hw_09

lisa liubovich

2024-04-04

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(broom)
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.3.2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

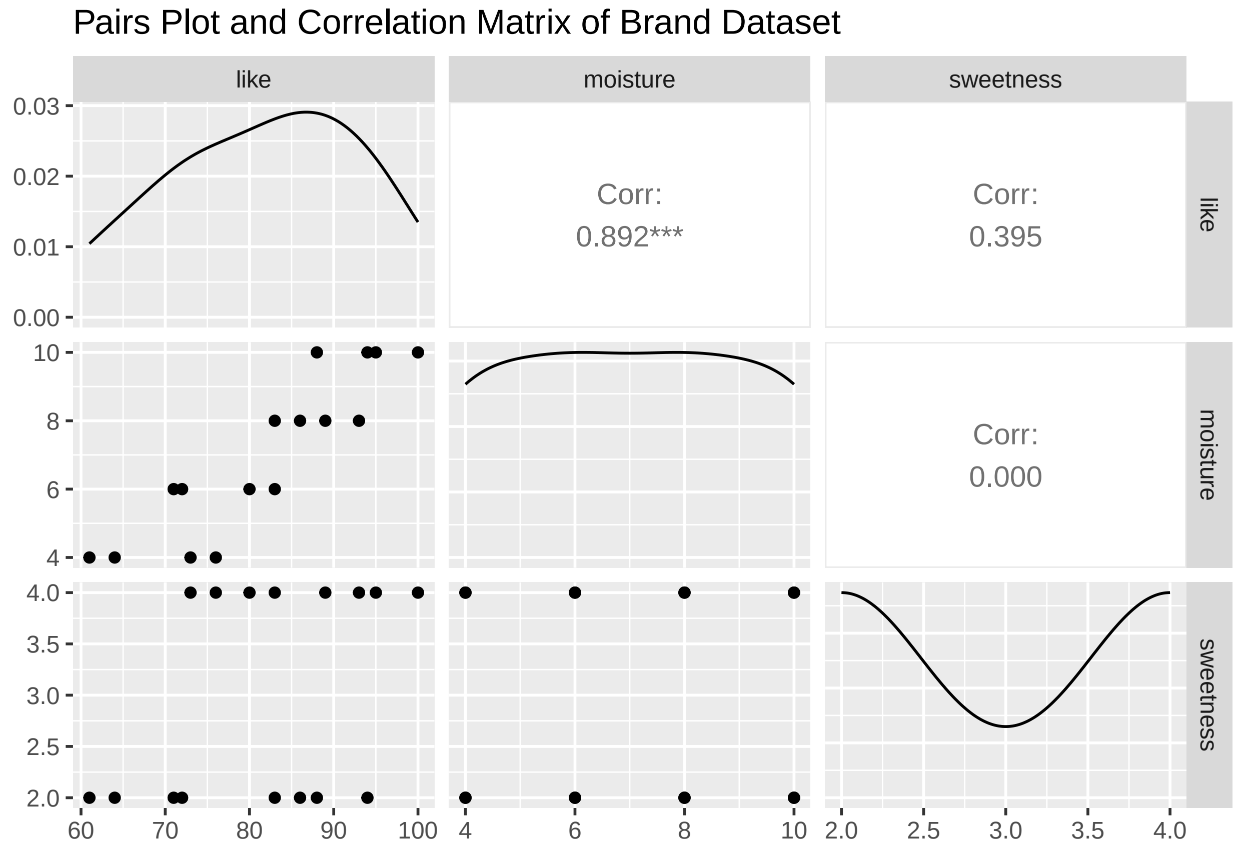
Brand Preference

1.

```
brand <- read_csv("https://dcgerard.github.io/stat_415_615/data/brand.csv")
```

```
## Rows: 16 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (3): like, moisture, sweetness
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
ggpairs(brand, title = "Pairs Plot and Correlation Matrix of Brand Dataset")
```



These diagnostic aids provide information about the relationship between like and moisture like and sweetness, moisture and sweetness, and like and sweetness. Specifcally, the correlation matrix tells us the correlation between like and moisture is 0.892, indicating a strong positive linear association. The correlation between sweetness and like is 0.395, which indicates a weak positive linear association. The correlation between moisture and sweetness is 0, indicating no linear association.

2.

```
lm_brand1 <- lm(like ~ moisture, data = brand)
tidy(lm_brand1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic      p.value
##   <chr>          <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)    50.8      4.39      11.6  0.0000000152
## 2 moisture        4.42     0.598      7.40 0.00000336
```

```
lm_brand2 <- lm(like ~ moisture + sweetness, data = brand)
tidy(lm_brand2)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic       p.value
##   <chr>          <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept)    37.7      3.00      12.6  0.0000000120
## 2 moisture        4.42     0.301     14.7  0.00000000178
## 3 sweetness       4.37     0.673      6.50 0.0000201
```

Model one (just moisture):

$Y_i = \beta_0 + \beta_1\ X_{i1} + \varepsilon_i \rightarrow Y_i = 50.755 + 4.425X_{i1}$

Model two (moisture and sweetness):

$Y_i = \beta_0 + \beta_1\ X_{i1} + \beta_2\ X_{i2} + \varepsilon_i \rightarrow Y_i = 37.650 + 4.425X_{i1} + 4.375X_{i2}$

3.

The estimated regression coefficient for moisture content for both models is 4.425. The coefficient didn't change when adding another predictor, indicating that these predictors are not highly correlated and multicollinearity is not present.

4.

```
Anova(lm_brand1)
```

```
## Anova Table (Type II tests)
##
## Response: like
##            Sum Sq Df F value    Pr(>F)
## moisture  1566.45  1  54.751 3.356e-06 ***
## Residuals  400.55 14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(lm_brand2)
```

```
## Anova Table (Type II tests)
##
## Response: like
##            Sum Sq Df F value    Pr(>F)
## moisture  1566.45  1 215.947 1.778e-09 ***
## sweetness  306.25  1  42.219 2.011e-05 ***
## Residuals   94.30 13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSR(X_1) = 1566.45 (the amount of variability in Y explained by X1)

SSE(X_1) = 306.25 + 94.30 = 400.55 (amount of variability not explained by X1)

SSR(X_2) = 306.25 (the amount of variability in Y explained by X2)

SSE(X_2) = 1566.45 + 94.30 = 1660.75 (variability in Y not explained by X2)

SSR(X_1, X_2) = 1566.45 +306.25 = 1872.7 (variability in Y explained by X1 and X2

SSE(X_1, X_2) = 94.30 (variability in Y that is not explained by X1 or X2)

SSR($X_1|X_2$) = SSE(X_2) - SSE(X_1, X_2) = 1660.75 - 94.30 = 1566.45 (the difference in variability explained by adding X2 to the model)

SSTO = 1566.45 + 306.25 + 94.30 = 1967 (total variation in Y)

Yes, SSR(X_1) equals SSR($X_1|X_2$) in this scenario.

5.

The correlation matrix in question one shows that moisture and sweetness have no correlation, which makes sense given that a) the estimated regression coefficient for moisture didn't change when sweetness was added to the model and b) the extra sums of squares stayed the same. If moisture and sweetness were highly correlated, we would see a change in estimated regression coefficient of moisture and a difference between SSR(X_1) and SSR($X_1|X_2$). But we can see that the relative reduction in the variation in Y by including X1 is the same no matter if X2 is in the model.

6.

```
# R squared Y1:
1566.45 / 1967
```

```
## [1] 0.796365
```

```
# R squared Y2:
306.25 / 1967
```

```
## [1] 0.155694
```

```
# R squared Y12:
1872.7/ 1967
```

```
## [1] 0.952059
```

```
# R squared Y 1 given 2:
1566.45/1660.75
```

```
## [1] 0.9432184
```

```
# R squared Y 2 given 1:
306.25/400.55
```

```
## [1] 0.7645737
```

R^2_{Y1} = SSR(X_1) / SSTO = 1566.45 / 1967 = 0.796

- Proportion of variability in Y explained by X1

R^2_{Y2} = SSR(X_2) / SSTO = 306.25 / 1967 = 0.156

- Proportion of variability in Y explained by X2

R^2_{Y12} = SSR(X_1, X_2) / SSTO = 1872.7 / 1967 = 0.952

- Proportional of variability in Y explained by X1 and X2

$R^2_{Y1|2}$ = SSR($X_1|X_2$) / SSE(X_2) = 1566.45/1660.75 = 0.943

- Proportion of variability not explained by X2 that is explained by X1

$R^2_{Y2|1}$ = (SSE(X_1) - SSE(X_1, X_2))/ SSE(X_1) = (400.55 - 94.30) / 400.55 = 306.25/400.55 = 0.765

- Proportion of variability not explained by X1 that is explained by X2

R^2 = SSR / SSTO = 0.952

- Proportion of total variation in Y associated with the use of the set of variables X1 and X2

Interpretation: 79.6% of the variation in like (Y) can be explained by moisture (X1). Including sweetness (X2) results in an additional 76.5% reduction. 15.6% of the variation in like (Y) can be explained by sweetness (X2). Including moisture (X1) results in an additional 94.3% reduction. 95.2% of the variation in like (Y) is explained by both moisture (X1) and (X2).

Conceptual Exercise

1.

I wouldn't necessarily say that the difficulty of determining effects is "unimportant" because of the high coefficient of multiple determination. Yes, the high R^2 indicates that a large proportion of the variation in the dependent variable is explained by the independent variables collectively but I would still exercise caution in interpreting the coefficients individually.