STAC67: Regression Analysis

Lecture 15

Sohee Kang

Mar. 10, 2021

Example

Y: Speed of innovation, X_1: size of a insurance firm, X_2: type of firm

0 : mutual

1 : stock

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

Initial data:

	Y	X_1	X_2
1	17	151	Mutual
2	26	92	Mutual
3	21	175	Mutual
⋮	⋮	⋮	⋮
10	16	238	Mutual
11	28	164	Stock
12	15	272	Stock
13	11	295	Stock
14	38	68	Stock
⋮	⋮	⋮	⋮
20	14	246	Stock

Recoded data:

	Y	X_1	X_2
1	17	151	0
2	26	92	0
3	21	175	0
⋮	⋮	⋮	⋮
10	16	238	0
11	28	164	1
12	15	272	1
13	11	295	1
14	38	68	1
⋮	⋮	⋮	⋮
20	14	246	1

$$E(Y) = \begin{cases} \beta_0 + \beta_1 X_1 \\ (\beta_0 + \beta_2) + \beta_1 X_1 \end{cases}$$

$H_0 : \beta_2 = 0$

is testing whether type of firm has effect on innovation.

Eg. whether they have same intercept.

R codes

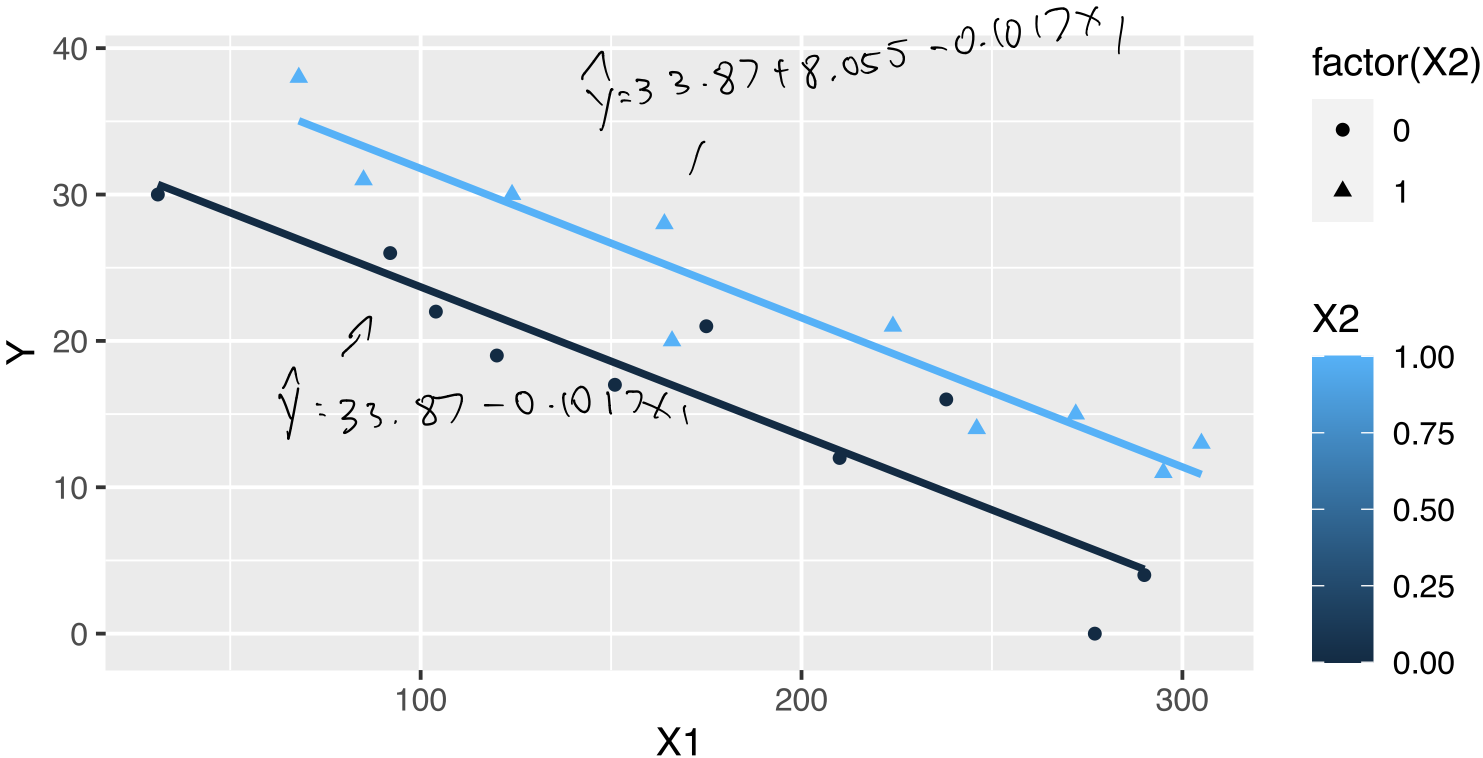
```r
Innovation = read.table("Table8-2.txt", header=F,  col.names=c("Y","X1","X2"))
fit = lm(Y~X1 + X2, data=Innovation)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = Innovation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6915 -1.7036 -0.4385  1.9210  6.3406
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.874069   1.813858  18.675 9.15e-13 ***
## X1          -0.101742   0.008891 -11.443 2.07e-09 ***
## X2           8.055469   1.459106   5.521 3.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.221 on 17 degrees of freedom
## Multiple R-squared:  0.8951, Adjusted R-squared:  0.8827
## F-statistic:  72.5 on 2 and 17 DF,  p-value: 4.765e-09
```

R codes

```r
library(ggplot2)
ggplot(data=Innovation, aes(x=X1, y=Y, color=X2, shape=factor(X2))) + geom_point() + geom_smooth(method='lm', f
```

```
## `geom_smooth()` using formula 'y ~ x'
```



$\hat{y} = 33.87 + 8.055 - 0.1017x_1$

$\hat{y} = 33.87 - 0.1017x_1$

Review

- A test of whether same line is appropriate for all levles of X_2 can be done by fitting the reduced model with X_1 only and comparing the residual sum of squres as we did before.

Example) Y: tool wear X_1 : tool speed
X_2: tool model (4 models)

- Quantify the qualitative predictor, 4 levels:

$D_1 \begin{cases} 1 & TM = 1 \\ 0 \end{cases}$ $D_2 \begin{cases} 1 & TM = 2 \\ 0 \end{cases}$

$D_3 \begin{cases} 1 & TM = 3 \\ 0 \end{cases}$ Basline Var TM = 4

X_2	D_1	D_2	D_3
M_1	1	0	0
M_2	0	1	0
M_3	0	0	1
M_4	0	0	0

Interpretation

$$y = \beta_0 + \beta_1 X_1 + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_3 + \varepsilon$$

$$E(y) = \begin{cases} \beta_0 + \beta_1 X_1 & TM = 4 \\ \beta_0 + \beta_2 + \beta_1 X_1 & TM = 1 \\ \beta_0 + \beta_3 + \beta_1 X_1 & TM = 2 \\ \beta_0 + \beta_4 + \beta_1 X_1 & TM = 3 \end{cases}$$

β_2: How much higher/lower the response function for tool model in the M1, than one for M4 for any given level of tool speed

β_3, β_4 defined similarly

$H_0: \beta_2 = \beta_3 = \beta_4 = 0$

H_a: at least one $\beta_i \neq 0$

$$F = \frac{(SSE_R - SSE_F)/(5-2)}{MSE_F} \sim F(3, n-5)$$

More general setting

- One categorical variable and multiple continous variables, the interpretation is same: the only thing that changes for different level of the categorical variable is the **intercept** of the model.
- The effect of continous covariates is assumed to be the same for all levels of categorical variables.
- If there are more than one categorical variables then there are more groups and more different intercepts.
- **Additive Structure**: A regression model with p predictor variables is **additive** (or contain additive effects) if it can be written

$$Y_i = f_1(X_{i1}) + f_2(X_{i2}) + \ldots + f_p(X_{ip})$$

  for any functions f_k, i.e. there is no intearaction terms (cross product terms)
- We have made two assumptions so far

1. level of categorical variable only alter the intercept of the model, not its slope on any of the continuous covariates.
2. the change in the intercept is additive over multiple categorical variables.

Interaction

- However,for example, One's initial salary depends on eduction level (categorical) and that increases are with year of experiene (continuous) and also depends on education level
- Interaction Term: can be used to expand the linear model to deal with these siuations : take product of two variables

Interaction of 2 categorical variables

X: , Z_1 Z_2:

Interaction

- If both categorical variables have 3 categories:
- Including interactions increases the number of parameters to etimate.
- In many cases, the additive structure fits the data well.

Interaction between a categorical and a continous variable

Example: Insurance Innovation Example

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 D_i + \beta_3 D_i X_{i1} + \epsilon_i$$

- The estimated model:
- Test whether the effect of firm size changes with the firm type.
- Conclusion:

R codes:

```
Innovation = read.table("Table8-2.txt", header=F,  col.names=c("Y","X1","X2"))
fit = lm(Y~X1*X2, data=Innovation)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X1 * X2, data = Innovation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7144 -1.7064 -0.4557  1.9311  6.3259
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.8383695  2.4406498  13.864 2.47e-10 ***
## X1          -0.1015306  0.0130525  -7.779 7.97e-07 ***
## X2           8.1312501  3.6540517   2.225   0.0408 *
## X1:X2       -0.0004171  0.0183312  -0.023   0.9821
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.32 on 16 degrees of freedom
## Multiple R-squared:  0.8951, Adjusted R-squared:  0.8754
## F-statistic: 45.49 on 3 and 16 DF,  p-value: 4.675e-08
```

Example

- Samples of male atheletes from the National Basketball Association (NBA), National Hockey League (NHL), and English premier (Football) Leagure are obtained, and the relationship between players' Weight (Y) and Height (X) is measured.
- NBA $=$ NHL $=$
- Full Model:

1. Test the identitiy of three regression functions

R codes

```r
Player = read.csv("sample.csv", header=T)
fit = lm(Weight ~ Height + NBA + NHL, data=Player)
fit2 = lm(Weight~Height, data=Player)
summary(fit2)
```

```
##
## Call:
## lm(formula = Weight ~ Height, data = Player)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.504 -13.353  -1.287  13.128  36.647
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -278.5315    37.9272  -7.344 7.75e-10 ***
## Height         6.3585     0.5071  12.539  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17 on 58 degrees of freedom
## Multiple R-squared:  0.7305, Adjusted R-squared:  0.7259
## F-statistic: 157.2 on 1 and 58 DF,  p-value: < 2.2e-16
```

```r
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: Weight
```

Example: Interaction Model

- Full Model:

  $$Y_i = \beta_0 + \beta_1 Height + \beta_2 NBA + \beta_3 NHL + \beta_4 Height * NBA + \beta_5 Height * NHL + \epsilon_i$$

2) Test the equality of slopes of three regression models:

R codes

```
fit2 = lm(Weight ~ Height*NBA + Height*NHL , data=Player)
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: Weight
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## Height      1  45416   45416 306.2009 < 2.2e-16 ***
## NBA         1   2157    2157  14.5413 0.0003541 ***
## NHL         1   5587    5587  37.6672 1.026e-07 ***
## Height:NBA  1    944     944   6.3644 0.0146215 *
## Height:NHL  1     57      57   0.3867 0.5366565
## Residuals  54   8009     148
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
\begin{verbatim}
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -201.914     73.134  -2.761  0.00786 **
Height         5.136      1.032   4.976 6.98e-06 ***
NBA         -184.599     98.395  -1.876  0.06605 .
NHL          105.640    119.624   0.883  0.38110
Height:NBA     2.513      1.326   1.895  0.06345 .
Height:NHL    -1.020      1.641  -0.622  0.53666
\end{verbatim}
```

R codes

```
library(ggplot2)
ggplot(data=Player, aes(x=Height, y=Weight, color= factor(League), shape=factor(League))) + geom_point() + geom_
```

```
## `geom_smooth()` using formula 'y ~ x'
```