Housing Price Prediction: Statistical approach

Розбір позначень в інформативному повідомленні про регресійну модель.

Залишок (residual) u_i: $$y_i = \hat{y}_i + \hat{u}_i.$$

Оцінка дисперсії похибок (residual standard error, підправлена оцінка ММП) [Woolridge ст. 33]:

$$SSR = \frac{1}{n-p}\sum_{i=1}^{N} \hat{u}_i^{\,2}, де\ p - кількість\ коефіцієнтів\ регресійної\ моделі.$$

Зазвичай p = n_features + 1, оскільки враховується вільний член (intercept). Тому треба звертати увагу на формули, які ми використовуємо – там за p може братись кількість признаків, тоді потрібно відняти ще одиницю (в нас кількіість признаків позначена k). Можна використовувати таку формулу [Woolridge ст. 88]:

$$SSR = \frac{1}{n-k-1}\sum_{i=1}^{N} \hat{u}_i^{\,2}, де\ k - кількість\ признаків.$$

Оцінка коефіцієнту детермінації (multiple R-squared):

$$R^2 = 1 - \frac{SSR}{SST}, де\ SSR - оцінка\ дисперсії\ похибок, SST - дисперсія\ цільового\ признака\ y.$$

Підправлений коефіцієнт детермінації (adjusted R-squared):

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$$

F-статистика в загальному випадку [Woolridge ст. 129]:

$$F \equiv \frac{(\mathrm{SSR}_r - \mathrm{SSR}_{ur})/q}{\mathrm{SSR}_{ur}/(n-k-1)},$$

F-статистика для перевірки рівності нулю усього вектора коефіцієнтів, за виключенням вільного члена. Це перевірка регресійної моделі на значимість [Woolridge ст. 135]:

$$F = \frac{\|X\hat{\alpha}\|}{k \cdot SSR} = \frac{\|X\hat{\alpha}\|}{\frac{k}{n-k-1}\sum_{i=1}^{N}\hat{u}_i^2}.$$

Довірчі інтервали

Хай дане спостереження x_h. Потрібно дати інтервальну оцінку для M(y| x = x_h) та y|x=x_h.

Для задання довіри прогнозування в регресійній моделі використовується два типи довірчих інтервалів: ДІ умовного матсподівання M(y| x = x_h) та прогнозувальний інтервал, що встановлює межі для майбутніх можливих значень y.

a) Довірчий інтервал (confidence interval) для середнього значення відклику за даним спостереженням *x_h*. Відповідає на запитання: “Яким є матсподівання умовного розподілу *y* за даним значенням *x = x_h*?”. Він будується на основі лише *standard error of fit*.

b) Прогнозувальний інтервал (prediction interval) — інтервал, який із заданим рівнем довіри буде містити значення відклику *y* для заданого значення спостереження *x = x_h*. Це також вид довірчого інтервалу. Питання, на яке він відповідає: “Яким може бути відклик y|x=x_h із заданим рівнем довіри?”. Він будується на основі *standart error of fit* та *standard error of prediction*.

Якщо ДІ дає оцінку якомусь параметру популяції, то ПІ дає оцінку області, в яку значення y потрапляє з деякою ймовірністю.

Різниця між описаними довірчим інтервалом і прогнозувальним інтервалом, з формулами: https://online.stat.psu.edu/stat501/lesson/3/3.3

Обрахунок в R: https://rpubs.com/aaronsc32/regression-confidence-prediction-intervals

Популярне пояснення довірчого та прогнозувального інтервалу в регресії: https://towardsdatascience.com/how-confidence-and-prediction-intervals-work-4592019576d8

Довірчі інтервали для коефіцієнтів лінійної регресії

Обчислюються згідно формул.

Обрахунок в R: https://rpubs.com/aaronsc32/regression-confidence-prediction-intervals

Довірчі інтервали для справжньої лінії регресії

```
> summary(model)

Call:
lm(formula = SalePrice ~ OverallQual, data = X)

Residuals:
    Min      1Q  Median      3Q     Max
-168790  -21881   -5844   17512  138119

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -143.6     5057.9  -0.028    0.977
OverallQual  29715.9      809.1  36.726   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42290 on 1453 degrees of freedom
Multiple R-squared:  0.4814,    Adjusted R-squared:  0.481
F-statistic:  1349 on 1 and 1453 DF,  p-value: < 2.2e-16
```

Плотинг в R:

https://rpubs.com/Bio-Geek/71339

https://www.r-graph-gallery.com/line-plot.html

Вибір структури регресійної моделі

http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/

1. Крокова регресія (Stepwise regression). Її недоліки описані тут.
   План: вибираємо структуру моделі на половині train набору. На іншій половині фітимо модель.

   *The essential **problems with stepwise methods** have been admirably summarized by Frank Harrell (2001) in Regression Modeling*
   Strategies, and can be paraphrased as follows:
   1. R^2 values are biased high
   2. The F statistics do not have the claimed distribution.
   3. The standard errors of the parameter estimates are too small.
   4. Consequently, the confidence intervals around the parameter estimates are too narrow.
   5. p-values are too low, due to multiple comparisons, and are difficult to correct.
   6. Parameter estimates are biased away from 0.
   7. Collinearity problems are exacerbated.