project_step_2

Lucy Zhao & Keon Dibley

2023-11-05

Dataset: California Housing Prices Data Source: Kaggle

The processed dataset **california_housing** has **200 observations**, **1 response variable**, **13 quantitative variables** and **2 categorical variables**(which we created). Each observation represents a block of California houses and has attributes like the median house value of that block, its median income, and the population of the block. From this dataset, we chose **Median_House_Value** as the response variable and used other variables as predictors to predict Median_House_Value.

Also, The raw dataset only includes quantitative variables and some of these variables are too detailed, like latitude and distance to LA/SD/SJ/SF, so we set a standard to classify different values into different levels. For example, we classified all the observations into north/south CA according to latitude. Later on, we'll dive into whether using the raw data or using the level after assignment can better fit a model.(For categorical variables "north_south_CA" & "near_top4_big_city", we gave them assignments according to their level.)

But in this step, we will only explore the SLM for each quantitative predictor:

$$Y = \beta_0 + \beta_1 x$$

We tested the hypothesis below for the significance of each variable as a predictor of Median House Value.:

- $H_0 : \beta_1 = 0$
- $H_a : \beta_1 \neq 0$

Then, we need to find $\hat{\beta}_1$,and (1-α)CI for $\hat{\beta}_1$, if $\hat{\beta}_1$ falls into CI, then the corresponding variable is not significant. Also, we list R^2 to show the proportion of variance in the response explained by the model. Below is the table showing the info above (we set $\alpha = 0.05$).

Table 1: Estimation of each SLM

Predictor	Estimate	CI-lwr	CI-upr	If significant	R^2
Median_Income	45013.09	-5994.72	5994.72	Yes	0.53
Median_Age	2157.44	-1412.18	1412.18	Yes	0.04

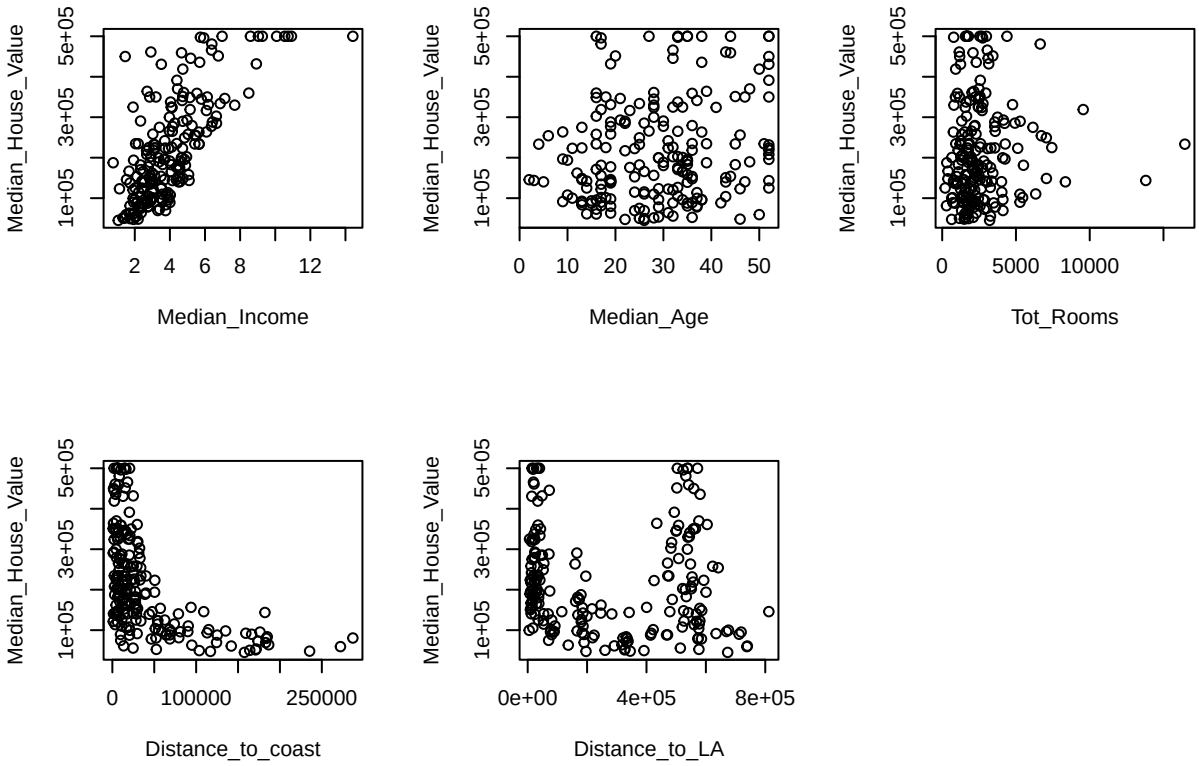
Predictor	Estimate	CI-lwr	CI-upr	If significant	R^2
Tot_Rooms	3.25	-8.90	8.90	No	0.00
Tot_Bedrooms	-7.23	-44.14	44.14	No	0.00
Population	-12.76	-17.95	17.95	No	0.01
Households	1.88	-48.84	48.84	No	0.00
Latitude	-3439.27	-8406.36	8406.36	No	0.00
Longitude	-10444.30	-8325.65	8325.65	Yes	0.03
Distance_to_coast	-1.24	-0.28	0.28	Yes	0.28
Distance_to_LA	-0.04	-0.07	0.07	No	0.01
Distance_to_SanDiego	0.00	-0.06	0.06	No	0.00
Distance_to_SanJose	-0.08	-0.08	0.08	No	0.02
Distance_to_SanFrancisco	-0.06	-0.07	0.07	No	0.02

With this table, we find that Median Income, Distance to Coast, Median Age, and Longitude, are significant variables at $\alpha = 0.05$. However, Median Age and Longitude have very low R-squared values, so they are not very useful in explaining the variance of the response variable. If we set $\alpha = 0.025$, these predictors are not significant, while Median Income and Distance to Coast still are. This suggests a linear relationship between Median Income and Median House Value, as well as between Distance to Coast and Median House Value.

Many of the other predictors are not significant, but it doesn't mean they don't have a relationship with the response. So what's quite important is that we neeed to look at the scatter plots of each predictor and the response, then we can transform each predictor to better fit their relationship with the Median House Value.

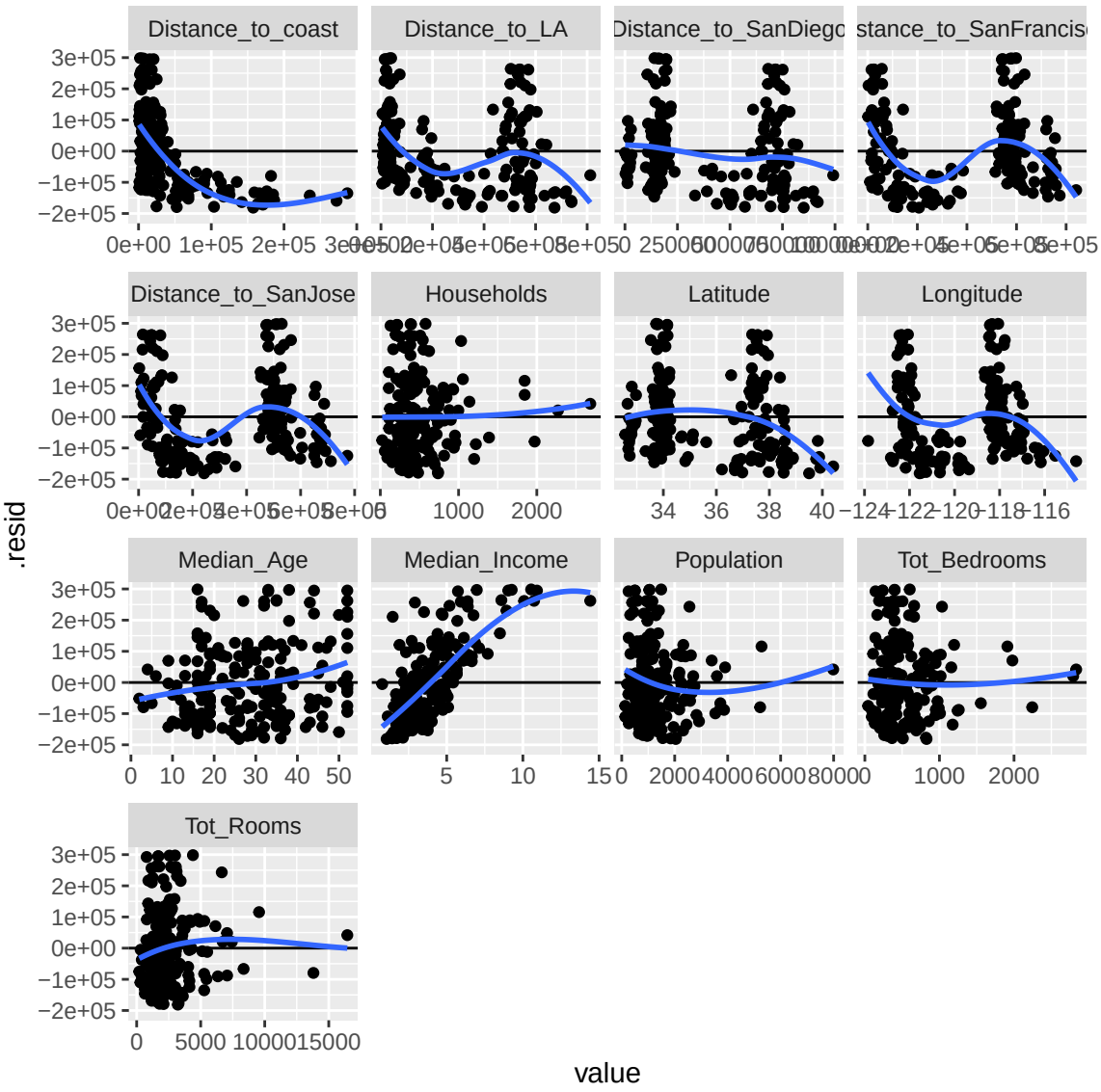
Through the scatter plots, we can see the groups below share similar relationships with response variable, so later we can use similar transforms on predictor variables within a group. This also indicates that predictors within a group might be dependent.
(1)Tot_Rooms & Tot_Bedrooms & Population & Households
(2)Latitude & longitude & Distance to LA/SD/SJ/SF(they all have 2 or 3 peeks in the plots)

We only show 5 of all the quantitative predictors.



(Figure 1: Median_House_Value vs 5 different predictors)

About how to transform each predictor, we also need to focus on residuals vs predictor of each SLM, which is showed below.



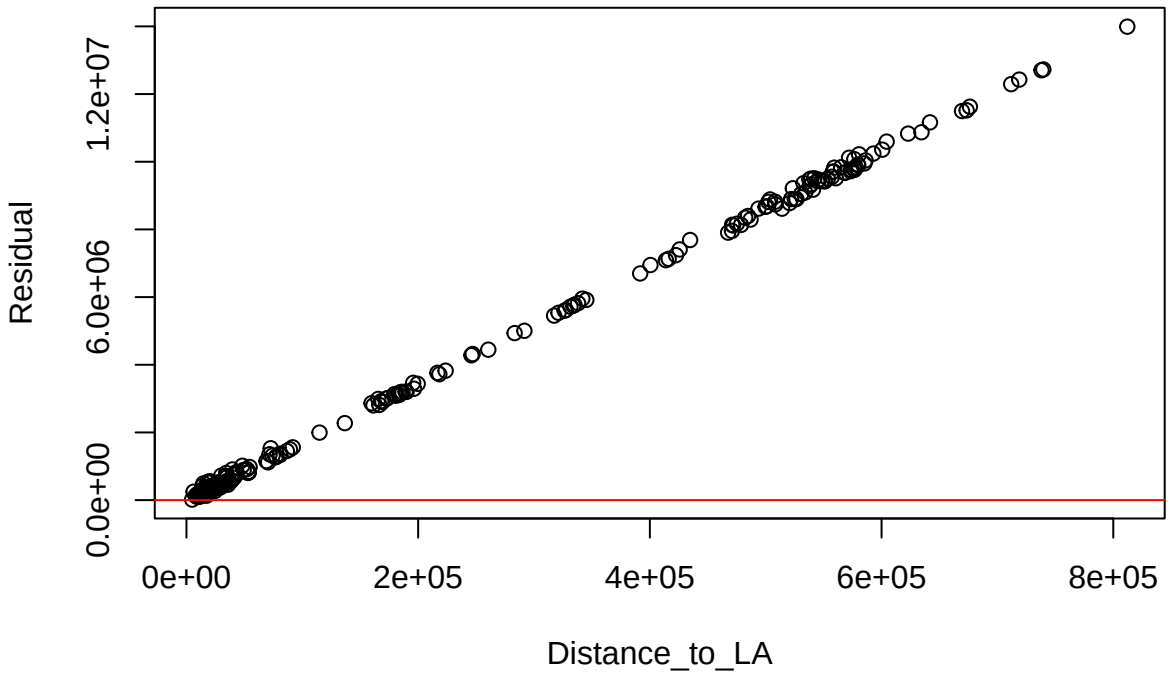
(Figure 2: Residuals vs all predictors)

All the predictors can be transformed and then check whether it turns into significant or whether R^2 of the model increase. Here we just take one as an example.

We can see from the residual that each predictor in group(2)(ie. distance to LA/SD/SF/SJ) might can be transformed to a cubic equation to better fit model with the response variable.

But when we start to work on distance to LA, we find as the degree of polynomial grows, R^2 grows as well, this is amazing and we think of Taylor series to explain it. Since many functions about x can be written into Taylor series(ie.$f(x) = \sum_{k=0}^{\infty} a_k x^k$), so when the degree of polynomial grows, it's more similar to Taylor series, thus can better explain y, thus has higher R^2.

What we are doing above is to find $f(x)$ which we transform x into, then $\hat{y} = f(x)\hat{\beta}$. Then we draw residual vs predictor distance to la again to see what's different(showed below).



(Figure 3: Residual vs Distance to LA after transformation)

Through the residual plot, we can see that $y - \hat{y} = ax + \varepsilon$, which means the residual has a perfect linear relationship with predictor x. We can include this part into the model of y, then we get $y = ax + f(x)\hat{\beta} + \varepsilon$, so once we figure out the value of a(=17.2), we can perfectly fit the model between y and x.(In this part, y refers to Median_House_Value, x refers to Distance_to_LA)

Then, distance to the other cities as well as other predictor variables can be transformed through this way. We'll not gonna include here.

Concluding Statement:

As statisticians, we need to research further the external factors that impact house value because these quantitative variables don't fully explain California Housing Value. We are interested in painting a complete picture of California Housing statistics, and we need to examine all contributing factors to do so.