Regression-based forecast

(Jing Li, Miami University)
Big picture

1. OLS regression produces best linear (parametric) forecast of $y$ using other predictors $x$

2. We need to address several issues
   (a) Nonlinearity
   (b) Outlier
   (c) Multicollinearity
   (d) Mean vs Median
Linear regression model

Recall our goal is using $x$ to predict $y$

$$y = f(x) + u$$

The fundamental theorem of forecasting implies that we need to replace $f(x)$ with conditional mean $E(y|x)$. Next, if we assume the conditional mean is a linear function, i.e.,

$$E(y|x) = \beta_0 + \beta_1 x \quad (\text{Assumption of Linearity}) \quad (1)$$

then we get the linear regression model

$$y = \beta_0 + \beta_1 x + u \quad (2)$$

Remarks

1. Regression is a model, which is as good as its assumptions

2. The forecast based on (2) is good when

   (a) $x$ has strong predictive power (or $u$ does not matter much)

   (b) $E(y|x)$ is indeed linear
Assumption of Linearity

Under the linearity assumption, we have

\[
\frac{dE(y|x)}{dx} = \beta_1 = \text{constant}
\]  

(3)

This linear assumption implies that when \( x \) changes by one unit, the average \( y \) will change by a fixed amount. Two counter examples

1. Eating pizza: the first slice tastes best

2. Slowing depreciation of car value
Ordinary Least Squares (OLS) Estimation

We use OLS to estimate $\beta_0$ and $\beta_1$. That is, their estimates minimize residual sum squares (RSS)

$$\hat{\beta}_0, \hat{\beta}_1 = \text{argmin}_{b_0, b_1} \sum_i (y_i - b_0 - b_1 x_i)^2$$

(4)

Next we can compute the predicted value as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad \text{(predicted value)}$$

(5)

and the predicting error is

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad \text{(predicting error or residual)}$$

(6)
\( \hat{\beta}_0 + \hat{\beta}_1 x \) is called the best linear predictor (BLP), where \( \hat{\beta}_0, \hat{\beta}_1 \) are obtained via OLS. In short, OLS produces the best linear forecast (not necessarily the best forecast). This fact explains the popularity of OLS in forecasting analysis.
A story of tracking Comet Ceres (line-fitting, or OLS)—the comet moves on a track. But due to measurement error (randomness, error term), the observed positions deviate from the track. OLS is proposed by Gauss to estimate the blue dash track, and he uses the blue dash line to predict the future position.
Example: predicting house price

A real estate investor buys an undervalued house and sell an overvalued house. The investor uses house data to run OLS, compute the predicted price, and compare to the actual price

```r
> address = "https://www.fsb.muohio.edu/lij14/400_house.txt"
> data = read.table(url(address), header=T)
> data[1:3,]

   age area baths rprice
1  48 1660  1   60000
2  83 2612  2   40000
3  58 1144  1   34000

> # define variables
> price = data[,4]
> age = data[,1]
> area = data[,2]
> baths = data[,3]
```
Remarks

1. $y$ is the price, the fourth column in the data file

2. common sense tells us the age of house (age), its size (area), and the number of bathrooms (baths) matter. So we try using them as predictors

3. The first house (observation) is 48 year old, has 1660 square feet, one bathroom, and its price is 60000

4. A starting point is a simple regression that only uses age as the predictor

5. Due to depreciation, our prior belief is that age affects price negatively

6. The error term can include important factors such as city, school district, maintenance condition, the interior design, etc. We do not have data for those factors. If we believe, say, school district matters most for house price, then a regression that ignores school district will not have strong predictive power
Scatter plot and OLS fitted line

> plot(age, price)
> abline(lm(price ~ age), col = "red", lty=2, lwd=4)
Remarks

1. Price is on the vertical axis and age is on the horizontal axis. Each point represents a house.

2. The red dash OLS fitted line is downward-sloping, consistent with the prior belief that aging leads to depreciation.

3. There are several very old houses with age greater than 100. It seems that the OLS does a bad job tracking those old houses (with big gap between points and fitted line).

4. There is a house with zero age, and its price is about 300000. That house really stands out in the upper left corner, so is labeled as “outlier.”

5. Next we use R `lm` function to show the estimation results. To learn more about that command, google ”R, lm”
OLS Estimation of simple regression

> summary(lm(price~age))

Call:
  lm(formula = price ~ age)

Residuals:
   Min     1Q    Median     3Q    Max
-63124  -22187   -5499  13887  210201

Coefficients:
     Estimate Std. Error t value  Pr(>|t|)
(Intercept)   89799.41    1996.48    44.979 < 2e-16 ***
age           -337.49     53.71    -6.283  1.09e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31290 on 319 degrees of freedom
Multiple R-squared:  0.1101,  Adjusted R-squared:  0.1073
F-statistic: 39.48 on 1 and 319 DF,  p-value: 1.086e-09
OLS Estimation of simple regression

> reg1 = lm(price~age)
> summary(reg1)$coefficient

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 89799.4076 | 1996.48362 | 44.978785 | 3.837920e-140 |
| age            | -337.4944   | 53.71305   | -6.283285 | 1.086304e-09  |

> summary(reg1)$sigma

[1] 31290.83

> summary(reg1)$adj.r.squared

[1] 0.1073413

1. The intercept $\hat{\beta}_0$ is 89799.41

2. The slope $\hat{\beta}_1 = -337.49$, and is statistically significant with t value $-6.283$. Increasing age by one year is associated with price reducing by 337.49
Measurement of performance

1. Adjusted R squared implies that this simple regression only explains about 11 percent variation of house price, thus age does not have strong predictive power. This finding is consistent with the fact that in the scatter plot the points are widely distributed around the fitted line (the fit is not tight).

2. The residual standard error is 31290. To put that number into perspective, the standard deviation of price is 33119. Again, the message is that the model does not explain much about price.

3. Later we will compare performance of models using residual standard error and adjusted R squared.
Revisiting linearity assumption—try quadratic regression

1. Given that the house depreciates at decreasing rate, we next try a multiple regression with a squared term

\[ y = \beta_0 + \beta_1 x + \beta_2 x^2 + u \]  

(quadratic regression)  

(7)

2. This is still a parametric model. It assumes the conditional mean takes the form of a quadratic function

3. A big t value for \( \beta_2 \) is an evidence for nonlinearity

4. For the current example, we expect \( \beta_1 < 0, \beta_2 > 0 \) (an upward-facing parabola with a turning point at the bottom)
Adding squared term (quadratic regression)

> # quadratic regression
> agesq = age^2
> reg2 = lm(price~age+agesq)
> summary(reg2)$coefficient

| Estimate   | Std. Error | t value | Pr(>|t|)   |
|------------|------------|---------|-----------|
| (Intercept)| 98512.37    | 1904.19 | 51.73     |
| age        | -1457.47    | 115.19  | -12.65    |
| agesq      | 8.29       | 0.78    | 10.62     |

> summary(reg2)$sigma

[1] 26929.55

> summary(reg2)$adj.r.squared

[1] 0.3388358
Remarks

1. The signs of estimated $\hat{\beta}_1 = -1457.472891 < 0, \hat{\beta}_2 = 8.292846 > 0$ confirm our projection

2. The negative $\hat{\beta}_2$ implies decreasing depreciation rate

3. We see smaller residual standard error 26929.55 and bigger adjusted R squared 0.3388358, both implying the quadratic model has better in-sample fitting than the simple regression

4. The nonlinear relationship can be seen from the plot of fitted value against $x$
Plot of predicted price from the quadratic regression
Does quadratic model produce better forecast? Not necessarily. If we want to predict price for a house with age less than 100, we better run the simple regression without the squared term. Our concern is that the quadratic model may suffer overfitting if majority of houses are young.
Overfitting—try cubic regression

Let’s go one-step further and run the cubic regression

```r
> # cubic regression
> agecu = age^3
> reg3 = lm(price~age+agesq+agecu)
> summary(reg3)$coefficient

        Estimate     Std. Error     t value   Pr(>|t|)
(Intercept)  1.009953e+05  2001.6920532  50.454987  1.567361e-153
     age -2.117128e+03  220.0443769  -9.621370  2.153801e-19
    agesq  2.240136e+01   4.1080350   5.453059  9.978114e-08
    agecu -6.020896e-02   0.0172224  -3.495967  5.397027e-04

> summary(reg3)$sigma

[1] 26466.61

> summary(reg3)$adj.r.squared

[1] 0.3613721
```
Overfitting

Despite of the statistically significant $\beta_3$ for the cubic term, there are three red flags for possible overfitting issue:

1. The residual standard error decreases by a very small amount relative to the quadratic regression

2. Adjusted R squared increases only marginally

3. The plot of fitted value indicates another turning point, which may be non-existing. Moreover, the first turning point in cubic regression happens much earlier than the turning point in the quadratic regression
Plot of a overfitted cubic regression
Lesson: just because a predictor is statistically significant does not necessarily mean it is useful for forecasting. It just means it helps improve in-sample fitting. Keep in mind the overfitting issue.
Multicollinearity—redundant information

1. Multicollinearity arises when predictors are highly correlated

2. As a result, those predictors are imprecisely estimated with big standard errors and small t values

3. We need to drop the redundant predictors

4. For instance, we first generate a new variable builtyear, which is perfectly correlated with age. Then lm function refuses to report its estimate

5. Next we change one value, but builtyear and age are still highly correlated (correlation coefficient is -0.9998533). The age and builtyear now both have small t values (-0.07818348 and 0.46088796). Both predictors are imprecisely estimated
Example of Multicollinearity

> builtyear = 1980-age
> summary(lm(price~age+agesq+builtyear))$coefficient

|                | Estimate  | Std. Error | t value | Pr(>|t|)   |
|----------------|-----------|------------|---------|------------|
| (Intercept)    | 98512.37  | 1904.18    | 51.73   | 6.69e-157  |
| age            | -1457.47  | 115.19     | -12.65  | 5.33e-30   |
| agesq          | 8.29      | 0.78       | 10.62   | 9.79e-23   |

> summary(lm(price~age+agesq+builtyear))$coefficient

|                | Estimate  | Std. Error | t value | Pr(>|t|)   |
|----------------|-----------|------------|---------|------------|
| (Intercept)    | -23794046| 5376396   | -0.44   | 0.65       |
| age            | -2115469 | 2705775   | -0.08   | 0.94       |
| agesq          | 8325787  | 7854185   | 10.60   | 1.13e-22   |
| builtyear      | 1251481  | 2715369   | 0.46    | 0.65       |

> cor(age, builtyear)

[1] -0.9998533
Lesson: do not include predictors that are highly correlated with each other. Some of them are redundant.
Outliers—noise maker

1. Outliers differ from the majority of observations, and cause the OLS estimation to be imprecise

2. Outliers are observations with Studentized Residuals greater than 1.96 in absolute value

3. In general, removing outliers can result in better fitting
A story of outlier
We use age, bathrooms and area as predictors. The results using all observations are below

```r
> reg4a = lm(price ~ age + baths + area)
> summary(reg4a)$sigma
[1] 21341.8
> summary(reg4a)$adj.r.squared
[1] 0.5847466
```
Detecting outliers—we find that the biggest outlier is a house with age 0 and price greater than 300000. That house is labeled as “outlier” in the first scatter plot.

```r
> library(MASS)
> ehat = studres(reg4a)
> data[abs(ehat) == max(abs(ehat)),]

   age area baths rprice
157  0 3770   3 3e+05
```
Estimation without outliers

```r
> index = which(abs(ehat)<1.96)
> reg4 = lm(price[index]~age[index]+baths[index]+area[index])
> summary(reg4)$coef

                     Estimate Std. Error  t value     Pr(>|t|) 
(Intercept) 16339.94945  3134.42567   5.213060 3.478268e-07 
age[index]  -250.91652   29.29527  -8.565088 5.882498e-16 
baths[index] 11632.92413  1570.40699   7.407586 1.331265e-12 
area[index]  20.68098    1.72512   11.988136 2.737291e-27 

> summary(reg4)$sigma
[1] 13758.88 

> summary(reg4)$adj.r.squared
[1] 0.7470752
```
Remarks

1. Removing outliers reduces the residual standard error from 21341.8 to 13758.9
2. The adjusted R squared increases from 0.58 to 0.75
3. In this case, removing outlier leads to significant improvement
4. Next we define \texttt{yhat} function that generates the predicted house price for given information about age, baths and area
Forecasting

\[
\text{yhat} = \text{function}(\text{ag}, \text{ba}, \text{ar}) \{
+ \text{fitted} = \text{reg4}$\text{coef}[1] + \text{reg4}$\text{coef}[2] \times \text{ag} + \text{reg4}$\text{coef}[3] \times \text{ba} + \text{reg4}$\text{coef}[4] \times \text{ar}
+ \text{cat("predicted price is", fitted,"\n")}
+ \}
\]

\[
> \text{yhat}(40,1,2000) \\
\text{predicted price is 59298.17} \\
> \text{yhat}(0,1,2000) \\
\text{predicted price is 69334.83}
\]
Mean vs Median

Symmetric Distribution

Mean
Median
Mode

Skewed Distribution

Mode
Median
Mean
Outliers causes the distribution to be asymmetric (skewed). Other than dropping outliers, we can estimate the conditional median using Quantile regression, which is less sensitive to outliers than conditional mean. Then the predicted price can be computed based on the quantile regression.
quantile regression

> library(moments)
> library(L1pack)
> skewness(price)

[1] 1.375726

> lad(price~age+baths+area, data=data, method = "EM")$coef

(Intercept)   age   baths   area
18879.7946   -261.6854   9747.6507   21.1412

> summary(lm(price~age+baths+area))$coef

             Estimate Std. Error   t value Pr(>|t|)   
(Intercept)   9959.99  4484.12 2.22117  0.027047
age           -201.37  40.7700 -4.93912  0.000019
baths         13641.6 2298.85  5.93410  0.000008
area          21.5844  2.3834  9.05613  0.000001
Remarks

1. The distribution of house price is not symmetric since its skewness is not close to 0

2. We see that the signs of coefficients are the same across regression and quantile regression, but the magnitude differ
Exercise

Suppose the age predictor is unavailable.

1. Find the predicted price for a house with area= 2000 and baths= 3 using regression and quantitle regression.

2. Explain whether it is a good idea to include squared baths in the regression.

3. Suppose there is a predictor called bedroom, which equals baths plus 1. Explain whether it is a good idea to include both baths and bedroom in the regression.