Assuming the linear function $\hat{y} = a_1 x + a_0$

$$\% \underset{var}{exp} = \frac{\text{Exp. var.}}{\text{Tot. var.}} = \frac{\Sigma(\hat{y}-\bar{y})^2}{\Sigma(y-\bar{y})^2} = \frac{\Sigma(a_1 x + a_0 - \bar{y})^2}{\Sigma y'^2}$$

$$= \frac{\Sigma(a_1 x + \bar{y} - a_1\bar{x} - \bar{y})^2}{\Sigma y'^2} = \frac{\Sigma(a_1 x')^2}{\Sigma y'^2}$$

Substitute for a_1

$$= \frac{\Sigma\left(\frac{x'y'}{x'^2}\right)^2 \Sigma x'^2}{\Sigma y'^2}$$

$$= \frac{\Sigma(x'y')^2}{\Sigma y'^2 \, \Sigma x'^2}$$

Divide by N to get

$$\% \text{ exp. variance} = \boxed{\frac{(\overline{x'y'})^2}{\overline{y'^2}\;\overline{x'^2}} = r^2}$$

In words, the square of the covariance divided by the variances is equal to the explained variance.

the correlation coefficient is just the square root of the explained variance.

$$\boxed{r = \frac{\overline{x'y'}}{\sqrt{\overline{x'^2}}\sqrt{\overline{y'^2}}} = \frac{\overline{x'y'}}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}}$$

Properties of r^2 and r.

1) r^2 is the % variance explained by linear fit $\hat{y}$

2) r^2 always is between 0 and 1

3) r varies from -1 to 1

$r > 0 \rightarrow$ positively correlated (variables relate in the same way)

$r < 0 \rightarrow$ negatively correlated (variables relate in the opposite way)

Relationship of r to a_1

$$a_1 = r \cdot \frac{\sigma_y}{\sigma_x}$$

Correlation coef. x ratio of std. deviations.

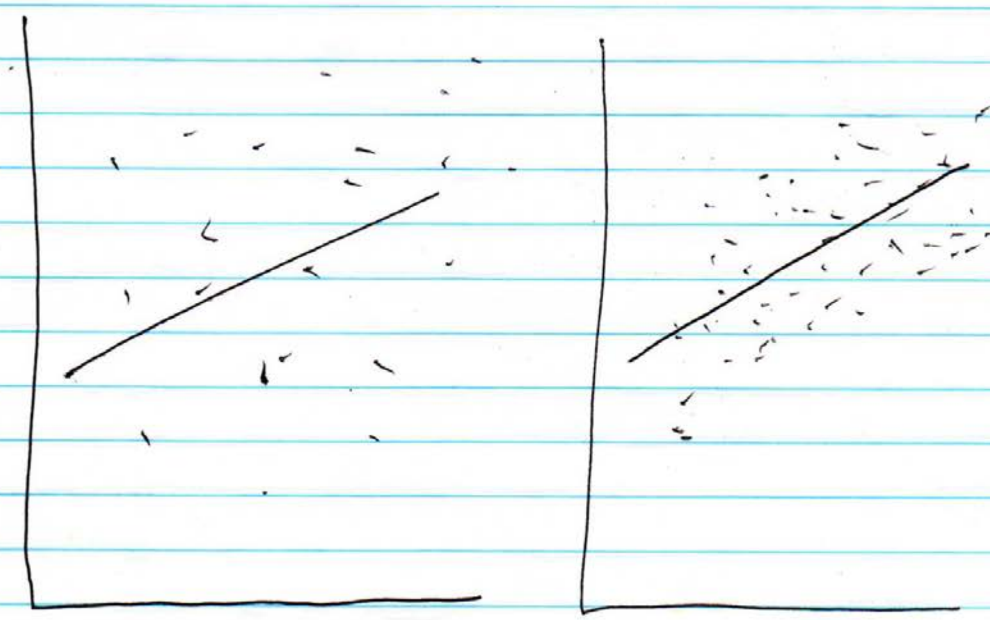
What r and a_1 give physically

$$a_1 = \frac{\overline{x'y'}}{\overline{x'^2}}$$ Change in y per change in x

$$r = \frac{\overline{x'y'}}{\sigma_x \sigma_y}$$ How good the regression fit is (the spread)

So a large value in a_1 does not necessarily mean a large correlation coefficient.

Consider the following scatter plots with the same regression line (a_1)



Weakly correlated
Lots of scatter

Highly correlated
Tightly clustered.

The quantity $\sqrt{1-r^2}$ has a special physical meaning too:

$1-r^2 \rightarrow$ Unexplained variance
$\sqrt{1-r^2} \rightarrow$ Root mean square error (RMSE)

CAVEATS ON CORRELATION:

- Only works for linear relationships
- Does not reveal quadrature relationships (when things are 90° out of phase)
- We've assumed the data (x_i, y_i) are linearly independent (i.e. not autocorrelated) → A big one!!
- Correlation DOES NOT ESTABLISH PHYSICAL CAUSE AND EFFECT (e.g. I think this is a major handicap ~~of~~ any statistically-based forecasting technique. Typically find some "cool" relationship between predictor(s) and predictand(s), establish a methodology that works in hindcast mode, apply in operational forecasting.

Hurricane forecasting: Bill Gray bases forecast on things like

- QBO
- African rainfall
- Atlantic SST
- El Niño

How do these physically interact to give more or less hurricanes? Stats. don't say...

Quote from Bill Gray at my master's defense: " Why do want to waste your time on modeling? "

My perspective : A sound and robust approach to physical understanding + predictability should include stat. & physically based methods - the latter is next semester! !

Sampling theory of correlation.

- When we calculate the sample correlation coefficient r, we are estimating the correlation coefficient assuming an infinite sample size (ρ). ρ is the "true" correlation coefficient.

$\boxed{\text{Case } \rho = 0}$

- When the true correlation coefficient is zero, then the sampling distribution of r will follow the t-distribution.

Statistic is :

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Student's t-distribution with $\nu = N-2$.

Example: $N = 10$ $\quad r = 0.6$.

Does r differ from $\rho = 0$ at the 95% level?

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = 2.12.$$

2-tailed test (95%)

$t_{crit} = t_{0.975} = 2.31$

X Not satisfied.

1-tailed test is satisfied

$\therefore$ State that the null hypothesis (H_0) that $\rho = 0$ cannot be rejected against alternative hypothesis that $\rho \neq 0$ (H_1).

When to use 1-sided vs. 2-sided test:

1-sided: Have a prior expectation that correlation should be positive or negative.

2-sided: No prior expectation that correlation should be of either sign.

Case $\rho \neq 0$

When the true correlation coefficient is <u>not</u> expected to be zero, distribution for $\rho \neq 0$ is skewed;

Use Fisher's z-transformation to convert distribution of r to normal:

$$z = \frac{1}{2} \ln \left\{ \frac{1+r}{1-r} \right\} \Rightarrow z\text{-statistic}$$

This statistic is normal with mean:

$$\mu_z = \frac{1}{2} \ln \left\{ \frac{1+\rho_0}{1-\rho_0} \right\}$$

Std. deviation

$$\sigma_z = \frac{1}{\sqrt{N-3}}$$

Example from Hartmann:

$N = 21 \quad r = 0.8$ → What is 95% conf. interval?

$$z = \frac{1}{2} \ln \left\{ \frac{1+0.8}{1-0.8} \right\} = 1.098$$

$$z - 1.96\,\sigma_z < \mu_z < z + 1.96\,\sigma_z$$

$$0.6366 < \mu_z < 1.560.$$

Transforming this back in terms of correlation:

$0.56 < \rho < 0.92$ → 95% conf. interval.

Also a similar test for differences in correlation, similiar to t-test.