An Introduction to Regression

Dr. Austin Brown

Kennesaw State University

Note

- Some of the contents of these slides have been adapted from materials created by Dr Taasoobshirazi (thanks Dr T!).

Table of Contents

Introduction

- Statistics, in general, is the study of variability. In the sport of baseball, for example, not every team has the exact same number of regular season wins.
  - Natural variability exists.

- One of the foundational premises of science is to try and understand why things vary. From this seemingly simple question, ultimately, the field of statistics was born.

Introduction

- One of the tools that scientists across a wide array of disciplines use is called **regression analysis**. (Note, when people typically say "regression" they are referring to linear regression, but be mindful that other types of regression methods also exist).

- Effectively, with regression, we have some outcome variable of interest (say, MLB team wins or systolic blood pressure or amount of nightly REM sleep) that we know, based on our experience/expertise/review of the literature, may be associated with some other measurable explanatory variables.
  - For the baseball example, I know that runs scored and earned run average are likely related to team wins.

Introduction

- While we'll get into the weeds of exactly how to fit and assess a regression model later on, suffice it to say that the results of the analysis will help us better answer the question: "Why and how does this observable phenomenon vary?"
- Somewhat implicit in that question is causality. We have to be careful with jumping immediately to that conclusion and will discuss this more throughout the semester.

Visualizing the Relationship Between Variables

- Okay, aside from all this waxing poetic about the scientific method, let's look at this a bit more pragmatically: I have some variables I'm interested in, where do I start?
- For those of you who have had courses with me before know, I'm a big proponent of data visualization as a way of describing data in all phases of an analysis. So let's start there!
- Let's say I want to visually examine the relationship between MLB team wins and runs scored. Team wins will be my outcome and runs scored will be my predictor.

Visualizing the Relationship Between Variables

- Typically, the way I interpret a scatterplot is by answering four basic questions:

1. What is the form of the relationship (linear/nonlinear)?
2. What is the direction of the relationship (positive, negative, not clear)?
3. What is the strength of the relationship (weak, moderate, strong)?
4. Are there any unusual characteristics (clustering, outliers, etc)?

- In our baseball example, what can we say?

Visualizing the Relationship Between Variables

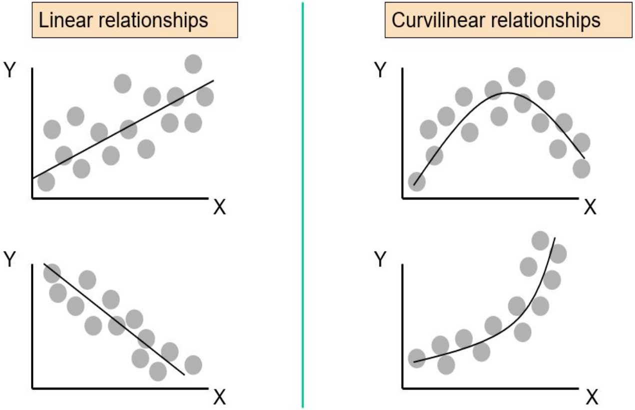

Figure 1: From: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall, c/o Dr. Taasoobshirazi

Visualizing the Relationship Between Variables

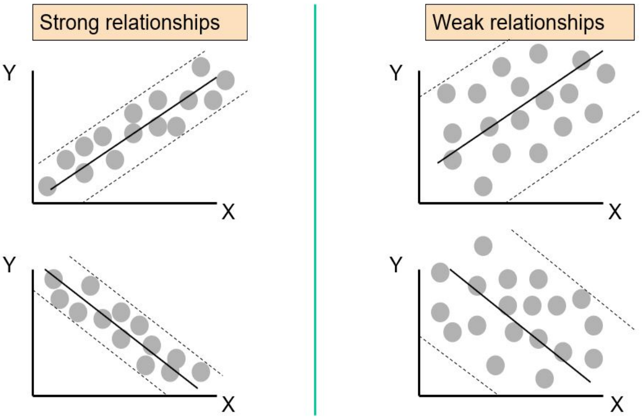

Figure 2: From: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall, c/o Dr. Taasoobshirazi

Visualizing the Relationship Between Variables

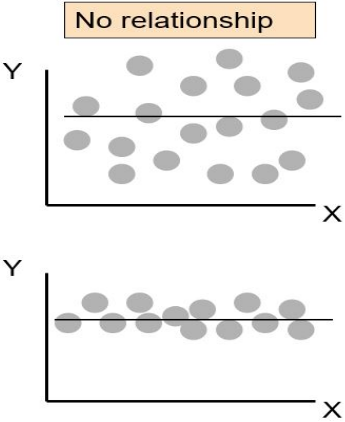

Figure 3: From: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall, c/o Dr. Taasoobshirazi

Covariance & Correlation

- We have so far learned some general methods for interpreting a scatterplot, and thus, the relationship between two quantitative variables.
- While this is a useful skill (especially in the exploratory phase of an analysis), it is a bit subjective. What someone considers a moderately strong linear relationship, someone else may interpret somewhat differently.
- There is resultingly a need to quantify the relationship between two variables. We do this through **correlation** and **covariance**.

Covariance & Correlation

- Sample covariance between two random variables, say X and Y, is defined as:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Covariance tells us how two variables literally "co-vary" or move together. (Also for you linear algebra fans out there it and regular variance are examples of inner products, which is a handy property).

Covariance & Correlation

- If two variables are positively related, then covariance will also be positively signed and if the converse is true, then covariance will be negatively signed.

- It is sometimes implied that a covariance of 0 means that the two variables are independent, which is typically true except in cases where the data are strongly nonlinear (e.g., a parabolic relationship).

Covariance & Correlation

- In mostly every manuscript I've read that does regression analysis or some sort of correlational analysis, I don't think I have ever seen the covariance between two variables reported.
- Why? Because while we can interpret the sign of covariance to help us understand the relationship, the actual value of covariance is a bit more challenging (consider the units!).
- Thus, having a unitless measure of the strength and direction of the relationship between two quantitative variables is much more useful.

Covariance & Correlation

- This is where **Pearson's Product-Moment Correlation Coefficient** (more colloquially, correlation) becomes a valuable tool.

- Pearson's Correlation (for a sample) is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Covariance & Correlation

- More succinctly, it is the covariance between X and Y divided by the product of the standard deviations of X and Y:

$$r = \frac{s_{xy}}{s_x s_y}$$

- What's nice about Pearson's correlation is that it is:
  1. Unitless
  2. Bounded between -1 and $+1$.

Covariance & Correlation

- A value of r that is positive implies a positive relationship and a negatively signed value implies the opposite, just like covariance (makes sense as standard deviation must be positive...the sign of r comes from s_{xy}).

- If $r = 0$, the two variables are said to be "uncorrelated" (not necessarily independent).

- It is important to point out that the Pearson correlation coefficient measures the strength and direction of a *linear* relationship (i.e., if a strong nonlinear relationship is present, this coefficient won't be able to detect it very effectively).

Covariance & Correlation

- If our research hypothesis states that, based on our previous knowledge/experience, the correlation should be non-zero (i.e., significantly different from zero), then how do we go about testing that hypothesis?

- The traditional test used for testing the significance of a correlation coefficient tests the hypotheses:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

Covariance & Correlation

- Assuming that our observations were randomly sampled from normal distributions (important assumption, but often ignored), the test statistic we use is:

$$t_{Stat} = r\sqrt{\frac{n-2}{1-r^2}} \sim t(n-2)$$

- Obviously, most of the time we don't do this by hand. Let's see how to run these tests using R.

Regression Concepts

- Now, as alluded to previously, the purpose of regression analysis is to build a model (in this case a linear model) which describes the relationship between a response or outcome variable and one or more predictor or explanatory variables.

- In the case of simple linear regression (meaning we only have one predictor variable), the form of the model is typically given as:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Here, y_i is an observed value of the response variable, β_0 is the y-intercept, β_1 is the slope parameter, x_i is an observed value of the explanatory variable which is paired with y_i, and ε_i is the random error term.

Regression Concepts

- Why do we need a random error term? Let's consider the penguins example.
- If we were to draw a line through the data to approximate the relationship between bill length and bill depth, not every single observation will perfectly fall along that line.
- The random error (sometimes also called the "residual term") accounts for the deviation an observed value is from the line we fit through our data.

Regression Concepts

- What are our assumptions?

1. An approximate linear relationship between the response and the predictor(s)
2. Observations are randomly sampled (independence of observations)
3. The residuals are normally, independently, and identically distributed as:

$$\varepsilon_i \sim N(0, \sigma^2)$$

- Notice, the assumption of normality is not upon the response variable.

Regression Concepts

- Assuming the explanatory variable, X, is fixed, then what the simple linear regression equation (which can be generalized to multiple linear regression) is saying is that:
- We have a random variable, ε, that has a mean of 0, to which we're adding a fixed constant $\beta_0 + \beta_1 x_i$.
- As you may recall (or will learn) from math stats, if you simply add a constant to a random variable, only its mean changes. Its variance and its overall distributional form will stay the same.

Regression Concepts

- Let's recall the basic principle of regression worded in a slightly different manner.
- Based on my prior knowledge of the data, I believe that observed variability in my response variable (e.g., regular season wins) can be explained by (or depends on) my predictor variable (e.g., runs scored).
- This reconceptualizes regression as a technique for conditioning my response on a set of predictors.

Regression Concepts

- All of this is to say that the way regression (and all generalized linear models) work is that they model/estimate the conditional mean of the response variable given values of my predictor variable:

$$E[y_i|x_i] = \beta_0 + \beta_1 x_i$$

- This is why when we check assumptions (which we'll discuss later on), we check them for the residuals, not the response (the residuals should have a constant mean and variance).

Regression Concepts

- The conditional variance is:

$$Var[y_i|x_i] = Var[\beta_0 + \beta_1 x_i + \varepsilon_i] = \sigma^2$$

- Which, as we've discussed, is the variance of the residuals and why particular importance needs to be placed on residual analysis when validating a model.

Regression Concepts

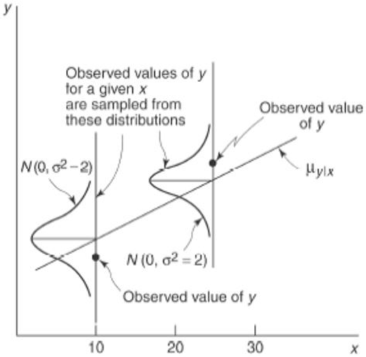

Figure 1.2 How observations are generated in linear regression.

Figure 4: Page 3 of your text

Regression Concepts

- The values that fall along the line, which are literally estimated conditional means, I will refer to as “predicted” or “fitted” values and denote them, $\hat{y}_i$.
- The closer the points are to falling exactly on the line, the better of a job our predictor variable does at explaining the variability in the response and vice versa.

Regression Concepts

- There are a few practical considerations we should be aware of before utilizing regression methods:

1. Be wary of extrapolation.
2. Your model is only as good as the data you've collected or are using.
3. There are lots of spurious correlations out there which means just because you can fit a model doesn't mean you should.
4. **CORRELATION DOES NOT IMPLY CAUSATION**.