Chapter 7:

Correlation and Simple Linear Regression

Statistics for Engineers and Scientists

Fourth Edition

William Navidi

Introduction

- Scientists and engineers often collect data in order to determine the nature of the relationship between two quantities.
- For example, a chemical engineer may run a chemical process several times in order to study the relationship between the concentration of a certain catalyst and the yield of the process.
- Each time the process is run, the concentration x and the yield y are recorded.

Introduction

- In many cases, the ordered pairs of measurements fall approximately along a straight line when plotted.
- In those situations, the data can be used to compute an equation for the line that “best fits” the data.
- This line can be used for various things. In our catalyst versus yield experiment, the line could be used to predict the yield that will be obtained the next time the process is run with a specific catalyst concentration.

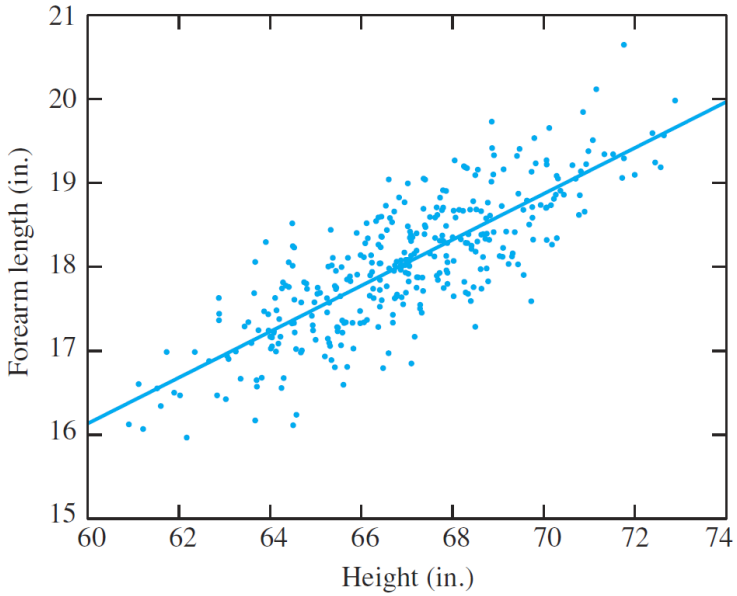
Section 7.1: Correlation

- One of the earliest applications of statistics was to study the variation in physical characteristics in human populations (e.g., forearm length versus height).
- The quantity called the **correlation coefficient** is a way of describing how closely related two variables are.

Section 7.1: Correlation

- We look at the direction of the relationship, positive or negative, the strength of the relationship, and then we find a line that best fits the data.
- In computing correlation, we can only use quantitative data.

Example



- This is a plot of height vs forearm length for men.
- We say that there is a positive association between height and forearm length. This is because the plot indicates that taller men tend to have longer forearms.
- The slope is roughly constant throughout the plot, indicating that the points are clustered around a straight line.
- The line superimposed on the plot is a special line known as the least-squares line.

Correlation Coefficient

- The degree to which the points in a scatterplot tend to cluster around a line reflects the strength of the linear relationship between x and y.
- The **correlation coefficient** is a numerical measure of the strength of the linear relationship between two quantitative variables.
- The correlation coefficient is usually denoted by the letter r.

Computing *r*

- Let $(x_1, y_1),\ldots,(x_n, y_n)$ represent n points on a scatterplot.
- Compute the means and the standard deviations of the x's and y's.
- Then convert each x and y to standard units. That is, compute the z-scores:

$$(x_i - \bar{x})/s_x \qquad (y_i - \bar{y})/s_y$$

Computing *r*

- The correlation coefficient is the average of the products of the *z*-scores, except that we divide by *n* - 1 instead of *n*.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Computational Formula

- This formula is easier for calculations by hand:

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

Comments

- In principle, the correlation coefficient can be calculated for any set of points.
- In many cases, the points constitute a random sample from a population of points.
- In this case, the correlation coefficient is called the **sample correlation**, and it is an estimate of the population correlation.

Properties of *r*

- It is a fact that *r* is always between –1 and 1.
- Positive values of *r* indicate that the least squares line has a positive slope. The greater values of one variable are associated with greater values of the other.
- Negative values of *r* indicate that the least squares line has a negative slope. The greater values of one variable are associated with lesser values of the other.

More Comments

- Values of r close to –1 or 1 indicate a strong linear relationship.
- Values of r close to 0 indicate a weak linear relationship.
- When r is equal to –1 or 1, then all the points on the scatterplot lie exactly on a straight line.

More Comments

- If the points lie exactly on a horizontal or vertical line, then r is undefined.

- If $r \neq 0$, then x and y are said to be **correlated**. If $r = 0$, then x and y are **uncorrelated**.

- For the scatterplot of height vs. forearm length, $r = 0.80$.

More Properties of *r*

- An important feature of *r* is that it is unitless. It is a pure number that can be compared between different samples.

- *r* remains unchanged under each of the following operations:
  - Multiplying each value of a variable by a positive constant.
  - Adding a constant to each value of a variable.
  - Interchanging the values of *x* and *y*.

- If $r = 0$, this does not imply that there is not a relationship between *x* and *y*. It just indicates that there is no *linear* relationship.

More Properties of *r*

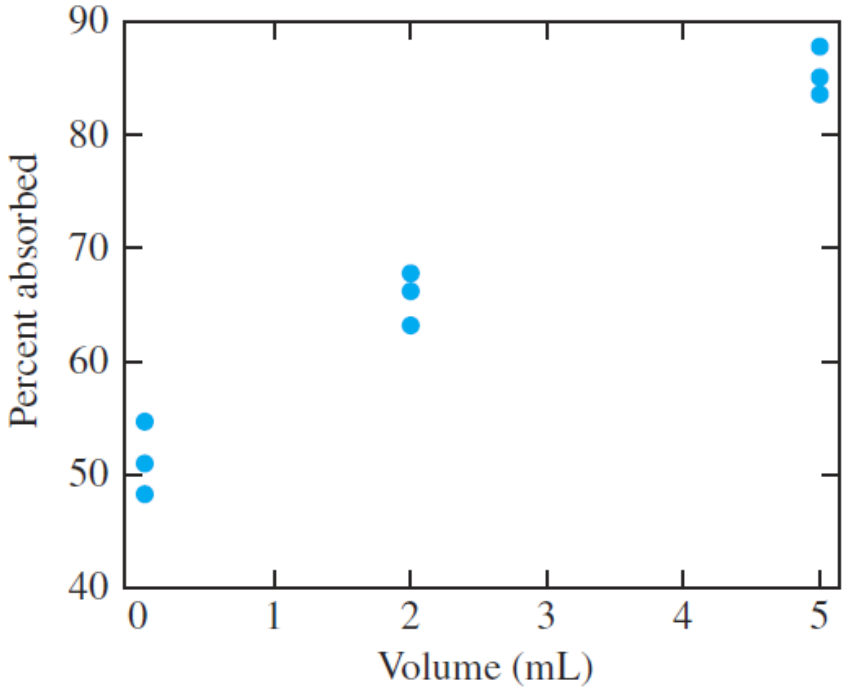
- Outliers can greatly distort *r*, especially, in small data sets, and present a serious problem for data analysts.

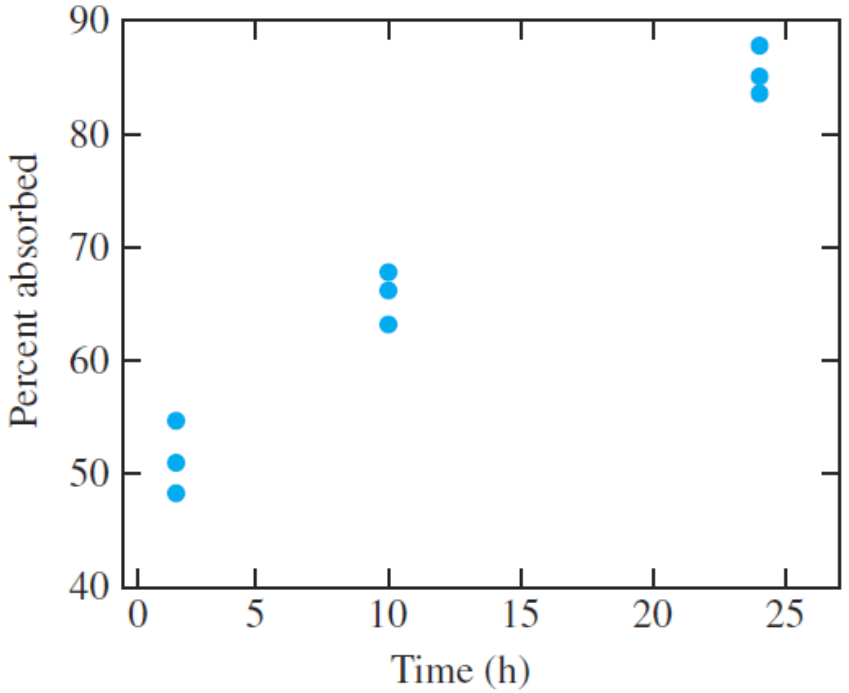
- Correlation is not causation. For example, vocabulary size is strongly correlated with shoe size, but this is because both increase with age. Learning more words does not cause feet to grow or vice versus. Age is **confounding** the results.

Example 1

An environmental scientist is studying the rate of absorption of a certain chemical into skin. She places differing volumes of the chemical on different pieces of skin and allows the skin to remain in contact with the chemical for varying lengths of time. She then measures the volume of chemical absorbed into each piece of skin. The scientist plots the percent absorbed against both volume and time. She calculates the correlation between volume and absorption and obtains $r = 0.988$. She concludes that increasing the volume of the chemical causes the percentage absorbed to increase. She then calculates the correlation between time and absorption, obtaining $r = 0.987$. She concludes that increasing the time that the skin is in contact with the chemical causes the percentage absorbed to increase as well. Are these conclusions justified?

Example 1





Inference on the Population Correlation

- If the random variables X and Y have a certain joint distribution called a **bivariate normal distribution**, then the sample correlation r can be used to construct confidence intervals and perform hypothesis tests on the population correlation, ρ.
- The following results make this possible.

Distribution

Let $(x_1, y_1),\ldots, (x_n, y_n)$ be a random sample from the joint distribution of X and Y and let r be the sample correlation of the n points. Then the quantity

$$W = \frac{1}{2}\ln\frac{1+r}{1-r}$$

is approximately normal with mean

$$\mu_W = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}$$

and variance

$$\sigma_W^2 = \frac{1}{n-3}.$$

Example 2

In a study of reaction times, the time to respond to a visual stimulus (x) and the time to respond to an auditory stimulus (y) were recorded for each of 10 subjects. Times were measured in ms.

x	161	203	235	176	201	188	228	211	191	178
y	159	206	241	163	197	193	209	189	169	201

Find a 95% confidence interval for the correlation between the two reaction times.

Example 2 cont.

Find the *P*-value for testing H_0: $\rho \leq 0.3$ versus H_1: $\rho > 0.3$.

Test the hypothesis H_0: $\rho \leq 0$ versus H_1: $\rho > 0$.

The Least-Squares Line

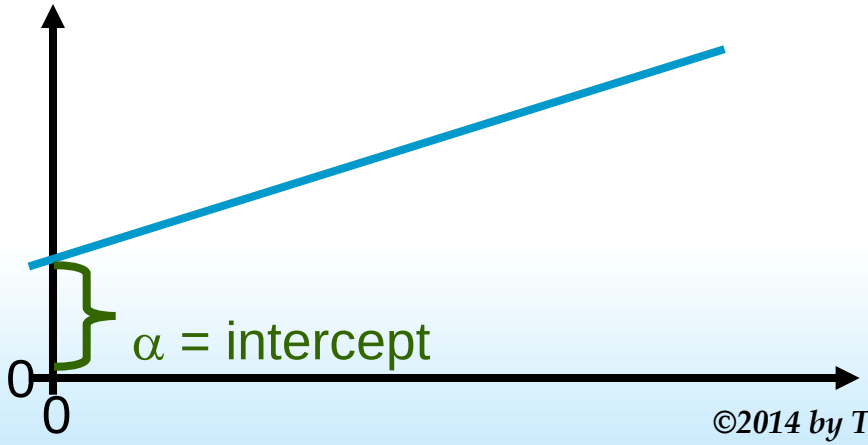
- When two variables have a linear relationship, the scatterplot tends to be clustered around a line known as the least-squares line

Linear Deterministic Model

- Denote x as **independent variables** and y as **dependent variable**. For example, x=age and y=%fat in the first example; x=advertising expenditure, y=sales in the second example

- We want to find how y depends on x, or how to predict y using x

- One of the simplest deterministic mathematical relationship between two variables x and y is a linear relationship $y = \alpha + \beta x$



α: intercept

β: slope

Reality

- In the real world, things are never so clean!
  - Age influences fatness. But it is not the sole influence. There are other factors such as gender, body type and random variation (e.g. measurement error)
  - Other factors such as time of year, state of economy and size of inventory, besides the advertising expenditure, can influence the sale

- Observations of (x, y) do not fall on a straight line

As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality. **Albert Einstein**

Probabilistic Model

- Probabilistic model:
  y = deterministic model + random error
- Random error represents random fluctuation from the deterministic model
- The probabilistic model is assumed for the population
- **Simple linear regression model**:
  $y = \alpha + \beta x + \varepsilon$
- Without the random deviation ε, all observed points (x, y) points would fall exactly on the deterministic line. The inclusion of ε in the model equation allows points to deviate from the line by random amounts.

Section 7.2: The Least-Squares Line

- The line that we are trying to fit is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.
- The variable y_i is the dependent variable, the x_i is the independent variable, and β_0 and β_1 are called the regression coefficients, and ε_i *i*s called the error. We only know the values of x and y, we must estimate β_0 and β_1.
- This is what we call **simple linear regression**.
- We use the data to estimate these quantities.

Using the Data

- We write the equation of the least-square line as

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The quantities $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the **least-squares coefficients**.

- The least-squares line is the line that fits the data "best."

Residuals

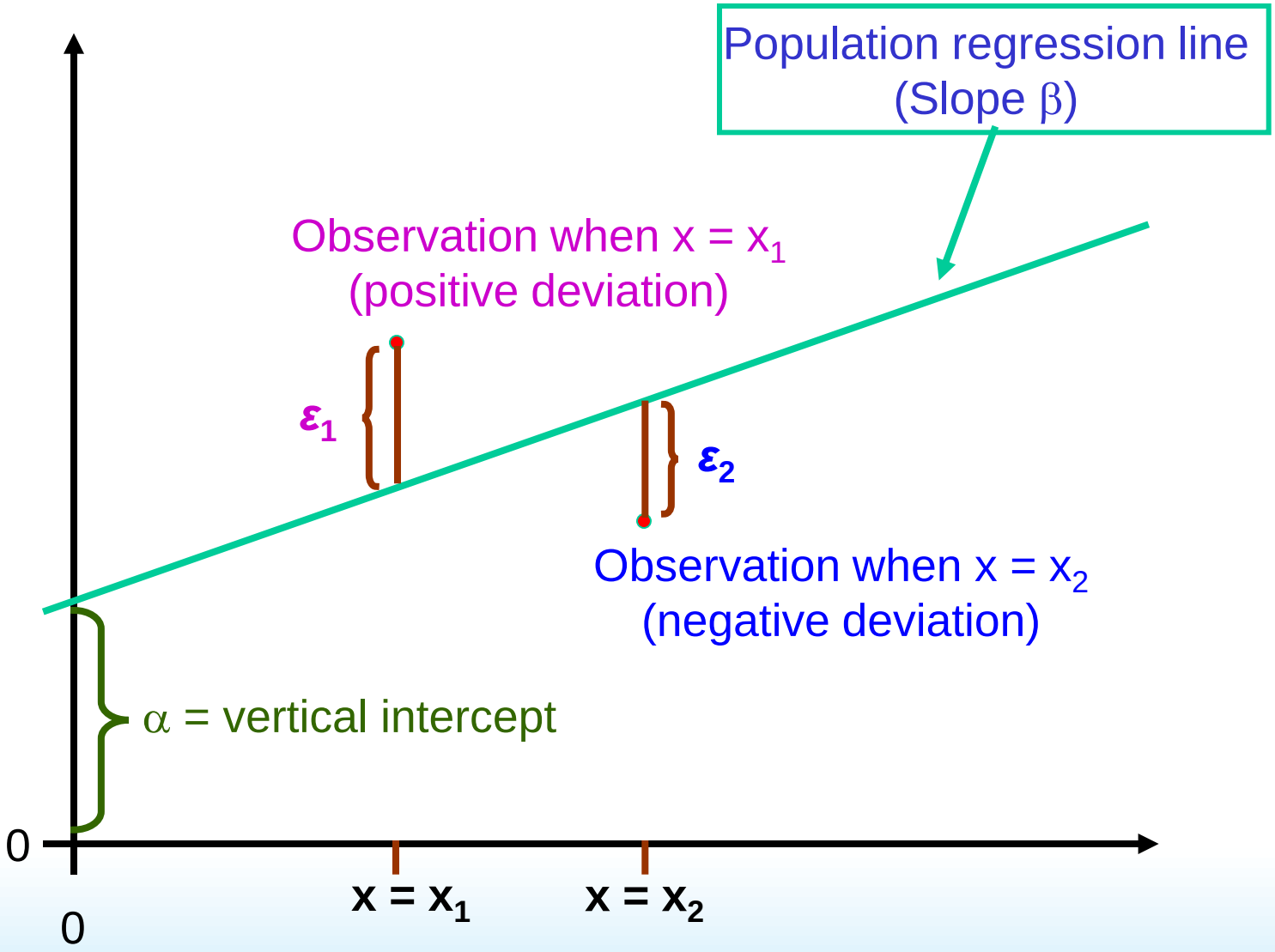
- For each data point, (x_i, y_i) the vertical distance to the point $(x_i, \hat{y}_i)$ on the least squares line is $e_i = y_i - \hat{y}_i$. The quantity $\hat{y}_i$ is called the fitted value and the quantity e_i is called the residual associated with the point (x_i, y_i).

- The residual e_i is the difference between the value y_i observed in the data and the fitted value predicted by the least-squares line.

Residuals

- Points above the least-squares line have positive residuals, and points below the line have negative residuals.
- The closer the residuals are to 0, the closer the fitted values are to the observations and the better the line fits the data.
- The least-squares line is the one that minimizes the sum of squared residuals.

Simple Linear Regression Model



Finding the Equation of the Line

- To find the least-squares line, we must determine estimates for the slope β_0 and β_1 intercept that minimize the sum of the squared residuals.

- These quantities are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \ .$$

Note: The true values of β_0 and β_1 are unknown.

Some Shortcut Formulas

The expressions on the right are equivalent to those on the left, and are often easier to compute:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-n\bar{x}$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-n\bar{y}$$

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

Example 3

Using the data in the table below, compute the least-squares estimates of the spring constant and the unloaded length of the spring. Write the equation of the least-squares line. Estimate the length of the spring under a load of 1.3 lb.

Weight (lb) x	Measured Length (in.) y	Weight (lb) x	Measured Length (in.) y
0.0	5.06	2.0	5.40
0.2	5.01	2.2	5.57
0.4	5.12	2.4	5.47
0.6	5.13	2.6	5.53
0.8	5.14	2.8	5.61
1.0	5.16	3.0	5.59
1.2	5.25	3.2	5.61
1.4	5.19	3.4	5.75
1.6	5.24	3.6	5.68
1.8	5.46	3.8	5.80

Cautions

- Do not extrapolate the fitted line (such as the least-squares line) outside the range of the data. The linear relationship may not hold there.

- We learned that we should not use the correlation coefficient when the relationship between x and y is not linear. The same holds for the least-squares line. When the scatterplot follows a curved pattern, it does not make sense to summarize it with a straight line.

- If the relationship is curved, then we would want to fit a regression line that contain squared terms.

Another Representation of the Line

- Another way to compute an estimate of β_1 is

$$\hat{\beta}_1 = r\frac{s_y}{s_x}$$

- The units of $\hat{\beta}_1$ are the same as the units of y/x.

- The slope is proportional to the correlation coefficient.

Another Representation of the Line

- The least-squares line can be rewritten as

$$\hat{y} - \bar{y} = r\frac{s_x}{s_y}(x - \bar{x})$$

so the line passes through the center of the mass of the scatterplot with slope $r(s_y/s_x)$.

Measures of Goodness of Fit

- A goodness-of-fit statistic is a quantity that measures how well a model explains a given set of data.
- The quantity r^2 is the square of the correlation coefficient and we call it the **coefficient of determination**.

$$r^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- The **proportion of variance in *y* explained by regression** is the interpretation of r^2.

Sums of Squares

- $\sum_{i=1}^{n}(y_i - \hat{y})^2$ is the error sums of squares and measures the overall spread of the points around the least-squares line.

- $\sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sums of squares and measures the overall spread of the points around the line $y = \bar{y}$.

- The difference $\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y})^2$ is called the regression sum of squares and measures the reduction in the spread of points obtained by using the least-squares line rather than $y = \bar{y}$.

Sums of Squares

- The coefficient of determination r^2 expresses the reduction as a proportion of the spread around $y = \bar{y}$.

- Clearly, the following relationship holds:

  Total sum of squares = regression sum of squares + error sum of squares

Section 7.3: Uncertainties in the Least-Squares Coefficients

- the errors εi create *uncertainty in the estimates* $\beta 0$ *and* $\beta 1$.
- *It is intuitively clear that if the* εi *tend* to be small in magnitude, the points will be tightly clustered around the line, and the uncertainty in the least-squares estimates $\beta 0$ *and* $\beta 1$ *will be small*.
- *On the other hand, if* the εi *tend to be large in magnitude, the points will be widely scattered around the line*,and the uncertainties (standard deviations) in the least-squares estimates $\beta 0$ *and* $\beta 1$ *will* be larger.

- Assume we have *n data points (x1, y1), . . . , (xn, yn), and we plan to fit the leastsquares* line.
- In order for the estimates *β1 and β0 to be useful, we need to estimate just* how large their uncertainties are. In order to do this, we need to know something about
- the nature of the errors εi .
- *We will begin by studying the simplest situation, in which* four important assumptions are satisfied.

Section 7.3: Uncertainties in the Least-Squares Coefficients

Assumptions for Errors in Linear Models:

In the simplest situation, the following assumptions are satisfied:

1. The errors $\varepsilon_1,\ldots,\varepsilon_n$ are random and independent. In particular, the magnitude of any error ε_i does not influence the value of the next error ε_{i+1}.
2. The errors $\varepsilon_1,\ldots,\varepsilon_n$ all have mean 0.
3. The errors $\varepsilon_1,\ldots,\varepsilon_n$ all have the same variance, which we denote by σ^2.
4. The errors $\varepsilon_1,\ldots,\varepsilon_n$ are normally distributed.

- When the sample size is large, the normality assumption (4) becomes less important.
- Mild violations of the assumption of
- constant variance (3) do not matter too much, but severe violations should be corrected.

- Under these assumptions, the effect of the εi *is largely governed by the magnitude* of the variance $\sigma 2$, *since it is this variance that determines how large the errors are* likely to be.
- Therefore, in order to estimate the uncertainties in $\beta 0$ *and* $\beta 1$, *we must first* estimate the error variance $\sigma 2$.
- *Since the magnitude of the variance is reflected in the*

degree of spread of the points around the least-squares line, it follows that by measuring this spread, we can estimate the variance.

Specifically, the vertical distance from each data point *(xi , yi) to the least-squares line is given by the residual ei*

- The spread of the points around the line can be measured by the sum of the squared residuals
- The estimate of the error variance *σ2 is the quantity* *s2* given by

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

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

Distribution

In the linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, under assumptions 1 through 4, the observations $y_1, \ldots, y_n$ are independent random variables that follow the normal distribution. The mean and variance of y_i are given by

$$\mu_{y_i} = \beta_0 + \beta_1 x_i$$

$$\sigma^2_{y_i} = \sigma^2$$

The slope represents the change in the mean of y associated with an increase in one unit in the value of x.

More Distributions

Under assumptions 1 – 4:

- The quantities $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed random variables.

- The means of $\hat{\beta}_0$ and $\hat{\beta}_1$ are the true values β_0 and β_1, respectively.

More Distributions (cont.)

- The standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated with

$$s_{\hat{\beta}_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \quad \text{and} \quad s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where $s = \sqrt{\frac{(1-r)^2\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-2}}$ is an estimate of the error standard deviation σ.

Example 2 cont.

Using the data in Example 2, calculate s, $s_{\hat{\beta}_1}$, $s_{\hat{\beta}_0}$.

Notes

1. Since the quantity $\sum_{i=1}^{n}(x_i - \bar{x})^2$ appears in the denominators of $s_{\hat{\beta}_0}$ and $s_{\hat{\beta}_1}$, it follows that the more spread out the x's are, the smaller the uncertainties in $\hat{\beta}_0$ and $\hat{\beta}_1$ will be.
2. Use caution: if the range of *x* values extends beyond the range where the linear model holds, the results will not be valid.
3. The quantities $\left(\hat{\beta}_0 - \beta_0\right)/s_{\hat{\beta}_0}$ and $\left(\hat{\beta}_1 - \beta_1\right)/s_{\hat{\beta}_1}$ have Student's *t* distribution with $n - 2$ degrees of freedom.

Section 7.4: Checking Assumptions and Transforming Data

- We stated some assumptions for the errors. Here we want to see if any of those assumptions are violated.

- The single best diagnostic for least-squares regression is a plot of residuals versus the fitted values, sometimes called a **residual plot**.

More of the Residual Plot

- When the linear model is valid, and assumptions 1 – 4 are satisfied, the plot will show no substantial pattern. There should be no curve to the plot, and the vertical spread of the points should not vary too much over the horizontal range of the data.

- A good-looking residual plot does not by itself prove that the linear model is appropriate. However, a residual plot with a serious defect does clearly indicate that the linear model is inappropriate.

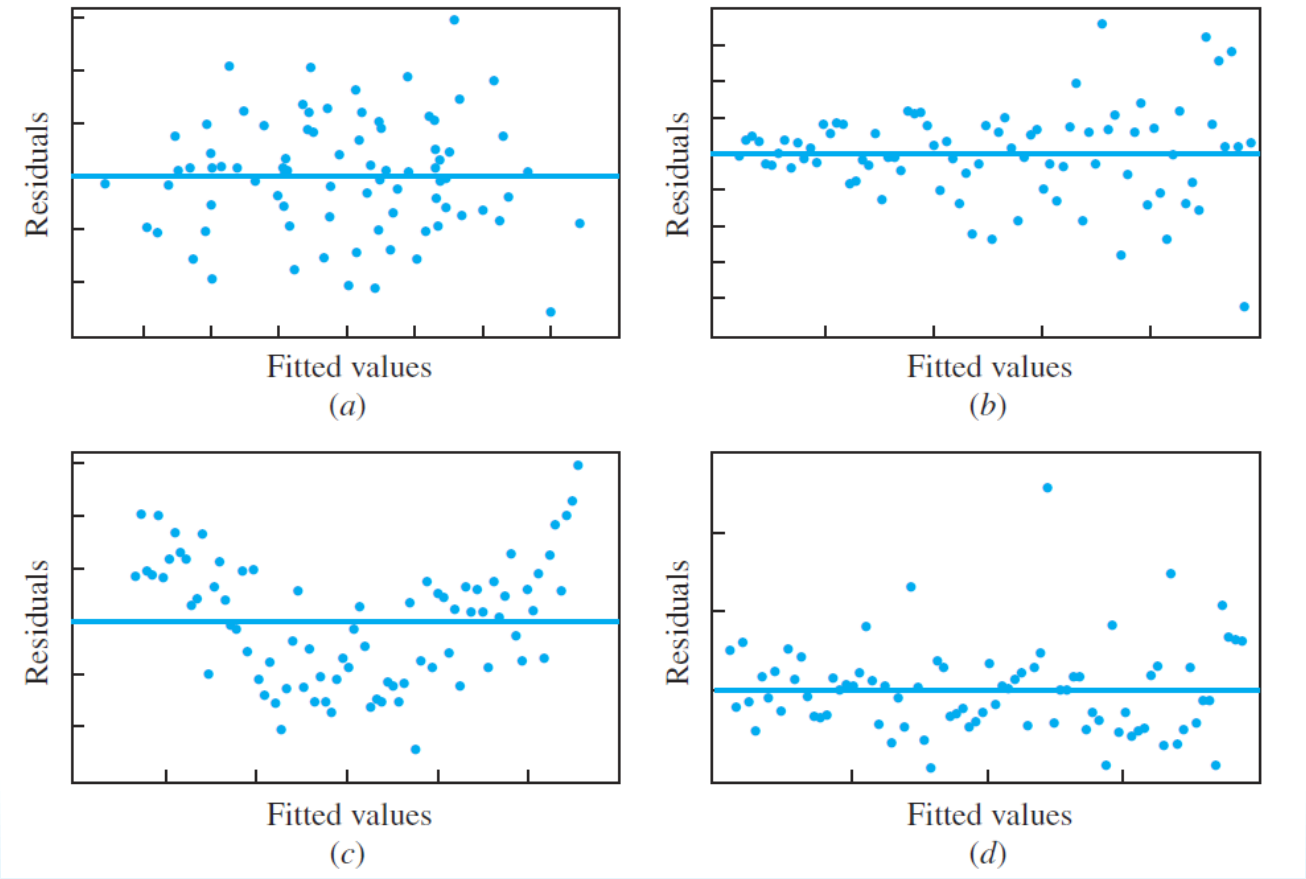
Residual Plots

A: No noticeable pattern

B: Heteroscedastic

C: Trend

D: Outlier



Residuals versus Fitted Values

If the plot of residuals versus fitted values

- Shows no substantial trend or curve, and
- Is **homoscedastic**, that is, the vertical spread does not vary too much along the horizontal length of plot, except perhaps near the edges.

then it is *likely*, but not *certain*, that the assumptions of the linear model hold.

However, if the residual plot *doe*s show a substantial trend or curve, or is **heteroscedastic**, it is certain that the assumptions of the linear model do *not* hold.

Transforming the Variables

- If we fit the linear model $y = \beta_0 + \beta_1 x + \varepsilon$ and find that the residual plot exhibits a trend or pattern, we can sometimes fix the problem by raising x, y, or both to a power.

- It may be the case that a model of the form $y^a = \beta_0 + \beta_1 x^b + \varepsilon$ fits the data well.

- Replacing a variable with a function of itself is called **transforming** the variable. Specifically, raising a variable to a power is called a **power transformation**.

Don’t Forget

- Once the transformation has been completed, then you must inspect the residual plot again to see if that model is a good fit.
- It is fine to proceed through transformations by trial and error.
- It is important to remember that power transformations don’t always work.

Caution

- When there are only a few points in a residual plot, it can be hard to determine whether the assumptions of the linear model is met.

- When one is faced with a sparse residual plot that is hard to interpret, a reasonable thing to do is to fit a linear model, but to consider the results tentative, with the understanding that the appropriateness of the model has not been established.

Outliers

- **Outliers** are points that are detached from the bulk of the data.
- Both the scatter plot and the residual plot should be examined for outliers.
- The first thing to do with an outlier is to determine why it is different from the rest of the points.

Outliers

- Sometimes outliers are caused by data-recording errors or equipment malfunction. In this case, the outlier can be deleted from the data set. In this case, you may present results that do not include the outlier.

- If it cannot be determined why there is an outlier, then it is not wise to delete it. Here the results presented, should be the ones from analysis with the outlier included in the data set.

Influential Point

- If there are outliers that cannot be removed from the data set, then the best thing to do is fit the whole data set and then remove the outlier and fit a line to the data set.

- If none of the outliers upon removal make a noticeable difference to the least-squares line or to the estimated standard deviation of the slope and intercept, then use the fit with the outliers included.

Influential Point

- If one or more outlier does make a difference, then the range of values for the least-squares coefficients should be reported. Avoid computing confidence and prediction intervals and performing hypothesis tests.

- An outlier that makes a considerable difference to the least-squares line when removed is called an **influential point**.

Influential Point

- In general, outliers with unusual *x* values are more likely to be influential than those with unusual *y* values, but every outlier should be checked.

- Some authors restrict the definition of outliers to points that have unusually large residuals.

Comments

- Transforming the variables is not the only method for analyzing data when the residual plot indicates a problem.

- There is a technique called weighted least squares regression. The effect is to make the points whose error variance is smaller have greater influence in the computation of the least-squares line.

- When the residual plot shows a trend, this sometimes indicates that more than one independent variable is needed to explain the variation in the dependent variable.

Comments

- If the relationship is nonlinear, then a method called nonlinear regression can be applied.

- If the plot of residuals versus fitted values looks good, it may be advisable to perform additional diagnostics to further check the fit of the linear model. A time series plot is used to see if time should be included in the model. A normal probability plot can be used to check the normality assumption.

Independence of Observations

- If the plot of residuals versus fitted values looks good, then further diagnostics may be used to further check the fit of the linear model.

- A time order plot of the residuals versus order in which observations were made.

- If there are trends in this plot, then x and y may be varying with time. In this case, adding a time term to the model as an additional independent variable.

Normality Assumption

- To check that the errors are normally distributed, a normal probability plot of the residuals can be made.

- If the plot looks like it follows a rough straight line, then we can conclude that the residuals are approximately normally distributed.

Summary

We discussed

- correlation
- least-squares line / regression
- uncertainties in the least-squares coefficients
- confidence intervals and hypothesis tests for least-squares coefficients
- checking assumptions
- residuals