Multiple Linear Regression

Comparing Regression Lines for Two Groups

We want to relate course evaluation scores (Y) to the beauty score assigned to the instructor (X_1) and the gender of (female v. male).

Questions we could ask:

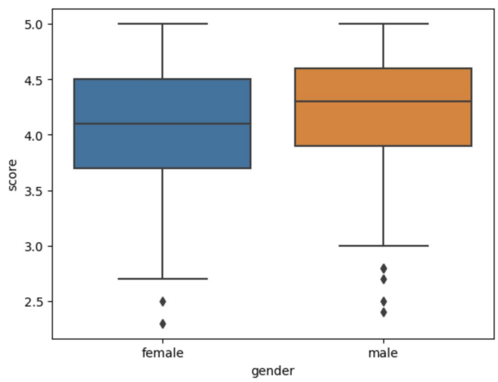
1. Is there a difference in the **mean** *course evaluation score* between *female and male instructors*?
2. Is there a linear relationship between course evaluation scores and beauty scores?
3. Is there a difference in the mean course evaluation score between female and male instructors, after accounting for the beauty score?
4. Is the relationship between beauty score and course evaluation score the same for male and female instructors?

Comparing Regression Lines for Two Groups

We want to relate course evaluation scores (Y) to the beauty score assigned to the instructor (X_1) and the gender of (female v. male).

1. Is there a difference in the **mean** *course evaluation score* between *female and male instructors*?

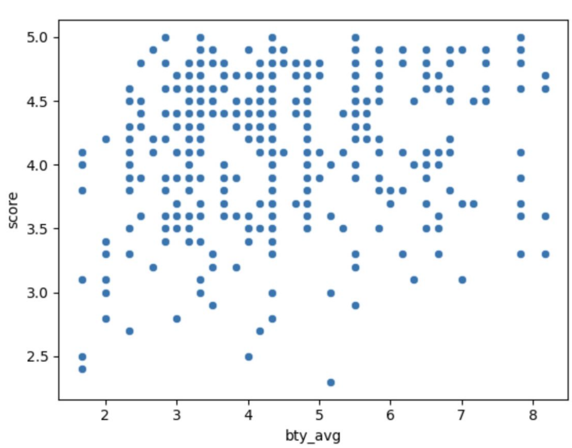
↳ t-test



Comparing Regression Lines for Two Groups

We want to relate course evaluation scores (Y) to the beauty score assigned to the instructor (X_1) and the gender of (female v. male).

2. Is there a linear relationship between course evaluation scores and beauty scores?



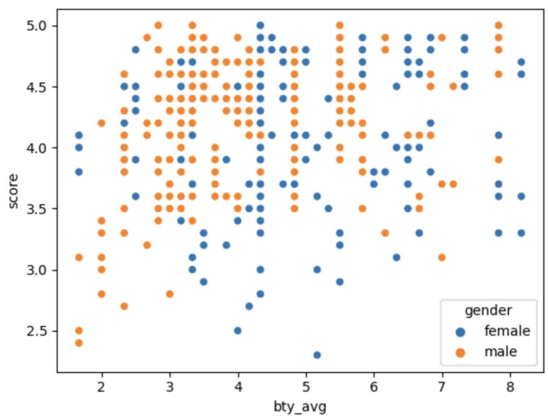
inference on B_1

$Y = B_0 + B_1 X$

Comparing Regression Lines for Two Groups

We want to relate course evaluation scores (Y) to the beauty score assigned to the instructor (X_1) and the gender of (female v. male).

3. Is there a difference in the mean course evaluation score between female and male instructors, after accounting for the beauty score?



2 lines - same slope
diff. intercepts

The Regression Model: Common (Parallel) Slope

Let Gender = $\begin{cases} 1, & \text{Male} \\ 0, & \text{Female} \end{cases}$

$\varepsilon \sim N(0, \sigma_\varepsilon)$

Then,

$$\text{Score} = B_0 + B_1(\text{gender}) + \beta_2(\text{beauty}) + \varepsilon$$

We have two lines:

- Female $\quad \text{score} = B_0 + B_2(\text{beauty}) + \varepsilon$
- Male $\quad \text{score} = (B_0 + B_1) + B_2(\text{beauty}) + \varepsilon$

Interpretation:

- β_0: intercept for female instructors
- β_1: shift in intercept for males
- β_2: common relationship b/w beauty + score

Interpretation

$\hat{y} = 3.7473 + 0.0742 * \text{beauty}$ (female)

$\hat{y} = (3.7473 + 0.1724) + 0.0742 * \text{beauty}$ (male)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  score   R-squared:                       0.059
Model:                            OLS   Adj. R-squared:                  0.055
Method:                 Least Squares   F-statistic:                     14.45
Date:                Tue, 25 Jul 2023   Prob (F-statistic):           8.18e-07
Time:                        16:28:11   Log-Likelihood:                -360.37
No. Observations:                 463   AIC:                             726.7
Df Residuals:                     460   BIC:                             739.1
Df Model:                           2
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          3.7473      0.085     44.266      0.000       3.581       3.914
gender[T.male]     0.1724      0.050      3.433      0.001       0.074       0.271
bty_avg            0.0742      0.016      4.563      0.000       0.042       0.106
==============================================================================
Omnibus:                       30.145   Durbin-Watson:                   1.277
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               34.960
Skew:                          -0.672   Prob(JB):                     2.56e-08
Kurtosis:                       2.925   Cond. No.                         17.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
              df      sum_sq   mean_sq          F    PR(>F)
gender       1.0    2.260213  2.260213   8.086320  0.004659
bty_avg      1.0    5.819173  5.819173  20.819135  0.000006
Residual   460.0  128.574955  0.279511        NaN       NaN
```

$H_0: B_2 = 0$

$H_a: B_2 \neq 0$

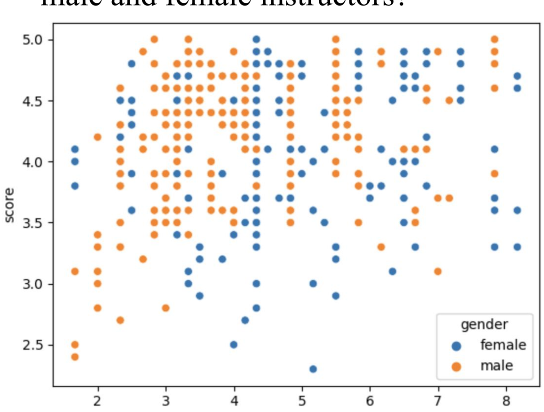
reject null

evidence to support relationship b/w beauty + score after accounting for gender

Comparing Regression Lines for Two Groups

We want to relate course evaluation scores (Y) to the beauty score assigned to the instructor (X_1) and the gender of (female v. male).

4. Is the relationship between beauty score and course evaluation score the same for male and female instructors?



Slope + intercept can both vary by gender

The Regression Model: Different Slopes

Let Gender = $\begin{cases} 1, & \text{Male} \\ 0, & \text{Female} \end{cases}$ $\qquad \varepsilon \sim N(0, \sigma_\varepsilon)$

Then,

$$Score = \beta_0 + \beta_1(Gender) + \beta_2(Beauty) + \beta_3(Gender * Beauty) + \varepsilon$$

We have two lines:

- Male $\quad Score = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)(Beauty) + \varepsilon$
- Female $\quad Score = \beta_0 + \beta_2(Beauty) + \varepsilon$

Interpretation:

- $H_0\colon \beta_3 = 0 \Rightarrow$ lines are // the relationship b/w beauty + score is the same b/w genders
- $H_0\colon \beta_1 = 0 \Rightarrow$ no diff in avg score after accounting for beauty

Interpretation

$\hat{y} = 3.9501 + 0.0306 * beauty$ (F)

$\hat{y} = (3.9501 + -0.1835) + (0.0306 + 0.0796) * beauty$ (M)

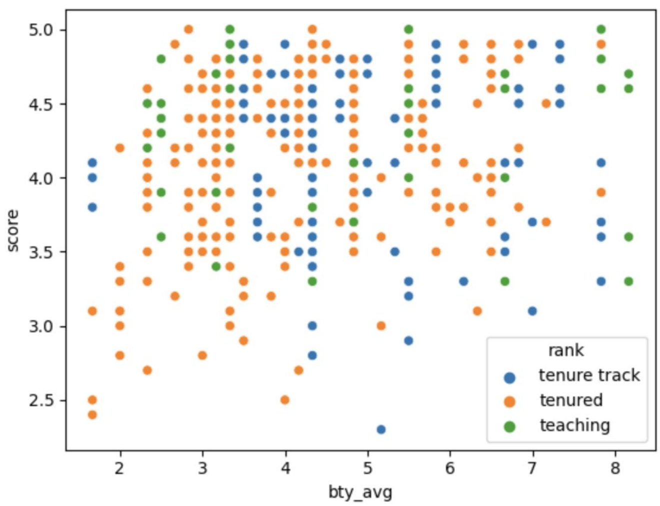
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  score   R-squared:                       0.071
Model:                            OLS   Adj. R-squared:                  0.065
Method:                 Least Squares   F-statistic:                     11.74
Date:                Tue, 25 Jul 2023   Prob (F-statistic):           2.00e-07
Time:                        18:00:28   Log-Likelihood:                -357.35
No. Observations:                 463   AIC:                             722.7
Df Residuals:                     459   BIC:                             739.3
Df Model:                           3
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept                   3.9501      0.118     33.475      0.000       3.718       4.182
gender[T.male]             -0.1835      0.153     -1.196      0.232      -0.485       0.118
bty_avg                     0.0306      0.024      1.277      0.202      -0.017       0.078
bty_avg:gender[T.male]      0.0796      0.032      2.452      0.015       0.016       0.143
==============================================================================
Omnibus:                       26.631   Durbin-Watson:                   1.282
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               30.276
Skew:                          -0.624   Prob(JB):                     2.67e-07
Kurtosis:                       2.890   Cond. No.                         42.3
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
                   df      sum_sq   mean_sq          F    PR(>F)
gender            1.0    2.260213  2.260213   8.174440  0.004442
bty_avg           1.0    5.819173  5.819173  21.046010  0.000006
bty_avg:gender    1.0    1.662530  1.662530   6.012817  0.014574
Residual        459.0  126.912425  0.276498        NaN       NaN
```

$H_0\colon \beta_3 = 0$

$H_a\colon \beta_3 \neq 0$

evidence to support diff slopes b/w gender

What if the categorical variable has more than 2 levels?



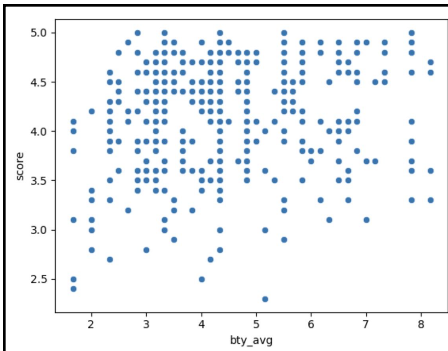
$$isTT = \begin{cases} 1 & \text{tenure track} \\ 0 \end{cases}$$

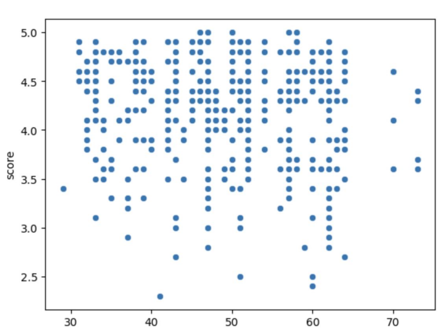
$$isTN = \begin{cases} 1 & \text{tenured} \\ 0 \end{cases}$$

teaching is baseline

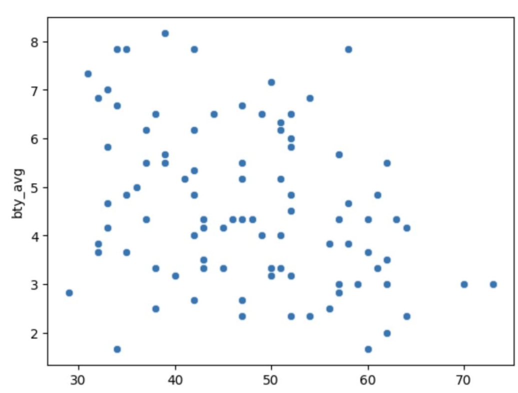
$$y = \beta_0 + \beta_1(isTT) + \beta_2(isTN) + \beta_3(beauty) + \varepsilon$$

$$\varepsilon \sim N(0, \sigma_\varepsilon)$$





What if I want to use more than one quantitative variable?



$$y = \beta_0 + \beta_1(beauty) + \beta_2(age) + \varepsilon$$

$$\varepsilon \sim N(0, \sigma_\varepsilon)$$

Effects of Multicollinearity

- If predictors are highly correlated amongst themselves, then the estimated regression coefficients and tests can be:

  unreliable

- The regression tests can be difficult to interpret individually

  how do I understand the effect of X_1 on Y holding X_2 constant when X_1 + X_2 are highly correlated?

- One variable alone might work just as well as many

- Explore the potential for multicollinearity by examining scatterplots of the response and the predictors (matrix plot)

Variance Inflation Factor (VIF)

- The variance of the coefficients of correlated predictors is inflated

- The Variance Inflation Factor (VIF) reflects the association b/w a predictor + all other predictors

- For each Predictor X_i, regress X_i onto the other predictors. Record R_i^2.
  Then, for the i^{th} predictor,

$$VIF_i = \frac{1}{1 - R_i^2}$$

- Be suspicious of multicollinearity when VIF > 5

  $R_i^2 > 80\%$