


Линейные модели (часть 1)

Outline

- $\text{lm}(y \sim x)$
 - Модели с множеством предикторов
 - $\text{lm}(y \sim x_1 + x_2 + x_3)$
 - Если x – не число, а фактор?
 - Взаимодействия факторов
 - ANOVA
 - Предсказания по модели и валидация моделей:
 - Предсказание y по $x_1 \dots x_n$, зная модель
 - Сравнение и валидация моделей, переобученность
 - Обобщенные линейные модели (glm)
 - Нелинейный метод наименьших квадратов (не lm/glm!)
 - Подбор параметров в формуле по экспериментальным данным
- 
- сегодня

Корреляция

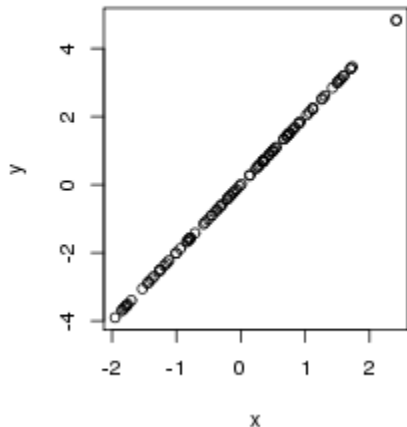
- Измеряет линейную зависимость между переменными
- Не означает причинно-следственной связи

> cor(x, y)

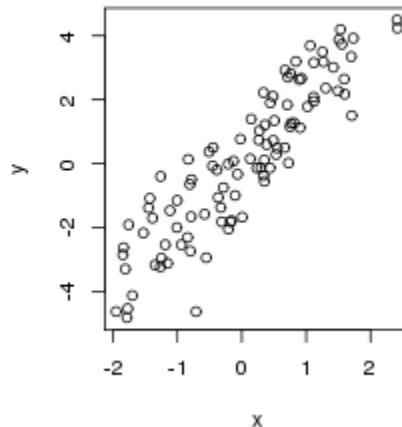
> cor(x, y, method="spearman") #ранговая корреляция

> cor(m)

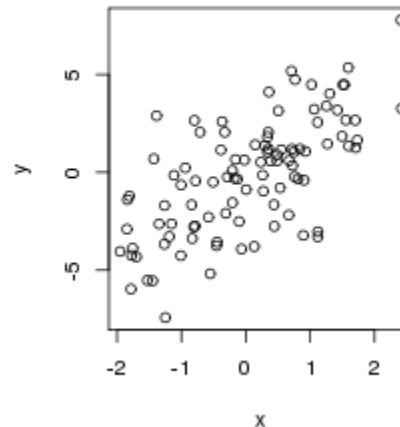
cor= 1



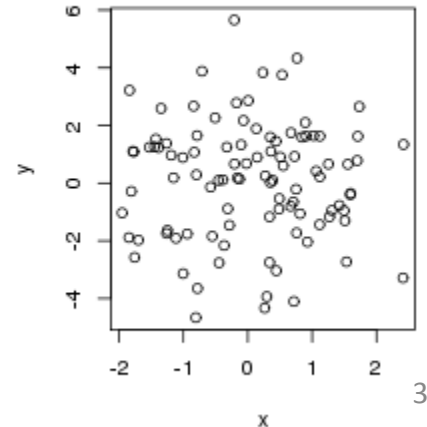
cor= 0.9



cor= 0.663



cor= -0.0161



Dataset

- Цены на ноутбуки (реальный прайс-лист)
- Что определяет цену, как она зависит от разных параметров?
- Как предсказать цену, зная параметры ноутбука?

```
> laptop=read.csv("laptop_price.csv")  
> head(laptop)
```

	Manufacturer	Model	Processor	Memory_Gb	HDD_Gb	HDD_type	Price_RUR
1	Acer	Aspire	i3-3110M	4	500	HDD	16400
2	Acer	Aspire	i3-3120M	4	500	HDD	16500
3	Acer	Aspire	i5-3230M	4	500	HDD	18500
4	Acer	Aspire	C-70	2	500	HDD	12000
5	Acer	Aspire	C-70	2	500	HDD	12000
6	Acer	Aspire	1007U	2	500	HDD	11300
7	Acer	Aspire	i5-2467M	4	240	SSD	33800

	Screen_size_inch	Battery_capacity_mAh	OS	color
1	15.6	4400	win8	black
2	15.6	4400	win8	black
3	15.6	4400	win8	black
4	11.6	2500	win8	turquoise
5	11.6	2500	win8	black
6	11.6	5000	win8	turquoise
7	13.3	3260	win7HP	silver

Корреляция

Если на входе – матрица, cor вычисляет корреляции между всеми колонками матрицы

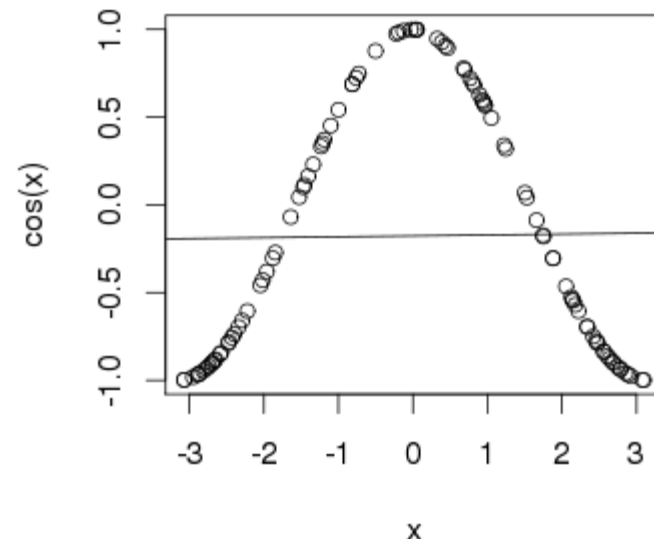
```
> m=as.matrix(laptop[,c("Memory_Gb", "HDD_Gb",  
  "Screen_size_inch", "Battery_capacity_mAh")])  
> cor(m)
```

	Memory_Gb	HDD_Gb	Screen_size_inch	Battery_capacity_mAh
Memory_Gb	1.000	0.6741	0.5259	0.2282
HDD_Gb	0.674	1.0000	0.5156	0.0568
Screen_size_inch	0.526	0.5156	1.0000	-0.0329
Battery_capacity_mAh	0.228	0.0568	-0.0329	1.0000

Корреляция

- Измеряет линейную зависимость между переменными
- Не означает причинно-следственной связи
- Отсутствие корреляции не означает независимость

```
> x=runif( 100, 0-pi, pi )  
> plot( x, cos(x) )  
> abline( lm( cos(x)~x ) )  
> cor( cos(x), x )  
[1] 0.0145
```



Корреляция

```
> cor.test(laptop$Price_RUR, laptop$Memory_Gb)
```

Pearson's product-moment correlation

data: laptop\$Price_RUR and laptop\$Memory_Gb

t = 13.3, df = 304, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.532 0.674

sample estimates:

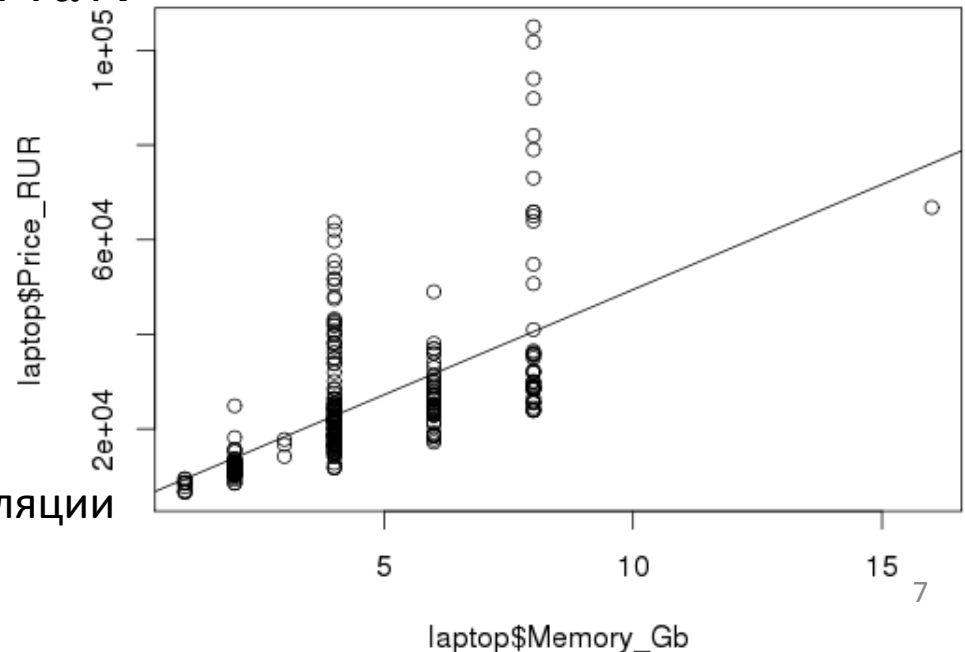
cor

0.608

```
> 0.608^2
```

```
[1] 0.37
```

R2 равен квадрату коэффициента корреляции



Регрессия

Основная идея: наблюдаемые значения
зависимой переменной – измерения,
которые содержат шум

$$y = f(x, b) + e$$

b_i – параметры модели

x_i – предикторы (независимые переменные)

e – ошибка (все, что мы не можем измерить
и учесть в модели)

Мат. ожидание $E[e] = 0$

Линейная регрессия

$$y = f(x, b) + e$$

$$f(x, b) = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

b_i – параметры модели

x_i – предикторы (независимые переменные)

e – ошибка (все, что мы не можем измерить и учесть в модели)

e распределено нормально!

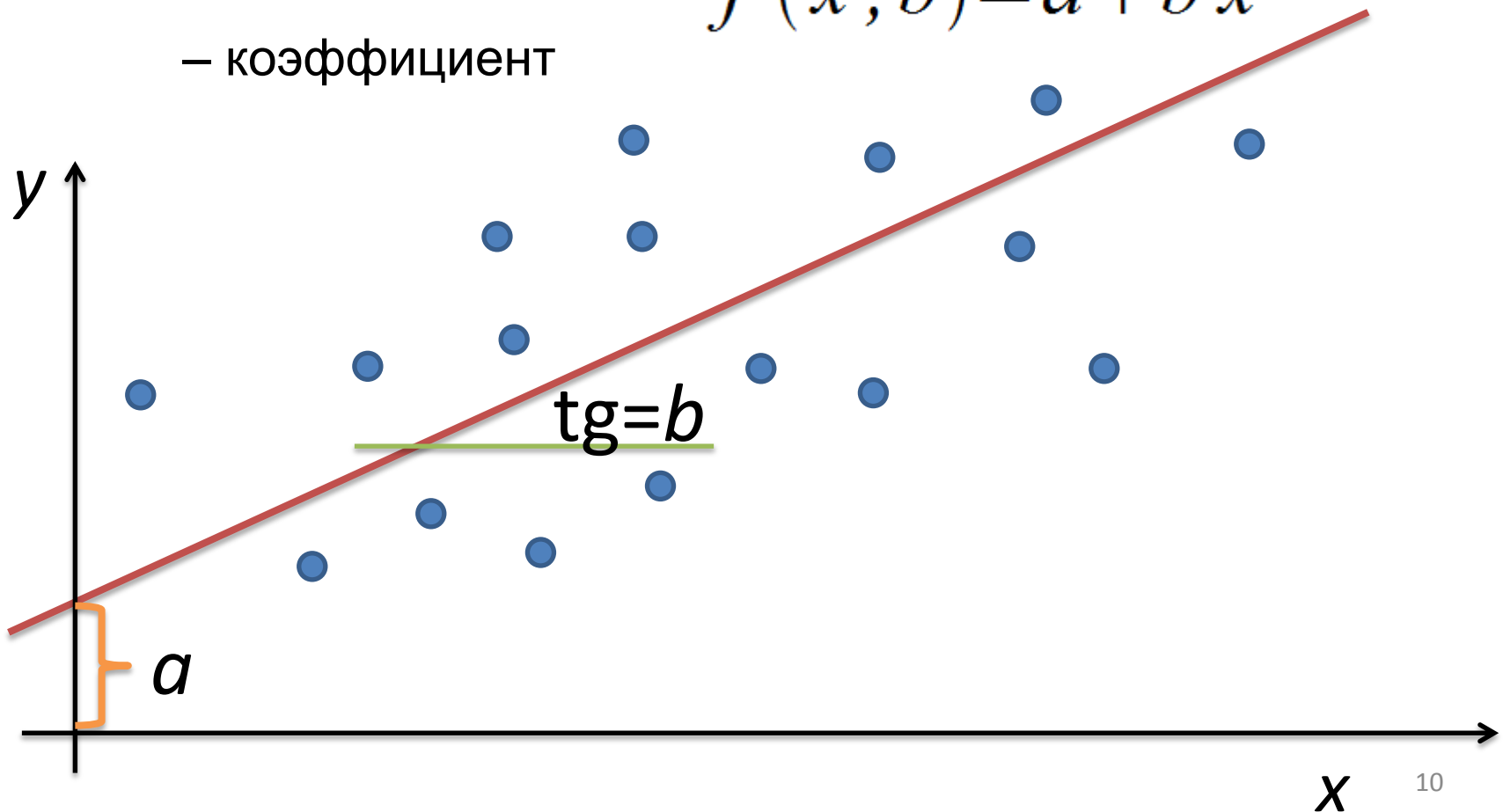
Линейная регрессия с 1 переменной

В случае одной независимой переменной

– константа

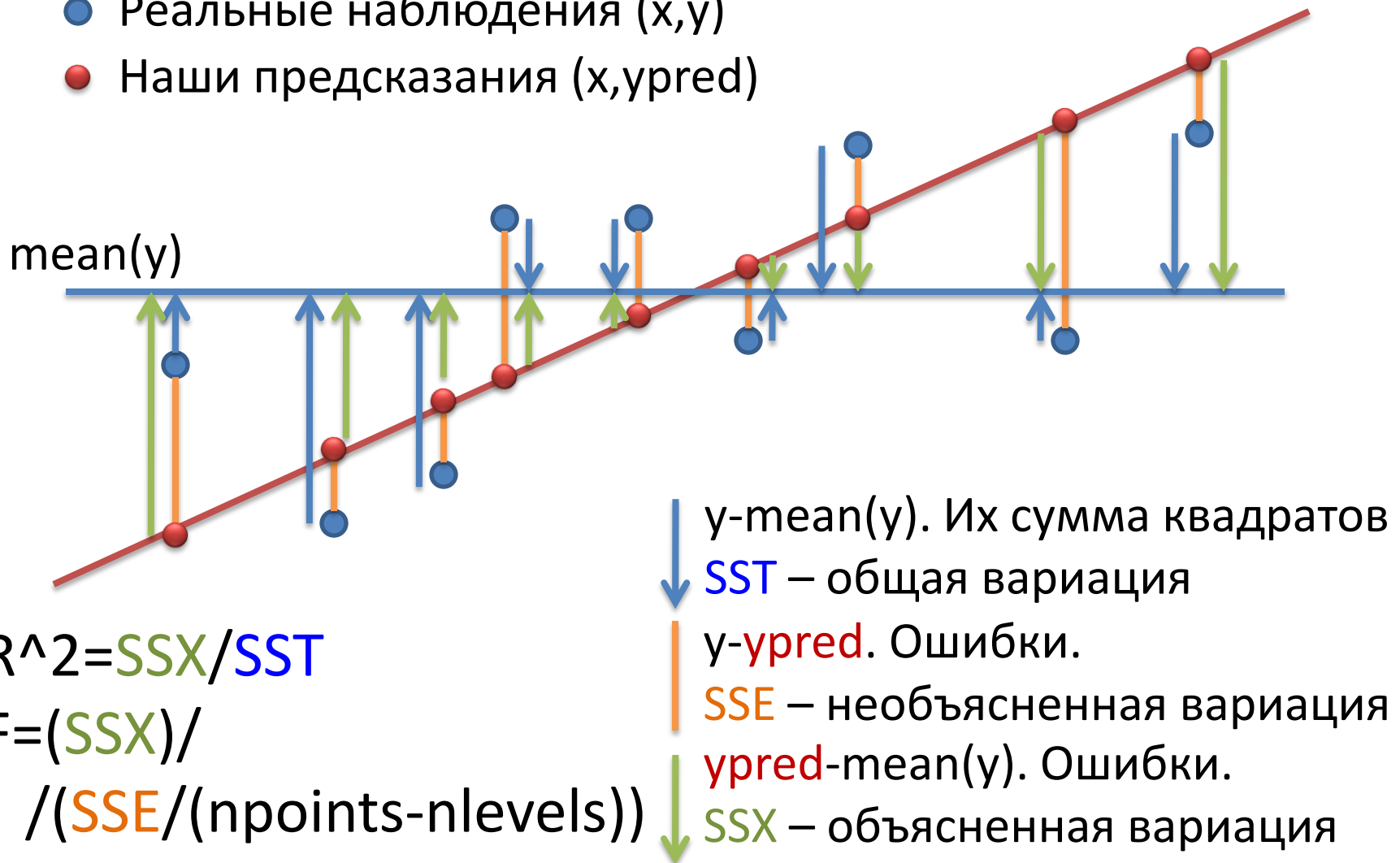
– коэффициент

$$f(x, b) = a + b x$$



$$SST = SSX + SSE$$

- Реальные наблюдения (x,y)
- Наши предсказания (x,ypred)

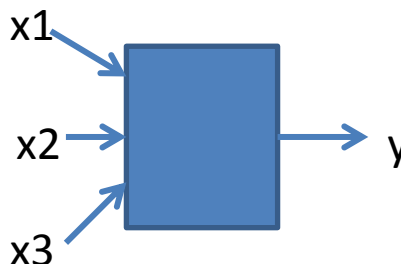


- $R^2 = SSX / SST$
- $F = (SSX) /$
- $/(SSE / (n_{\text{points}} - n_{\text{levels}}))$

Для чего нужны линейные модели?

Входные данные

y	x1	x2	x3



Значимость каждой переменной:

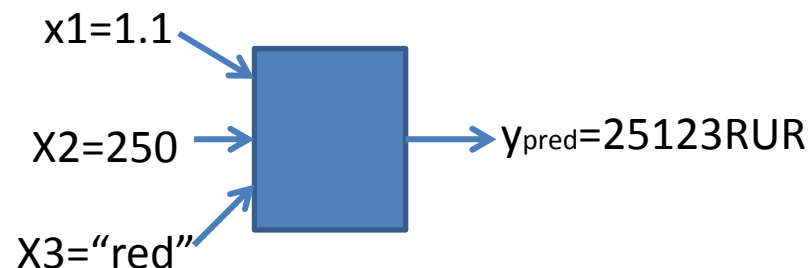
Предсказание y по x

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	9503.178	1733.057	5.483	9.10e-08	***
Memory_Gb	6232.204	421.098	14.800	< 2e-16	***
HDD_Gb	-26.604	3.275	-8.123	1.34e-14	***
Colorblue	-2496.492	4012.744	-0.622	0.5343	
...					
Colorred	1685.698	2736.167	0.616	0.5383	
Colorsilver	8617.956	1679.963	5.130	5.33e-07	***
...					

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10330 on 289 degrees of freedom
Multiple R-squared: 0.578, Adjusted R-squared: 0.5547
F-statistic: 24.74 on 16 and 289 DF, p-value: < 2.2e-16

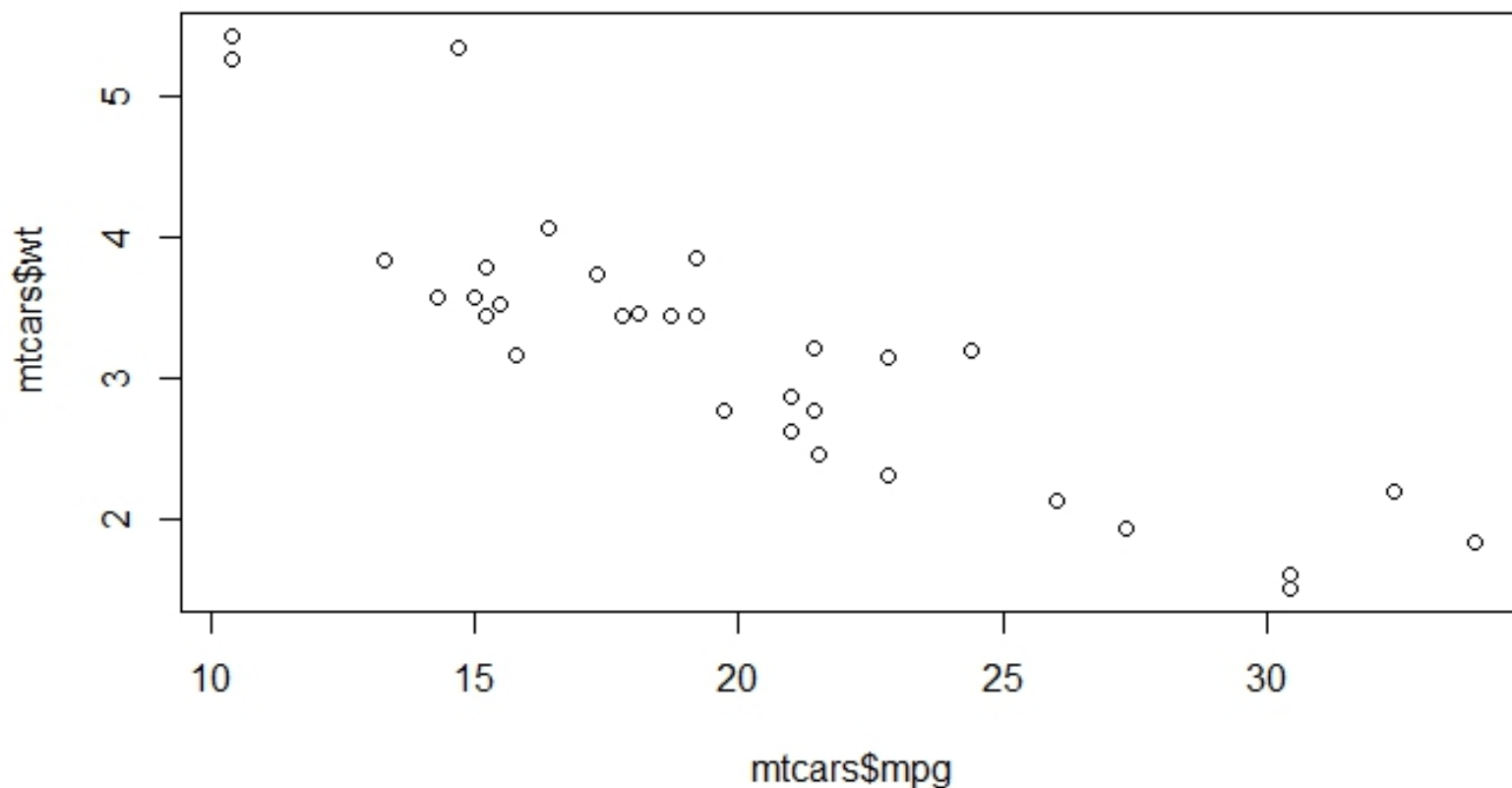


Линейная регрессия. Допущения

- **Подходящий тип данных** – зависимая переменная непрерывна, независимая непрерывна или категориальна
- **Независимость** – Каждое значение зависимой переменной не зависит от остальных ее значений
- **Линейность** – между зависимой и независимой переменными линейная зависимость (проверяем графически)
- **Распределение** – переменные распределены приблизительно нормально и не имеют выбросов
- **Гомоскедастичность** – ошибки предсказания постоянны на всем промежутке данных
- **Независимость и нормальность ошибок**
- **Мультиколлинеарность (для множественной регрессии)** – предикторы не должны сильно коррелировать

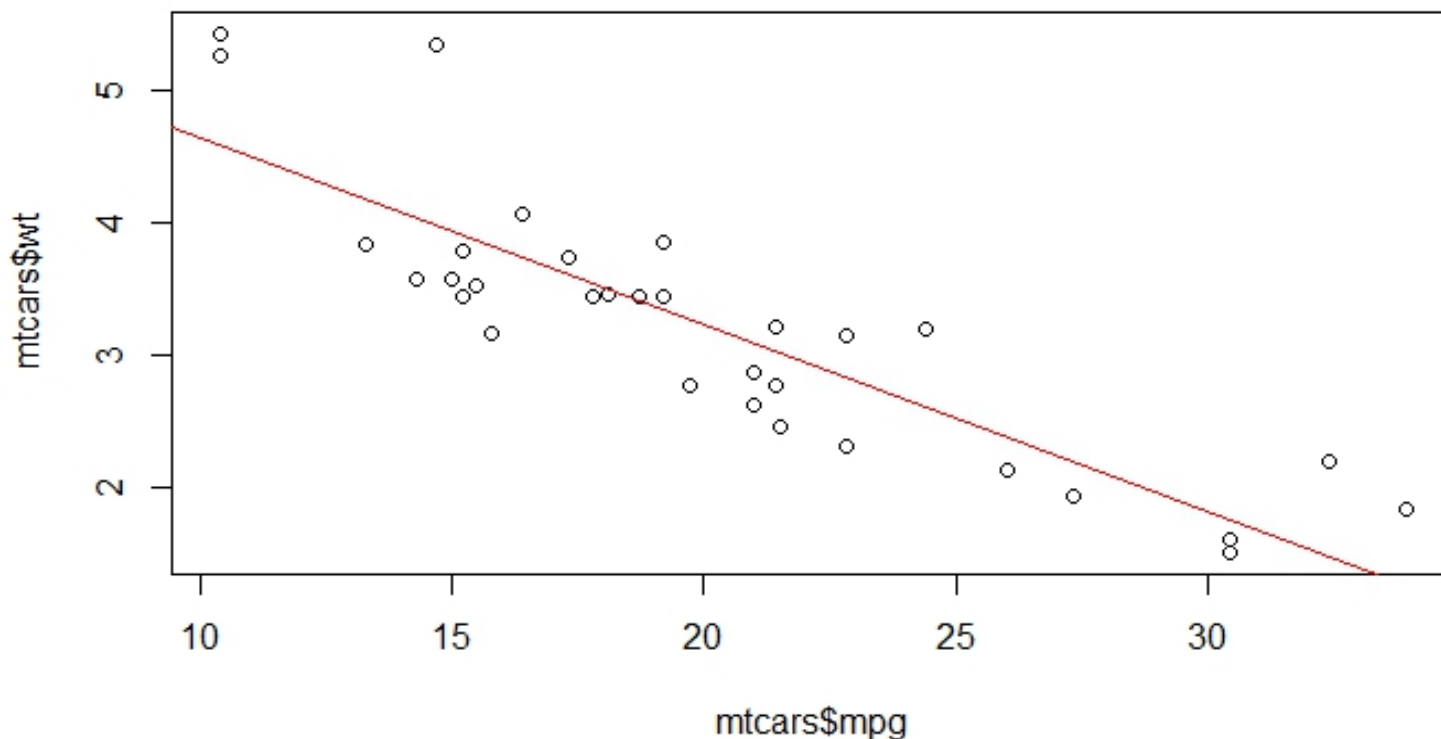
Модель с 1 предиктором. Пример 1

```
> plot(mtcars$mpg, mtcars$wt)
```



Модель с 1 предиктором. Пример 1

```
> plot(mtcars$mpg, mtcars$wt)  
> lm1<-lm(mtcars$wt~mtcars$mpg)  
> abline(lm1, col='red')
```



Модель с 1 предиктором. Пример 1

```
> lm1<-lm(mtcars$wt~mtcars$mpg)
```

```
> lm1
```

Call:

```
lm(formula = mtcars$wt ~ mtcars$mpg)
```

Coefficients:

(Intercept)	mtcars\$mpg
-------------	-------------

6.0473

-0.1409

коэффициент

константа

Модель с 1 предиктором. Пример 1

Коэффициенты:

```
> lm1$coefficients
```

```
(Intercept)    mtcars$mpg  
    6.047255    -0.140862
```

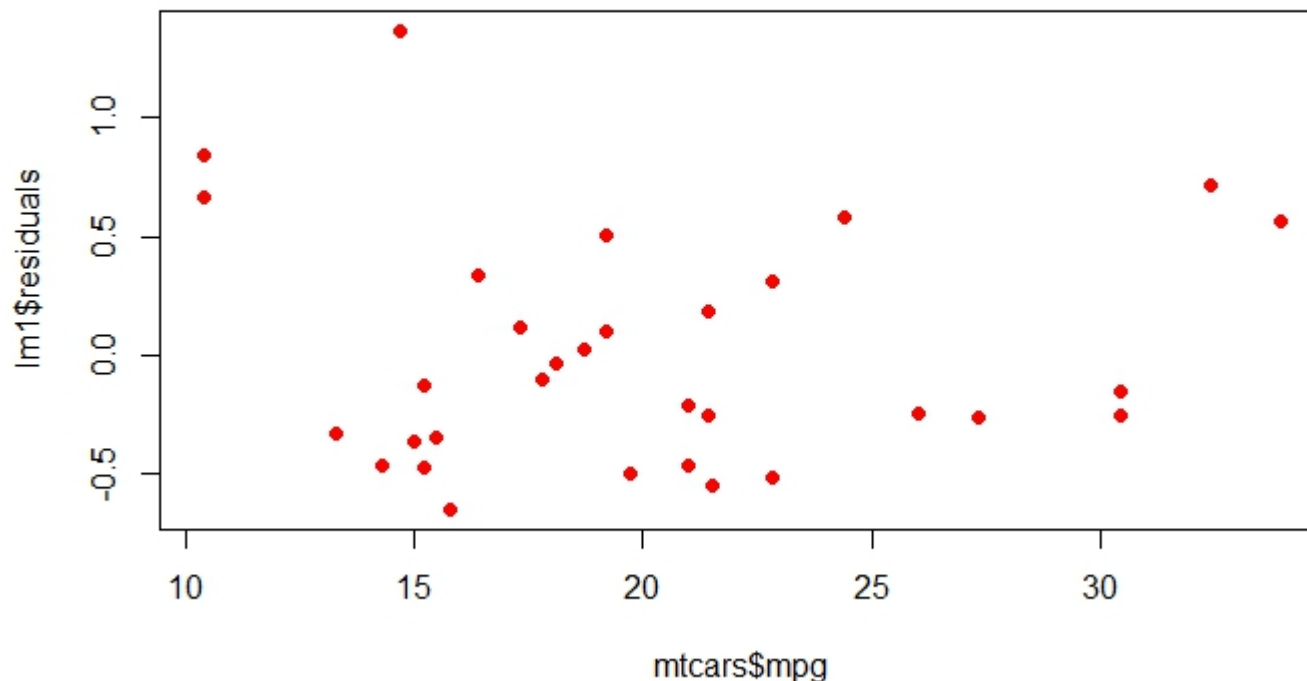
```
> lm1$coefficients[1]
```

```
(Intercept)  
    6.047255
```

Гомоскедастичность остатков

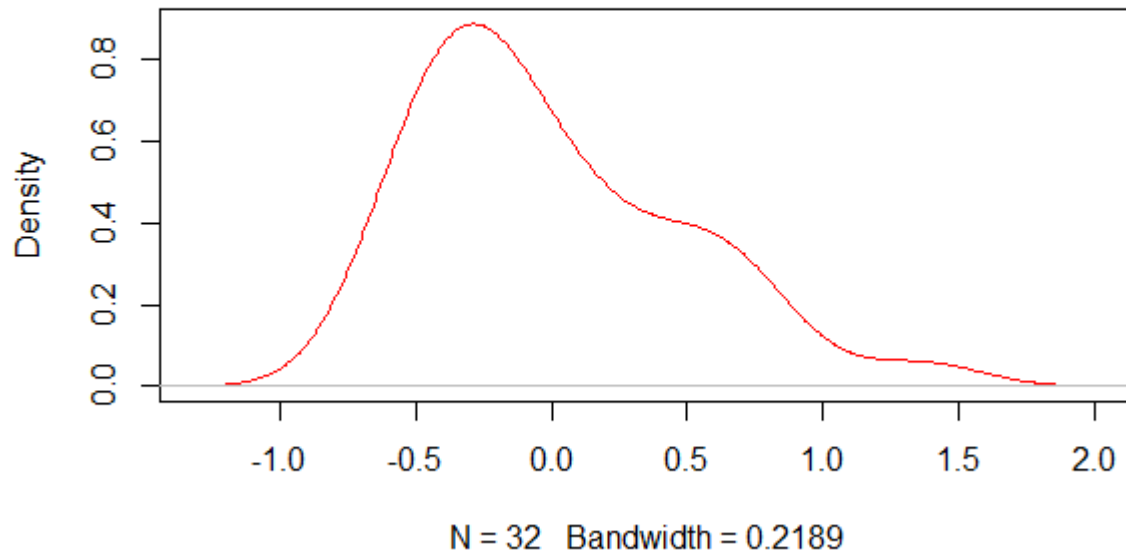
Остатки = y – предсказанные y

```
> plot(mtcars$mpg, lm1$residuals, pch=19, col='red')
```



Распределение остатков

```
>plot(density(lm1$residuals), col='red', main="")
```



Оценка качества линейной регрессионной модели

```
> summary(lm1)
```

```
Call:
```

```
lm(formula = mtcars$wt ~ mtcars$mpg)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-0.6516	-0.3490	-0.1381	0.3190	1.3684

Квантили для остатков

(остаток=отклонение наблюдаемого значения от модели)

В идеале должны быть симметричны относительно 0

Оценка качества линейной регрессионной модели

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.04726	0.30869	19.590	< 2e-16	***
mtcars\$mpg	-0.14086	0.01474	-9.559	1.29e-10	***

оценка параметра



t-статистика (оценка/стандарная ошибка)



p-value



Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Доверительные интервалы для параметров модели

Даем параметру не точечную (одно значение), а интервальную оценку с заданным уровнем надежности.

Такая оценка предпочтительна при небольшом объеме выборки

```
> confint(lm1)
```

```
> confint(lm1, level=0.95)
```

	2.5 %	97.5 %
(Intercept)	5.4168245	6.6776856
mtcars\$mpg	-0.1709569	-0.1107671

Уровень надежности (95%) означает вероятность того, что значение параметра попадет в этот интервал.

Оценка качества линейной регрессионной модели

Доля объясненной дисперсии
(чем ближе к 1, тем лучше)

> summary(lm1)

...

$$R^2 = \frac{Var(\hat{Y})}{Var(Y)} = SSX/SST$$

Multiple R-squared: 0.7528, Adjusted R-squared: 0.7446

F-statistic: 91.38 on 1 and 30 DF, p-value: 1.294e-10

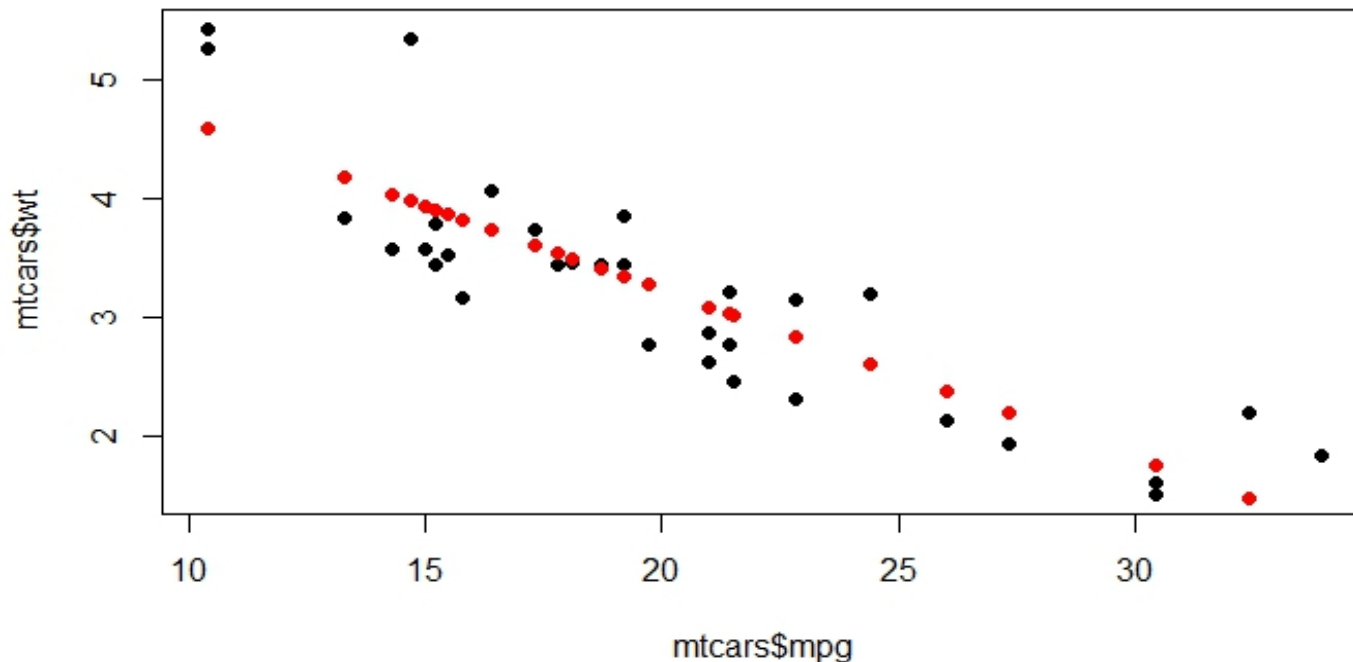
**F-статистика (отношение
объясненной дисперсии к
ошибочной)**

P-value для всей модели

$$F = \frac{Var(\hat{Y})}{Var(error)} = (SSX) / (SSE / (npoints - nlevels))$$

Предсказания модели = fitted values

```
> plot(mtcars$mpg, mtcars$wt, pch=19)
> lm1<-lm(mtcars$wt~mtcars$mpg)
> points(mtcars$mpg, lm1$fitted, pch=19,
  col='red')
```



Задание 1

- Создайте датасет с аллигаторами:

```
alligator = data.frame( lnLength = c(3.87, 3.61,  
  4.33, 3.43, 3.81, 3.83, 3.46, 3.76, 3.50, 3.58,  
  4.19, 3.78, 3.71, 3.73, 3.78), lnWeight = c(4.87,  
  3.93, 6.46, 3.33, 4.38, 4.70, 3.50, 4.50, 3.58,  
  3.64, 5.90, 4.43, 4.38, 4.42, 4.25) )
```

- Постройте линейную модель зависимости веса от длины аллигаторов
- Какую долю дисперсии объясняет модель?

Построение модели с несколькими предикторами

Шаг 1. Модель с 1 предиктором

```
> L_M=lm(Price_RUR ~ Memory_Gb, data=laptop)
> summary(L_M)
```

...

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5023	1651	3.04	0.0026	**
Memory_Gb	4442	333	13.34	<2e-16	***

...

Residual standard error: 12300 on 304 degrees of freedom

Multiple R-squared: 0.369, Adjusted R-squared: 0.367

F-statistic: 178 on 1 and 304 DF, p-value: <2e-16

Корреляция и регрессия

```
> cor.test(laptop$Price_RUR, laptop$Memory_Gb)
```

Pearson's product-moment correlation

data: laptop\$Price_RUR and laptop\$Memory_Gb

t = 13.3, df = 304, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.532 0.674

sample estimates:

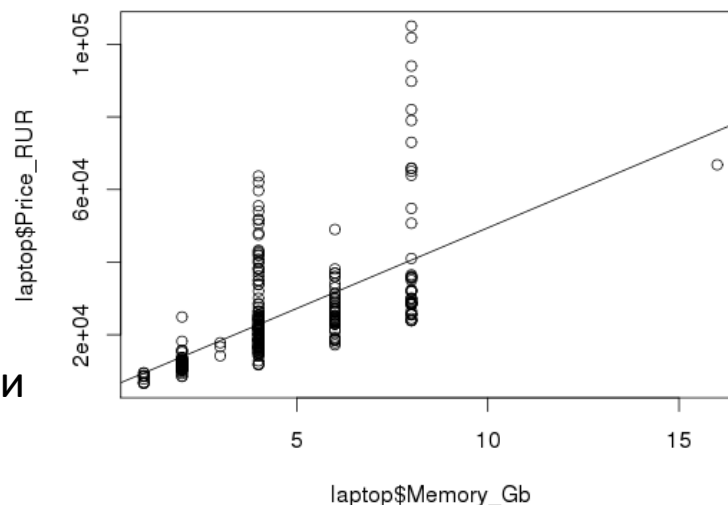
cor

0.608

```
> 0.608^2
```

```
[1] 0.37
```

R2 равен квадрату коэффициента корреляции



Несколько предикторов

- Как цена ноутбука зависит от объема памяти, объема жесткого диска и размера дисплея?
- Предикторы разделены через +

```
> l_MSH=lm(Price_RUR ~ Memory_Gb + Screen_size_inch +  
  HDD_Gb, data=laptop)
```

```
> summary(l_MSH)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	16475.481	4437.238	3.713	0.000244	***
Memory_Gb	7266.167	406.667	17.868	< 2e-16	***
Screen_size_inch	-511.390	350.186	-1.460	0.145237	
HDD_Gb	-31.022	3.305	-9.387	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10650 on 302 degrees of freedom

Multiple R-squared: 0.5308, Adjusted R-squared: 0.5262

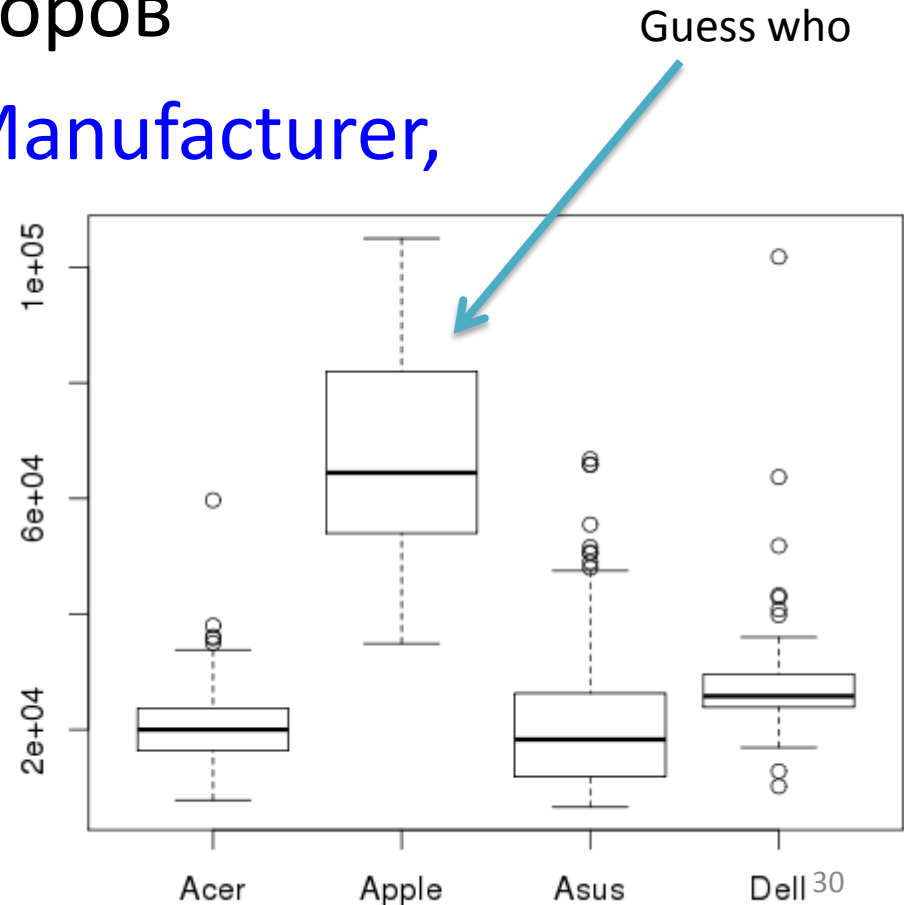
F-statistic: 113.9 on 3 and 302 DF, p-value: < 2.2e-16

Если x – фактор? Пример

- Уже умеем искать значимые отличия при разных уровнях факторов

```
> boxplot(Price_RUR ~ Manufacturer,  
data=laptop)
```

Верно ли, что для хотя бы одного уровня фактора наблюдаем отличия?



Если x – фактор? Модель

- Почему одна переменная превратилась в несколько?

```
> L_M=lm(Price_RUR ~ Manufacturer,  
data=laptop)  
> summary(L_M)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	21198.6	1415.0	14.981	< 2e-16	***
ManufacturerApple	46078.5	3527.5	13.063	< 2e-16	***
ManufacturerAsus	206.2	1714.4	0.120	0.904357	
ManufacturerDell	7427.6	2079.1	3.573	0.000411	***

Residual standard error: 12090 on 302 degrees of freedom

Multiple R-squared: 0.3958, Adjusted R-squared:

0.3898

F-statistic: 65.95 on 3 and 302 DF, p-value: < 2.2e-16

Если x – фактор?

- Влияет ли цвет ноутбука на его цену?
- Модель, если x – число: $y_i = \alpha x_{1i} + \beta x_{2i} + \varepsilon_i$
- Если x – фактор, то такая запись не подходит. Вместо этого:

$$y_i = \alpha_1 I(x_{1i}=black) + \alpha_2 I(x_{1i}=white) + \dots + \varepsilon_i$$

Коэффициент

(подбираются при построении модели)

Индикатор (равен 1, если x – черный цвет, иначе 0)

Если две факторные переменные?

$$y_i = \alpha_1 I(x_{1i}=black) + \alpha_2 I(x_{1i}=white) + \dots + \\ + \beta_1 I(x_{2i}=Apple) + \beta_2 I(x_{2i}=ASUS) + \dots + \varepsilon_i$$

Числа + факторы

- **Шаг 1.** Как зависит цена ноутбука от размера жесткого диска?

#Построим lm с 1 переменной

```
> l1=lm(Price_RUR ~ HDD_Gb, data=laptop)
```

```
> summary(l1)
```

```
...
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	21238.584	2027.553	10.475	<2e-16	***
HDD_Gb	6.913	3.410	2.027	0.0435	*

```
---
```

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
...
```

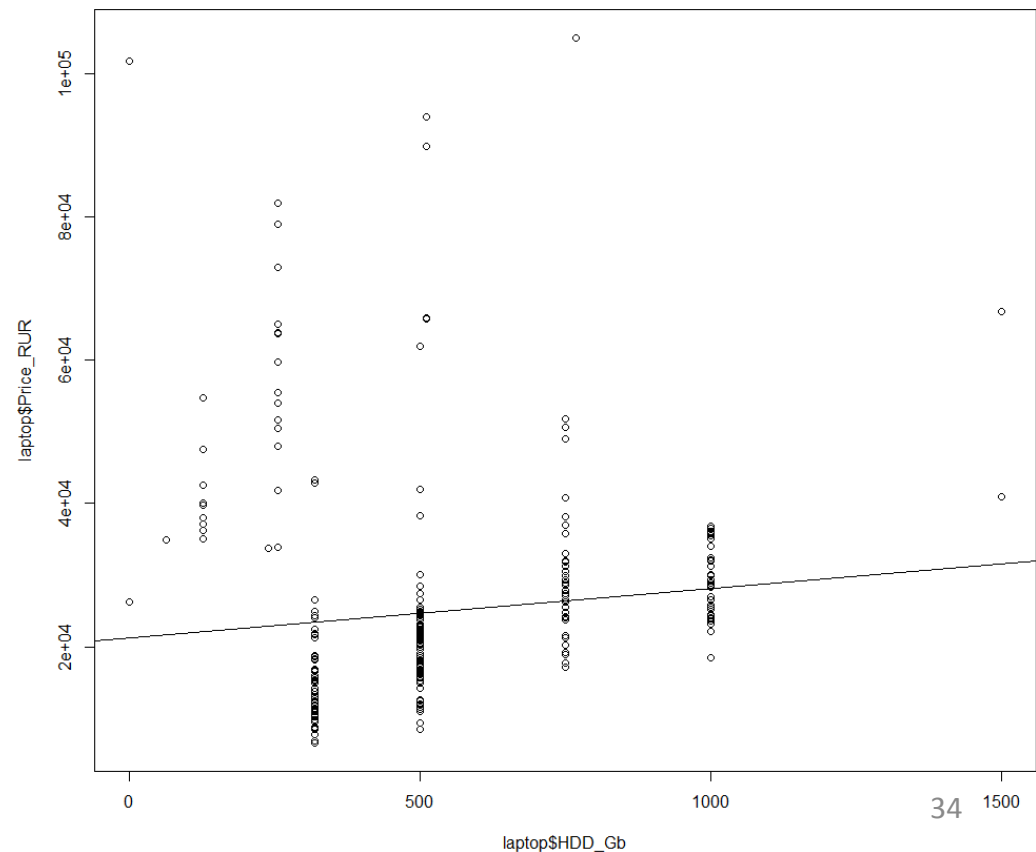
#0.043 – на грани порога значимости

Числа + факторы

#Нарисуем scatterplot

```
> plot(laptop$HDD_Gb, laptop$Price_RUR)
```

```
> abline(l1)
```

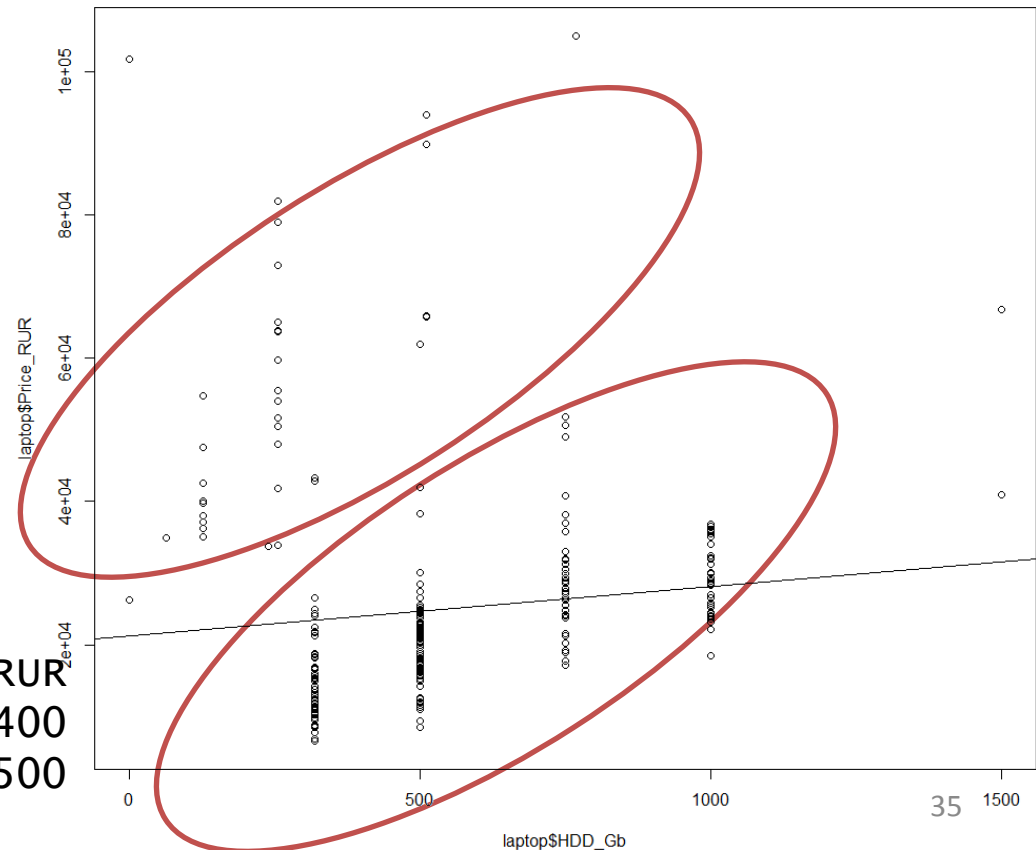


Числа + факторы

#Нарисуем scatterplot

```
> plot(laptop$HDD_Gb, laptop$Price_RUR)  
> abline(l1)
```

Видно 2 группы
Это знак, что мы
чего-то не учли



Memory_Gb	HDD_Gb	HDD_type	Price_RUR
4	500	HDD	16400
4	500	HDD	18500

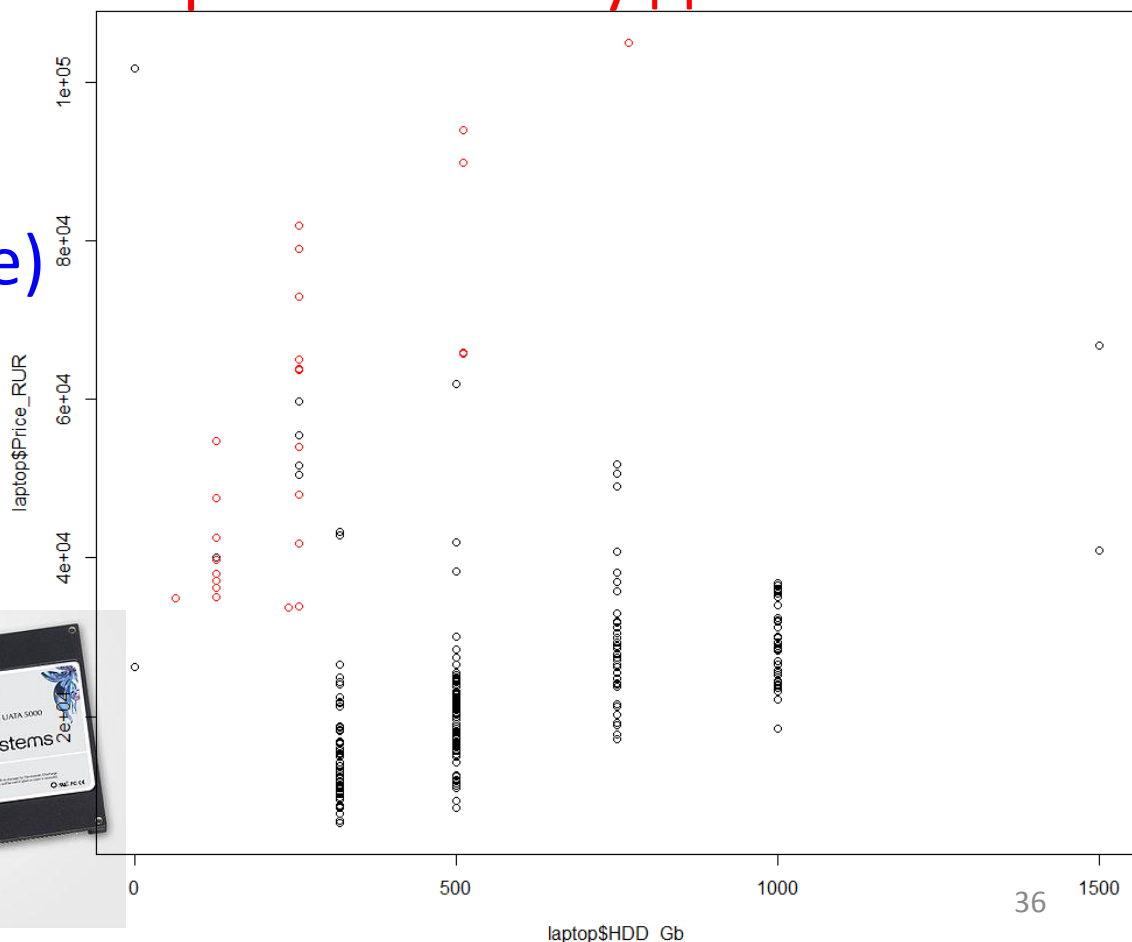
Числа + факторы

Шаг 2. Фактор, который мы не учли – тип накопителя, HDD или SSD. Вторые сильно дороже

Нарисуем scatterplot и покрасим по типу диска

```
> plot(laptop$HDD_Gb,  
      laptop$Price_RUR,  
      col=laptop$HDD_type)
```

Похоже, мы правы



Числа + факторы

#Добавим тип диска как переменную в модель

```
> l2=lm(Price_RUR ~ HDD_Gb + HDD_type, data=laptop)
```

```
> summary(l2)
```

...

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	10741.160	1594.347	6.737	8.14e-11	***
HDD_Gb	20.290	2.591	7.830	8.27e-14	***
HDD_typeSSD	40797.575	2442.199	16.705	< 2e-16	***

...

Значимость улучшилась

Наклон прямой будет одинаковым, но среднее между группами - отличается

Числа + факторы

#Нарисуем scatterplot и две
регрессионные прямые (для
каждого значения фактора)

```
> l2$coeff  
(Intercept)      HDD_Gb  HDD_typeSSD  
10741.15975      20.28962  40797.57542
```

```
> plot(laptop$HDD_Gb,  
      laptop$Price_RUR,  
      col=laptop$HDD_type)
```

```
> abline(l2$coeff[1], l2$coeff[2],  
        col="black")
```

```
> abline(l2$coeff[1]+l2$coeff[3],  
        l2$coeff[2], col="red")
```

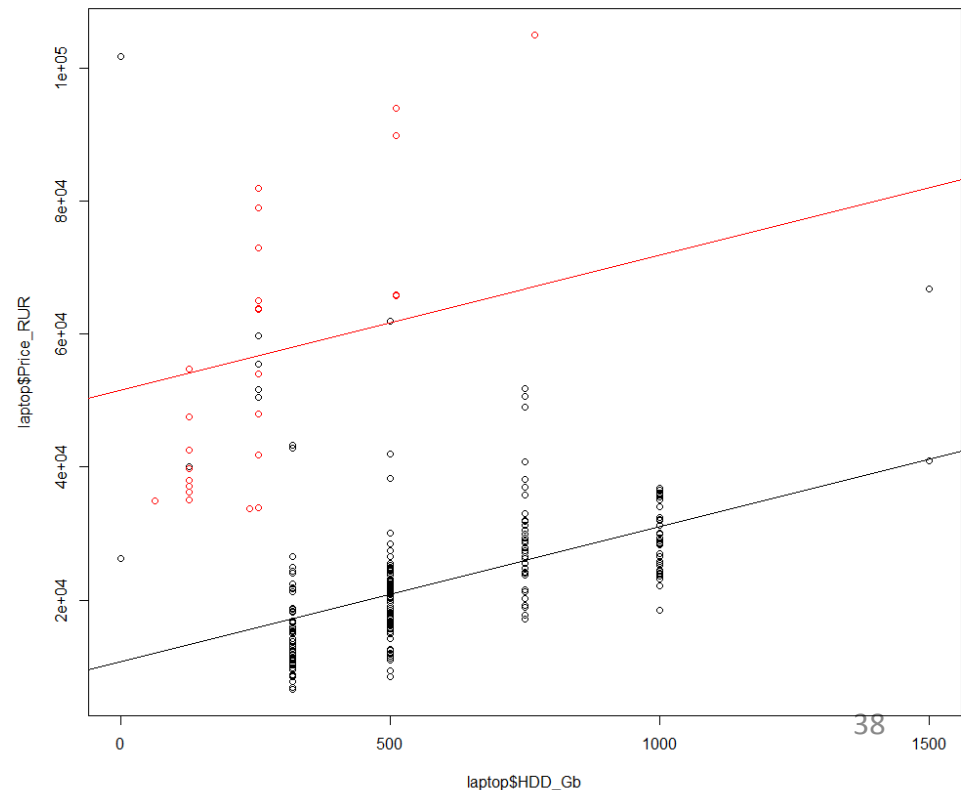
$Price = 10741 + 20 * HDD_Gb + 40797 * I(type = SSD)$

if $type \neq SSD$:

$Price = 10741 + 20 * HDD_Gb + 40797 * 0$

if $type = SSD$:

$Price = 10741 + 20 * HDD_Gb + 40797 * 1$
 $= (10741 + 40797) + 20 * HDD_Gb$

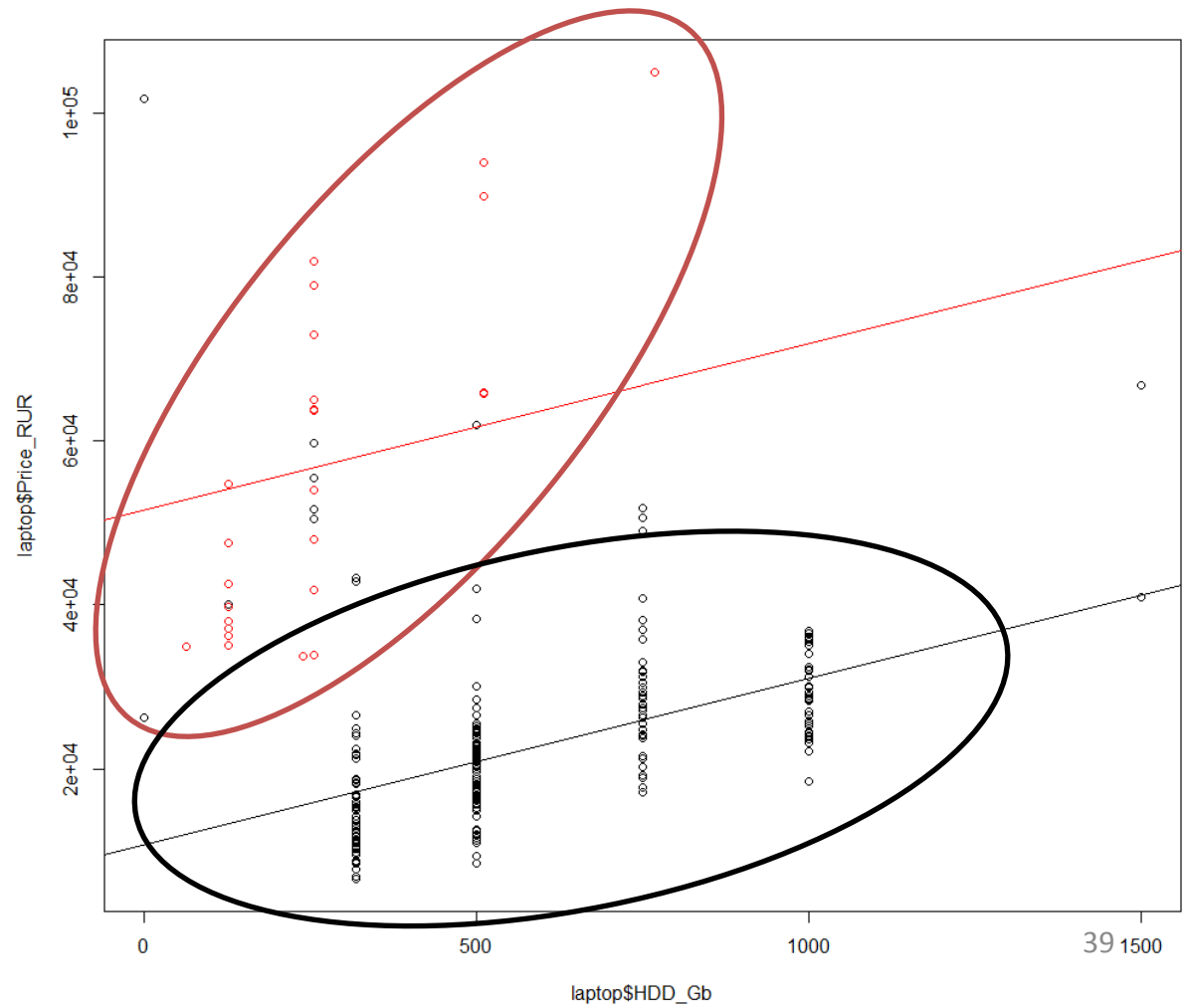


Числа + факторы.

Взаимодействие переменных

Шаг 3.

- Похоже, что наклоны для двух групп тоже отличаются, другими словами, каждый 1Gb SSD стоит дороже каждого 1Gb HDD.
- Как это учесть?



Числа + факторы.

Взаимодействие переменных

- Как в формуле сделать разные наклоны для разных групп факторов?

- Было: $Price = a + b * HDD_Gb + c * I(type = SSD)$

- Надо:

$Price = a +$

$+ (b1 * I(type = SSD) + b2 * I(type = HDD)) * HDD_Gb +$

$+ c * I(type = SSD)$

$= a +$

$+ (b1 * I(type = SSD) + b2 * (1 - I(type = SSD))) * HDD_Gb +$

$+ c * I(type = SSD)$

...преобразуем формулу...

Числа + факторы.

Взаимодействие переменных

```
> l3=lm(Price_RUR ~ HDD_Gb + HDD_type + HDD_Gb:HDD_type, data=laptop)
> l3=lm(Price_RUR ~ HDD_Gb*HDD_type , data=laptop)
> summary(l3)
```

Эквивалентные записи:

$a*b := a + b + a:b$

Call:

```
lm(formula = Price_RUR ~ HDD_Gb * HDD_type, data = laptop)
```

Residuals:

Min	1Q	Median	3Q	Max
-21886	-6049	-1461	2885	89344

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	12430.529	1525.776	8.147	9.97e-15	***
HDD_Gb	17.270	2.488	6.941	2.38e-11	***
HDD_typeSSD	18232.081	4265.934	4.274	2.58e-05	***
HDD_Gb:HDD_typeSSD	80.870	12.874	6.281	1.17e-09	***

Взаимодействие
значимо

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10480 on 302 degrees of freedom

Multiple R-squared: 0.5457, Adjusted R-squared: 0.5412

F-statistic: 120.9 on 3 and 302 DF, p-value: < 2.2e-16

Числа + факторы.

Взаимодействие переменных

```
> l3$coeff
```

(Intercept)	HDD_Gb	HDD_typeSSD	HDD_Gb:HDD_typeSSD
12430.52872	17.26953	18232.08144	80.86960

Нарисуем регрессионные прямые для каждого из значений факторной переменной

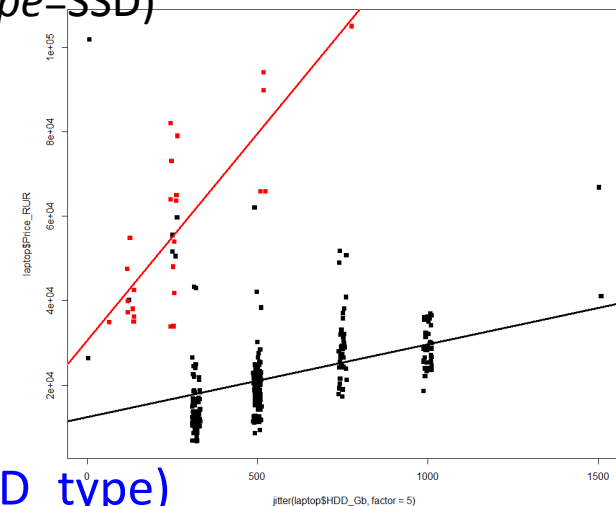
$Price = 12430 + 17 * HDD_Gb + 18232 * I(type = SSD) + 81 * HDD_Gb * I(type = SSD)$

if type ≠ SSD:

$Price = 12430 + 17 * HDD_Gb + 18232 * 0 + 81 * HDD_Gb * 0$

if type = SSD:

$Price = 12430 + 17 * HDD_Gb + 18232 * 1 + 81 * HDD_Gb * 1 =$
 $= (12430 + 18232) + (17 + 81) * HDD_Gb$



```
> plot(laptop$HDD_Gb, laptop$Price_RUR, col=laptop$HDD_type)
```

```
> abline(l3$coeff[1], l3$coeff[2])
```

```
> abline(l3$coeff[1]+l3$coeff[3], l3$coeff[2]+l3$coeff[4], col="red")
```

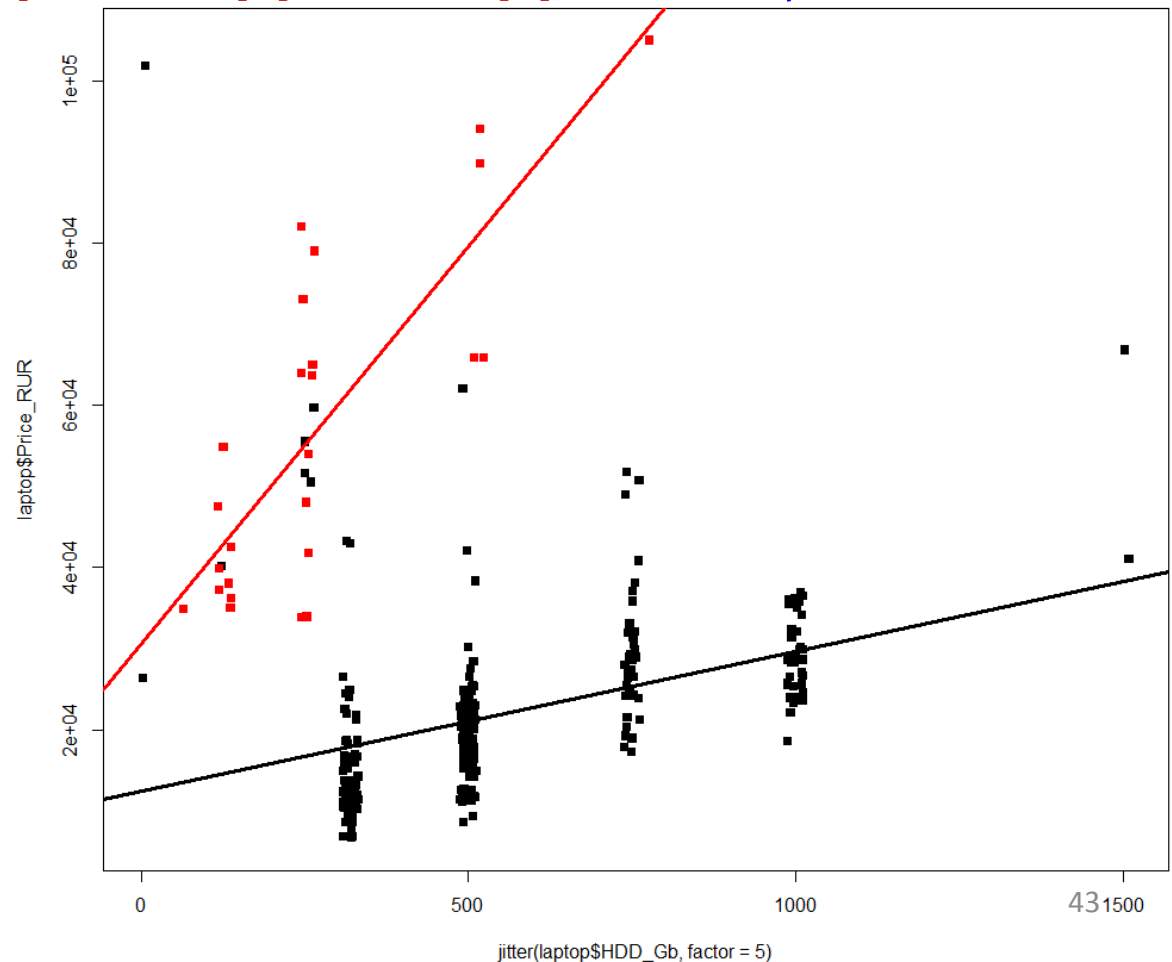
Числа + факторы.

Взаимодействие переменных

```
> plot(laptop$HDD_Gb, laptop$Price_RUR, col=laptop$HDD_type)
```

```
> abline(l3$coeff[1], l3$coeff[2])
```

```
> abline(l3$coeff[1]+l3$coeff[3], l3$coeff[2]+l3$coeff[4], col="red")
```



Обозначения в формулах

$$a * b = a + b + a : b$$

- $y \sim x + 0$
- $y \sim x - 1$

– **x1** удаляет предиктор x1 из модели

- $y \sim a * b - a$
- $y \sim b + a : b$

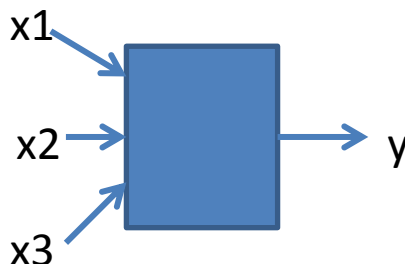
- $y \sim .$ **# . – все остальные переменные**

- $l(a+b), l(a*b)$ **#защитить арифметические операторы**

Для чего нужны линейные модели?

Входные данные

y	x1	x2	x3



Значимость каждой переменной:

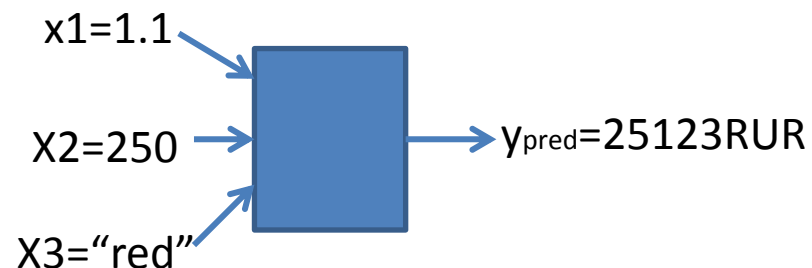
Предсказание y по x

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	9503.178	1733.057	5.483	9.10e-08 ***	
Memory_Gb	6232.204	421.098	14.800	< 2e-16 ***	
HDD_Gb	-26.604	3.275	-8.123	1.34e-14 ***	
Colorblue	-2496.492	4012.744	-0.622	0.5343	
...					
Colorred	1685.698	2736.167	0.616	0.5383	
Colorsilver	8617.956	1679.963	5.130	5.33e-07 ***	
...					

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10330 on 289 degrees of freedom
Multiple R-squared: 0.578, Adjusted R-squared: 0.5547
F-statistic: 24.74 on 16 and 289 DF, p-value: < 2.2e-16



predict

```
> L3 = lm(formula = Price_RUR ~ HDD_Gb +  
HDD_type + HDD_Gb:HDD_type, data = laptop)  
> newlaptops=data.frame( HDD_Gb=c(200, 1000,  
500), HDD_type=c("SSD", "HDD", "HDD"))
```

```
> newlaptops
```

	HDD_Gb	HDD_type
1	200	SSD
2	1000	HDD
3	500	HDD

```
> predict(L3, newlaptops)
```

	1	2	3
	50290.44	29700.06	21065.29

Модель

Dataframe с x-координатами
новых точек, для которых
делается предсказание y.

Названия колонок должны
соответствовать
предикторам модели

Вектор предсказанных
значений y