## How well does the model fit?

It is easy to estimate the regression parameters. It is not really too difficult to test if the slope deviates from zero.
Unfortunately, a significant regression does not mean that the model fits the data well.
The degree to which the model fits can be measured several ways, though we will only discuss two of them in detail here.

## Confidence intervals on coefficients

One means to assess fit is to look at the confidence intervals around a regression line.
As you remember, confidence intervals depend upon standard errors.

- upper $= \bar{x} + t_{\alpha,df} \times s_{\bar{x}}$
- lower $= \bar{x} - t_{\alpha,df} \times s_{\bar{x}}$

## Standard error of regression coefficients

The standard error of the regression coefficient is:

$$s_{b_1} = \sqrt{\frac{MS_{\text{error}}}{\Sigma(X - \bar{X})^2}}$$

The degrees of freedom are: $n - 2$

## Conf interval of coefficients

- upper $= b_1 + t_{\alpha,df} \times s_{b_1}$
- lower $= b_1 - t_{\alpha,df} \times s_{b_1}$

## Confidence intervals around $\hat{Y}$

An important question occurs when considering a regression:
How good is the line at predicting y-values?
To answer this one has to calculate the confidence intervals of $\hat{Y}$ for different values of $X$.
Std error:

$$s_{\hat{Y}_i} = \sqrt{MS_{\text{error}} \left[ \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma(X_i - \bar{X})^2} \right]}$$

degrees of freedom $= n - 2$

## Confidence intervals around $\hat{Y}$ cont

The actual confidence limits are:

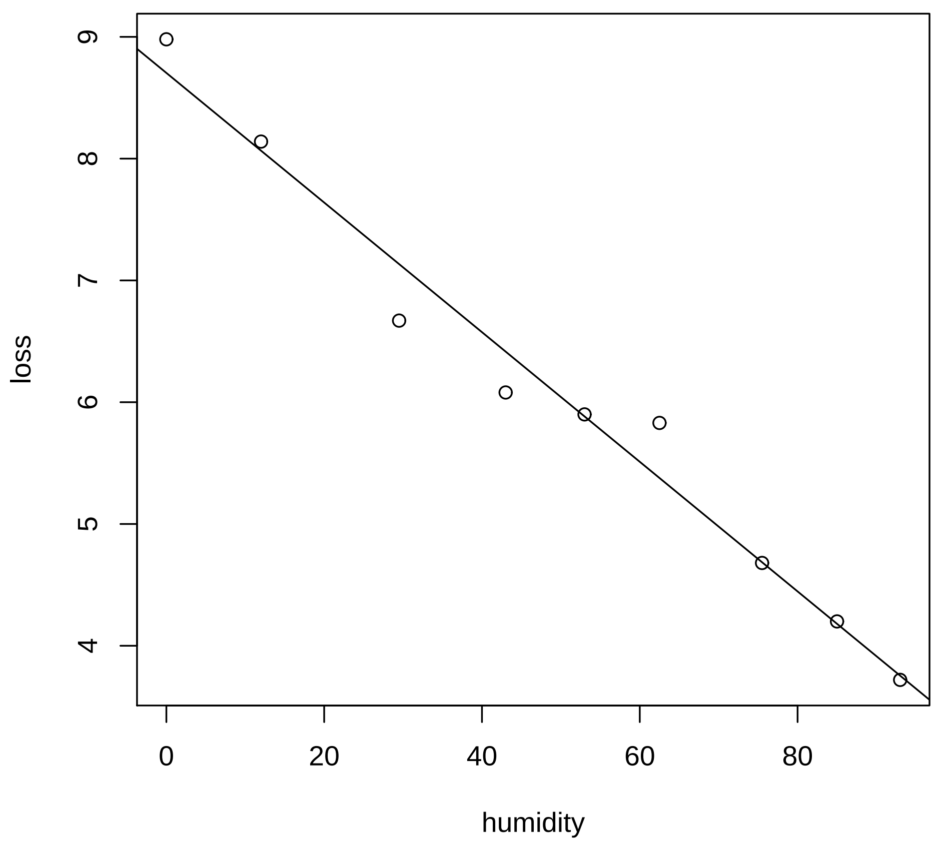
- upper $= \hat{Y}_i + t_{\alpha,df} * s_{\hat{Y}_i}$
- lower $= \hat{Y}_i - t_{\alpha,df} * s_{\hat{Y}_i}$

**Example**

---

This example will use the *Tribolium* data

```
> humidity <- c(0, 12, 29.5, 43, 53, 62.5, 75.5, 85, 93)
> loss <- c(8.98, 8.14, 6.67, 6.08, 5.9, 5.83, 4.68, 4.2, 3.72)
> plot(loss ~ humidity)
> abline(coef(lm(loss ~ humidity)))
```




**Standard error of $b_1$**

---

$$s_{b_1} = \sqrt{\frac{MS_{\text{error}}}{\Sigma(X - \bar{X})^2}}$$

```
> xdev <- humidity - mean(humidity)
> ydev <- loss - mean(loss)
> b1 <- sum((xdev) * (ydev))/sum(xdev^2)
> b0 <- mean(loss) - mean(humidity) * b1
> print(paste(b0, b1))

[1] "8.7040273046679 -0.0532221515810607"

> y.hat <- b0 + b1 * humidity
> ss.explained <- sum((y.hat - mean(loss))^2)
> ss.unexplained <- sum((y.hat - loss)^2)
> stderr.b1 <- sqrt((ss.unexplained/7)/sum(xdev^2))
> stderr.b1

[1] 0.003256028
```

**Conf interval around $b_1$**

---

```
> lower <- b1 - qt(0.95, 7) * stderr.b1
> upper <- b1 + qt(0.95, 7) * stderr.b1
> lower

[1] -0.05939095

> upper

[1] -0.04705335
```

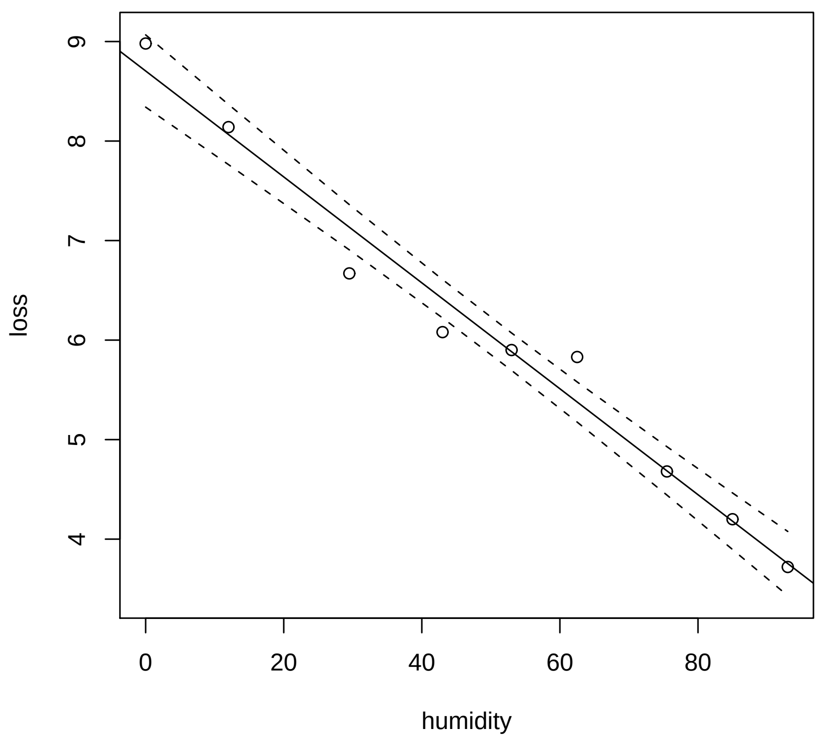
**Conf intervals around $\hat{Y}$**

---

```
> std.yhat <- sqrt((ss.unexplained/7) * ((1/9) + (xdev^2/sum(xdev^2))))
> std.yhat

[1] 0.19156450 0.15938192 0.12001995 0.10177221 0.09925249 0.10646043 0.12831164
[8] 0.14992958 0.17037718

> upper <- y.hat + qt(0.95, 7) * std.yhat
> lower <- y.hat - qt(0.95, 7) * std.yhat
> upper

[1] 9.066961 8.367323 7.361361 6.608290 6.071295 5.579340 4.928851 4.464198
[9] 4.077160
```

```
> lower
[1] 8.341093 7.763400 6.906587 6.222659 5.695212 5.175945 4.442658 3.896091
[9] 3.431574
```

**Plot of error band**




$R^2$

$R^2$ is also known as the proportion of the variance explained by the linear model. The easiest way to calculate this statistic is based upon exactly this idea:

$$R^2 = \frac{SS_{\text{model}}}{SS_{\text{model}} + SS_{\text{error}}}$$

However, $R^2$ is also the square of the correlation coefficient (not the best way to calculate though)
$R^2$ ranges from 0->1. High values indicate a better fit.

**Example**

```
> R.squared <- ss.explained/(ss.explained + ss.unexplained)
> R.squared
```

```
[1] 0.9744696
```

```
> lm.summary <- summary(lm(loss ~ humidity))
> lm.summary$r.squared
```

```
[1] 0.9744696
```

### Adjusted $R^2$

---

The adjusted $R^2$ takes into account the amount of information in the data (ie it depends upon degrees of freedom). In general, you should use adjusted $R^2$.

$$R^2 = \frac{\frac{SS_{\text{model}}}{df_{\text{model}}}}{\frac{SS_{\text{model}}+SS_{\text{error}}}{df_{\text{total}}}}$$

### What to report

---

To be complete, when looking at a regression relationship it is important to report:

- The regression equation
- The results of a test of $H_0$: $b_1 = 0$
- Estimates of error around the coefficients
- $R^2$

### Correlation

---

Correlation measures the association between quantitative variables.
A better way of saying this is that correlation measures the degree to which the variables *co-vary* and presents it on a (-1,1) interval.

### Covariance

---

If correlation depends upon co-variation, it would probably be a good idea to define covariance.
This is not too difficult. Just as sample variance can be defined as a sum of squares divided by degrees of freedom, covariance can be defined as the product of the deviations divided by the degrees of freedom.
Variance:

$$s^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$$

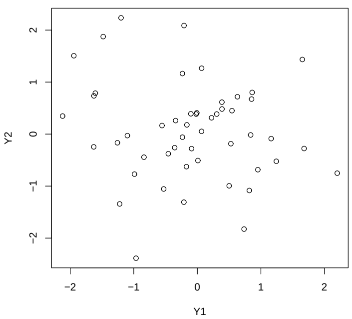
Covariance:

$$s_{12} = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

## What does covariance look like?

First, no covariance.

```
> Y1 <- rnorm(50)
> Y2 <- rnorm(50)
> s12 <- sum((Y1 - mean(Y1)) * (Y2 - mean(Y2)))/49
> s12

[1] -0.2671901
```

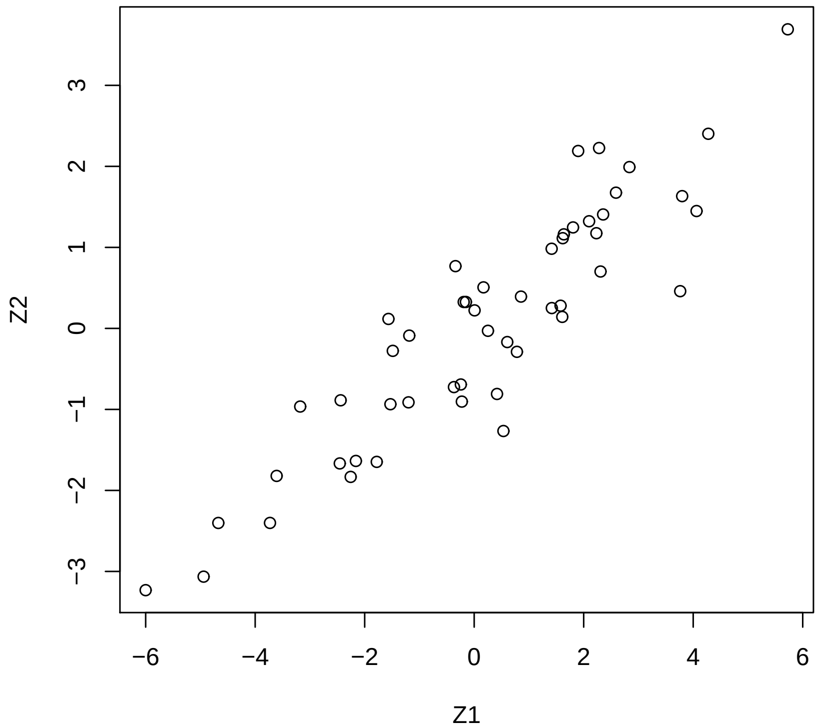



## What does covariance look like?

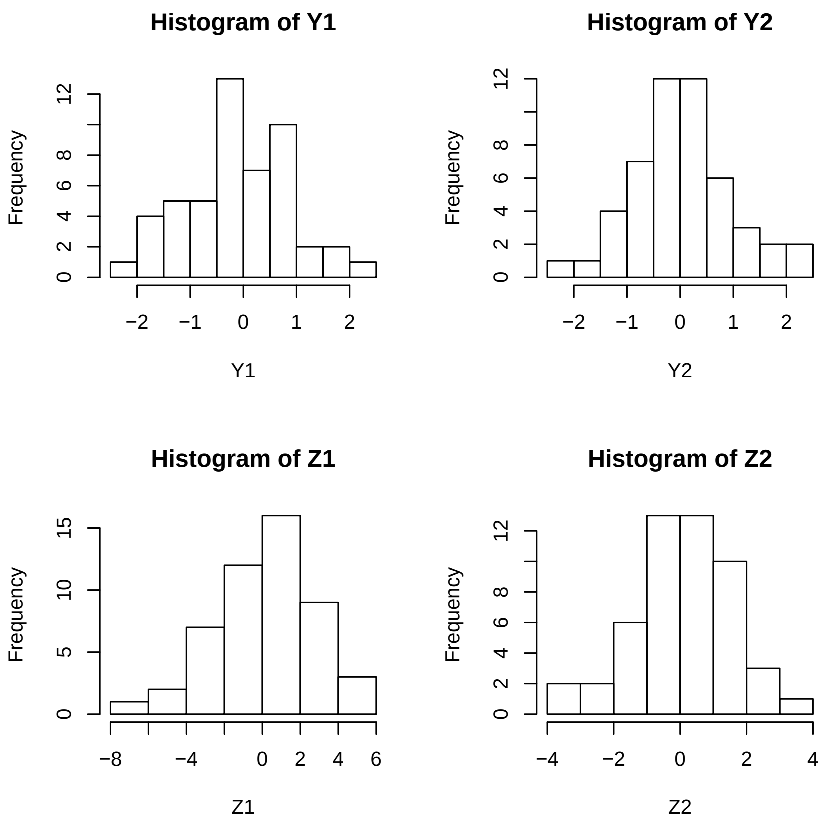
Now with covariance.

```
> library(MASS)
> Z <- mvrnorm(50, c(0, 0), Sigma = matrix(c(6, 3, 3, 2), 2, 2,
+     byrow = T))
> Z1 <- Z[, 1]
> Z2 <- Z[, 2]
> s12 <- sum((Z1 - mean(Z1)) * (Z2 - mean(Z2)))/49
> s12

[1] 3.369007
```

## What does covariance look like?

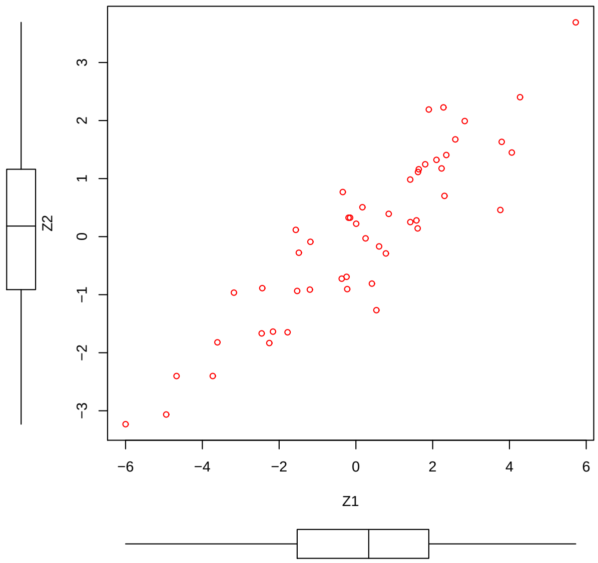



**Can univariate dists say anything?**




## Nice scatterplot

```
> library(car)
> scatterplot(Z2 ~ Z1, smooth = F, reg.line = NULL)
```



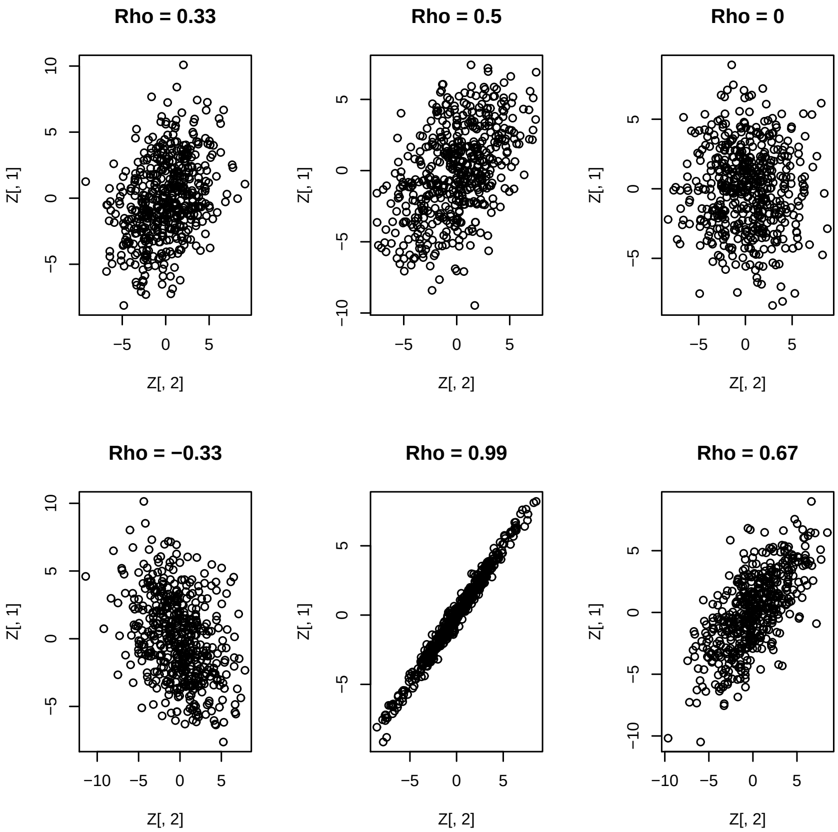

## Back to correlation

After that discussion of covariance lets get to the definition of correlation.
The Pearson product-moment correlation between two variables is

$$r_{12} = \frac{s_{12}}{s_1 s_2}$$

It is the covariance divided by the standard deviations of each variable. This results in a scaled, unit-less measure.
This is an estimate of the population parameter, $\rho$

## Scatterplots with different $\rho$




## Correlation vs regression

There are clearly similarities between correlation and regression. The big difference is the difference between *Causality* and *Association*.

- Causal relationship shows association
- Association does not necessarily imply causality
- Association is often a situation where two variables are affected by the same process. For example the association between leg-length and arm-length.