Our first regression modeling practice

R workshop 2023

2023-11-02

Warning: package 'ggplot2' was built under R version 4.3.1

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.2 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.0  
✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
✔ purrr 1.0.1   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors  
  
Attaching package: 'janitor'  
  
  
The following objects are masked from 'package:stats':  
  
 chisq.test, fisher.test

Warning: package 'Ecdat' was built under R version 4.3.1

Loading required package: Ecfun

Warning: package 'Ecfun' was built under R version 4.3.1

Attaching package: 'Ecfun'  
  
The following object is masked from 'package:base':  
  
 sign  
  
  
Attaching package: 'Ecdat'  
  
The following object is masked from 'package:datasets':  
  
 Orange

[1] 420 17

Rows: 420  
Columns: 17  
$ distcod <int> 75119, 61499, 61549, 61457, 61523, 62042, 68536, 63834, 62331…  
$ county <fct> Alameda, Butte, Butte, Butte, Butte, Fresno, San Joaquin, Ker…  
$ district <fct> Sunol Glen Unified, Manzanita Elementary, Thermalito Union El…  
$ grspan <fct> KK-08, KK-08, KK-08, KK-08, KK-08, KK-08, KK-08, KK-08, KK-08…  
$ enrltot <int> 195, 240, 1550, 243, 1335, 137, 195, 888, 379, 2247, 446, 987…  
$ teachers <dbl> 10.90, 11.15, 82.90, 14.00, 71.50, 6.40, 10.00, 42.50, 19.00,…  
$ calwpct <dbl> 0.5102, 15.4167, 55.0323, 36.4754, 33.1086, 12.3188, 12.9032,…  
$ mealpct <dbl> 2.0408, 47.9167, 76.3226, 77.0492, 78.4270, 86.9565, 94.6237,…  
$ computer <int> 67, 101, 169, 85, 171, 25, 28, 66, 35, 0, 86, 56, 25, 0, 31, …  
$ testscr <dbl> 690.80, 661.20, 643.60, 647.70, 640.85, 605.55, 606.75, 609.0…  
$ compstu <dbl> 0.34358975, 0.42083332, 0.10903226, 0.34979424, 0.12808989, 0…  
$ expnstu <dbl> 6384.911, 5099.381, 5501.955, 7101.831, 5235.988, 5580.147, 5…  
$ str <dbl> 17.88991, 21.52466, 18.69723, 17.35714, 18.67133, 21.40625, 1…  
$ avginc <dbl> 22.690001, 9.824000, 8.978000, 8.978000, 9.080333, 10.415000,…  
$ elpct <dbl> 0.000000, 4.583333, 30.000002, 0.000000, 13.857677, 12.408759…  
$ readscr <dbl> 691.6, 660.5, 636.3, 651.9, 641.8, 605.7, 604.5, 605.5, 608.9…  
$ mathscr <dbl> 690.0, 661.9, 650.9, 643.5, 639.9, 605.4, 609.0, 612.5, 616.1…

save(cashool,file=“data/caschool.RData”)

distcod county district grspan enrltot teachers  
401 69526 Santa Clara Los Gatos Union Elementary KK-08 2768 141.4  
402 69252 Santa Barbara Montecito Union Elementary KK-06 535 31.0  
 calwpct mealpct computer testscr compstu expnstu str avginc  
401 0.4481 2.7921 540 689.10 0.1950867 5566.468 19.57567 34.1595  
402 0.0800 0.7200 84 691.05 0.1570093 7039.949 17.25806 43.2300  
 elpct readscr mathscr  
401 1.589595 691.6 686.6  
402 1.495327 693.5 688.6

distcod county district grspan enrltot teachers  
1 75119 Alameda Sunol Glen Unified KK-08 195 10.90  
2 61499 Butte Manzanita Elementary KK-08 240 11.15  
3 61549 Butte Thermalito Union Elementary KK-08 1550 82.90  
4 61457 Butte Golden Feather Union Elementary KK-08 243 14.00  
5 61523 Butte Palermo Union Elementary KK-08 1335 71.50  
6 62042 Fresno Burrel Union Elementary KK-08 137 6.40  
 calwpct mealpct computer testscr compstu expnstu str avginc  
1 0.5102 2.0408 67 690.80 0.3435898 6384.911 17.88991 22.690001  
2 15.4167 47.9167 101 661.20 0.4208333 5099.381 21.52466 9.824000  
3 55.0323 76.3226 169 643.60 0.1090323 5501.955 18.69723 8.978000  
4 36.4754 77.0492 85 647.70 0.3497942 7101.831 17.35714 8.978000  
5 33.1086 78.4270 171 640.85 0.1280899 5235.988 18.67133 9.080333  
6 12.3188 86.9565 25 605.55 0.1824818 5580.147 21.40625 10.415000  
 elpct readscr mathscr  
1 0.000000 691.6 690.0  
2 4.583333 660.5 661.9  
3 30.000002 636.3 650.9  
4 0.000000 651.9 643.5  
5 13.857677 641.8 639.9  
6 12.408759 605.7 605.4

[1] "distcod" "county" "district" "grspan" "enrltot" "teachers"  
 [7] "calwpct" "mealpct" "computer" "testscr" "compstu" "expnstu"   
[13] "str" "avginc" "elpct" "readscr" "mathscr"

Warning: There was 1 warning in `summarise()`.  
ℹ In argument: `across(where(is.numeric), median, na.rm = TRUE)`.  
Caused by warning:  
! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.  
Supply arguments directly to `.fns` through an anonymous function instead.  
  
 # Previously  
 across(a:b, mean, na.rm = TRUE)  
  
 # Now  
 across(a:b, \(x) mean(x, na.rm = TRUE))

testscr str avginc elpct mealpct calwpct  
1 654.45 19.72321 13.7278 8.777634 41.7507 10.52045

testscr\_Mean testscr\_Variance testscr\_SD testscr\_Min testscr\_Max testscr\_N  
1 654.1565 363.0301 19.05335 605.55 706.75 420  
 str\_Mean str\_Variance str\_SD str\_Min str\_Max str\_N avginc\_Mean  
1 19.64043 3.578952 1.891812 14 25.8 420 15.31659  
 avginc\_Variance avginc\_SD avginc\_Min avginc\_Max avginc\_N elpct\_Mean  
1 52.21348 7.22589 5.335 55.328 420 15.76816  
 elpct\_Variance elpct\_SD elpct\_Min elpct\_Max elpct\_N mealpct\_Mean  
1 334.3751 18.28593 0 85.53972 420 44.70524  
 mealpct\_Variance mealpct\_SD mealpct\_Min mealpct\_Max mealpct\_N calwpct\_Mean  
1 735.6778 27.12338 0 100 420 13.24604  
 calwpct\_Variance calwpct\_SD calwpct\_Min calwpct\_Max calwpct\_N  
1 131.2129 11.45482 0 78.9942 420

# A tibble: 6 × 6  
 variable n mean variance min max  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 testscr 420 654. 363. 606. 707.   
2 str 420 19.6 3.58 14 25.8  
3 avginc 420 15.3 52.2 5.34 55.3  
4 elpct 420 15.8 334. 0 85.5  
5 mealpct 420 44.7 736. 0 100   
6 calwpct 420 13.2 131. 0 79.0

# A tibble: 6 × 6  
 Variable Observations Mean Variance Min Max  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 testscr 420 654. 363. 606. 707.   
2 str 420 19.6 3.58 14 25.8  
3 avginc 420 15.3 52.2 5.34 55.3  
4 elpct 420 15.8 334. 0 85.5  
5 mealpct 420 44.7 736. 0 100   
6 calwpct 420 13.2 131. 0 79.0

# A tibble: 6 × 6  
 Variable Observations Mean Variance Min Max  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 testscr 420 654. 363. 606. 707.   
2 str 420 19.6 3.58 14 25.8  
3 avginc 420 15.3 52.2 5.34 55.3  
4 elpct 420 15.8 334. 0 85.5  
5 mealpct 420 44.7 736. 0 100   
6 calwpct 420 13.2 131. 0 79.0

Table 1: Descriptive Analysis of caschool

| Variable | Observations | Mean | Variance | Min | Max |
| --- | --- | --- | --- | --- | --- |
| testscr | 420.00 | 654.16 | 363.03 | 605.55 | 706.75 |
| str | 420.00 | 19.64 | 3.58 | 14.00 | 25.80 |
| avginc | 420.00 | 15.32 | 52.21 | 5.34 | 55.33 |
| elpct | 420.00 | 15.77 | 334.38 | 0.00 | 85.54 |
| mealpct | 420.00 | 44.71 | 735.68 | 0.00 | 100.00 |
| calwpct | 420.00 | 13.25 | 131.21 | 0.00 | 78.99 |

Data summary

|  |  |
| --- | --- |
| Name | caschool\_sm |
| Number of rows | 420 |
| Number of columns | 6 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 6 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| testscr | 0 | 1 | 654.16 | 19.05 | 605.55 | 640.05 | 654.45 | 666.66 | 706.75 | ▂▆▇▅▁ |
| str | 0 | 1 | 19.64 | 1.89 | 14.00 | 18.58 | 19.72 | 20.87 | 25.80 | ▁▃▇▃▁ |
| avginc | 0 | 1 | 15.32 | 7.23 | 5.34 | 10.64 | 13.73 | 17.63 | 55.33 | ▇▃▁▁▁ |
| elpct | 0 | 1 | 15.77 | 18.29 | 0.00 | 1.94 | 8.78 | 22.97 | 85.54 | ▇▂▁▁▁ |
| mealpct | 0 | 1 | 44.71 | 27.12 | 0.00 | 23.28 | 41.75 | 66.86 | 100.00 | ▆▇▆▅▅ |
| calwpct | 0 | 1 | 13.25 | 11.45 | 0.00 | 4.40 | 10.52 | 18.98 | 78.99 | ▇▃▁▁▁ |

Warning: package 'modelsummary' was built under R version 4.3.1

Attaching package: 'rstatix'

The following object is masked from 'package:janitor':  
  
 make\_clean\_names

The following object is masked from 'package:stats':  
  
 filter

# A tibble: 1 × 8  
 .y. group1 group2 n1 n2 statistic df p  
\* <chr> <chr> <chr> <int> <int> <dbl> <dbl> <dbl>  
1 testscr 0 1 182 238 -4.00 418 0.0000752

# A tibble: 4 × 8  
 .y. group1 group2 n1 n2 statistic df p  
 <chr> <chr> <chr> <int> <int> <dbl> <dbl> <dbl>  
1 testscr 0 1 182 238 -4.00 418 0.0000752  
2 testscr 0 1 44 64 -1.05 106 0.296   
3 testscr 0 1 50 54 -1.73 102 0.0872   
4 testscr 0 1 61 44 -0.705 103 0.483

| group1 | group2 | n1 | n2 | statistic | df | p |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 182 | 238 | -4.00 | 418 | 0.000 |
| 0 | 1 | 44 | 64 | -1.05 | 106 | 0.296 |
| 0 | 1 | 50 | 54 | -1.73 | 102 | 0.087 |
| 0 | 1 | 61 | 44 | -0.70 | 103 | 0.483 |

Call:  
lm(formula = testscr ~ str, data = caschool\_sm)  
  
Coefficients:  
(Intercept) str   
 698.93 -2.28

Call:  
lm(formula = testscr ~ str, data = caschool\_sm)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-47.727 -14.251 0.483 12.822 48.540   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 698.9330 9.4675 73.825 < 2e-16 \*\*\*  
str -2.2798 0.4798 -4.751 2.78e-06 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 18.58 on 418 degrees of freedom  
Multiple R-squared: 0.05124, Adjusted R-squared: 0.04897   
F-statistic: 22.58 on 1 and 418 DF, p-value: 2.783e-06

# A tibble: 5 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 676. 8.00 84.5 1.34e-263  
2 str -1.19 0.398 -2.98 3.03e- 3  
3 elpctlow 11.2 2.10 5.33 1.59e- 7  
4 elpctmedium 11.4 2.05 5.55 5.00e- 8  
5 elpctvery high -16.3 2.07 -7.89 2.69e- 14

|  | (1) | (2) |
| --- | --- | --- |
| (Intercept) | 698.93 (9.47)\*\*\* | 675.85 (8.00)\*\*\* |
| str | -2.28 (0.48)\*\*\* | -1.19 (0.40)\*\* |
| elpctlow |  | 11.22 (2.10)\*\*\* |
| elpctmedium |  | 11.38 (2.05)\*\*\* |
| elpctvery high |  | -16.32 (2.07)\*\*\* |
| Num.Obs. | 420 | 420 |
| R2 | 0.051 | 0.393 |
| R2 Adj. | 0.049 | 0.388 |
| AIC | 3650.5 | 3468.6 |
| BIC | 3662.6 | 3492.9 |
| Log.Lik. | -1822.250 | -1728.305 |
| F | 22.575 | 67.297 |
| RMSE | 18.54 | 14.82 |

# estimate the multiple regression model

Loading required package: carData

Attaching package: 'carData'

The following object is masked from 'package:Ecdat':  
  
 Mroz

Attaching package: 'car'

The following object is masked from 'package:dplyr':  
  
 recode

The following object is masked from 'package:purrr':  
  
 some

Call:  
lm(formula = testscr ~ mealpct, data = caschool\_sm)  
  
Coefficients:  
(Intercept) mealpct   
 681.4395 -0.6103

Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':  
  
 as.Date, as.Date.numeric

Linear hypothesis test  
  
Hypothesis:  
str = 0  
mealpct = 0  
  
Model 1: restricted model  
Model 2: testscr ~ str + elpct + mealpct  
  
 Res.Df RSS Df Sum of Sq F Pr(>F)   
1 416 94241   
2 414 34338 2 59903 361.12 < 2.2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Correlations

Warning: package 'corrr' was built under R version 4.3.1

Attaching package: 'corrr'

The following object is masked from 'package:skimr':  
  
 focus

Non-numeric variables removed from input: `elpct`  
Correlation computed with  
• Method: 'pearson'  
• Missing treated using: 'pairwise.complete.obs'

# A tibble: 4 × 5  
 term testscr str mealpct calwpct  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 testscr NA -0.226 -0.869 -0.627   
2 str -0.226 NA 0.135 0.0183  
3 mealpct -0.869 0.135 NA 0.739   
4 calwpct -0.627 0.0183 0.739 NA

## gt, gtsummary, gtExtras

Non-numeric variables removed from input: `elpct`  
Correlation computed with  
• Method: 'pearson'  
• Missing treated using: 'pairwise.complete.obs'

| term | testscr | str | mealpct | calwpct |
| --- | --- | --- | --- | --- |
| testscr | NA | -0.23 | -0.87 | -0.63 |
| str | -0.23 | NA | 0.14 | 0.02 |
| mealpct | -0.87 | 0.14 | NA | 0.74 |
| calwpct | -0.63 | 0.02 | 0.74 | NA |

Non-numeric variables removed from input: `elpct`  
Correlation computed with  
• Method: 'pearson'  
• Missing treated using: 'pairwise.complete.obs'

Table 1: Correlation between variables

| term | testscr | str | mealpct | calwpct |
| --- | --- | --- | --- | --- |
| testscr | NA | -0.23 | -0.87 | -0.63 |
| str | -0.23 | NA | 0.14 | 0.02 |
| mealpct | -0.87 | 0.14 | NA | 0.74 |
| calwpct | -0.63 | 0.02 | 0.74 | NA |

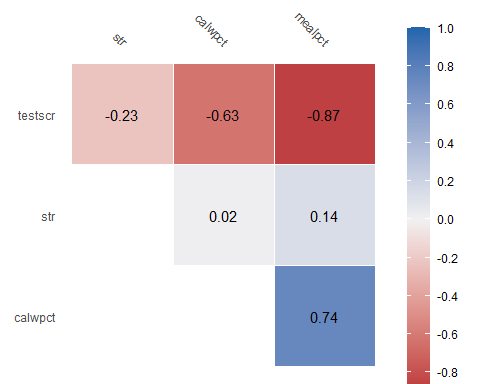
## theme of a table

Non-numeric variables removed from input: `elpct`  
Correlation computed with  
• Method: 'pearson'  
• Missing treated using: 'pairwise.complete.obs'

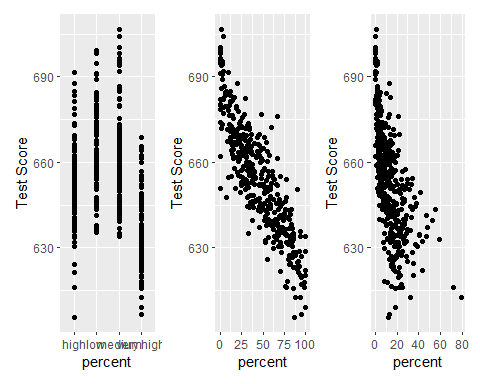
Table 1: Correlation between variables

| term | testscr | str | mealpct | calwpct |
| --- | --- | --- | --- | --- |
| testscr | NA | -0.23 | -0.87 | -0.63 |
| str | -0.23 | NA | 0.14 | 0.02 |
| mealpct | -0.87 | 0.14 | NA | 0.74 |
| calwpct | -0.63 | 0.02 | 0.74 | NA |

Non-numeric variables removed from input: `elpct`  
Correlation computed with  
• Method: 'pearson'  
• Missing treated using: 'pairwise.complete.obs'

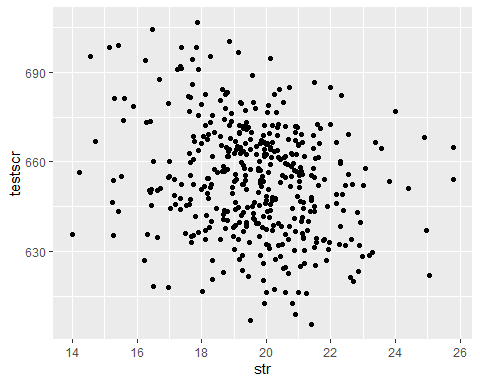


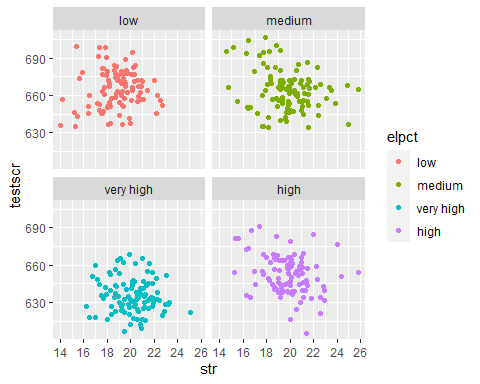
## Plots (Arrange plots easily)



## Regression Models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | (1) | (2) | (3) | (4) | (5) |
| (Intercept) | 698.933\*\*\* | 675.853\*\*\* | 697.720\*\*\* | 688.687\*\*\* | 697.751\*\*\* |
|  | (9.467) | (7.999) | (4.955) | (6.553) | (4.961) |
| str | -2.280\*\*\* | -1.186\*\* | -1.024\*\*\* | -1.325\*\*\* | -1.031\*\*\* |
|  | (0.480) | (0.398) | (0.243) | (0.323) | (0.244) |
| elpctlow |  | 11.223\*\*\* | 3.241\* | 10.546\*\*\* | 3.346\* |
|  |  | (2.104) | (1.320) | (1.709) | (1.349) |
| elpctmedium |  | 11.377\*\*\* | 3.791\*\* | 8.607\*\*\* | 3.830\*\* |
|  |  | (2.048) | (1.284) | (1.674) | (1.289) |
| elpctvery high |  | -16.320\*\*\* | -0.784 | -10.327\*\*\* | -0.853 |
|  |  | (2.068) | (1.393) | (1.728) | (1.406) |
| mealpct |  |  | -0.560\*\*\* |  | -0.551\*\*\* |
|  |  |  | (0.021) |  | (0.031) |
| calwpct |  |  |  | -0.809\*\*\* | -0.024 |
|  |  |  |  | (0.055) | (0.061) |
| Num.Obs. | 420 | 420 | 420 | 420 | 420 |
| R2 | 0.051 | 0.393 | 0.774 | 0.601 | 0.774 |
| R2 Adj. | 0.049 | 0.388 | 0.772 | 0.596 | 0.771 |
| AIC | 3650.5 | 3468.6 | 3055.5 | 3294.6 | 3057.3 |
| BIC | 3662.6 | 3492.9 | 3083.8 | 3322.8 | 3089.6 |
| Log.Lik. | -1822.250 | -1728.305 | -1520.740 | -1640.283 | -1520.663 |
| F | 22.575 | 67.297 | 283.989 | 124.784 | 236.198 |
| RMSE | 18.54 | 14.82 | 9.04 | 12.02 | 9.04 |





── Attaching packages ────────────────────────────────────── tidymodels 1.1.0 ──

✔ broom 1.0.5 ✔ rsample 1.2.0  
✔ dials 1.2.0 ✔ tune 1.1.1  
✔ infer 1.0.4 ✔ workflows 1.1.3  
✔ modeldata 1.1.0 ✔ workflowsets 1.0.1  
✔ parsnip 1.1.0 ✔ yardstick 1.2.0  
✔ recipes 1.0.8

Warning: package 'broom' was built under R version 4.3.1

Warning: package 'recipes' was built under R version 4.3.1

Warning: package 'rsample' was built under R version 4.3.1

── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
✖ infer::chisq\_test() masks rstatix::chisq\_test()  
✖ scales::discard() masks purrr::discard()  
✖ rstatix::filter() masks dplyr::filter(), stats::filter()  
✖ recipes::fixed() masks stringr::fixed()  
✖ dials::get\_n() masks rstatix::get\_n()  
✖ dplyr::lag() masks stats::lag()  
✖ infer::prop\_test() masks rstatix::prop\_test()  
✖ car::recode() masks dplyr::recode()  
✖ car::some() masks purrr::some()  
✖ yardstick::spec() masks readr::spec()  
✖ recipes::step() masks stats::step()  
✖ infer::t\_test() masks rstatix::t\_test()  
• Dig deeper into tidy modeling with R at https://www.tmwr.org

here() starts at D:/RepTemplates/econometric23

Table 1: **Linear Regression Analysis**

**Model: displ~year+hwy+cyl**

| variable | Estimate | Standard Error | T-value | P Value |
| --- | --- | --- | --- | --- |
| (Intercept) | -27.69 | 13.55 | -2.04 | 0.04 |
| year | 0.01 | 0.01 | 2.06 | 0.04 |
| hwy | -0.03 | 0.01 | -4.06 | 0.00 |
| cyl | 0.65 | 0.03 | 22.18 | 0.00 |
| R-Squared: 0.883 | | | | |
| Adj-r-squared: 0.881 | | | | |
| Prob(F-statistic: 0 | | | | |
| Source: mpg | | | | |