

Statistical packages - report 1

Urszula Grochocińska, Marcin Mazurkiewicz

December 2, 2018

Contents

1	Introduction	1
2	Aim of analysis	1
3	Data preprocessing	1
4	Model fitting	4
4.1	Continuous variables	4
4.2	Additional categorical variables	5
4.3	Model with additional interactions	6
5	Other models	7
6	Check model assumptions	7
7	Outliers handling	8
8	Order models	8
9	Check significance of parameters	8
10	Conclusions	10

1 Introduction

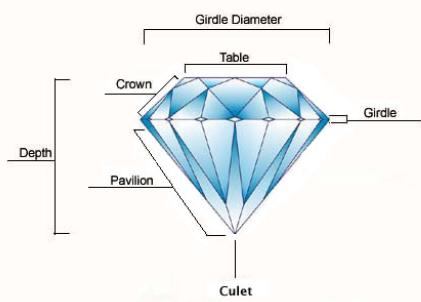


Figure 1: Anatomy of diamond. source:
<https://www.everything-wedding-rings.com>

For this report we have decided to analyse "Diamonds", dataset about the features and prices of stones which we have found on Kaggle [<https://www.kaggle.com/shivam2503/diamonds>]. It contains 54940 records and each one has 11 columns with different variables. First one is irrelevant, because it's just index number. Then we have numerical variables - x, y, z that describe size of diamond in each of three dimensions. There is also weight given in carats. Today, a carat is equal to exactly 0.2 grams). Carat weight is unrelated to the similar sounding karat, which refers to gold's purity. Next we have two variables describing percentage relation between appropriate measures within the diamond. Depth is the height of a diamond, measured from the culet to the table, divided by its average girdle diameter. Table is the width of the diamond's table expressed as a percentage of its average diameter. The last numerical variable is price. Coming to discrete variables we have cut describing the quality of diamond's cut: Fair, Good, Very Good, Premium, Ideal. The cut describes the symmetry proportioning

and polish of the diamond. Then there is color of diamond ranged descending from D to J. The later letter in the alphabet the more yellow. And the last is clarity of particular diamond with possible values: FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3 (ordering from best to worse). It refers to the absence of inclusions and blemishes.

2 Aim of analysis

Jewelry is a kind of art made of expensive raw material. These art is based on precision and integrity but because of the fact that it is an art sometimes the defects increase the price. These defects make diamond unique and uniqueness is the main part of art. In our analyze we would like to check which features influence the price and at which level. Besides we are interested how important are all features that are influenced by jeweler(shape, cut) and that what are natural (color, clarity)

3 Data preprocessing

This, first step of working with data we started from removing unnecessary variables. We remove the index column, because it is irrelevant information. Then we looked for missing data. Dataset didn't contain any NaNs, but some values were impossible to occur in reality. To handle such situations we changed every zero from x, y or z column to NA, because this will make our upcoming work easier. We also checked if other numerical columns contain zeros, but they didn't. Next, we focused on ensuring that categorical variables don't contain typos. In order to do that we just looked at all values they take, but there wasn't need any cleaning. We didn't deal with outliers, because we will do it within next steps of analysis.

```
> diamonds_data <- diamonds_data[2:11] #data cleaning - removing unnecessary variables
> #changing 0 to NAs
> diamonds_data$z[diamonds_data$z==0] = NA
> diamonds_data$x[diamonds_data$x==0] = NA
> diamonds_data$y[diamonds_data$y==0] = NA
> any(is.na(diamonds_data)) #check if there are any nulls in our dataset
[1] TRUE

> diamonds_data_omited <- na.omit(diamonds_data)
> #checking for typos
> summary(diamonds_data$cut)

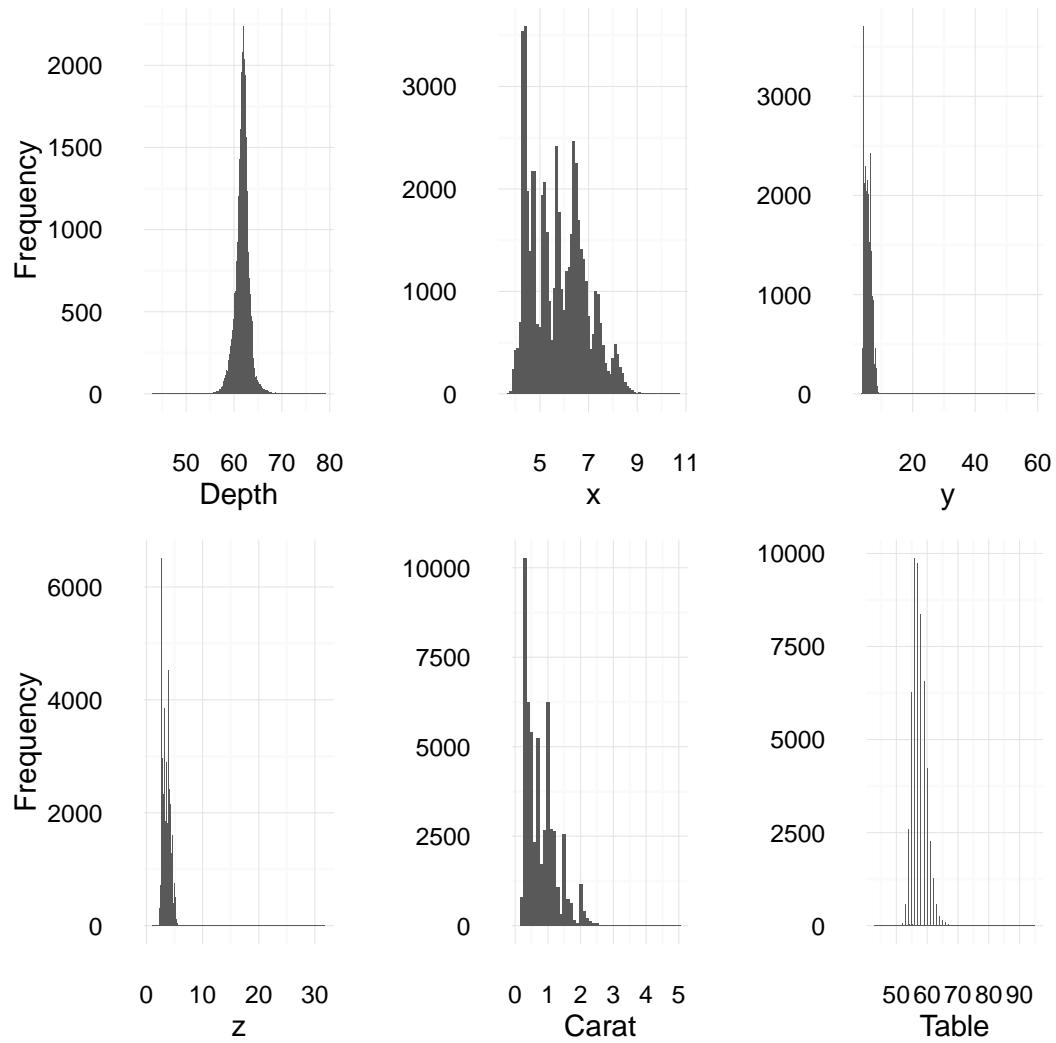
  Fair      Good      Ideal     Premium   Very Good
 1610       4906     21551      13791      12082

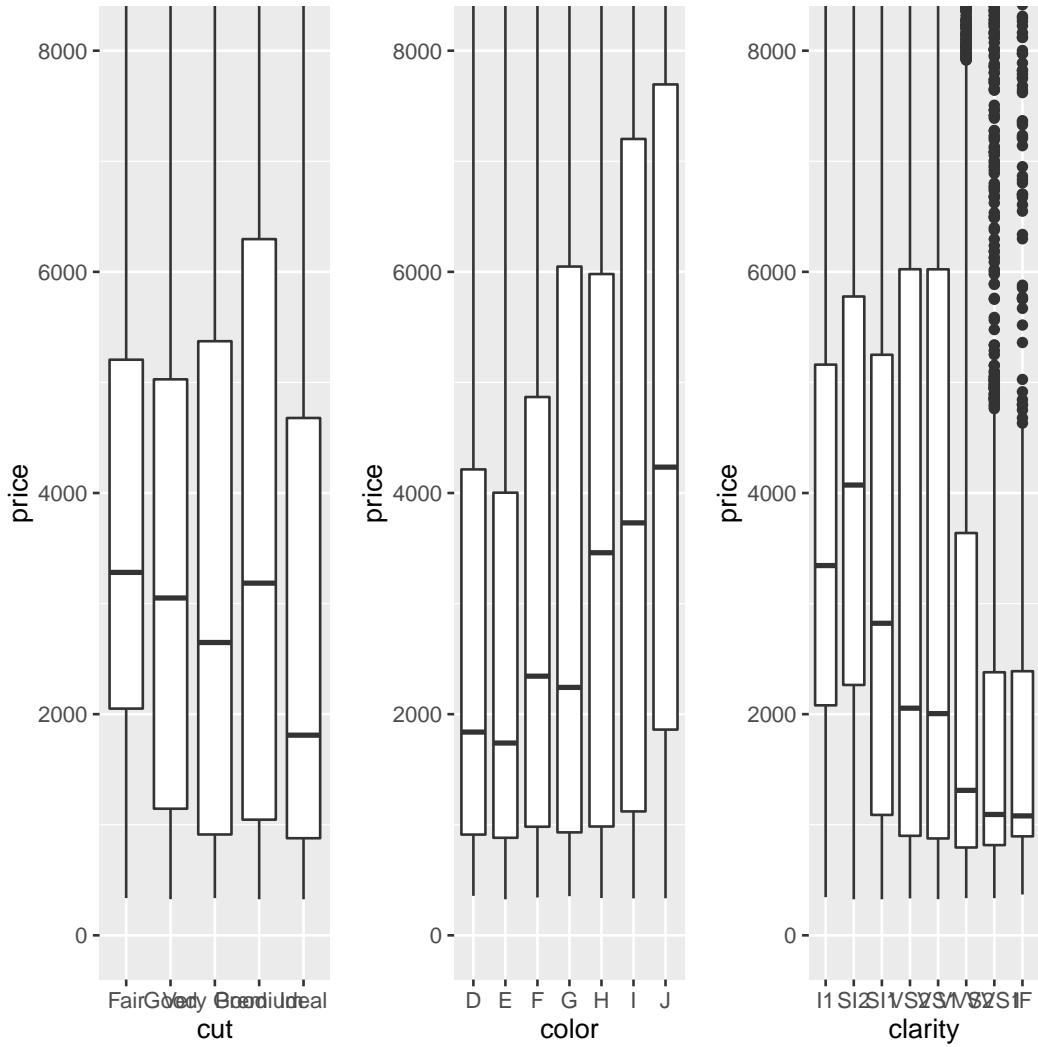
> summary(diamonds_data$color)

  D      E      F      G      H      I      J
 6775   9797  9542  11292  8304  5422  2808

> summary(diamonds_data$clarity)

  I1      IF      SI1      SI2      VS1      VS2      VVS1      VVS2
 741    1790  13065   9194   8171  12258   3655   5066
```





4 Model fitting

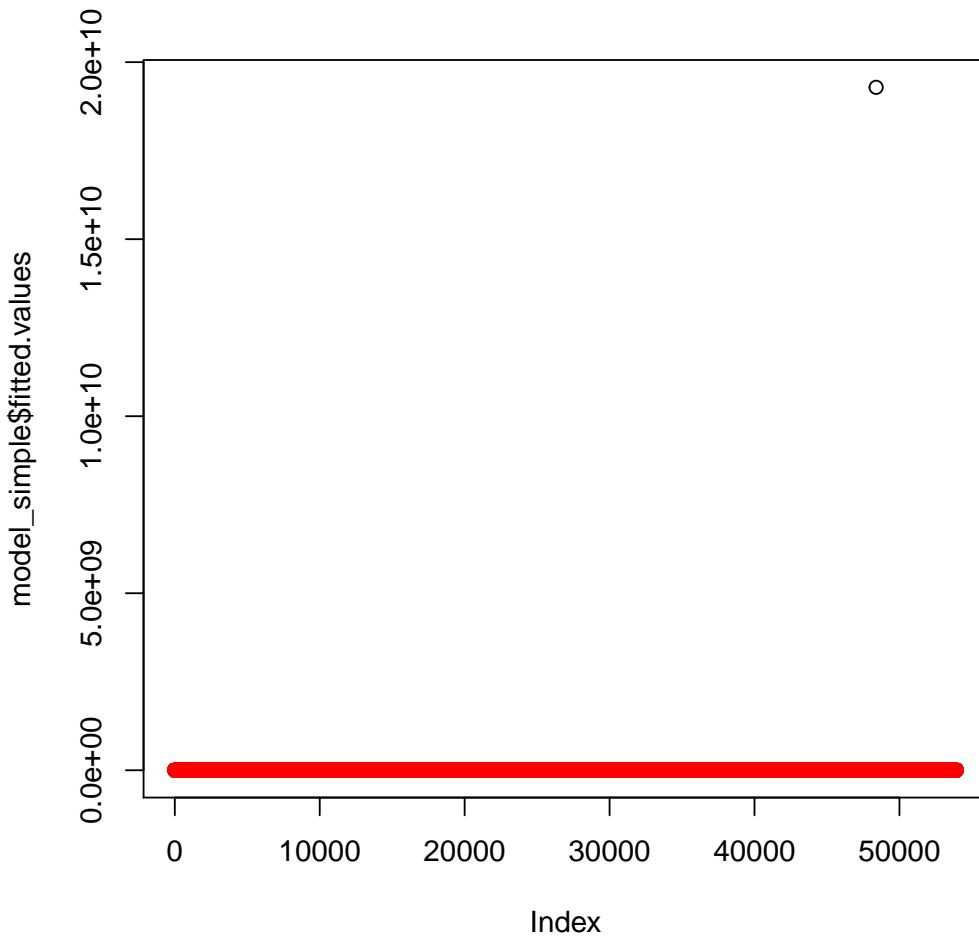
4.1 Continuous variables

To fit model to data we use function `glmulti` from package `glmulti`. To reduce complexity of model at this step we decided to choose maximum number of used variables to 3. We checked models with different distribution families. These were Gaussian, Poisson and Gamma. It turned out that Gamma model provides the best fit under the BIC criteria. The formula of this model is as follows:

$$\text{cost} = -1.0777\text{carat} + 0.5714z + 0.9717x + 1.0882. \quad (1)$$

```
> model_simple <- glm(formula = price~z+x+carat, data = diamonds_data_omited,
+                       family = Gamma(link = 'log'))
> plot(model_simple$fitted.values)
> points(diamonds_data_omited$price, col = 'red')
```

color(C_i)	E	F	G	H	I	J	
coefficient	-214.2	-282.6	-486.8	-993.2	-1477.5	-2397.4	
clarity(D_i)	IF	SI1	SI2	VS1	VS2	VVS1	VVS2
coefficient	5707.1	3922.3	2952.6	4878.2	4559.3	5339.5	5266.3



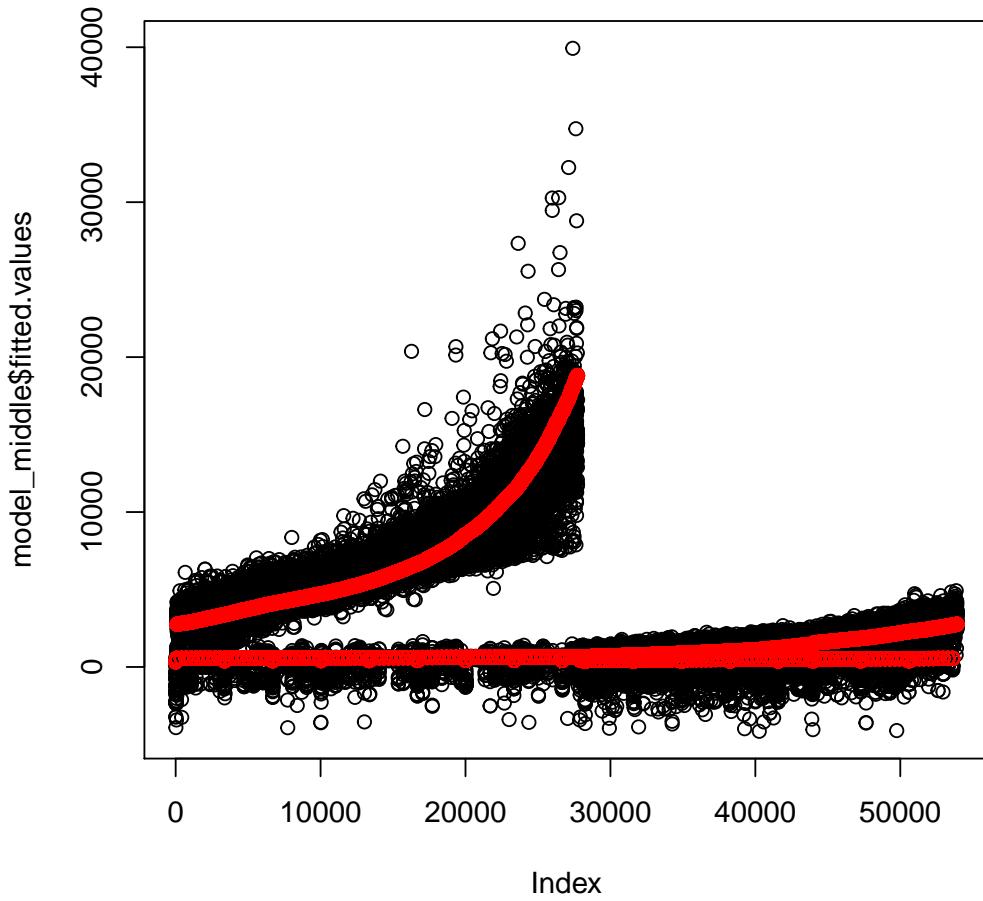
A the plot we can clearly see that there outliers, but we will handle it in next sections.

4.2 Additional categorical variables

In case of additional categorical variables we used very similar approach as in continuous variables. But this time we increased maximum number of variables to 4 and considered Gaussian family. The best model we have found is:

$$\text{cost} = -1016.9x + 11192.5\text{carat} + C_i + D_i - 2826.4, \quad (2)$$

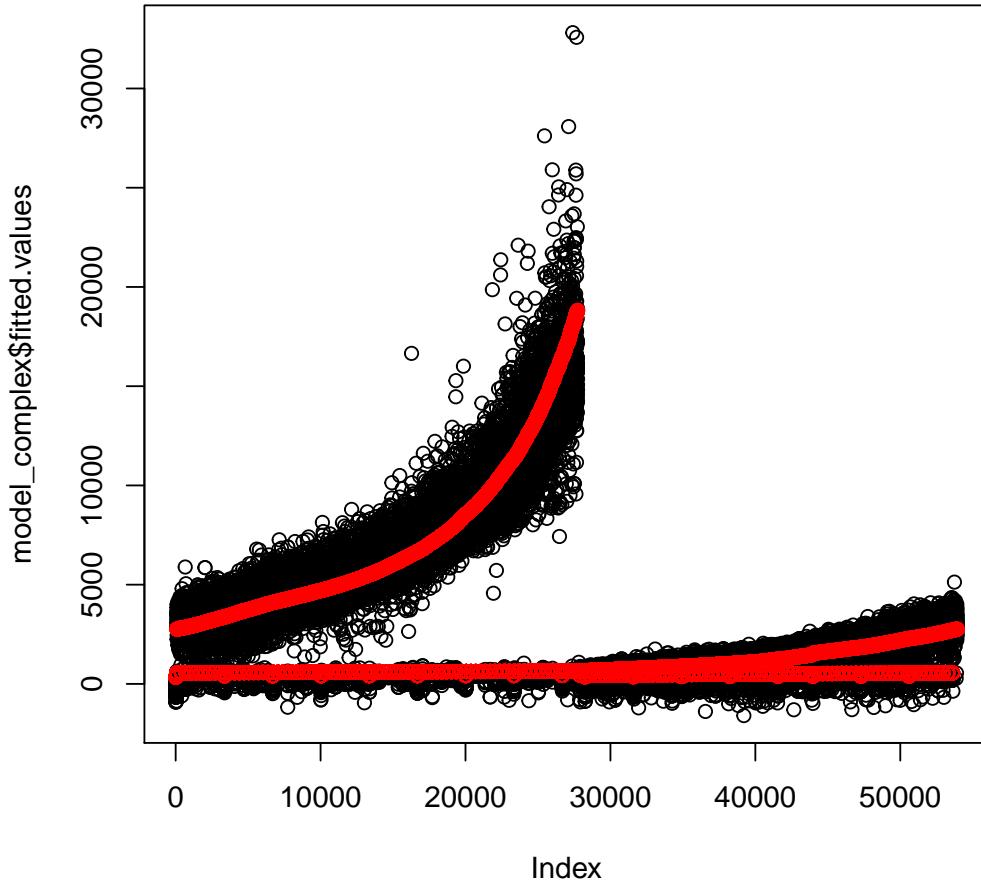
carat · color(C_i)	E	F	G	H	I	J	
coefficient	-219.1	-342.1	-737.5	-1373.5	-1920.8	-2786.8	
carat · clarity(D_i)	IF	SI1	SI2	VS1	VS2	VVS1	VVS2
coefficient	6891.3	3657.7	2664.9	4883.4	4448.8	6154.9	5788.6



4.3 Model with additional interactions

To find model with additional interactions we use very similar approach as in previous cases. Maximum number of variables is 4 and the family is Gaussian. That's how we find the following model:

$$\text{cost} = -6912.8\text{carat} + 1527.5\text{carat} \cdot x + C_i + D_i - 783.9, \quad (3)$$



5 Other models

To find those three models from previous section we used `glmulti` function from `glmulti` package. Thus we take into consideration all models with specific restrictions. We set maximum model size to 3 in case with only continuous variables and to 4 in the rest of cases.

We used `glmulti` function to find all of considered models. But it turned out, that in second and third case the best model doesn't include `z` variable. That arouse our curiosity if model excluding this variable is significantly better than with it. Under BIC criteria best model with categorical variables with additional `z` is only 0.02 % better than without it. Although it gives better BIC it contains more variables so it may be considered as worse model.

6 Check model assumptions

We checked if fitted models fulfill appropriate assumptions. It turns out that none of them do, but we will assume that they do. On the plots below we can see that none of them has normally distributed residuals. Chart 'Residuals vs Fitted' present the distribution of residuals which should be placed around 0. In each of our model the assumption about homogeneity is not met. The normal Q-Q shows if your residuals are normally distributed. Residuals should go around the diagonal line but in our cases they are not. Cook's distance for residuals helps to detect outliers. They are visible in our cases and we try to handle them in next step of our analysis. Scale location - to describe!!!! [Tu

	number of variables	AIC	difference between the model and the best model
Model with only continuous variables without outliers	3	848300	0.0%
Model with only continuous variables with outliers	3	848800	-0.06%
Model with continuous and categorical variables	4	912700	-7.6%
Model with interactions	4	884100	-4.2%

bym chciała, żeby pojawiły się tylko wykresy o których wspominamy w tekście] assumptions.png assumptions.pdf assumptions.jpg assumptions.mps assumptions.jpeg assumptions.jbig2 assumptions.jb2 assumptions.PNG assumptions.PDF assumptions.JPG assumptions.JPG assumptions.JBIG2 assumptions.JB2 assumptions.eps

7 Outliers handling

To find Outliers of our fits we use outlierTest function. It provides us a list of most significant outliers. We decided to delete outliers that are too far from rest of the data. This decision was made because of the fact that this observations are much different than others in only chosen characteristics and it could be the mistake. After that we can try to fit the models again (three lines below but using dataset with deleted outliers - different for each glm or fit by glmulti again, but maybe not necessarily).

8 Order models

It turns out that the best fit we managed to make is with the simplest model with just continuous variables. It becomes even better when we delete outliers. The worst fit of those three models was obtained for model with continuous and categorical variables, but without interactions. The most complex model with additional interactions gave fit somewhere between two previous ones.

9 Check significance of parameters

To check if all parameters in our models are significant we use function summary called at the best model that we find. Than we look at the significance codes and we can find out what parameters are significant. The all summaries are shown below:

```
> summary(model_simple_out)

Call:
glm(formula = price ~ z + x + carat, family = Gamma(link = "log"),
     data = diamonds_data_omited_simplout)

Deviance Residuals:
    Min      1Q      Median      3Q      Max 
-5.4945 -0.1971 -0.0367  0.1362  2.8305 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.088645  0.019882  54.76 <2e-16 ***
z            0.571280  0.007689  74.30 <2e-16 ***
x            0.971664  0.006364 152.69 <2e-16 ***
carat       -1.077577  0.011994 -89.85 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.07442222)
```

```

Null deviance: 52790.7  on 53889  degrees of freedom
Residual deviance: 3696.1  on 53886  degrees of freedom
AIC: 848304

Number of Fisher Scoring iterations: 16

> summary(model_middle)

Call:
glm(formula = price ~ color + clarity + x + carat, family = gaussian(),
     data = diamonds_data_omited)

Deviance Residuals:
    Min      1Q  Median      3Q      Max 
-21911.2   -592.4   -177.9    383.6   10661.4 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2826.45     93.97  -30.08 <2e-16 ***
colorE       -214.17    18.14  -11.80 <2e-16 ***
colorF       -282.59    18.35  -15.40 <2e-16 ***
colorG       -486.83    17.96  -27.11 <2e-16 ***
colorH       -993.16    19.09  -52.02 <2e-16 ***
colorI      -1477.46    21.46  -68.83 <2e-16 ***
colorJ      -2397.39    26.50  -90.46 <2e-16 ***
clarityIF    5707.09    51.01 111.89 <2e-16 ***
claritySI1   3922.27    43.75  89.66 <2e-16 ***
claritySI2   2952.57    44.01  67.09 <2e-16 ***
clarityVS1   4878.18    44.59 109.41 <2e-16 ***
clarityVS2   4559.32    43.90 103.86 <2e-16 ***
clarityVVS1  5339.55    47.12 113.31 <2e-16 ***
clarityVVS2  5266.32    45.85 114.87 <2e-16 ***
x           -1016.88    21.46  -47.38 <2e-16 ***
carat       11192.46    50.74 220.59 <2e-16 ***

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

(Dispersion parameter for gaussian family taken to be 1313862)

```

```

Null deviance: 8.5723e+11  on 53919  degrees of freedom
Residual deviance: 7.0822e+10  on 53904  degrees of freedom
AIC: 912687

Number of Fisher Scoring iterations: 2

> summary(model_complex)

Call:
glm(formula = price ~ carat + carat:x + carat:color + carat:clarity,
     family = gaussian(), data = diamonds_data_omited)


```

```

Deviance Residuals:
    Min      1Q  Median      3Q      Max 
-14794.9   -401.1   -26.9    332.4   8974.6 
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-783.901	17.108	-45.82	<2e-16 ***
carat	-6912.758	86.359	-80.05	<2e-16 ***
carat:x	1527.515	9.482	161.10	<2e-16 ***
carat:colorE	-219.073	18.503	-11.84	<2e-16 ***
carat:colorF	-342.142	17.934	-19.08	<2e-16 ***
carat:colorG	-737.484	17.290	-42.65	<2e-16 ***
carat:colorH	-1373.460	17.299	-79.40	<2e-16 ***
carat:colorI	-1920.826	18.141	-105.89	<2e-16 ***
carat:colorJ	-2786.799	19.940	-139.76	<2e-16 ***
carat:clarityIF	6891.256	42.282	162.98	<2e-16 ***
carat:claritySI1	3657.671	24.449	149.61	<2e-16 ***
carat:claritySI2	2664.851	24.041	110.84	<2e-16 ***
carat:clarityVS1	4883.444	25.999	187.83	<2e-16 ***
carat:clarityVS2	4448.760	24.833	179.15	<2e-16 ***
carat:clarityVVS1	6154.864	34.457	178.62	<2e-16 ***
carat:clarityVVS2	5788.636	29.482	196.34	<2e-16 ***

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

(Dispersion parameter for gaussian family taken to be 773362.4)

Null deviance: 8.5723e+11 on 53919 degrees of freedom
Residual deviance: 4.1687e+10 on 53904 degrees of freedom
AIC: 884111

Number of Fisher Scoring iterations: 2

For two other models we do the same analysis. It turns out that in all models that were fitted by us all parameters are significant. It means that we shouldn't get rid of them.

10 Conclusions