Topic 3: Correlation and Regression*

September 1 and 6, 2011

In this section, we shall take a careful look at the nature of **linear relationships** found in the data used to construct a scatterplot. The first of these, **correlation**, examines this relationship in a symmetric manner. The second, **regression**, considers the relationship of a **response variable** as determined by one or more **explanatory variables**. Correlation focuses primarily of association, while regression is designed to help make predictions. Consequently, the first does not attempt to establish any cause and effect. The second is a often used as a tool to establish causality.

1 Covariance and Correlation

The **covariance** measures the linear relationship between a pair of quantitative measures

$$x_1, x_2, \ldots, x_n \quad \text{and} \quad y_1, y_2, \ldots, y_n$$

on the same sample of n individuals. Beginning with the definition of variance, the definition of covariance is similar to the relationship between the norm $||v||$ or a vector v and the inner product $\langle v, w\rangle$ of two vectors v and w.

$$\text{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

A positive covariance means that the terms $(x_i - \bar{x})(y_i - \bar{y})$ in the sum are more likely to be positive than negative. This occurs whenever the x and y variables are more often both above or below the mean in tandem than not. Note that the covariance of x with itself $\text{cov}(x,x) = s_x^2$ is the variance of x.

Exercise 1. *Explain in words what a negative covariance signifies, what a covariance near 0 signifies.*

We next look at several exercises that call for algebraic manipulations of the formula for covariance or closely related functions.

Exercise 2. *Derive the alternative expression for the covaraince:*

$$\text{cov}(x,y) = \frac{1}{n-1}\left(\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}\right).$$

Exercise 3. $\text{cov}(ax+b, cy+d) = ac\cdot\text{cov}(x,y)$. *How does a change in units (say from centimeters to meters) affect the covariance?*

The **correlation**, r, is the covariance of the standardized versions of x and y.

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

Exercise 4. $r(ax+b, cy+d) = \pm r(x,y)$. *How does a change in units (say from centimeters to meters) affect the correlation? The plus sign occurs if* $a \cdot c > 0$ *and the minus sign occurs if* $a \cdot c < 0$.

Sometimes we will drop (x, y) if there is no ambiguity.

Exercise 5. *Show that*

$$s_{x+y}^2 = s_x^2 + s_y^2 + 2\text{cov}(x,y) = s_x^2 + s_y^2 + 2rs_xs_y. \tag{1}$$

Give the analogy between this formula and the law of cosines. In particular if the two observations are uncorrelated we have the **Pythagorean identity**

$$s_{x+y}^2 = s_x^2 + s_y^2. \tag{2}$$

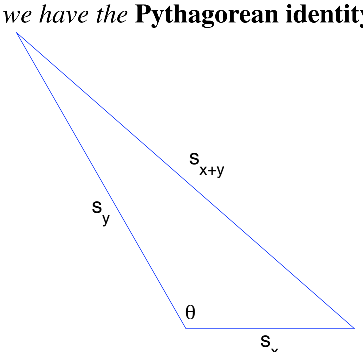


Figure 1: The analogy of the sample standard deviations and the law of cosines in equation (1). Here, the corrrelation $r = -\cos\theta$.

We will now look to uncover some of the properties. The next steps are to show that the correlation is always a number between -1 and 1 and to determine the relationship between the two variables in the case that the correlation takes on one of the two possible extreme values.

Exercise 6 (Cauchy-Schwarz inequality)**.** *For two sequences* $a_1, \cdots, a_n$ *and* $b_1, \ldots, b_n$, *show that*

$$\left(\sum_{i=1}^n a_ib_i\right)^2 \le \left(\sum_{i=1}^n a_i^2\right)\left(\sum_{i=1}^n b_i^2\right). \tag{3}$$

(Hint: Consider the expression $\sum_{i=1}^n (a_i + b_i\zeta)^2 \ge 0$ *as a quadratic expression in the variable* ζ *and consider the discriminant in the quadratic formula.) If the discriminant is zero, then we have equality in (3) and we have that* $\sum_{i=1}^n (a_i + b_i\zeta)^2 = 0$ *for exactly one value of* ζ.

We shall use inequality (3) by choosing $a_i = x_i - \bar{x}$ and $b_i = y_i - \bar{y}$ to obtain

$$\left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right)^2 \le \left(\sum_{i=1}^n (x_i - \bar{x})^2\right)\left(\sum_{i=1}^n (y_i - \bar{y})^2\right),$$

$$\left(\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right)^2 \le \left(\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2\right)\left(\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2\right),$$

$$\text{cov}(x,y)^2 \le s_x^2 s_y^2$$

$$\frac{\text{cov}(x,y)^2}{s_x^2 s_y^2} \le 1$$

Consequently, we find that

$$r^2 \le 1 \quad \text{or} \quad -1 \le r \le 1.$$

When we have $|r| = 1$, then we have equality in (3). In addition, for some value of ζ we have that

$$\sum_{i=1}^n ((x_i - \bar{x}) + (y_i - \bar{y})\zeta)^2 = 0.$$

The only way for a sum of nonnegative terms to add to give zero is for each term in the sum to be zero, i.e.,

$$(x_i - \bar{x}) + (y_i - \bar{y})\zeta = 0, \quad \text{for all } i = 1, \ldots, n.$$

Thus x_i and y_i are linearly related.

$$y_i = \alpha + \beta x_i.$$

In this case, the sign of r is the same as the sign of β.

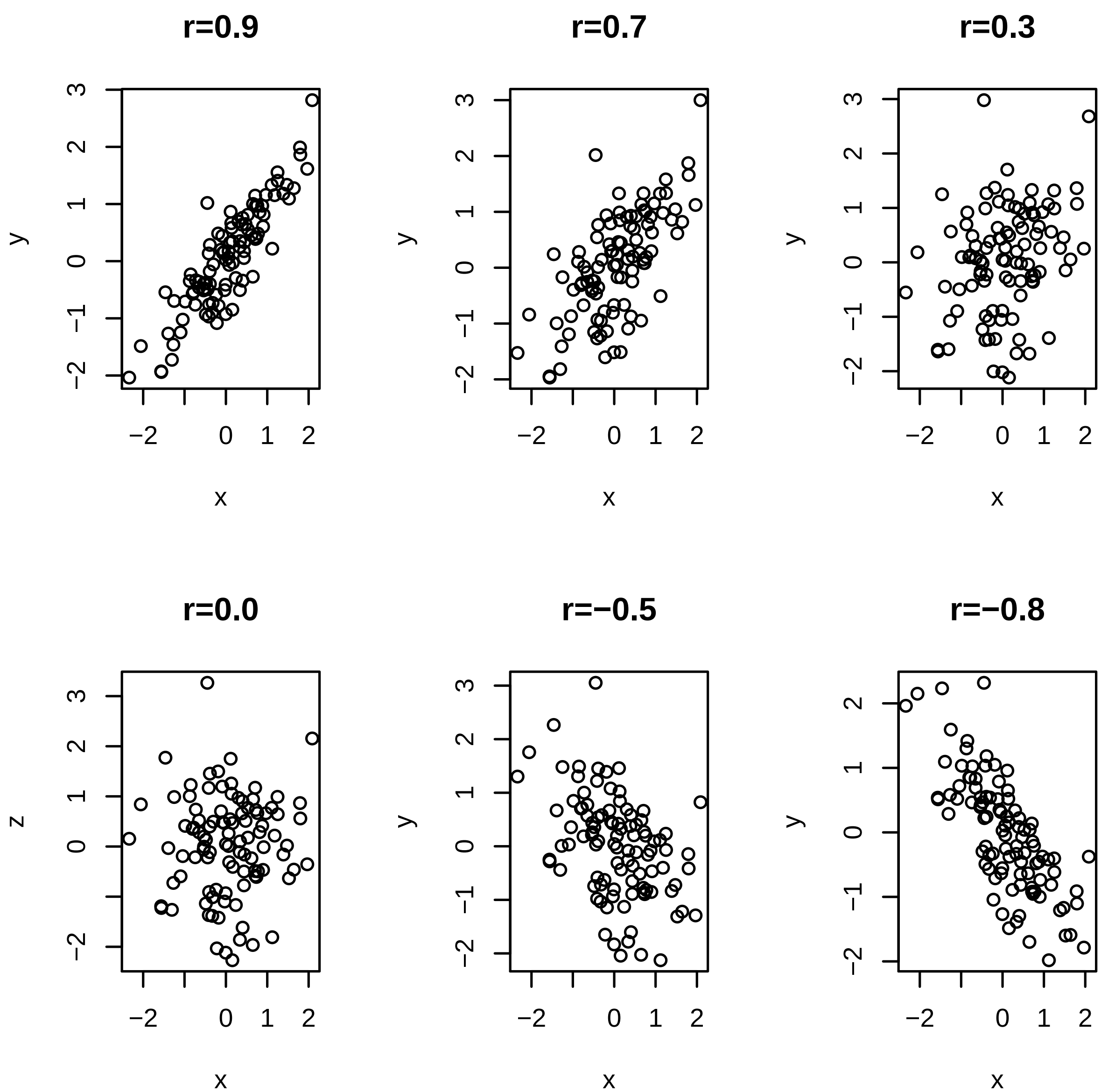
Exercise 7. *For a second proof that $-1 \le r \le 1$. Use equation (1) with x and y standardized observations.*

We can see how this looks for simulated data. Choose a value for `r` between -1 and $+1$.

```
>x<-rnorm(100)
>z<-rnorm(100)
>y<-r*x + sqrt(1-r^2)*z
```

For the moment, the object of this simulation is to obtain an intuitive feeling for differing values for correlation. We shall soon see that this is the simulation of pairs of normal random variables with the desired correlation. From the discussion above, we can see that the scatterplot would lie on a straight line for the values $r = \pm 1$.

Scatterplots showing differing levels of the correlation r



[t!]

For the *Archeopteryx* data on bone lengths, we have the correlation

```
> cor(femur, humerus)
[1] 0.9941486
```

Thus, the data land very nearly on a line with positive slope.

For the banks in 1974, we have the correlation

```
> cor(income,assets)
[1] 0.9325191
```

2 Linear Regression

Covariance and correlation are measures of linear association.

We now turn to situations in which the value of the first variable x_i will be considered to be **explanatory** or **predictive**. The corresponding observation y_i, taken from the input x_i, is called the **response**. For example, can we **explain** or **predict** the income of banks from its assets. In this case, *assets* is the explanatory variable and income is the response?

In **linear regression**, the response variable is linearly related to the explanatory variable, but is subject to deviation or to error. We write

$$y_i = \alpha + \beta x_i + \epsilon_i. \tag{4}$$

Our goal is, given the data, the x_i's and y_i's, to find α and β that determines the line having the best fit to the data. The principle of **least squares regression** states that the best choice of this linear relationship is the one that minimizes the square in the vertical distance from the y values in the data and the y values on the regression line. This choice reflects the fact that the values of x are set by the experimenter and are thus assumed known. Thus, the "error" appears only in the value of the response variable y.

This principle leads to a minimization problem for

$$S(\alpha,\beta) = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2.$$

In other words, given the data, determine the values for α and β that minimizes S. Take partial derivatives to find that

$$\frac{\partial}{\partial\beta} S(\alpha,\beta) = -2\sum_{i=1}^n x_i(y_i - \alpha - \beta x_i)$$
$$\frac{\partial}{\partial\alpha} S(\alpha,\beta) = -2\sum_{i=1}^n (y_i - \alpha - \beta x_i)$$

Set these two equations equal to 0 and call the solutions $\hat\alpha$ and $\hat\beta$.

$$0 = \sum_{i=1}^n x_i(y_i - \hat\alpha - \hat\beta x_i) = \sum_{i=1}^n x_i y_i - \hat\alpha \sum_{i=1}^n x_i - \hat\beta \sum_{i=1}^n x_i^2 \tag{5}$$

$$0 = \sum_{i=1}^n (y_i - \hat\alpha - \hat\beta x_i) = \sum_{i=1}^n y_i - n\hat\alpha - \hat\beta \sum_{i=1}^n x_i \tag{6}$$

Multiply these equations by the appropriate factors to obtain

$$0 = n\sum_{i=1}^n x_iy_i - n\hat\alpha\sum_{i=1}^n x_i - n\hat\beta\sum_{i=1}^n x_i^2 \tag{7}$$

$$0 = \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right) - n\hat\alpha\sum_{i=1}^n x_i - \hat\beta\left(\sum_{i=1}^n x_i\right)^2 \tag{8}$$

Now subtract the equation (8) from equation (7) and solve for $\hat\beta$.

$$\hat\beta = \frac{n\sum_{i=1}^n x_iy_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \frac{\sum_{i=1}^n (x_i-\bar x)(y_i-\bar y)}{\sum_{i=1}^n (x_i-\bar x)^2} = \frac{\text{cov}(x,y)}{\text{var}(x)}. \tag{9}$$

Next, divide equation (6) by n to obtain

$$\bar y = \hat\alpha + \hat\beta\bar x \tag{10}$$

The relation (10) states that the center of mass $(\bar x, \bar y)$ lies on the regression line. This can be used to determine $\hat\alpha$.

Thus, from equations (9) and (10), we obtain the **regression line**

$$\hat y_i = \hat\alpha + \hat\beta x_i.$$

We call $\hat y_i$ the **fit** for the value x_i.

Example 8. *Let's begin with 6 points and derive by hand the equation for regression line.*

x	-2	-1	0	1	2	3
y	-3	-1	-2	0	4	2

Add the x and y values and divide by $n = 6$ to see that $\bar x = 0.5$ and $\bar y = 0$.

x_i	y_i	$x_i - \bar x$	$y_i - \bar y$	$(x_i - \bar x)(y_i - \hat y)$	$(x_i - \bar x)^2$
-2	-3	-2.5	-3	7.5	6.25
-1	-1	-1.5	-1	1.5	2.25
0	-2	-0.5	-2	1.0	0.25
1	0	0.5	0	0.0	0.25
2	4	1.5	4	6.0	2.25
3	2	2.5	2	5.0	6.25
total		0	0	$\text{cov}(x,y) = 21/5$	$\text{var}(x) = 17.50/5$

Thus,

$$\hat\beta = \frac{21}{17.5} = 1.2 \quad \textit{and} \quad 0 = \hat\alpha + 1.2\times 0.5 = \hat\alpha + 0.6 \quad \textit{or} \quad \hat\alpha = -0.6$$

Fits, however, are rarely perfect. The difference between the fit and the data is an estimate $\hat\epsilon_i$ for the error ϵ_i. It is called the **residual**. So,

$$\text{RESIDUAL}_i = \text{DATA}_i - \text{FIT}_i = y_i - \hat y_i$$

$$\text{DATA}_i = \text{FIT}_i + \text{RESIDUAL}_i, \quad \textit{or} \quad y_i = \hat y_i + \hat\epsilon_i.$$

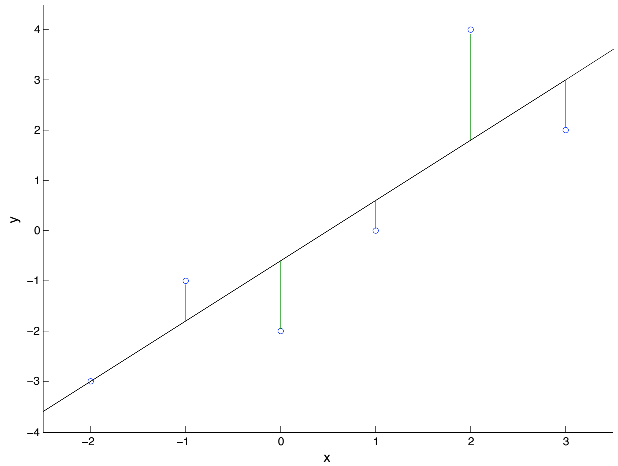
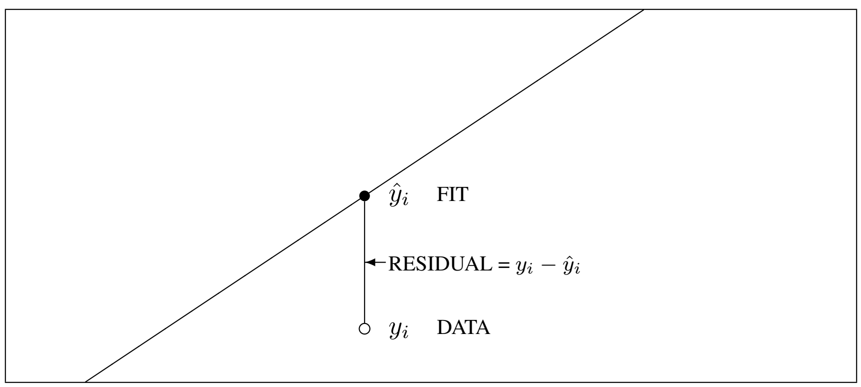


Figure 2: Scatterplot and the regression line for the six point data set below. The regression line is the choice that minimizes the square of the vertical distances from the observation values to the line, indicated here in green. Notice that the total length of the positive residuals (the lengths of the green line segments above the regression line) is equal to the total length of the negative residuals. This property is derived in equation (11).



We can rewrite equation (6) with $\hat{\epsilon}_i$ estimating the error in (4).

$$0 = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i) = \sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}\hat{\epsilon}_i \tag{11}$$

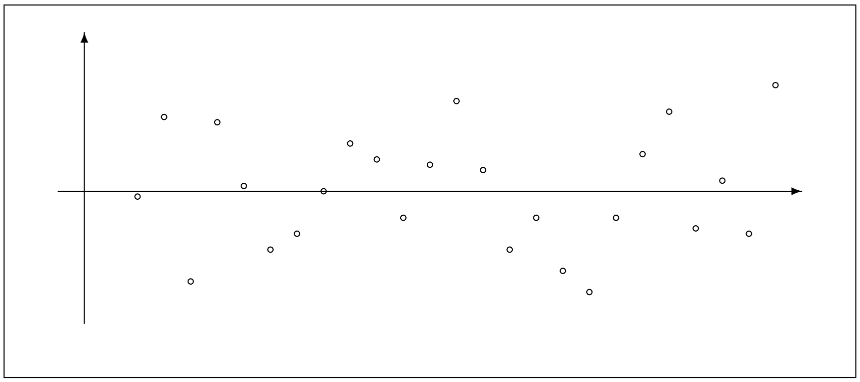
to see that the sum of the residuals is 0. Thus, we started with a criterion for a line of best fit, namely, least squares, and discover that as a consequence of this that the regression line also has the property that the sum of the residual values is 0. This is illustrated in Figure 2.

Let's check this property for the example above.

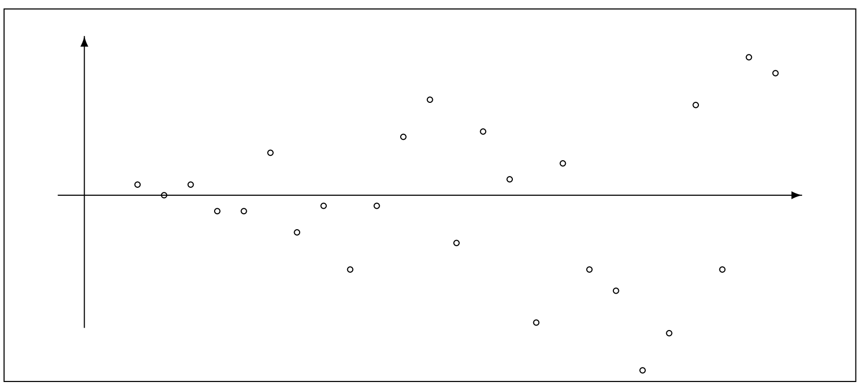
x_i	y_i	$\hat{y}_i$	$\hat{y}_i - y_i$
-2	-3	-3.0	0
-1	-1	-1.8	0.8
0	-2	-0.6	-1.4
1	0	0.6	-0.6
2	4	1.8	2.2
3	2	3.0	-1.0
	total		0

Generally speaking, we will look at a **residual plot**, the plot of the residuals versus the explanatory variable, to assess the appropriateness of a regression line. Specifically, we will look for circumstances in which the explanatory variable and the residuals have no systematic pattern.

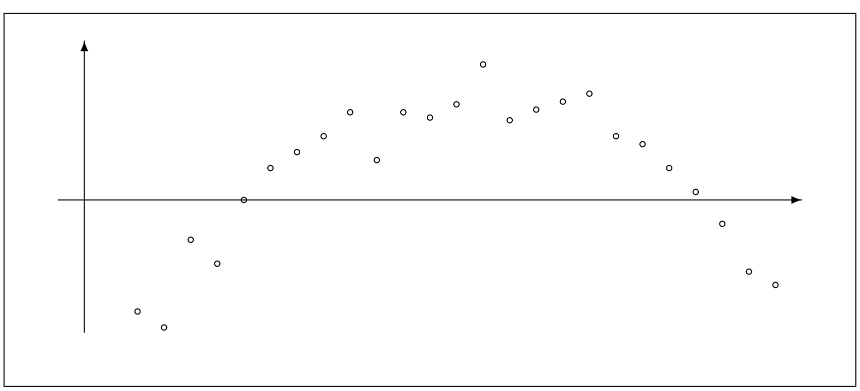
We next show three examples of the residuals plotting against the value of the explanatory variable.



Regression fits the data well - homoscedasticity



Prediction is less accurate for large x, an example of heteroscedasticity



Data has a curve. A straight line fits the data poorly.

For any value of x, we can use the regression line to estimate or **predict** a value for y. We must be careful in using this prediction outside the range of x. This **extrapolation** will not be valid if the relationship between x and y is not known to be linear in this extended region.

Example 9. *For the 1974 bank data set, the regression line*

$$income = 7.680 + 4.975 \cdot assets.$$

So, each dollar in assets brings in about $5 income.

For a bank having 10 billion dollars in assets, the predicted income is 56.430 billion dollars. However, if we extrapolate this down to very small banks, we would predict nonsensically that a bank with no *assets would have an income of 7.68 billion dollars. This illustrates the caution necessary to perform a reliable prediction through an extrapolation.*

In addition for this data set, we see that three banks have assets much greater than the others. Thus, we should consider examining the regression lines omitting the information from these three banks. If a small number of observations has a large impact on our results, we call these points **influential***.*

Obtaining the regression line in R *is straightforward:*

```
> lm(income~assets)

Call:
lm(formula = income ~ assets)

Coefficients:
(Intercept)       assets
      7.680        4.975
```

Example 10 (regression line in standardized coordinates)**.** *Sir Francis Galton was the first to use the term* **regression** *in his study* Regression towards mediocrity in hereditary stature. *The rationale for this term and the relationship between regression and correlation can be best seen if we convert the observations into standardized cooordinates.*

First, write the regression line to point-slope form.

$$\hat{y}_i - \bar{y} = \hat{\beta}(x_i - \bar{x}).$$

Because

$$\hat{\beta} = \frac{\text{cov}(x,y)}{\text{var}(x)} = \frac{rs_xs_y}{s_x^2} = \frac{rs_y}{s_x},$$

we can re-write the point slope form as

$$\hat{y}_i - \bar{y} = \frac{rs_y}{s_x}(x_i - \bar{x}) \quad or \quad \frac{\hat{y}_i - \bar{y}}{s_y} = r\frac{x_i - \bar{x}}{s_x}, \quad \hat{y}_i^* = rx_i^*. \tag{12}$$

where the asterisk is used to indicate that we are stating our observations in standardized form. In words, if we use standardized coordinates, then the slope of the regression line is the correlation.

For Galton's example, let's use the height of a male as the explanatory variable and the height of his adult son as the response. If we observe a correlation $r = 0.6$ and consider a man whose height is 1 standard deviation above the mean, then we predict *that the son's height is 0.6 standard deviations above the mean. If a man whose height is 0.5 standard deviation below the mean, then we* predict *that the son's height is 0.3 standard deviations below the mean. In either case, our prediction for the son is a height that is closer to the mean then the father's height. This is the "regression" that Galton had in mind.*

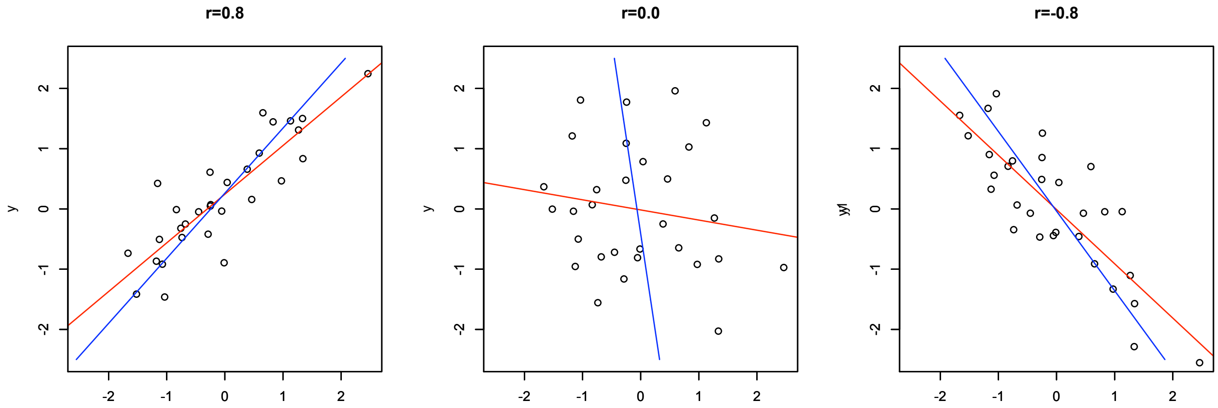


Figure 3: Scatterplots of standardized varibles and their regression lines. The red lines show the case in which x is the explanatory variable and the blue lines show the case in which y is the explanatory variable.

For the discussion above, we can see that if we reverse the role of the explanatory and response variable, then we change the regression line. This should be intuitively obvious since in the first case, we are minimizing the total square vertical distance and in the second, we are minimizing the total square horizontal distance. In the most extreme circumstance, cov$(x, y) = 0$. In this case, the value x_i of an observation is no help in predicting the response variable. Thus, as the formula states, when x is the explanatory variable the regression line has slope 0 - it is a horizontal line through $\bar{y}$. When y is the explanatory variable, the regression is a vertical line through $\bar{x}$. Intuitively, if x and y are uncorrelated, then the best prediction we can make for y_i given the value of x_i is just the sample mean $\bar{y}$ and he best prediction we can make for x_i given the value of y_i is the sample mean $\bar{x}$.

More formally, the two regression equations are

$$\hat{y}_i^* = rx_i^* \quad \text{and} \quad \hat{x}_i^* = ry_i^*.$$

These equations have slopes r and $1/r$. This is shown by example in Figure 2.

Exercise 11. *Compute the regression line for the 6 pairs of observations above assuming that y is the explanatory variable.*

Let's again write the regression line in point slope form

$$\text{FIT}_i - \bar{y} = \hat{y}_i - \bar{y} = r\frac{s_y}{s_x}(x_i - \bar{x}).$$

Using the quadratic identity for variance we find that

$$s_{FIT}^2 = r^2 \frac{s_y^2}{s_x^2} s_x^2 = r^2 s_y^2 = r^2 s_{DATA}^2 \quad \text{or} \quad r^2 = \frac{s_{FIT}^2}{s_{DATA}^2}$$

When the straight line fits the data well, the FIT and the RESIDUAL are uncorrelated and the magnitude of the residual does not depend on the value of the explanatory variable. We have, in this circumstance, from equation (2), the Pythagorean identity, that

$$s_{DATA}^2 = s_{FIT}^2 + s_{RESID}^2 = r^2 s_{DATA}^2 + s_{RESIDUAL}^2$$

$$s_{RESIDUAL}^2 = (1 - r^2) s_{DATA}^2.$$

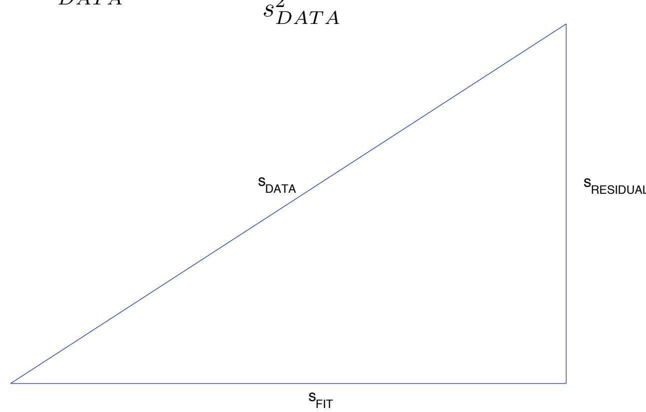


Figure 4: The relationship of the standard deviations of the DATA, the FIT, and the RESIDUALS in the case that the FIT and the RESIDUALS are uncorrelated. $s_{DATA}^2 = s_{FIT}^2 + s_{RESID}^2 = r^2 s_{DATA}^2 + (1 - r^2) s_{DATA}^2$. In this case, we say that r^2 of the variation in the response variable is due to the fit and the rest $1 - r^2$ is due to the residuals.

Thus, r^2 of the variance in the data can be explained by the fit. As a consequence of this computation, many statistical software tools report r^2 as a part of the linear regression analysis. In the case the remaining $1 - r^2$ of the variance in the data is found in the residuals.

Exercise 12. *For some situations, the circumstances dictate that the line contain the origin. Use a least squares criterion to derive the slope of the regression line.*

Remark 13 (logarithms)**.** *In the next example, we will work with* **logarithms***. We will use both* $\log$*, the base 10* **common logarthm***, and* $\ln$*, the base* e **natural logarithm***. Common logarithms help us see orders of magnitude. For example, if* $\log y = 5$*, then we know that* $y = 10^5 = 100,000$*. if* $\log y = -1$*, then we know that* $y = 10^{-1} = 1/10$*. We will use natural logarithms to show instantaneous rates of growth. For example. consider the differential equation*

$$\frac{dy}{dt} = ky.$$

We are saying that the instantaneous rate of growth of y *is proportional to* y *with constant of proportionality* k*. The solution to this equation is*

$$y = y_0 e^{kt} \qquad \textit{or} \qquad \ln y = \ln y_0 + kt$$

where y_0 *is the initial value for* y*. This gives a linear relationship between* $\ln y$ *and* t*. The two values of logarithm have a simple relationship. If we write*

$$x = 10^a. \textit{ Then } \log x = a \textit{ and } \ln x = a \ln 10.$$

Thus, by substituting for a*, we find that*

$$\ln x = \log x \cdot \ln 10 = 2.3026 \log x.$$

In R*, the command for the natural logarithm of* `x` *is* `log(x)`*. For the common logarithm, it is* `log(x,10)`*.*

Example 14. *In the data on world oil production, the relationship between the explanatory variable and response variable is nonlinear but can be made to be linear with a simple transformation, the common logarithm. Call the new response variable* logbarrel*. The explanatory variable remains* year*. With these variables, we can use a regression line to help describe the data.*

Regression is the first example of a class of statistical models called **linear models***. Thus,* R *uses the command* `lm`*. Here is the output.*

```
> summary(lm(logbarrel~year))

Call:
lm(formula = logbarrel ~ year)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25562 -0.03390  0.03149  0.07220  0.12922

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.159e+01  1.301e+00  -39.64   <2e-16 ***
year         2.675e-02  6.678e-04   40.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.1115 on 27 degrees of freedom
Multiple R-Squared: 0.9834,Adjusted R-squared: 0.9828
F-statistic:  1604 on 1 and 27 DF,  p-value: < 2.2e-16
```

Note that the output states $r^2 = 0.9828$*. Thus, the correlation is* $r = 0.9914$ *is very nearly one and so the data lies very close to the regression line.*

For world oil production, we obtained the relationship

$$\widehat{\log(barrel)} = -0.05159 + 0.02675 \cdot year.$$

If we rewrite the equation in exponential form, we obtain

$$\widehat{barrel} = A10^{0.02675 \cdot year} = Ae^{\hat{k} \cdot year}.$$

Thus, $\hat{k}$ *gives the instantaneous growth rate that best fits the data. This is obtained by converting from a common logarithm to a natural logarithm.*

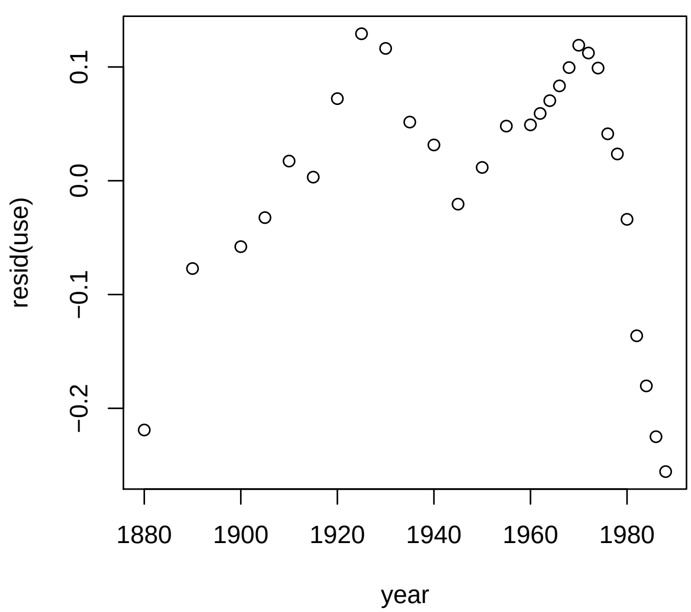
$$\hat{k} = 0.02675 \ln 10 = 0.0616$$

.

Consequently, the use of oil sustained a growth of 6% per year over a span of a hundred years.

Next, we will look for finer scale structure by examining the residual plot.

```
> use<-lm(logbarrel~year)
> plot(year,resid(use))
```



Exercise 15. *Remove the data points after the oil crisis of the mid 1970s, find the regression line and the instantaneous growth rate that best fits the data. Look at the residual plot and use fact about American history to explain why the residuals increase until 1920's, decrease until the early 1940's and increase again until the early 1970's.*

Example 16 (Michaelis-Menten Kinetics)**.** *In this example, we will have to use a more sophisticated line of reasoning to create a linear relationship between a explanatory and response variable. Consider the chemical reaction in which an enzyme catalyzes the action on a substrate.*

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \overset{k_2}{\to} E + P \tag{13}$$

Here

- E *is the amount of free enzyme.*
- S *is the substrate.*
- ES *is the substrate-bound enzyme.*
- P *is the product.*
- $V = d[P]/dt$ *is the production rate.*

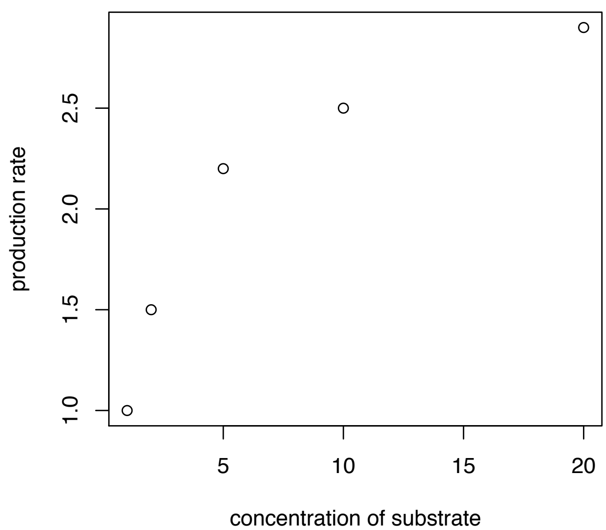
and the numbers above or below the arrows gives the reaction rates. Using the symbol $[\cdot]$ *to indicate* **concentration***, our goal is to relate the production rate* V *to the substrate concentrtion* $[S]$*.*

The **law of mass action** *turns the chemical reactions in (13) into differential equations. In particular, the reactions, focusing on the substrate-bound enzyme and the product gives the equations*

$$\frac{d[ES]}{dt} = k_1[E][S] - [ES](k_{-1} + k_2) \quad \textit{and} \quad \frac{d[P]}{dt} = k_2[ES] \tag{14}$$

Let's look at data,

$[S]$ (mM)	1	2	5	10	20
V (nmol/sec)	1.0	1.5	2.2	2.5	2.9



If we wish to use linear regression, then we will have to transform the data. In this case, we will develop the Michaelis-Menten transformation applied to situations in which the concentration of the substrate-bound enzyme (and hence also the unbound enzyme) change much more slowly than those of the product and substrate.

$$0 \approx \frac{d[ES]}{dt}$$

In words, the substrate-bound enzyme is nearly in steady state. Using the law of mass action equation (14) for $d[ES]/dt$, we can rearrange terms to conclude that

$$[ES] \approx \frac{k_1[E][S]}{k_{-1}+k_2} = \frac{[E][S]}{K_m}. \tag{15}$$

The ratio $K_m = (k_{-1}+k_2)/k_1$ of the rate of loss of $[ES]$ to its production is called the **Michaelis constant***.* *Enzyme, E_0, is either free or bound to the substrate. Its total concentration is, therefore,*

$$[E_0] = [E] + [ES], \quad \textit{and, thus} \quad [E] = [E_0] - [ES]$$

Combine this with (15) and solve for $[ES]$,

$$[ES] \approx \frac{([E_0]-[ES])[S]}{K_m}, \quad [ES] \approx [E_0]\frac{[S]}{K_m+[S]}$$

Under this approximation, the production rate of the product is:

$$V = \frac{d[P]}{dt} = k_2[ES] = k_2[E_0]\frac{[S]}{K_m+[S]} = V_{\max}\frac{[S]}{K_m+[S]}$$

Here, $V_{\max} = k_2[E_0]$ is the maximum production rate. (To see this, let the substrate concentration $[S] \to \infty$.) The **Lineweaver-Burke double reciprocal plot** *provides a useful method for analysis of the Michaelis-Menten equation:*

$$V = V_{\max}\frac{[S]}{K_m+[S]}.$$

Taking the reciprocal gives

$$\frac{1}{V} = \frac{K_m+[S]}{V_{\max}[S]} = \frac{K_m}{V_{\max}}\frac{1}{[S]} + \frac{1}{V_{\max}} \tag{16}$$

Thus, we have a linear relationship between

$$\frac{1}{V}\text{, the response variable, and } \frac{1}{[S]}\text{, the explanatory variable}$$

subject to experimental error.

For the data,

$[S]$ (mM)	1	2	5	10	20
V (nmol/sec)	1.0	1.5	2.2	2.5	2.9

The regression line is

$$\frac{1}{V} = 0.3211 + \frac{1}{[S]}0.6813.$$

Here are the R *commands*

```
> S<-c(1,2,5,10,20)
> V<-c(1.0,1.5,2.2,2.5,2.9)
> Sinv<-1/S
> Vinv<-1/V
> lm(Vinv~Sinv)

Call:
lm(formula = Vinv ~ Sinv)

Coefficients:
(Intercept)         Sinv
     0.3211       0.6813
```

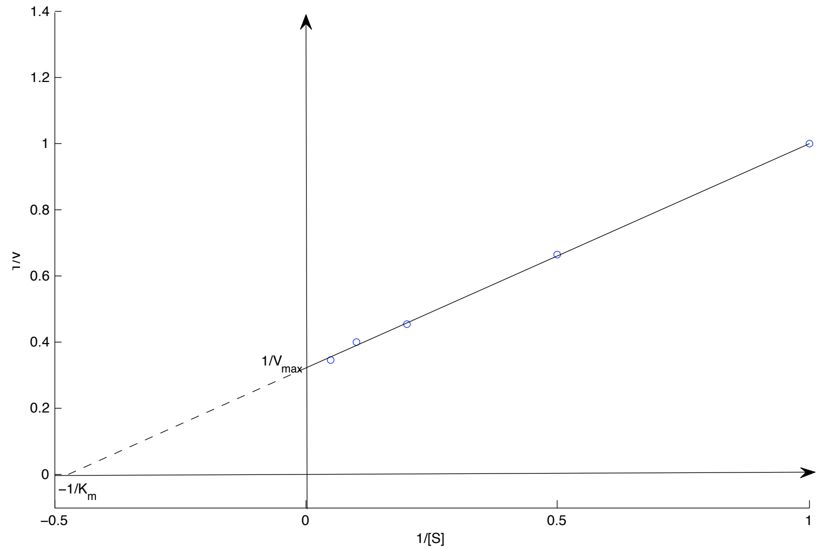


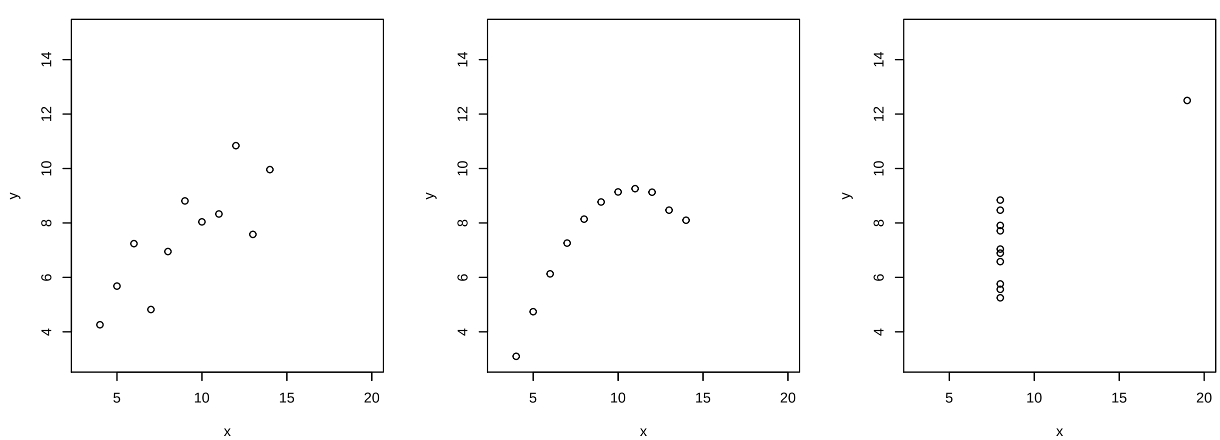
Figure 5: Lineweaver-Burke double reciprocal plot for the data presented above. The y-intercept gives the reciprocal of the maximum production. The dotted line indicates that negative concentrations are not physical. Nevertheless, the x-intercept give the negative reciprocal of the Michaelis constant.

Using (16), we fid that $V_{\max} = 3.1141$ and $K_m = 0.4713$. This method is not used as much as before. The measurements for small values of the concentration (and thus large value of $1/[S]$) are more variable. Look in the next section for an alternative approach.

Example 17 (Frank Amscombe)**.** *Consider the three data sets:*

x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.47	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
x	8	8	8	8	8	8	8	8	8	8	19
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

t



Each of these data sets has a regression line $\hat{y} = 3 + 0.5x$ and correlations between 0.806 and 0.816. However, only the first is a suitable data set for linear regression. This example is meant to emphasize the point that software will happily compute a regression line and an r^2 value, but further examination of the data is required to see if this method is appropriate for any given data set.

3 Extensions

We will discuss briefly two extensions - the first is a least squares criterion between x and y that is **nonlinear** in the parameters $\beta = (\beta_0, \ldots, \beta_k)$. Thus, the model is

$$y_i = g(x_i|\beta) + \epsilon_i.$$

The second considers situations with more than one explanatory variable.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i. \tag{17}$$

This brief discussion does not have the detail necessary to begin to use these methods. It serves primarily as an invitation to begin to consult resources that more fully develop these ideas.

3.1 Nonlinear Regression

Here, we continue using estimation of parameters using a **least squares** criterion.

$$SS(\beta) = \sum_{i=1}^{n} (y_i - g(x_i|\beta))^2.$$

For most choices of $g(x|\beta)$ the solution to the nonlinear least square criterion cannot be expressed in closed form. Thus, a numerical strategy for the solution $\hat{\beta}$ is necessary. This generally begins with some initial guess of parameter values and an iteration scheme to minimize $SS(\beta)$. Such a scheme is likely to use local information about the first and second partial derivatives of g with respect to the β_i.

The R command `gnls` for **general nonlinear least squares** is used to accomplish this. As above, you should examine the residual plot to see that it has no structure.

For, example, if the Lineweaver-Burke method for Michaelis-Mentens kinetics yields structure in the residuals, then linear regression is not considered a good method. Under these circumstances, one can next try to use the parameter estimates derived from Lineweaver-Burke and use this in a nonlinear least squares regression using

$$SS(V_{max}, K_m) = \sum_{j=1}^{n} \left(V_j - V_{max} \frac{[S]_j}{K_m + [S]_j} \right)^2$$

for data $(V_1, [S]_1), (V_2, [S]_2), \ldots (V_n, [S]_n)$.

3.2 Multiple Linear Regression

Before we start with multiple linear regression, we first recall a couple of concepts from linear algebra.

- Suppose we have a d-dimensional vector a of known values and a $d \times d$ matrix C and we want to determine the vectors x that satisfy

  $$a = Cx.$$

  This equation could have no solutions, a single solution, or infinite number of solutions. If the matrix C is invertible with inverse C^{-1}. Then we have a single solution

  $$x = C^{-1}a.$$

- The **transpose** of a matrix is obtained by reversing the rows and columns of a matrix. We use a superscript T to indicate the **transpose** of a matrix. Thus, the ij entry of a matrix C is the ji entry of its transpose, C^T.

  Example 18.
  $$\begin{pmatrix} 2 & 1 & 3 \\ 4 & 2 & 7 \end{pmatrix}^T = \begin{pmatrix} 2 & 4 \\ 1 & 2 \\ 3 & 7 \end{pmatrix}$$

- A square matrix C is invertible if and only if its determinant $\det(C) \neq 0$. For a 2×2 matrix
  $$C = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$
  $\det(c) = ad - bc$ and the matrix inverse
  $$C^{-1} = \frac{1}{\det(C)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Exercise 19. $(Cx)^T = x^T C^T$

Exercise 20. *For*
$$C = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix},$$
find $\det(C)$ *and* C^{-1}.

In multiple linear regression, we have more than one predictor or explanatory random variable. Thus can write (17) in matrix form
$$y = X\beta + \epsilon$$

- $y = (y_1, y_2, \ldots, y_n)^T$ is a column vector of responses,
- X is a matrix of predictors,
$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}. \tag{18}$$
- $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ is a column vector of parameters, and
- $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$ is a column vector of "errors".

Exercise 21. *Show that the least squares criterion*
$$SS(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - \beta_k x_{ik})^2. \tag{19}$$
can be written in matrix form as
$$SS(\beta) = (\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta).$$

To minimize S, we take the gradient and set it equal to 0.

Exercise 22. *Check that the gradient is*
$$\nabla_\beta SS(\beta) = -2(\mathbf{y} - X\beta)^T X. \tag{20}$$

Based on the exercise above, the value $\hat{\beta}$ that minimizes S is

$$(\mathbf{y} - X\hat{\beta})^T X = 0, \quad \mathbf{y^T} X = \hat{\beta}^T X^T X.$$

Taking the transpose of this last equation

$$X^T X\hat{\beta} = X^T \mathbf{y}.$$

If $X^T X$ is invertible, then we can multiply both sides of the equation above by $(X^T X)^{-1}$ to obtain an equation for the parameter values $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n)$ in the least squares regression.

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}. \tag{21}$$

This gives the regression equation

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

Example 23 (ordinary least squares regression)**.** *In this case,*

$$X^T X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

and

$$X^T y = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}.$$

The determinant of $X^T X$ is

$$n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 = n(n-1)\text{var}(x).$$

and thus

$$(X^T X)^{-1} = \frac{1}{n(n-1)\text{var}(x)} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

and

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y} = \frac{1}{n(n-1)\text{var}(x)} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}.$$

For example, for the second row, we obtain

$$\frac{1}{n(n-1)\text{var}(x)} \left(\left(-\sum_{i=1}^n x_i\right) \left(\sum_{i=1}^n y_i\right) + n\sum_{i=1}^n x_i y_i \right) = \frac{n(n-1)\text{cov}(x,y)}{n(n-1)\text{var}(x)} = \frac{\text{cov}(x,y)}{\text{var}(x)}$$

as seen in equation (9).

Example 24. *The choice of $x_{ij} = x_i^j$ in (18) results in* **polynomial regression**

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \epsilon_i.$$

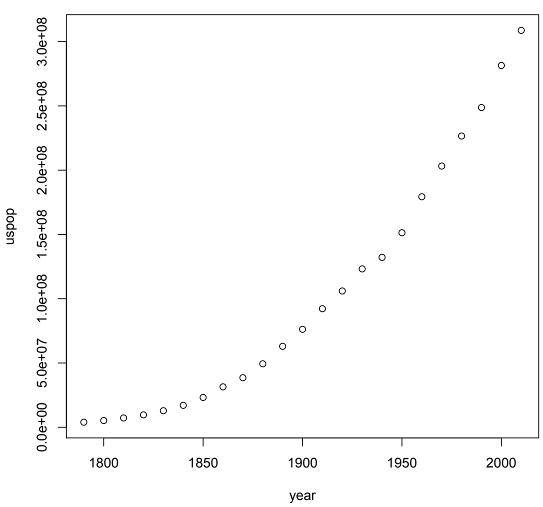
in equation (17).

Example 25 (US population). *Below are the census populations*

year	census population	year	census population	year	census population	year	census population
1790	3,929,214	1850	23,191,876	1910	92,228,496	1970	203,211,926
1800	5,236,631	1860	31,443,321	1920	106,021,537	1980	226,545,805
1810	7,239,881	1870	38,558,371	1930	123,202,624	1990	248,709,873
1820	9,638,453	1880	49,371,340	1940	132,164,569	2000	281,421,906
1830	12,866,020	1890	62,979,766	1950	151,325,798	2010	308,745,538
1840	17,069,453	1900	76,212,168	1960	179,323,175		

To analyze this in R *we enter the data:*

```
> uspop<-c(3929214,5236631,7239881,9638453,12866020,17069453,23191876,31443321,
+ 38558371,49371340,62979766,76212168,92228496,106021537,123202624,132164569,
+ 151325798,179323175,203211926,226545805,248709873,281421906,308745538)
> year<-c(0:22)*10 +1790
> plot(year,uspop)
> loguspop<-log(uspop,10)
> plot(year,loguspop)
```



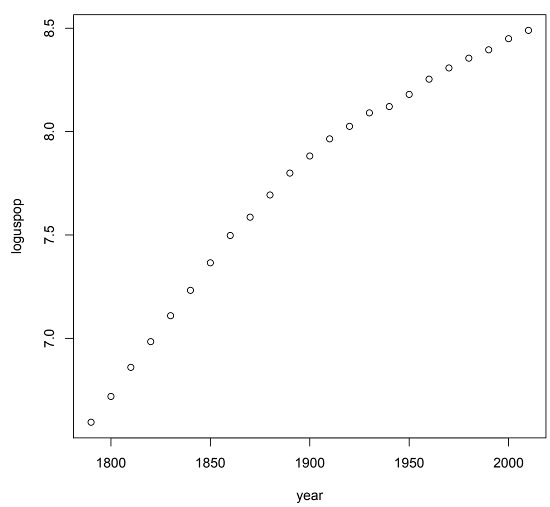


Figure 6: (a) United States census population from 1790 to 2010 and (b) it base 10 logarithm.

Note that the logarithm of the population still has a bend to it, so we will perform a quadratic regression on the logarithm of the population. In order to keep the numbers smaller, we shall give the year minus 1790, the year of the first census for our explanatory variable.

$$\log(uspopulation) = \beta_0 + \beta_1(year - 1790) + \beta_2(year - 1790)^2.$$

```
> year1<-year-1790
> year2<-year1^2
> lm.uspop<-lm(loguspop~year1+year2)
```

```
> summary(lm.uspop)

Call:
lm(formula = loguspop ~ year1 + year2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.037387 -0.013453 -0.000912  0.015281  0.029782

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.582e+00  1.137e-02  578.99   <2e-16 ***
year1        1.471e-02  2.394e-04   61.46   <2e-16 ***
year2       -2.808e-05  1.051e-06  -26.72   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.01978 on 20 degrees of freedom
Multiple R-squared: 0.999,Adjusted R-squared: 0.9989
F-statistic:  9781 on 2 and 20 DF,  p-value: < 2.2e-16
```

The R *output shows us that*

$$\hat{\beta}_0 = 6.587 \quad \hat{\beta}_1 = 0.1189 \quad \hat{\beta}_2 = -0.00002905.$$

So, taking the the regression line to the power 10, we have that

$$\widehat{uspopulation} = 3863670 \times 10^{0.1189(year-1790)-0.00002905(year-1790)^2}$$

In Figure 7, we show the residual plot for the logarithm of the US population.

```
> resid.uspop<-resid(lm.uspop)
> plot(year,resid.uspop)
```

4 Answers to Selected Exercises

1. Negative covariance means that the terms $(x_i - \bar{x})(y_i - \bar{y})$ in the sum are more likely to be negative than positive. This occurs whenever one of the x and y variables is above the mean, then the other is likely to be below.

2. We expand the product inside the sum.

$$\begin{aligned}\text{cov}(x,y) &= \frac{1}{n-1}\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^n x_iy_i - \bar{y}\sum_{i=1}^n x_i - \bar{x}\sum_{i=1}^n y_i + n\bar{x}\bar{y}\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^n x_iy_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}\right) = \frac{1}{n-1}\left(\sum_{i=1}^n x_iy_i - n\bar{x}\bar{y}\right)\end{aligned}$$

The change in measurements from centimeters to meters would divide the covariance by 10,000.

3. We rearrange the terms and simplify.

$$\begin{aligned}\text{cov}(ax+b, cy+d) &= \frac{1}{n-1}\sum_{i=1}^n ((ax_i+b) - (a\bar{x}+b)((cy_i+d) - (c\bar{y}-d)) \\ &= \frac{1}{n-1}\sum_{i=1}^n (ax_i - a\bar{x})(cy_i - c\bar{y}) = ac\cdot\frac{1}{n-1}\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y}) = ac\cdot\text{cov}(x,y)\end{aligned}$$

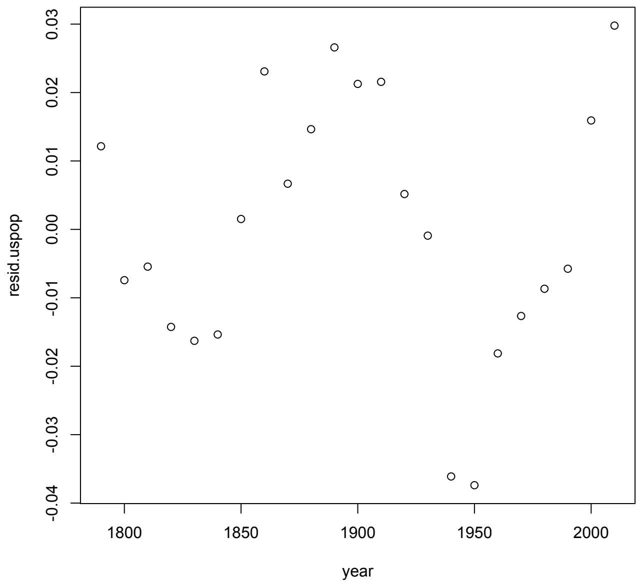


Figure 7: Residual plot for US population regression.

4. Assume that $a \neq 0$ and $c \neq 0$. If $a = 0$ or $c = 0$, then the covariance is 0 and so is the correlation.

$$r(ax+b, cy+d) = \frac{\text{cov}(ax+b, cy+d)}{s_{ax+b}s_{cy+d}} = \frac{ac \cdot \text{cov}(x,y)}{|a|s_x \cdot |c|s_y} = \frac{ac}{|ac|}\frac{\cdot \text{cov}(x,y)}{s_x \cdot s_y} = \pm r(x,y)$$

We take the plus sign if the sign of a and c agree and the minus sign if they differ.

5. First we rearrange terms

$$\begin{aligned} s^2_{x+y} &= \frac{1}{n-1}\sum_{i=1}^n((x_i+y_i)-(\bar{x}+\bar{y}))^2 = \frac{1}{n-1}\sum_{i=1}^n((x_i-\bar{x})+(y_i-\bar{y}))^2 \\ &= \frac{1}{n-1}\sum_{i=1}^n(x_i-\bar{x})^2 + 2\frac{1}{n-1}\sum_{i=1}^n(x_i-\bar{x})(y_i-\bar{y}) + \frac{1}{n-1}\sum_{i=1}^n(y_i-\bar{y})^2 \\ &= s_x^2 + 2\text{cov}(x,y) + s_y^2 = s_x^2 + 2rs_xs_y + s_y^2 \end{aligned}$$

For a triangle with sides a, b and c, the law of cosines states that

$$c^2 = a^2 + b^2 - 2ab\cos\theta$$

where θ is the measure of the angle opposite side c. Thus the analogy is

s_x corresponds to a, s_y corresponds to b, s_{x+y} corresponds to c, and r corresponds to $-\cos\theta$

Notice that both r and $\cos\theta$ take values between -1 and 1.

6. Using the hint,

$$0 \le \sum_{i=1}^n (a_i + b_i\zeta)^2 = \sum_{i=1}^n a_2^2 + 2\left(\sum_{i=1}^n a_i b_i\right)\zeta + \left(\sum_{i=1}^n b_i^2\right)\zeta^2 = A + B\zeta + C\zeta^2$$

For a quadratic equation to always take on non-negaitive values, we must have a negative discriminant

$$0 \ge B^2 - 4AC = 4\left(\sum_{i=1}^n a_i b_i\right)^2 - 4\left(\sum_{i=1}^n a_i^2\right)\left(\sum_{i=1}^n b_i^2\right).$$

$$\left(\sum_{i=1}^n a_i^2\right)\left(\sum_{i=1}^n b_i^2\right) \ge \left(\sum_{i=1}^n a_i b_i\right)^2$$

7. If x and y are standardized observations, then $s_x = s_y = 1$ Now, using equation (1), we have that

$$0 \le s_{x+y} = 1 + 1 + 2r = 2 + 2r \text{ or } -2 \le 2r \text{ and } r \ge -1.$$

For the second inequality, use the similar identity to (1) for the difference in the observations

$$s_{x-y}^2 = s_x^2 + s_y^2 - 2rs_xs_y.$$

Then,

$$0 \le s_{x-y} = 1 + 1 - 2r = 2 - 2r \text{ or } 2r \le 2 \text{ and } r \le 1.$$

Thus, correlation must always be between -1 and 1.

8. The least squares criterion becomes

$$S(\beta) = \sum_{i=1}^n (y_i - \beta x_i)^2.$$

The derivative with respect to β is

$$S'(\beta) = -2\sum_{i=1}^n x_i(y_i - \beta x_i).$$

$S'(\beta) = 0$ for the value

$$\hat\beta = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

11. Use the subscript y in $\hat\alpha_y$ and $\hat\beta_y$ to emphasize that y is the explanatory variable. We still have $\bar x = 0.5, \bar y = 0$.

y_i	x_i	$y_i - \bar y$	$x_i - \bar x$	$(x_i - \bar x)(y_i - \hat y)$	$(y_i - \bar y)^2$
-3	-2	-3	-2.5	7.5	9
-1	-1	-1	-1.5	1.5	1
-2	0	-2	-0.5	1.0	4
0	1	0	0.5	0.0	0
4	2	4	1.5	6.0	16
2	3	2	2.5	5.0	4
total		0	0	cov$(x,y) = 21/5$	var$(y) = 34/5$

So, the slope $\hat{\beta}_y = 34/21$ and

$$\bar{x} = \hat{\alpha}_y + \hat{\beta}_y \bar{y}, \quad 1/2 = \hat{\alpha}_y.$$

Thus, to predict x from y, the regression line is $\hat{x}_i = 1/2 + 34/21 y_i$. This differs from the line used to predict y from x.

12. In this circumstance, the y-intercept is 0 and so the least square criterion become minimizing

$$SS(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2.$$

Take a derivative with respect to β.

$$SS(\beta) = -2 \sum_{i=1}^{n} (y_i - \beta x_i).$$

Call $\hat{\beta}$ the value satisfying $S'(\hat{\beta}) = 0$ and check that it is a minimum. Then

$$\hat{\beta} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} = \frac{(\sum_{i=1}^n y_i)/n}{(\sum_{i=1}^n x_i)/n} = \frac{\bar{y}}{\bar{x}}.$$

17. The i-th component of $(Cx)^T$ is

$$\sum_{j=1}^{n} C_{ij} x_j.$$

Now the i-th component of $x^T C^T$ is

$$\sum_{j=1}^{n} x_j C^T_{ji} = \sum_{j=1}^{n} x_j C_{ij}.$$

18. The j-component of $\mathbf{y} - X\beta$,

$$(\mathbf{y} - X\beta)_j = y_i - \sum_{i=0}^{k} \beta_i x_{ik} = y_i - \beta_0 - x_{i1}\beta_1 - \cdots - \beta_k x_{ik}.$$

Now, $(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta)$ is the dot product of $\mathbf{y} - X\beta$ with itself. This gives (19).

19. We check that $(Cx)_j = (x^T C^T)_j$ for each component j. The j-th component of column vector Cx

$$(Cx)_j = \sum_{i=1}^{d} C_{ji} x_i$$

is the j-th component of row vector $(Cx)^T$ Now, the j-th component of row vector $x^T C^T$,

$$(x^T C^T)_j = \sum_{i=1}^{d} x_i^T C_{ij}^T = \sum_{i=1}^{d} C_{ji} x_i = (Cx)_j.$$

20. $\det(C) = 4 - 6 = -2$ and

$$C^{-1} = \frac{1}{-2} \begin{pmatrix} 4 & -3 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} -2 & 3/2 \\ 1 & -1/2 \end{pmatrix}.$$

22. Write $x_{i0} = 1$ for all i, then we can write (19) as

$$SS(\beta) = \sum_{i=1}^{n} (y_i - x_{i0}\beta_0 - x_{i1}\beta_1 - \cdots - \beta_k x_{ik})^2.$$

Then,

$$\begin{aligned}\frac{\partial}{\partial \beta_j} S(\beta) &= -2\sum_{i=1}^{n} (y_i - x_{i0}\beta_0 - x_{i1}\beta_1 - \cdots - \beta_k x_{ik})x_{ij} \\ &= -2\sum_{i=1}^{n} (y_i - (X\beta)_i)x_{ij} = -2((y - X\beta)^T X))_j.\end{aligned}$$

This is the j-th coordinate of (20).