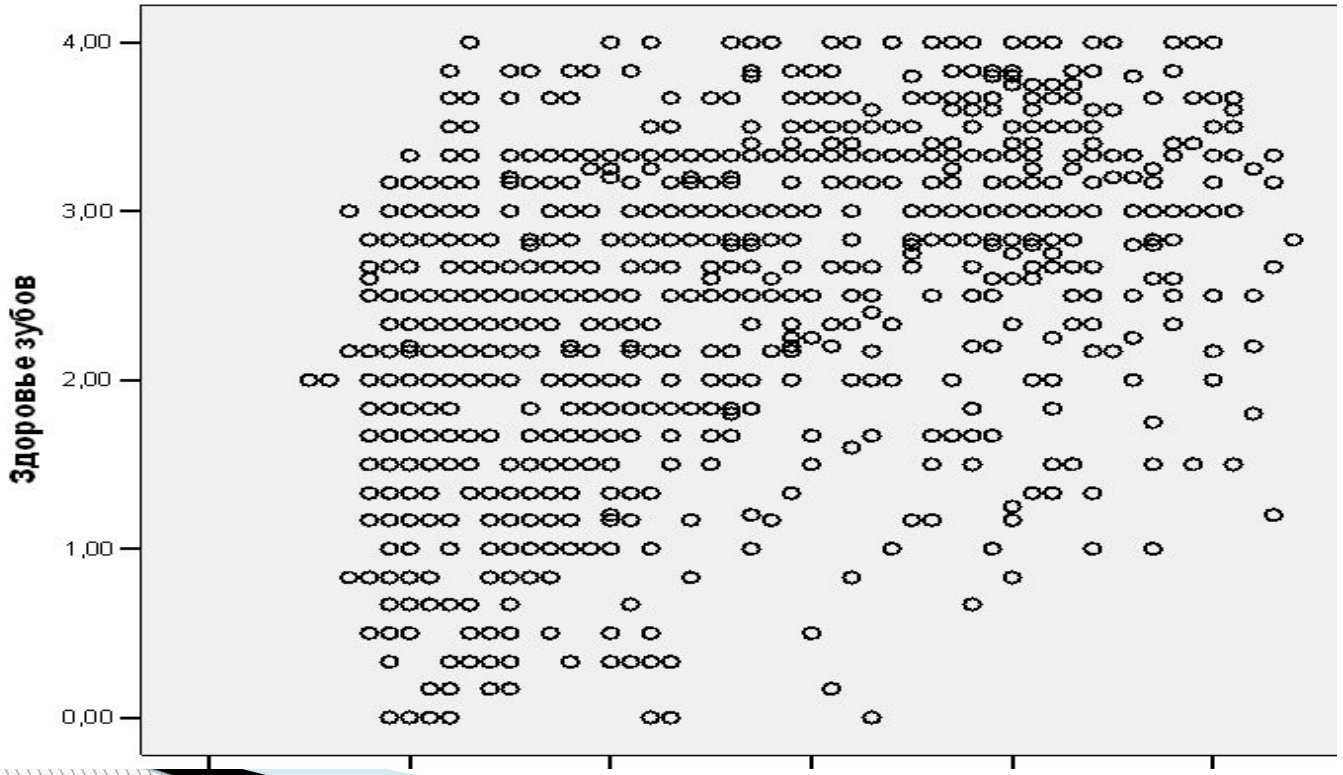
Построить уравнение парной линейной регрессии для переменных «Возраст» и «Заболевания зубов» (измеренной по пятибалльной шкале, где 0 - здоровые зубы, а 4 – наибольшая степень развития заболевания)

Проверка причинно-следственной связи

1. Теоретически мы должны доказать, что изучение связи между причиной и следствием имеет смысл.
2. Причина всегда по времени должна предшествовать следствию.
3. Причина должна коррелировать со следствием.

Рассмотрим корреляцию переменных «Возраст» и «Заболевания зубов»

Chart /Графики→ Scatterdot .../Рассеяние.точки → Simple/простая диаграмма рассеяния



Проверка на наличие корреляции возраста и заболевания зубов

Analyze/ Анализ → Correlation/Корреляции → Bivariate/Парные

Correlations

		Возраст	Заболевание зубов
Возраст	Pearson Correlation	1	,425**
	Sig. (2-tailed)	.	,000
	N	1130	1130
Заболевание зубов	Pearson Correlation	,425**	1
	Sig. (2-tailed)	,000	.
	N	1130	1130

**. Correlation is significant at the 0.01 level (2-tailed).

Построение парной линейной регрессии

Выполнение команды:

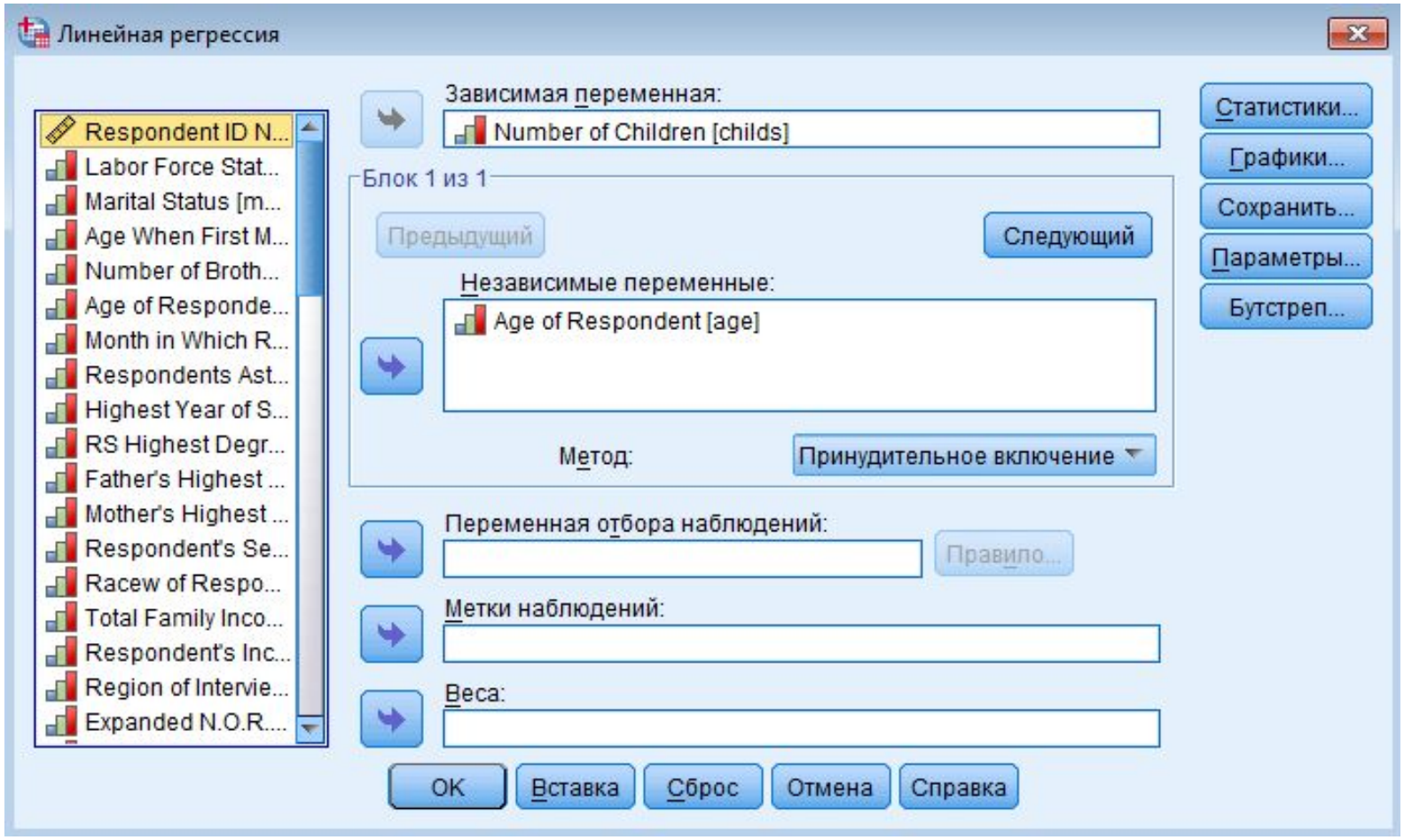
Analyze/Анализ Regression/Регрессия Linear/Линейная

В поле Dependent

Имя зависимой переменной

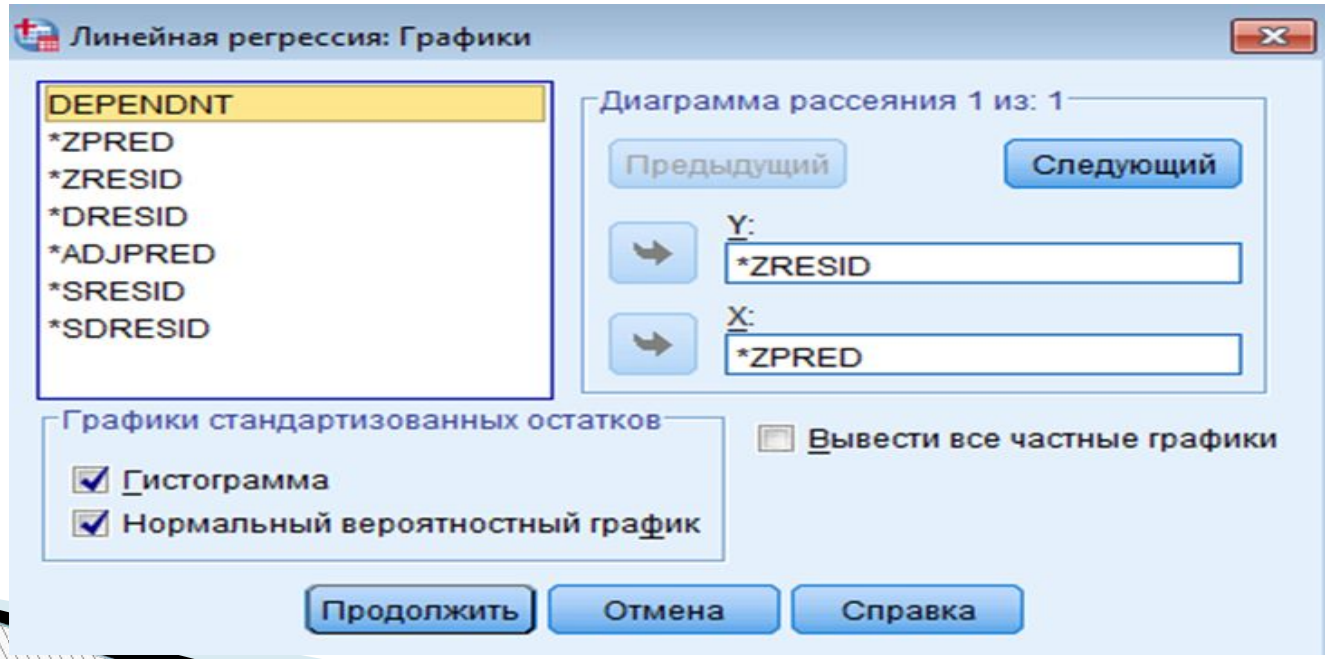
В поле Independent(s)

Имя независимой переменной OK


Линейная регрессия
Respondent ID N...
Labor Force Stat...
Marital Status [m...
Age When First M...
Number of Broth...
Age of Responde...
Month in Which R...
Respondents Ast...
Highest Year of S...
RS Highest Degr...
Father's Highest ...
Mother's Highest ...
Respondent's Se...
Racew of Respo...
Total Family Inco...
Respondent's Inc...
Region of Intervie...
Expanded N.O.R....
Зависимая переменная:
Number of Children [childs]
Блок 1 из 1
Предыдущий
Следующий
Независимые переменные:
Age of Respondent [age]
Метод:
Принудительное включение
Переменная отбора наблюдений:
Правило...
Метки наблюдений:
Веса:
OK
Вставка
Сброс
Отмена
Справка
Статистики...
Графики...
Сохранить...
Параметры...
Бутстреп...

Дополнительные настройки

- Кнопка «Статистики/Statistics» - активизируем вычисление теста Дарбина-Уотсона;
- Кнопка «Графики/Plots» - помечаем вывод в отчет графиков стандартизованных остатков (Гистограмма, Нормальный вероятностный график), а также задаем Диаграмму рассеяния стандартизованных предсказанных значений (ZRESID по оси X) и стандартизованных остатков (ZPRED по оси У)



Результаты выполнения команд регрессионного анализа

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. ошибка	Бета		
1	(Константа)	1,295	,071		18,220	,000
	Возраст	,033	,002	,452	17,006	,000

a. Зависимая переменная: Заболевания зубов

- y=1,295+0,033x

Анализ качества регрессионной модели

Сводка для модели[b]

Модель	R	R квадрат	Скорректированный R квадрат	Стд. ошибка оценки	Дурбин-Уотсон
1	,452[a]	,204	,203	,83156	1,845

a. Предикторы: (константа) Возраст

b. Зависимая переменная: Заболевания зубов

Нормальный P-P график для регрессии Стандартизированный остаток

Зависимая переменная: Заболевания зубов

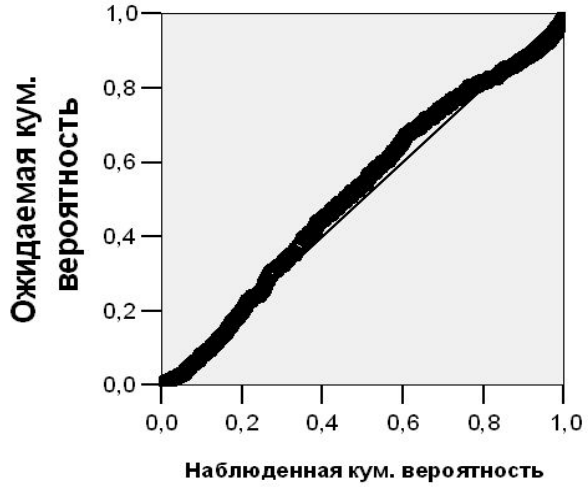
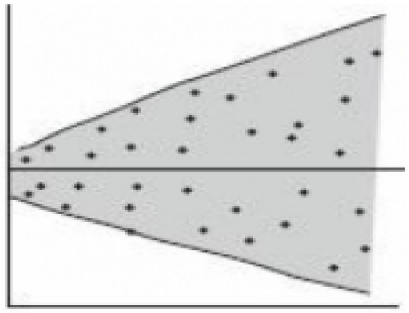
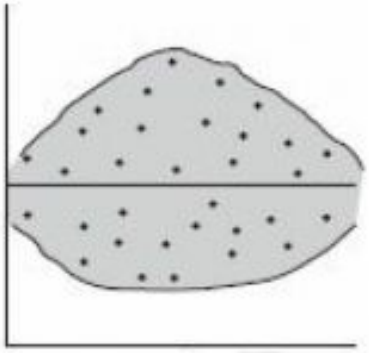


Диаграмма рассеяния стандартных остатков и стандартизированных предсказанных значений, проверка гомоскедактичности

Диаграмма рассеяния остатков



явление гомоскедактичности отсутствует



Остатки гомоскедактичные

Множественная линейная регрессия

В большинстве задач следствие не может быть объяснено одной единственной причиной; как правило, приходится изучать влияние на него нескольких причин одновременно. Для исследования такой множественной связи используется ***уравнение множественной линейной регрессии***:

Пример:

Построить уравнение множественной линейной регрессии для зависимой переменной «Заболевания зубов» и независимых переменных «Возраст», «Периодичность чистки зубов».

Множественная линейная регрессия

Выполнение команды:
Analyze Regression Linear

В поле Dependent
Имя зависимой переменной

В поле Independent(s)
Имена независимых переменных

Дополнительные вычисления аналогичны парной регрессии

Выбор метода анализа

- В случае множественной регрессии можно использовать установленный по умолчанию метод Enter (включения всех переменных в модель одновременно)
- или специальный пошаговый метод Stepwise (модель строиться не для всех исходных причин сразу, а пошагово в модель включаются новые причины, оговоренные в условии)

Корреляционная таблица

Correlations

		Возраст	Здоровье зубов	Количество смен зубных щеток за год	Сколько раз в день чистят зубы?
Возраст	Pearson Correlation	1	,452**	,008	-,049
	Sig. (2-tailed)	.	,000	,776	,100
	N	1130	1130	1130	1130
Здоровье зубов	Pearson Correlation	,452**	1	-,251**	-,359**
	Sig. (2-tailed)	,000	.	,000	,000
	N	1130	1130	1130	1130
Количество смен зубных щеток за год	Pearson Correlation	,008	-,251**	1	,359**
	Sig. (2-tailed)	,776	,000	.	,000
	N	1130	1130	1130	1130
Сколько раз в день чистят зубы?	Pearson Correlation	-,049	-,359**	,359**	1
	Sig. (2-tailed)	,100	,000	,000	.
	N	1130	1130	1130	1130

**. Correlation is significant at the 0.01 level (2-tailed).

Результаты множественной линейной регрессии (метод Enter)

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,461	,107		22,972	,000
	Возраст	,033	,002	,439	18,080	,000
	количество смен зубных щеток за год	-,020	,008	-,153	-2,582	,000
	Сколько раз в день чистят зубы?	-,258	,049	-,585	-10,847	,000

a. Dependent Variable: Здоровье зубов

Уравнение множественной регрессии
y=2,461+0,033возраст - 0,05щетки – 0,528чистки

Стандартизованное уравнение множественной регрессии
y=0,439возраст - 0,153щетки – 0,282чистки

Качество множественной линейной регрессии. Метод Enter

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,581[a]	,338	,336	,75898

a. Predictors: (Constant), Сколько раз в день чистят зубы?, Возраст, Количество смен зубных щеток за

Нормальный P-P график для регрессии Стандартизированны

Зависимая переменная: Заболевания зубов

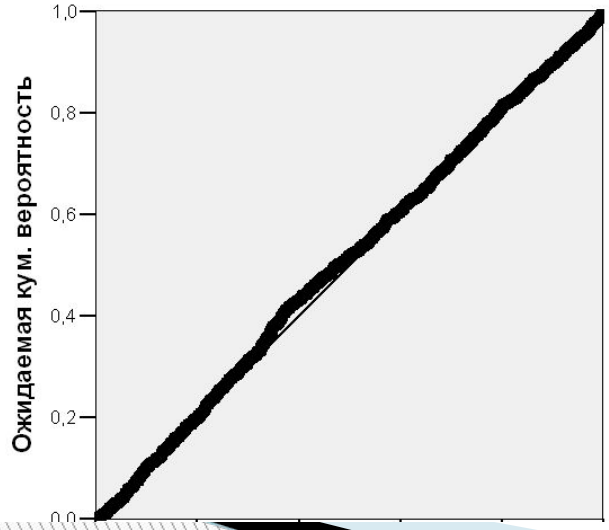
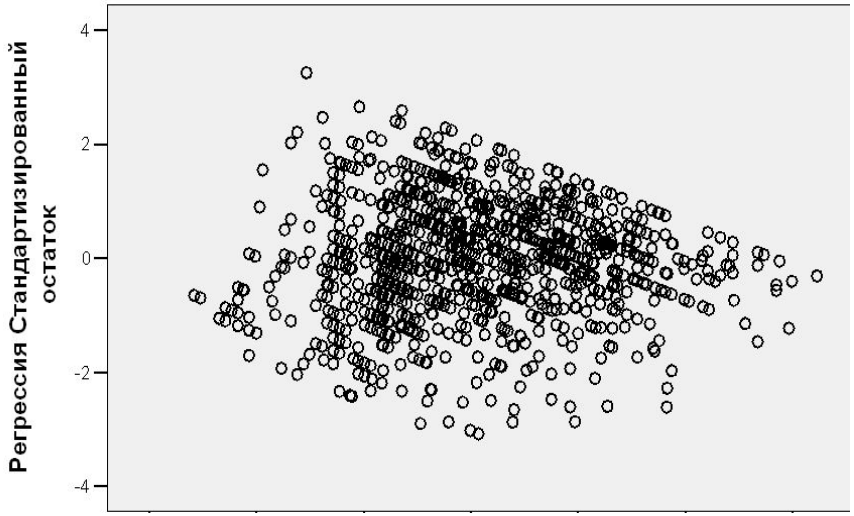


Диаграмма рассеяния

Зависимая переменная: Заболевания зубов



Результаты множественной линейной регрессии (метод Stepwise)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,452[a]	,204	,203	,83156
2	,564[b]	,318	,316	,77027
3	,581[c]	,338	,336	,75898

a. Predictors: (Constant), Возраст

b. Predictors: (Constant), Возраст, Сколько раз в день чистят зубы?

c. Predictors: (Constant), Возраст, Сколько раз в день чистят зубы?, Количество смен зубных щеток за год

Коэффициенты множественной линейной регрессии (метод Stepwise)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	1,295	,071		18,220	,000
	Возраст	,033	,002	,452	17,006	,000
2	(Constant)	2,480	,109		22,815	,000
	Возраст	,032	,002	,435	17,666	,000
	Сколько раз в день чистят зубы?	-,631	,046	-,337	-13,666	,000
3	(Constant)	2,461	,107		22,975	,000
	Возраст	,032	,002	,439	18,086	,000
	Сколько раз в день чистят зубы?	-,528	,049	-,282	-10,847	,000
	Количество смен зубных щеток за год	-,050	,008	-,153	-5,865	,000

a. Dependent Variable: Здоровье зубов