

Regression - Modeling a response

- We often wish to construct a *model* to
 - Explain the association between two or more variables
 - Predict the outcome of a variable given values of other variables.
- Regression is a technique that allows us to do so.
- A regression model has three components:
 - An **outcome** variable: y
 - **Predictor** variables or **covariates**: x_1, x_2, \dots, x_k
 - A **random error**: ϵ .

Regression - Salary example

Data on 20 mid-level managers in insurance industry in 2003: number of employees supervised and annual salary in \$1,000.

Employees	28	31	38	38	43	47	48
Salary	99.4	102.4	136.3	127.3	157.5	121.5	173.4

Employees	49	53	56	56	56	60	60
Salary	197.8	200.5	176.9	219.9	223.4	219.0	224.4

Employees	60	60	70	72	78	81
Salary	247.2	237.5	207.3	214.9	242.9	262.5

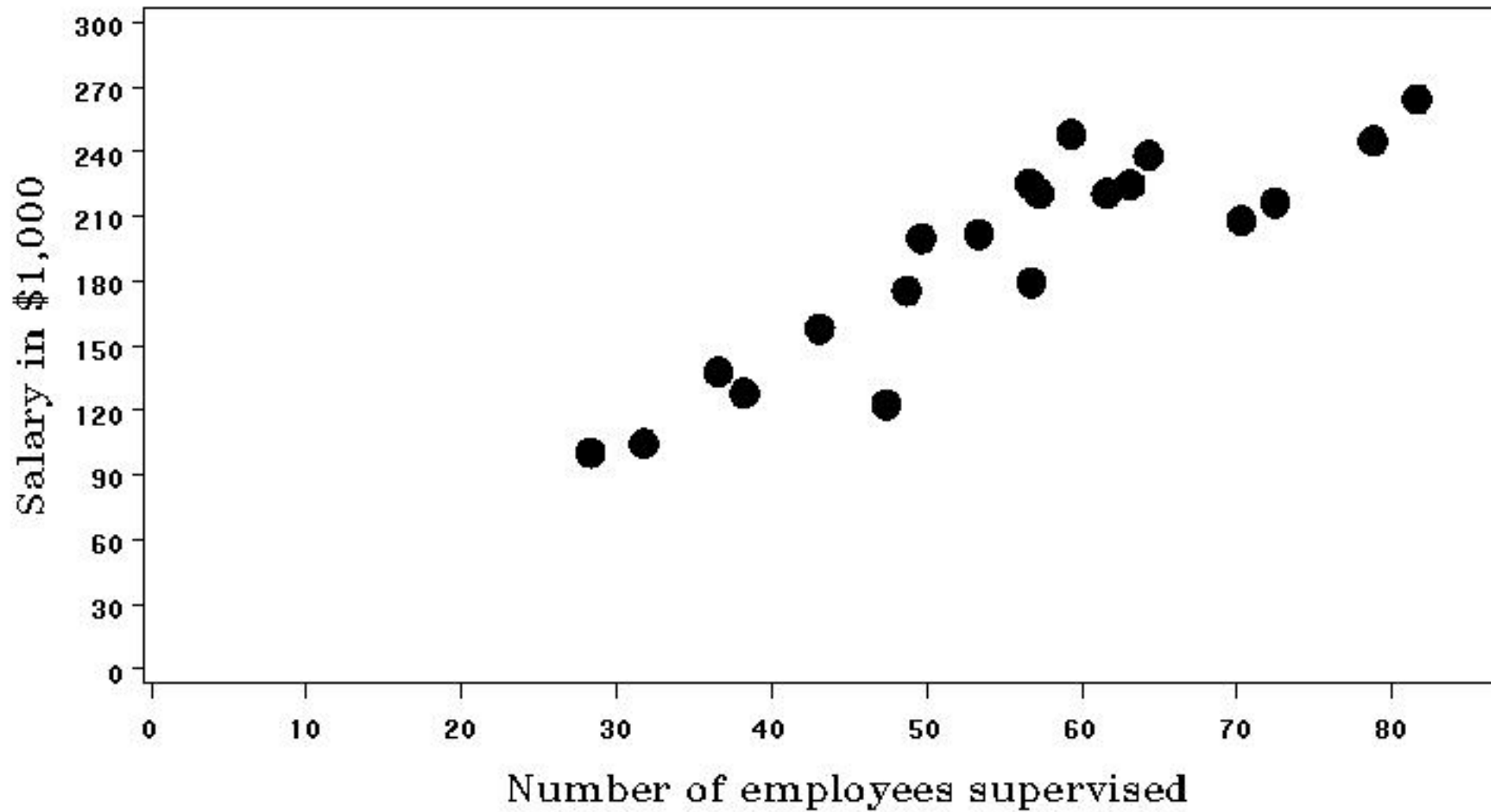


Figure 1: Data on mid-level executives

Example (cont'd)

- Data on graph are salaries (in \$1,000) and number of employees supervised by 20 mid-level managers in the insurance industry in 2003.
- Questions of interest:
 - Is annual compensation associated to personnel supervised?
 - If a manager's supervision responsibilities increase by 30 employees, can she predict what her new salary might be?

Example (cont'd)

- A few things to note:
 - Salary appears to increase with number of employees supervised. There is a *positive* association between salary and supervision responsibilities.
 - A straight line through the points provides a good summary of the relationship between salary and number of supervisees.
 - Two managers with the same number of supervisees do not necessarily make exactly the same salary, so factors other than number of supervisees also determine salary.

Regression - Formal introduction

- Every regression model has the following form:

$$y = E(y) + \epsilon$$

- $E(y)$ is the systematic part of the model which establishes the relationship between y and the predictors x_1, x_2, \dots
- ϵ is a random deviation or random error. Note:

$$\epsilon = y - E(y).$$

Regression - Formal intro (cont'd)

- In a *linear regression* model, $E(y)$ is a linear function of the predictors.
- In a *simple linear regression* model there is one predictor x . The regression relation is just a straight line of the form:

$$E(y) = \beta_0 + \beta_1 x,$$

so that

$$\begin{aligned} y &= E(y) + \epsilon \\ &= \beta_0 + \beta_1 x + \epsilon. \end{aligned}$$

- Back to salary example.

Regression - Salary example

- In example, managers supervising less than 40 employees earn between \$99,430 and \$136,325. Those who supervise more than 70 earn \$207,306 to \$262,515.
- Three managers supervise 56 employees and make \$176,990, \$219,895 and \$223,390.
- Salary appears to depend on number of employees supervised: expected salary increases with number of supervisees.
- A manager with no employees to supervise still earns some money.
- All of this can be summarized as:

$$E(\text{Salary}) = \beta_0 + \beta_1 \text{ employees}$$

Regression - Definitions

- From earlier page:

$$E(\text{Salary}) = \beta_0 + \beta_1 \text{ employees}$$

- β_0 is the **intercept** of the regression equation: the salary (in \$1,000) that can be expected by a manager with no employees to supervise.
- β_1 is the *regression coefficient*: it is the expected increase in salary (in \$1,000) when the number of employees increases by 1.
- If, for example, $\beta_0 = 25$ and $\beta_1 = 3$ then:
 - A manager with no supervisees can expect to earn $25 + 3 \times 0 = \$25$.
 - A manager with 30 supervisees can expect to earn $25 + 3 \times 30 = \$115$.
 - A manager with 56 supervisees can expect to earn $25 + 3 \times 56 = \$193$.

Regression - Definitions (cont'd)

- A positive β_1 indicates a positive association between y and x : as x increases, so does y .
- A negative β_1 indicates a negative association between y and x : as x increases, y decreases. An example might be median household salary and proportion of household income spent on food.
- A β_1 close to zero indicates no association between y and x : y might increase or decrease when x increases. An example might be divorce rate and oil prices.
- See graphical examples of each of the three possibilities.
- Errors ϵ are the deviations of the observations from the regression line (more soon).

Regression - Definitions (cont'd)

- *Simple* linear regression: outcome y is linear function of one predictor x . For i th unit:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

- *Multiple* linear regression: outcome y is linear function of two or more predictors x_1, x_2, \dots, x_k :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i.$$

Regression - Definitions (cont'd)

- In model with k predictors, there are $k + 1$ coefficients: $\beta_0, \beta_1, \dots, \beta_k$.
- Intercept β_0 is the expected value of y when all predictors x_1, x_2, \dots, x_k take on the value zero.
- Meaning of the multiple regression coefficients later in the course.
- See Figure 2.3 in textbook for example of regression with two predictors.

Regression - The random error

- The regression line (in simple linear regression) or the regression surface (in multiple linear regression) indicates the expected or average association between the outcome and the predictor(s).
- Actual, individual observations deviate up or down from the regression line.
- Deviations of observations from the regression relationship occur because the model is not perfect: there is variability in the y that is not explainable by the predictor(s).
- In the case of the salary example, we saw that several managers who supervise the same number of employees have different salaries. This is because number of supervisees is but one factor associated to salary.

Regression - Random errors (cont'd)

- We include all (known) explanatory factors in the model. Unexplainable effects on the outcome variable that induce variability of the y around the regression relation will remain.
- Errors are defined as deviations: $\epsilon_i = y_i - E(y_i)$, after accounting for all known effects on y .
- Typically, errors are assumed to be normal: $\epsilon_i \sim N(0, \sigma^2)$.
- The larger σ^2 , the larger the deviations of the observations from their expectation.
- On the graph, the ϵ_i are the *vertical* distances of observations from the regression line (or surface).

Regression - Population vs sample

- As in Chapter 1, we talk about a *true* regression relation and an *estimated* regression relation.
- The true regression relation is knowable only if we can obtain values of y and the predictors for each unit in the entire population.
- The estimated regression relation is what we obtain from a sample: for each unit in our sample, measure y and predictors, and from those measurements obtain *sample estimates* of all parameters in the model.

Regression - Population vs sample

- We use b_0, b_1, \dots, b_k to denote the sample estimates of the population quantities $\beta_0, \beta_1, \dots, \beta_k$.
- We use $\hat{\sigma}^2$ or MSE (Mean Squared Error) to denote the sample estimate of the population parameter σ^2 , the variance of the errors.
- We will see later that as in the case of \bar{y} , the sample estimates of intercept and regression coefficients are random and have their own sampling distributions.
- As before, randomness arises from the fact that the sample from which we obtain sample estimates is generated by a random sampling process.

Regression - Steps to fit model

- We fit a regression model when we wish to determine whether there is an association between a response variable and one or more predictors.
- Steps are:
 1. Hypothesize a relationship between y and one or more predictors x_1, x_2, \dots
 2. Draw a random sample of size n from the relevant population.
 3. Collect data on a random sample of units $(y_i, x_{i1}, x_{i2}, \dots, x_{ik})$: one outcome variable and k predictor variables on the i th unit, with $i = 1, \dots, n$ in a sample of units of size n .
 4. 'Fit' the model: obtain sample estimates for the population parameters $\beta_0, \beta_1, \dots, \beta_k, \sigma^2$.
 5. Check the fit of model and if necessary, reformulate.
 6. Test hypotheses and draw conclusions.

Regression - Collecting data

- Two types of studies for obtaining sample of size n :
 1. Experimental: the values of x_1, x_2, \dots are fixed in advance according to some experimental design. Example: x_1 is amount of fertilizer, x_2 is amount of irrigation and y is yield of corn. Experiment consists of applying pre-determined amounts of water and fertilizer to different plots and measuring yields.
 2. Observational: values of predictors are *not* set in advance. Example: sample 20 mid-level managers in insurance industry to determine association between number of supervisees and salary.
- In business, economics, social sciences, observational studies more common than experimental studies.

Regression - Observational vs experimental studies

- The mechanics of fitting a regression model are the same in both types of studies.
- Inferences that can be drawn from results are more limited in observational studies. In particular, we **cannot** establish cause-effect relationships, only associations.
- How large a sample? As in Chapter 1, sample size will depend on desired *reliability* of estimates b_0, b_1, \dots, b_k .
- Determining sample size precisely in observational studies is not straight forward.
- Rule of thumb: $n = 10 \times k$ (10 times as many observations as there are regression coefficients to estimate).