در فرایند تولید هر محصول در صنایع غذایی یکی از مهمترین بخشهای پروسه ی تولید، کنترل کیفیت و ارائه ی بهترین محصول به مصرف کننده است. در هر محصول عوامل گوناگونی در کیفیت نهایی می توانند اثر گذار باشند، از جمله عوامل محیطی (مانند آب و هوا) محل کشت مواد اولیه (در اینجا میوه) ، تغییرات اقلیمی، نوع خاک، نگهداری و حمل مواد اولیه، کیفیت تجهیزات کارخانه ی صنعتی محل تولید، دمای محیط حین رخداد واکنشها، میزان نور و روشنایی و موارد این چنینی، که همه ی اینها منجر به دشواری فرایند کنترل کیفیت میشوند.

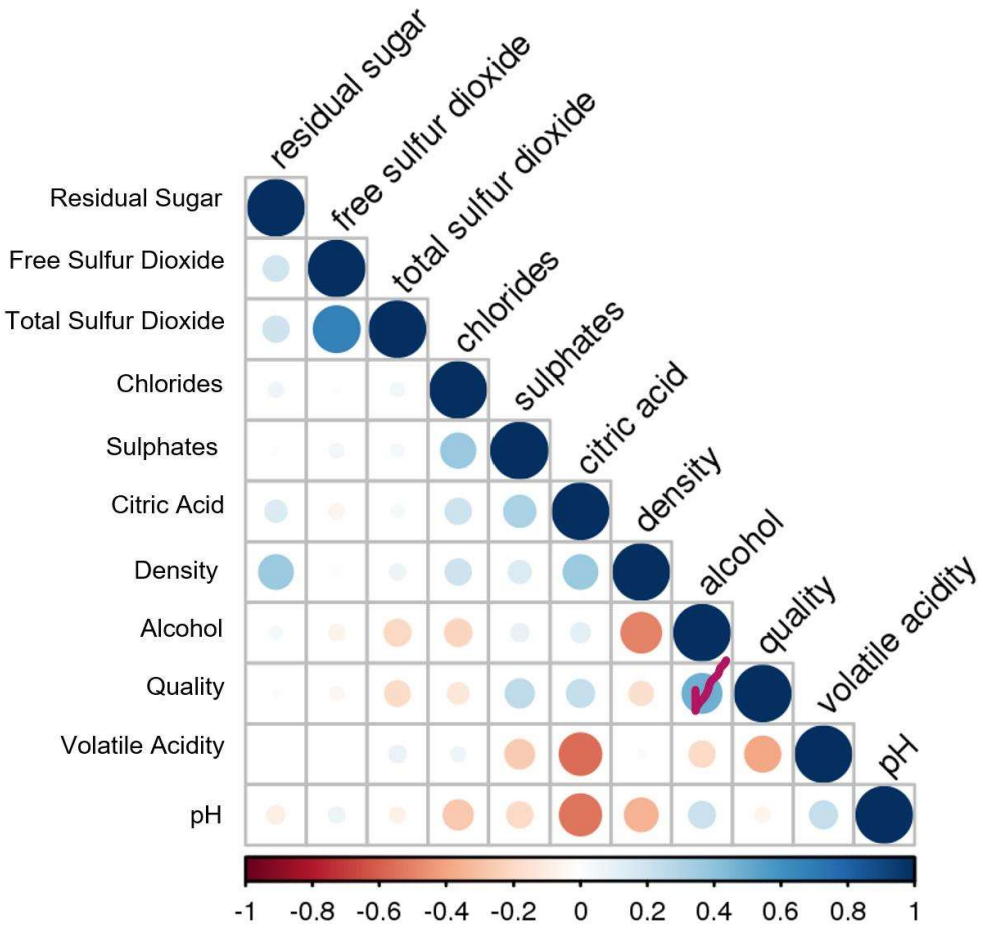
حقیقتی که وجود دارد این است که اغلب تاثیرات عوامل ذکر شده که بسیار اثرگذارند قابل سنجش و یا محاسبه نیست (به جز شراط کارخانه در صنعت). در حالی که محصول صنایع غذایی مستقیما به مصرف کننده می رسد و اثر زیادی در سلامت افراد دارد.

در این مجموعه دیتای ارائه شده، کیفیت محصول از طریق تست توسط متخصصین نوشیدنی مورد نظر ارائه شده است. همچنین دیگر متغیر های فیزیوکمیکال از طریق انالیز نمونه ها در ازمایشگاه بدست امده اند.

این دیتاست توسط دانشگاه University of Minho, Guimarães از کشور پرتغال ارائه شده است و هدف از تحلیل و آنالیز آن افزایش کیفیت نوشیدنی vinho verde میباشد.

با بدست آوردن مدل رگرسیونی این دیتاست میتوان حالت بهینه برای با کیفیت ترین محصول را یافت و کیفیت نمونه های جدید کارخانه را از دیدگاه مصرف کننده تخمین زد.

مهمترین متغیر توضیحی‌ای که برای برازش این قسمت انتخاب کرده‌ایم اتانول می‌باشد. با توجه به جدول ضرایب همبستگی که در فاز اول رسم کردیم، اتانول بیشترین تاثیر را در کیفیت نوشیدنی خواهد داشت.

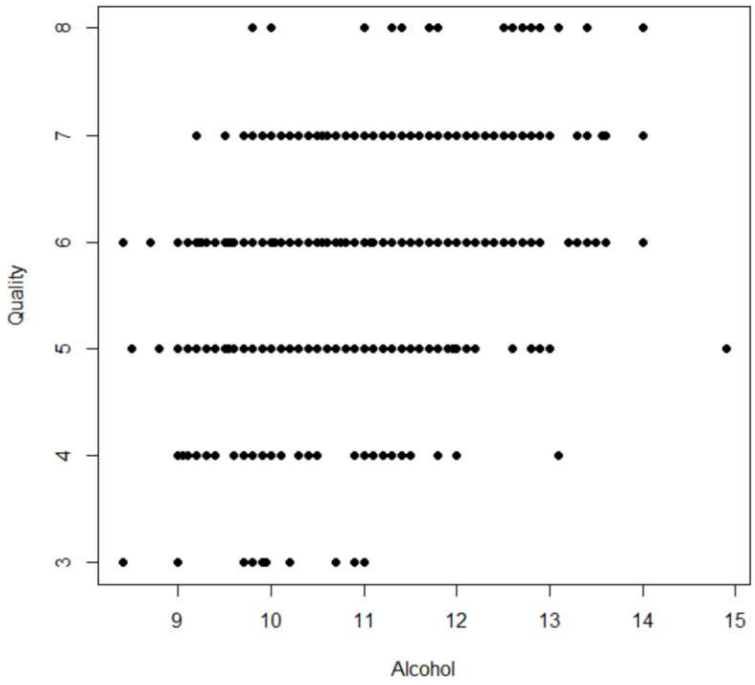


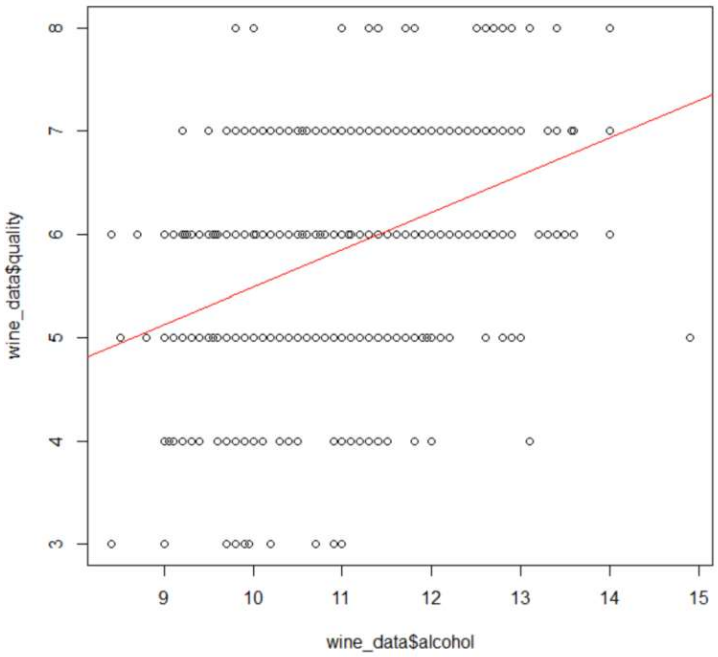
از جهت دیگر اتانول موجود در مایع شرایط را برای انجام واکنشهای شیمیایی بیشتر و بهتر فراهم می کند و فرایند تولید کامل تر طی خواهد شد. معاده ی مدل براز داده شده به صورت زیر حاصل شد:

$$y = \beta_0 + \beta_1 x$$

$$\beta_0 = 1.87497 \quad , \quad \beta_1 = 0.36084$$

نمودار پراکنش داده ها (میزان کیفیت بر حسب میزان اتانول)





تفسیر ضرایب رگرسیونی

اگر میزان حجمی که نمونه از اتانول به خود اختصاص داده است صفر باشد، کیفیت نوشیدنی نمونه، ۱.۸۷۴۹۷ خواهد بود، و به ازای افزایش هر درصد اتانول موجود در نمونه، کیفیت نمونه ۰.۳۶۰۸۴ واحد افزایش خواهد یافت.

نتیجه ی آزمون معنا داری

باتوجه به رد شدن H_0 نتیجه میگیریم که اولا، رابطه‌ای قوی و خطی بین x و y انتخابی وجود دارد و ثانیا، علاوه بر رابطه‌ی خطی، رابطه‌ای بهتری نیز وجود دارد که رابطه ی x و y را دقیق تر توصیف می کند. (که به دلیل پیچیدگی محاسبات از این رابطه استفاده میکنیم.)

ضریب تعیین

$$R^2 = 0.2267$$

ضریب تعیین نسبتی از تغییر پذیری کل پاسخ است که مدل رگرسیونی آن را با x توضیح می دهد. از آنجایی که این ضریبت همواره بین صفر و یک است، هرچه به یک نزدیکتر باشد برای ما مطلوبیت بیشتری دارد.

آزمون صفر بودن عرض از مبدا

برای β_0 داریم : $T = 10.73$ <2e-16

برای β_1 داریم: $T = 21.64$ <2e-16

در نتیجه H_0 رد می شود.

رگرسیون عبوری از مرکز

$$y = \beta_1 x \quad , \quad \beta_1 = 0.538870$$

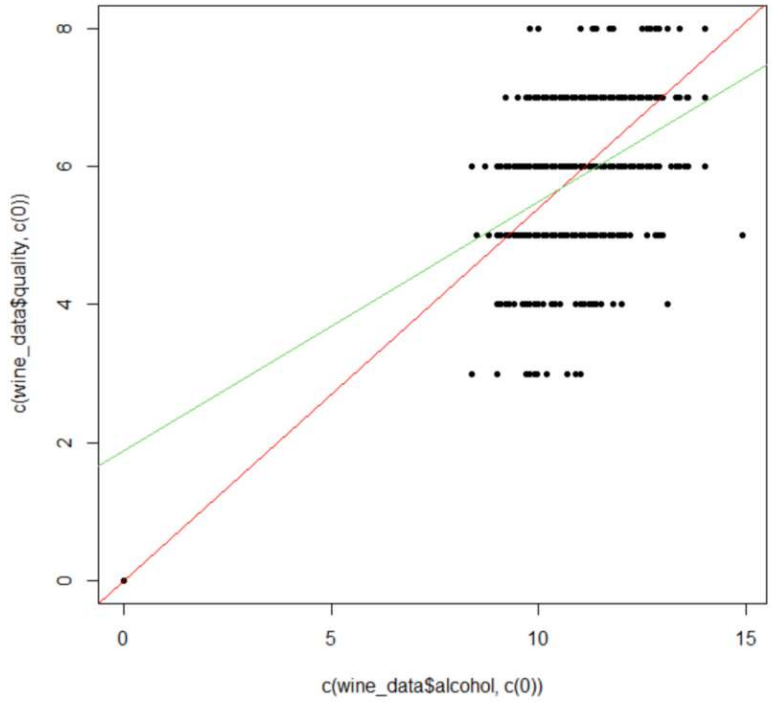
آنالیز واریانس

```
Analysis of Variance Table

Response: quality
           Df Sum Sq Mean Sq F value    Pr(>F)
alcohol     1  50970   50970   94272 < 2.2e-16 ***
Residuals 1598    864       1
```

در جدول آنالیز واریانس *ANOVA* عدد موجود در ستون آخر (از سمت چپ) میزان حمایت داده ها را از درستی فرض H_0 نشان می دهد. اولین ستون (Df) نیز درجه ی آزادی را نشان می دهد که برای مدل عبوری از مرکز یک می باشد.

نمودار پراکنش مدل عبوری از مرکز

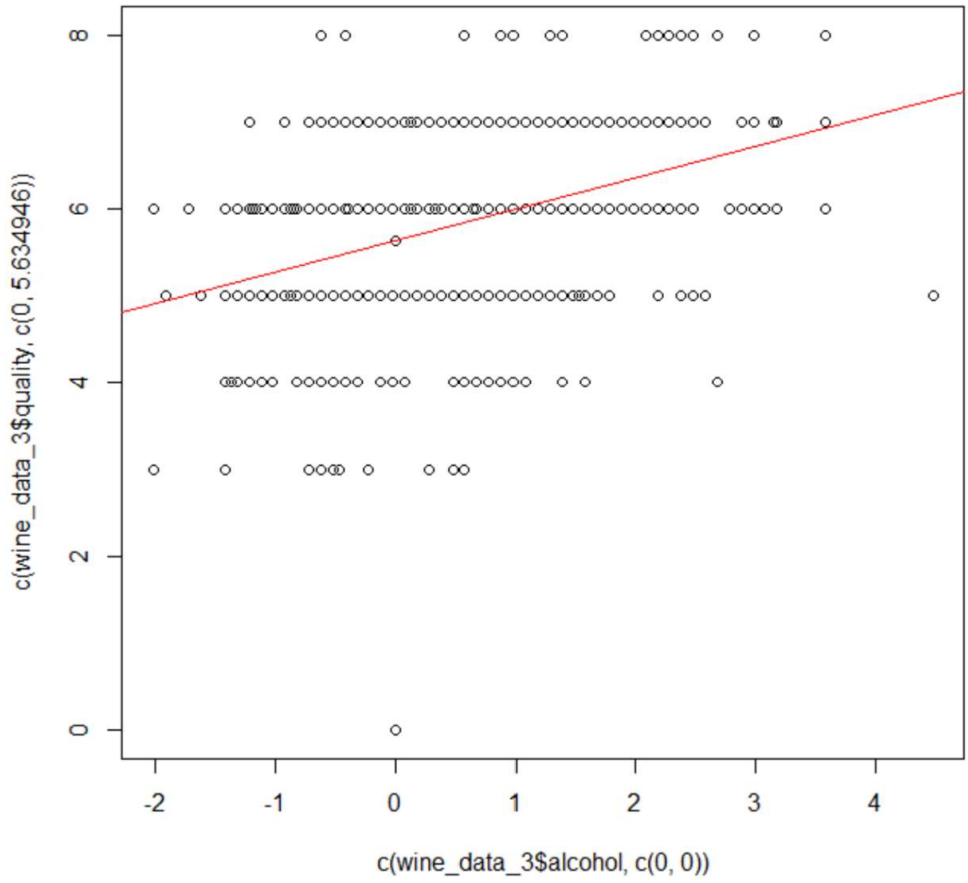


با توجه به معاده ی حاصل از برازش و خطوط نمودار، مدل با عرض از مبدا ارجحیت دارد. فارغ از اینکه حاصل برازش اولیه (مناسبترین مدل خطی فعلی) دارای عرض از مبدا بود، در حالت ثانویه، واریانس داده ها افزایش یافته که مطلوب ما نیست.

رگرسیون مرکزی شده

$$y = \beta_0 + \beta_1 x \quad , \quad \beta_{0\,new} = 5.63495, \ \beta_{1\,new} = 0.36084$$

نمودار پراکنش مدل مرکزی شده



این مدل با اولین مدل برازش داده شده تفاوتی ندارد، جز اینکه در اینجا نمودار در راستای محور طولها به سمت چپ –به دلیل مرکزی شدن داده ها– شیفت داده شده است. اما اگر قرار به انتخاب باشد مدل اولیه را ترجیح میدهیم چرا که اعداد را به طور درست و دقیق داریم. در این مدل باید اعداد را بعدها برای محاسبات با میانگین داده ها جمع کنیم تا عدد واقعی حاصل شود.

تفسیر ضرایب مدل

در مدل رگرسیون مرکزی شده، عرض از مبدا حتما تفسیر دارد و در اینجا این است که اگر میزان درصد اتانول برابر با مقدار میانگین باشد، کیفیت آن چیزی حدود 5.634946 خواهد بود. و به ازای افزایش هر درصد اتانول، کیفیت نوشیدنی 0.36084 واحد افزایش خواهد داشت.

مهمترین متغیر دودویی و برازش مدل رگرسیونی

در این دیتاست متغیر دودویی موجود نیست لذا این قسمت از برنامه قابل برازش نیست.

مدل رگرسیون خطی چندگانه با متغیرهای توضیحی

معادله‌ی مدل برازش داده شده

معادله‌ی مدل برازش داده شده به شرح زیر می باشد:

quality ~ constant + fixed.acidity + volatile.acidity + citric.acid + residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH + sulphates + ethanol

constant = 2.197e+01 ; fixed acidity = 2.499e-02; volatile acidity = -1.084e+00; citric acid = -1.826e-01; residual sugar = 1.633e-02; chloride = -1.874e+00; free sulfur dioxide = 4.361e-03; total sulfur dioxide = -3.265e-03; density = -1.788e+01; ph = -4.137e-01; sulphates = 9.163e-01;ethanol = 2.762e-01

تفسیر ضرایب

constant = 20.97

یعنی به ازای صفر بودن تمامی متغیرها، میزان کیفیت نوشیدنی برابر با ۲۰.۹۷ خواهد بود.

برداشتی که از باقی ضرایب $\beta_i X_i$ می‌توان کرد این است که برای هر کدام، به ازای ثابت ماندن مابقی ضرایب، با هر تغییر واحد X_i در میزان کیفیت نوشیدنی به طور متوسط به میزان β_i تغییر خواهد کرد. برای مثال:

ضریب اسیدیته ثابت برابر با ۰.۰۲۴۹۹ می‌باشد، یعنی اگر تمامی متغیرها ثابت باشند، میزان تغییر کیفیت نوشیدنی به ازای هر واحد تغییر در اسیدیته ثابت، *0.02499* واحد خواهد بود.

آزمون معناداری کلی مدل

فرض آزمون معناداری کلی:

$$H_0 = \beta_1 = \cdots = \beta_i = 0;\ H_1 = o.w.$$

$$F = 81.408 > F(k, n-p) = 1.79$$

در نتیجه فرض H_0 رد می شود.

تفسیر آزمون معنا داری

فرض اولیه رد نشد، پس به همه‌ی متغیرها در حضور سایرین نیاز داریم و هیچ یک قابل حذف نیستند و حضور تمامی آنها معنادار است.

آزمون معناداری ضرایب رگرسیون جزئی

فرض آزمون ضرایب رگرسیون جزئی

$$H_0: \beta_j = 0;\ H_1: \beta_j \neq 0;\ j=1,\ldots,k$$

```
Analysis of Variance Table

Response: quality
                       Df Sum Sq Mean Sq  F value    Pr(>F)
fixed.acidity           1  16.04  16.038  38.1924 8.132e-10 ***
volatile.acidity        1 143.57 143.573 341.9062 < 2.2e-16 ***
citric.acid             1   0.02   0.024   0.0581  0.809535
residual.sugar          1   0.16   0.158   0.3764  0.539600
chlorides               1  13.06  13.062  31.1057 2.868e-08 ***
free.sulfur.dioxide     1   2.97   2.974   7.0828  0.007861 **
total.sulfur.dioxide    1  30.09  30.093  71.6631 < 2.2e-16 ***
density                 1  61.31  61.310 146.0054 < 2.2e-16 ***
pH                      1   7.15   7.154  17.0358 3.859e-05 ***
sulphates               1  55.70  55.697 132.6366 < 2.2e-16 ***
alcohol                 1  45.67  45.672 108.7643 < 2.2e-16 ***
Residuals            1587 666.41   0.420
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

تفسیر آزمون معناداری ضرایب رگرسیون جزئی

با توجه به مقادیر مختلف $p\text{-}value$ درمی‌یابیم که پارامترهای شکر باقیمانده و اسید سیتریک از میزان $\alpha = 0.05$ بیشتر است و آزمون فرض برای آنها رد نمی‌شود. یعنی در صورت وجود باقی متغیرها می‌توان از آنها چشم‌پوشی کرد و همچنان مدل قابل قبولی بدست آورد.

حذف همزمان دو متغیر توضیحی

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.66294 -0.36559 -0.05027  0.46695  1.89579

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.779e+01  1.328e+01   2.847 0.004476 **
citric.acid           5.460e-01  1.186e-01   4.603 4.49e-06 ***
residual.sugar        1.619e-02  1.385e-02   1.169 0.242608
chlorides            -2.913e+00  4.033e-01  -7.222 7.89e-13 ***
free.sulfur.dioxide   7.689e-03  2.182e-03   3.524 0.000437 ***
total.sulfur.dioxide -4.744e-03  7.102e-04  -6.680 3.29e-11 ***
density              -3.384e+01  1.320e+01  -2.563 0.010455 *
pH                   -5.307e-01  1.365e-01  -3.888 0.000105 ***
sulphates             1.121e+00  1.129e-01   9.926  < 2e-16 ***
alcohol               2.650e-01  2.259e-02  11.732  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6638 on 1589 degrees of freedom
Multiple R-squared:  0.3282,    Adjusted R-squared:  0.3244
F-statistic: 86.25 on 9 and 1589 DF,  p-value: < 2.2e-16


Analysis of Variance Table

Response: quality
                      Df Sum Sq Mean Sq  F value Pr(>F)
citric.acid            1  53.41  53.405 121.2050 <2e-16 ***
residual.sugar         1   0.37   0.375   0.8509 0.3564
chlorides              1  33.15  33.150  75.2350 <2e-16 ***
free.sulfur.dioxide    1   1.03   1.032   2.3417 0.1261
total.sulfur.dioxide   1  50.83  50.835 115.3717 <2e-16 ***
density                1  68.35  68.351 155.1260 <2e-16 ***
pH                     1   0.42   0.421   0.9565 0.3282
sulphates              1  73.81  73.810 167.5139 <2e-16 ***
alcohol                1  60.64  60.642 137.6286 <2e-16 ***
Residuals           1589 700.14   0.441
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

در این قسمت دو متغیر *fixed acidity* و *volatile acidity* را حذف کرده‌ایم و

$$SSR = 342.0214,\ MSE=0.4406$$

$$F = 387.77>>>>>F(2,\ 1589)$$

فرض H_0 قطعا رد می‌شود و نمیتوان این متغیرها را حذف کرد و به حضور حداقل یکی از آنها نیاز داریم و حضورشان معنا دار است.

فاصله پیشبینی مشاهده‌ی جدید

با توجه به خروجی نرم افزار:

pred

fit lwr upr

1 1.764963 0.093243 3.436683

با ۹۵٪ اطمینان، به طور متوسط پیشبینی جدید حداقل ۰.۰۹۳۲۴۳ و حداکثر *3.436683* خواهد بود.

بهترین مدل قابل برازش

در این مدل تمامی متغیرها به غیر از *volatile acidity* و *ethanol* را حذف کرده‌ایم که جدول آنالیز واریانس *ANOVA* برای این مدل به شرح زیر می‌باشد:

```
Analysis of Variance Table

Response: quality
                  Df Sum Sq Mean Sq F value    Pr(>F)
volatile.acidity   1 158.97 158.967  356.44 < 2.2e-16 ***
alcohol            1 171.40 171.402  384.32 < 2.2e-16 ***
Residuals       1596 711.80   0.446
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

آزمون معنا داری برای این مدل، با فرض صفر بودن مابقی ضرایب:

$$F = 11.6 > F(r = 9, n-p)$$

فرض در اینجا رد می‌شود پس متغیرهای باقیمانده برای تخمین میزان کیفیت حضوری معنا دار دارند.

اخرین متغیر حذف شده را که نسبت به بقیه متغیر های حذف شده SSR بیشتری داشت به مدل اضاافه میکنیم و ازمون صفر بودن ضرایب متغیر های حذف شده را دوباره انجام میدهیم:

$$F = 7.26> F(r = 8, n-p)$$

یک بار دیگر با اضافه کردن متغیر حذف شده فرض را بررسی میکنیم:

$$F = 5.64> F(r = 7, n-p) = 2.01$$

میبینیم که با الفا برابر ۰.۰۵ این متغیر ها با ارزش هستند.

حال سعی میکنیم از ابتدا یکی یکی ضرایب متغیر هارا صفر قرار دهیم و فرض صفر بودن را بررسی کینیم:

با استفاده از فرض صفر بودن تعدادی از ضرایب یکی یکی متغیر هارا از مدل حذف کرده و ازمون معنا داری را انجام میدهیم:

فرض صفر بودن ضریب citric.acid :

$F = 1.46 < F(r = 1, n-p) = 3.84$

پس متغیر citric.acid را حذف میکنیم.

فرض صفر بودن ضرایب citric.acid و residual.sugar :

$F = 1.32 < F(r = 2, n-p) = 3.00$

بنابراین این دو متغیر را از مدل حذف میکنیم.

فرض صفر بودن ضرایب citric.acid و residual.sugar و free.sulfur.dioxid :

$F = 2.81 > F(r = 3, n-p) = 2.61$

این فرض رد میشود.

در نتیجه فقط میتوانیم دو citric.acid و residual.sugar را از مدل حذف کنیم.

بهترین مدل بعد از حذف این دو متغیر بدست می اید.