Introduction to Scatterplots and Regression

• regression analysis is the study of how a response of interest depends upon one or more potential predictors

• a response is some variable that varies over a population of interest and that we have access to for at least certain members of the population (‘the data we want to predict or explain’)
  — ‘response’ is also referred to as the ‘dependent variable’

• a predictor is another variable that varies over the same population and that we think might be connected in some way to the response (‘how we want to predict or explain the data’)
  — other names for predictors are ‘explanatory variables’ and ‘independent variables’ (latter can cause confusion)
Examples of Responses and Potential Predictors: I

- adult height of daughters as predicted by height of mothers
  - Q: what might the population of interest be here?
  - Q: what might be some other potential predictors?
- short-term (1 year) measurements of wind speed at proposed wind-farm location as predicted by wind speed at nearby reference sites, for which long-term (50+ years) measurements are available
  - Q: what is the population of interest here?
  - Q: other potential predictors?
Examples of Responses and Potential Predictors: II

- interval of time to next eruption of Old Faithful Geyser (Yellowstone National Park, Wyoming) as predicted by length of current eruption
  - Q: population of interest?
  - Q: other potential predictors?

- weight of a gold coin (manufactured with a standard weight of 7.9876 g and taken from circulation in Manchester, England) as predicted by its age in decades
  - Q: population of interest?
  - Q: other potential predictors?
Uses for Regression Analysis

• summarize sets of data
• extract interesting patterns/relationships in data
• predict future values of response for given values of predictors
• predict result of certain types of interventions
• formulate and test scientific hypotheses
• introduce discipline into ‘eye-balling’ patterns in data
• determine useful/important predictors for a given response
• help design forthcoming scientific experiments
• identify unusual data points (discordant with bulk of data, an example being Florida votes in 2000 presidential election)
  — above is by no means an exhaustive list!!!
Linear Least Squares Regression

- regression analysis is a minor industry (*many* different twists!)
- simplest form is linear regression with parameters estimated by least squares (the focus of Stat 423)
- simplest form of linear least squares regression involves one response and one predictor
- will review one response/one predictor case first before getting into complications (of which there are many!) introduced when dealing with one response/multiple predictors
- given a set of data \((x_i, y_i), i = 1, 2, \ldots, n\), containing values for response \(Y\) and predictor \(X\), starting point in linear regression analysis is a *scatterplot*, i.e., a plot of \(y_i\)'s versus \(x_i\)'s (here \(n\) is the number of cases or units we have in all)
Inheritance of Height

• as a first example, consider using a mother’s height to predict a daughter’s adult height (the response), a classic example from which the term ‘regression’ arose
  – $X$ here is the mother’s height (mheight)
  – $Y$ is the daughter’s height (dheight)

• data consist of $n = 1375$ cases collected by K. Pearson from 1893 to 1898 in the United Kingdom

• side issues to ponder in view of extract on next overhead
  – what population was Pearson considering?
  – was Pearson’s sample a random sample from this population?
Professor KARL PEARSON, of University College, London, would esteem it a great favour if any persons in a position to do so, would assist him by making one set (or if possible several sets) of anthropometric measurements on their own family, or on families with whom they are acquainted. The measurements are to be made use of for testing theories of heredity, no names, except that of the recorder, are required, but the Professor trusts to the *bona fides* of each recorder to send only correct results. Each family should consist of a father, mother, and at least one son or daughter, not necessarily the eldest. The sons or daughters are to be at least 18 years of age, and measurements are to be made on not more than two sons and two daughters of the same family. If more than two sons or two daughters are easily accessible, then not the tallest but the eldest of those accessible should be selected. To be of real service the whole series ought to contain 1000–2000 families, and therefore the Professor will be only too grateful if anyone will undertake several families for him.
Scatterplot of Daughters’ Versus Mothers’ Heights
Scatterplot with Jittering to Eliminate Overplotting

![Scatterplot with Jittering](image-url)
Scatterplot with Jittering Using R’s jitter Function
Scatterplot with R’s Jittering Only Where Needed
What Does Scatterplot Tell Us?: I

- range of heights seems similar for mothers and daughters
- if height of daughters were *independent* of their mothers’ heights, would expect heights of daughters to be scattered at random about their mean value, so will augment plot by adding a horizontal line indicating this mean value
- if each daughter had a height *exactly* equal to her mother (i.e., \( Y = X \) so that \( Y \) and \( X \) would be *functionally dependent*), would expect points to fall on a 45° line, so will augment plot with this line also
Scatterplot with R’s Jittering Only Where Needed
What Does Scatterplot Tell Us?: II

- hypotheses of independence and functional dependence don’t seem viable
- let’s consider mothers with heights 57”, 58”, . . . , 68”, and, for each height, show sample mean of daughters on scatterplot as red filled circles (not many mothers shorter than 57” and taller than 68”, so we’ll let these data speak for themselves)
Scatterplot with R’s Jittering Only Where Needed
What Does Scatterplot Tell Us?: III

• mean heights of daughters seem to ‘regress’ toward overall mean on the average
  — shorter moms tend to have daughters taller than themselves
  — taller moms tend to have shorter daughters than themselves
• useful summary of increase in mean height of daughters with increasing mothers’ height is a so-called regression line
• more later on how to best set intercept & slope of this line
• some tentative conclusions
  — daughters’ mean heights vary linearly with mothers’ heights, but with lots of individual variation (scatter about the line)
  — can rule out independence or functional dependence
Atmospheric Pressure & Boiling Point of Water: I

- as a second example, consider historical data discussed in 1857 article by physicist James Forbes

- data consists of co-located measurements of atmospheric pressure (by a barometer) and of boiling point of water (by a thermometer) in 17 different locations in the Alps and Scotland (differing in altitude)

- scientific question of interest: how well can pressure be predicted by boiling point of water?

- side issues to ponder
  - observational rather than controlled experiment
  - what population was Forbes considering?
  - was Forbes’ sample a random sample from this population?
Scatterplot of Pressure Versus Boiling Point
Scatterplot of $\log_{10}(\text{Pressure})$ Versus Boiling Point
Atmospheric Pressure & Boiling Point of Water: II

- note: lines were added to scatterplots to help show that relationship between responses (either pressure of log thereof) and predictor (boiling point) is indeed linear to a good approximation (will discuss later how these lines were determined)
- Forbes had physical theory saying that log(pressure) is linearly related to boiling point (i.e., a functional relationship):
  \[ \log (\text{pressure}) = \beta_0 + \beta_1 \times \text{boiling point}, \]
  which is an example of a transformation of response \( Y \)
- points in scatterplot do not lie exactly on a straight line, but this might be due to measurement noise rather than a defect in the physical theory
- scatterplot indicates one vertically separated point, also called an outlier; i.e., a case markedly different from other cases
Length of Smallmouth Bass at Different Ages

• as a third example, consider lengths of $n = 439$ smallmouth bass caught in West Bearskin Lake in Minnesota, with ages of fish gotten from examination of fish scales (and ages restricted to eight years or less)

• scientific question of interest: how does length of fish (response) depend on age (predictor)?

• side issues to ponder
  — cross-sectional observational data rather than longitudinal
  — what’s the population here?
  — is this a random sample from the population under study?

• following plots suggest that average length of fish varies approximately linearly with age
Scatterplot of Length Versus Age
Scatterplot of Length Versus Jittered Age
Predicting the Weather

- fourth example addresses question: can early season snowfall (Sept–Dec) predict snowfall in next six months (Jan–June)?
- data consist of 93 years of measured snowfalls (in inches) from Fort Collins, Colorado
- scatterplot does not indicate a strong relationship between the predictor (early snowfall) and the response (late snowfall)
  - horizontal line indicates sample mean of late snowfalls
- can use linear regression analysis to see if we can reject null hypothesis that early & late season snowfalls are unrelated versus alternative that they are linearly related with a nonzero slope
- scatterplot here resembles a *null plot* (response has mean and variance that do not dependent on predictor)
Scatterplot of Late Snowfall Versus Early Snowfall

![Scatterplot](image_url)
Monthly Soil Temperatures in Mitchell, Nebraska

- fifth example is one with a scatterplot that is seemingly another null plot: average monthly soil temperature (centigrade) at depth of 20 cm in Mitchell, Nebraska
  - response is temperature
  - predictor is month count
- lesson: default settings for scatterplot can be misleading!
Monthly Soil Temperature Versus Month Count

Average soil temperature

Months after January 1976
Monthly Soil Temperature Versus Month Count

Average soil temperature

Months after January 1976
Anscombe’s Cartoon Scatterplots

- next 3 examples of scatterplots are due to Anscombe (1973) and are of artificially constructed predictors $X$ & responses $Y$ with $n = 11$ cases showing

  [1] a predictor/predictor relationship that is *not* linear;
  [2] a vertically separated point (an *outlier*); and
  [3] a horizontally separated point (a *leverage point*)

- note: will revisit Anscombe’s examples later (more to them than meets the eye!)
Nonlinear Scatterplot

![Nonlinear Scatterplot Diagram]
Scatterplot with Outlier
Scatterplot with Leverage Point

Predictor

Response

ALR–13, 4

I–32
Mean Functions: I

- reconsider (unjittered) heights of mothers \(X\) & daughters \(Y\)
- scatterplot of \(Y\) versus \(X\) depicts how to form bivariate distribution of these two variables:

\[
\Pr(X = x, Y = y) = \frac{\text{\# of mothers of height } x \text{ with daughter of height } y}{n},
\]

with the \(n = 1,375\) units taken to be the entire population

- interpretation: if we were to pick a particular mother/daughter at random (with all \(n\) units being given an equal chance of being picked), then \(\Pr(X = 64, Y = 65)\) is the probability that we would get a 64” mother with a 65” daughter

- mothers go from 55” to 71” & daughters from 55” to 73”, so

\[
\sum_{x=55}^{71} \sum_{y=55}^{73} \Pr(X = x, Y = y) = 1
\]
Mean Functions: II

- univariate distribution of daughters' heights is given by

$$\Pr(Y = y) = \sum_{x=55}^{71} \Pr(X = x, Y = y) = \frac{\# \text{ of daughters with height } y}{n}$$
Mean Functions: III

• mean $E(Y)$ of univariate distribution = sample mean $\bar{y}$ of $y_i$’s:

$$E(Y) = \sum_{y=55}^{73} y \times \Pr(Y = y) = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$  

(unassigned exercise: verify)
Jittered Scatterplot with Cases $x_i = 64$ in Red
Mean Functions: IV

- univariate distribution of daughters’ heights with 64” mothers

\[
\Pr(Y = y \mid X = 64) = \frac{\# \text{ of daughters with height } y \text{ & mother of height } 64''}{\# \text{ of mothers of height } 64''}
\]
Mean Functions: V

• $\Pr(Y = y \mid X = 64)$ is called \textit{conditional distribution} of $Y$ given $X = 64$

• can argue that

$$
\Pr(Y = y \mid X = 64) = \frac{\Pr(Y = y, X = 64)}{\Pr(X = 64)},
$$

which is one form of the celebrated \textit{Bayes theorem}

• conditional distribution is a special case of a univariate distribution, so it has a mean, to be denoted by $\mathbb{E}(Y \mid X = 64)$
Mean Functions: VI

- mean of $\Pr(Y = y \mid X = 64)$ determined in usual manner:

$$E(Y \mid X = 64) = \sum_{y=55}^{73} y \times \Pr(Y = y \mid X = 64)$$
Mean Functions: VII

• considering cases other than $x = 64$ leads us to *mean function*:

$$
E(Y \mid X = x) = \sum_{y=55}^{73} y \times \Pr(Y = y \mid X = x),
$$

which is a function that depends on value of mother’s height $x$

• conditional distribution $\Pr(Y = y \mid X = x)$ and $E(Y \mid X = x)$ for heights $x = 57, 58, \ldots, 68$ are shown on plots that follow
Conditional Distribution and Mean for $X = 68$
Jittered Scatterplot with $E(Y \mid X = x)$
Mean Functions: VIII

• above demonstration assumes that population of interest is just the $n = 1,375$ cases; i.e., bivariate distribution for $X$ and $Y$ given by empirical bivariate distribution of these cases

• assuming now that $n = 1,375$ cases are a sample from a well-defined population, mean function shown previously is an estimate of unknown mean function $E(Y \mid X = x)$ for this population

• given response $Y$ and predictor $X$, use of simple linear regression requires the key (and strong!) assumption that the unknown mean function for population takes the simple form

$$E(Y \mid X = x) = \beta_0 + \beta_1 x,$$

i.e., a straight line parameterized by intercept $\beta_0$ and slope $\beta_1$
Variance Functions: I

• recall that variance for random variable $Y$ with expected value (mean) $E(Y)$ is defined as

$$\text{Var}(Y) = E\{(Y - E(Y))^2\}$$

• reconsidering heights of mothers ($X$) & daughters ($Y$) and taking $n = 1,375$ units to be entire population of interest, have

$$\text{Var}(Y) = \sum_{y=55}^{73} (y - E(Y))^2 \times \Pr(Y = y) = \frac{1}{n} \sum_{i=1}^{n} (y_i - E(Y))^2,$$

(unassigned exercise: verify last equality), where, as before,

$$E(Y) = \sum_{y=55}^{73} y \times \Pr(Y = y) = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$
Variance Functions: II

• conditional variance is variance of response $Y$ given that predictor $X$ is fixed at $X = x$

• will denote conditional variance by $\text{Var}(Y \mid X = x)$

• for mother/daughter example,

$$\text{Var}(Y \mid X = x) = \sum_{y=55}^{73} (y - \text{E}(Y \mid X = x))^2 \times \text{Pr}(Y = y \mid X = x),$$

where

$$\text{E}(Y \mid X = x) = \sum_{y=55}^{73} y \times \text{Pr}(Y = y \mid X = x)$$
Variance Functions: III

- variance function $\text{Var}(Y \mid X = x)$ for mothers/daughters
Variance Functions: IV

• variance function for mothers/daughters example is relatively constant

• returning to point of view that $n = 1,375$ cases are a sample from a well-defined population, might be willing to assume that unknown variance function is the same for all $x$:

$$\text{Var}(Y \mid X = x) = \sigma^2,$$

where $\sigma^2$ is an unknown constant

• above is frequently assumed in linear regression

• in theory, $\sigma^2$ is nonnegative (i.e., $\sigma^2 \geq 0$), but will usually take it to be positive (i.e., $\sigma^2 > 0$) to rule out special case $\sigma^2 = 0$ (of little practical interest)