Московский авиационный институт (национальный исследовательский университет)
Институт №3.
«Системы управления, информатика и электроэнергетика»
Кафедра №304
«Автоматизированные системы обработки информации и управления»
Отчет по Лабораторной работе № 2.
по учебной дисциплине
«Анализ и обработка данных»
на тему
«Множественный регрессионный анализ»

Группа М30-207Б
Выполнили:
Гордеев Н.М.
Макаров Е.

Принял:
Игнатов Н.А.

МОСКВА 2020

Содержание

Обозначение предметной области

Предметная область – Поддержанные мотоциклы марки SUZUKI, HONDA.

Зависимая переменная Y – цена мотоцикла.

Независимые переменные:

1) Год начала выпуска модели
2) Объем двигателя, куб. см
3) Максимальная мощность (лс/об/мин)
4) Масса мотоцикла кг
5) Пробег км
6) Длина мм
7) Ширина мм
8) Высота мм
9) Высота по седлу мм
10) Максимальная скорость км/ч
11) Максимальный крутящий момент Нм

Расчетные формулы:

Математическое ожидания и дисперсия ошибки

$$\widetilde{m}_E = \frac{\sum_{i=1}^{m}(\widetilde{y}_i - (\widetilde{a}_0 + \sum_{k=1}^{n}\widetilde{a}_k\widetilde{x}_{i,k}))}{m}, \widetilde{D}_E = \frac{\sum_{i=1}^{m}(\widetilde{y}_i - (\widetilde{a}_0 + \sum_{k=1}^{n}\widetilde{a}_k\widetilde{x}_{i,k}))^2}{m-n-1}.$$

Дисперсия адекватности

$$\widetilde{D}_{ad} = \frac{\sum_{l=1}^{L}(\widetilde{y}_{Nl} - (\widetilde{a}_0 + \sum_{k=1}^{n}\widetilde{a}_k\widetilde{x}_{Nl,k}))^2}{L}.$$

Коэффициент детерминации

$$R^2 = 1 - \frac{\widetilde{D}_E}{\widetilde{D}_Y}$$

Расчетные формулы для проверки гипотез

Статистическая гипотеза о равенстве нулю коэффициента детерминации ($H_0 : R^2 = 0$) проверяется с помощью критерия

$$Z = \frac{R^2}{1-R^2}\frac{(m-n-1)}{n},$$

подчиненного F-распределению с числом степеней свободы числителя $q_1 = n$ и числом степеней свободы знаменателя $q_2 = m-n-1$, критическая область – правосторонняя (конкурентная гипотеза $H_1 : R^2 > 0$).

$$\widetilde{P}_a = \widetilde{D}_E(\widetilde{X}^T\widetilde{X})^{-1}. \qquad \widetilde{\sigma}_{ak} = \sqrt{\widetilde{D}_{ak}}:$$

$$\widetilde{D}_{ak} = (\widetilde{D}_E(\widetilde{X}^T\widetilde{X})^{-1})_{k,k}. \qquad \widetilde{Z}_k = \frac{\widetilde{a}_k}{\widetilde{\sigma}_{ak}}$$

Для проверки статистической значимости коэффициента регрессии a_k используется статистическая гипотеза о равенстве данного коэффициента нулю (то есть, равенство нулю математического ожидания его оценки), H_0: $a_k = 0$ с конкурентной гипотезой H_1: $a_k <> 0$.

Статистический критерий

$$\widetilde{Z}_k = \frac{\widetilde{a}_k}{\widetilde{\sigma}_{ak}}$$

подчинен t-распределению с числом степеней свободы q = m – n – 1.

Для проверки адекватности регрессионной модели используется статистическая гипотеза о равенстве дисперсий ошибки в области построения и за ее пределами (то есть, о статистической незначимости их отличий): H_0: $D_{ad} = D_E$ с конкурентной гипотезой H_1: $D_{ad} > D_E$.

Статистический критерий

$$\widetilde{Z} = \frac{\widetilde{D}_{ad}}{\widetilde{D}_E}$$

подчинен F-распределению с числом степеней свободы числителя $q_1 = L$ и знаменателя $q_2 = m-n-1$.

Данные

Основная выборка

Название	GSF750 Bandit	VT400 Shadow	VT400 Shadow	PC800 Pacific Coast	VT750 Shadow Aero	VT1100 Shadow Sabre	C50 Boulevard	VT1100 Shadow Sabre	NC700SA	NC700SA	VTX1300C	M50 Boulevard	VT750 Shadow Spirit
Цена	195	197	219	219	308	330	348	359	381	385	396	407	425
Год	1997	1998	2002	1993	2004	2002	2006	2004	2012	2013	2006	2009	2012
Максимальная мощ	77	33	33	56,5	44	67	51,7	67	50	50	74,8	52	44
Максимальная скор	180	145	145	172	164	172	177	172	160	160	178	172	177
Макс крут момент	64	34	34	65,7	62	91	69	91	61	61	123	68	62
Объем двигателя	750	400	400	800	750	1100	800	1100	700	700	1300	800	750
Масса мотоцикла	201	242	242	280	245	251	277	251	215	215	106	247	243
Пробег	1622	14692	26758	29277	11472	26031	26887	15998	4935	14246	16120	6298	22765
Длина	2090	2455	2455	2290	2510	2380	2510	2380	2195	2195	2400	2370	2430
Ширина	770	945	945	820	940	880	970	880	760	760	920	920	835
Высота	1095	1050	1050	1370	1090	1220	1105	1220	1130	1130	1110	1125	1130
Высота по седлу	795	695	695	765	660	730	700	730	790	790	697	700	655

Название	F800R	NC750SA	VTX1800S	VT750 Shadow Phantom	R1200R	VT750 Shadow Phantom	R1200GS	GL1500 Valkyrie Interstate	CBR600F ABS	VTX1800F	VFR1200F ABS	CB1000R
Цена	425	436	439,5	452	458	496	496	512	528	568	594	616
Год	2013	2015	2002	2010	2007	2013	2004	2000	2011	2007	2010	2013
Максимальная мощность	66	54	106	45	81	45	105	100	102	106	111	125
Максимальная скорость	200	160	189	250	200	250	200	210	220	189	250	228
Макс крутящий момент	86	62	162,7	66	119	66	115	134	65	162,7	111	99
Объем двигателя	800	750	1800	750	1200	750	1200	1500	600	1800	1200	1000
Масса мотоцикла	202	216	359	184	231	184	223	309	211	351	268	217
Пробег	13362	24781	31987	18486	22247	17857	31009	12039	14187	33186	20384	16590
Длина	2145	2215	2630	2405	2145	2405	2210	2525	2150	2420	2250	2105
Ширина	860	775	990	825	872	825	935	980	685	940	755	785
Высота	1235	1350	1155	1090	1285	1090	1450	1185	1150	1130	1220	1095
Высота по седлу	790	790	695	650	800	650	850	740	800	701	810	825

Дополнительная выборка

Название	NV400 Shadow Slasher	VT400 Shadow	VT1100 Shadow Spirit	VT1100 Shadow Spirit	VFR800-2	VFR800	VTX1300C	VT750 Shadow Aero	VTX1300R	C50 Boulevard	VTX1800C	CB1100A ABS	CBR600F ABS
Цена	205	219	308	337	385	392	399	429	436	436	496	496	518
Год	2001	2003	2000	2005	2005	2009	2006	2013	2008	2013	2006	2010	2011
Максимальная мощ	33	33	52	52	106	51,7	74,8	44	74,8	51,7	106	89,7	102
Максимальная скор	145	145	172	172	230	177	178	164	178	177	189	220	220
Макс крут момент	34	34	91	91	69	69	123	62	123	69	162,7	91	65
Объем двигателя	400	400	1100	1100	800	800	1300	750	1300	800	1800	1100	600
Масса мотоцикла	196	223	255	255	213	292	296	241	305	292	340	248	211
Пробег	14387	16389	25186	12734	24907	15724	8560	9087	21380	5521	33736	19897	20987
Длина	2310	2455	2380	2380	2120	2510	2400	2510	2575	2510	2455	2200	2150
Ширина	795	945	880	880	750	970	920	920	960	970	930	830	685
Высота	1060	1105	1220	1220	1195	1105	1110	1125	1125	1105	1125	1130	1150
Высота по седлу	645	695	730	730	809	700	697	660	685	700	695	780	800

Название	VTX1800R	M109R Boulevard
Цена	572	634
Год	2004	2008
Максимальная мощ	106	125
Максимальная скор	189	206
Макс крут момент	162,7	160
Объем двигателя	1800	1800
Масса мотоцикла	359	315
Пробег	23802	28375
Длина	2630	2480
Ширина	990	875
Высота	1155	1185
Высота по седлу	695	705

Статистические характеристики

*Для расчетов статистических характеристик использовались встроенные формулы в Exel

Мат ожидание =СУММ(соответствующая строка из вектора наблюдений)/25

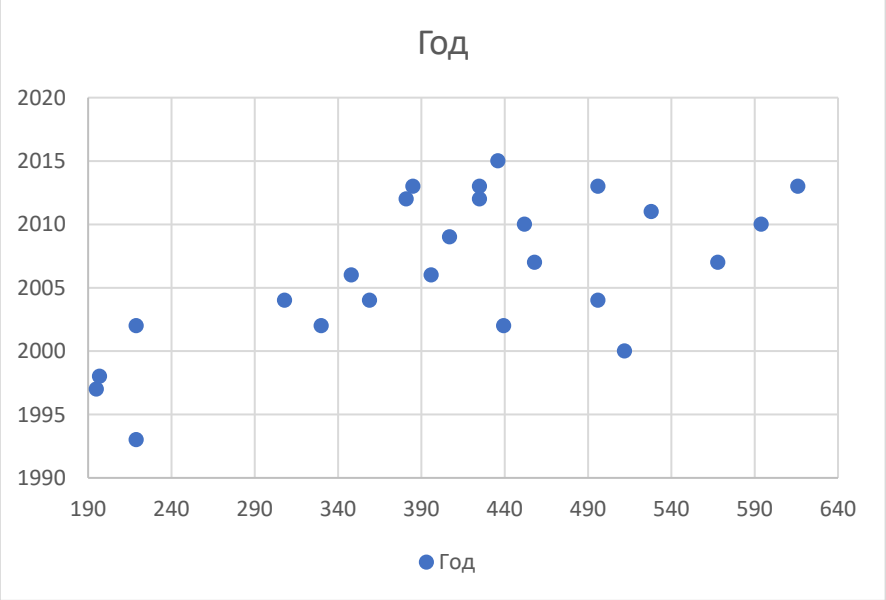
Дисперсия =ДИСП.В(соответствующая строка из вектора наблюдений)

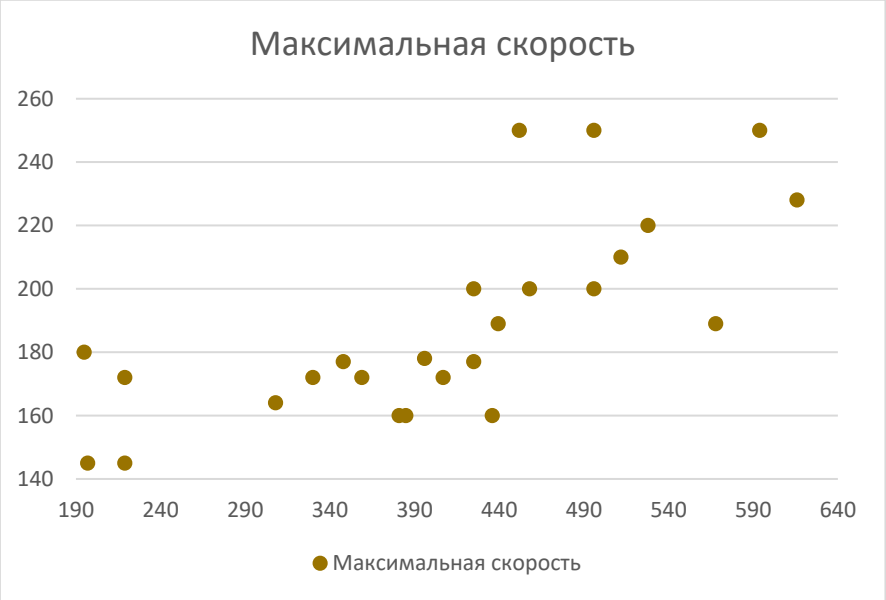
	Основная выборка		Дополнительная выборка	
	Мат ожидание	Дисперсия	Мат ожидание	Дисперсия
Цена	407,58	13990,41	417,47	14218,70
Год	2006,52	34,18	2006,80	15,89
Максимальная мощность	69,84	746,27	73,45	923,20
Максимальная скорость	188,80	963,08	184,13	652,98
Максимальный крутящий момент	85,36	1222,45	93,76	1888,82
Объем двигателя	948,00	136975,00	1056,67	224595,24
Масса мотоцикла	238,80	2752,17	269,40	2421,40
Пробег	18928,64	72440540,57	18711,47	62675986,27
Длина	2330,60	22752,75	2404,33	22985,24
Ширина	862,88	7145,28	886,67	7759,52
Высота	1170,40	10504,00	1141,00	2136,43
Высота по седлу	740,12	3656,03	715,07	2259,64

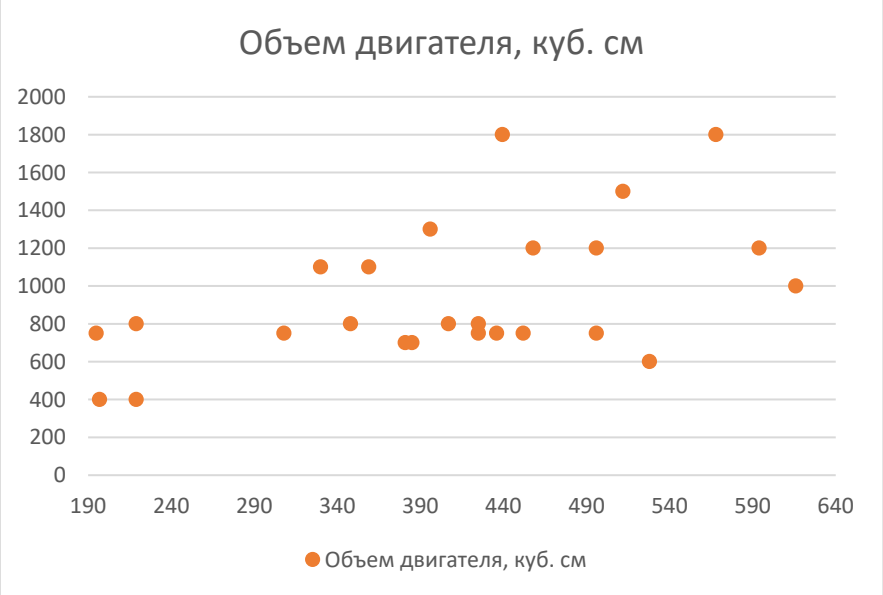
Коэффициент корреляции =КОРРЕЛ(соответствующая строка из вектора наблюдений; строка из вектора наблюдений, соответствующая столбцу матрицы ковариации)

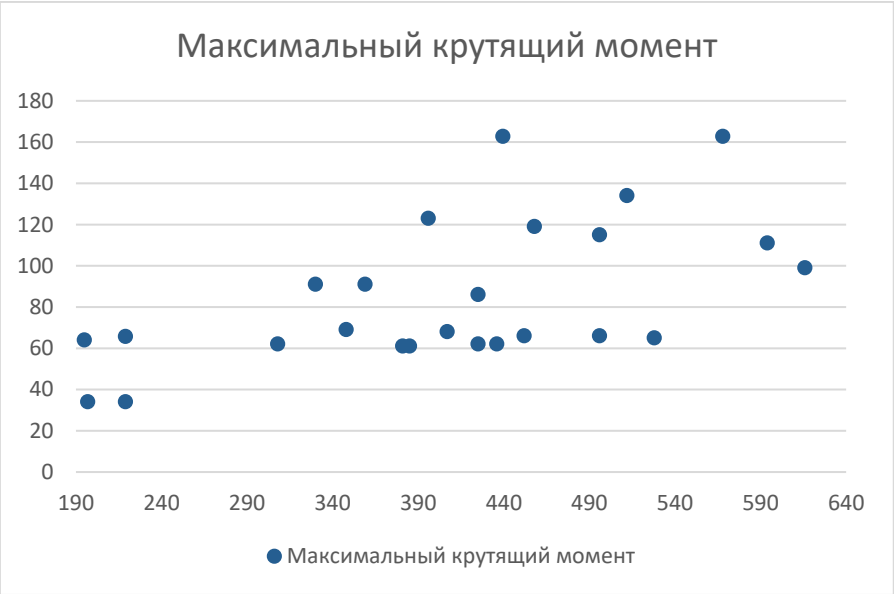
Матрица коэффициентов корреляции												
		Ц	Г	ММ	МКМ	ОД	ММ	П	Д	Ш	В	ВпС
Цена	1	0,63	0,69	0,72	0,58	0,51	0,09	0,16	-0,17	-0,20	0,16	0,25
Год	0,63	1	0,02	0,35	-0,06	-0,13	-0,35	-0,14	-0,33	-0,47	-0,09	0,11
Максимальная мощность	0,69	0,02	1	0,50	0,78	0,71	0,30	0,20	-0,25	-0,07	0,29	0,54
Максимальная скорость	0,72	0,35	0,50	1	0,33	0,27	-0,11	0,02	-0,17	-0,30	0,04	0,10
Максимальный крутящий момент	0,58	-0,06	0,78	0,33	1	0,98	0,44	0,39	0,16	0,35	0,27	0,14
Объем двигателя	0,51	-0,13	0,71	0,27	0,98	1	0,51	0,41	0,26	0,40	0,26	0,05
Масса мотоцикла	0,09	-0,35	0,30	-0,11	0,44	0,51	1	0,46	0,46	0,43	0,10	-0,10
Пробег	0,16	-0,14	0,20	0,02	0,39	0,41	0,46	1	0,32	0,35	0,41	-0,09
Длина	-0,17	-0,33	-0,25	-0,17	0,16	0,26	0,46	0,32	1	0,76	-0,35	-0,82
Ширина	-0,20	-0,47	-0,07	-0,30	0,35	0,40	0,43	0,35	0,76	1	-0,08	-0,49
Высота	0,16	-0,09	0,29	0,04	0,27	0,26	0,10	0,41	-0,35	-0,08	1	0,58
Высота по седлу	0,25	0,11	0,54	0,10	0,14	0,05	-0,10	-0,09	-0,82	-0,49	0,58	1

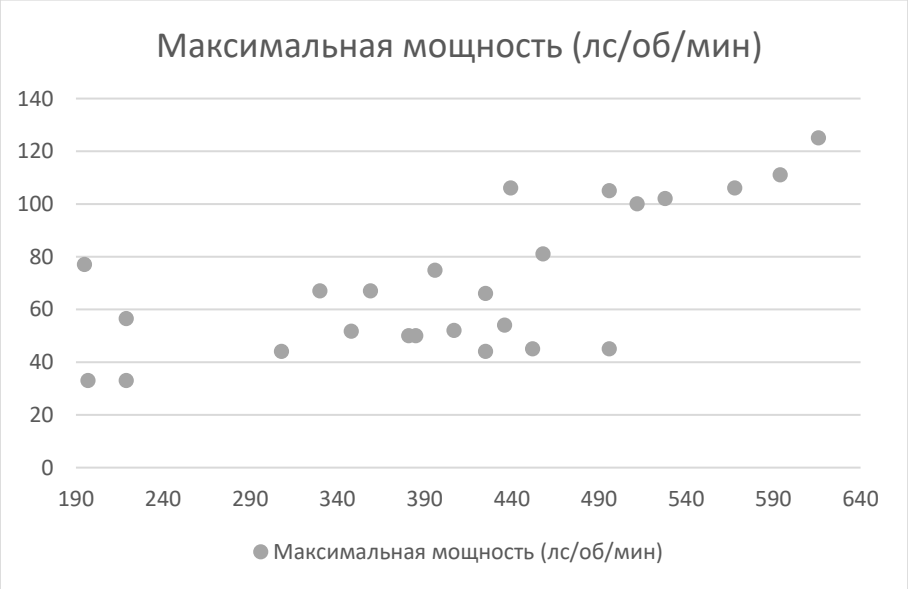
Двумерные диаграммы рассеяния каждой из независимых и зависимой переменных

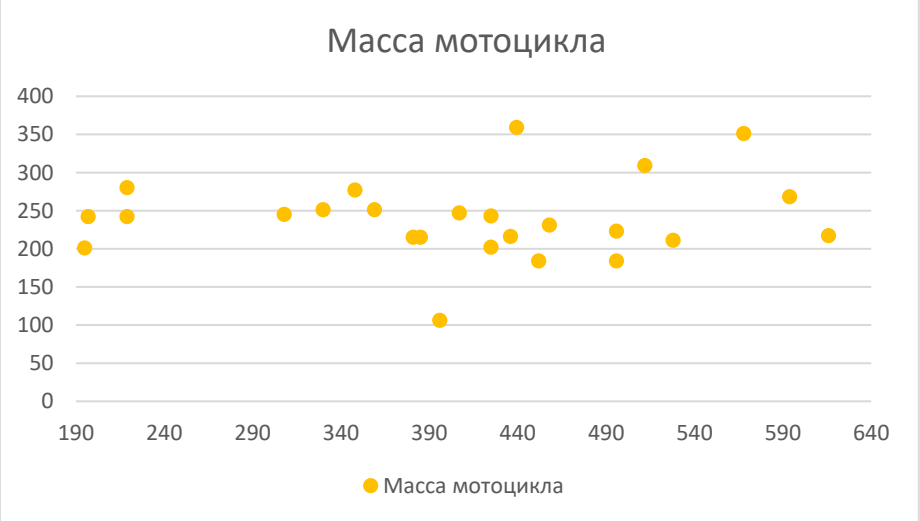


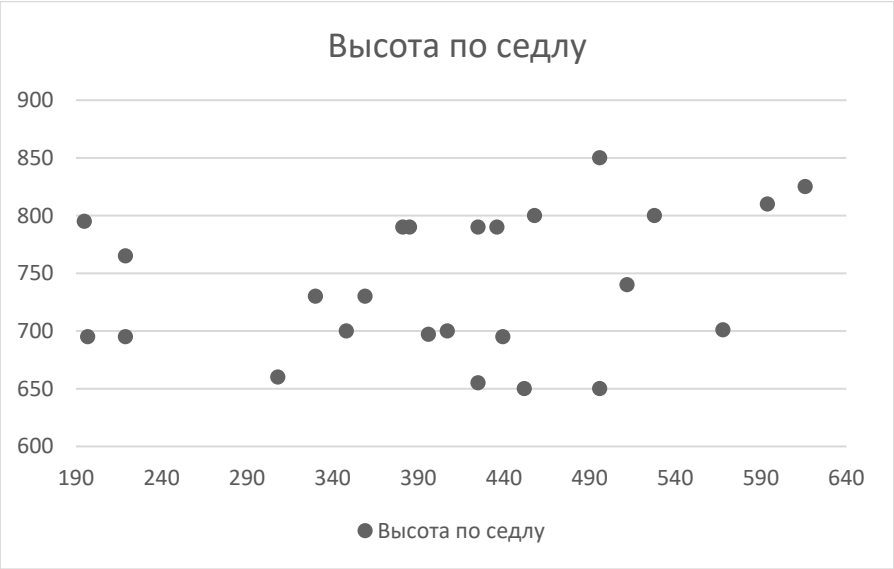




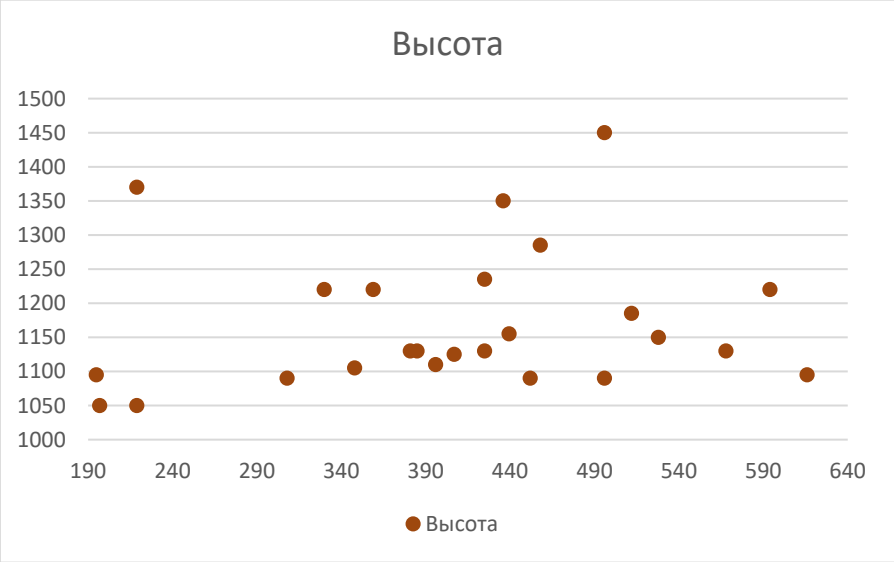

Максимальный крутящий момент
180
160
140
120
100
80
60
40
20
0
190
240
290
340
390
440
490
540
590
640
Максимальный крутящий момент

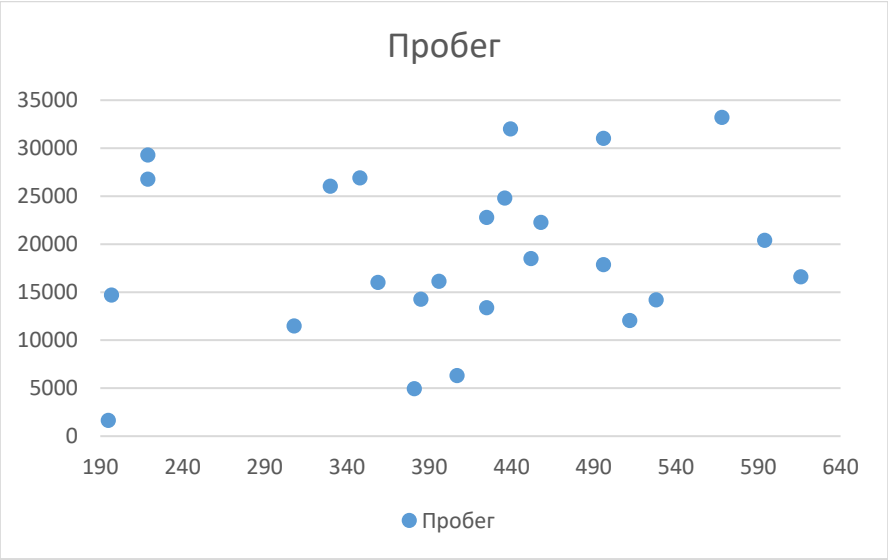

Максимальная мощность (лс/об/мин)
140
120
100
80
60
40
20
0
190
240
290
340
390
440
490
540
590
640
Максимальная мощность (лс/об/мин)

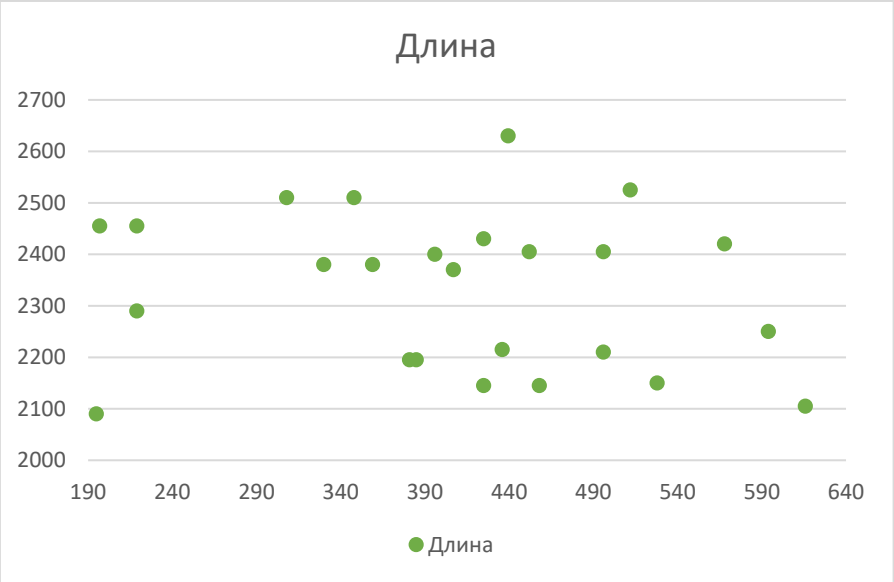

Масса мотоцикла
400
350
300
250
200
150
100
50
0
190
240
290
340
390
440
490
540
590
640
Масса мотоцикла


Высота по седлу
900
850
800
750
700
650
600
190
240
290
340
390
440
490
540
590
640
Высота по седлу


Ширина
1050
1000
950
900
850
800
750
190
240
290
340
390
440
490
540
590
640
Ширина


Высота
1500
1450
1400
1350
1300
1250
1200
1150
1100
1050
1000
190
240
290
340
390
440
490
540
590
640
Высота





Допущения о характере связи

На основе построенных диаграмм сделали допущения о линейном характере связей между независимыми и зависимой переменными и предложили уравнение множественной регрессии в виде $Y = a_0 + \sum_{i=1}^{11} a_i * x_i$

Коэффициенты уравнения множественной регрессии

"Коэффициенты уравнения регресии"							
Индекс	Все 11 переменных	Зелёные и желтые	Зелёные	Первые 4 переменные	Зелёные и желтые без Объ. Двиг	Зелёные и желтые без Объ. Двиг	Первые 3 + объём двигателя
a0	-27031,4084	-26799,58917	-27072,68785	-23100,22211	-22010,49645	-26796,4634	-26434,7407
a1	13,2834881	13,1668905	13,18696834	11,51578152	10,98330645	13,16519375	13,26441306
a2	2,494838931	2,442364132	2,275682081	1,629673621	1,383864332	2,439335144	2,244029392
a3	1,064038286	1,039743503	1,116516664	1,121529194	1,08041264	1,040235947	-0,351492955
a4	-0,45463986	0,1966559	0,176262792	-0,953468465	0,928042148	0,196522406	0,105793704
a5	0,038708011	-0,000628345	-0,000500397	0,165598035		-0,000627387	
a6	0,183373602	0,096753312	0,150606294			0,09678326	
a7	-0,00056703	0,106961291	0,105040284			0,106769922	
a8	0,085503542	0,181124253	0,150683752			0,180818699	
a9	0,131655433	-0,000210627				-0,193641213	
a10	0,174975408	-0,194538887					
a11	-0,18691091						

*Для подсчета коэффициентов уравнений воспользовались встроенной в Exel функцией

(a_0…a_n)=ЛИНЕЙН (Вектор цены; Массив наблюдений независимых переменных, по которым строится модель)

Год
Максимальная мощность
Максимальная скорость
Максимальный крутящий момент
Объем двигателя, куб. см
Масса мотоцикла
Пробег
Длина
Ширина
Высота
Высота по седлу

	Остаточная дисперсия	Коэффициент Детерминации	Коэф. Множ. Корреляции
Все 11 переменных	1079,52	0,922838483	0,960644827
Первые 5 переменных	1199,34	0,914274344	0,956176942
Первые 4 переменные	1231,11	0,912003476	0,954988731
Зелёные и жолтые	1006,28	0,928073476	0,963365702
Зелёные	880,50	0,937064178	0,968020753
Зелёные и желтые - объём дв	939,20	0,932868578	0,965851219
Первые 3 + объём двигателя	2043,05	0,853967473	0,924103605

Остаточная дисперсия =СУММПРОИЗВ(f;f)/(25-5-1), где f это вектор из f_i=СУММ(-СУММПРОИЗВ(независимые переменные i-го наблюдения (по которым строится модель); коэффициенты уравнения a_1…a_n); - коэффициент ур a_0; цена)

Коэф. Множ. Корреляции = (Коэффициент Детерминации)^(1/2)

Проверка статистической значимости коэффициентов уравнения множественной регрессии по t-критерию Стьюдента

Оценка матрицы ковариаций Pa коэффициентов уравнения линейной регрессии											Z	Достигнутый уровень значимости
0,05	0,08	-0,03	-0,05	0,00	0,01	0,00	-0,03	0,01	0,01	-0,08	244,992	0,00000
0,08	0,64	-0,15	-0,35	0,00	-0,02	0,00	-0,03	0,02	0,05	-0,22	3,87531	0,00191
-0,03	-0,15	0,13	-0,24	0,02	0,00	0,00	0,00	0,02	-0,02	0,07	8,01735	0,00000
-0,05	-0,35	-0,24	3,68	-0,32	0,12	0,00	0,11	-0,18	0,07	-0,05	-0,12361	0,90351
0,00	0,00	0,02	-0,32	0,03	-0,01	0,00	-0,01	0,01	-0,01	0,01	1,30019	0,21611
0,01	-0,02	0,00	0,12	-0,01	0,04	0,00	-0,01	0,00	0,00	-0,02	4,55683	0,00054
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-421,321	0,00000
-0,03	-0,03	0,00	0,11	-0,01	-0,01	0,00	0,02	-0,02	0,00	0,03	3,95669	0,00164
0,01	0,02	0,02	-0,18	0,01	0,00	0,00	-0,02	0,03	0,00	-0,01	4,41897	0,00069
0,01	0,05	-0,02	0,07	-0,01	0,00	0,00	0,00	0,00	0,02	-0,03	11,0794	0,00000
-0,08	-0,22	0,07	-0,05	0,01	-0,02	0,00	0,03	-0,01	-0,03	0,16	-1,18156	0,25855

$$\widetilde{P}_a = \widetilde{D}_E(\widetilde{X}^T\widetilde{X})^{-1}. \qquad \widetilde{\sigma}_{ak} = \sqrt{\widetilde{D}_{ak}}:$$

$$\widetilde{D}_{ak} = (\widetilde{D}_E(\widetilde{X}^T\widetilde{X})^{-1})_{k,k}. \qquad \widetilde{Z}_k = \frac{\widetilde{a}_k}{\widetilde{\sigma}_{ak}}$$

При проверке данной гипотезы малый достигнутый уровень значимости p означает малую вероятность (случайно) получить достигнутое или большее (по модулю) значение меры Z , значит, списать на случайность (влияние разброса) полученное значение $a_к$ нельзя, значит, гипотеза H_0 отклоняется, и коэффициент уравнения регрессии a_k значим (выделен зелёным цветом).

отличные от малого значение достигнутого уровня значимости p означает большую вероятность (случайно) получить достигнутое или большее (по модулю) значение меры Z , значит, полученное значение a_k, можно списать на случайность (влияние разброса) при $a_k = 0$, гипотеза H_0 принимается, и коэффициент уравнения регрессии a_k незначим (выделен красным цветом).

Желтым выделены коэффициенты, по которым сложно судить о статистической значимости.

Оценка дисперсии адекватности и проверка гипотезы об адекватности регрессионной модели

	М. О. Ошибки	Дисперсия адекватности	Дис. Ошибки / дис. адекватности	Достигнутый уровень значимости
Все 11 переменных	-12,96	1196,29	1,11	0,43
Первые 5 переменных	-3,97	1359,52	1,13	0,39
Первые 4 переменные	-0,93	1235,82	1,00	0,49
Зелёные и желтые	-11,90	1062,85	1,06	0,46
Зелёные	-10,29	1126,88	1,28	0,31
Зелёные и желтые - объём дв	-11,89	1062,12	1,13	0,41
Первые 3 + объём двигателя	2,61	1303,73	0,64	0,81

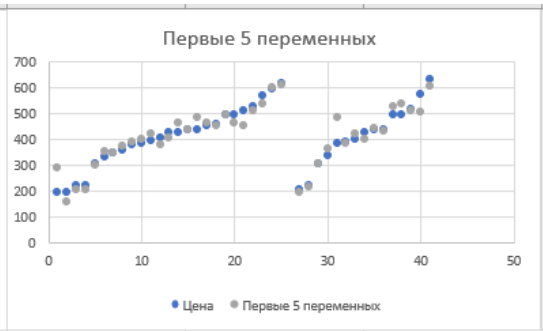
М. О. Ошибки =СУММ(f)/15

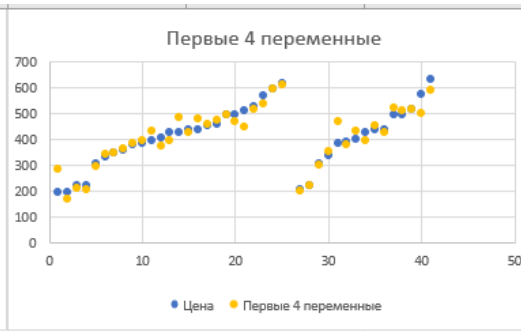
Дисперсия адекватности =СУММПРОИЗВ(f;f)/15

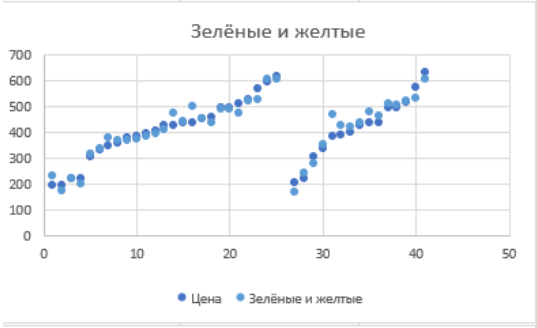
Где f – вектор, составленный из f_i=СУММ(-СУММПРОИЗВ(независимые переменные i-го наблюдения (по которым строится модель); коэффициенты уравнения $a_1 \ldots a_n$); - коэффициент ур a_0; цена)

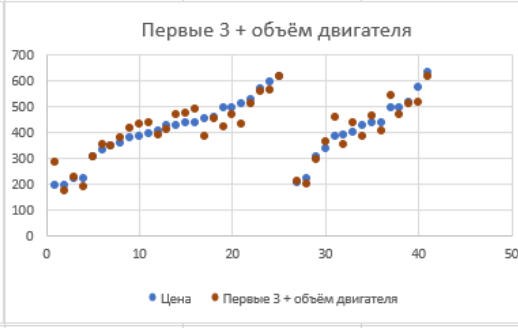
Пример расчета достигнутого уровня значимости: =F.РАСП.ПХ(Дис. адекватности / дис. Ошибки;15;25-5-1)
Где F.РАСП.ПХ – правостороннее F-распределение, (считает интеграл по плотности от 1 параметра до +беск)
Дис. адекватности / дис. Ошибки – x, от которого считается интеграл
15;25-5-1 – 2 параметра F распределения
F.РАСП.ПХ(1196,29/1079,52;15;19) = 0,43

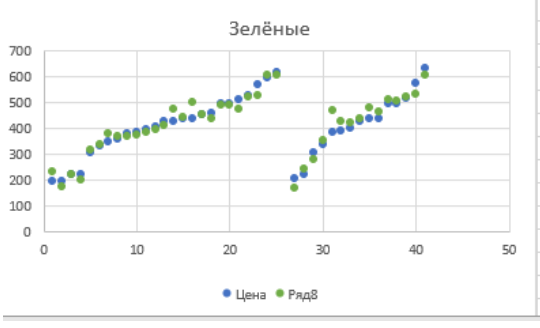
Графики, на которых исходная и дополнительная, (как продолжение) и значения регрессионных моделей











Сводная таблица

Кол-во факторов	Дисперсия ошибки	Коэффициент Детерминации	М. О. Ошибки (доп. Выборка)	Дисперсия адекватности	Дис. Ошибки / дис. адекватности	Достигнутый уровень значимости
11	1079,52	0,922838483	-12,96	1196,29	1,11	0,43
5	1199,34	0,914274344	-3,97	1359,52	1,13	0,39
4	1231,11	0,912003476	-0,93	1235,82	1,00	0,49
10	1006,28	0,928073476	-11,90	1062,85	1,06	0,46
8	880,50	0,937064178	-10,29	1126,88	1,28	0,31
9	939,20	0,932868578	-11,89	1062,12	1,13	0,41
4	2043,05	0,853967473	2,61	1303,73	0,64	0,81

Вывод

1) Как не посмотри, а пробег мотоцикла должен влиять на его цену, однако по матрице ковариации видно, что цена почти не зависит от пробега (коэф. Ков. = 0,16). Следовательно, делаем вывод, что либо продавцы на сайте https://www.drivebike.ru/ подкручивают пробег, либо снижают цену за изношенность исходя из степени наглости, а не показателей пробега.

2) Из той же матрице ковариации видно, что наибольшее влияние на цены мотоциклов оказывают: год выпуска модели, максимальная мощность, максимальная скорость, максимальный крутящий момент, объем двигателя.

3) По двумерным диаграммам рассеяния и матрице ковариации видно, что максимальный крутящий момент и объем двигателя связаны линейной зависимостью.

4) Все 7 моделей обладают отличной объясняющей способностью, что видно по коэф. детерминации.

5) Все 7 моделей показали отличное от малого значение достигнутого уровня значимости p означающее большую вероятность (случайно) получить достигнутое или большее значение меры Z, значит, полученное отличие отношения оценок дисперсий Z от 1 можно списать на случайность (влияние разброса) при $D_{ад} = D_E$, гипотеза H_0 принимается, что говорит об адекватности модели.

6) Так как все модели адекватны, и обладают высокой объясняющей способностью, наиболее “удобной” моделью является последняя, составленная из: года выпуска модели, максимальной мощности, максимальной скорости и объем двигателя.