

Linear Regression

Synferlo

May, 10

1 Simple Linear Regression

1.1 Estimation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = \rho_{xy} \frac{s_y}{s_x}, \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

\bar{x}, \bar{y} : sample means

s_x, s_y : sample standard deviation

ρ_{xy} : the estimate of correlation between X and Y based on the data.

1.2 Statistical Inference

Table 1: Price and Volume (unnormalized result)

	Coefs	SE	t-value	p-value
(Intercept)	2.342e+03	8.799e+01	26.62	$< 2e - 16$ ***
Volume	2.696e-07	5.252e-09	51.33	$< 2e - 16$ ***
Multiple R-squared:	0.5061	Adjusted R-squared:	0.506	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3803 on 2571 degrees of freedom

F-statistic: 2635 on 1 and 2571 DF, p-value: $< 2.2e - 16$

p -value and t -value for the coefs are the results of a two- tailed test based on t -distribution

with $\text{DOF} = 2$.

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

where $j = 0$ for the intercept β_0 , and $j = 1$ for the coef. of the volume.

1.2.1 R^2 and adjusted R^2

$$R^2 = \rho_{xy}^2$$

In R,

```
# compute correlation then square it.  
cor(df$Open_price, df$Volume, use = 'complete.obs')**2  
# [1] 0.5061467
```

You can check this 0.506 with the R-squared we've got from *summary(result)*. They are exactly the same.

R-squared tells us 50 percent of the variation in the price can be attributed to volume.

The adjusted R-squared is important ONLY IF you are using the coef of determination to assess the overall quality of the fitted model in terms of a balance between goodness of fit(GOF) and complexity.

1.2.2 Other summary output

Residual standard error is the estimated SE of the ε , i.e., $\hat{\sigma}$.

1.3 Categorical Predictor

Explanatory variables can be categorical.

1.3.1 Binary Variables

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

X can be either 0 or 1.

If so, the interpretation of β would be different. It's better to think of them like two intercepts, where β_0 provides the baseline of the response when $X = 0$, and β_1 represents the additive effect on the mean response if $X = 1$.

1.3.2 Multilevel variables

The categorical variables have more than two levels. For example, there can be many levels under education, e.g., high school, college, master, phd, etc.

Suppose there are k levels, then variable X can be written as

$$X = 1, 2, 3, \dots, k$$

$$X_{(1)} = 0, 1 \quad X_{(2)} = 0, 1 \quad \dots X_{(k)} = 0, 1$$

And we can write the model,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_{(2)} + \hat{\beta}_{(2)} X_{(3)} + \dots + \hat{\beta}_{k-1} X_{(k)}$$

We normally use $k - 1$ of the dummy variables. Also, each observation only satisfy one of the levels, i.e., if you are a Ph.D. student, then you cannot be a high school student. Hence, when $X_{(i)} = 1$, all others equals to zero.

So, if one observation belongs to level 3, then the model (the predicted mean) would be

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2$$

Since the reference level(that omitted dummy) is defined as 1, if an observation has values of zero for all other dummies, it implies the obs originally had $X = 1$.

$$\hat{y} = \hat{\beta}_0$$

Here we have the result of a model with a dummy has four levels (Heavy, Never, Occas, Regular), where Heavy is the reference.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	173.7720	3.1028	56.005	<2e-16 ***
SmokeNever	-1.9520	3.1933	-0.611	0.542
SmokeOccas	-0.7433	3.9553	-0.188	0.851
SmokeRegul	3.6451	4.0625	0.897	0.371

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The observation in the reference category Heavy is represented solely by $\hat{\beta}_0 = 173.7720$.

2 Multiple Linear Regression

2.1 Terminology

A lurking variable, is what we've learned of the omitted variable. It can lead to a omitted variable bias.

A nuisance or extraneous variable is a predictor of no interest, but has the potential to confound (mess up) relationships between other variables, and so affect your estimation. We are not interested in it, but we must add it to the model.

2.2 The model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

You have n obs and p explanatory variables. For each obs, the regression equation would be

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_p x_{p,i} + \varepsilon_i$$

where $i = 1, 2, \dots, n$ stand for the i^{th} obs.

Least-squared:

$$\min \left(\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_p x_{p,i}))^2 \right)$$

Matrix Form:

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where \mathbf{Y} and $\boldsymbol{\varepsilon}$ are $n \times 1$ matrices

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{p,1} \\ 1 & x_{1,2} & \cdots & x_{p,2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,3} & \cdots & x_{p,3} \end{bmatrix}$$

Matrix \mathbf{X} is $n \times (p + 1)$ dimension.

The OLS estimator $\hat{\beta}$:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{Y}$$

2.3 Interpretation of the coef

A coefficient for a specific variable should be interpreted as the change in the mean response for a one-unit increase in the variable, while holding all other variables constant.

2.4 Transformation

Two ways to approach the transformation: polynomial and logarithmic.

The transformation in general does not represent a universal solution to solving problems of nonlinearity in the trend, but it can at least improve how faithfully a linear model is able to represent those trends.

2.4.1 Polynomial

Add squared, cubic, and other polynomial terms to fit curved trend. For example,

$$\text{Income} = \beta_0 + \beta_1 \text{edu} + \beta_2 \text{edu}^2 + \beta_3 \text{exp} + \beta_4 \text{exp}^2 + \varepsilon$$

Use $I()$ function in R to add a polynomial term in $lm()$ function.

Code:

```
lm_poly_result = lm(norm_price ~ norm_volume + norm_supply
+ I(norm_supply**2), data = df)
summary(lm_poly_result)
```

If the effect of adding one term is not good, we can try to add another cubic term and compare these two models.

Table 2: Quadratic vs. Cubic term model

	Model 1	Model 2
(Intercept)	-0.024260 (-1.523)	-0.095378 *** (-5.557)
norm_volume	0.465420*** (28.338)	0.418280 *** (24.920)
norm_supply	0.397987*** (24.030)	0.525587 *** (25.473)
$I(norm_supply^2)$	0.024269* (2.446)	0.088337 *** (7.588)
$I(norm_supply^3)$		-0.030171 *** (-10.035)

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
, 1

Now we can see, by adding cubic term, the model performs better.

2.4.2 Logarithmic

Transforming to a logarithmic scale can help reduce the severity of heavily skewed data.

2.5 Interaction Terms

When estimating regression models, you always have to accompany interactions with the main effect of the relevant predictors.

In R, use a “:” specify an interaction term.

Code:

```
inter_lm_result = lm(
  norm_price ~ norm_volume + norm_supply + norm_volume : norm_supply ,
  data = df
)
summary(inter_lm_result)
```


3 Linear Model Selection

3.1 Goodness of fit vs. Complexity

GOF: refers to the goal of obtaining a model that best represents the relationships between the dependent and explanatory variables.

Complexity: how many terms (e.g., polynomial, etc) and additional functions in your model. The more you have, the more complex the model would be.

The principle of parsimony: The balancing act between GOF and complexity.

Our goal is to find a model that is as simple as possible without sacrificing too much GOF.

The model satisfies this notion is a parsimonious fit.

3.1.1 General Guideline

1. You CANNOT remove individual levels of a categorical variable in a given model. Suppose college, master, phd are under edu, but only phd are significant. You cannot remove the college and master. You can ONLY remove the entire categorical variable, i.e., edu.
2. If an interaction term is present in the fitted model, ALL lower- order interactions and main effects of the relevant variables MUST remain in the model. Suppose you add an interaction term, $edu \times exp \times age$, then the main effect (edu, exp, age) and all lower-order interaction terms ($edu \times exp, edu \times age, exp \times age$) should be shown in the model. Hence, in R, you'd better use $VR_1 * VR_2$ for interaction term, so that R will add all these terms for you. You will not miss any one of them.
3. Keep ALL lower-order polynomial terms in the model if the highest is deemed significant.

3.2 Model Selection Algorithms

3.2.1 Nested Comparisons: The Partial F-Test

It is the most direct way to compare several different models. It looks at two or more nested models. The less complex model is a reduced version of the more complex model.

$$\begin{aligned}\hat{y}_{\text{redu}} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p \\ \hat{y}_{\text{full}} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p + \dots + \hat{\beta}_q x_q\end{aligned}$$

Clearly, \hat{y}_{full} is more complex than \hat{y}_{redu} , so we say that \hat{y}_{redu} is nested within \hat{y}_{full} .

The **partial F-Test** tries to answer if it is worth it to include any additional variables. Its goal is to test whether include those extra $q - p$ terms in \hat{y}_{full} provide a statistically significant improvement GOF.

$$H_0 : \beta_{p+1} = \beta_{p+2} = \dots = \beta_q = 0$$

$$H_1 : \text{At least one of the } \beta_j \neq 0 \text{ (for } j = p, \dots, q)$$

If the p -value is less than the significant level, then we say it is worth it because at least one of those extra $q - p$ terms is non-zero.

F statistics:

$$F = \frac{(R_{\text{full}}^2 - R_{\text{redu}}^2)(n - q - 1)}{(1 - R_{\text{full}}^2)(q - p)}$$

It follows an F distribution with $df_1 = q - p$, $df_2 = n - q$ degree of freedom. The p -value is found as the upper-tail area from F as usual.