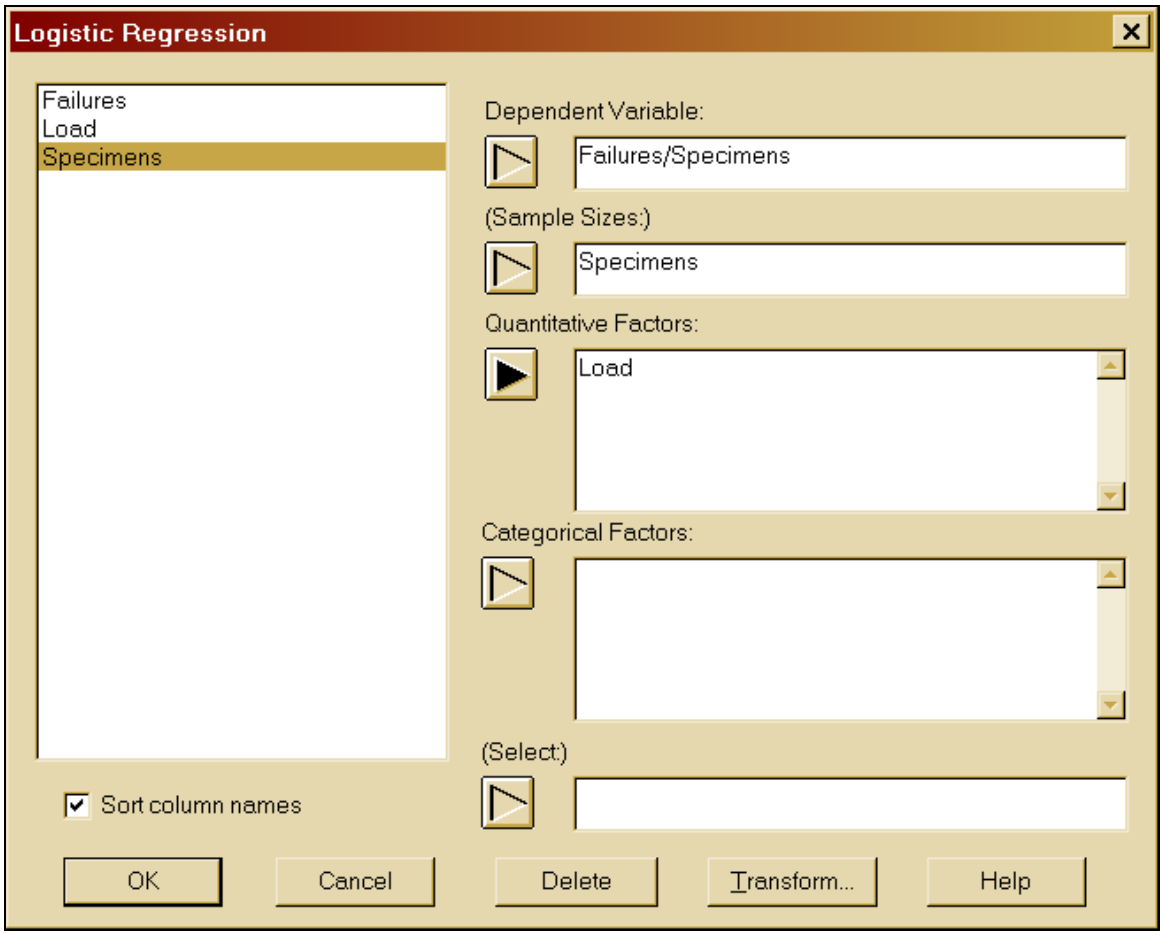
ЛАБОРАТОРНАЯ РАБОТА №9

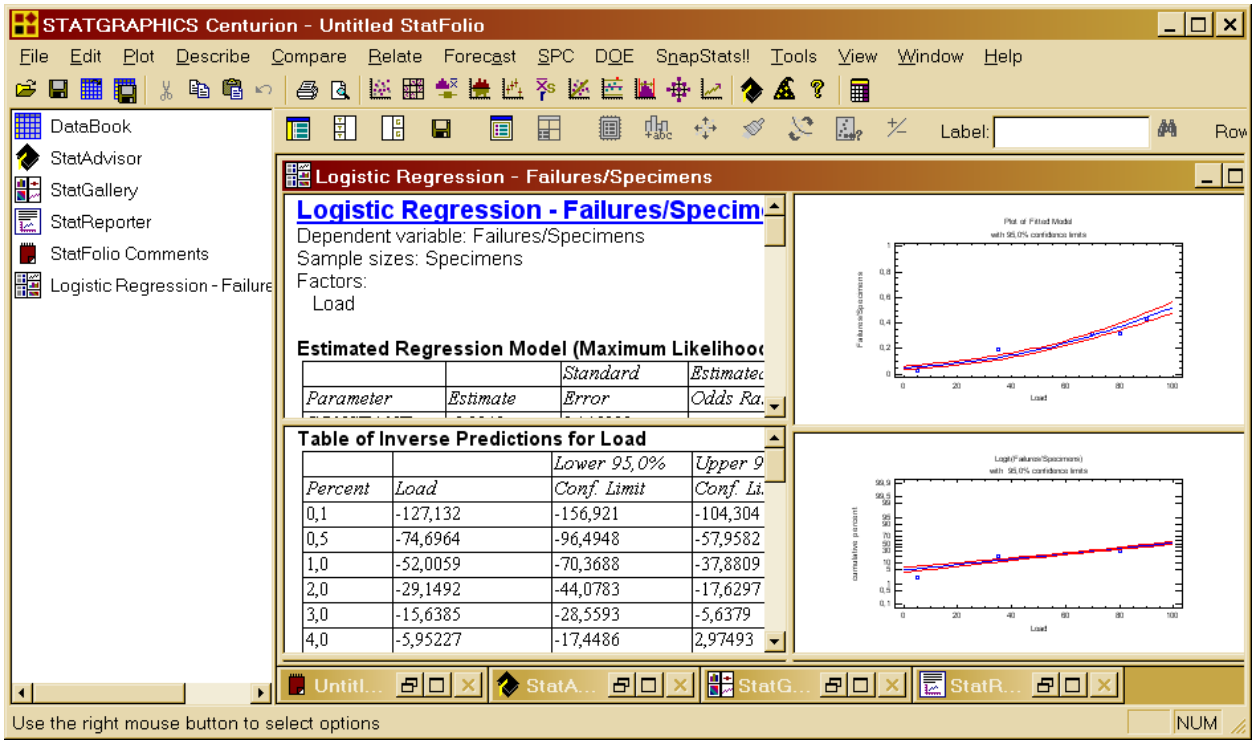
Логит и пробит модели в Ewiews и StatGraphics.

Задача. В файле fabric.sf содержатся данные о 2300 образцах (Spesimens), подвергнутых нагрузке (Load). В результате нагрузки некоторая часть образцов разрушается (Failures). Построить уравнение регрессии, в котором будет выведена зависимость отношения испорченных образцов к общему количеству в зависимости от нагрузки.

Задачи такого типа можно решать с помощью логистической регрессии. Откройте пакет StatGraphics. В строке меню выбираем Relate\ Attribute Data\ Logistic Regression



В задачах, где зависимая переменная – пропорция, надо заполнить поле (Sample Size), введите туда переменную **Load**. Нажмите OK. Раскроется окно с результатами логистической регрессии.



Рассмотрим эти результаты внимательнее. Само уравнение выглядит так:

$$\frac{Failures}{Specimens} = \frac{e^{(-2,99+0,03*Load)}}{1+e^{(-2,99+0,03*Load)}}$$

Logistic Regression - Failures/Specimens

Dependent variable: Failures/Specimens
Sample sizes: Specimens
Factors:
Load

Estimated Regression Model (Maximum Likelihood)

		Standard	*Estimated*
Parameter	*Estimate*	*Error*	*Odds Ratio*
CONSTANT	-2,9949	0,145939	
Load	0,0307699	0,00209432	1,03125

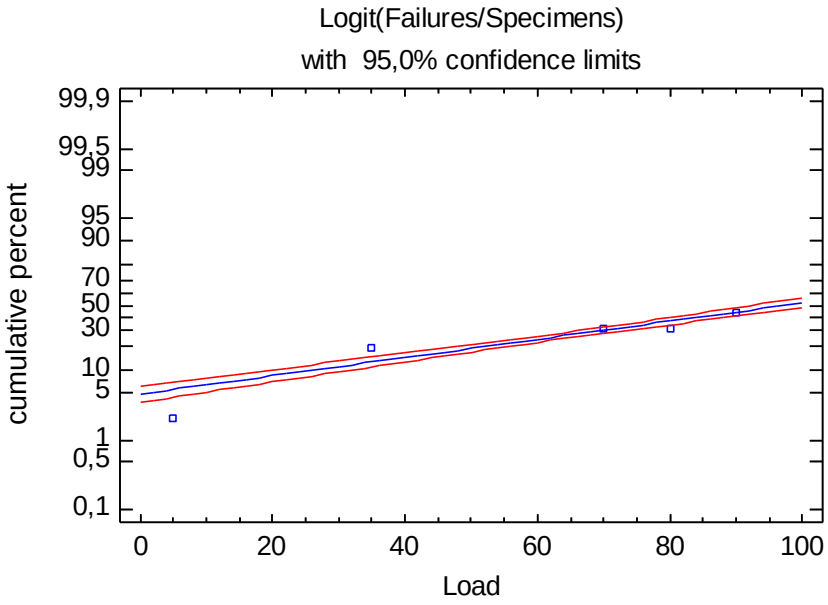
Analysis of Deviance

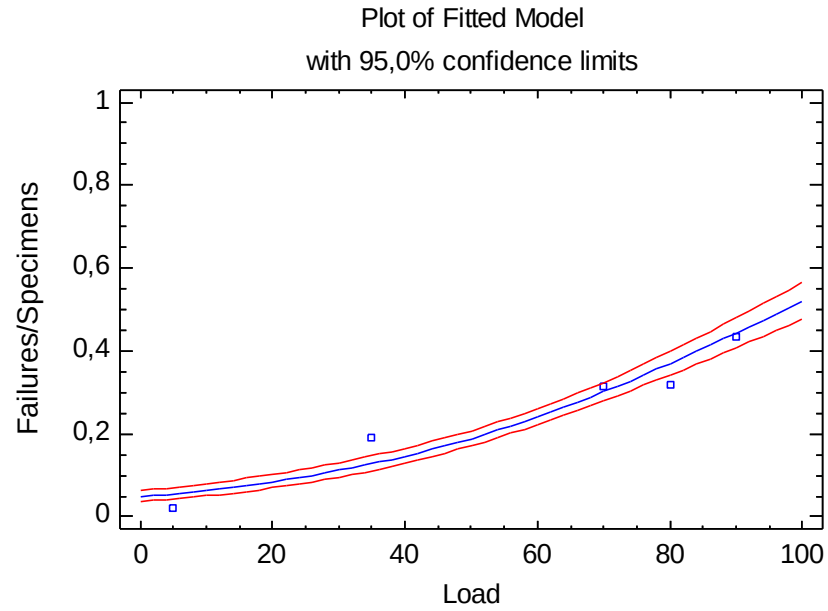
Source	*Deviance*	*Df*	*P-Value*
Model	283,056	1	0,0000
Residual	36,2181	3	0,0000
Total (corr.)	319,274	4	

Percentage of deviance explained by model = 88,6561
Adjusted percentage = 87,4033

Как видно из результатов, **Adjusted percentage = 87,4033**, следовательно, модель объясняет 87% всех изменений. Это – очень хороший результат. Посмотрим на **Analysis of Deviance**. *P-Value=0,00* для модели говорит о том, что модель хорошая. Но *P-Value=0,00* для остатков говорит о том, что остатки не свободны, следовательно, не все в модели мы учли.

Estimated Odds Ratio=1,03125 говорит о том, что при увеличении нагрузки на единицу, отношение увеличивается на 3%. Посмотрим на графическое изображение логистической регрессии.





Поскольку не все в модели осталось объясненным, перейдем к полулогарифмической модели, т.е. в качестве объясняющей переменной рассмотрим Log(Load). Самостоятельно введите данные, рассмотрим результаты

Estimated Regression Model (Maximum Likelihood)

		Standard	*Estimated*
Parameter	*Estimate*	*Error*	*Odds Ratio*
CONSTANT	-5,5784	0,368202	
log(Load)	1,13997	0,0892554	3,12667

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	313,886	1	0,0000
Residual	5,38828	3	0,1455
Total (corr.)	319,274	4	

Percentage of deviance explained by model = 98,3123
Adjusted percentage = 97,0595

Likelihood Ratio Tests

Factor	*Chi-Squared*	*Df*	*P-Value*
log(Load)	313,886	1	0,0000

Residual Analysis

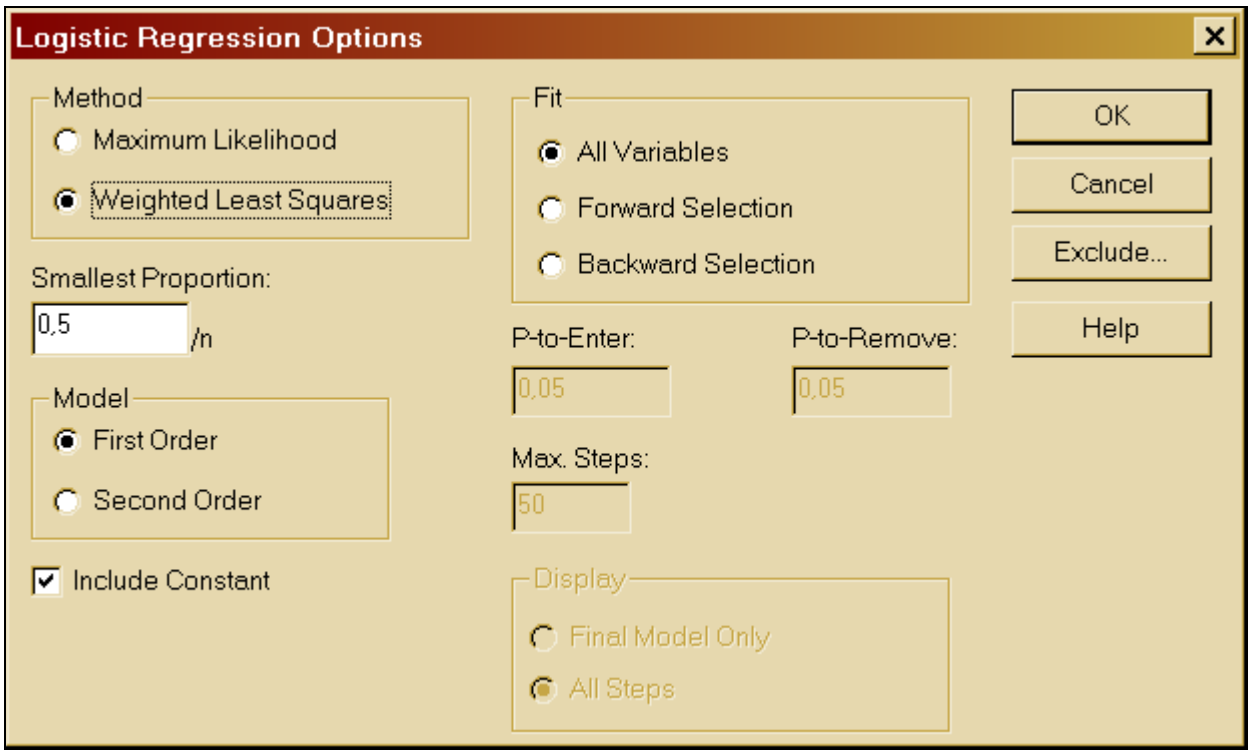
	Estimation	*Validation*
n	5	
MSE	0,0918328	
MAE	0,0223697	
MAPE	7,05518	
ME	-0,000356236	
MPE	-0,604615	

Сама модель здесь запишется в виде $\frac{Failures}{Specimens} = \frac{e^{(-2,99+0,03*\log(Load))}}{1+e^{(-2,99+0,03*\log(Load))}}$

Видно, что это уравнение объясняет 97% изменений, сама модель значима, а P-Value=0,1455 для остатков говорит о том, что в остатках нет зависимостей. Модель устраивает нас.

В случае, когда объясняемая переменная является отношением, можно применить другой метод – метод взвешенных квадратов.

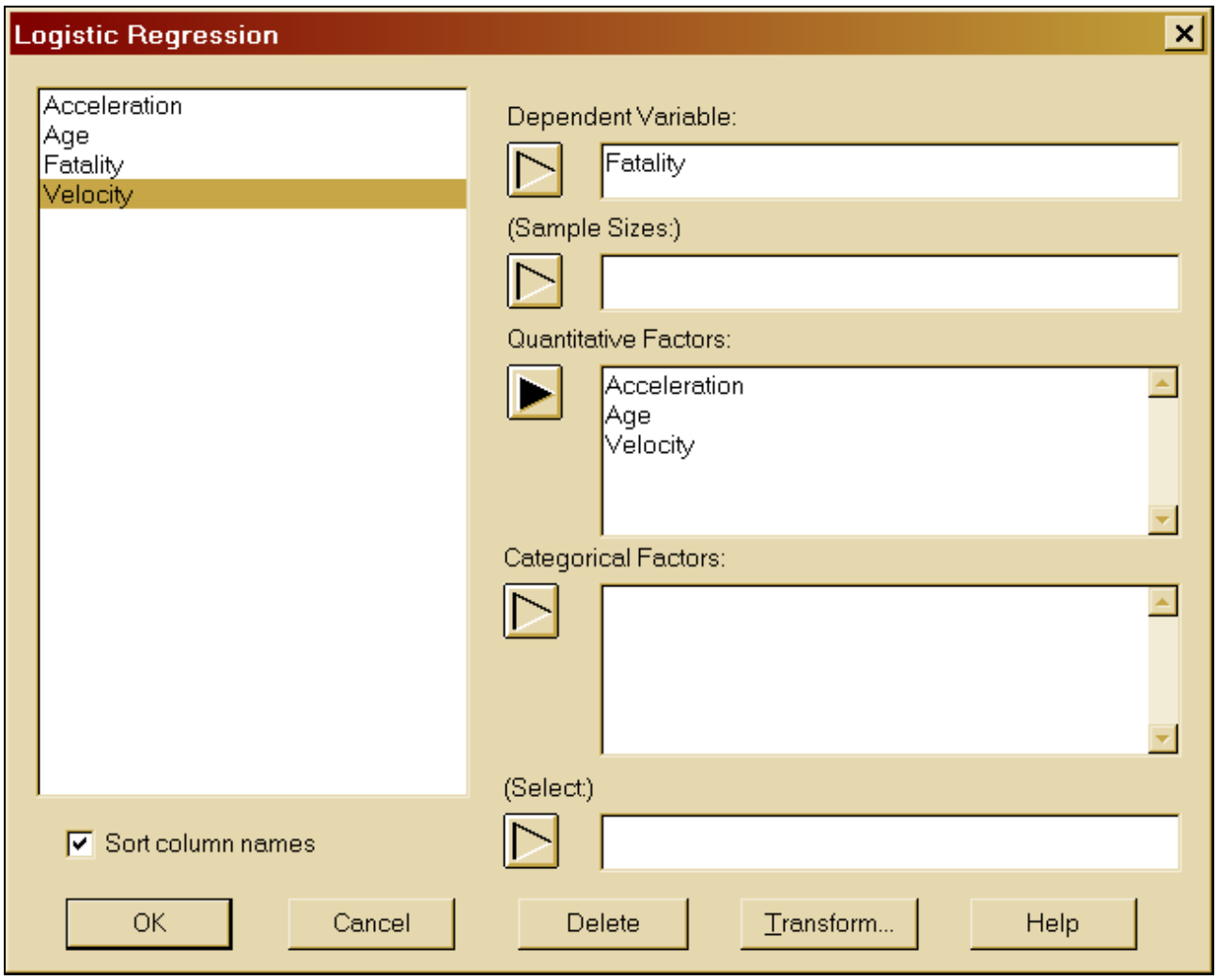
Щелкните правой кнопкой по окну анализа, выберите Analysis Options



В этом окне выберите Weighted Least Squares, нажмите ОК. Результаты рассмотрите самостоятельно.

Задача. В файле collisions.sf имеются данные об авариях автомобилей. Объясняемая переменная Fatality принимает два значения – 1, если травма была смертельной, 0 – в противном случае. Age – возраст автомобилистов, Velocity – скорость, Acceleration – ускорение.

В строке меню выбираем Relate\ Attribute Data\ Logistic Regression, введите данные



Нажмите ОК. Рассмотрим внимательно результаты.

Estimated Regression Model (Maximum Likelihood)

		Standard	*Estimated*
Parameter	*Estimate*	*Error*	*Odds Ratio*
CONSTANT	-15,0536	5,29655	
Acceleration	0,0161775	0,0144943	1,01631
Age	0,170905	0,0432274	1,18638
Velocity	0,146279	0,11218	1,15752

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	34,6964	3	0,0000
Residual	43,9759	54	0,8331
Total (corr.)	78,6723	57	

Percentage of deviance explained by model = 44,1024
Adjusted percentage = 33,9336

Likelihood Ratio Tests

Factor	*Chi-Squared*	*Df*	*P-Value*
Acceleration	1,3556	1	0,2443
Age	31,2312	1	0,0000
Velocity	1,83353	1	0,1757

Модель объясняет 33,9% всех изменений. Это, конечно, не много. Видно к тому же, что переменная **Acceleration** не является значимой. Кстати, если ее удалить, результат улучшится. Убедитесь в этом самостоятельно.

Используя логистическую регрессию, можно пытаться предсказывать результат. Например, в этой задаче введите в 59 строку следующие данные: Age=50, Acceleration=180, Velocity=60. Fatality оставьте незаполненным. Щелкните по кнопке **Tables**, выберите **Predictions.** Появится результат

Predictions for Fatality

	Observed	*Fitted*	*Lower 95,0%*	*Upper 95,0%*
Row	*Value*	*Value*	*Conf. Limit*	*Conf. Limit*
59		0,994375	0,91099	0,999673

Проверьте самостоятельно, что будет, если возраст будет 20 лет, если скорость 40? Результаты покажите преподавателю.

Можно посмотреть, какие из 58 данных кажутся StatGraphics'у «неправильными». Щелкните по кнопке **Tables**, выберите **Unusual Residuals**. Получился результат

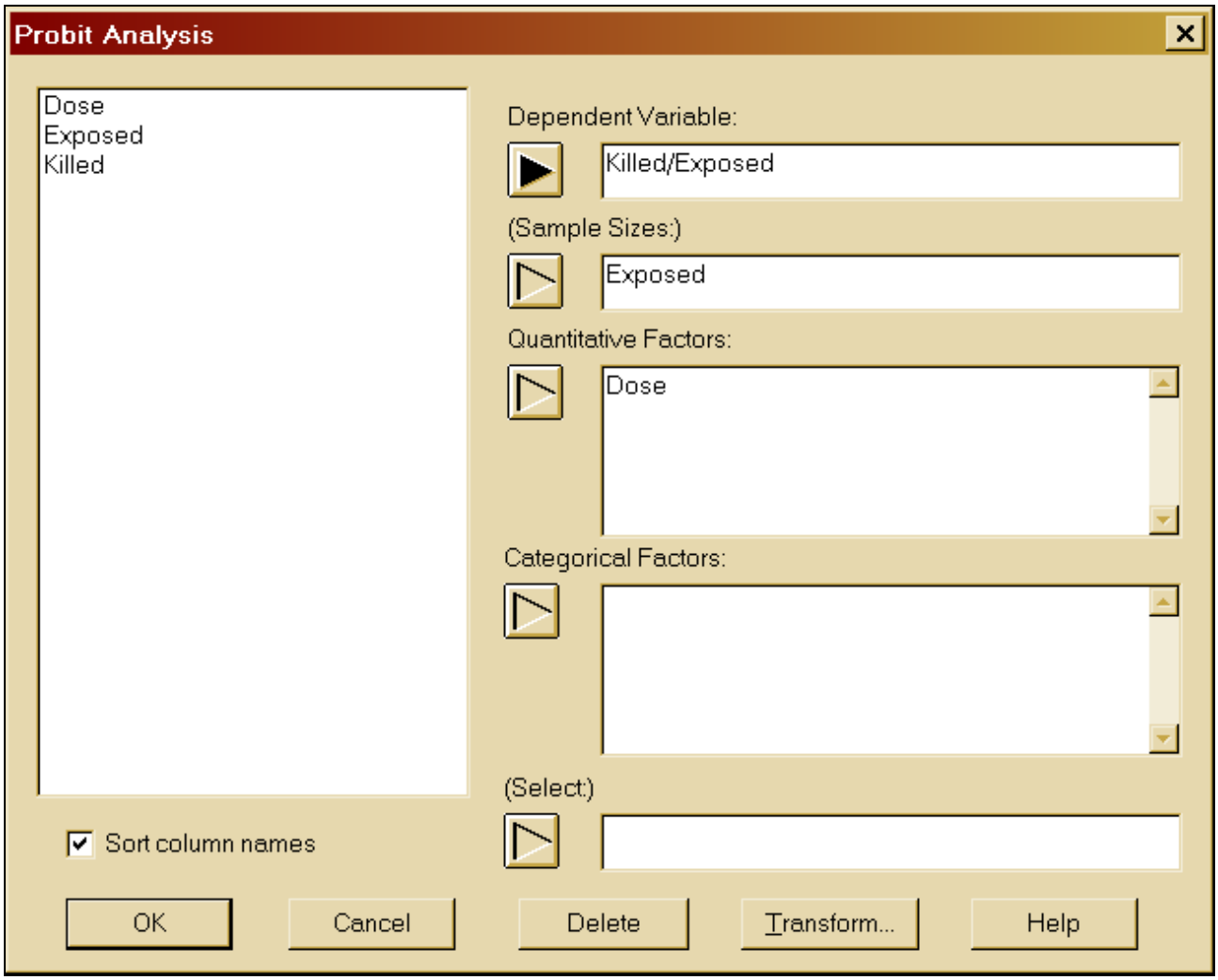
Unusual Residuals for Fatality

		Predicted Y	*Residual*	*Pearson Residual*	*Deviance Residual*
Row	*Y*				
23	0,0	0,952518	-0,952518	-4,48	-2,47
31	1,0	0,103655	0,896345	2,94	2,13
35	0,0	0,91865	-0,91865	-3,36	-2,24

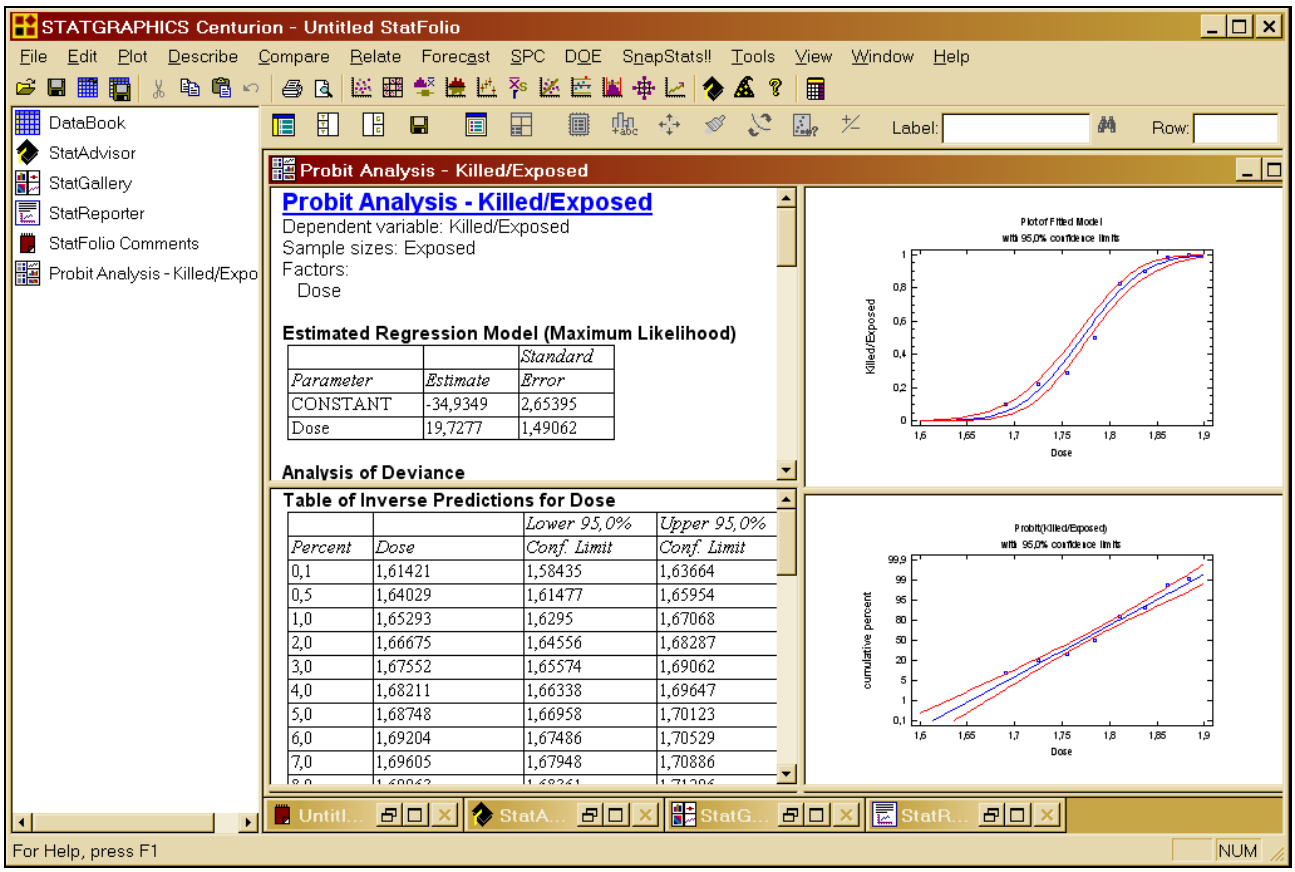
Видно, что StatGraphics'у кажутся неправильными результаты 23, 31 и 35 строк.

Задача. В файле beetled.sf приведены данные о действии химикатов на жуков. Исследовалось влияние разных доз на 481 экземпляр жуков. Часть из них под действием химикатов погибала. Воспользуемся Probit- анализом для исследования модели выживаемости жуков.

Решим эту задачу в пакете StatGraphics. В строке меню выберите Relate\ Attribute Data\ Probit Analysis. Введите данные



Получился результат

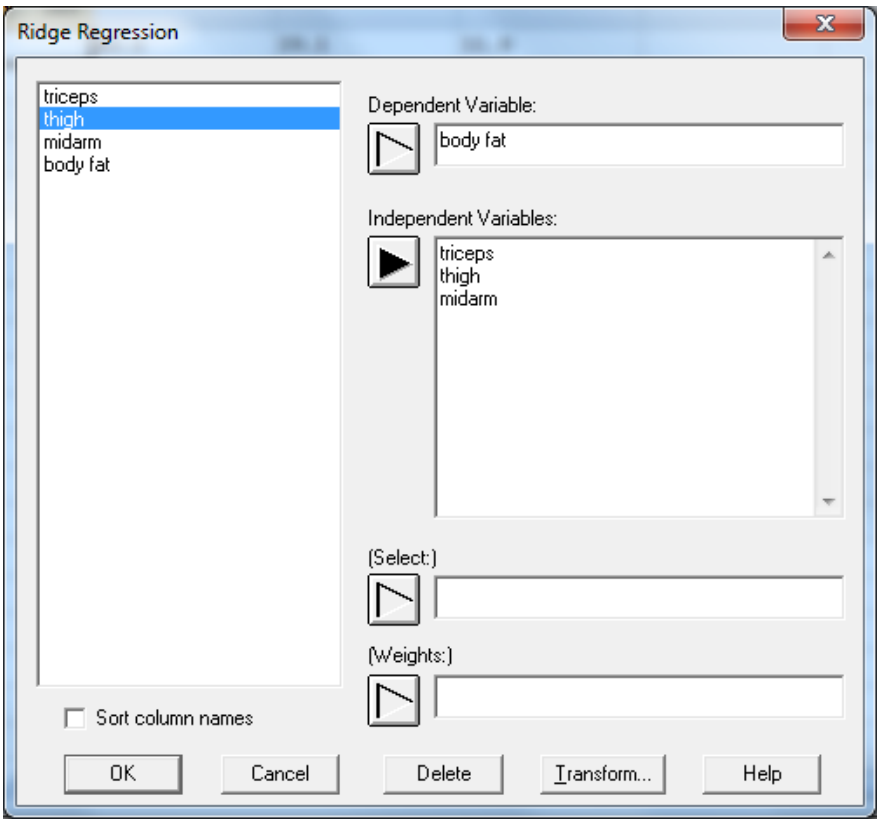


Рассмотрите самостоятельно результат. Покажите преподавателю.

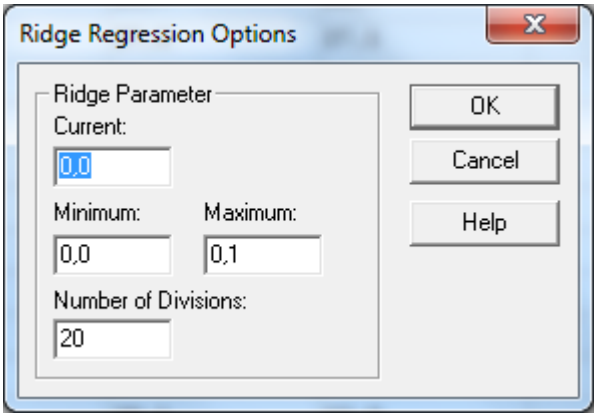
РИДЖ-РЕГРЕССИЯ

Ридж-регрессия дает нам один из способов избавления от мультиколлинеарности. Откройте файл bodyfat.sf6. в нем приведены данные измерений, произведенных у 20 женщин (индекс жира в теле bodyfat, трицепс (triceps), окружность бедра (thigh) и середина предплечья (midarm)). Построим регрессионную модель (Multiple Regression). Модель получается хорошая, R^2=76,41, но все коэффициенты статистически незначимы, поэтому пользоваться такой моделью, вообще говоря, нельзя.

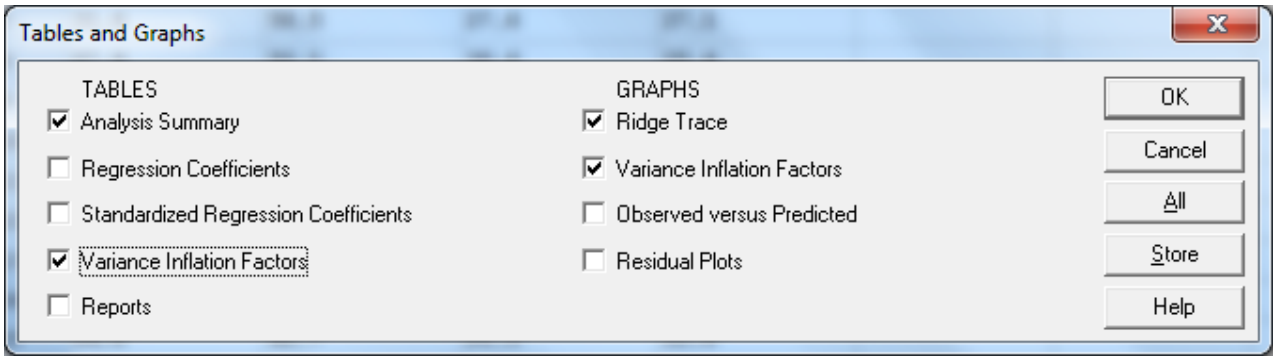
Запустим Ридж-регрессию. Relate/ Multiple Factors/ Ridge Regression.



нажмите OK. Раскроется окно



пока в этом окне оставим все как есть. Еще раз OK и выберите **Variance Inflation Factors**



OK,

Ridge Regression - body fat

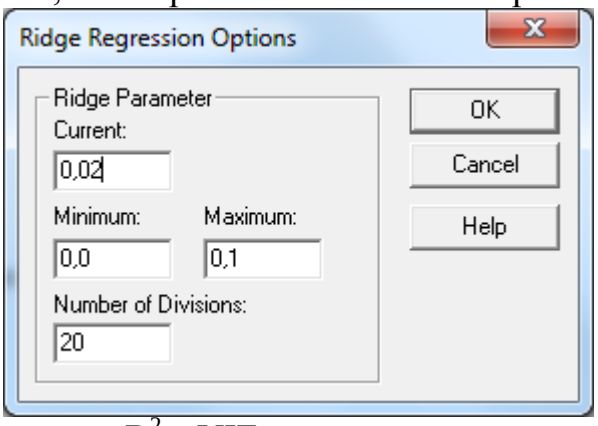
Dependent variable: body fat
Independent variables:
triceps (Neter, p. 385)
thigh
midarm
Number of complete cases: 20

Model Results for Ridge Parameter = 0,0

		Variance
		Inflation
Parameter	*Estimate*	*Factor*
CONSTANT	117,085	
triceps	4,33409	708,843
thigh	-2,85685	564,343
midarm	-2,18606	104,606

R-Squared = 80,1359 percent
R-Squared (adjusted for d.f.) = 76,4113 percent
Standard Error of Est. = 2,47998
Mean absolute error = 1,88563
Durbin-Watson statistic = 2,24291
Lag 1 residual autocorrelation = -0,167729

Здесь все как в обычной регрессии, но показатель VIF (Variance Inflation Factors) показывает наличие мультиколлинеарности. Попробуйте менять параметр λ. Для этого в окне щелкните правой кнопкой, выберите Analysis Options, раскроется окно, в котором можно менять параметр λ



Посмотрите, как изменилось R^2 и VIF.

ЗАДАНИЕ

1. В файле rusdasl1.xls приведены данные о больных сердечниках. Постройте уравнение логистической регрессии. Допустим, поступил больной 70 лет. У него была фибрилляция желудочков, задний ИМ, не было рецидива, появилась аневризма сердца, степень острой сердечной недостаточности – легочно-венозный застой, есть реперфузия при тромболитической терапии, хроническая сердечная недостаточность второй степени. Какой можно сделать прогноз о его выживаемости? А если у него передний ИМ? Сильно ли ухудшает выживаемость наличие рецидива?

2. В файле Диабет.xls имеются данные о 768 больных со следующими данными:

1. Число случаев беременности

2. Концентрация глюкозы;
3. Артериальное диастолическое давление, мм. рт. ст.;
4. Толщина кожной складки трехглавой мышцы, мм.;
5. 2-х часовой сывороточный инсулин;
6. Индекс массы тела;
7. Числовой параметр наследственности диабета;
8. Возраст, лет;
9. Зависимая переменная (1 – наличие заболевания, 0 – отсутствие).

Построить уравнение регрессии.

3. Постройте Ридж-регрессию для файла Питер. Что говорит о мультиколлинеарности?

Вопросы.

1. Для чего нужна Логит-регрессия?
2. Что такое мультиколлинеарность?
3. Почему в случае результирующей бинарной переменной нельзя пользоваться обычной МНК-регрессией?
4. Какой VIF говорит о мультиколлинеарности?
5. Что может указать на мультиколлинеарность?
6. Что надо посмотреть, чтобы говорить о качестве Логит-регрессии?