The simple linear regression model for a paired data set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

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

We **assume** the ε_i are a *random sample* from a mean 0 normal distribution with variance σ^2.

We can simulate data according to this model. First let's set some parameters:

```
> beta0 <- 0
> beta1 <- 2
> sigma <- 1
```

We will do 100 simulated data points. First we define places to store the data:

```
> y <- numeric(100)
```

Next we define our x values:

```
> x <- rep(1:10, 10)
```

(Look at what x is to see what that did.) Now we simulate in a for loop:

```
> for (i in 1:100) {
+     y[i] <- beta0 + beta1 * x[i] + rnorm(1, mean = 0, sd = sigma)
+ }
```

Then we plot with a regression line

```
> plot(y ~ x)
> abline(lm(y ~ x))
```

[This can be simplified typing wise with

```
> y <- rnorm(100, mean = beta0 + beta1 * x, sd = sigma)
```

]

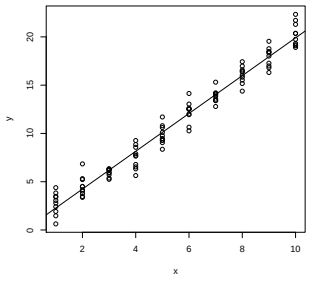

Figure 1: Plot of simulated data. Notice for each fixed x the y values are a normal sample with mean depending on x.

Question 0.1. Run two simulations. One with `sigma=1` and another with `sigma=5`. Comment on the differences and similarities between the two graphs you get.

For inference there are 3 possible parameters to estimate from the data: β_0, β_1, and σ. I'll use hats to indicate the estimates based on the sample, e.g. $\hat{\beta}_1$ is the estimate for the unknown value of β_1.

For the β's we have the following fact about the *sampling distribution*

$$T = \frac{\text{observed} - \text{expected}}{\mathsf{SE}} = \frac{\hat{\beta} - \beta}{\mathsf{SE}}$$

has a t distribution with $n-2$ degrees of freedom. (Just as with many other statistics we have encountered.)

This fact allows us to do

1. Confidence intervals for either β_0 or β_1:

$$\hat{\beta}_i \pm t^* \mathsf{SE}(\hat{\beta}_i)$$

2. Significance tests about either use T as a test statistic.

Now, how to find the estimates and the standard errors using the computer?

Let's look at the data stored in the `fat` data set. First download it:

```
> source("http://wiener.math.csi.cuny.edu/st/R/fat.R")
```

The data set has many variables, including measurements of neck and wrist for different people. The `names` functions lists them all

```
> names(fat)
```

```
 [1] "case"          "body.fat"      "body.fat.siri" "density"
 [5] "age"           "weight"        "height"        "BMI"
 [9] "ffweight"      "neck"          "chest"         "abdomen"
[13] "hip"           "thigh"         "knee"          "ankle"
[17] "bicep"         "forearm"       "wrist"
```

To regress the neck size on the wrist size means to fit a model with neck size as a *response* variable and wrist size as a *predictor* variable. The `lm` function does the work: (Just put the response on the left of the tilde and the predictor on the right.)

```
> res = lm(neck ~ wrist, data = fat)
> res

Call:
lm(formula = neck ~ wrist, data = fat)

Coefficients:
(Intercept)        wrist
      2.637        1.939
```

The estimate for β_0 (labeled `(Intercept)`) and β_1 (labeled `wrist`) are printed. To get more information we need to ask for it. The summary command is used to do so:

```
> summary(res)

Call:
lm(formula = neck ~ wrist, data = fat)

Residuals:
     Min       1Q   Median       3Q      Max
-7.99799 -1.08890 -0.01920  1.11489  7.05953

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.6370     2.0058   1.315     0.19
wrist         1.9394     0.1099  17.649   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.625 on 250 degrees of freedom
Multiple R-squared: 0.5548,        Adjusted R-squared: 0.553
F-statistic: 311.5 on 1 and 250 DF,  p-value: < 2.2e-16
```

The output contains

1. The "call" which just echoes your command
2. A summary of the Residuals. These have a median of -0.01920. The mean is 0, why?

3. The important summary of the coefficients. This is where the action is. More below

4. A summary of the standard error. The value 1.625 is found from
$$\sqrt{\frac{1}{n-2}\sum e_i^2}$$
Or in R

```
> sqrt(sum(resid(res)^2)/(252 - 2))

[1] 1.625288
```

5. That value of `R-squared` might sound familiar, especially if you type `cor(neck, wrist)^2`

The coefficients have not only the estimates, but also the standard errors computed. Additionally a two-sided significance test for whether β is 0 is performed and summarized with an observed value (`t value`) and a p-value, `Pr(>|t|)`.

From the output we see that the estimate $\hat{\beta}_0 = 2.6370$ is *not* statistically significant from 0. (Why?) Yet the smaller value $\hat{\beta}_1 = 1.9364$ *is* statistically significant from 0.

What about the natural question prompted by the following passage from *Gulliver's Travels* by Jonathan Swift (in the giant's voice)

> Then they measured my right Thumb, and desired no more; for by a mathematical Computation, that twice round the Thumb is once round the Wrist, and so on to the Neck and the Waist, and by the help of my old Shirt, which I displayed on the Ground before them for a Pattern, they fitted me exactly.

That is, in the language of MTH 214, is the data consistent with
$$H_0 : \beta_1 = 2, \quad H_A : \beta_1 \neq 2?$$

Here we can use T as the test statistic. The observed value is

```
> Tobs = (1.9364 - 2)/0.1099
> Tobs

[1] -0.5787079
```

The p-value is computed with the t-distribution and $n-2$ degrees of freedom:

```
> pt(Tobs, df = 250)

[1] 0.2816536
```

That is the difference between the 1.9364 and the hypothesized 2 is not statistically significant.

Okay, your turn.

Question 0.2. Use T to find a 95% confidence interval for β_0 based on $\hat{\beta}_0$. You can use $t^* = 1.96$ (why?) or see what it is yourself with `qt(.975, df = 250)`.

Question 0.3. Is twice around the neck once around the waist? Let's model by the `hip` values using the `waist` values as a predictor.

Do two-sided tests of the following

$$H_0 : \beta_0 = 0$$
$$H_0 : \beta_1 = 2$$

What are your p-values?

Question 0.4. The BMI is a simple measurement of fitness: weight divided by height squared. Larger BMIs mean a person is heavier than taller. The body fat is a much more precise measurement of fitness, but is much harder to measure. Does the BMI predict the body fat well? and if so, how does one convert?

Model the variable `body.fat` using `BMI` as a predictor.

Find a 95% CI for β_0.

Perform a two-sided significance test of

$$H_0 : \beta_1 = 1.5$$

Comment on this proposed relationship between BMI and body fat:

To compute body fat from BMI do the following triple the BMI, divide by 2 and subtract 20.

Question 0.5. Does body fat depend on height? Model body fat using height as a predictor and perform a one-sided significance test of $H_0 : \beta_1 = 0$ against $H_A : \beta_1 < 0$. Is the difference significant at the $\alpha = 0.05$ level?

Question 0.6. For a model of ankle size modeled by wrist size find 95% CIs for β_0 and β_1. Is 0 in the first one? Is 1 in the latter? If so, comment on the summary that wrist and ankle sizes are the same *on average*.