Chapter 12 Relationships Between Quantitative Variables: Regression and Correlation

We start with Chapter 3.1 on plotting and modeling the relationship between two QUANTITATIVE variables.

Main idea: We wish to study the relationship between two quantitative variables. Generally one variable is the **response variable**, denoted by y. The response variable measures the outcome of the study and is also referred to as the *dependent variable*. The other variable is the **explanatory variable**, denoted by x. It is the variable that is thought to explain the changes we see in the response variable. The explanatory variable is also referred to as the *independent variable*.

The first step in examining the relationship is to use a graph – a **scatterplot** – to display the relationship. We will look for an overall pattern and see if there are any departures from this overall pattern. If a **linear** relationship appears to be reasonable from the scatterplot, we will take the next step of finding a **model** (an equation of a line) to summarize the relationship. The resulting equation may be used for **predicting** the response for various values of the explanatory variable.

Example: How do children grow?
The pattern of growth varies from child to child, so we can best understand the general pattern by following the average height of a number of children. The following table presents the mean heights of 161 children in Kalama village aging 18 to 29 months.

Age (month)	Height (cm)
18	76.1
19	77.0
20	78.1
21	78.2
22	78.8
23	79.7
24	79.9
25	81.1
26	81.2
27	81.8
28	82.8
29	83.5

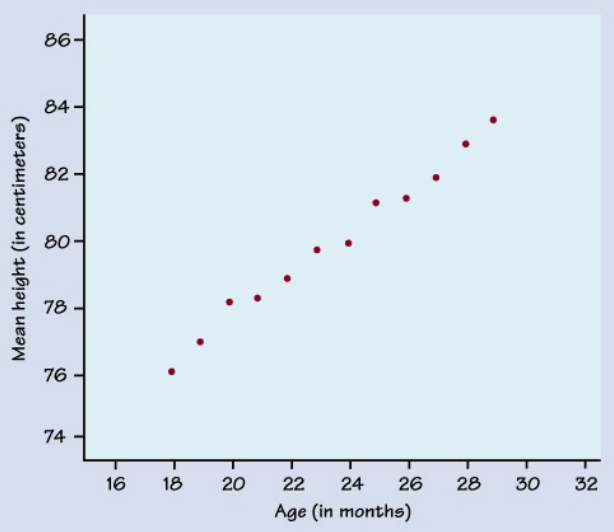
We want to know how children grow, i.e., how high a child would be (on average) if we know its age. In other words, we hope to see the relationship between age and height, and possibly develop a model to predict the height of a child based on its age.

Response (dependent) variable (y) is _____

Explanatory (independent) variable (x) is _____

Plotting the Data: Scatterplots

The first step is to examine the data graphically with a scatterplot.



Interpret the scatterplot by looking for

- overall pattern
  - **form** (clusters, lines, curves, etc.)
  - **direction** of association (positive or negative)
  - **strength** of association (weak, moderate, strong)
- deviations from the overall pattern (**outliers**)

What do you see?

Comments on scatterplots

- Each individual in the data appears as a point on the graph.
- Always plot the explanatory variable (x) on the x axis and the response variable on the y axis.
- Can use colors or symbols to see the effect of a categorical variable.

Section 12.5 Correlation and Association

We want to add a numerical measure to supplement the graph. We use *correlation*.

The sample correlation coefficient r (or just correlation r)

- measures the direction and strength of the linear relationship between two quantitative variables.

Notation

- Data on n individuals: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- (sample) means
$$\bar{x} = \frac{1}{n}\sum x_i \quad \text{and} \quad \bar{y} = \frac{1}{n}\sum y_i$$
- (sample) standard deviations
$$s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2} \quad \text{and} \quad s_y = \sqrt{\frac{1}{n-1}\sum (y_i - \bar{y})^2}$$

correlation coefficient
$$r = \frac{1}{n-1}\sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Example Kalama children growth, $n = 12$

First from the 12 observations of age

18 19 20 21 22 23 24 25 26 27 28 29

we find
$$\bar{x} = \frac{18 + 19 + \cdots + 29}{12} = 23.5$$
$$s_x = \sqrt{\frac{(18 - 23.5)^2 + (19 - 23.5)^2 + \cdots + (29 - 23.5)^2}{12 - 1}} = 3.6$$

Next from the 12 observations of height

76.1 77.0 78.1 78.2 78.8 79.7 79.9 81.1 81.2 81.8 82.8 83.5

we find
$$\bar{y} = \frac{76.1 + 77.0 + \cdots + 83.5}{12} = 79.85$$
$$s_y = \sqrt{\frac{(76.1 - 79.85)^2 + (77.0 - 79.85)^2 + \cdots + (83.5 - 79.85)^2}{12 - 1}} = 2.3$$

Finally, we compute r as

Properties of correlation

- $r > 0$ indicates _____ association and $r < 0$ indicates _____ association.
- r is always a number between -1 and 1, i.e., $-1 \leq r \leq 1$.
- r close to 0 indicates a _____ relationship.
- r close to 1 indicates a _____ linear relationship.
- r close to -1 indicates a _____ linear relationship.
- r measures the strength of only the linear relationship. It does not measure curved relationships.
- Like the mean and standard deviation, r is _____
- r has no unit. It does not change when we change the units of measures of x, y, or both.
- **Correlation does not necessarily imply causation.**
- Distinguish correlation ρ between two random variables.

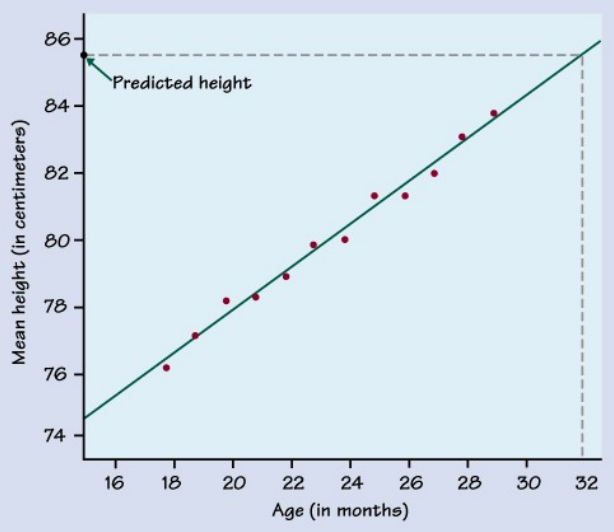
See Fig 12.5.2 (page 542): How the correlation r measures the strength and direction of linear association.

See Fig. 12.5.3 (page 543): What the correlation r does NOT tell?

Section 12.3 Choosing the Best Line

Recall: In the Kalama children growth example, we saw a strong and positive linear relationship between age and height. We want to find a **model** to summarize the relationship and to **predict** the growth of a child. The model is called a **regression line**, a straight line that describe how y changes as x changes.

Here is the scatterplot with a regression line for the Kalama data.



The regression line is

$$\text{height} = 64.93 + 0.635 \times \text{age}$$

What does the value 0.635 tell us?

How would you predict the height for a child of 32 month old?

Q: How to find the regression line?

Idea: To find a line that comes as close as possible to all points. We use the least-squares methods: make the sum of squares of the prediction errors as small as possible.

Least-squares regression

Recall **Notation**

- Data on n individuals: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- (sample) means $\bar{x}$ and $\bar{y}$
- (sample) standard deviations s_x and s_y
- correlation r

Equation of the Least-squares regression line of y on x is

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

with **slope**

$$\hat{\beta}_1 = r\frac{s_y}{s_x} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

and **intercept**

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

NOTE: we use $\hat{y}$ to denote a **predicted** (or **ideal**) response. The difference between observed response y and predicted response $\hat{y}$ is called prediction **error** or **residual**.

Example: Kalama Data

Recall

$$\bar{x} = 23.5, \quad \bar{y} = 79.85, \quad s_x = 3.6, \quad s_y = 2.3, \quad r = 0.99.$$

We can compute the slope $\hat{\beta}_1$ and intercept $\hat{\beta}_0$ as

Note: The sum of prediction errors

$$\sum(\text{error})^2 = \sum(y_i - a - bx_i)^2$$

is minimized if slope $b = \hat{\beta}_1 = r\frac{s_y}{s_x}$ and intercept $a = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$.

Interpreting the regression line:

- The **slope** $\hat{\beta}_1$ is the _____ at which the predicted response $\hat{y}$ changes along the line as x changes.
- If x increases by 1 unit, $\hat{y}$ would change by _____ units.
- The **intercept** $\hat{\beta}_0$ is the predicted $\hat{y}$ when $x = 0$. We are usually not interested in the intercept.

NOTES:

- The square of correlation, r^2, is the _____ of variation in the response that is explained by the least-squares regression.
- Prediction far outside the range of x (called **extrapolation**) is often _____

Section 12.4 Formal Inference for the Simple Linear Model

Recall: To study the relationship between two quantitative variables, we first use a scatterplot to examine the relationship. We look for an overall pattern and see if there are any departures from this overall pattern. If a **linear** relationship appears to be reasonable from the scatterplot, we take the next step of finding a model (a **regression line**) to summarize the relationship. Then we use the regression line for predicting the response for various values of the explanatory variable.

In this section we will assess the significance of the linear relationship and make some confidence intervals for our estimations and predictions under certain assumptions.

Example: The fat content of the human body has physiological and medical importance. The fat content is found by measuring body density. High fat content corresponds to low body density. Body density is hard to measure directly. Research suggests that skinfold thickness can accurately predict body density.

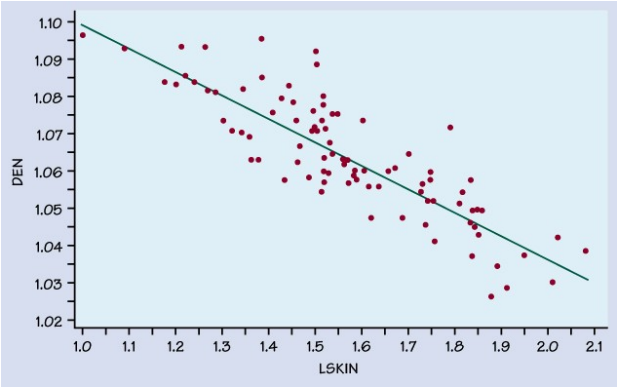
We wish to study the relationship between **body density** and **skinfold thickness**. We are interested in seeing if there is a relationship between the two variables and possibly developing a model to **predict** the body density based on the skinfold thickness.

Response (dependent) variable = y = _____

Explanatory (independent) variable = x = _____

Examine the data graphically with a scatterplot.

Here is the scatterplot for body density (DEN) and logarithm of the skinfold thickness (LSKIN).



Describe the scatterplot:

Does a linear relationship between DEN and LSKIN seem plausible based on the scatterplot?

Note: The correlation $r = -0.8488$.

Here is the summary from Stata.

Variable	Obs	Mean	Std. Dev.	Min	Max
lskin	92	1.567935	.2159596	1	2.08
density	92	1.064033	.0160599	1.026	1.096

Find the regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

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

intercept $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} =$

Predict the body density for LSKIN=1.0.

Note: We observed a value of $y =$ _____ when $x = 1.0$ (see scatterplot).

The difference between the observed response and predicted response is the _____

The Simple Linear Model

The statistical model for simple linear regression assumes that for each value of x, the observed values of the response y are normally distributed with some mean (that may depend on x in a linear way) and a standard deviation σ that does not depend on x. These assumptions can be summarized as:

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

where $\beta_0 + \beta_1 x_i$ is the mean response when $x = x_i$. The deviation ϵ_i are assumed to be independent and have $N(0, \sigma)$ distribution.

The parameters of the model are β_0, β_1, and σ.

Our goals will be:

- Estimate the unknown parameters $(\beta_0, \beta_1, \sigma)$ based on some data.
- Assess if the linear relationship is statistically significant.
- Provide confidence intervals for our estimates/predictions.
- Understand and check the assumptions of our model.

We know how to estimate β_0 and β_1.

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

estimated intercept $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

and the least squares regression line (or estimated regression function) is

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

ESTIMATE OF σ^2

We need to estimate the common population variance σ^2. A variance is computed as a sum of the squared distances and divided by the degrees of freedom. In regression, the residuals are the distance of the observations from their estimated mean. So we will compute the sum of the squared residuals and divide by an appropriate degrees of freedom.

Estimate of $\sigma^2 = s^2 = \frac{\sum (e_i)^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$ = Mean Square Error (or Residuals) = MSE

The estimate of σ is given by

$$s = \sqrt{s^2} = \sqrt{\text{MSE}},$$

The **degrees of freedom** of s is _____

In practice, we use software to do the computation.

Here is the **Stata command**

```
regress density lskin
```

Stata output

```
      Source |       SS       df       MS              Number of obs =      92
-------------+------------------------------           F(  1,    90) =  231.89
       Model |  .016908557     1  .016908557           Prob > F      =  0.0000
    Residual |   .00656234    90  .000072915           R-squared     =  0.7204
-------------+------------------------------           Adj R-squared =  0.7173
       Total |  .023470897    91  .000257922           Root MSE      =  .00854

------------------------------------------------------------------------------
     density |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lskin |   -.063119   .0041449   -15.23   0.000    -.0713536   -.0548844
       _cons |   1.162999   .0065596   177.30   0.000     1.149967    1.176031
------------------------------------------------------------------------------
```

(a) Estimated slope $\hat{\beta}_1$=

(b) Estimated intercept $\hat{\beta}_0$=

(c) Estimated regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$:

(d) Estimated standard deviation s=

With our variance estimate in hand, we can now turn to some inference procedures. We first assess if the linear relationship is statistically significant.

SIGNIFICANT LINEAR RELATIONSHIP?

Consider the following hypotheses:

$H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$

What happens if the null hypothesis is true?

There are a number of ways to test this hypothesis. One way is through a t-test statistic (think about why it is a t and not a z test).

The general form for a t test statistic is:

$$t = \frac{\text{point estimate} - \text{hypothesized value}}{\text{standard error of the point estimate}}$$

We have our point estimate for β_1, it is $\hat{\beta}_1$. And we have the hypothesized value of 0. So we need the standard error for $\hat{\beta}_1$, which can be read from the Stata output.

The standard error for $\hat{\beta}_1 = \text{se}(\hat{\beta}_1) =$

Is there a significant (non-zero) linear relationship between DEN and LSKIN? That is, test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.

Observed t statistic $= \frac{\hat{\beta}_1}{\text{se}_{\hat{\beta}_1}} =$

Degrees of freedom $= n - 2 =$

p-value $=$

Decision: (circle one) Reject H_0 or Accept H_0 at a 5% level.

Conclusion: (circle one) There IS or IS NOT a significant linear relationship between DEN and LSKIN.

Think about it: Would a 95% confidence interval for the true slope β_1 contain the value of 0?

Find the 95% CI for true slope β_1:

Compute the interval and check your answer:

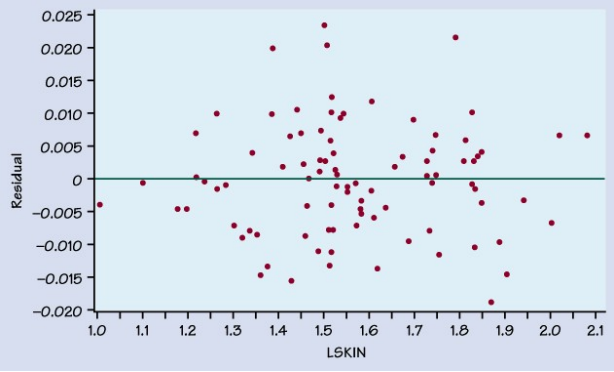
CI for β_1: $\hat{\beta}_1 \pm t^* \text{se}(\hat{\beta}_1)$ (note: df = n-2 for the t^* value)

Model Checking: How to check the Assumptions?

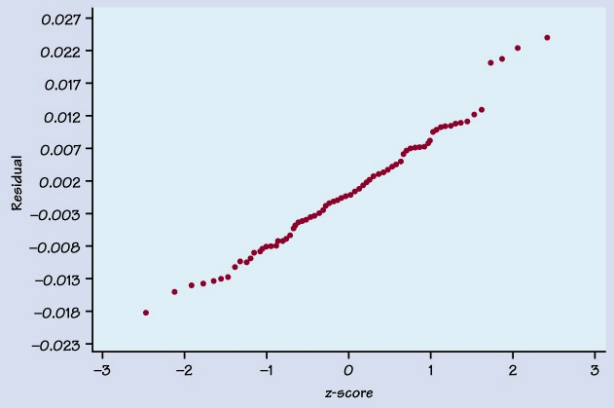
The residuals are a random sample from a normal population with mean 0 if the assumptions hold.

- Look at the residual plot to see if there is any pattern.
- Look at the normal quantile plot of the residuals to see if all points are close to a straight line.

Residual plot : residuals vs. the explanatory variable (x).



The normal quantile plot of the residuals:



What do you think about the assumptions of the simple linear regression model?

Summary for Simple Linear Regression

Model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where ϵ_i are assumed to be independent and have $N(0, \sigma)$ distribution.

Estimates: slope $\hat{\beta}_1 = r\frac{s_y}{s_x}$, intercept $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, and regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Estimate of σ : $s = \sqrt{s^2} = \sqrt{\text{MSE}}$

Test [$H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$] : $t = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}$ (note: df $= n - 2$)

CI for β_1: $\hat{\beta}_1 \pm t^* \text{se}(\hat{\beta}_1)$ (note: df $= n - 2$ for the t^* value)

Note: You need to know where to find the estimates, SE, the t statistics and p-values for the significance of the linear relationship from Stata outputs.