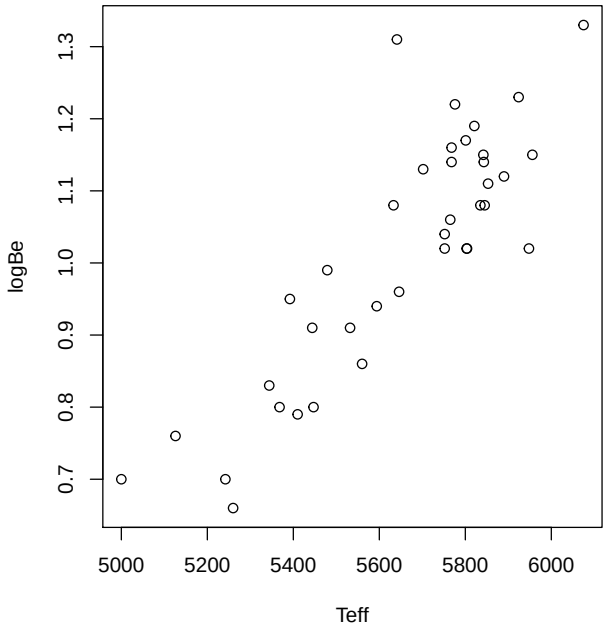
STAT 515 fa 2023 Lec 18

Simple Linear Regression

Karl Gregory

Studying the relationship between two variables

You have likely seen many plots like the one below, in which the values of one variable are plotted against the values of another. This is called a scatterplot. The data to make the plot are pairs of numbers $(x_1, Y_1), \ldots, (x_n, Y_n)$. For this plot the x values were temperatures of stars and the Y values were the natural log of the beryllium abundance in the corresponding stars. The data are taken from [1].



Plots like these can be helpful for depicting the relationship between two random variables X and Y. In addition, they can help researchers make predictions; for example, there is no star in the data set with temperature exactly 5200, but based if one were to draw a straight

line through the points in the scatterplot, one could use the height of the line at $x = 5200$ to make a reasonable guess about the log of beryllium abundance in such a star.

In this lecture we will introduce what is called the correlation coefficient, which describes the strength and direction of a linear relationship between random variables, and then introduce simple linear regression analysis. The latter is a way to draw "the best" straight line through the data and to make statistical inferences—conclusions to which we can attach levels of confidence—about the linear relationship between the variables X and Y.

Pearson's correlation coefficient

When we suspect that two variables might be linearly related (related in such a way that one could draw a straight line through a scatterplot of their values), we often compute what is called *Pearson's correlation coefficient*, which we will denote by r_{xY}. This is a measure describing the strength and direction of a linear relationship between two variables.

Definition: Pearson's correlation coefficient

For data pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$, Pearson's correlation coefficient is defined as

$$r_{xY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}}.$$

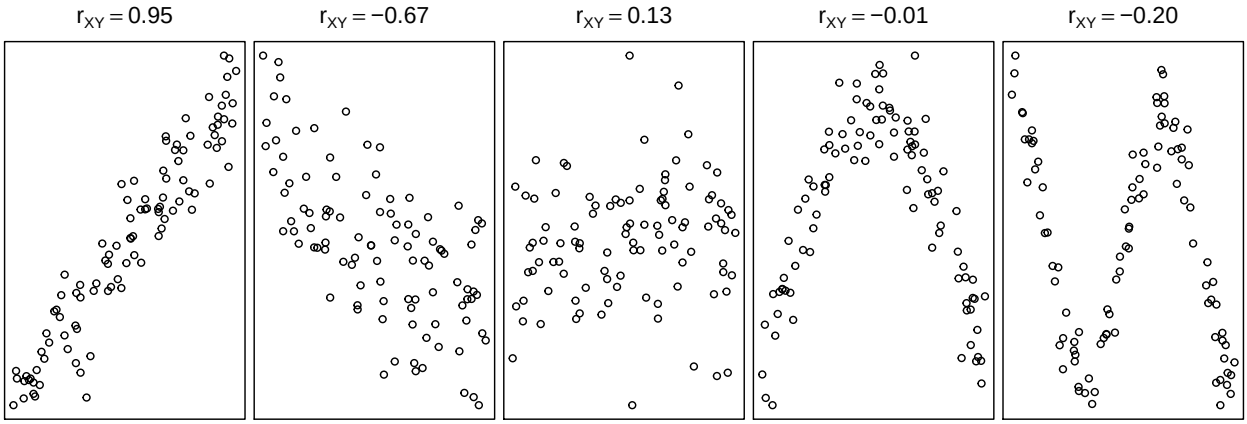
It may not be clear from the formula for r_{xY} what this quantity is supposed to tell us. One fact about this value, however, which is also not immediately clear from the formula, is that r_{xY} cannot take a value outside the interval $[-1, 1]$. Let's put this fact in a big box:

Result: Possible values for Pearson's correlation coefficient

Pearson's correlation coefficient r_{xY} must take a value in $[-1, 1]$.

Back to looking at the formula: If we look for a moment at the numerator of r_{xY}, we see that if x_i and Y_i tend at the same time to exceed their respective means $\bar{x}_n$ and $\bar{Y}_n$ and tend at the same time to fall below their respective means, the correlation coefficient should take a positive value. However, if x_i tends to be below its mean while Y_i is above its mean and above its mean when Y_i is below its mean, then r_{xY} will probably take a negative value. If x_i and Y_i tend to fall above and below their respective means without regard to one another, then r_{xY} does not know whether to be positive or negative and so takes a value close to zero.

The plot below shows a few scatterplots along with the value of Pearson's correlation coefficient computed on the data.



We may notice a few things:

1. The value of r_{xY} is close to zero when the relationship is weak (middle panel.
2. The value of r_{xY} is close to zero when the relationship is nonlinear (two rightmost panels).
3. The value of r_{xY} is positive when the scatterplot slopes upward (two leftmost panels).
4. The value of r_{xY} is negative when the scatterplot slopes upward (two leftmost panels).
5. the value of r_{xY} is greater *in magnitude* the more closely the points are scattered around a straight line (three leftmost panels).

As a summary, we may say that Pearson's correlation coefficient describes the strength and direction of *linear* relationships. Very strong non-linear relationships, like the ones in the two rightmost panels of the plot, do not correspond to large values of Pearson's correlation coefficient. Therefore, the quantity r_{xY} is only appropriate for describing linear relationships. If a scatterplot suggests a non-linear relationship—if you cannot draw a straight line through the data points—then it is not appropriate to use Pearson's correlation coefficient to describe the relationship.

In statistics, the word "correlation" almost always means Pearson's correlation, so a statistician would never say something like "what you are saying is correlated with what so and so is saying." A statistician will not use the word "correlated" unless it is to describe the strength and direction of the linear relationship between two variables.

The following code generates some data $(x_1, Y_1), \ldots, (x_n, Y_n)$ and computes Pearson's correlation coefficient between x and Y.

```
n <- 100
X <- runif(n,0,10)
Y <- X + rnorm(n)
cor(X,Y)
```

Simple linear regression

The simple linear regression model is a mathematical description of how Y is related to X. For data $(x_1, Y_1), \ldots, (x_n, Y_n)$ it assumes

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

for $i = 1, \ldots, n$, where

- $x_1, \ldots, x_n$ are fixed real numbers
- $Y_1, \ldots, Y_n$ are independent random variables
- β_0 and β_1 are unknown constants called *regression coefficients*
- $\varepsilon_1, \ldots, \varepsilon_n$ are iid *errors* with $\mathbb{E}\varepsilon_i = 0$ and $\operatorname{Var}\varepsilon_i = \sigma^2$ for $i = 1, \ldots, n$.

The model states that for the value x_i, the value of Y_i will be equal to the height of the line $\beta_0 + \beta_1 x_i$ plus or minus some random amount. This random amount is often called an *error*, but there is nothing really erroneous going on; we just don't expect all the Y values to fall exactly on the line $\beta_0 + \beta_1 x$, so we allow them to be different by the amount ε, and we assume that the average of many values of ε will be zero. All of this means we expect the Y values to bounce more or less around the line, falling above and below it at random.

Having written down the linear regression model, the next task for the statistician is to use the data to estimate the unknown quantities involved. There are three: The parameters β_0 and β_1 giving the intercept and slope of the line, and the parameter σ^2 giving the variance of the error terms.

We first focus on estimating β_0 and β_1. Reasonable guesses at the values of β_0 and β_1 might be obtained by drawing, with a ruler, a line through the scatterplot which runs as much as possible through the center of the points. Although this may give reasonable results, two different people might draw two different lines, and it would hardly possible to analyze the statistical properties of such a line. We would like to have a way of drawing the line such that we can depend on its on-average being the right one, or in the sense that if we collected more and more data our method of line-drawing would eventually lead us to the true line governing the true relationship between Y and x. To this end, we will introduce a criterion for judging the quality of any line drawn through the points and use calculus to find the slope and intercept values which produce the best possible line—according to the criterion.

The most commonly used criterion for judging the quality of straight lines drawn through scatterplots is called the *least-squares criterion.* It is equal to the sum of squared vertical distances between the data points on the scatterplot and the line drawn. If the line $a + bx$ is drawn, for some slope b and intercept a, the least squares criterion is the sum of $Y_i - (a + bx_i)$, $i = 1, \ldots, n$. Under this criterion, one can find the best choices of a and b. We present the result in the next box:

Result: Least squares estimators of regression coefficients

Provided $\sum_{i=1}^n (x_i - \bar{x}_n)^2 > 0$, the function

$$Q_n(\beta_0, \beta_1) := \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 x_i)]^2$$

is (uniquely) minimized at

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{x}_n$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} = r_{xY} \cdot \frac{s_Y}{s_x},$$

where $s_Y^2 = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ and $s_x^2 = (n-1)^{-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$.

So, to compute the least-squares estimators of the regression coefficients β_0 and β_1, one first obtains the slope as $\hat{\beta}_1 = r_{xY} s_Y / s_x$ and then the intercept as $\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{x}_n$.

Example. We can plot the beryllium data and overlay the least-squares regression line using the following R code:

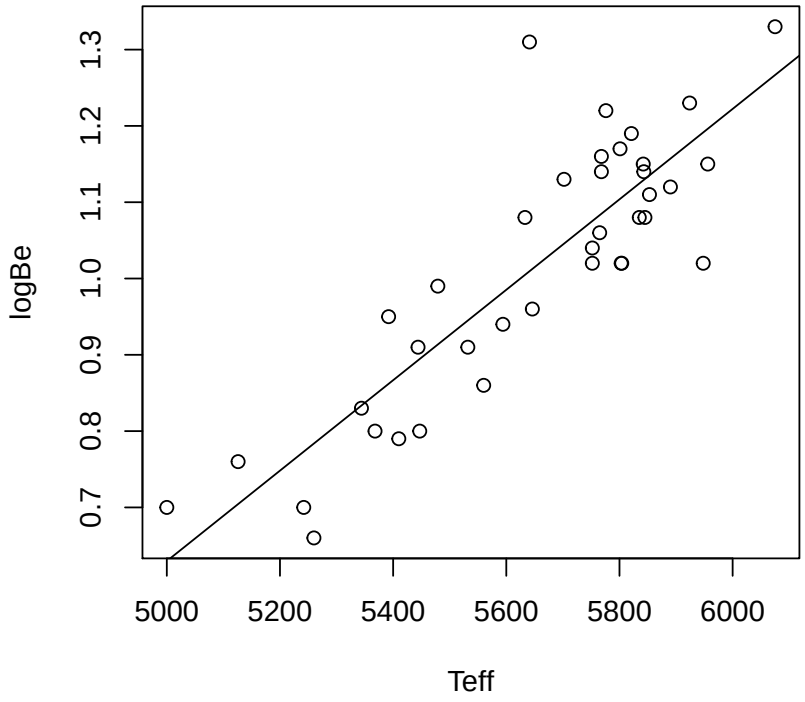
```
# load the data
load(url("https://people.stat.sc.edu/gregorkb/data/beryllium.Rdata"))

# pull x and Y from the beryllium data frame
x <- beryllium$Teff
Y <- beryllium$logN_Be

# compute the least-squares regression coefficients
x_bar <- mean(x)
b1 <- cor(x,Y) * sd(Y) / sd(x)
b0 <- mean(Y) - b1*x_bar

# make a scatterplot with the least-squares line overlaid
plot(Y ~ x , xlab="Teff", ylab = "logBe")
abline(b0,b1)
```

The above code produces the plot



Statistical inference in simple linear regression

It is nice to be able to draw the best line through a scatterplot of points. We know that the least-squares line is, of any line we could possibly draw, the one that minimizes the sum of the squared vertical distances between the points and itself.

Remember the whole idea behind statistics, though? If we were to collect another data set studying the same phenomenon, we would get different data; so what can we learn from this one single data set? The line we drew through the scatterplot is really a "guess" at where the true line would be. If we could go on collecting data on all the stars in the universe, we could then draw the "true" or "population" level line. But we only have the data in our sample.

So what can we learn from it?

As we have done before with population means and variances and proportions, we will formulate some hypotheses about the true best line underneath the data and use the least-squares line to test those hypotheses. Fundamentally, we are interested in knowing whether there exists a linear relation between our X and our Y variable. The true best line is specified

by the unknown values of β_0 and β_1. Of these parameters, it is β_1 which carries information about the relationship between X and Y. We therefore concetrate on making statistical inferences about β_1; that is, we try to learn what we can about β_1 from the data.

What we can learn about β_1 from the data depends on the behavior of out estimator $\hat{\beta}_1$– specifically how far away from the true β_1 we should expect it to be. This information is contained in the sampling distribution of $\hat{\beta}_1$, which we shall present after giving names to some important quantities that arise after we draw the least-squares line.

Definition: Fitted values, residuals

Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the least-squares estimators of β_0 and β_1. Then the values

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \ldots, n$$

are called the *fitted values*, and the values

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i, \quad i = 1, \ldots, n$$

are called the *residuals*.

Having defined the fitted values and the residuals, we may present an estimator for the other unknown quantity in the linear regression model, which is the error term variance σ^2. We are going to have to take a guess at what this value is in order to make any inferences about β_1 based on the data, and the estimator defined next is the best guess we can make ("best" according to some criteria that are beyond the scope of this course!).

Definition: Estimator of the error term variance

An unbiased estimator of the error term variance σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

The word "unbiased" means $\mathbb{E}\hat{\sigma}^2 = \sigma^2$, so that the sampling distribution of the estimator $\hat{\sigma}^2$ is centered in a nice way around the value it is trying to estimate.

Now we present a result about the sampling distribution of $\hat{\beta}_1$:

Result: Sampling distribution of $\hat{\beta}_1$

Provided $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2)$, we have

$$\hat{\beta}_1 \sim \text{Normal}(\beta_1, \sigma^2/S_{xx}) \quad \text{and} \quad (n-2)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2},$$

where $S^2_{xx} = \sum_{i=1}^n (x_i - \bar{x}_n)^2$, from which we have

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{n-2}.$$

We can use the above to construct a $(1-\alpha) \times 100\%$ confidence interval for β_1.

Result: Confidence interval for β_1

If $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2)$, a $(1-\alpha) \times 100\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2}\hat{\sigma}/\sqrt{S_{xx}}.$$

Example. Continuing the example with the beryllium data, we can build a 95% confidence interval for β_1 with the following code:

```
n <- length(Y)
Sxx <- sum((x - x_bar)^2)
sigma.hat <- sqrt( sum(e_hat^2)/(n-2))

lo <- b1 - qt(.975,n-2) * sigma.hat / sqrt(Sxx)
up <- b1 + qt(.975,n-2) * sigma.hat / sqrt(Sxx)
```

The confidence interval is $(0.00047, 0.00071)$.

We can also use the `confint()` function on the output of the `lm()` function, like this:

```
confint(lm(Y ~ x))
```

It will print a confidence interval for β_0 as well as β_1.

Assumptions of simple linear regression

Implicit in the setup of the simple linear regression model are the assumptions listed here:

Assumption: Assumptions of simple linear regression

(A.1) The responses are Normally distributed around the regression line.

To check: Look at a QQ plot of the residuals.

(A.2) The responses have the same variance for all values of the covariate.

To check: Look at the residuals versus fitted values plot.

(A.3) The covariate and the response are linearly related.

To check: Look at the residuals versus fitted values plot.

(A.4) The responses are independent from each other.

Cannot check: Trust the experimental design/beyond scope of course.

I haven't quite finished typing the rest of this note. See the lecture slides!!

References

[1] Nuno C Santos, G Israelian, RJ García López, M Mayor, R Rebolo, S Randich, A Ecuvillon, and C Domínguez Cerdeña. Are beryllium abundances anomalous in stars with giant planets? *Astronomy & Astrophysics*, 427(3):1085–1096, 2004.