Inference topics we've covered so far...

- Identified estimators for common parameters
- Discussed the sampling distributions of estimators
- Introduced ways to judge the "goodness" of an estimator (bias, MSE, etc.)
- Used confidence intervals and hypothesis testing to make inferences about one mean, one proportion, or differences between two means
- Used maximum likelihood estimation to identify potential estimators for a parameter

None of the methods that we've discussed so far allow us to model the relationship (correlation) between two variables.

Describing the linear relationship

- Imagine that you have two quantitative variables that are correlated
- Think back to our coefficient of correlation, ρ. Now instead of just measuring the strength of the linear relationship, we want to get a more specific idea of what the relationship is
- In particular, we might want to predict values of one variable given the other variable.
- If the relationship between the 2 variables is not linear, sometimes appropriate transformations of the data may yield a more linear relationship.

Deterministic linear relationships

If the relationship between the two variables is a linear one, we can try to quantify/describe it based on the equation for a line If the correlation between the two variables is known to be perfect, or very close to it, we might use a deterministic model

- Then, we are saying that for every value of the independent variable X, we can exactly (or very close to it) predict the dependent variable Y
- This model may be appropriate for well-established laws of science, say: `momentum=mass*velocity`
- If I take an object, cause it to move with various velocities, then I can predict its momentum (with perhaps small measurement errors).
- This can be modeled by

  $$Y_i = \beta_0 + \beta_1 X_i$$

  where (X_i, Y_i) are (velocity, momentum) measurements, and β_0 and β_1 are the intercept and slope for the line that describes the relationship.

Probabilistic linear models

What if the correlation between two variables is not perfect? What if there seems to be a scatter cloud of points that has a general linear trend?

- We can express the relationship using the model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- This means that for every observation (X_i, Y_i), Y_i is a linear function of X_i plus some random "noise" given by ε_i.
- The noise is assumed to have mean 0 and be uncorrelated from data point to data point. This means that the errors (the noise that scatters data points around the underlying trend), are independent and are fall symmetrically around the mean of zero
- We cannot exactly predict Y for every X, but we can say what the expected value for Y given X is. Then, we are modeling the means of the Y_is given the changing X_is

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

Example of two correlated variables

We may be interested in predicting a student's second midterm score, given his/her first midterm score. We could look at a sample of such scores to help us determine how these grades are correlated.

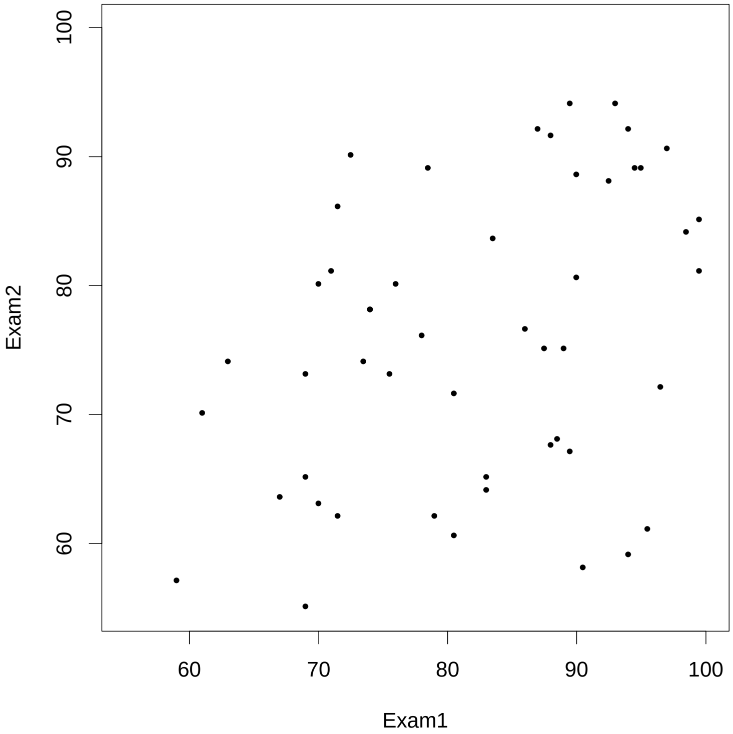

Figure 1: Sample of scores on the two midterms

Interpretations of estimates

- Since the data we have is just a sample from the population, we can only estimate the slope and intercept of the "true" regression line.
- These estimates are denoted $\hat{\beta}_0$ and $\hat{\beta}_1$, and the fitted values that they yield are denoted $\hat{y}$
$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
- $\hat{\beta}_1$ gives the estimated change in the dependent variable associated with a one unit change in the independent variable
- $\hat{\beta}_0$ gives the estimated value of the dependent variable when independent variable is 0

How to fit the line?

- We want the line to be as close to the data points as possible, but since there is so much variation from a strict linear pattern, the methodology for doing this is unclear.
- How do we measure the distances between the data points and the line? We could use vertical distance, horizontal distance, or closest distance (perpendicular approach from each data point to the fitted line).

Least squares approach

- We choose to minimize the sum of the squared vertical distances between each data point and the regression line.
- These distances are called the errors, and denoted by $e_i = y_i - \hat{y}_i$
- The approach is called "least squares" since it minimizes the sum of the squared distances. (The sum of the vertical distances, not squared, is 0.)
- This provides us with the following least squared estimates for the slope and intercept of the regression line:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

Back to our grades data

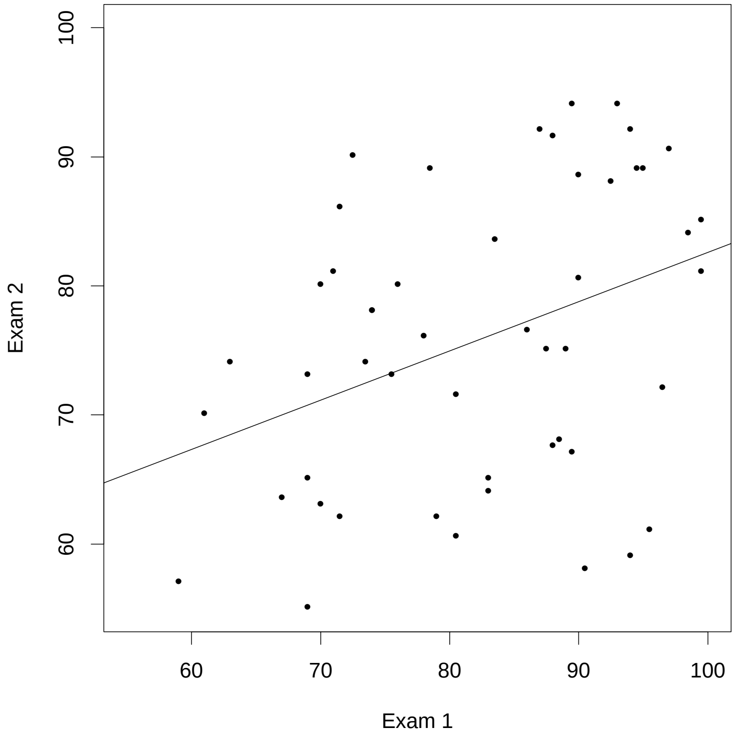

Figure 2: Fitted regression line: $\hat{y} = 44.396 + 0.382x$

$\sum(x_i - \bar{x})(y_i - \bar{y}) = 2290.73 \qquad \sum(x_i - \bar{x})^2 = 5996.445$

$\bar{x} = 82.31 \qquad \bar{y} = 75.84$

When are these estimates good?

When certain conditions are met, we can say that our least squares method yields good estimates $\hat{\beta}_1$ and $\hat{\beta}_0$ of β_1 and β_0. These are called the Gauss-Markov conditions.

- $E(\varepsilon_i) = 0$ for all i
- $Var(\varepsilon_i) = \sigma^2$ for all i
- $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$

These assumptions are also necessary for us to make statements about the mean and variance of the estimates and for further inference about the model parameters.