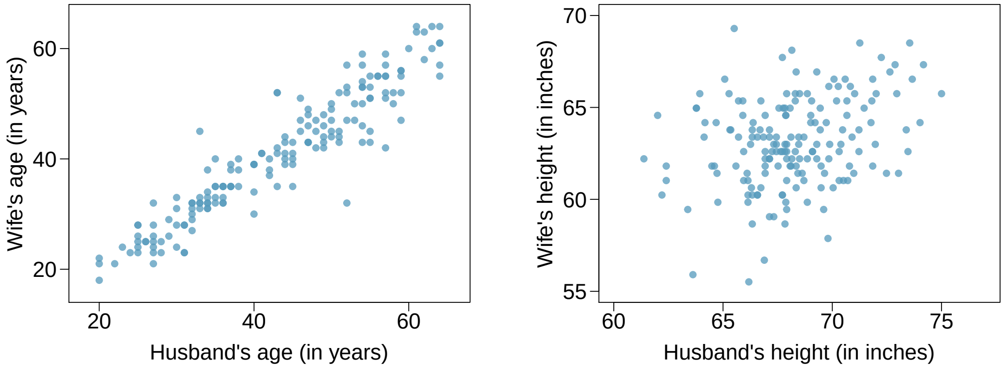
Linear regression

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.
   The association between husbands' and wives' ages is positive, moderate, and linear. There are a few potential outliers: a man in his early- to-mid 30s marrying a woman in her mid 40s, a man in his early 40s marrying a woman in her early-to-mid 50s, and a man in his early 50s marrying a woman in her early 30s.
   b. Describe the relationship between husbands' and wives' heights.
   The association between husbands' and wives' heights is weak but positive.
   c. Which plot shows a stronger correlation? Explain your reasoning.
   The age plot, where there is a more evident linear trend in the data.
   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?
   No. Correlation is unaffected by such rescaling.

2. Match the calculated correlations to the corresponding scatterplot.
   a. $r = 0.49$

$$r = 0.49 \rightarrow 2$$

   b. $r = -0.48$

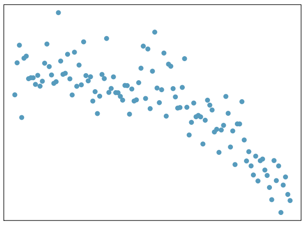
$$r = -0.48 \rightarrow 4$$

   c. $r = -0.03$

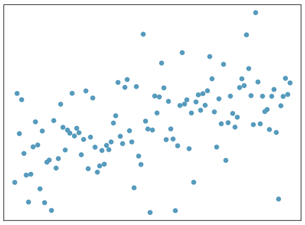
$$r = -0.03 \rightarrow 3$$

   d. $r = -0.85$

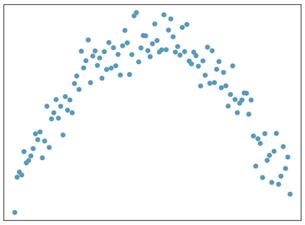
$$r = -0.85 \rightarrow 1$$



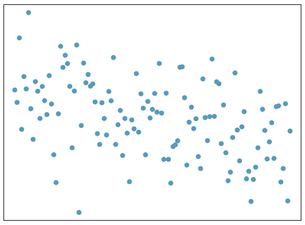
(1)



(2)



(3)



(4)

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

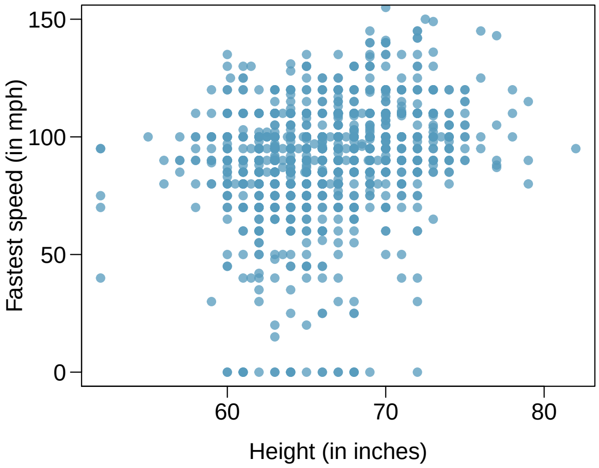
   a. A correlation coefficient of -0.90 indicates a stronger linear relationship than a correlation coefficient of 0.5.

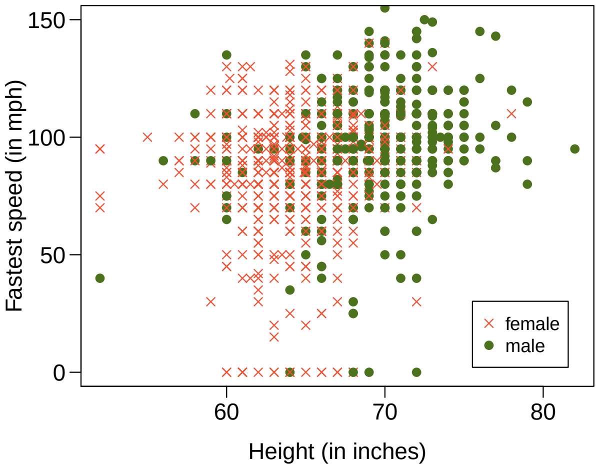
      True.

   b. Correlation is a measure of the association between any two variables.

      False, correlation is a measure of the linear association between any two numerical variables.

2. 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.

The relationship is positive, weak, and possibly linear. However, there do appear to be some anomalous observations along the left where several students have the same height that is notably far from the cloud of the other points. Additionally, there are many students who appear not to have driven a car, and they are represented by a set of points along the bottom of the scatterplot.

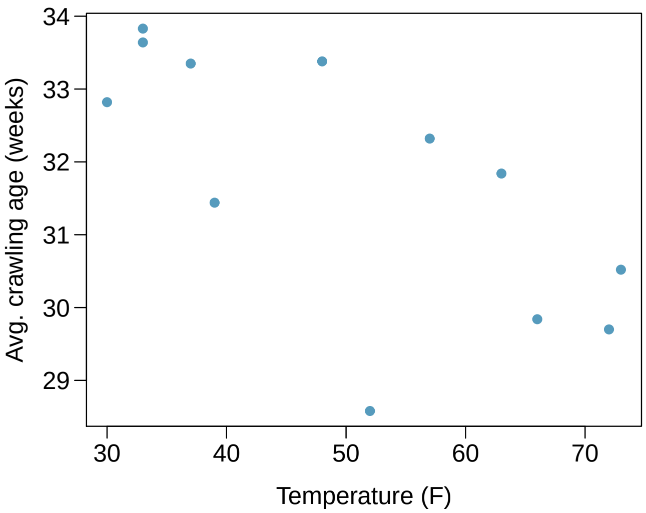
b. Why do you think these variables are positively associated?

There is no obvious explanation why simply being tall should lead a person to drive faster. However, one confounding factor is gender. Males tend to be taller than females on average, and personal experiences (anecdotal) may suggest they drive faster. If we were to follow-up on this suspicion, we would find that sociological studies confirm this suspicion.

c. What role does gender play in the relationship between height and fastest driving speed?

Males are taller on average and they drive faster. The gender variable is indeed an important confounding variable.

3. 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.

The relationship is moderate, negative, and linear. There is also an outlying month when the average temperature is about 53°F and average crawling age of about 28.5 weeks.

b. How would the relationship change if temperature was measured in degrees Celsius (° C) and age was measured in months?

Changing the units will not change the form, direction or strength of the relationship between the two variables. After all, if higher temperatures measured in °F are associated with lower average crawling age measured in weeks, higher temperatures measured in °C will be associated with lower average crawling age 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?

Since changing units doesn't affect correlation, R = -0.70.

4. 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

A is higher, the more the scatter the lower the correlation coefficient, and hence the higher the uncertainty around the regression line.

5. 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.

First calculate the slope.

$$b_1 = \frac{s_y}{s_x} R = \frac{113}{99} \times 0.636 = 0.726$$

Next, make use of the fact that the regression line passes through the point $(\bar{x}, \bar{y})$.

$$\begin{aligned} \bar{y} &= b_0 + b_1 \times \bar{x} \\ 129 &= b_0 + 0.726 \times 107 \\ b + 0 = 129 - 0.726 \times 107 \\ &= 51 \end{aligned}$$

The regression line can be written as:

$$\widehat{\text{travel time}} = 51 + 0.726 \times \text{distance}$$

b. Interpret the slope and the intercept in this context.

- $b_1 =$ For each additional mile in distance, the model predicts an additional 0.726 minutes in travel time.
- $b_0 =$ When the distance traveled is 0 miles, the travel time is expected to be 51 minutes. It does not make sense to have a travel distance of 0 miles. Here, the y-intercept serves only to adjust the height of the line and is meaningless by itself.

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.

   $R^2 = 0.636^2 = 0.40$. Approximately 40% of the variability in travel time is accounted for by the model, i.e. explained by the distance traveled.

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.

   In order to calculate the predicted time it takes to travel from Santa Barbara to Los Angeles, we plug in 103 for distance in the model:

$$\widehat{\text{travel time}} = 51 + 0.726 \times \text{distance} = 51 + 0.726 \times 103 \approx 126 \text{ minutes}$$

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.

   The residual can be calculated as $e_i = y_i - \hat{y}_i = 168 - 126 = 42$ minutes. A positive residual means that the model underestimates the travel time.

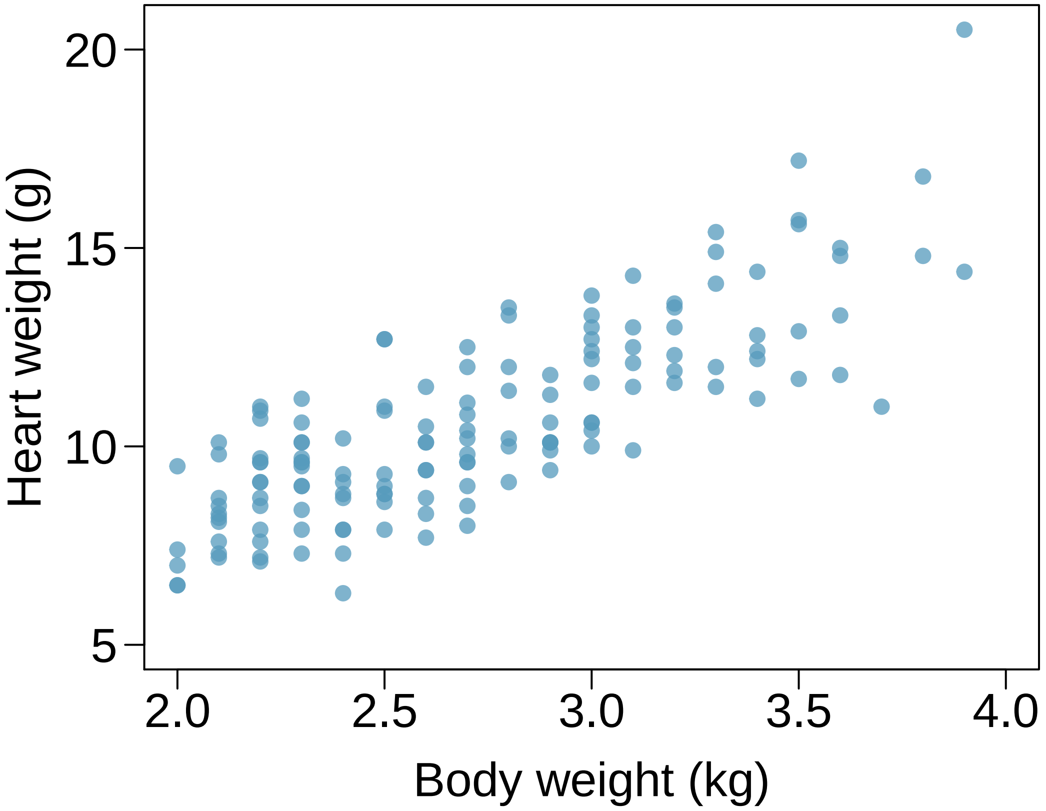
f. Suppose Amtrak is considering adding a stop to the Coast Starlight 500 miles away from Los Angeles. Would it be appropriate to use this linear model to predict the travel time from Los Angeles to this point?

   No, we should not use this linear model to predict the travel time for a distance of 500 miles as this would be extrapolation. The data we used to create the model is for approximately 10 to 350 miles of distance. The linear model may no longer hold outside the range of the data.

6. 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.

$$\widehat{\text{heart weight}} = -0.357 + 4.034 \times \text{body wt}$$

b. Interpret the intercept.

Expected heart weight of cats with 0 body weight is -0.357. This is obviously not a meaningful value, it just serves to adjust the height of the regression line.

c. Interpret the slope.

For each additional kilogram in cats' weights, we expect their hearts to be heavier by 4.034 grams, on average.

d. Interpret R^2.

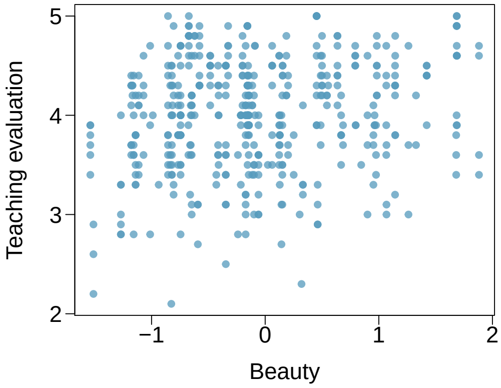
Body weight explains 64.66% of the variability in murder rates weights of cats' hearts.

e. Calculate the correlation coefficient.

$$\sqrt{0.6466} = 0.8041$$

7. 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.

The slope can be calculated as follows:

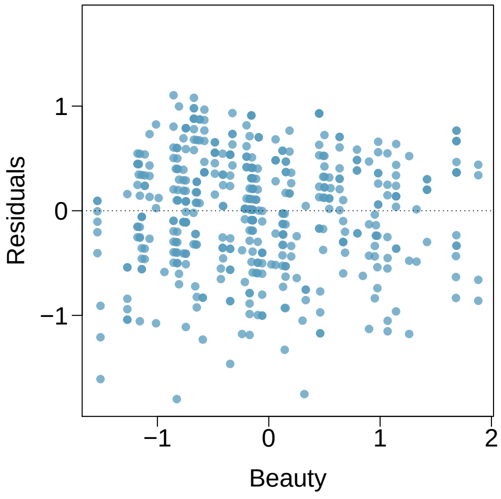
$$\begin{aligned} \bar{y} &= b_0 + b_1\bar{x} \\ 3.9983 &= 4.010 + b_1 \times -0.0883 \\ -0.0117 &= b_1 \times -0.0883 \\ b_1 &= 0.133 \end{aligned}$$

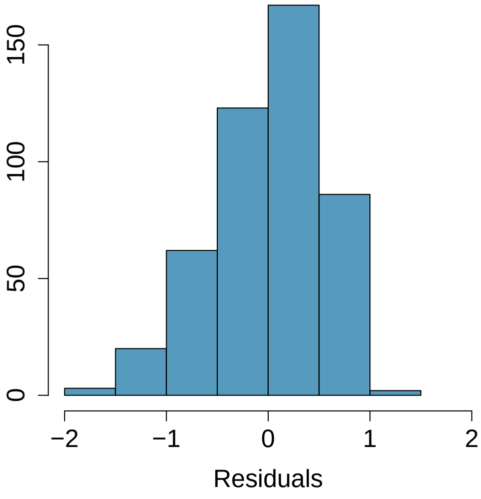
b. Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning.

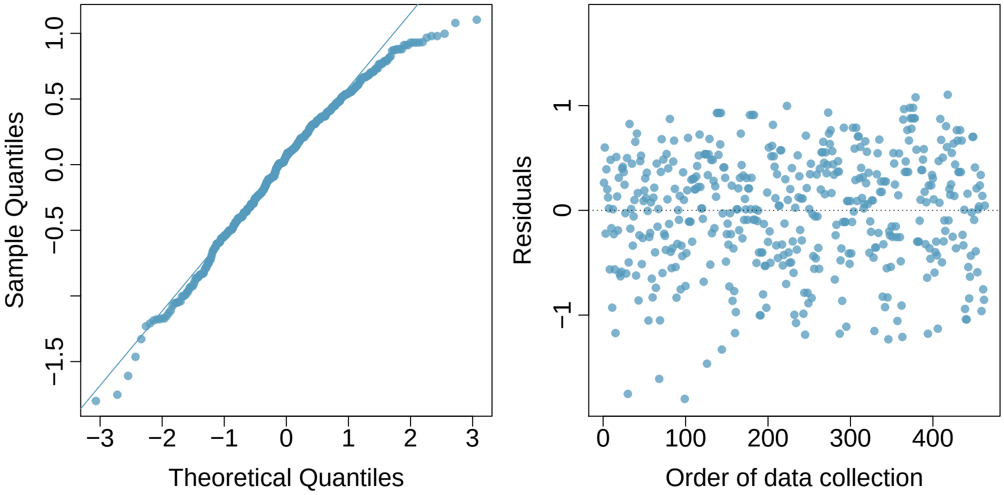
- $H_0 : \beta_{\text{beauty}} = 0$
- $H_A : \beta_{\text{beauty}} > 0$

With a t-score of 4.13 and a p-value of approximately 0, we reject H_0. The data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive.

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







1. Nearly normal residuals: The distribution of residuals look unimodal but somewhat left skewed, this assumption may not be satisfied.
2. Constant variance of residuals: The residuals vs. beauty score plot shows residuals that are randomly distributed around 0.
3. Independent observations/residuals: We are told the sample is random, so the observations/residuals are independent.
4. Linear relationship: Beauty score seems to be linearly associated with teaching evaluation score, or at least, the relationship doesn't appear non-linear.

8. 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.

   With a *t*-statistic of 2 and a p-value < 0.05, we reject the null hypothesis that there is no difference in the sample means. The treatment appears to be associated with on average a 0.57 increase in the outcome of interest. Since the study is a randomized controlled trial, we assume it was implemented appropriately and therefore the difference in means is the causal effect of the treatment on the outcome. NOTE: some students may mention this, but I don't expect many. It's not a camp on research design.

b. Identify the similarities between the results of the *t*-test and the linear regression model.

   - The estimated intercept $\hat{\beta}_0$ is the same as the mean of Y when $X_D = 0$.
   - The slope estimate $\hat{\beta}_1$ is the difference between $\bar{Y}|X_D = 1$ and $\bar{Y}|X_D = 0$; that is, $(\bar{Y}|X_D = 1) - (\bar{Y}|X_D = 0)$, which implies that
   - The mean of Y when $X_D = 1$ (which is the same as the expected value of Y given $X_D = 1$) is equal to $\hat{\beta}_0 + \hat{\beta}_1$
   - The estimated standard error of $\hat{\beta}_1$ is equal to the standard error in the difference of means.
   - The *t*-score for $\hat{\beta}_1 = 0$ is exactly the same as the *t*-statistic for the difference of means, as is the level of significance, 95% confidence interval, etc.

c. Do you need to estimate a linear regression model under these circumstances? Why or why not?

   There is little reason to do a regression of Y on X_D alone, since a *t*-test is in many respects simpler and more easily understood. That said, there is nothing "wrong" with including dichtomous Xs in a multivariate regression. In fact, if we ran an observational study or needed to condition on multiple independent variables, we should run a multivariate regression.