Simple Linear Regression

Inference for the Slope

Linear relationships between two variables can be strong or weak, depending on how much of the variability the regression line explains. We'll now look at answering the question: *Is a linear relationship statistically significant?*

EXAMPLE 1

An anthropologist is interested in how accurately she can predict age at death from human skeletal remains. She examined 400 skeletons, and estimated the age at death for each skeleton. Our variable of primary interest is

$$\text{error in age estimation} = \text{estimated age - actual age.}$$

We'll consider how the error in age estimation is related to body mass index, or BMI.

Before investigating the relationship between BMI and error in age estimation, we need to decide if one of these variables is playing the role of response variable, and the other the role of predictor variable. Since we interested in learning how BMI affects the error in age estimation it's clear that BMI is our predictor and error in age estimation is our response.

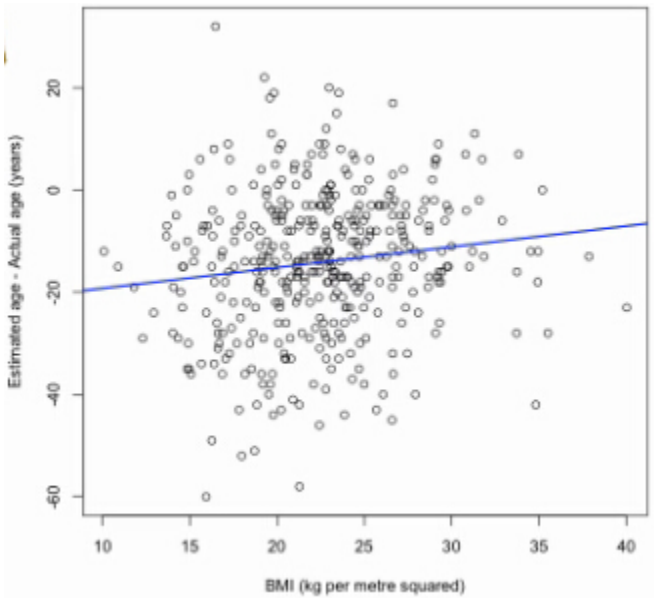

Figure 1: Scatterplot of error in age estimation versus BMI (Example 1).

The scatterplot of error in age estimation versus BMI is pretty noisy and it's not clear if

there's a relationship. If we use the method of least squares to estimate the equation of line to describe this relationship, we get

$$\widehat{\text{error in age estimation}} = 23.28 + 0.41 \times \text{BMI}$$

For this regression line, R^2 is 0.019. So only 1.9% of the variation in the response variable is being explained by its linear relationship with BMI. So is there really a relationship between error in age estimation and BMI? Or could the fact that we got a positive slope here just be due to chance? We can answer this question by carrying out inference for the slope of the regression line.

Let's review the process of statistical inference. In the real world, we have some observed data. And in the theoretical world, we have some models which may be useful in describing the real world. In regression, our model is the straight line model, to describe the relationship between the response and the predictor variable. The statistical models include the variation that we naturally have in our observed data. All of our data won't lie perfectly on the straight line. It will be randomly scattered about it. To account for this noise, we add an error term to the straight line. So our model is

$$y = \beta_0 + \beta_1 x + \epsilon$$

β_0 is a parameter. It is the theoretical world value of the intercept. Similarly β_1 is our slope parameter. We can never observe these parameters, but we can get estimates for them from our data. We've already seen how to estimate the intercept and slope using the method of least squares, and we call these estimates b_0 and b_1. ϵ is our random error giving how much each point differs from the line in the vertical direction.

Our model says that y and x are related by a linear relationship. Our data points are randomly scattered about that line, with that random scatter captured by the ϵ in the model. We typically model the random scatter of the ϵ's as having a normal distribution, so for a given value of x, we get a distribution of values of y, centred on the linear model, with a bell-shape, with the same standard deviation., regardless of the value of x. We estimate the β's from our data, and get the b's, and we'd like to infer what we can about what the β's really are, from what we've observed.

The usual question in regression is: *Is* $\beta_1 = 0$*?* If it is, our model is a horizontal line and the line gives the same value of y for every value of x. Then we'd say that there is no linear relationship between y and x. If our goal was to understand if x and y were linearly related, we'd say they weren't. If our goal was to use x to predict y, we'd say it's not a good predictor of y.

We're now in the usual position for statistical inference where we want to make conclusions about the parameter in the theoretical world, β_1, the slope of our model. The inferences will be based on our observed estimate of β_1, which is b_1. b_1 is a statistic, as it is calculated from our data. For every sample of data we might observe, we'll get a slightly different value of this statistic, and the different values it could be, in all possible samples, has a distribution

which is its sampling distribution. Our theoretical world model assumed that our data were scattered about a line, with the scatter determined by a probability distribution. We'll assume that this probability distribution is normally distributed. If the errors are normally distributed, then our observed values of the response variables and estimators calculated from them are also normally distributed.

b_1 is an unbiased estimator of β_1, so $\mathrm{E}(b_1) = \beta_1$. For its standard devation., we'll estimate it from the data by statistical software. The estimate of the standard deviation of the sampling distribution of is called its **standard error**, which we'll denote as S.E.(b_1).

Let's now carry out a statistical test for the slope. If it is 0, we have evidence that there is actually no linear relationship between our y and our x.

$$\text{Hypotheses: } H_0 : \beta_1 = 0 \text{ versus } H_A : \beta_1 \neq 0,$$

using a 2-sided alternative. Test statistic:

$$\frac{\text{estimate}-H_0 \text{ value}}{\text{S.E.(estimate)}} = \frac{b_1 - 0}{\text{S.E.}_{(b_1)}} = \frac{b_1}{\text{S.E.}_{(b_1)}}$$

Once we have our value of the test statistic, and assuming that the null hypothesis is true, that is the true slope is 0, we calculate our P-value, the probability of observing the value of the test statistic we got, or a value more extreme. The test statistic has a t-distribution, since we estimated the standard of b_1 by its standard error, that is the value we estimate using our data. And in this case the t-distribution has $n - 2$ degrees of freedom where n is the number of data values we have and we lose 2 degrees of freedom because we had to estimate 2 parameters, the slope and the intercept.

Going back to our relationship between error in age estimation and BMI of our skeletons (Example 1), we're not sure if this observed linear relationship is just due to chance so we'll carry out a test of

$$H_0 : \beta_1 = 0 \text{ versus } H_A : \beta_1 \neq 0,$$

$b_1 = 0.41$, $S.E.(b_1) = 0.15$, which gives us a test statistic $t = \frac{0.41}{0.15} = 2.7$. $n = 400$, so $df = n - 2 = 398$. Our p-value will be calculated from a t-distribution with 398 df. And since we're carrying out a 2-sided test, our P-value will be

$$P(T > 2.7) + P(T < -2.7) = 0.006 \text{ where T has a } t\text{-distribution with 398 degrees of freedom}$$

(using statistical software). So although our data seemed quite noisy, there is still quite strong evidence that our slope is something other than 0 and we can conclude that there is a linear relationship between error in age estimation and BMI.

In order to get a range of plausible values for what the slope might be, for an idea of how accurate our estimate of the slope is, we can construct a confidence interval for the slope. A $100(1 - \alpha)\%$ confidence interval for a parameter often has the form

$$\text{estimate} \pm \text{critical value} \times \text{S.E.(estimate)}$$

In regression, a $100(1-\alpha)\%$ confidence interval for β_1 is

$$b_1 \pm t_{n-2,\alpha/2}\text{S.E.}(b_1)$$

Suppose we wanted a 95% confidence interval for the slope for our skeleton example. For this example, $n = 400$ so we need the value with probability 0.025 in the tails from a t-distribution with 398 degrees of freedom, which is 1.966 (using a computer). This gives a 95% confidence interval for the slope:

$$0.41 \pm 1.966 \times 0.15 = (0.12, 0.71)$$

Note that this confidence interval doesn't include 0, so we're again quite confident that there is a linear relationship.

Example 2
For the linear relationship between average crawling age for babies and temperature, we can carry out a test of

$$H_0 : \beta_1 = 0 \text{ versus } H_A : \beta_1 \neq 0$$

The relevant statistics are

$$b_1 = 0.14, \text{ S.E.}(b_1) = 0.045, \text{ test statistic} = -3, \text{ degrees of freedom} = 12 - 2 = 10$$

Resulting in P-value = 0.011. So we have moderate evidence that the slope is not 0.

Here we were more confident that there is a linear relationship from the plot than we were for the noisy skeleton data, but there are two factors that result in the fact that the evidence against a zero slope is only moderate for the babies crawling data:

1. We had one point that didn't fit the data well, which contributes a lot to our estimate of the variability of our errors, making it larger, and thus making the standard error of the estimated slope larger.
2. Our sample size is only 12, so we have less power to detect differences from 0 in the slope than we did with the much larger sample size in the skeleton example.

We now have methodology for carrying out inference on the slope of the regression line. Inference on the intercept ollows the same pattern.

We should always be careful about how strongly we can state our conclusions. Just because there is a statistically significant relationship between y and x, that doesn't mean that changes in x **cause** y to change. Do lower temperatures cause babies to take longer to learn to crawl, or could it be other factors associated with temperature such as time spent outdoors or type of clothing worn that results in the relationship?

There are many reasons we may see evidence of a linear relationship between two variables:

- Change in the explanatory variable cause changes in the response variable.
- Changes in the response variable cause changes in the explanatory variable.
- The explanatory variable contributes to, but is not the only cause of, the response variable.
- There is a third variable confounding the relationship.
- There is a common cause.
- Both variables are changing over time.
- The relationship is just a coincidence.