Tutorial 5

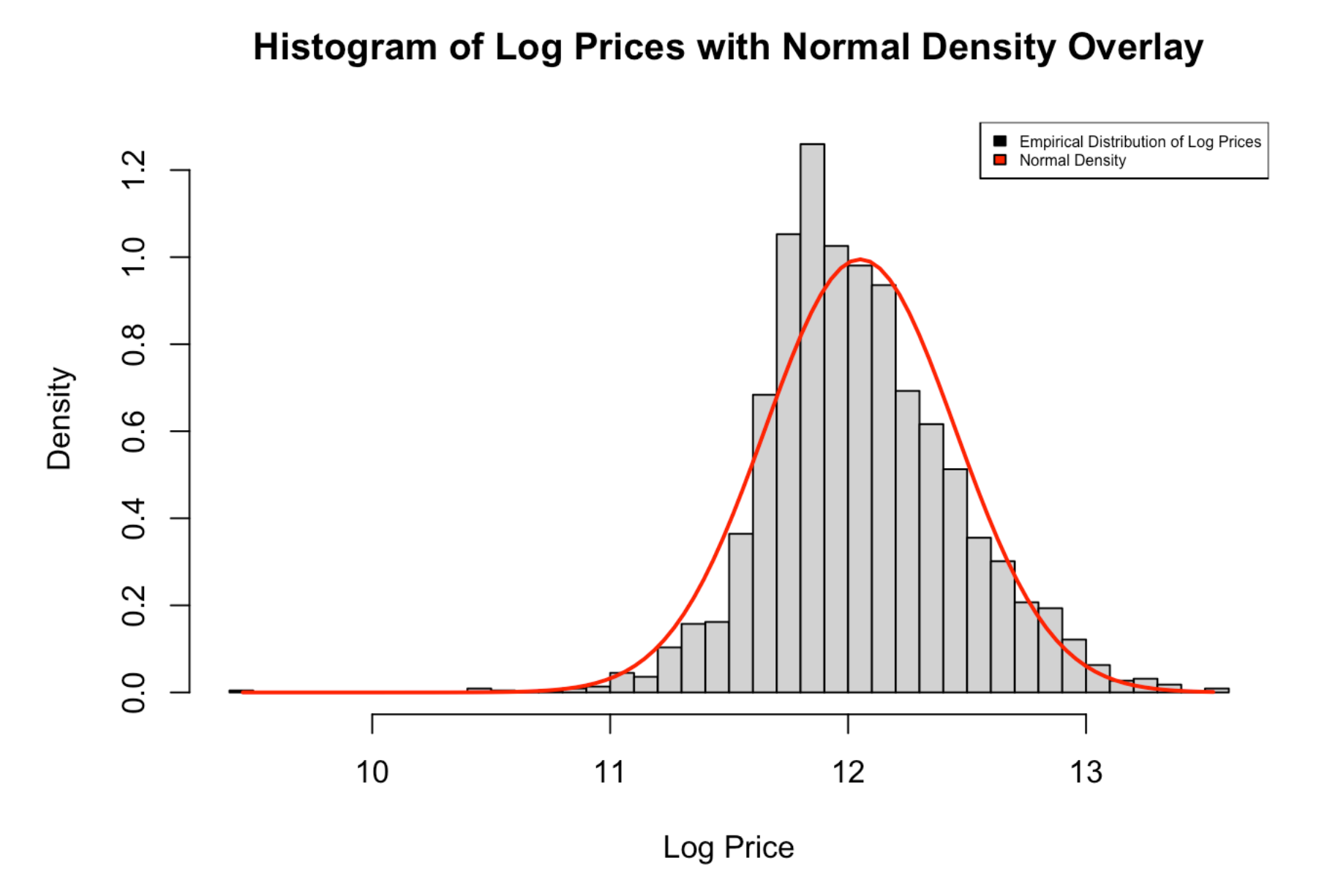
B158582

2023-11-18

Question 1: Produce a histogram showing the empirical distribution of log prices. Overlay the plot with a Normal density with mean equal to the sample average of log prices and variance equal to the sample variance of log prices. How can you use this plot to guide or validate the specification of the prediction model for price?

The statistical methods we have studied assume data is normally distributed. Hence, before applying any statistical methods, it should be confirmed whether the data is normally distributed or not. If the data was not normally distributed, it would be useful to apply a log transformation.

In fact, price was not normally distributed, so a log transformation was applied to fix the skewness, which yielded a much more normal distribution (as shown below).



Question 2: Perform Principal Components Analysis (PCA) on the numerical features in your data. Why focus only on the numerical features? How many Principal Components (PC) do you find? Why do you find exactly that number of PCs? What is the variance of the first PC, i.e. PC1? What is the variance of the last PC? What is the share of the total feature variance is captured by the first PC? Produce a scree plot of your PCA (for the plot, you may restrict attention to the first, say, 10 PCs). Briefly comment on the scree plot.

Why focus only on the numerical features?

PCA is a linear transformation method that finds the directions in which data variance is the highest. This is achieved by computing the covariance matrix, which requires numerical data. Categorical variables don't have a meaningful covariance and incorporating them can introduce noise into the analysis.

As PCA is sensitive to the units of the variables, variables must be scaled, which can be done for numerical variables.

How many Principal Components (PC) do you find? Why do you find exactly that number of PCs?

I found 26 PCs, which is the same amount of numerical features used for the PCA. This is because the PCA algorithm utilizes the following to identify the number of PCs:

$$\text{Number of PCs} = \begin{cases} n & \text{if } n \leq p, \\ p & \text{otherwise.} \end{cases}$$

where n is the number of observations in the data and p is the number of numerical features / regressors.

What is the variance of the first PC, i.e. PC1? What is the variance of the last PC? What is the share of the total feature variance is captured by the first PC?

If both are rounded to two decimal points, the variance of PC1 is 6.31 and the variance of PC26 is 0. This implies that PC26 carries virtually no information and is of little to no interest. In contrast, PC1 captures 24.3% of the total feature variance, indicating it could be a substantial determinant in housing prices.

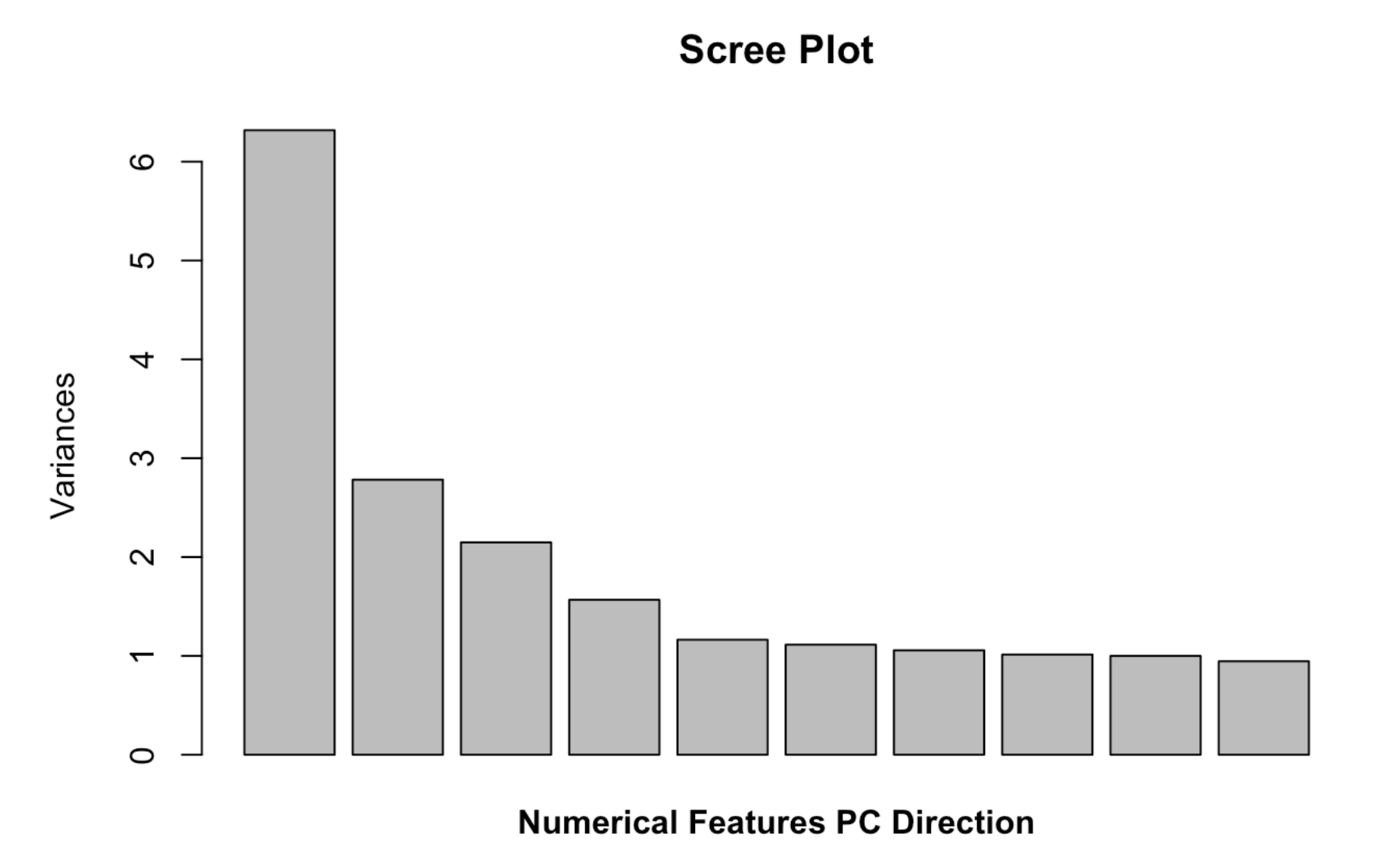
However, it must be noted that the PCs are capturing the variance within the numerical features, not across all variables (e.g. the non-numerical, categorical variables). Thus PC1 captures 24.3% of the variance across the 26 numerical features.

Briefly comment on the scree plot.

The scree plot shows PC1 has the largest variance and the first four PCs explain 49% of total variance, implying they could offer substantial information about the housing prices. But PC1 still stands out, as it carries more than double the variance than PC2, triple the variance of PC3, and quadruple the variance of PC4.

The first ten PCs explain 75% of total feature variance, but PC4 to PC10 are of lower importance (i.e. each approximately capture 4% of the total feature variance) and are unlikely to be the focus of the analysis.

Although it might be tempting to make a preliminary judgement about how many PCs are influential based on the scree plot, it is rarely clear how small the variance ought be before a PC is not worth tracking. This is especially true if some of the most complicated PCs are difficult to interpret and are required for forming an understanding of the PCs with the highest variance. So the scree plot is rather an exploratory tool.



Question 3: List the rotations for the first five PCs in your PCA. Provide a brief interpretation of each of the first three PCs based on the rotations.

As a note, the numbers in the brackets are loadings on the mentioned rotations.

Interpretation of PC1: Inverse measure of the overall size of the home

PC1 penalizes total area (-0.3), first and second floor area (-0.3 and -0.1 respectively), the total amount of rooms (-0.3), garage space (-0.3), and both lot frontage and area (both -0.2). All these are measures of the size of a home that are independent of quality. For instance, a home could be small (e.g. one floor and no garage), but have high-quality amenities. As such, I believe PC1 represents the overall size (e.g. square feet) of a home; a high/low PC1 means a small/large home.

It also makes sense that this would represent 24% of the total feature variance, as home size is a key concern for many buyers (i.e. it varies on the demand-side). Among other features, it will vary based on income, demographics (e.g. age or amount of children) or location preferences (e.g. younger people prefer living in the city). The size also varies across supply because of zoning regulations, geographic restrictions or historical legacies.

A high PC1 home is a small, one-floor apartment with no garage and no lot area/frontage.

Interpretation of PC2: The number of rooms and space available for living

PC2 rewards area (0.3), the amount of bedrooms (0.4) and rooms (0.4), second floor square footage (0.4), kitchen size (0.2), and the size of the enclosed porch (0.2). The common feature between these is that they are living areas: bedrooms and rooms for living are often placed on the second floor and the total area must be larger to fit these rooms.

In support, PC2 penalizes garages (-0.1). If PC2 would have been emphasizing space in general, it would not have a negative loading on garages. But garages are not living spaces, they are storage spaces, like basements (-0.1 on square footage).

It makes sense that PC2 would capture 11% of the total feature variance because there is variation on the demand-side (e.g. across income, amount of children, relationship status, location preferences etc) and supply-side (e.g. agencies offer bungalows, apartments, studios, family houses or large villas etc). It is likely smaller because size is more important than the amount of rooms - rooms often gain in usefulness with size.

A home high in PC2 would be a family home with two floors, numerous bedrooms, a large kitchen, and an enclosed porch.

Interpretation of PC3: Storage areas and non-living areas

PC3 rewards lot frontage and area (0.2 and 0.3 respectively), a finished basement (0.4), overall basement size (0.2), pool area size (0.2), porch size (0.2), and first floor size (0.3). These are mostly storage and non-living areas. The particularly large emphases on a finished basement and punishment for an unfinished basement (-0.2) reinforce the hypothesis that PC3 captures storage/non-living areas and outdoor "extras". For example, a bungalow (a one-floor house) with a basement, pool, and porch would score well in PC3.

It is sensible that this reflects 8% of total feature variance, as people need to store many of their things in the place they live and the market supply is responsive to this.

```
##                 PC1  PC2  PC3  PC4  PC5
## area           -0.3  0.3  0.0  0.1  0.0
## Lot.Frontage   -0.2  0.1  0.2 -0.1 -0.1
## Lot.Area       -0.2  0.1  0.3  0.0 -0.1
## Year.Built     -0.2 -0.3 -0.2  0.1  0.0
## Year.Remod.Add -0.2 -0.2 -0.3  0.1 -0.2
## Mas.Vnr.Area   -0.2  0.0  0.1  0.1  0.2
## BsmtFin.SF.1   -0.2 -0.2  0.4  0.3  0.2
## Bsmt.Unf.SF    -0.1  0.1 -0.2 -0.6 -0.2
## Total.Bsmt.SF  -0.3 -0.1  0.2 -0.3  0.0
## X1st.Flr.SF    -0.3 -0.1  0.3 -0.2  0.0
## X2nd.Flr.SF    -0.1  0.4 -0.3  0.4  0.0
## Low.Qual.Fin.SF  0.0  0.1  0.0 -0.1 -0.4
## Bedroom.AbvGr  -0.1  0.4 -0.1  0.0  0.1
## Kitchen.AbvGr   0.0  0.2  0.0 -0.2  0.6
## TotRms.AbvGrd  -0.3  0.4 -0.1  0.0  0.1
## Fireplaces     -0.2  0.1  0.2  0.2 -0.2
## Garage.Yr.Blt  -0.3 -0.3 -0.3  0.1 -0.1
## Garage.Cars    -0.3 -0.1 -0.2 -0.1  0.1
## Garage.Area    -0.3 -0.1 -0.1 -0.1  0.0
## Wood.Deck.SF   -0.1  0.0  0.0  0.2  0.0
## Open.Porch.SF  -0.2  0.0 -0.1  0.1 -0.2
## Enclosed.Porch  0.1  0.2  0.1 -0.1 -0.3
## X3Ssn.Porch     0.0  0.0  0.0  0.0  0.0
## Screen.Porch    0.0  0.0  0.2  0.1 -0.1
## Pool.Area      -0.1  0.1  0.2  0.1 -0.4
## Misc.Val        0.0  0.0  0.1  0.0  0.0
```

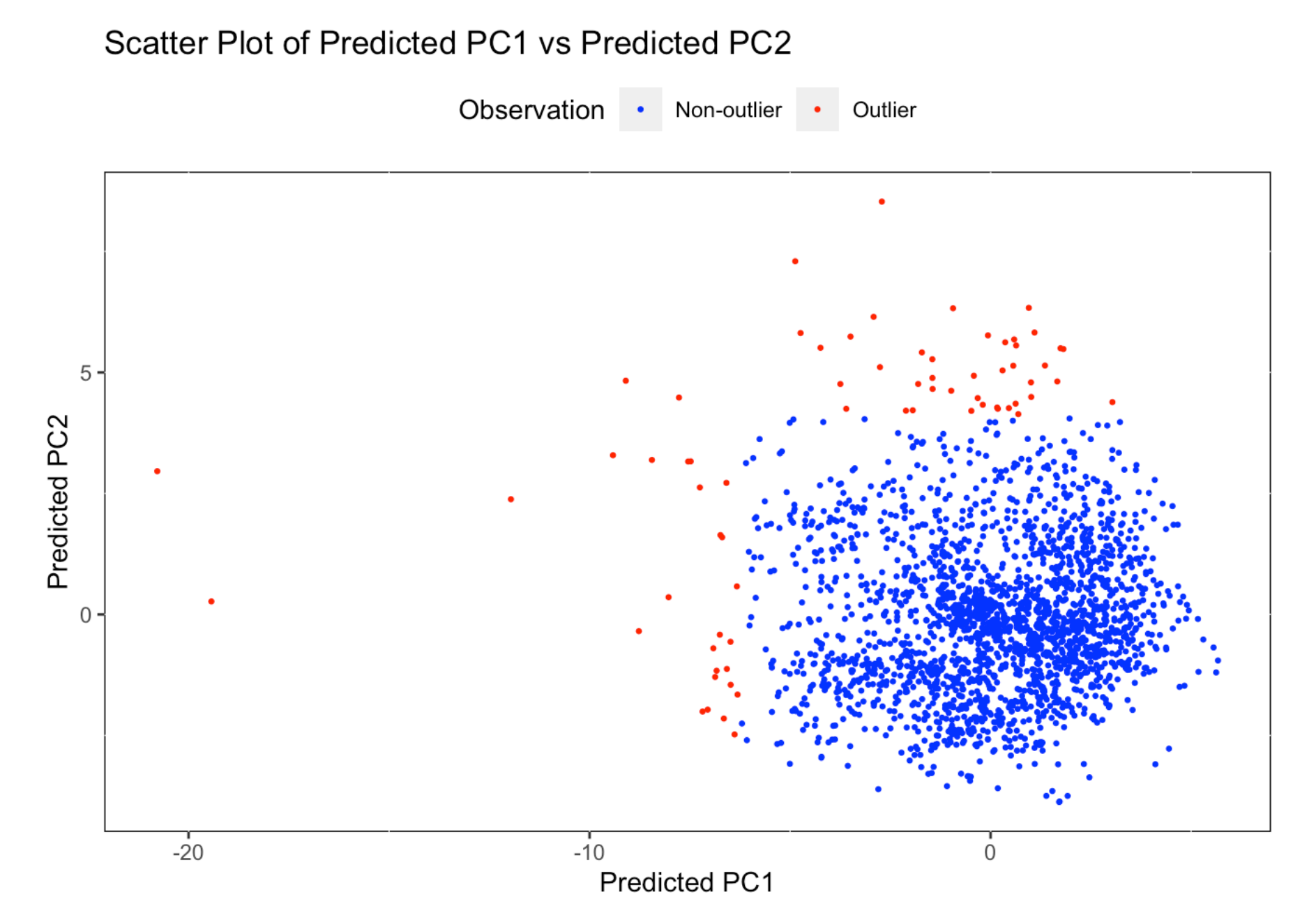
Question 4:Predict the PCs for each observation. Produce a scatter plot of the predicted PC1 against the predicted PC2 that supports (or questions, as the case may be) the interpretation you gave PC1 in question 3. Comment on the plot.

Under the interpretation of Question 3, a high/low PC1 is a small/large home and a high/low PC2 is a home with many/few rooms.

If we plot the predicted PC1 against PC2, we observe some substantial outliers. Outliers are identified in red using the conventional definition - an outlier is a point which falls more than 1.5 times the interquartile range above the third quartile or below the first quartile.

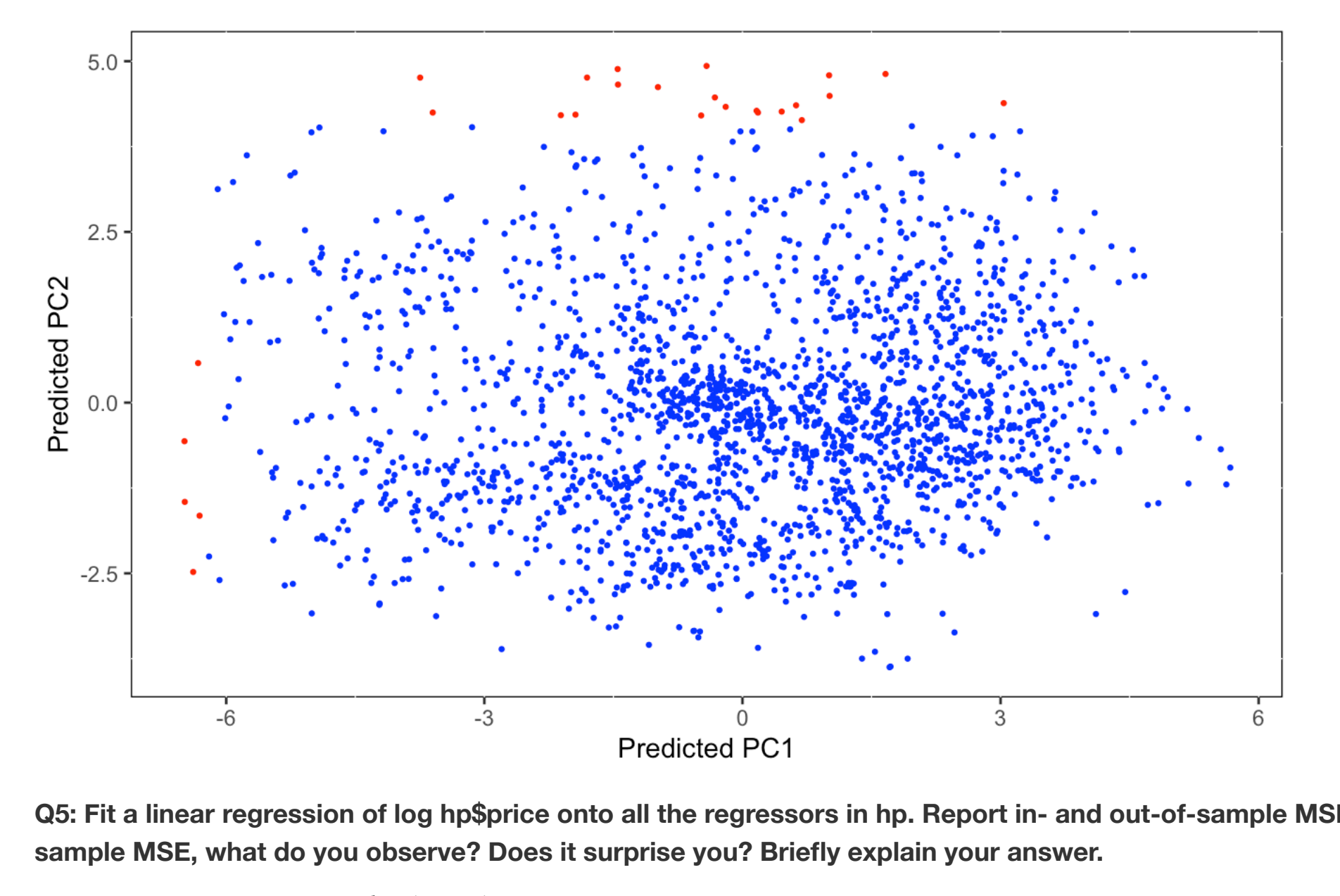
The PC1 outliers all represent uncharacteristically large homes (i.e. very negative PC1) and never small homes (i.e. very positive PC1). This is because there a minimum legal home size, but no maximum limitation. This would imply that the interpretation of PC1 from question 3 is sensible.

The PC2 outliers are less extreme than PC1 outliers but also only represent homes with many rooms. As for PC1, the amount of rooms is theoretically unbounded but it is unlikely to be an extreme amount as rooms lose their purpose if they are duplicative. There is also a minimum amount of rooms set by studio apartments (i.e. 1-2 rooms). This also helps validate the interpretation for PC2.



Although useful for verifying our understanding of the PCs, the extreme outliers hinder interpretation of most of the observations. In turn, the outliers are mostly removed and it becomes clearer there is no correlation between the PCs. This is expected, as PCs should be **orthogonal**.

If PC1 places a negative weight on the overall size of a home and PC2 emphasizes the number of rooms for living, the plot tells us that one can buy a small/large house with many/few rooms. But there is a clustering around the center (point (0,0)), so most homes are of a moderate size with a moderate amount of rooms. This implies there is a mix of smaller family apartment, studios, two-floor family villas or bungalows. As this is a very reasonable conclusion, the **interpretations for PC1 and PC2 are quite plausible**.



Q5: Fit a linear regression of log hp$price onto all the regressors in hp. Report in- and out-of-sample MSE. Comparing in- and out-of-sample MSE, what do you observe? Does it surprise you? Briefly explain your answer.

Fitting a linear regression of $log(price)$ on all regressors exposes us to the possibility of over-fitting by creating a highly flexible model. This model would reproduce in-sample idiosyncrasies that do not extend to the rest of the data. Thus, we should expect that our out-of-sample (OOS) MSE is considerably larger than our in-sample (IS) MSE. However, this turns out not to be true with IS MSE being near to OOS MSE.

The log-transformed IS MSE is 0.0163 and the log-transformed OOS MSE is 0.0229 (see table below). This is not very useful for interpretation, so I took the square root of each to find the IS RMSE and OOS RMSE. I then exponentiate and algebraically manipulate these to find the mean percentage IS and OOS errors.

For IS MSE:

$$\text{IS RMSE:} \quad \sqrt{0.0163} = 0.1276$$

$$\text{Exponentiating yields:} \quad e^{0.1276} = 1.136 \approx 1.14$$

$$\text{Mean percentage IS prediction error is then:} \quad (1.14 - 1) \times 100 = 14\%$$

For OOS MSE:

$$\text{OOS RMSE:} \quad \sqrt{0.0229} = 0.1513$$

$$\text{Exponentiating yields:} \quad e^{0.1513} = 1.16$$

$$\text{Mean percentage OOS prediction error is then:} \quad (1.16 - 1) \times 100 = 16\%$$

The mean percentage error of our IS prediction is 14% , which is lower than the 16% mean OOS prediction error. This makes sense, as the model was trained on the in-sample data, so it likely reproduced some features only observed within the sample. However, the difference in errors is not as large, implying we do not suffer from substantial overfitting.

14% and 16% mean prediction errors are unsatisfactory from the view of someone trying to sell their home, indicating room for improvement.

```
##                     Model     IS.MSE    OOS.MSE
## 1 Full Linear Regression 0.01627639 0.02295418
```

Q6: Fit a linear regression of log hp$price onto all the regressors in hp interacted with Neighborhood. Make a table where you report in- and out-of-sample MSE for this regression and the ones from the regression fitted in 5. Compare in- and out-of-sample MSE of the two fitted models. What do you observe? Does it surprise you? Briefly explain your answer.

We created the interaction with neighborhood because we suspect that there might be a clustering of features by neighborhood. This is reasonable as geographic conditions, zoning laws, and historical legacies often mean that only certain housing types can be built in a given area. Furthermore, social groups (e.g. high income professionals or middle-income families) tend to cluster together and demand similar housing. If this neighborhood-clustering analysis holds true, we could expect a lower OOS MSE. However we are interacting and regressing on all variables, risking over-fitting. We would then have a lower IS MSE, but higher OOS MSE than in question 5.

In fact, we have a lower IS MSE (i.e. we reproduce the data-wide and neighborhood-specific idiosyncrasies) and a similarly high OOS MSE. Using the same transformation as in question 5, we observe that we have a mean percentage IS prediction error of 13% and mean percentage OOS prediction error of 16%. The decline in IS MSE is probably explained by reproducing the idiosyncrasies created by the additional interaction variables. The OOS MSE is both similar to the previous OOS MSE and the current IS MSE, indicating we probably have not over-fit. But the error is still likely quite unsatisfactory for a potential house-buyer or seller.

A penalized regression (e.g. the lasso algorithm) could reduce the amount of variables in our model to only those that are most important, reducing IS and OOS MSE. This could then highlight our neighborhood-clustering features, potentially revealing interesting results. We test this in question 7.

```
##                     Model     IS.MSE    OOS.MSE
## 1 Full Linear Regression 0.01627639 0.02295418
## 2       Interaction Model 0.01489974 0.02281294
```

Q7: Fit a linear regression of log hp$price onto all the regressors in hp interacted with Neighborhood using the lasso algorithm. Report out-of-sample MSE at the out-of-sample MSE-minimizing penalty weight. Compare this out-of-sample MSE to those you reported in 5 and 6. What do you observe? Does it surprise you? Briefly explain your answer.

The lasso algorithm yields the smallest OOS MSE: a mean prediction error of 15.9% and .0218 if log-transformed, but only slightly less than the two previous models. The marginal improvement likely occurs as the lasso algorithm utilizes penalized regression that prevents over-fitting using a penalty that discourages complex models. At the cost of introducing bias, this can decrease variance and stabilizes out-of-sample predictions. However, our models likely do not suffer from substantial over-fitting, as the OOS MSEs are quite similar to IS MSEs. Thus, the over-fitting correction was minor at best.

One could argue that penalized regression deals with multicollinearity, thus our OOS MSE declined. However, multicollinearity often blows up OOS MSEs, which has not been the case with our previous models.

Nonetheless, it seems that the penalized regression yielded a small but unsatisfactory improvement in OOS MSE.

```
##                        Out.of.Sample.MSE
## Full Linear Regression        0.02295418
## Interaction Model             0.02281294
## Lasso Model                   0.02185305
```

Question 8: The command hpPC <- data.frame(hp,as.data.frame(Z[,1:ncol(Z)-1])) omits the last Principal Component. Why do omit that last component in our prediction model?

The last PC's share of variance is zero, so it carries no useful information for our analysis.

In practical terms, if a PC has no variance, it means that the corresponding linear combination of original variables does not contribute meaningfully to the overall variability of the data. Dropping such a PC would not impact the information captured by the remaining principal components, and it may be considered redundant.

However, we could also drop other PCs. For instance, each PC from PC21 only has 1% share of total feature variance. Alternatively, our analysis could operate with the first 13 PCs, which cover roughly 85% of the total feature variance.

Question 9: Fit a linear regression of log hp$price onto the PC-features in hpPC by OLS. Do not include any other features, nor should you include any interaction terms. Report in- and out-of-sample MSE. Make a table where you report in- and out-of-sample MSE for this regression and the ones from the regressions fitted in 5 and in 6. Compare in- and out-of-sample MSE for the PCR with those of the fitted models from 5 and 6. What do you observe? Does it surprise you? Briefly explain your answer.

The PCR OLS model has by far the largest IS and OOS MSE (see table below). After transformation, we find it has a mean IS prediction error of 18.5% and mean OOS prediction error of 20.4%.

This could be because when I executed PCA, it only used the 26 numerical features specified in *x*. Thus the 25 PCs only capture the variation within *x*. However, the price is determined by 71 variables, of which many are non-numerical, categorical, and material for a home's price. This results in a high IS MSE because the 26 numerical features do a poor job (relative to other models) in explaining full variation in the train set. Similarly, it performs even worse in the test set, as for the same reason and possibly reproducing some test-set-specific idiosyncrasies.

```
##                        In.Sample.MSE Out.of.Sample.MSE
## Full Linear Regression    0.01627639        0.02295418
## Interaction Model         0.01489974        0.02281294
## OLS PCR Model             0.02866095        0.03458749
```

Question 10: Fit a linear regression of log hp$price onto the PC-features in Z as well as the regressors in hp interacted with Neighborhood using the lasso algorithm. Make a table where you report out-of-sample MSE at the out-of-sample MSE-minimizing penalty weight for this lasso and the lasso fitted in 7. What do you observe? Does it surprise you? Briefly explain your answer

This model has the lowest OOS MSE out of all models (0.02173 log-transformed and 15.8% mean prediction error). This could be for the following reasons:

1. It remedies the issue with the model in question 9 by adding the remaining non-numerical and neighborhood-interacted regressors. This means the PCs are able to capture the variance caused by all the other regressors, improving IS MSE and OOS MSE.
2. It benefits from the same gains as the lasso model in question 7, namely penalized regression reducing over-fitting by undergoing the bias-variance trade-off (explained in q7). This should contribute to improving OOS MSE.

However, the improvement is extremely small as the lasso has a 15.9% mean prediction error, only 0.1% more. If the data had more numerical features, the PCA and PCR could be more useful and the MSEs would be smaller.

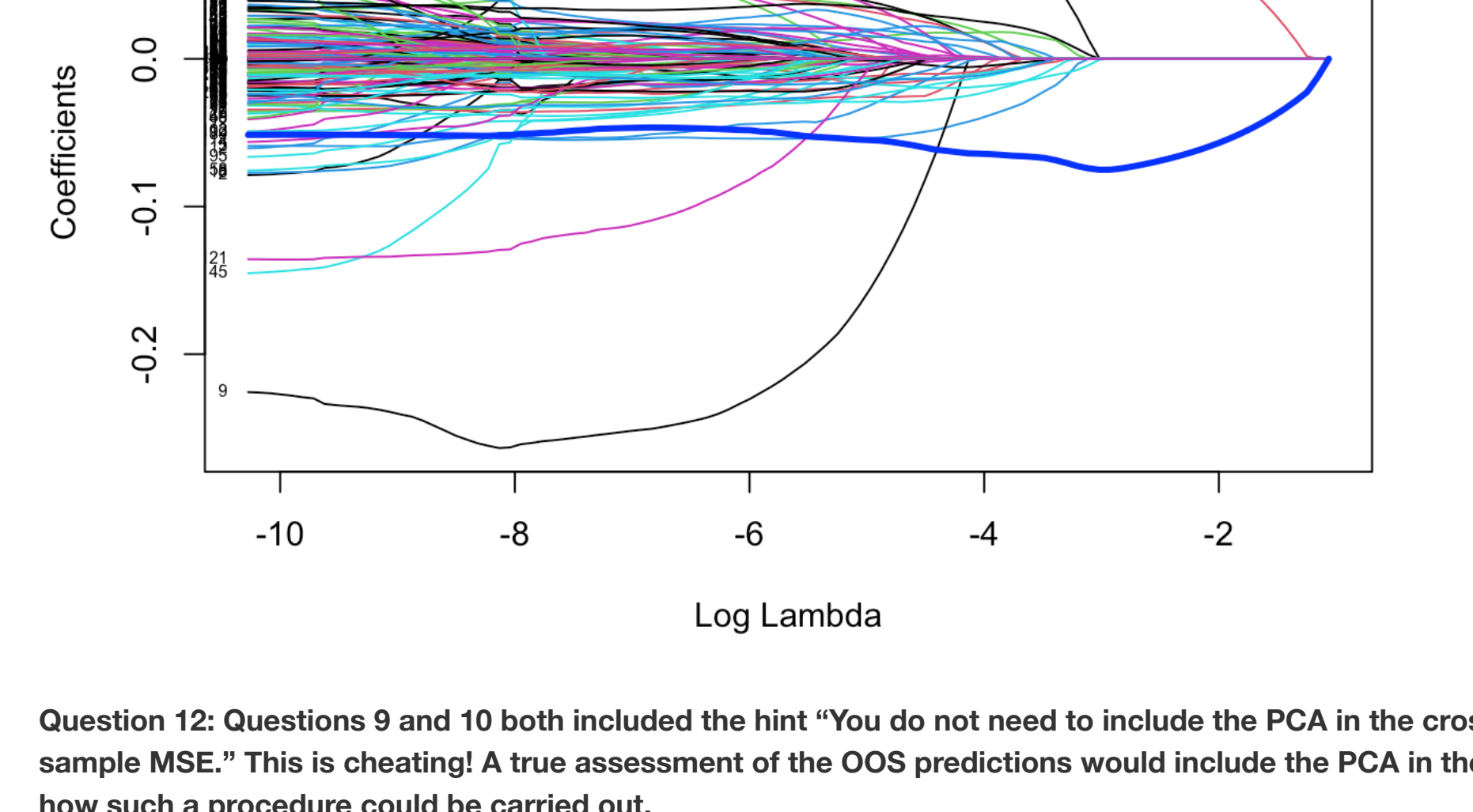
```
##                 Out.of.Sample.MSE
## Lasso Model            0.02185305
## PCR Lasso Model        0.02173700
```

Question 11: Plot the coefficients along the regularization path (i.e. as a function of the log penalty weight log λ) for the PCR-lasso from 10. Highlight the coefficient path for the PC1-feature in the PCR-lasso. What do you observe? Does it surprise you? Briefly explain your answer.

As is typical for regularization paths, the model complexity increases as the penalty declines. PC1 composes the uni-variate model and remains one of the most negative coefficients as the penalty loosens. However, some of the most impactful variables surface in more complex models (e.g. the ninth variable is the most negative and 138th most positive).

We see that PC1 is relatively stable compared to other significant features that tend to blow up and then decline (e.g. feature 9) or have paths that frequently fluctuate (e.g. feature 108).

PCA also takes the longest to become zero. This is consistent with it having the largest share of total feature variance and can be expected to compose the uni-variate model. If we assume that PC1 inversely reflects a home's size (i.e. the larger the PC1, the smaller the home), this suggests that the primary/first consideration of a buyer is the size of a home.



Question 12: Questions 9 and 10 both included the hint "You do not need to include the PCA in the cross validation to get the out-of-sample MSE." This is cheating! A true assessment of the OOS predictions would include the PCA in the cross validation. Briefly outline how such a procedure could be carried out.

For a split-sample validation approach:

1. I would divide the data set into a training and test set (e.g. a 50/50 train/test split)
2. I would execute the PCA on the training set to identify the PCs.
3. I would apply the PCA transformation to the training set to reduce dimensionality, so that my training data is expressed in terms of the identified PCs.
4. I would train my model (e.g. a Lasso regression) on the PCA-transformed trained data.
5. I would transform the test set using the PCA that were fitted on the training set.
6. I would evaluate the performance of the PCA-transformed test data set by calculating the OOS MSE.