Exercise 14 - Model diagnostics

Zoltan Kekecs

8 May 2020

Table of Contents

Abstract 2
Data management and descriptive statistics 2
Load data about housing prices in King County, USA 2
Check the dataset 2
Model diagnostics 2
Build the model 3
Dealing with outliers 3
Identifying extreme cases 3
Identifying extreme cases with high leverage (Influential cases) 4
Assumptions of linear regression 6
Normality 7
What happens if the assumption of normality is violated? 9
What to do if the assumption of normality if violated? 9
Bootstrapping 10
Linearity 12
What happens if the assumption of linearity is violated? 13
What to do if the assumption of linearity if violated? 13
Homoscedasticty 14
What happens if the assumption of homoscedasticity is violated? 15
What to do if the assumption of homoscedasticity if violated? 15
No multicollinearity 18
What happens in the case of multicollinearity? 19
What to do in the case of multicollinearity? 19

Abstract

This exercise will show you how to check whether the assumptions of linear regression hold true for your model, what is the consenquence of the violation of the assumptions, and what to do in case of an assumption being violated.

Data management and descriptive statistics

Load data about housing prices in King County, USA

In this exercise we will use the same housing price dataset as in exercises 12.

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

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

You can download the dataset from github:

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

Check the dataset

As usual, you should always check the dataset for coding errors or data that does not make sense, by eyeballing the data through the data view tool, checking descriptive statistics and through data visualization.

Model diagnostics

Whenever we arrive at a model that we use for statistical inference, we need to make sure that the assumptions of the linear regression hold true for the model we are working with.

Thus, you have to do model diagnostics for all important models in your analyses. Usually we would do model diagnostics on the final model we arrive at in the end of model selection.

In some cases, if you do result-based post hoc model selection, it might be important to run model diagnostics on the interim steps as well, the models based on which you reached your final model composition.

If you make any adjustments to the data, or the model based on the model diagnostics, remember that you have to re-run model diagnostics on the new model/new dataset.

Build the model

We first build a linear regression model to predict the price of the apartment by using only sqft_living and grade as predictors. We are going to run model diagnostics on this model. We can do this in **Analyze > Regression > Linear** as usual. However, we will ask for some additional statistics and saved values. In the save menu ask for Cook's distance, unstandardized predicted values, and unstandardized residuals to be saved. We should also ask in the Plots menu for 1) a plot where ZRESID (standardized residuals) is on the Y axis and ZPRED (standardized predicted values) is on the X axis, and 2) a Normal probablility plot. In the Statistics menu we should ask for confidence intervals, covariance matrix, and collinearity diagnostics.

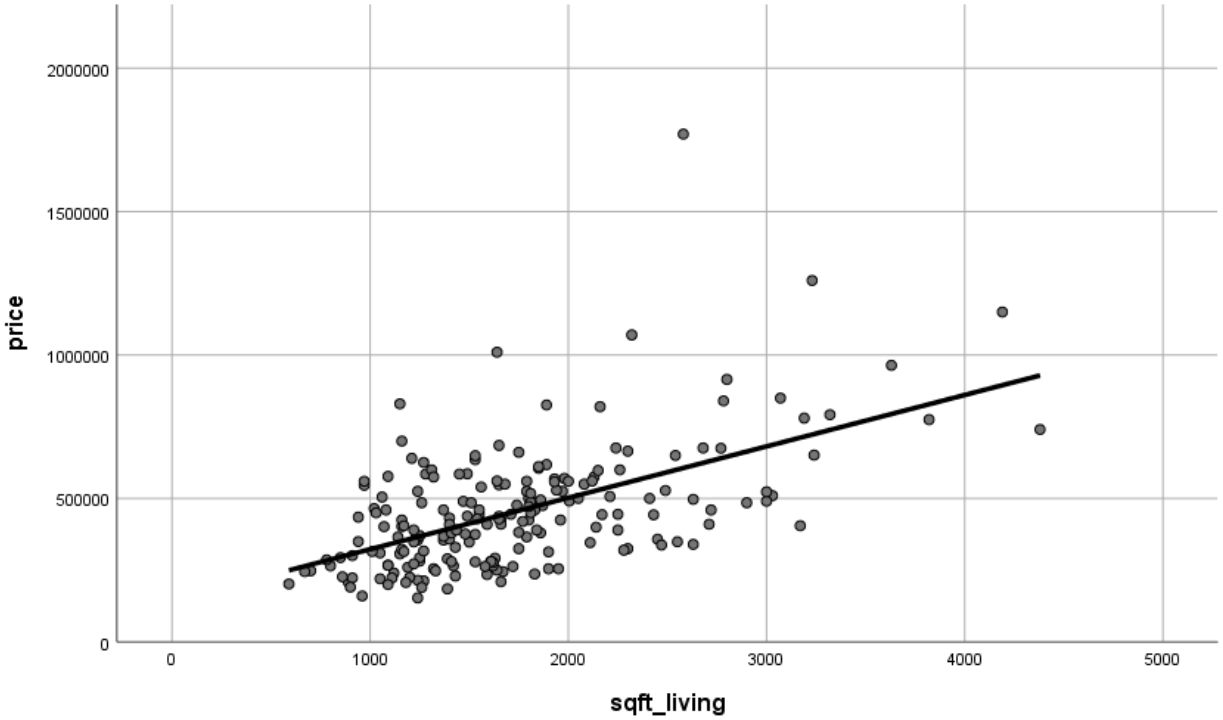
```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) BCOV R ANOVA COLLIN TOL
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT price
  /METHOD=ENTER sqft_living grade
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS NORMPROB(ZRESID)
  /SAVE PRED COOK RESID.
```

Dealing with outliers

Identifying extreme cases

Cases with extreme values can be identified by visualizing the data and the relationship of the outcome and the predictors one by one.

For example here we visualize price and sqarefootage data on a scatterplot.



Notice that most apartment sales had a final sales price below 1 million USD. However, there were a few cases where the price was higher than that. These might be considered extreme cases, especially the one with the sales price of 1.7 million USD.

However, it might not be a huge problem if we have a few of these cases in the dataset if we have a lot of data to counteract their effects, especially if they don't have high leverage.

Identifying extreme cases with high leverage (Influential cases)

Outlier cases that are extreme on both the Y (outliers in terms of the outcome variable) and the X axes (high leverage cases) have a great influence on the regression fit (regression line). Highly influential cases can be identified by looking at the scatterplot with the regression line, and by using Cook's distance. Cases that are close to the middle of the regression line on the scatterplot have less leverage while cases at the ends have more. Cases that have high residual error and high leverage are influential cases. Cook's distance is a number that quantifies this, taking into account the combination of "extremeness" and leverage.

Summary statistics (min and max value) of Cook's distance can be found in the Residual statistics table of the regression output. Here we can see that the highest Cook's distance is 0.28.

'Residuals Statistics[a]

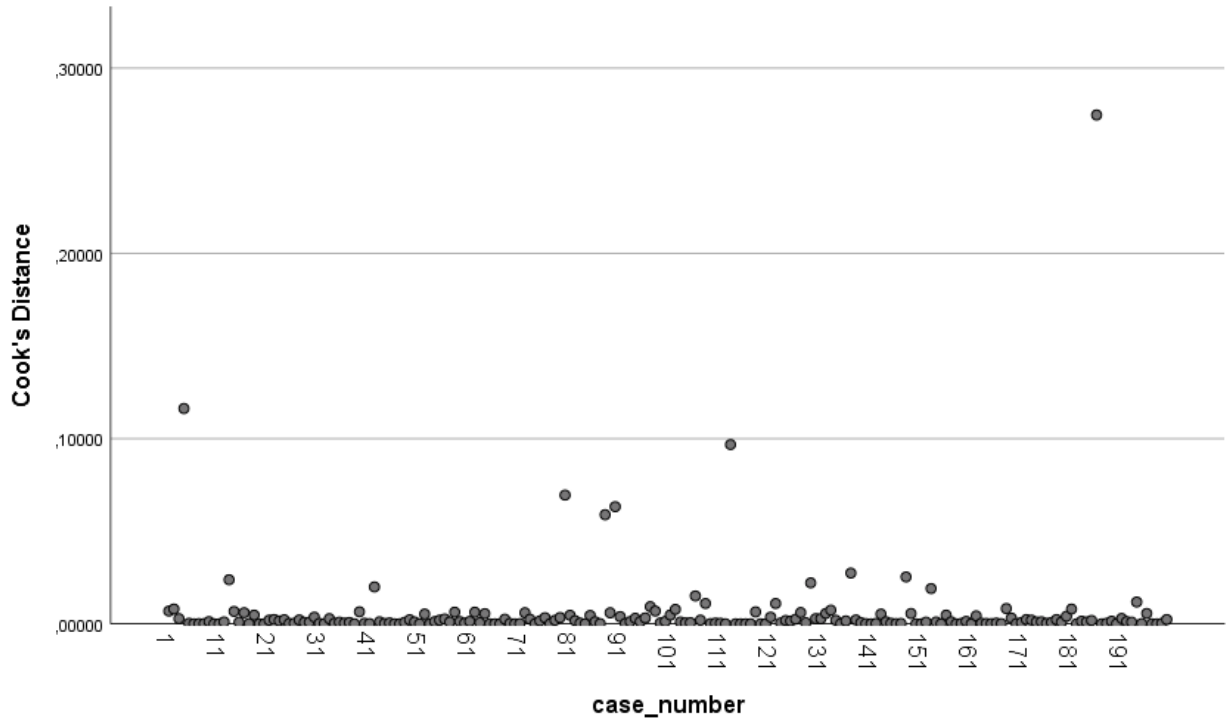
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	205329,17	898473,81	453610,89	126370,238	200
Std. Predicted Value	-1,965	3,520	,000	1,000	200

Standard Error of Predicted Value	12754,943	60397,676	19540,245	7232,182	200
Adjusted Predicted Value	202667,45	883399,56	453467,91	126244,219	200
Residual	-371917,094	1120747,875	,000	169214,588	200
Std. Residual	-2,187	6,590	,000	,995	200
Stud. Residual	-2,232	6,651	,000	1,004	200
Deleted Residual	-387481,125	1141626,750	142,980	172226,656	200
Stud. Deleted Residual	-2,255	7,534	,007	1,041	200
Mahal. Distance	,124	24,103	1,990	2,782	200
Cook's Distance	,000	,275	,006	,023	200
Centered Leverage Value	,001	,121	,010	,014	200

a. Dependent Variable: price

We asked for Cook's distance to be saved, so now we can also use the Chart builder to produce a scatterplot with the Cook's distance on the Y axis and the case number (ID) on the X axis.

```
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=case_number COO_1
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: case_number=col(source(s), name("case_number"), unit.category())
  DATA: COO_1=col(source(s), name("COO_1"))
  GUIDE: axis(dim(1), label("case_number"))
  GUIDE: axis(dim(2), label("Cook's Distance"))
  SCALE: linear(dim(2), include(0))
  ELEMENT: point(position(case_number*COO_1))
END GPL.
```



It can vary from based on the data and the statistical question what should be considered a problematic case, but there are some rules of thumb. Some say that cases with Cook's distance > 1 are influential, while some use a threshold of Cook's distance > 4/N. (In our case this is 4/200 = 0.02). Again others say that cases with very different Cook's distance than most cases in the dataset are potentially problematic.

We have no cases that would have a Cook's distance higher than 1, but we have several where it is higher than 0.02. So we might have a problem with outliers according to this criteria. We can find these cases by sorting the dataset based on Cook's distance.

Lets test whether the assumptions of multiple regression hold true, and determine if we need to do anything with these cases.

Assumptions of linear regression

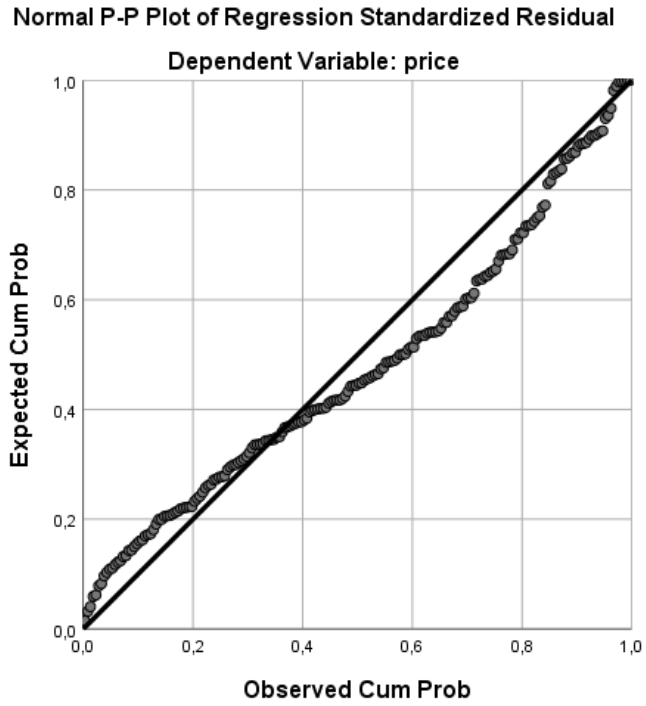
- **Normality**: The residuals of the model must be normally distributed
- **Linearity**: Ther relationship between the outcome variable and the predictor(s) must be linear
- **Homoscedasticity**: The variance of the residuals are similar at all values of the predictor(s)
- **No Multicollinearity**: None of the predictors can be linearly determined by the other predictor(s)

Normality

The residuals of the model must be normally distributed.

Note that here we talk about the distribution of the prediction error (residulas) of the model, and not the distribution of the individual predictors or the outcome.

We can check this assumption by looking at the Normal PP plot of the standardized residuals that we requested in the plots menu when we built the regression model, and checking if the cases are alligned wit the theoretical diagonal. If the cases divert from the diagonal (the line in the middle of the plot), the assumption of normality might be violated.

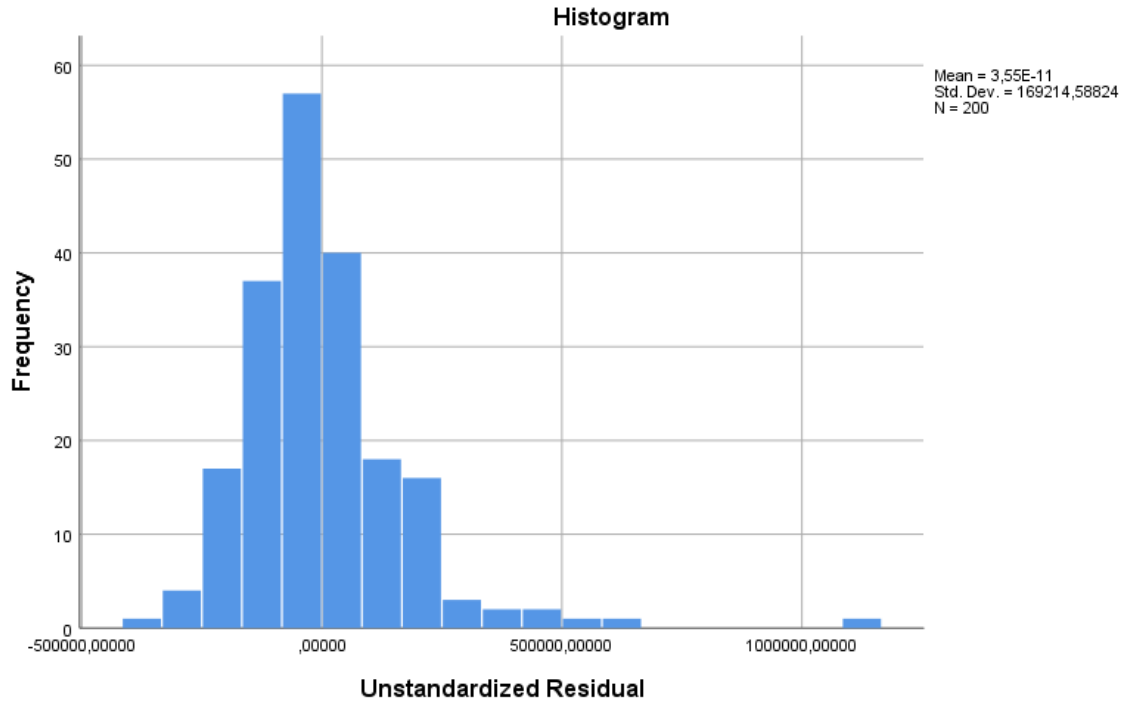


We should also go to **Analyze > Descriptive Statistics > Explore**, enter the Unstandardized residuals (which we asked for to be saved in the Save menu of the regression) as the dependent variable and in the Plots menu ask for normality plots and tests and a histogram. We should look at the histogram of the unstandardized residuals. We should see a roughly bell shaped (normal) distribution if the assumption of normality is met. Also in the Explore output, Skew and kurtosis > 1 can indicate the violation of the assumption of normality.

Descriptives

		Statistic	Std. Error
Unstandardized Residual	Mean	,0000000	11965,27828209

	95% Confidence Interval for Mean	Lower Bound	-23595,0088122	
		Upper Bound	23595,0088122	
	5% Trimmed Mean		-11761,3842990	
	Median		-23118,9569198	
	Variance		28633576873,589	
	Std. Deviation		169214,588241 05	
	Minimum		-371917,09909	
	Maximum		1,12075E+6	
	Range		1492664,98474	
	Interquartile Range		168569,41573	
	Skewness		2,085	,172
	Kurtosis		10,139	,342



```
EXAMINE VARIABLES=RES_1
 /PLOT BOXPLOT HISTOGRAM NPPLOT
 /COMPARE GROUPS
 /STATISTICS DESCRIPTIVES
 /CINTERVAL 95
 /MISSING LISTWISE
 /NOTOTAL.
```

Notice that the plots and the statistics indicate a deviation from the assumption of normality mainly driven by a few irregular cases.

What happens if the assumption of normality is violated?

The estimates and confidence intervals might be less accurate if the assumption of normality is violated. The effect of this depend on the sample size as well. For large sample sizes (N>500) the effect is very minor, while for smaller samples (N around 100) the effect is greater. For example in the study of Lumley, Diehr, Emerson and Chen (2002), the effect of extreme violation of normality (skewness = 8.8, kurtosis = 131) was that the 95% confidence intervals contained the true population mean in only 93.6% of cases instead 95% when N was 500, and 91.3% of the cases when N was 65.

The bottom line is that if the assumption of normality is violated, the confidence intervals and the p values are less reliable, but they are still informative.

Reference:

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. Annual review of public health, 23(1), 151-169.

What to do if the assumption of normality if violated?

1. you can interpret the results more cauciously, such as using 99% CI instead of 95% CI, or $p < 0.01$ as a threshold for significance (You can adjust this in the Statistics menu)

2. if the deviation from normality is driven by a few cases, you might want to try to exclude the outliers. If you are doing formal hypothesis testing, exclusion of variables should not be based on the p-value. Rules for exclusion can be pre-registered, or a sensitivity analysis can be presented, meaning that you conduct the same analysis twice on data including and excluding the problematic cases, and compare the results of the two analysis to see the influence of exclusion of outliers.

3. you can try transforming the predictors and/or the outcome variable to achieve more normally distributed residuals. But if you do this, be mindful when you interpret the

unstandardized coefficients that they will be in the transformed units. The same goes for the error terms, so RSS of a model on transformed variables cannot be copared with that of one with untransformed variables.

4. you can use bootstrapping to get a robust estimate of the confidence intervals.

In our case, since non-normality is the result of a few extreme cases, we can build a model with these cases excluded to see if this fixes the problem. We exclude cases 186 and 113 here because they showed up as extreme cases according to the Cook's distance, and they also showed up as the biggest outliers on the box-plot of the unstandardized residuals in the Explore output.

Here we refit the model without these cases two cases and recheck the assumption of normality. Notice that the residuals look much more normal in the model witout these outliers.

When we compare the two models in a sensitivity analysis, the exclusion of the outlier did not change the statistical inference, the previously significant predictors are still significant, and the overall model F test is significant in both cases. The adjusted R^2 improved considerably because our regression line now fits the remaining data much better. However, model fit should be assessed on new data or a test set as well to get a more accurate picture of true prediction efficiency, because similar outliers might be present in new data as well, and our model will not be very good at predicting these cases.

We are going to work with the new model without the outliers in the upcoming tests of the assumptions.

Bootstrapping

[If you are running SPSS 25, bootstrapping might not work for you. This is a known issue by IBM SPSS developers. This public interim fix might fix it for you:

https://www-945.ibm.com/support/fixcentral/swg/selectFixes?product=ibm%2FInformation+Management%2FSPSS+Statistics&fixids=25.0-IM-S25STATC-ALL-IF003&source=SAR&function=fixId&parent=SPSS

]

In case you needed a robust esitmate of the confidence intervals, you can use bootstrapping to get them. Bootstrapping means that we randomly sample from our own observations and rebuild the model on this random sample. We repeat this many times, and drow conclusions about the confidence limits based on this multitude of results.

You can ask for a bootstrap sampled regression in the Bootsrap menu in the regression tab. Be sure that nothing is checked in the Save menu, because the bootstrapping will run the regression thousands of times, and it will not start until nothing is saved.

Compare the regular and bottstrapped confidence intervals of the model coefficients.

```
BOOTSTRAP
 /SAMPLING METHOD=SIMPLE
 /VARIABLES TARGET=price INPUT=  sqft_living grade
 /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
 /MISSING USERMISSING=EXCLUDE.
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS CI(95) R ANOVA COLLIN TOL
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT price
 /METHOD=ENTER sqft_living grade
 /SCATTERPLOT=(*ZRESID ,*ZPRED)
 /RESIDUALS NORMPROB(ZRESID).
```

Bootstrap for Coefficients

Model		B	Bootstrap[a] Bias	Std. Error	Sig. (2-tailed)	95% Confidence Interval Lower	Upper
1	(Constant)	-164126,218	-447,163	79392,530	,044	-332386,077	-10033,490
	sqft_living	108,949	-1,052	21,083	,001	69,196	150,549
	grade	57141,257	277,091	12553,646	,001	32599,859	82810,153

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

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	Collinearity Statistics Tolerance	Collinearity Statistics VIF

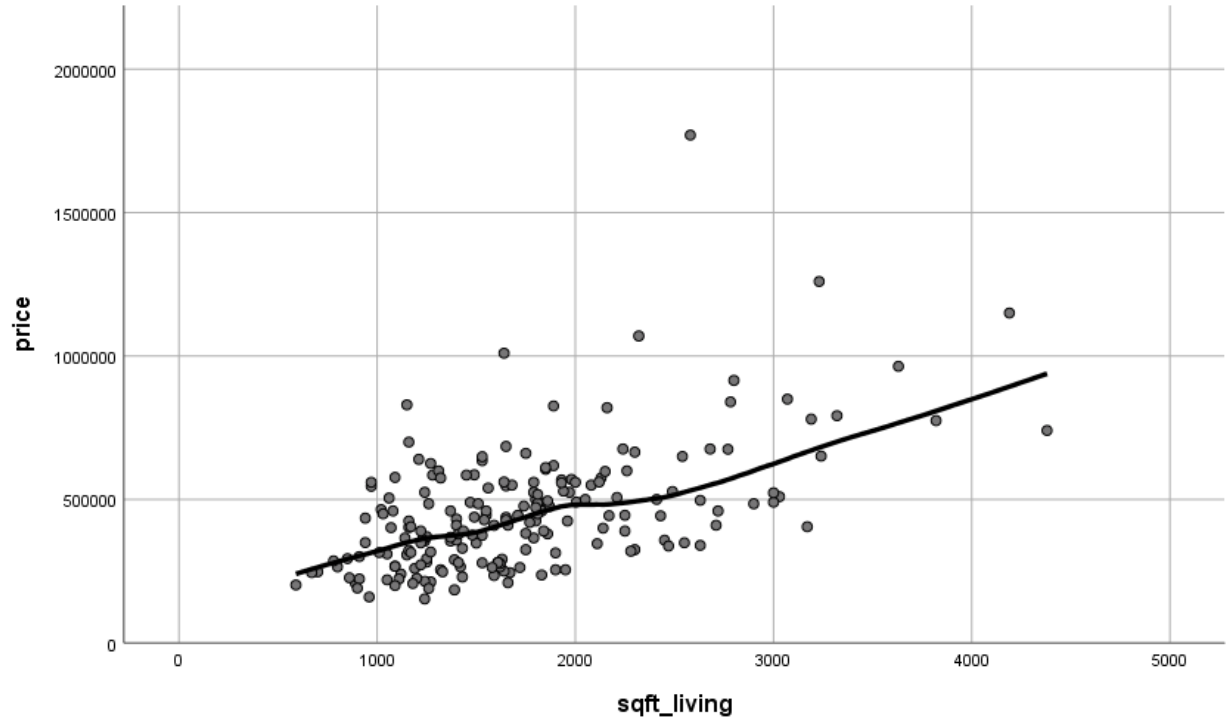
1	(Constant)	-164126,218	81165,631		-2,022	,045	-324201,403	-4051,032		
	sqft_living	108,949	20,918	,389	5,208	,000	67,695	150,203	,541	1,849
	grade	57141,257	13653,242	,312	4,185	,000	30214,278	84068,235	,541	1,849

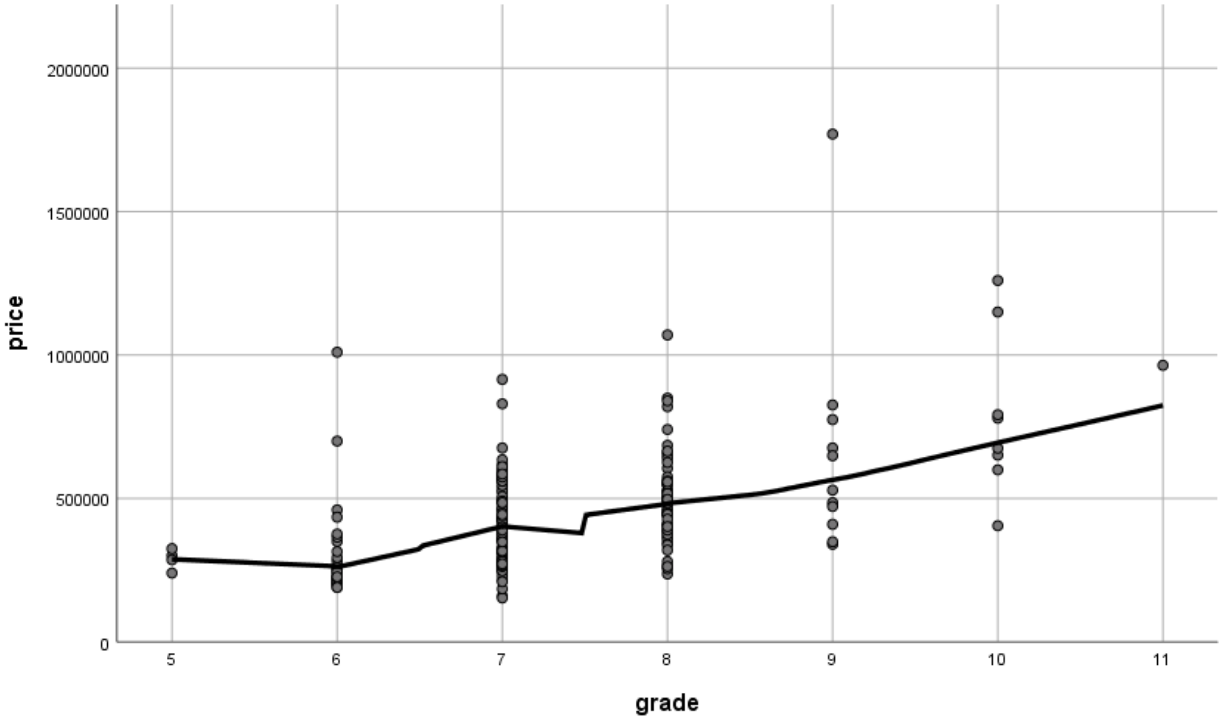
a. Dependent Variable: price

Linearity

The relationship between the outcome variable and the predictor(s) must be linear.

In the residual-predicted value plot that we asked for in the regression output, we should see a roughly linear relationship between the predictions and the residuals (the regression line we fit on this scatterplot should be flat and linear). We can also look at the scatterplots of the dependent and the predictor variables one-by-one, and we should also see a linear relationship there. We can fit a line on these scatterplots by double clicking the plot, bringing up the elements menu, and clicking on Fit Line at Total. Here we can select loess, so we can see a potential non-linear relationship. Small deviations from linearity are fine.





What happens if the assumption of linearity is violated?

If there is nonlinear relationships among the outcome and the predictors, the predictions of the model may be off, resulting in lower prediction accuracy. Also, the model coefficients will be also unreliable when used for prediction, and even though the standardized coefficients and the t-test p-values for some predictors would indicate no significant added predictive value for a predictor, in reality the predictor might still be related to the outcome, just in a non-linear way.

What to do if the assumption of linearity if violated?

If the assumption of linearity is violated, you can try to account for the non-linearity in the relationship by making your model more flexible.

1. you can include a higher order term of the predictor that seems to have a non-linear relationship with the outcome. (See exercise 13 on how to do this.) Usually, including the second and third order term should be enough to address curved relationship. You should avoid adding too much flexibility to reduce overfitting.

2. If higher order terms do not capture the relationship well, you might have to experiment with non-linear regression. You can learn more about nonlinear regression for example from this book: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: springer. It is freely available at: http://www-bcf.usc.edu/~gareth/ISL/

Homoscedasticty

In regression we assume that the standard deviations of the error terms (residuals) are constant and do not depend on the value of the predictors. So for example the variance of the residuals should be the same for 600 sqft apartments and 3000 sqft apartments.

We can check this assumption by looking at the plot of the standardized residuals and the predicted values, where we should see roughly equal variation at all values of the predicted values. There are also good statistical tests to check this assumption. We can perform the Breush-Pagan test in SPSS. In order to do this, we would need to create a new variable, containing the sqared values of the unstandardized residuals. (This can be done in the **Transform > Compute variable** tab, where we can compute the squared of the unstandardized resudials which we have saved during the regression analysis as a new variable.) Once we have this new variable, we can build a new regression model identical to the previous one, but instead of price as a dependent variable, we will enter the squared residuals as a dependent variable. The predictors should remain the same.

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

We need to look at the ANOVA F-test result. If it is significant, we have significant heteroscedasticity. This is the situation in our case as well:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1345921149669 6224000000,00 0	2	6729605748348 112000000,000	5,438	,005[b]
	Residual	2413195475049 41300000000,0 00	195	1237536141050 981000000,000		

	Total	2547787590016 37540000000,000	197			

a. Dependent Variable: Residuals_sqared

b. Predictors: (Constant), grade, sqft_living

What happens if the assumption of homoscedasticity is violated?

If there is heteroscedasticity, the prediction can be off, meaning that the model overall would be less efficient in predicting the outcome. This being said, we can still use the model for prediction, it will be just less good predicting new data.

More importantly, the model coefficients and their confidence intervals are no longer accurate.

So if we are interested in predicting the outcome, we could still do that with some success even if there is heteroscedasticity. However, interpreting the individual coefficients and our confidence in their added predictive value is no longer possible.

What to do if the assumption of homoscedasticity if violated?

We have a couple of options to deal with heteroscedasticity.

1. **Transformation**. If we are mainly interested in more accurate predictions, we can do some transformation on the outcome and/or the predictors to make them more normally distributed, and thus, homogenize the variability in some parts of the dataset. For example in the example below we apply log() transformation to the outcome, price, in the **Transform > Compute variables** tab.
   We can now refit the regression model with this now normalized variable as a dependent variable.
   Now we can run the Breush-Pagan test again (remember that you will have to save the residuals and compute their sqared value again). This now does not show a significant violation of homoscedasticity.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,001	2	,000	,481	,619[b]
	Residual	,125	195	,001		
	Total	,126	197			

a. Dependent Variable: RES_sq_afterlog

b. Predictors: (Constant), grade, sqft_living

However in this case, we have to keep in mind when we interpret the model coefficient and the predicted (fitted) values that we are no longer predicting price, we the predictions are about log(price). So the predicted values will have to be transformed back with the exponential function to be meaningful again, and the model coefficents now also have to be interpreted by keeping in mind that they refer to changes in log(price).

2. **Using robust estimators**. If it is important to keep the outcome on its original scale to make the interpretation of the model coefficients more intuitive, we can use robust estimators to establish the Heteroscedasticity-consistent (HC) standard errors and use them to compute confidence intervals, and corrected p-values of the model coefficients. These new statistics are robust to the violation of homoscedasticity, so we call them robust estimates. In the example below we use the Huber-White Sandwich Estimator is used to get these robust values. We can get this robust estimation of the coefficients if we run Generalized Linear Models instead of the Linear Regression. The Generalized linear model can be run at **Analyze > Generalized linear model > Generalized linear model**.

   There we need to set up our analysis with the price as the dependent variable and sqft_living and grade as the covariates (because they are continuous variables). If we had some categorical predictors, we would need to enter them as factors. In the model tab you will have to specify which elements we want to include in the model (if we want to include interactions, here is the place to do this). Most importantly, in the Estimation tab in the Covariance matrox box we should select Robust estimators, which will use the Huber-White estimator.

   (For small samples (N around 50) you might have to use some other method to correct the stndard error, for example Bell-McCaffrey estimates. For more details see the paper by Imbens and Kolesar (2016), and their R function at: https://github.com/kolesarm/Robust-Small-Sample-Standard-Errors) Reference: Imbens, G. W., & Kolesar, M. (2016). Robust standard errors in small samples: Some practical advice. Review of Economics and Statistics, 98(4), 701-712.

```
GENLIN price WITH sqft_living grade
  /MODEL sqft_living grade INTERCEPT=YES
 DISTRIBUTION=NORMAL LINK=IDENTITY
  /CRITERIA SCALE=MLE COVB=ROBUST PCONVERGE=1E-006(ABSOLUTE)
SINGULAR=1E-012 ANALYSISTYPE=3(WALD)
   CILEVEL=95 CITYPE=WALD LIKELIHOOD=FULL
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

The confidence intervals returned by this model is not that different from the original model. Note that the model coefficients art the same between the two models, the robust estimator only affected the CIs.

Original model:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-164126,218	81165,631		-2,022	,045	-324201,403	-4051,032
	sqft_living	108,949	20,918	,389	5,208	,000	67,695	150,203
	grade	57141,257	13653,242	,312	4,185	,000	30214,278	84068,235

a. Dependent Variable: price

Robust estimator with generalized linear model:

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-164126,218	81784,5648	-324421,019	-3831,416	4,027	1	,045
sqft_living	108,949	21,2242	67,350	150,548	26,350	1	,000
grade	57141,257	13001,0546	31659,658	82622,855	19,317	1	,000
(Scale)	20207041441,115[a]	2030884078,2830	16594102111,757	24606605470,604			

Dependent Variable: price

Model: (Intercept), sqft_living, grade

a. Maximum likelihood estimate.

3. **Separate models for different parts of the dataset**. Another approach you might take is to visually inspect the plots of the relationships, to see if you can identify clusters with different variance. For example, in our case you might notice that the variances start to be larger at around 2000sqft apartment size.

We can create separate datasets for the part of the dataset containing 2000sqft or smaller apartments, and apartments larger than that, and fit separate models on these.

No multicollinearity

We assume that none of the predictors can be linearly determined by the other predictor(s). This usually mean that the predictors should not be too highly correlated with each other, but there are some more subtle cases where even though the predictors are not super highly correlated pair-by-pair, but a combination of multiple predictors can still accurately determine another predictor.

So to evaluate whether this assumption holds true, we compute the variance inflation factor (VIF) by looking at the VIF values displayed in the Coefficients table in the regression output. (This is included in this table because we asked for the **collinearity diagnostics** in the **Statistics menu of the Regression analysis tab**.)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-164126,218	81165,631		-2,022	,045	-324201,403	-4051,032		
	sqft_living	108,949	20,918	,389	5,208	,000	67,695	150,203	,541	1,849
	grade	57141,257	13653,242	,312	4,185	,000	30214,278	84068,235	,541	1,849

a. Dependent Variable: price

There is no real consensus about what vif values indicates problematic multicollinearity, some advocate for vif values of 10 or larger are problematic (for example Montgomery and Peck ,1992). However, a more conservative approach is to explore and treat multicollinearity if VIF is above 3 (an approach suggested by Zuur, Ieno, and Elphick, 2010).

In the Lab report, use the VIF threshold of 3 to identify problematic multicollinearity!!!

Reference: Montgomery, D.C. & Peck, E.A. (1992) Introduction to Linear Regression Analysis. Wiley, New York. Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data

exploration to avoid common statistical problems. Methods in ecology and evolution, 1(1), 3-14.

In the example above multicollinearity is not problematic.

What happens in the case of multicollinearity?

In regression we would often like to know the main effect of each predictor, and whether this effect (or in other words, whether the added predictive value of a certain predictor) is statistically significant or not. The model coefficient indicates the size and direction of the effect of increasing the value of the predictor by one point, when all other predictors are held constant. However, if correlation is high among the predictors, there are fewer cases in the dataset where one predictor has high values, while the others are low. This makes it difficult for the model to estimate the independent effect of the predictors that are highly correlated.

The bottom line is, that in the case of multicollinearity, the model coefficients and the t-test of their individual predictive value will become less reliable. Futhermore, the model coefficients will be very unstable, and will show big changes and even sign changes (a positive effect would turn into a negative effect) if we change the model a bit, such as by adding or removing a new predictor.

Fortunately, multicollinearity does not affect prediction accuracy. So if prediction accuracy is our only concern, we might not have to deal with multicollinearity. But if we are interested in the coefficients as well and drawing inference about the individual predictors and their effect and predictive value, we have to address it.

What to do in the case of multicollinearity?

There are two types of multicollinearity:

Structural multicollinearity, where multicollinearity is a result of entering a predictor into the model that is derived from one or more of the other predictors. For example higher order terms (e.g. grade and grade^2), or interaction terms (e.g. long*lat).

Data multicollinearity, where multicollinearity is present in the data itself rather than being an artifact of our model.

To demonstrating the two types of multicollinearity and how to address them, we will build some new models.

Example of handling structural multicollinearity

First, lets build a model where in addition to the sqft_living and grade, we also enter geolocation information into the podel as predictors.

In the first model we only add the main effect of these predictors (no interaction term). Notice that longitude has a has a negative coefficient indicating that further east we go the lower the price of the apartment which makes sense if you look at the map of King County with Seattle being in the west of the county. Latitude has a positive coefficient, indicating

that the further north we go the higher the price. The added predicitive value of latitude is significant in this model. Also notice that there is no multicollinearity in the model according to vif.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-34784848,888	6671700,337		-5,214	,000	-47943654,794	-21626042,982		
	sqft_living	119,984	17,829	,428	6,730	,000	84,819	155,149	,531	1,885
	grade	52667,738	11563,650	,288	4,555	,000	29860,386	75475,090	,537	1,861
	lat	534994,869	62567,537	,401	8,551	,000	411590,935	658398,804	,976	1,025
	long	-75175,016	52622,316	-,068	-1,429	,155	-178963,677	28613,644	,960	1,041

a. Dependent Variable: price

Now, we add an interaction term as well (see exercise 13 for how to include an interaction term in linear regression models), where we include the interaction of longitude and latitude in the model.

Notice that in this case, SPSS automatically excludes one of the predictors from the model because of high multicollinearity (in this case, latitude is excluded).

Excluded Variables[a]

Model		Beta In	T	Sig.	Partial Correlation	Collinearity Statistics: Tolerance	Collinearity Statistics: VIF	Collinearity Statistics: Minimum Tolerance
1	lat	-78,554[b]	-1,876	,062	-,134	1,208E-6	827573,498	8,883E-7

a. Dependent Variable: price

b. Predictors in the Model: (Constant), INT_lat_long, grade, long, sqft_living

Notice that coefficients and the signs undervent a violent shift. Longitude now has a positive coefficient.

Structural multicollinearity in this case is due to the fact that the interaction term long*lat is derived from both long and lat predictors, so they are now highly correlated.

This can be fixed by standardizing the variables. A good method to use here is “centering”, that is, subtracting the mean of the variable from the values of the variable. By doing this, we can still preserv the original scale of the variables, and thus, the coefficients will mean the same thing as before centering. (We could use Z transformation as well, but that would change the interpretation of the coefficients, because the predictors would be transformed to a scale between -1 to +1.)

So lets center longitude and latitude, and re-build the model.

```
COMPUTE lat_cntr=lat - 47.1764.
EXECUTE.
COMPUTE long_cntr=long + 122.455.
EXECUTE.
COMPUTE INT_long_lat_cntr=lat_cntr * long_cntr.
```

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-165409,796	68698,127		-2,408	,017	-300909,740	-29909,852		
	sqft_living	120,629	17,716	,430	6,809	,000	85,686	155,571	,530	1,885
	grade	54356,799	11522,671	,297	4,717	,000	31629,524	77084,074	,534	1,872
	lat_cntr	516629,754	62915,396	,387	8,211	,000	392535,647	640723,861	,952	1,050

	long_cntr	-90333,447	52891,500	-,081	-1,708	,089	-194656,456	13989,563	,938	1,066
	INT_long_lat_cntr	-861819,768	456978,566	-,089	-1,886	,061	-1763162,687	39523,152	,946	1,057

a. Dependent Variable: price

Notice that mutlicollinearity is gone, and the coefficients are now back to their original signs and roughly the same effect size as before entering the interaction. Latitude has a significant added predictive value again. But also notice that the R^2 was not at all effected by the removal of multicollinearity! So the prediction efficientcy is the same, but the coefficients are more reliable now and their interpretation is more straightforward.

Example of handling data multicollinearity

Now lets build a model where in addition to the sqft_living and grade, we also use sqft_above as a predictor (the squarefootage of the area of the apartment above ground level).

The vif is now above 3.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-215231,978	80407,724		-2,677	,008	-373817,521	-56646,436		
	sqft_living	192,153	31,675	,685	6,066	,000	129,682	254,625	,223	4,474
	grade	68614,498	13706,009	,375	5,006	,000	41582,581	95646,415	,509	1,966
	sqft_above	-114,882	33,499	-,388	-3,429	,001	-180,950	-48,813	,223	4,494

a. Dependent Variable: price

We can explore the reason for this by looking at the correlation matrix in the regression output.

The correlation matrix clearly indicates that the correlation of sqft_living and sqft_above is very high.

Coefficient Correlations[a]

Model			sqft_above	grade	sqft_living
1	Correlations	sqft_above	1,000	-,244	-,766
		grade	-,244	1,000	-,236
		sqft_living	-,766	-,236	1,000
	Covariances	sqft_above	1122,173	-112071,652	-812,748
		grade	-112071,652	187854688,563	-102246,689
		sqft_living	-812,748	-102246,689	1003,310

a. Dependent Variable: price

Compare the coefficients in the model summary of the two models. What can you observe? Do the coefficients in the new model make sense?

Old model (without sqft_above):

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-164126,218	81165,631		-2,022	,045	-324201,403	-4051,032		
	sqft_living	108,949	20,918	,389	5,208	,000	67,695	150,203	,541	1,849
	grade	57141,257	13653,242	,312	4,185	,000	30214,278	84068,235	,541	1,849

a. Dependent Variable: price

New model:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-215231,978	80407,724		-2,677	,008	-373817,521	-56646,436		
	sqft_living	192,153	31,675	,685	6,066	,000	129,682	254,625	,223	4,474
	grade	68614,498	13706,009	,375	5,006	,000	41582,581	95646,415	,509	1,966
	sqft_above	-114,882	33,499	-,388	-3,429	,001	-180,950	-48,813	,223	4,494

a. Dependent Variable: price

Because of the multicollinearity, we cannot trust the coefficients or the t-tests of the predictors.

You have a number of options to deal with mutlicollinearity:

1. removing highly correlated predictors
2. linearly combining predictors, such as using the mean of the two values for each observation (e.g. for case 3, sqft_living is 1060, sqft_above is 960, so the mean for this case would be 1010). However, it is not clear how we would interpret this value in this particular case of predicting housing prices.
3. you could use some other statistical method entirely (e.g. partial least squares regression or principal components analysis)

Here I suggest using method 1., relaizing that there is hardly any difference in the two variables sqft_living and sqft_above, and when they are different it is due to the apartment having a basement. So we should just settle with using one of these variables, the one that makes more sense for the types of apartments the price of which we want to be able to predict. In this case I personally would keep sqft_living in the model, because I have a theory that the size of the living area is what ultimately influences the price, whether the living area is above or below ground, and the information about the apartment having a basement can be added by using the has_basement variable in a later model. However, those more familiar with the literature of predicting housing price might use previous research results to make this decision. Or the decision might be made based on practical

grounds, for example the fact that size of the living area is more accessable information than size of the area above ground, so we would be able to use the model practically in more cases for prediction, if we chose sqft_living as a predictor.

Notice that I did not use model comparison to make a decision about which varialbe to exclude! This decision, like all model selection decisions, should be made based on theoretical grounds, or based on previous research results or practical issues.

You can find some additional resources about multicollinearity on these links:

https://statisticalhorizons.com/multicollinearity

http://blog.minitab.com/blog/understanding-statistics/handling-multicollinearity-in-regression-analysis

http://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/