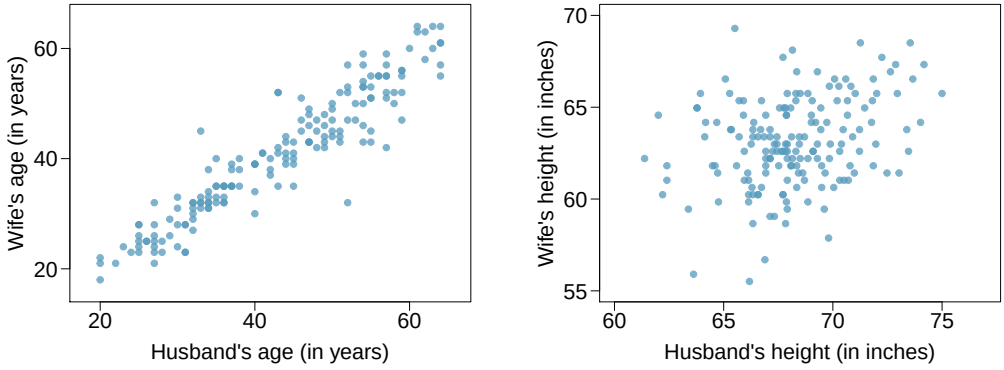
# Linear regression

Computational Mathematics and Statistics Camp
University of Chicago
September 2018

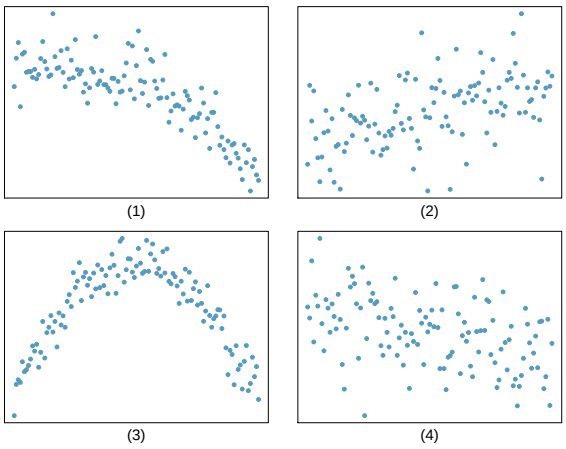
1. The Great Britain Office of Population Census and Surveys once collected data on a random sample of 170 married couples in Britain, recording the age (in years) and heights (converted here to inches) of the husbands and wives. The scatterplot on the left shows the wife's age plotted against her husband's age, and the plot on the right shows wife's height plotted against husband's height.




   a. Describe the relationship between husbands' and wives' ages.
   b. Describe the relationship between husbands' and wives' heights.
   c. Which plot shows a stronger correlation? Explain your reasoning.
   d. Data on heights were originally collected in centimeters, and then converted to inches. Does this conversion affect the correlation between husbands' and wives' heights?

2. Match the calculated correlations to the corresponding scatterplot.

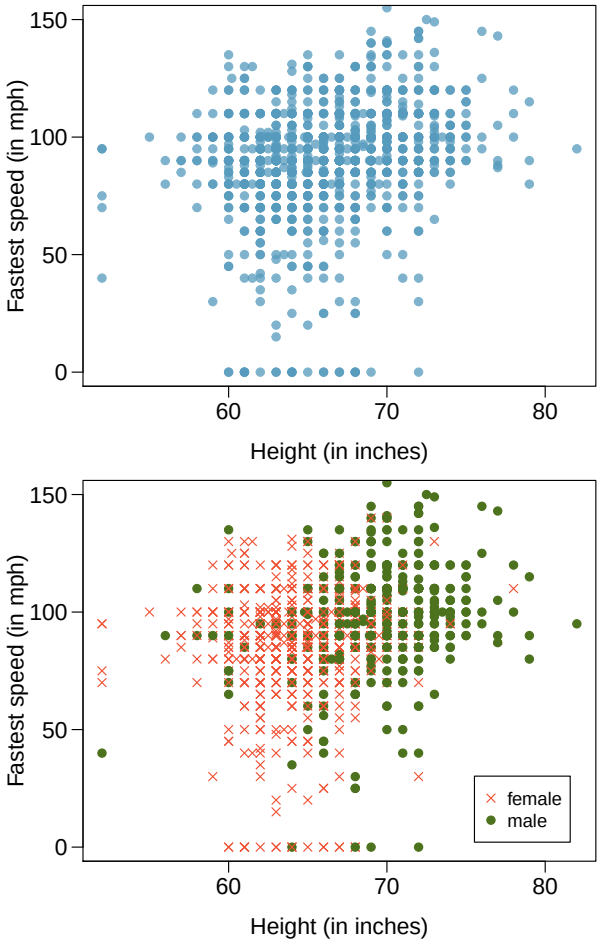
   a. $r = 0.49$
   b. $r = -0.48$
   c. $r = -0.03$
   d. $r = -0.85$



3. Determine if the following statements are true or false. If false, explain why.

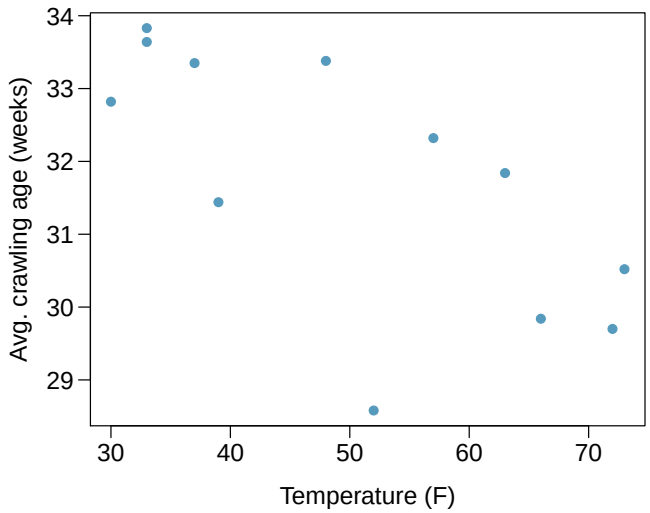
    a. A correlation coefficient of -0.90 indicates a stronger linear relationship than a correlation coefficient of 0.5.
    b. Correlation is a measure of the association between any two variables.

4. 1,302 UCLA students were asked to fill out a survey where they were asked about their height, fastest speed they have ever driven, and gender. The scatterplot on the left displays the relationship between height and fastest speed, and the scatterplot on the right displays the breakdown by gender in this relationship.




    a. Describe the relationship between height and fastest speed.
    b. Why do you think these variables are positively associated?
    c. What role does gender play in the relationship between height and fastest driving speed?

5. A study conducted at the University of Denver investigated whether babies take longer to learn to crawl in cold months, when they are often bundled in clothes that restrict their movement, than in warmer months. Infants born during the study year were split into twelve groups, one for each birth month. We consider the average crawling age of babies in each group against the average temperature when the babies are six months old (that's when babies often begin trying to crawl). Temperature is measured in degrees Fahrenheit (° F) and age is measured in weeks.




a. Describe the relationship between temperature and crawling age.
b. How would the relationship change if temperature was measured in degrees Celsius (° C) and age was measured in months?
c. The correlation between temperature in ° F and age in weeks was $r = -0.70$. If we converted the temperature to ° C and age to months, what would the correlation be?

6. Determine if a) or b) is higher or if they are equal. Explain your reasoning.

   For a regression line, the uncertainty associated with the slope estimate, $b_1$, is higher when

   a. there is a lot of scatter around the regression line or
   b. there is very little scatter around the regression line

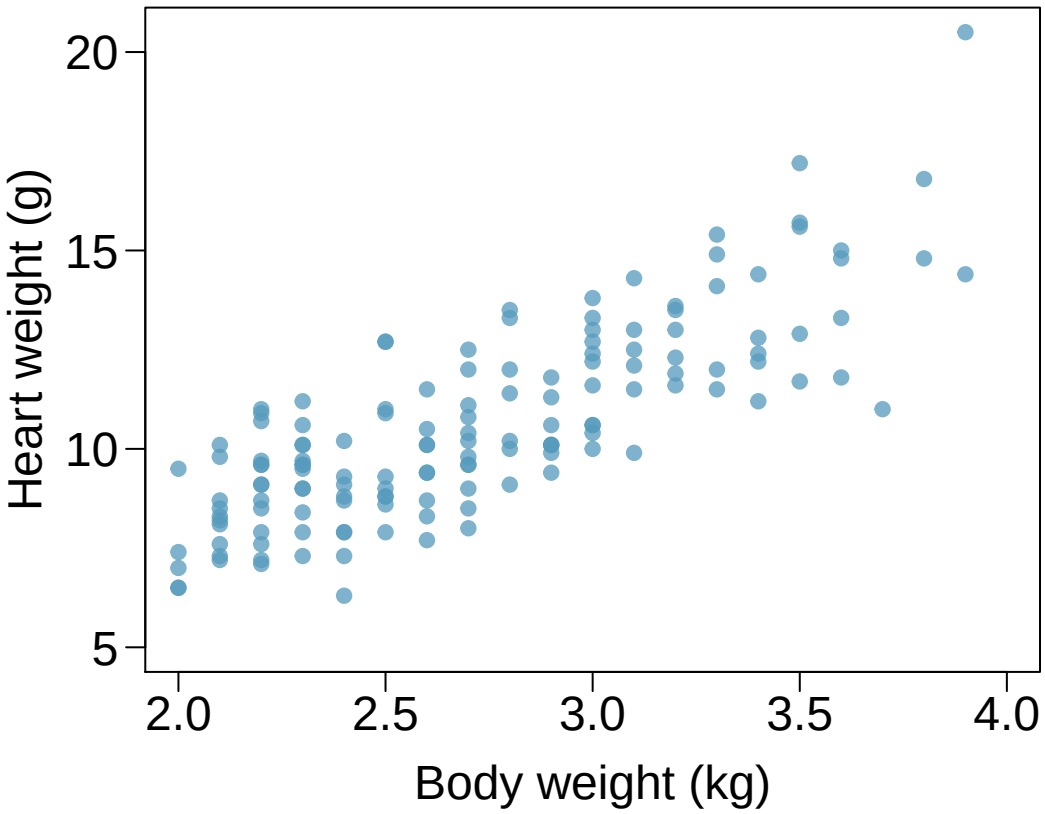
7. The Coast Starlight Amtrak train runs from Seattle to Los Angeles. The mean travel time from one stop to the next on the Coast Starlight is 129 mins, with a standard deviation of 113 minutes. The mean distance traveled from one stop to the next is 108 miles with a standard deviation of 99 miles. The correlation between travel time and distance is 0.636.

   a. Write the equation of the regression line for predicting travel time.
   b. Interpret the slope and the intercept in this context.
   c. Calculate $R^2$ of the regression line for predicting travel time from distance traveled for the Coast Starlight, and interpret $R^2$ in the context of the application.
   d. The distance between Santa Barbara and Los Angeles is 103 miles. Use the model to estimate the time it takes for the Starlight to travel between these two cities.
   e. It actually takes the Coast Starlight about 168 mins to travel from Santa Barbara to Los Angeles. Calculate the residual and explain the meaning of this residual value.
   f. Suppose Amtrak is considering adding a stop to the Coast Starlight 500 miles away from Los Angeles. Would it be appropriate to use this linear model to predict the travel time from Los Angeles to this point?

8. The following regression output is for predicting the heart weight (in g) of cats from their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cats.\[2mm]

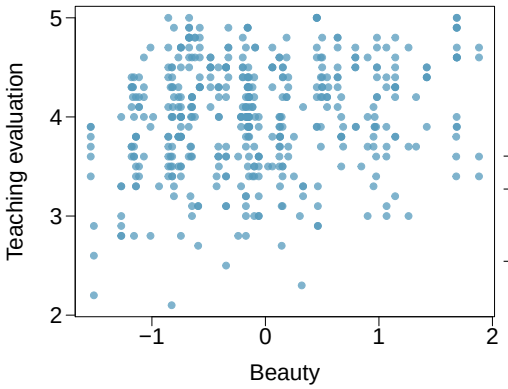
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -0.357 | 0.692 | -0.515 | 0.607 |
| body wt | 4.034 | 0.250 | 16.119 | 0.000 |

$$s = 1.452, \quad R^2 = 64.66\%, \quad R^2_{adj} = 64.41\%$$




a. Write out the linear model.
b. Interpret the intercept.
c. Interpret the slope.
d. Interpret $R^2$.
e. Calculate the correlation coefficient.

9. Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. Researchers at University of Texas, Austin collected data on teaching evaluation score (higher score means better) and standardized beauty score (a score of 0 means average, negative score means below average, and a positive score means above average) for a sample of 463 professors. The scatterplot below shows the relationship between these variables, and also provided is a regression output for predicting teaching evaluation score from beauty score.

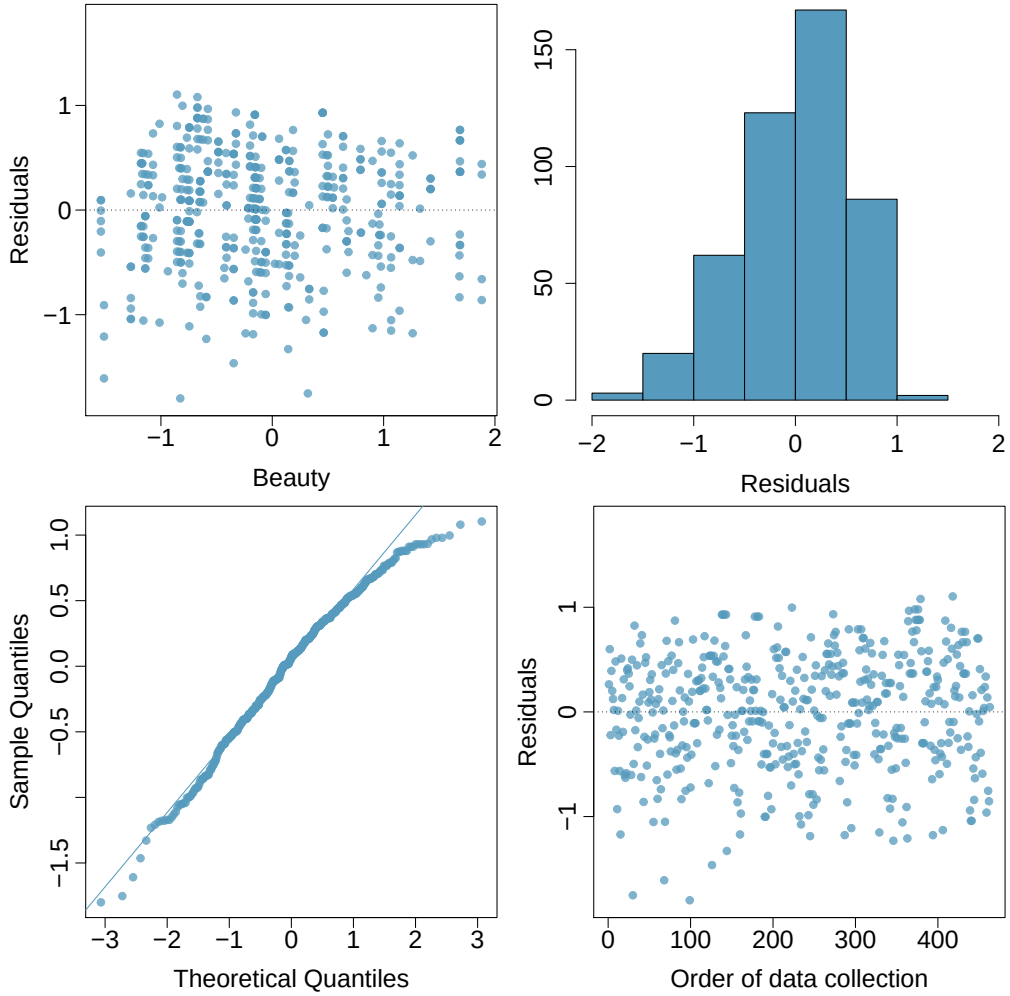



| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.010 | 0.0255 | 157.21 | 0.0000 |
| beauty | | 0.0322 | 4.13 | 0.0000 |

a. Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.

b. Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning.

c. List the conditions required for linear regression and check if each one is satisfied for this model based on the following diagnostic plots.




10. Dichotomous covariates present some challenges in interpretation in the context of regression models.

Consider a model such as:

$$Y_i = \beta_0 + \beta_1 X_{Di} + u_i$$

where $X_D$ indicates that $X$ is a dichotomous variable, coded either zero or one. Consider a situation where you have run a randomized experiment on a sample of 2000 individuals and obtained the following results for a two-sample $t$-test, where Control = 0 and Treatment = 1:

```
##
##  Welch Two Sample t-test
##
## data:  value by var
## t = 2, df = 2000, p-value = 0.04
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.0191 0.9619
## sample estimates:
##   mean in group control mean in group treatment
##                    3.51                    3.02
```

The results of a linear regression on the same data yield the following:

```
##
## Call:
## lm(formula = value ~ var, data = sim_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.055  -3.677   0.053   3.618  20.536
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.510      0.170   20.65   <2e-16 ***
## vartreatment   -0.491      0.240   -2.04    0.041 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.37 on 1998 degrees of freedom
## Multiple R-squared:  0.00208,    Adjusted R-squared:  0.00158
## F-statistic: 4.16 on 1 and 1998 DF,  p-value: 0.0414
```

a. In a few sentences, interpret the results of the $t$-test.
b. Identify the similarities between the results of the $t$-test and the linear regression model.
c. Do you need to estimate a linear regression model under these circumstances? Why or why not?