Introduction to Linear Regression

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Introduction to Linear Regression

Introduction

In this module, we discuss an extremely important technique in statistics — Linear Regression.

Linear regression is very closely related to correlation, and is extremely useful in a wide range of areas.

Introduction

We begin by loading some data relating height to shoe size and drawing the scatterplot for the male data.

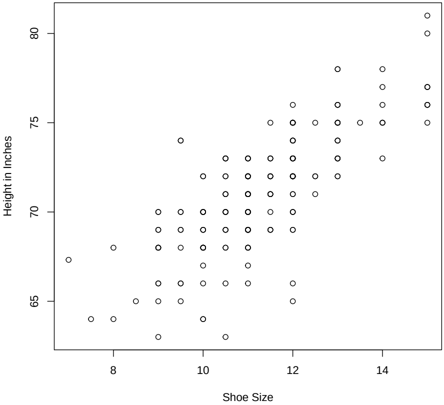
```
> all.heights <- read.csv("shoesize.csv")
```

```
> male.data <- all.heights[all.heights$Gender == "M", ]  #Select males
> attach(male.data)  #Make Variables Available
```

Introduction

Next, we draw the scatterplot. The points align themselves in a linear pattern.

```
> # Draw scatterplot
> plot(Size, Height, xlab = "Shoe Size", ylab = "Height in Inches")
```



Introduction

```
> cor(Size, Height)

[1] 0.7677
```

The correlation is an impressive 0.77. But how can we characterize the relationship between shoe size and height?

In this case, linear regression is going to prove very useful.

Fitting a Straight Line

Introduction

If data are scattered around a straight line, then the relationship between the two variables can be thought of as being represented by that straight line, with some “noise” or error thrown in.

We know that the correlation coefficient is a measure of how well the points will fit a straight line. But *which straight line is best*?

Fitting a Straight Line

Introduction

The key to understanding this is to realize the following:

1. Any straight line can be characterized by just two parameters, a *slope* and an *intercept*, and the equation for the straight line is $Y = b_0 + b_1 Xa$, where b_1 is the slope and b_0 is the intercept.
2. Any point can be characterized *relative to a particular line* in terms of two quantities: (a) where its X falls on a line, and (b) how far its Y is from the line in the vertical direction.

Let's examine each of these preceding points.

Fitting a Straight Line

Characteristics of a Straight Line

The slope multiplies X, and so any change in X is multiplied by the slope and passed on to Y. Consequently, the slope represents "the rise over the run," the amount by which Y increases for each unit increase in X.

The intercept is, of course, the value of Y when $X = 0$.

So if you have the slope and intercept, you have the line.

Fitting a Straight Line

Characteristics of a Straight Line

Suppose we draw a line — any line — in a plane.

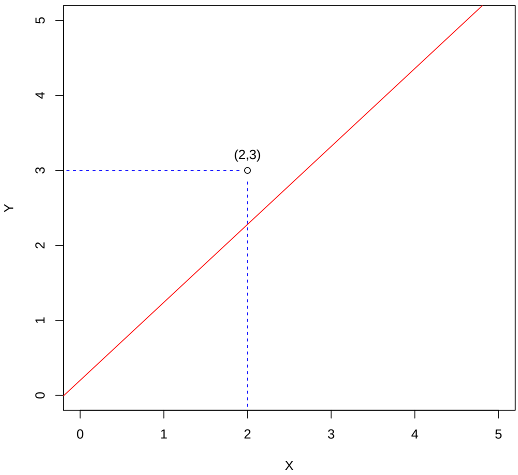
Then consider a point — any point — with respect to that line.

What can we say? Let's use a concrete example.

Suppose I draw the straight line whose equation is $Y = 1.04X + 0.2$ in a plane, and then plot the point $(2, 3)$ by going over to 2 on the X-axis, then up to 3 on the Y-axis.

Fitting a Straight Line

Characteristics of a Straight Line



Fitting a Straight Line

Characteristics of a Straight Line

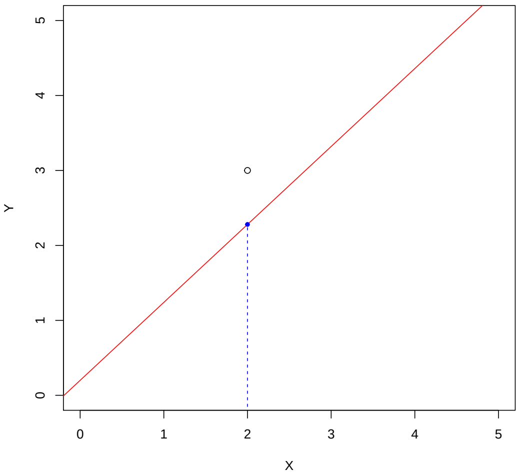
Now suppose I were to try to use the straight line to predict the Y value of the point only from a knowledge of the X value of that point.

The X value of the point is 2. If I substitute 2 for X in the formula $Y = 1.04X + 0.2$, I get $Y = 2.28$.

This value lies on the line, directly above X. I'll draw that point on the scatterplot in blue.

Fitting a Straight Line

Characteristics of a Straight Line



Fitting a Straight Line

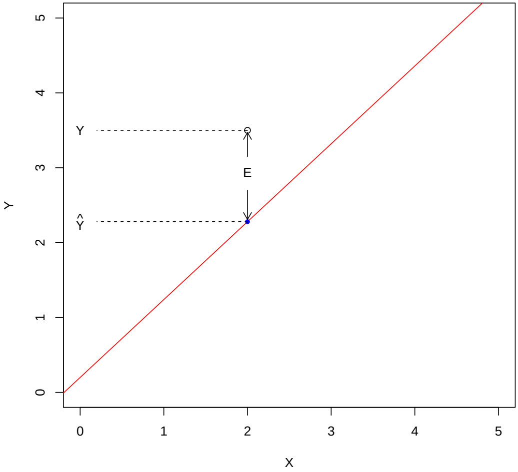
Characteristics of a Straight Line

The Y value for the blue point is called the "predicted value of Y," and is denoted $\hat{Y}$.

Unless the actual point falls on the line, there will be some error in this prediction. The error is the discrepancy in the vertical direction from the line to the point.

Fitting a Straight Line

Characteristics of a Straight Line



Fitting a Straight Line

Regression Notation

Now, let't generalize!

We have just shown that, for *any* point with coordinates (X_i, Y_i), relative to *any* line $Y = b_0 + b_1 X$, I may write

$$\hat{Y}_i = b_0 + b_1 X_i \quad (1)$$

and

$$Y_i = \hat{Y}_i + E_i \quad (2)$$

with E_i defined tautologically as

$$E_i = Y_i - \hat{Y}_i \quad (3)$$

But we are not looking for *any* line. We are looking for the *best* line. And we have many points, not just one. And, by the way, what *is* the best line, and how do we find it?

Fitting a Straight Line

The Least Squares Solution

It turns out, there are many possible ways of characterizing how well a line fits a set of points.

However, one approach seems quite reasonable, and has many absolutely beautiful mathematical properties.

This is the *least squares criterion* and the *least squares solution* for b_1 and b_0.

Fitting a Straight Line

The Least Squares Solution

The least squares criterion states, *the best-fitting line for a set of points is that line which minimizes the sum of squares of the E_i for the entire set of points.*

Remember, the data points are there, plotted in the plane, nailed down, as it were. The only thing free to vary is the line, and it is characterized by just two parameters, the slope and intercept.

For any slope b_1 and intercept b_0 I might choose, I can compute the sum of squared errors. And for any data set, the sum of squared errors is uniquely defined by that slope and intercept.

The sum of squared errors is thus a *function* of b_1 and b_0.

What we really have is a problem in minimizing a function of two unknowns.

This is a routine problem in first-year calculus. We won't go through the proof of the least squares solution, we'll simply give you the result.

Fitting a Straight Line

The Least Squares Solution

The solution to the least squares criterion is as follows

$$b_1 = r_{y,x}\frac{s_y}{s_x} = \frac{s_{y,x}}{s_x^2} \tag{4}$$

and

$$b_0 = M_y - b_1 M_x \tag{5}$$

Note: If X and Y are both in Z score form, then $b_1 = r_{y,x}$ and $b_0 = 0$.

Thus, once we remove the metric from the numbers, the very intimate connection between correlation and regression is revealed!

Predicting Height from Shoe Size

Creating a Fit Object

We could easily construct the slope and intercept of our regression line from summary statistics. But R actually has a facility to perform the entire analysis very quickly and automatically. You begin by producing a *linear model fit object* with the following syntax.

```
> fit.object <- lm(Height ~ Size)
```

R is an *object oriented language*. That is, objects can contain data and when general functions are applied to an object, the object "knows what to do." We'll demonstrate on the next slide.

Predicting Height from Shoe Size

Examining Summary Statistics

R has a generic function called `summary`. Look what happens when we apply it to our fit object.

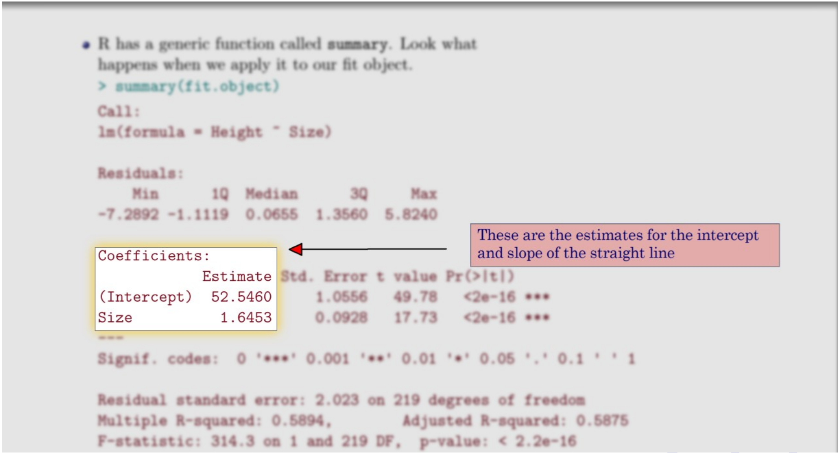
```
> summary(fit.object)


Call:
lm(formula = Height ~ Size)

Residuals:
   Min     1Q Median     3Q    Max
-7.289 -1.112  0.066  1.356  5.824

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.5460     1.0556    49.8   <2e-16 ***
Size          1.6453     0.0928    17.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.02 on 219 degrees of freedom
Multiple R-squared:  0.589, Adjusted R-squared:  0.588
F-statistic:  314 on 1 and 219 DF,  p-value: <2e-16
```

Predicting Height from Shoe Size

Examining Summary Statistics

The coefficients for the intercept and slope are perhaps the most important part of the output.

Here we see that the slope of the line is 1.6453 and the intercept is 52.5460.



- R has a generic function called **summary**. Look what happens when we apply it to our fit object.

```
> summary(fit.object)

Call:
lm(formula = Height ~ Size)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2892 -1.1119  0.0655  1.3560  5.8240

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.5460     1.0556   49.78   <2e-16 ***
Size          1.6453     0.0928   17.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.023 on 219 degrees of freedom
Multiple R-squared: 0.5894,      Adjusted R-squared: 0.5875
F-statistic: 314.3 on 1 and 219 DF,  p-value: < 2.2e-16
```

These are the estimates for the intercept and slope of the straight line

Predicting Height from Shoe Size

Examining Summary Statistics

Along with the estimates themselves, the program provides estimated standard errors of the coefficients, along with t statistics for testing the hypothesis that the coefficient is zero.

- R has a generic function called **summary**. Look what happens when we apply it to our fit object.

```
> summary(fit.object)

Call:
lm(formula = Height ~ Size)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2892 -1.1119  0.0655  1.3560  5.8240

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.5460     1.0556   49.78   <2e-16 ***
Size          1.6453     0.0928   17.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.023 on 219 degrees of freedom
Multiple R-squared: 0.5894,     Adjusted R-squared: 0.5875
F-statistic: 314.3 on 1 and 219 DF,  p-value: < 2.2e-16
```

Here we have the standard errors, t-values, p-values, and significance codes.

The legend for the codes shows that *** means significant at the 0.001 level.

Predicting Height from Shoe Size

Examining Summary Statistics

The program prints the R^2 value, also known as the coefficient of determination. When there is only one predictor, as in this case, the R^2 value is just $r^2_{x,y}$, the square of the correlation between height and shoe size.

The "adjusted R^2" value is an approximately unbiased estimator. With only one predictor, this can essentially be ignored, but with many predictors, it can be much lower than the standard R^2 estimate.

The F-statistic tests that $R^2 = 0$

When there is only one predictor, it is the square of the t-statistic for testing that $r_{x,y} = 0$.

Predicting Height from Shoe Size

Examining Summary Statistics

```
Call:
lm(formula = Height ~ Size)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2892 -1.1119  0.0655  1.3560  5.8240

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.5460     1.0556   49.78   <2e-16 ***
Size          1.6453     0.0928   17.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.023 on 219 degrees of freedom
Multiple R-squared: 0.5894,        Adjusted R-squared: 0.5875
F-statistic: 314.3 on 1 and 219 DF,  p-value: < 2.2e-16
```

The R-squared value, the Adjusted R-squared, and the F-test of the hypothesis that R-squared = 0

Predicting Height from Shoe Size

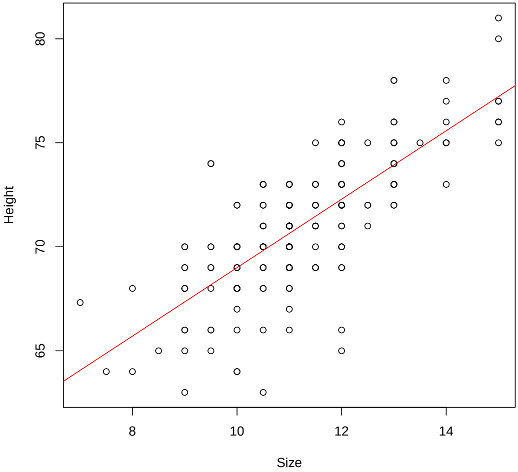
Drawing the Regression Line

Now we draw the scatterplot with the best-fitting straight line. Notice how we draw the scatterplot first with the `plot` command, then draw the regression line in red with the `abline` command.

```
> # draw scatterplot
> plot(Size, Height)
> # draw regression line in red
> abline(fit.object, col = "red")
```

Predicting Height from Shoe Size

Drawing the Regression Line



Predicting Height from Shoe Size

Using the Regression Line

We can now use the regression line to estimate a male student's height from his shoe size.

Suppose a student's shoe size is 13. What is his predicted height?

$$\hat{Y} = b_1X + b_0 = (1.6453)(13) + 52.5460 = 73.9349$$

The predicted height is a bit less than 6 feet 2 inches.

Of course, we know that not every student who has a size 13 show will have a height of 73.93. Some will be taller than that, some will be shorter. Is there something more we can say?

Predicting Height from Shoe Size

Thinking about Residuals

The predicted value $\hat{Y} = 73.93$ actually represents the average height of people with a shoe size of 13.

According to the most commonly used linear regression model, people with a shoe size of 13 actually have a normal distribution with a mean of 73.93, and a standard deviation called the "standard error of estimate."

This quantity goes by several names, and in R output is called the "residual standard error."

An estimate of this quantity is included in the R regression output produced by the `summary` function.

Predicting Height from Shoe Size

Thinking about Residuals

```
Call:
lm(formula = Height ~ Size)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2892 -1.1119  0.0655  1.3560  5.8240

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.5460     1.0556   49.78   <2e-16 ***
Size          1.6453     0.0928   17.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
                                        Residual Standard Error
Residual standard error: 2.023 on 219 degrees of freedom
Multiple R-squared: 0.5894,     Adjusted R-squared: 0.5875
F-statistic: 314.3 on 1 and 219 DF,  p-value: < 2.2e-16
```

Predicting Height from Shoe Size

Thinking about Residuals

In the population, the standard error of estimate is calculated from the following formula

$$\sigma_e = \sqrt{1 - \rho_{x,y}^2}\, \sigma_y \tag{6}$$

In the sample, we estimate the standard error of estimate with the following formula

$$s_e = \sqrt{\frac{n-1}{n-2}}\sqrt{1 - r_{x,y}^2}\, s_y \tag{7}$$

Partial Correlation

An Example

Residuals can be thought of as "The part of Y that is left over after that which can be predicted from X is partialled out."

This notion has led to the concept of *partial correlation*.

Let's introduce this notion in connection with an example.

Suppose we gathered data on house fires in the Nashville area over the past month. We have data on two variables — damage done by the fire, in thousands of dollars (`Damage`) and the number of fire trucks sent to the fire by the fire department (`Trucks`).

Here are the data for the last 10 fires.

Partial Correlation

An Example

	Trucks	Damage
1	0	8
2	0	9
3	1	33
4	1	38
5	1	27
6	2	70
7	2	94
8	2	83
9	3	133
10	3	135

Partial Correlation

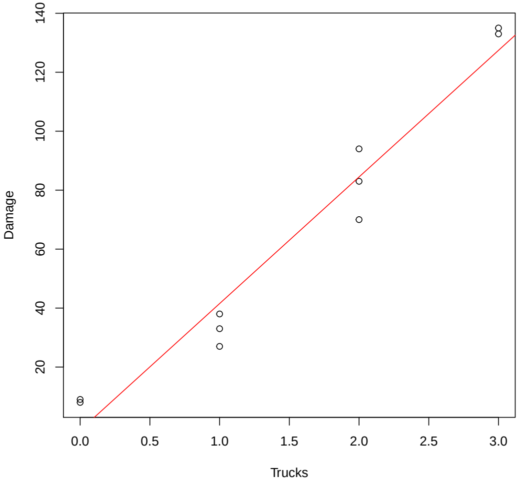
An Example

Plotting the regression line, we see that there is indeed, a strong linear relationship between the number of fire trucks sent to a fire, and the damage done by the fire.

```
> plot(Trucks, Damage)
> abline(lm(Damage ~ Trucks), col = "red")
```

Partial Correlation

An Example



Partial Correlation

An Example

The correlation between `Trucks` and `Damage` is 0.9779.

Does this mean that the damage done by fire can be reduced by sending fewer trucks?

Of course not. It turns out that the house fire records include another piece of information. Based on a complex rating system, each housefire has a rating based on the size of the conflagration. These ratings are in a variable called `FireSize`.

On purely substantive and logical grounds, we might suspect that rather than fire trucks causing the damage, that this third variable, `FireSize`, causes both more damage to be done and more fire trucks to be sent.

How can we investigate this notion statistically?

Partial Correlation

An Example

Suppose we predict `Trucks` from `FireSize`. The residuals represent that part of `Trucks` that isn't attributable to `Firesize`. Call these residuals $E_{\text{Trucks}|\text{FireSize}}$.

Then suppose we predict `Damage` from `Firesize`. The residuals represent that part of Damage that cannot be predicted from `FireSize`. Call these residuals $E_{\text{Damage}|\text{Firesize}}$.

The correlation between these two residual variables is called *the partial correlation* between `Trucks` and `Damage` with `FireSize` partialled out, and is denoted $r_{\text{Trucks},\text{Damage}|\text{FireSize}}$.

Partial Correlation

An Example

There are several ways we can compute this partial correlation.

One way is to compute the two residual variables discussed above, and then compute the correlation between them.

```
> fit.1 <- lm(Trucks ~ FireSize)
> fit.2 <- lm(Damage ~ FireSize)
> E.1 <- residuals(fit.1)
> E.2 <- residuals(fit.2)
> cor(E.1, E.2)

[1] -0.2163
```

Partial Correlation

An Example

Another way is to use the textbook formula

$$r_{x,y|w} = \frac{r_{x,y} - r_{x,w}r_{y,w}}{\sqrt{(1 - r^2_{x,w})(1 - r^2_{y,w})}} \tag{8}$$

```
> r.xy <- cor(Trucks, Damage)
> r.xw <- cor(Trucks, FireSize)
> r.yw <- cor(Damage, FireSize)
> r.xy.given.w <- (r.xy - r.xw * r.yw)/sqrt((1 - r.xw^2) * (1 - r.yw^2))
> r.xy.given.w

[1] -0.2163
```

Partial Correlation

An Example

The partial correlation is -0.216.

Once size of fire is accounted for, there is a negative correlation between number of fire trucks sent to the fire and damage done by the fire.

Review Questions

Question 1

Recall that $\hat{Y} = b_1 X + b_0$, with $b_1 = \rho_{YX}\sigma_Y/\sigma_X$.

The predicted scores in $\hat{Y}$ have a variance. Prove, using the laws of linear transformation, that this variance may be calculated as

$$\sigma^2_{\hat{Y}} = \rho^2_{YX}\sigma^2_Y \tag{9}$$

Review Questions

Question 1

Answer. The laws of linear transformation state that, if $Y = aX + b$, then $S_Y^2 = a^2 S_X^2$. In other words, additive constants can be ignored, and multiplicative constants "come straight through squared" in the variance.

Translating this idea to the Greek letter notation of the current problem, we find

$$
\begin{aligned}
\sigma_{\hat{Y}}^2 &= \sigma_{b_1 X + b_0}^2 && (10)\\
&= b_1^2 \sigma_X^2 && (11)\\
&= \left(\frac{\rho_{YX} \sigma_Y^2}{\sigma_X^2} \right) \sigma_X^2 && (12)\\
&= \rho_{YX}^2 \sigma_Y^2 && (13)
\end{aligned}
$$

Review Questions

Question 2

In the lecture notes on linear combinations, we demonstrate that the covariance of two linear combinations may be derived by taking the algebraic product of the two linear combination expressions, and then applying a straightforward conversion rule.

Using this approach, show that the covariance between Y and $\hat{Y}$ is equal to the variance of $\hat{Y}$ derived in the preceding problem.

Hint. The covariance of Y and $\hat{Y}$ is equal to the covariance of Y and $b_1 X + b_0$.

Review Questions

Question 2

Answer. $\sigma_{Y,b_1X+b_0} = \sigma_{Y,b_1X}$ because additive constants never affect covariances. Now, applying the multiplicative rule, we find that

$$\begin{aligned} \sigma_{Y,b_1X} &= b_1\sigma_{Y,X} && (14) \\ &= \frac{\rho_{Y,X}\sigma_Y}{\sigma_X}\rho_{Y,X}\sigma_Y\sigma_X && (15) \\ &= \rho^2_{Y,X}\sigma^2_Y && (16) \end{aligned}$$

Review Questions

Question 3

Sarah got a Z score of $+2.00$ on the first midterm. If midterm 1 and midterm 2 are correlated 0.75 and the relationship is linear, what is the predicted Z-score of Sarah on the second exam?

Review Questions

Question 3

Answer. When scores are in Z-score form, the formula for a predicted score is $\hat{Y} = \rho_{Y,X} X$, and so Sarah's predicted score is $(2)(0.75) = 1.50$. This is a classic example of regression toward the mean.