Exercise 12 - Multiple regression

Zoltan Kekecs

8 May 2020

Table of Contents

Abstract

This exercise will show you how multiple predictors can be used in the same regression model to achieve better prediction efficiency.

Data management and descriptive statistics

Load data about housing prices in King County, USA

In this exercise we will predict the price of apartments and houses.

We use a dataset from Kaggle containing data about housing prices and variables that may be used to predict housing prices. The data is about accomodations in King County, USA (Seattle and sorrounding area).

We only use a portion of the full dataset now containing information about N = 200 accomodations.

The .sav file can be downloaded from here:

https://github.com/kekecsz/SIMM32/blob/master/2020/Lab_2/House%20price%20King%20County.sav

Check the dataset

You should always check the dataset for coding errors or data that does not make sense.

View data in the data editor and display simple descriptive statistics and plots. You can find the commands for data exploration in the **Analyze > Descriptive Statistics tab**

Analyze > Descriptive Statistics tab > Frequencies

Analyze > Descriptive Statistics tab > Descriptives

Analyze > Descriptive Statistics tab > Explore

We are going to predict price of the apartment using the variables sqft_living (the square footage of the living area), and grade (overall grade given to the housing unit, based on King County grading system), so let’s focus on these variables.

Later we are also going to use a categorical variable, has_basement (whether the apartment has a basement or not) as well.

```
* Descriptives

FREQUENCIES VARIABLES=price sqft_living grade basement
  /ORDER=ANALYSIS.

DESCRIPTIVES VARIABLES=price sqft_living grade
  /STATISTICS=MEAN STDDEV MIN MAX KURTOSIS SKEWNESS.

EXAMINE VARIABLES=price sqft_living grade
  /PLOT BOXPLOT HISTOGRAM NPPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

Multiple regression

Fitting the regression model

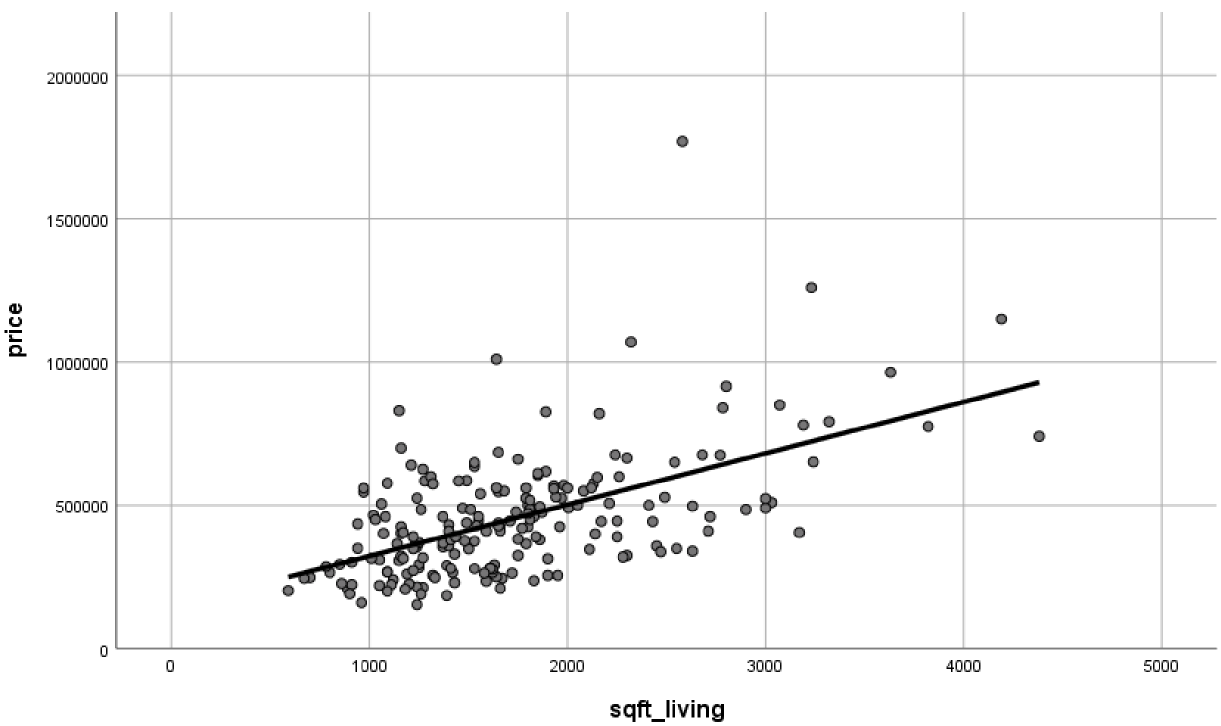
We fit a regression model with multiple predictors: sqft_living and grade. **Anaysis > Regression > Linear,** and let's ask for confidence intervals of regression coefficients in the

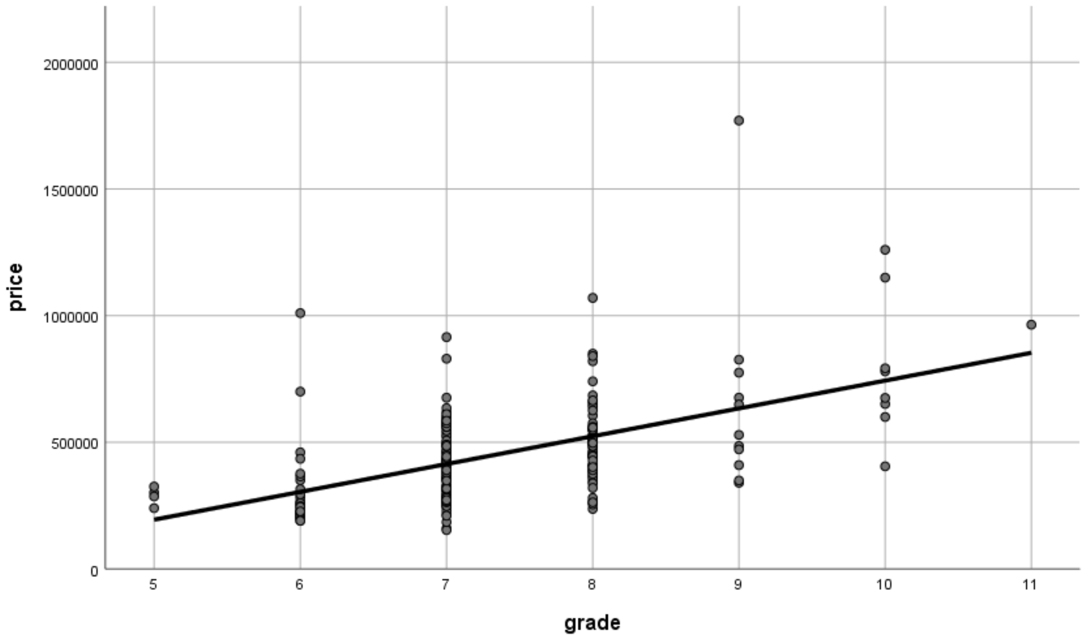
Statistics… button.

```
* Multiple regression


        REGRESSION
          /MISSING LISTWISE
          /STATISTICS COEFF OUTS CI(95) R ANOVA
          /CRITERIA=PIN(.05) POUT(.10)
          /NOORIGIN
          /DEPENDENT price
          /METHOD=ENTER sqft_living grade.
```

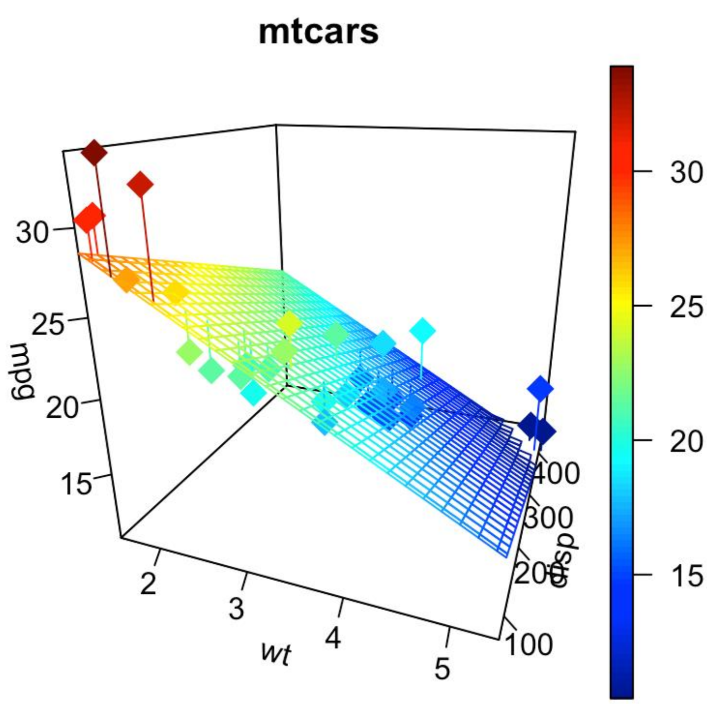
It is not trivial to visualize the regression equation in multiple regression. You can plot every simple regression separately, but that is not an accurate depiction of the prediction using the model.





Because in a multiple regression the regression "line" is actually a multidimensional plane.

Image from: http://www.sthda.com/sthda/RDoc/figure/3d-graphics/plot3d-regression-plane-1.png. This is just to demonstrate how a 3d scatterplot looks like with the regression plane overlayed, this plot is not a depiction of the data we use here.



Prediction

Again, we can compute predictions for specific values of predictors (new data), but we need to specify all predictor values (in this case, both sqft_living and grade of the apartment) to get a prediction. You can compute the predicted values in **Transform > Compute variable…** , by entering the regression formula based on the coefficients in the regression output.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95,0% Confidence Interval for B Lower Bound	95,0% Confidence Interval for B Upper Bound
1	(Constant)	-174389,862	95255,171		-1,831	,069	-362240,588	13460,864
	sqft_living	119,173	24,762	,374	4,813	,000	70,341	168,005
	grade	57352,786	16052,790	,278	3,573	,000	25695,416	89010,156

a. Dependent Variable: price

Based on this output, you can provide the following formula, given that you entered the new values for which you want to get a prediction to the variables called new_sqft_living and new_grade: -174389.86 + new_sqft_living * 119.17 + new_grade * 57352.79

COMPUTE predicted_value=-174389.86 + [new_sqft_living] * 119.17 + [new_grade] * 57352.79.

EXECUTE.

What to report in a publication

In a publication (and in the home assignment) you will need to report the following information about most types of regression analysis:

First of all, you will have to specify the regression model you built. For example:

"In a linar regression model we predicted housing price (in USD) with square footage of living area (in ft) and King County housing grade as predictors."

Next you will have to indicate the effectiveness of the model. You can do this by after a text summary of the results, giving information about the F-test of the whole model listed in the ANOVA table of the output, specifically, the F value, the degrees of freedom, and the p-value. Note that there are two degrees of freedom for the F test. You will need to provide the df listed in the "regression" and the "residual" lines within the ANOVA table. Also provide information about the model fit using the adjusted R squared from the Model Summary table.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Selection Criteria: Akaike Information Criterion	Amemiya Prediction Criterion	Mallows' Prediction Criterion	Schwarz Bayesian Criterion
1	,598[a]	,358	,352	170071,376	4820,567	,662	3,000	4830,462

a. Predictors: (Constant), grade, sqft_living

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3177917994185,416	2	1588958997092,708	54,935	,000[b]
	Residual	5698081797844,145	197	28924273085,503		
	Total	8875999792029,560	199			

a. Dependent Variable: price

b. Predictors: (Constant), grade, sqft_living

Don't forget to use APA guidelines when determining how to report these statistics and how many decimal places to report (2 decimals for every number except for p values, which should be reported up to 3 decimals).

"The multiple regression model was significantly better than the null model, explaining 35.15% of the variance in housing price (F (2, 197) = 54.98, p < .001, Adj. R^2 = 0.35)."

Furthermore, you will have to provide information about the regression equation and the predictors' added value to the model. You can do this by creating a table with the following information:

Regression coefficients with confidence intervals, and standardized beta values for each predictor, together with the p-values of the t-test. You can get all this information from the Coefficients table:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-174389,862	95255,171		-1,831	,069	-362240,588	13460,864
	sqft_living	119,173	24,762	,374	4,813	,000	70,341	168,005
	grade	57352,786	16052,790	,278	3,573	,000	25695,416	89010,156

a. Dependent Variable: price

The final table should look something like this:

Table 1. Regression coefficitions

	b	95% CI lb	95% CI ub	Std.Beta	p-value
Intercept	-174389,862	-362240,588	13460,864		,069
sqft_living	119,173	70,341	168,005	,374	,000
grade	57352,786	25695,416	89010,156	,278	,000

Interpreting the output

Interpretation of the regression coefficients

The interpretation of the regression coefficients of the predictors is: this is the amount by which the outcome variable's estimate would change if the predictor's value is increased by 1.

In our example the regression coefficient linked to sqft_living is 119,173. This means that an increas in the area of the apartment by 1 sqft results in an increased estimate in the price of the apartment by 119,173.

Interpretation of the estimate of the intercept

The coefficient of the intercept is a contant (different for each regression model) that is not dependent on the values of the predictors. It can be interpreted as if all the predictors in the model would have the value of zero (0), this would be the estimated value for the outcome. (Be careful that this often does not represent a true physical reality, if a 0 predictor value is meaningless, nevertheless, the mathematical interpretation stays the same.)

Interpretation of the standard beta

The benefit of the regression coefficient is that it is on the same metric as the predicted variable, making the influence of each predictor easy to interpret. However, we need to realize that this value is also dependent on the scale of the predictor. This makes it hard to directly compare the influence of predictors that use different scales just using the regression coefficient.

In order to be able to directly compare the predictive value contained by each predictor in the model, we can use the standardized Beta coefficient. This value is computed by refitting the model with the standardized predictors. Using the standardized Beta coefficient we can directly compare the predictive value of predictors within the context of the whole model. It is important to note that the predictive value of any given predictor in a model might be very different from its individual correlation with the outcome. This is because multiple predictors can explain the same portion of the variance, this way, the predictive value of any single predictor can be "masked" in the model by the predictive value of other predictors explaining the same portion of the variance.

Practice:

Experiment with different models based on your theories about what could influence housing prices.

Try to increase the adjusted R^2 above 52%. If you want to get access to the whole dataset or get ideas on which model works best, go to Kaggle, check out the top kernels, and download the data. https://www.kaggle.com/harlfoxem/housesalesprediction/activity