F19 STA 100 A01 Discussion 10

Yishan Huang

2019/12/03

Discussion Time: Tuesday 8:00 – 8:50 am, Haring Hall 1204.

Notes: https://github.com/Hahahuo-13316/sta100-a01-fall19

Office hour: Tuesday 12:00 – 1:00 pm, Mathematical Sciences Building 1117.

Email: yishuang@ucdavis.edu

Linear regression

- Sample: (x_i, y_i), $i = 1, \cdots, n$, n is sample size.
- Correlation coefficient:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}}.$$

  Where,

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

- Least square fitting: We use the line $y = b_0 + b_1 x$ to fit the observations, and find b_0 and b_1 to minimize $\sum(y_i - \hat{y}_i)^2$, where $\hat{y}_i = b_0 + b_1 x_i$, are the fitted value of y_i. The result is

$$b_1 = r \cdot \frac{s_y}{s_x} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \quad b_0 = \bar{y} - b_1\bar{x}.$$

In the following scatter plot, the solid line is the least square fitted line.

Use software to make inference on linear model

- Model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$.

```
mtcars
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE        16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL        17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC       15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
```

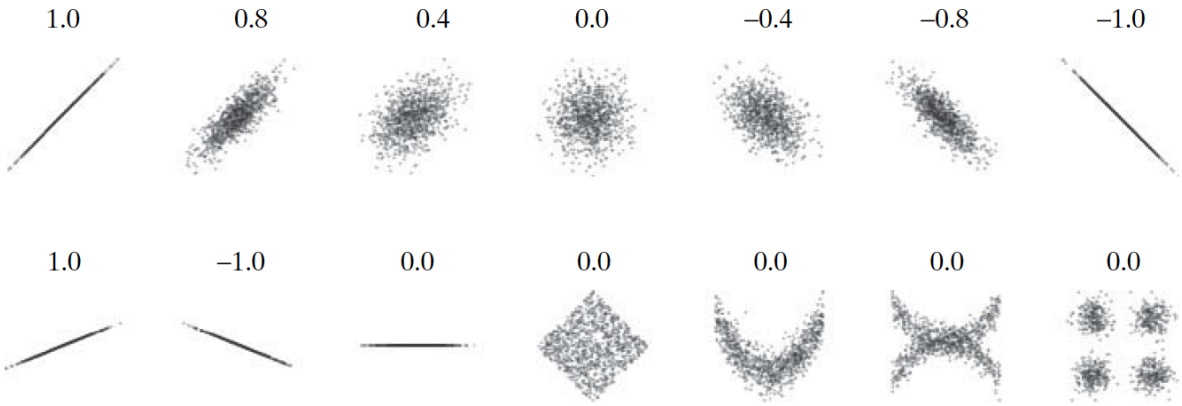

Figure 1: Correlation coefficient

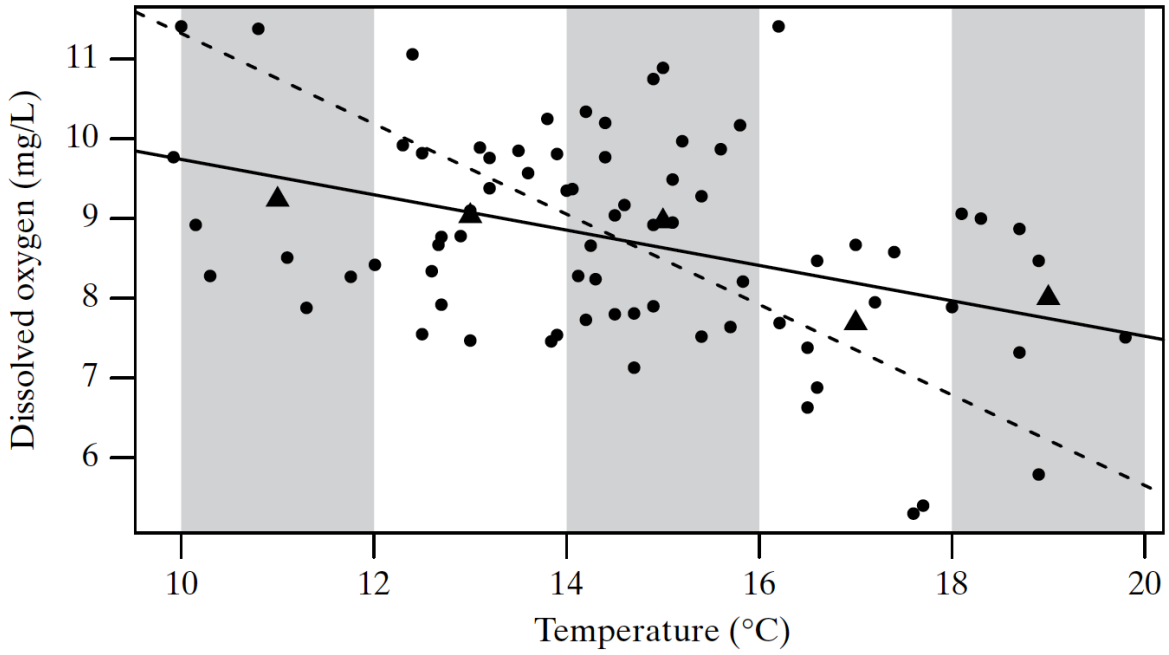

Figure 2: Scatter plot

```
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

```
fit <- lm(mpg ~ cyl + disp + hp + drat + wt + qsec, data = mtcars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + hp + drat + wt + qsec, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9682 -1.5795 -0.4353  1.1662  5.5272
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.30736   14.62994   1.798  0.08424 .
## cyl         -0.81856    0.81156  -1.009  0.32282
## disp         0.01320    0.01204   1.097  0.28307
## hp          -0.01793    0.01551  -1.156  0.25846
## drat         1.32041    1.47948   0.892  0.38065
## wt          -4.19083    1.25791  -3.332  0.00269 **
## qsec         0.40146    0.51658   0.777  0.44436
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.557 on 25 degrees of freedom
## Multiple R-squared:  0.8548, Adjusted R-squared:   0.82
## F-statistic: 24.53 on 6 and 25 DF,  p-value: 2.45e-09
```

Interpret linear regression summary in MINITAB

You may see the following summary in a MINITAB linear regression, taken from the website https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/regression/how-to/fit-regression-model/before-you-start/example/. It is quite similar to the summary in R.

```
Regression Equation
Rating = -0.756 + 0.1545 Conc + 0.2171 Ratio + 0.01081 Temp + 0.0946 Time
```

```
Term         Coef  SE Coef  T-Value  P-Value
Constant   -0.756    0.736    -1.03    0.314
Conc       0.1545   0.0633     2.44    0.022
Ratio      0.2171   0.0316     6.86    0.000
Temp      0.01081  0.00462     2.34    0.027
Time       0.0946   0.0546     1.73    0.094

Model Summary
       S    R-sq  R-sq(adj)
0.811840  72.92%     68.90%

Analysis of Variance
Source         DF       SS       MS  F-Value  P-Value
Regression      4  47.9096  11.9774    18.17    0.000
Error          27  17.7953   0.6591
Total          31  65.7049
```

- The coefficients in the regression equation is the estimated coefficients of our linear model, $\hat{\beta}_i$. We can use the regression equation to make prediction, by just plug in the new observed x values and gain the estimated y value on the left hand side.
- To test whether $\beta_i = 0$ for a specified $i \in \{0, 1, \cdots, k\}$, we can compare the p-value given by the corresponding row in the middle table, and the required significant level. We reject the null hypothesis $H_0 : \beta_i = 0$ if the corresponding p-value is less than the required significant level α. Otherwise, we do not reject H_0.
- **R^2 is the proportion of the variance in the dependent variable (Y) that is predictable from the independent variable(s) (X)**, and adjusted **R^2 is not**.

How to insert new predictors into a dataframe

```r
mtcars.new <- mtcars
mtcars.new$cyl.greater.4 <- as.integer(mtcars$cyl > 4)
mtcars.new[c("cyl", "cyl.greater.4")]
```

```
##                     cyl cyl.greater.4
## Mazda RX4             6             1
## Mazda RX4 Wag         6             1
## Datsun 710            4             0
## Hornet 4 Drive        6             1
## Hornet Sportabout     8             1
## Valiant               6             1
## Duster 360            8             1
## Merc 240D             4             0
## Merc 230              4             0
## Merc 280              6             1
## Merc 280C             6             1
## Merc 450SE            8             1
## Merc 450SL            8             1
## Merc 450SLC           8             1
## Cadillac Fleetwood    8             1
## Lincoln Continental   8             1
## Chrysler Imperial     8             1
```

```
## Fiat 128             4            0
## Honda Civic          4            0
## Toyota Corolla       4            0
## Toyota Corona        4            0
## Dodge Challenger     8            1
## AMC Javelin          8            1
## Camaro Z28           8            1
## Pontiac Firebird     8            1
## Fiat X1-9            4            0
## Porsche 914-2        4            0
## Lotus Europa         4            0
## Ford Pantera L       8            1
## Ferrari Dino         6            1
## Maserati Bora        8            1
## Volvo 142E           4            0
```

Last things: the materials not covered by the discussion notes

- stem-and-leaf plots:

```
a <- sort(round(runif(100) * 100))
a
```

```
##   [1]   1   1   2   3   5   5   5   5   6   6   7  10  11  13  13  15  17
##  [18]  18  18  20  20  21  22  23  24  24  27  28  31  33  34  35  35  36
##  [35]  37  39  40  44  44  45  48  49  49  50  51  51  52  53  55  56  56
##  [52]  58  60  60  60  60  61  63  65  66  66  66  66  66  69  69  69  70
##  [69]  71  73  74  75  76  76  76  76  76  77  80  81  82  82  83  83  84
##  [86]  86  86  86  89  89  90  92  94  96  97  97  97  98  98 100
```

```
stem(a)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    0 | 11235555667
##    1 | 01335788
##    2 | 001234478
##    3 | 13455679
##    4 | 0445899
##    5 | 011235668
##    6 | 000013566666999
##    7 | 01345666667
##    8 | 012233466699
##    9 | 024677788
##   10 | 0
```

- More hypothesis testing:
  - Two-sample t-test (and with large sample size, z-test) of the difference of population mean (Section 7.1 – 7.5);
  - Two-sample z-test for difference of proportion;