ANALYSIS OF VARIANCE (ANOVA)

ANOVA- ANCOVA

Analysis of Variance – Analysis of Covariance

- ANOVA: no covariate; one dependent variable
- ANCOVA: covariate/s; one dependent variable

are regression and the **general linear model**

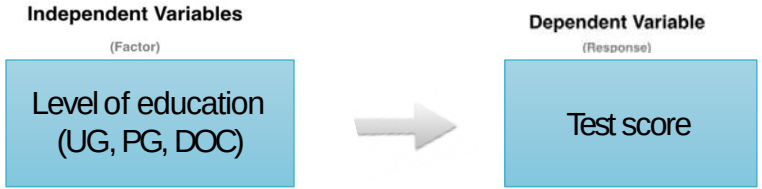
- The term "***factor***" (independent variable) refers to the variable that distinguishes this group membership →categorical var. Race, level of education, and treatment condition are examples of factors.

- The term "***response***" (dependent variable) is a continuous variable (i.e. scale or interval) that we test three or more groups for mean differences.

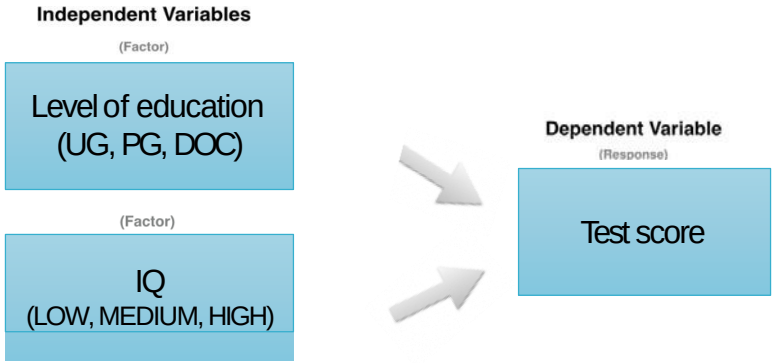
- ***Covariate*** is the term for the continuous independent variable (IV) used in ANCOVA and MANCOVA.

ANOVA

One way ANOVA example



Two way ANOVA example



Two main types of ANOVA:

- (1) "one-way" ANOVA compares levels (i.e. groups) of a single factor based on single continuous response variable
- (2) "two-way" ANOVA compares levels of two or more factors for mean differences on a single continuous response variable

- In practice, you will see one-way ANOVA more often and when the term ANOVA is generically used, it often refers to a one-way ANOVA.

ANOVA

- **Null hypothesis**

There are no mean differences between groups in the population

With hypothesis testing we are setting up a null-hypothesis – *the probability that there is no effect or relationship* – and then we collect evidence that leads us to either accept or reject that null hypothesis.

- **Research Question**

Do one or more grouping variables each with two or more levels differ in terms of their means on a metric dependent variable?

ANOVA Assumptions

- Observations should be independent (Independence)
- Normal distribution (Normality)
- Equal variances (homogeneity)

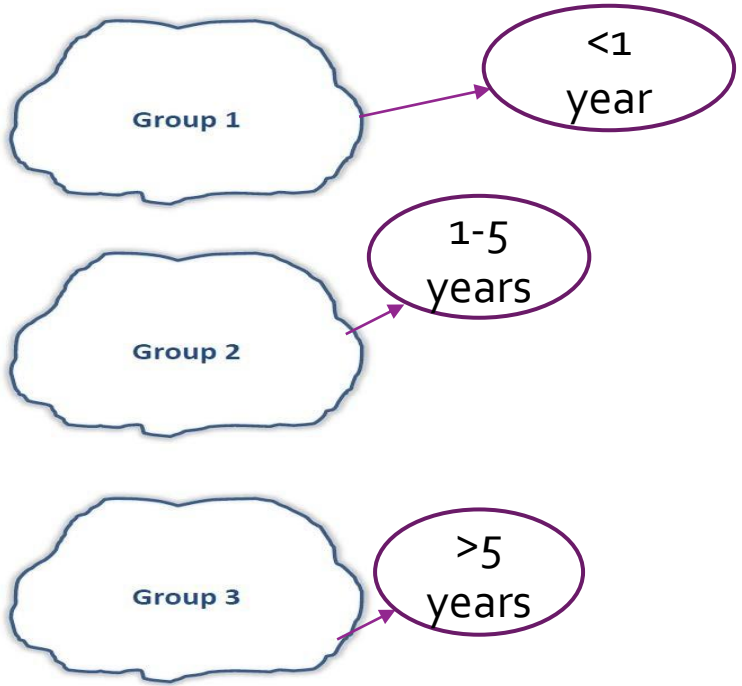
- *In our sample:*

Factor (IV)

x1 (customer type) – cat variable

Response (DV)

x6 (product quality) – cont variable



- More than 2 groups
- Hypothesis:

Ho: There will be no difference among types of customer on the product quality.

- One Way/ Two Way ANOVA ?

- Independence
- Normality
- Homogeneity

- Normal distribution:

(2ways)

- * Normality plot
- * KS test or SP test

    Kolmogorov-Smirnov (large sample size)

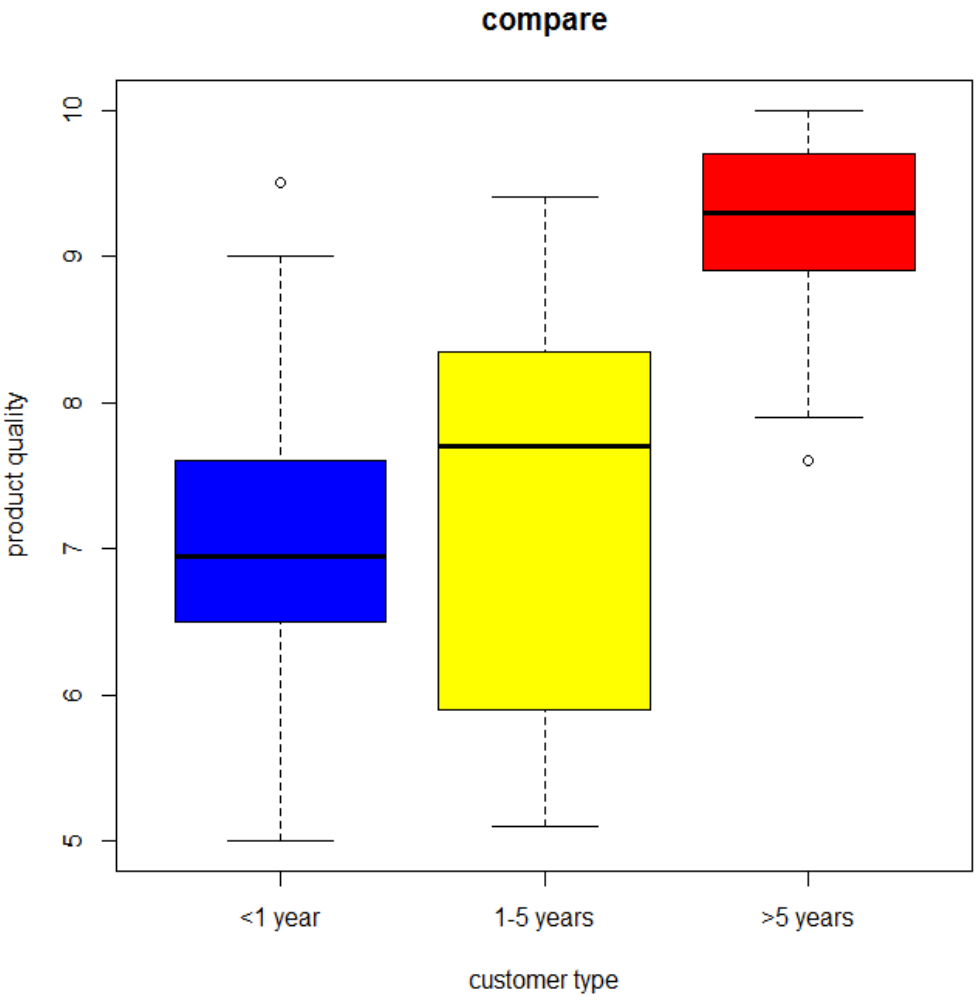
    Shapiro-Wilk (small sample size)

- *** We want to see **non-significant results**

- Normality plot:

Normality assumption is not satisfied

Outliers check: 2 outliers are identified via this plot



KS or SP TESTS

Shapiro-Wilk normality test

data: x$x6
W = 0.94985, **p-value = 1.824e-06**

Lilliefors (**Kolmogorov-Smirnov**) normality test

data: x$x6
D = 0.095142, p-value = 0.000148

Our sample includes 200 observations → small dataset, we better use SW for interpretation

→ We want to see **non-significant** result. However, p_value < 0.01 (level of significance)

→ SW tests give significant results →Normality assumption is not satisfied

Outliers

```
> print(outliersx6) [1] 9.5 9.5 7.6 7.6
```

```
> outliersx6
     id x1 x2 x3 x4 x5  x6  x7  x8  x9 x10 x11 x12 x13 x14 x15
13   13  1  1  0  0  1 9.5 5.6 4.6 6.9 5.0 6.9 6.6 7.6 6.5 5.3
58   58  3  1  0  0  0 7.6 3.6 5.2 5.8 5.6 6.6 5.4 4.4 6.7 6.4
104 104  1  1  1  0  1 9.5 5.6 0.4 5.5 5.6 6.9 6.6 7.6 3.8 4.4
196 196  3  1  1  0  0 7.6 3.6 2.1 5.2 4.8 6.6 5.4 4.4 4.3 6.8
    x16 x17 x18 x19 x20 x21  x22 x23 newvar
13  5.1 4.5 4.4 8.4 8.4 7.9 58.1   1      2
58  4.6 3.9 4.0 8.2 7.5 7.5 58.1   1      6
104 5.6 4.5 4.4 8.4 9.4 9.0 58.1   1      2
196 4.4 3.9 4.0 8.2 6.9 8.4 58.1   1      6
```

- All identified outliers should be removed!

- Homogeneity of variance:
- Using Levenne's test

Based on Mean
Test Statistic = 29.211, p-value = 7.763e-12
Based on Median
Test Statistic = 18.322, p-value = 5.039e-08
Based on trimmed mean
Test Statistic = 24.755, p-value = 2.565e-10

- We want to see **a non-significant result**
- All p_value are less than 0.01 (level of significance)→ the results are significant.
- →Homogeneity is not met

→ Not enough conditions to perform ANOVA.
However we still can do ANOVA by accepting assumption that variances not the same and choose **Tamhane's T2** instead of **LSD** or **Tukey**

One way ANOVA

*** ANOVA table**

TYPE I	Df	Mean Sq	F value	Pr(>F)
as.factor(x1)	2	180.6 90.29	**88.91**	**<2e-16 *****
Residuals		197 200.1	1.02	

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

TYPE III

Model:

x6 ~ as.factor(x1)

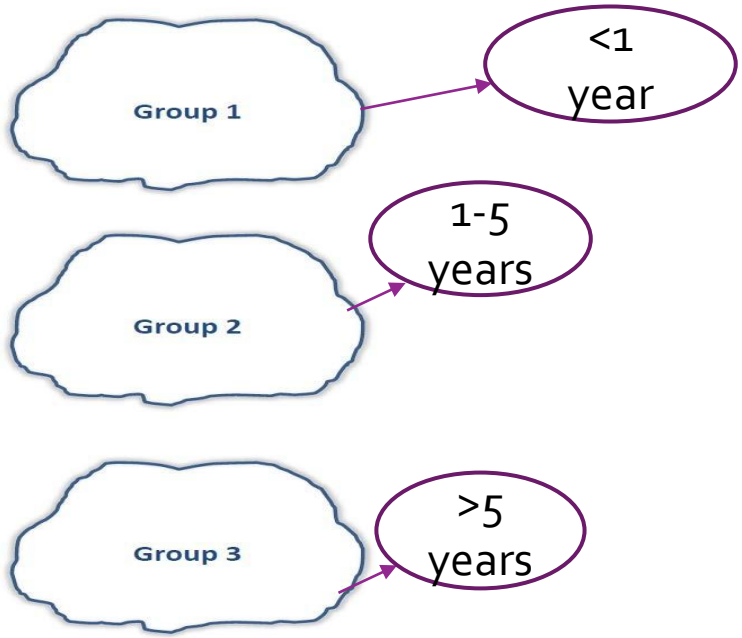
	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			200.06	6.06		
as.factor(x1)	2	180.57	**380.63**	130.70	**88.906**	**< 2.2e-16 *****

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

- *In our sample:*

Factor (IV)

x1 (customer type) – cat variable



Response (DV)

x6 (product quality) – cont variable

Looking at the F_value which is equal to 88.91 and p=0.000<0.001

Report the results

P=0.000<0.01, we can reject the null hypothesis and conclude that:

There is a statistically significant difference within groups of product quality as determined by one-way ANOVA ($F(2,197) = 88.91$, $p = .000$)

Multiple comparison

Pairwise comparisons using **Tamhane's T2-test** for unequal variances

data: x$x6 and as.factor(x$x1)
alternative hypothesis: two.sided
P value adjustment method: T2 (Sidak)
Ho

	t value	Pr(>|t|)
2 - 1 == 0	0.794	0.8134
3 - 1 == 0	14.446	<2e-16 ***
3 - 2 == 0	10.762	<2e-16 ***

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

The differences between groups in factor 1

P_value

Customer Type (Factor)

1: <1 year

2: 1-5 years

3: >5 years

***Multiple comparison**

In the relationship with X6 (product quality) there is not a significant difference between less than 1 year and 1 to 5 years predictors (p_value = 0.813) while there are significant differences between less than 1 year and over 5 years (p_value = 0.000); and 1 to 5 years and over 5 years (p_value = 0.000).

The mean difference between over 5 years and less than 1 year is the highest.

A Tamhane post hoc test revealed that the quality was statistically significantly increased after taking the customer type over 5 years) compared to the 1 to 5 years and less than 1 year.

2-way ANOVA

In our sample:

Factors:
x1 (customer type)
x2 (industry type- 0: magazine industry, 1: newsprint industry)

Response:
x6 (product quality)

* Assumption check:

1. Normality
2. Homogeneity of variances
3. Independence

- Normal distribution:

(2ways)

- * Normality plot
- * KS test or SP test

Kolmogorov-Smirnov (large sample size)

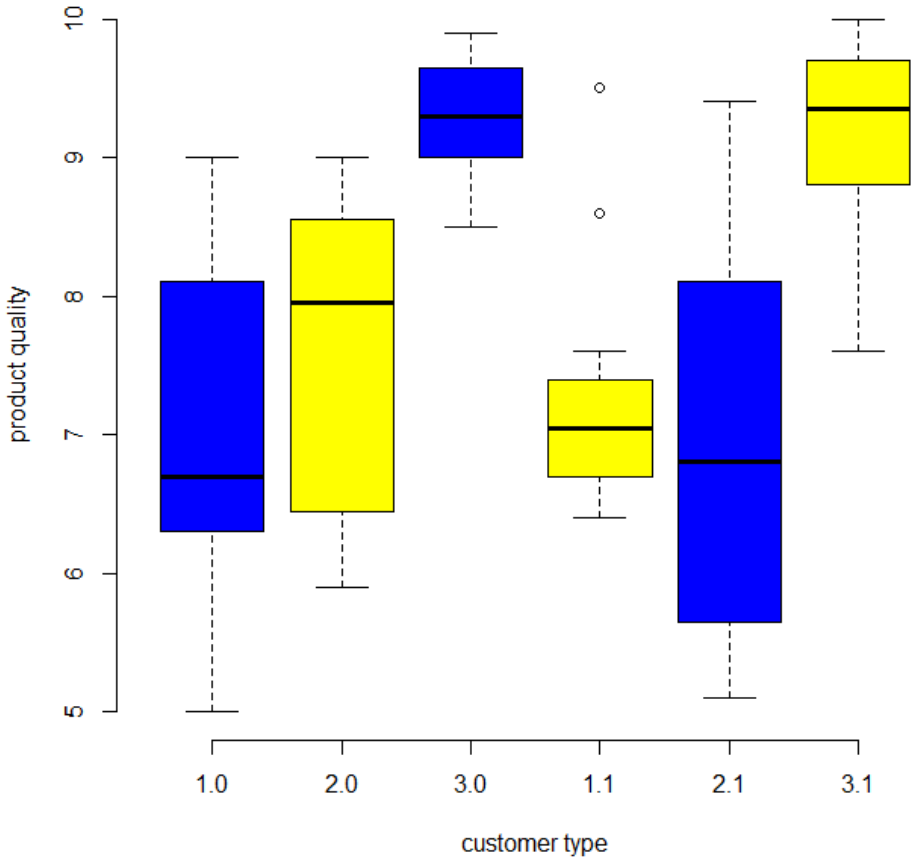
Shapiro-Wilk (small sample size)

*** We want to see non-significant results

- Normality plot:

Normality assumption is not satisfied (x1 blue, x2 yellow)

Outliers check: 2 outliers are identified via this plot



KS or SP TESTS

Shapiro-Wilk normality test

data: x$x6
W = 0.94985, p-value = 1.824e-06

→ We want to see **non-significant** result. However, p_value < 0.01 (level of significance)

→ SP tests give significant results →Normality assumption is not satisfied

- Homogeneity of variance:
- Using Levenne's test

Based on Mean
Test Statistic = 29.211, p-value = 7.763e-12
Based on Median
Test Statistic = 18.322, p-value = 5.039e-08
Based on trimmed mean
Test Statistic = 24.755, p-value = 2.565e-10

- We want to see a non-significant result
- All p_value are less than 0.01 (level of significance)→ the results are significant.

→Homogeneity is not met

→ Not enough conditions to perform ANOVA.
However we still can do ANOVA by accepting assumption that variances not the same and choose Tamhane's T2 instead of LSD or Tukey

- ANOVA Table

Main effect

Interaction effect

	Df	Sum Sq	Mean Sq	**F value**	**Pr(>F)**
as.factor(x1)	2	180.57	90.29	91.380	<2e-16 ***
as.factor(x2)	1	2.41	2.41	2.437	0.120
(x1):(x2)	2	5.97	2.99	3.023	0.051 .
Residuals			194	191.68	0.99

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

In these results, you can conclude the following, based on the p-values and a significance level of 0.05:

The p-value for x1 is 0.000, there is significant differences between levels of x1 in relationship with x6, which indicates that the levels of customer types are associated with different product quality.

•The p-value for x2 is 0.120 there is non-significant differences between levels of x2 in relationship with x6, which indicates that the levels of industry types are not associated with different product quality.

•The p-value for the interaction between x1*x2 is 0.051, there is significant interaction between x1 and x2 at 0.1 l.o.sig, which indicates that the relationship between customer types and product quality depends on the value of industry type.

Fit: aov(formula = x6 ~ as.factor(x1) + as.factor(x2) + as.factor(x1):as.factor(x2), data = x)

$`as.factor(x1)`

	diff	lwr	upr	p adj
2-1	**0.1615809**	-0.2472839	0.5704457	**0.6198593**
3-1	**2.0794118**	1.6767896	2.4820340	**0.0000000**
3-2	**1.9178309**	1.5089661	2.3266957	**0.0000000**

$`as.factor(x2)`

	diff	lwr	upr	p adj
1-0	**-0.2191765**	-0.4964235	0.05807059	**0.1205865**

Factor x1
In the relationship with X6 (product quality) there is not a significant difference between less than 1 year and 1 to 5 years predictors (p_value = 0.813) while there are significant differences between less than 1 year and over 5 years (p_value = 0.000); and 1 to 5 years and over 5 years (p_value = 0.000).

The mean difference between over 5 years and less than 1 year is the highest.

A Tukey post hoc test revealed that the quality was statistically significantly increased after taking the customer type over 5 years) compared to the 1 to 5 years and less than 1 year.

Factor x2:
Because the p=0.120>0.05, we can conclude that
There is non significant difference between magazine and newsprint industry.

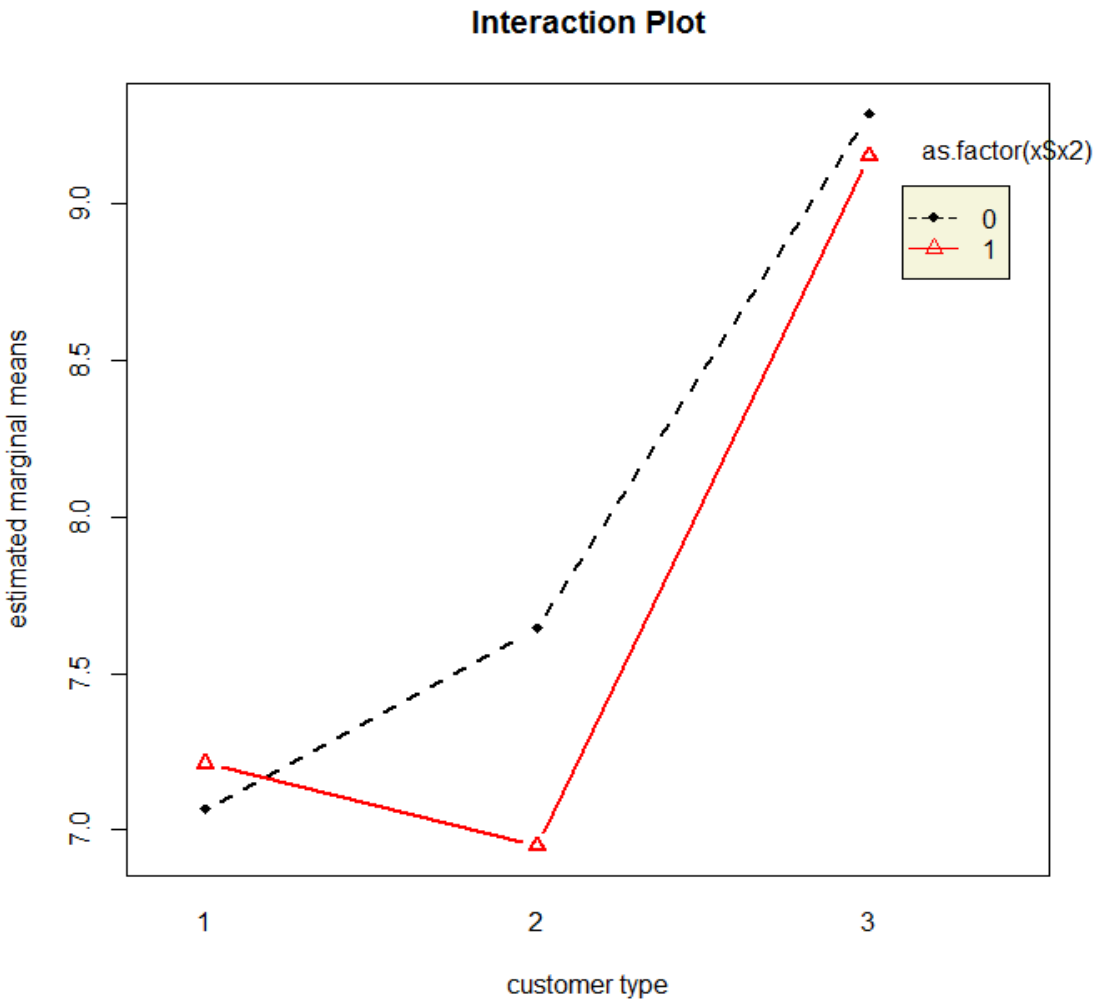
E.g.
There is significant different product quality of the magazine industry between level of customer >5 years and <1 year

Pairwise comparisons using **Tamhane's T2-test** for unequal variances

data: x$x6 and as.factor(x$x1):as.factor(x$x2)
alternative hypothesis: two.sided
P value adjustment method: T2 (Sidak)

	t value	Pr(>|t|)
1:1 - 1:0 == 0	0.601	0.99999
2:0 - 1:0 == 0	2.146	0.41881
2:1 - 1:0 == 0	-0.362	1.00000
3:0 - 1:0 == 0	10.522	**1.4355e-12** ***
3:1 - 1:0 == 0	8.978	**1.9178e-11** ***
2:0 - 1:1 == 0	1.859	0.65281
2:1 - 1:1 == 0	-0.899	0.99910
3:0 - 1:1 == 0	12.849	**< 2.22e-16** ***
3:1 - 1:1 == 0	10.301	**5.4956e-14** ***
2:1 - 2:0 == 0	-2.208	0.37993
3:0 - 2:0 == 0	8.270	**3.8605e-09** ***
3:1 - 2:0 == 0	6.818	**1.0959e-07** ***
3:0 - 2:1 == 0	8.772	**2.3922e-09** ***
3:1 - 2:1 == 0	7.763	**1.1491e-08** ***
3:1 - 3:0 == 0	-0.947	0.99834

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



As can be seen magazine industry produce the higher quality of product for customer type 2 and 3 while it has higher impact on product quality for customer type 3

In addition, customer type 3 has the highest influences on the product quality regardless of industry types.