step3

Lucy Zhao & Keon Dibley

2023-12-03

We randomly select 200 observations as train data and 50 observations as test data.

1. Choose two fitted model

From step1&2:
Highly correlated : "Households,Population,Tot_Bedrooms,Tot_Rooms".
Log : Consider taking log for predictors which has large values, like"Median_House_Value, Min_Distance, Distance_to_coast".
interaction term : **Median_Income:Median_Age** might be possible because people's median income will to some degree decide the age of the house they want to buy.

We fit two models(fit1 and fit2).
fit1: $ln(\hat{MedianHouseValue}) = -10.1 + 0.254 * MedianIncome - 0.194 * ln(MinDistance) - 0.19 * Logitude + 0.734 * I(northsouthCA = SOCAL) + 0.0138 * MedianAge - 0.00289 * MedianIncome * MedianAge$
Adjusted R-squared = 0.733, $\hat{\sigma}$ = 0.321, # of predictors = 6

fit2 :$ln(\hat{MedianHouseValue}) = -14.894 + 4.663 * MedianIncome - 1.491 * MedianIncome^2 - 0.125 * ln(MinDistance) - 0.162 * ln(Distancetocoast) + 0.004 * MedianAge$
Adjusted R-squared = 0.748, $\hat{\sigma}$ = 0.312, # of predictors = 5

Obviously, fit 2 is a better model, since it has higher Adjusted R-squared, lower $\hat{\sigma}$ and lower # of predictors which can make model simplier.

2. Use Forward selection using p-values

Start from only intercept, let $\alpha = 0.10$ be stopping criteria.

Till end, p-value always smaller than 0.10, so the full model fit2 is the best choice.

3. Justify chosen model

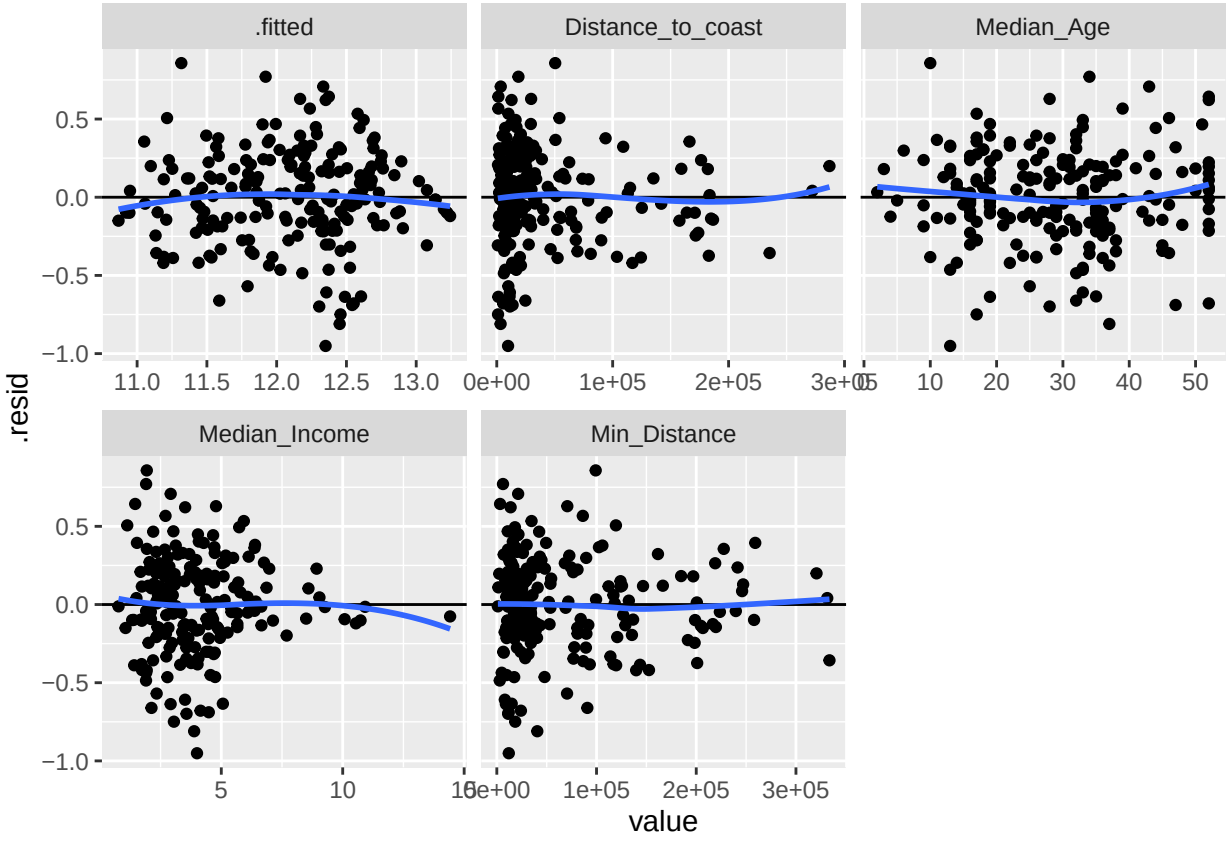
fit2 :

```
##                             Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)             14.893786577 0.26971565 55.220329 1.324183e-120
## poly(Median_Income, 2)1  4.663397036 0.32848287 14.196774  7.765803e-32
## poly(Median_Income, 2)2 -1.491285903 0.32385275 -4.604827  7.456809e-06
## log(Min_Distance)       -0.125311627 0.02579451 -4.858074  2.437267e-06
## log(Distance_to_coast)  -0.161831091 0.02205572 -7.337374  5.807833e-12
## Median_Age               0.004299827 0.00198194  2.169504  3.126031e-02
```

R-squared: 0.754
Adjusted R-squared: 0.748

Partial F test for each coefficient estimate: all p-value smaller than 0.05
Global F test for the model: p-value smaller than 0.05(<2e-16).
Very small p-value illustrates that the model is well fitted.

Use residual vs fitted values and predictors to double check:



(Figure 1: Residuals vs fitted values and predictors of model fit2)
In these residual plots, the points are scattered randomly around the residual=0 line, so all these estimates are unbiased.
Most plots have constant variance. Residual vs Distancetocoast and MinDistance seems to have changed variance as value of predictors grow, but this is because the number of observations decreases as value of predictors grow. In conclusion, the model is fitted well.

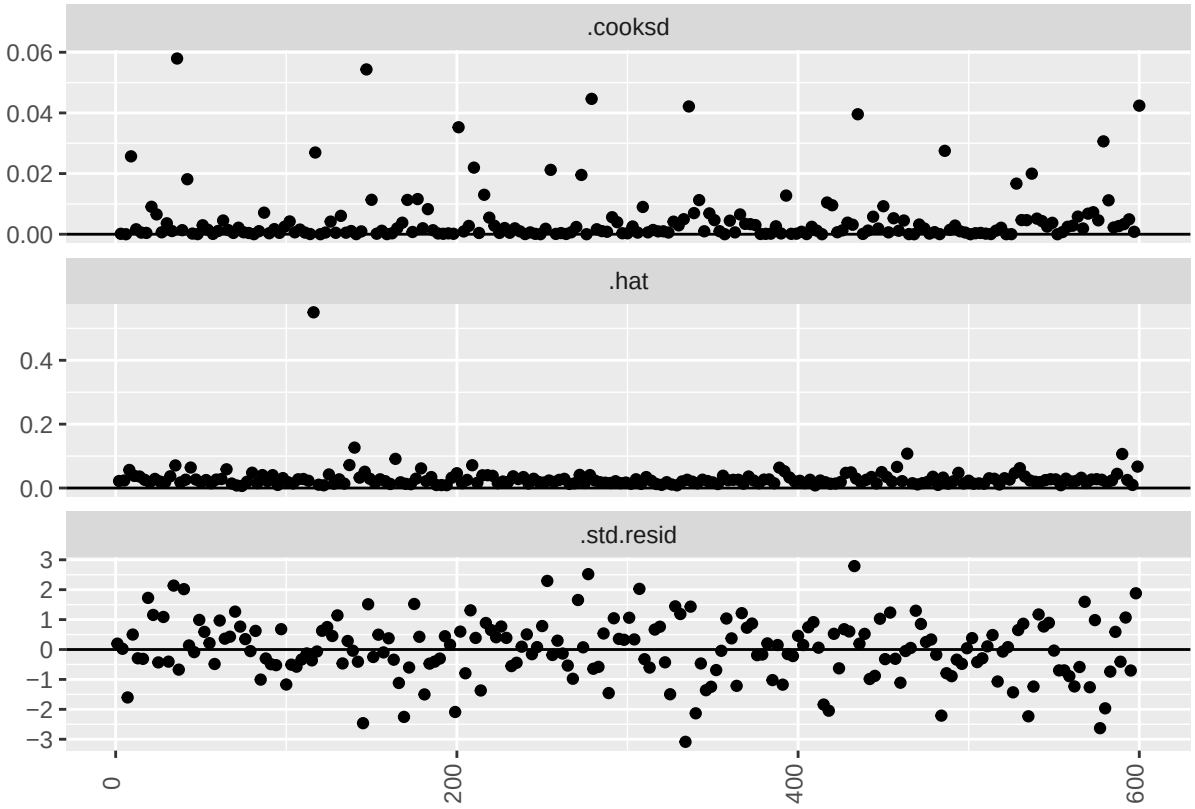
4. Interpret coefficients

From the above output, all coefficients are significant.
β_0 : Mean of ln(MedianHouseValue) when all other predictor variable values are 0.
β_1 : Change in ln(MedianHouseValue) when Median_Income changes in one unit while holding other variables in the model constant.
β_2 : Change in ln(MedianHouseValue) when Median_Income^2 changes in one unit while holding other variables in the model constant.
β_3 : Change in ln(MedianHouseValue) when log(Min_Distance) changes in one unit while holding other variables in the model constant.
β_4 : Change in ln(MedianHouseValue) when log(Distance_to_coast) changes in one unit while holding other variables in the model constant.
β_5 : Change in ln(MedianHouseValue) when Median_Age changes in one unit while holding other variables in the model constant.

5. R square on test data

R^2 **on test data** : -12380
Adjusted R^2 **on test data** : -12699
It's much more small than 0 because SS_{Res} is much more larger than SS_T on test data.
The variability explained by the model fit2 is 74.8% (Adjusted R^2 on train data)
So a high R^2 can not always guanrantee that the model will accurately describe the population, since sometimes model assumption fail for some observations.

6. Case influence statistics



(Figure 2 : Cook's Distance, h_{ii}, Internally Studentized Residual of each observation)
Leverage points: $h_{ii} > 2(p+1)/n = 0.06$ (No leverage point)
Outliers: Internally Studentized Residual > 3 (Only one outlier)
Influential points : Cook's Distance $> 4/n = 0.02$
The only outlier is not influential, so we don't need to re-fit a model without some certain data.

7. Interpretation of the final model

Our task is to estimate the MeanHouseValue of a block given some variables values.
There are many different variables provided in the raw data set.

Reasons for variables dropout:
some variables are highly correlated, so one can taken place of some or all the other ones(like "longitude and distancetocoast");
some variables are randomly distributed, having no relationship with response variable.

Reasons for variables significant:
Intuitively have a obvious relationship with house value;
Has a better way to illustrates house value.

Interpret our final model:

8. Prediction

9. Summary