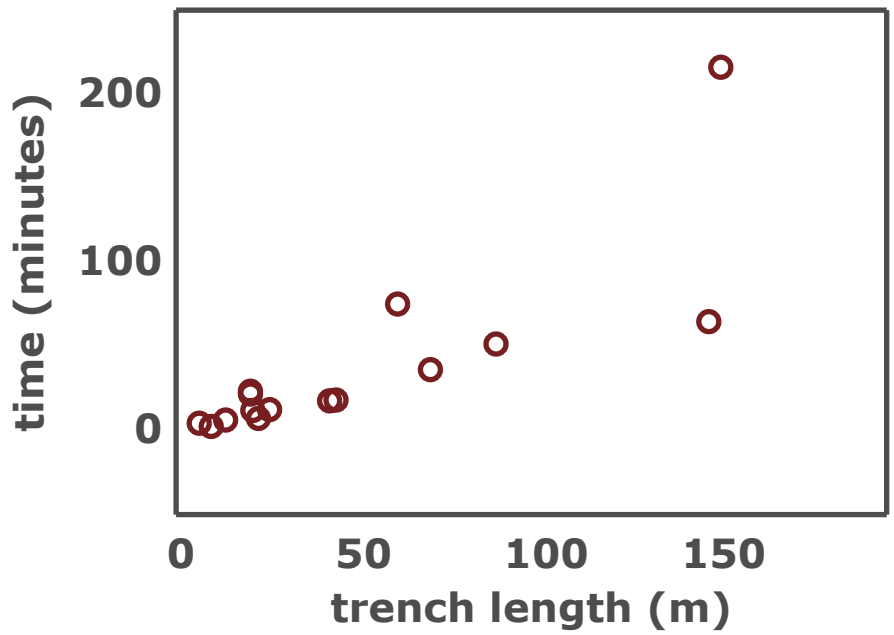
regression

Regression generates a statistical model ($\hat{y}$) for the predictand (y) based on one or more predictors (x).

Example:
How long does it take to dig a trench?

We measure N paired data points:
□ □ x□ (predictor)□ trench length
□ □ y□ (predictand)□digging time

We want to find a relationship between these two variables.



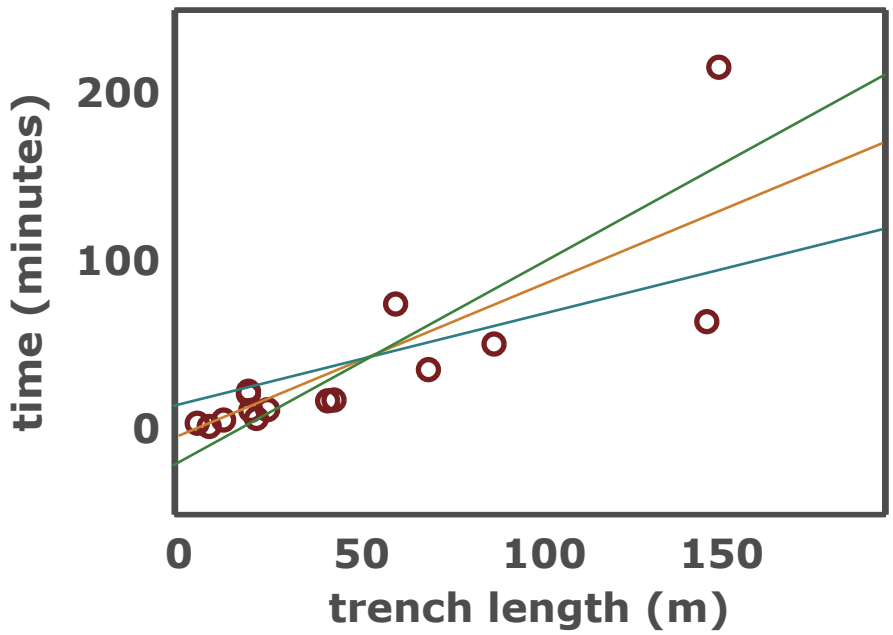
regression

Regression generates a statistical model ($\hat{y}$) for the predictand (y) based on one or more predictors (x).

Example:
How long does it take to dig a trench?

We measure N paired data points:
□ □ x□ (predictor)□ trench length
□ □ y□ (predictand)□digging time

We want to find a relationship between these two variables.



regression
linear least squares

How long does it take to dig a trench?

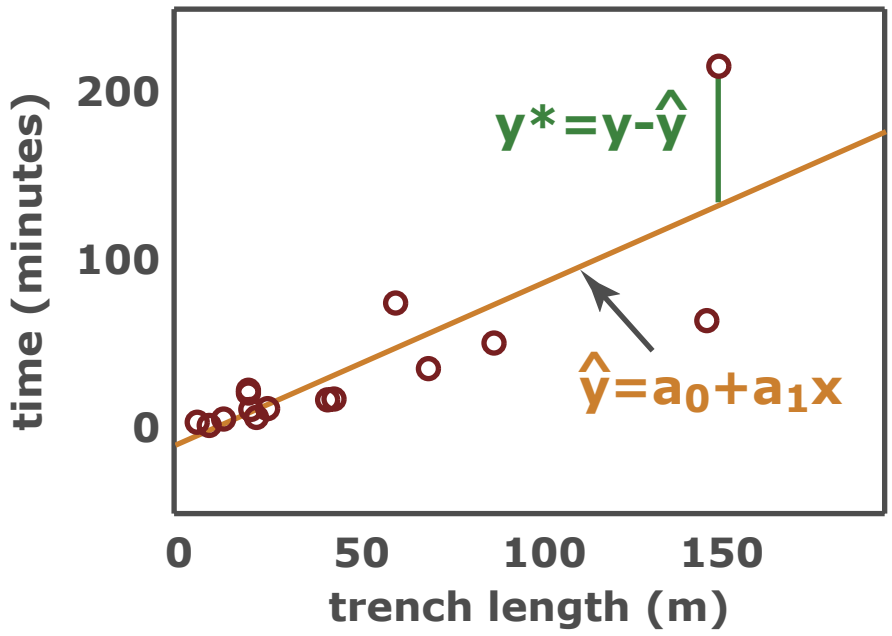
To estimate y based on known values of x, we can find a fit ($\hat{y}$)

$$\hat{y} = a_0 + a_1 x$$

that minimizes the error (ε or y*) in a least squares sense by minimizing

$$Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (\hat{y} - y_i)^2$$

We solve for the regression coefficients a_0 and a_1 that minimize Q.



regression
linear least squares

How long does it take to dig a trench?

To estimate y based on known values of x, we can find a fit ($\hat{y}$)

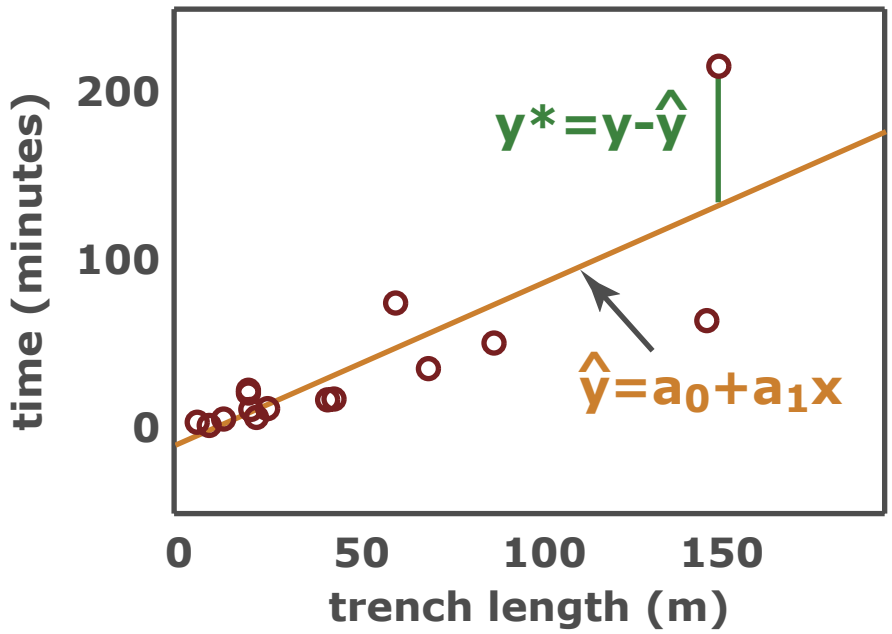
$$\hat{y} = a_0 + a_1 x$$

that minimizes the error (ε or y*) in a least squares sense by minimizing

$$Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (\hat{y} - y_i)^2$$

We solve for the regression coefficients a_0 and a_1 that minimize Q.

Note:
- Q is positive definite
- the minimization of Q is a
  - linear problem
- large errors are weighted
  - more heavily
- if x has uncertainties, results
  - will be biased



regression
linear least squares

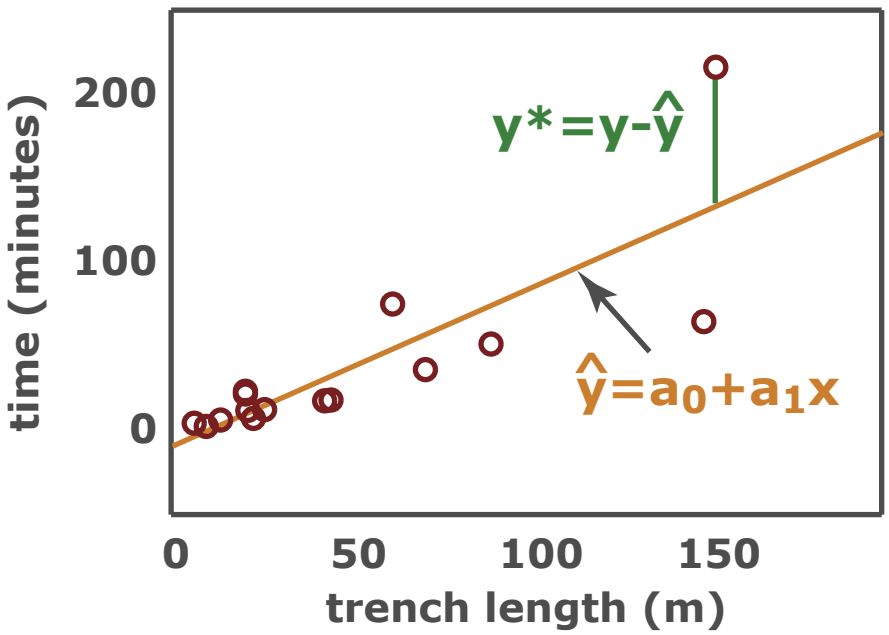
So we minimize: $Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (\hat{y} - y_i)^2$

$\frac{\partial Q}{\partial a_0} = 0$ **and** $\frac{\partial Q}{\partial a_1} = 0$

$$\rightarrow \frac{\partial Q}{\partial a_0} = 2a_0 N + 2a_1 \sum x_i - 2 \sum y_i = 0$$

$$\frac{\partial Q}{\partial a_1} = 2a_0 \sum x_i + 2a_1 \sum x_i^2 - 2 \sum x_i y_i = 0$$

$$\begin{bmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \overline{y} \\ \overline{xy} \end{bmatrix}$$ **in matrix form**



regression
linear least squares

So we minimize: $Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (\hat{y} - y_i)^2$

$\frac{\partial Q}{\partial a_0} = 0$ **and** $\frac{\partial Q}{\partial a_1} = 0$

$$\rightarrow \frac{\partial Q}{\partial a_0} = 2a_0 N + 2a_1 \sum x_i - 2 \sum y_i = 0$$

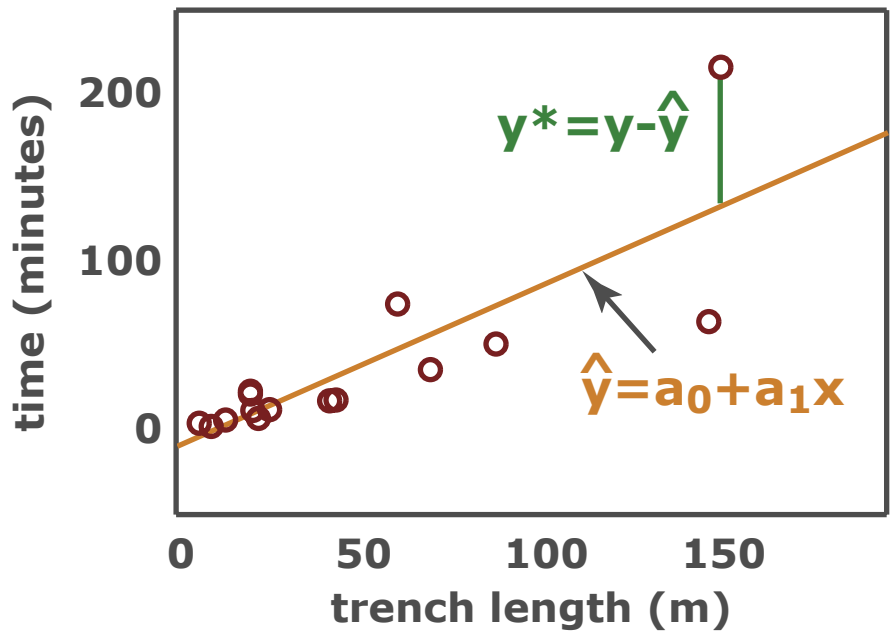
$$\frac{\partial Q}{\partial a_1} = 2a_0 \sum x_i + 2a_1 \sum x_i^2 - 2 \sum x_i y_i = 0$$

$$\begin{bmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \overline{y} \\ \overline{xy} \end{bmatrix}$$ **in matrix form**

Solutions:

$$a_0 = \overline{y} - a_1 \overline{x}$$

$$a_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\overline{x'y'}}{\overline{x'^2}}$$



regression
linear least squares

So we minimize: $Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (\hat{y} - y_i)^2$

$$\frac{\partial Q}{\partial a_0} = 0 \quad \text{and} \quad \frac{\partial Q}{\partial a_1} = 0$$

$$\rightarrow \frac{\partial Q}{\partial a_0} = 2a_0 N + 2a_1 \sum x_i - 2 \sum y_i = 0$$

$$\frac{\partial Q}{\partial a_1} = 2a_0 \sum x_i + 2a_1 \sum x_i^2 - 2 \sum x_i y_i = 0$$

$$\begin{bmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \overline{y} \\ \overline{xy} \end{bmatrix}$$ **in matrix form**

Solutions:

$$a_0 = \overline{y} - a_1 \overline{x}$$

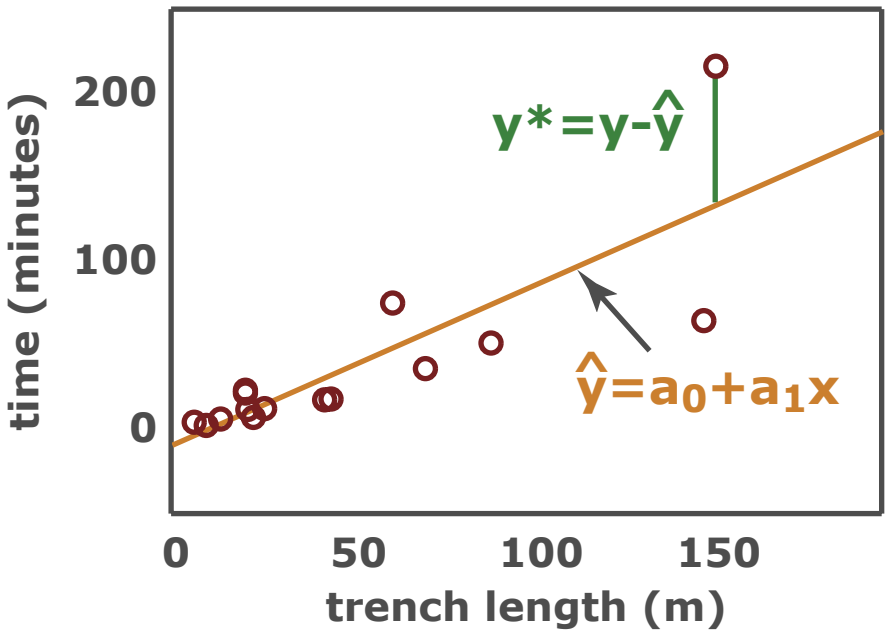
$$a_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\overline{x'y'}}{\overline{x'^2}}$$

covariance of y with x

$$\overline{x'y'} = \mathrm{cov}(x, y) = \frac{1}{N} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

variance of x

$$\overline{x'^2} = \mathrm{var}(x) = \frac{1}{N} \sum_{i=1}^{n} (x_i - \overline{x})^2$$



regression
correlation

How good is our least squares fit?

solution

$$a_0 = \overline{y} - a_1\overline{x}$$

$$a_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\overline{x'y'}}{\overline{x'^2}}$$

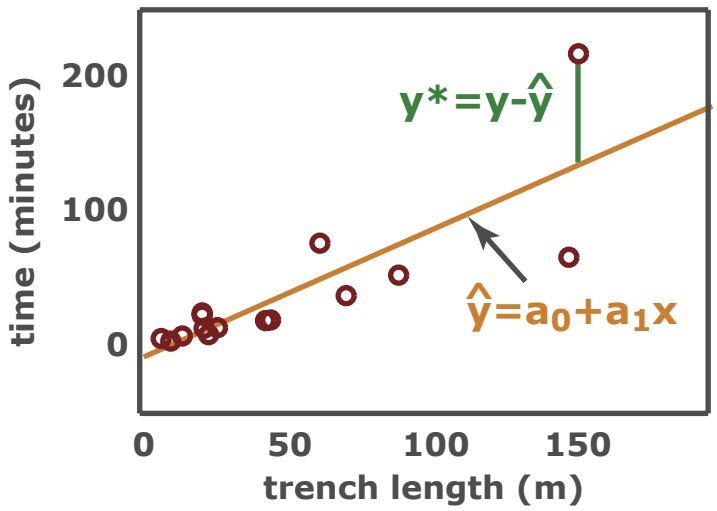
covariance of y with x

$$\overline{x'y'} = \mathrm{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

variance of x

$$\overline{x'^2} = \mathrm{var}(x) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Solution: a_0 = -9.6, a_1 = 0.93



regression
correlation

How good is our least squares fit?

$$\frac{\text{error (residual variance)}}{\text{total variance}} \longrightarrow \frac{\frac{1}{N}\sum_{i=1}^{n}(\hat{y}-y_i)^2}{\frac{1}{N}\sum_{i=1}^{n}(y_i-\overline{y})^2} = \frac{\frac{1}{N}\sum_{i=1}^{n}(\hat{y}-y_i)^2}{\overline{y'^2}}$$

solution

$$a_0 = \overline{y} - a_1\overline{x}$$

$$a_1 = \frac{\overline{xy}-\overline{x}\,\overline{y}}{\overline{x^2}-\overline{x}^2} = \frac{\overline{x'y'}}{\overline{x'^2}}$$

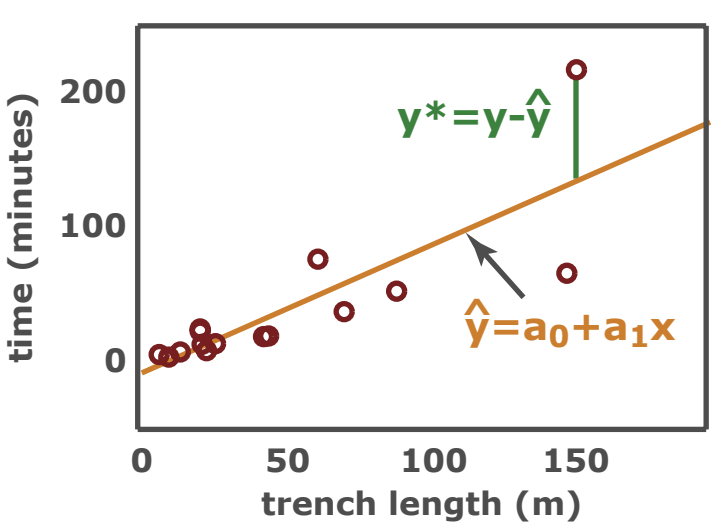
covariance of y with x

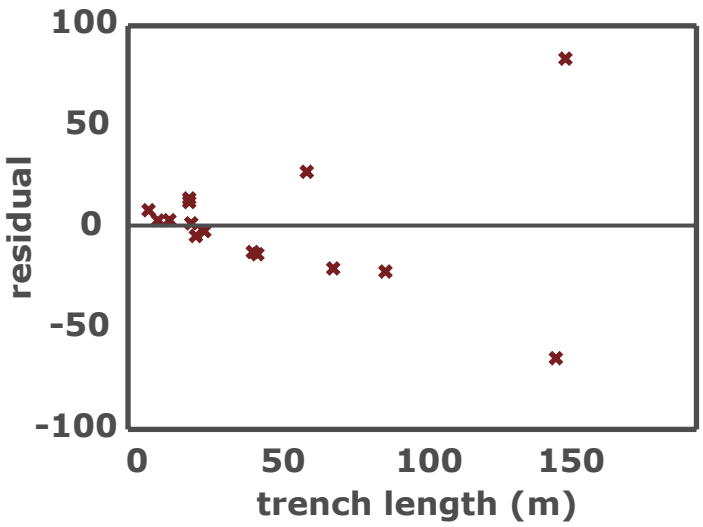
$$\overline{x'y'} = \mathrm{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$$

variance of x

$$\overline{x'^2} = \mathrm{var}(x) = \frac{1}{N}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

Solution: a_0 = -9.6, a_1 = 0.93





regression
correlation

How good is our least squares fit?

$$\frac{\text{error (residual variance)}}{\text{total variance}} \longrightarrow \frac{\frac{1}{N}\sum_{i=1}^{n}(\hat{y}-y_i)^2}{\frac{1}{N}\sum_{i=1}^{n}(y_i-\overline{y})^2} = \frac{\frac{1}{N}\sum_{i=1}^{n}(\hat{y}-y_i)^2}{\overline{y'^2}}$$

Or, we can write:

$$\overline{y'^2} = a_1^2\overline{x'^2} + \overline{y^{*2}} + 2a_1\overline{x'y^*}$$

$$1 = a_1^2\frac{\overline{x'^2}}{\overline{y'^2}} + \frac{\overline{y^{*2}}}{\overline{y'^2}}$$

☐ ☐ (total variance = % explained + % unexplained)

solution

$$a_0 = \overline{y} - a_1\overline{x}$$

$$a_1 = \frac{\overline{xy}-\overline{x}\,\overline{y}}{\overline{x^2}-\overline{x}^2} = \frac{\overline{x'y'}}{\overline{x'^2}}$$

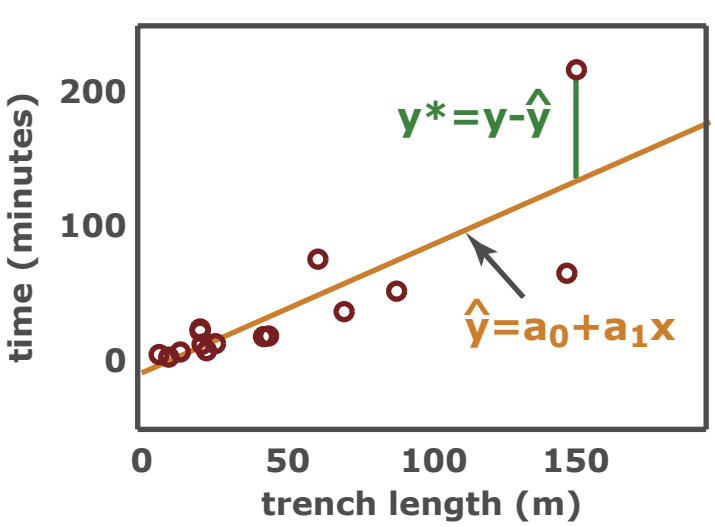
covariance of y with x

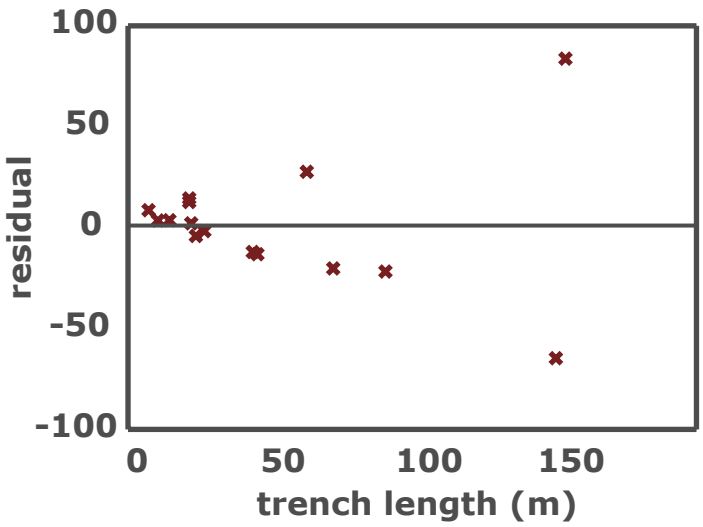
$$\overline{x'y'} = \mathrm{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$$

variance of x

$$\overline{x'^2} = \mathrm{var}(x) = \frac{1}{N}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

Solution: a_0 = -9.6, a_1 = 0.93





regression
correlation

How good is our least squares fit?

$$\frac{\text{error (residual variance)}}{\text{total variance}} \longrightarrow \frac{\frac{1}{N}\sum_{i=1}^{n}(\hat{y}-y_i)^2}{\frac{1}{N}\sum_{i=1}^{n}(y_i-\overline{y})^2} = \frac{\frac{1}{N}\sum_{i=1}^{n}(\hat{y}-y_i)^2}{\overline{y'^2}}$$

Or, we can write:

$$\overline{y'^2} = a_1^2\overline{x'^2} + \overline{y^{*2}} + 2a_1\overline{x'y^*}$$

$= 0$

$$1 = a_1^2\frac{\overline{x'^2}}{\overline{y'^2}} + \frac{\overline{y^{*2}}}{\overline{y'^2}}$$

☐ ☐ (total variance = % explained + % unexplained)

solution

$$a_0 = \overline{y} - a_1\overline{x}$$

$$a_1 = \frac{\overline{xy}-\overline{x}\,\overline{y}}{\overline{x^2}-\overline{x}^2} = \frac{\overline{x'y'}}{\overline{x'^2}}$$

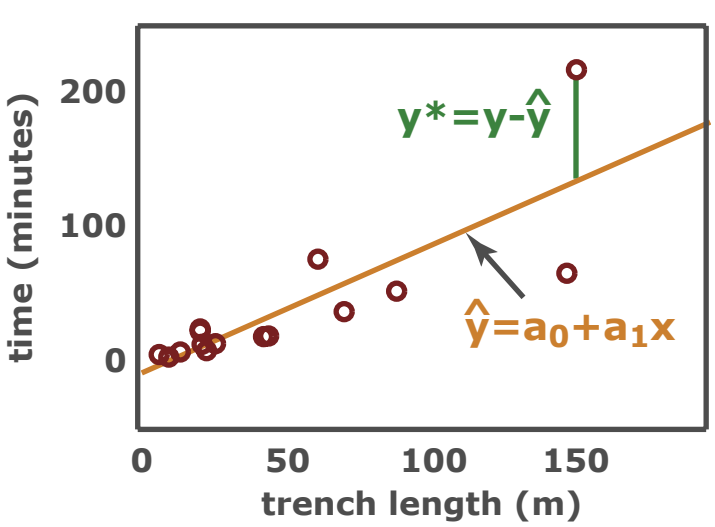
covariance of y with x

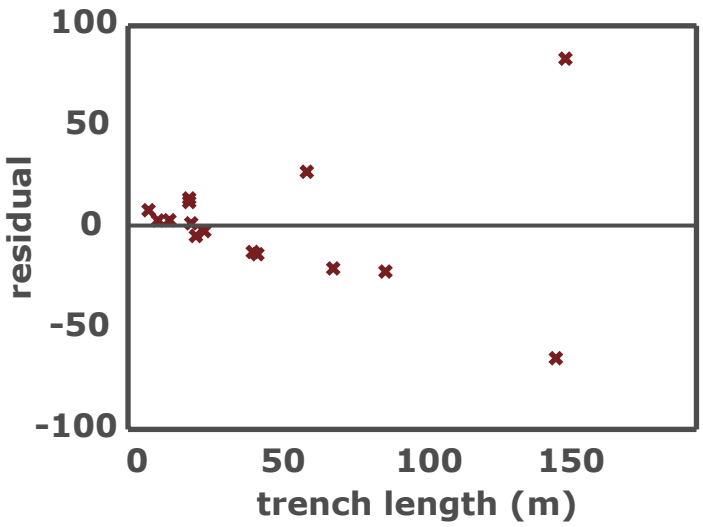
$$\overline{x'y'} = \mathrm{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$$

variance of x

$$\overline{x'^2} = \mathrm{var}(x) = \frac{1}{N}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

Solution: a_0 = -9.6, a_1 = 0.93





regression
correlation

How good is our least squares fit?

$$\overline{y'^2} = a_1^2\overline{x'^2} + \overline{y^{*2}} + 2a_1\overline{x'y^*}$$

$$1 = a_1^2\frac{\overline{x'^2}}{\overline{y'^2}} + \frac{\overline{y^{*2}}}{\overline{y'^2}}$$

$$r^2 = a_1^2\frac{\overline{x'^2}}{\overline{y'^2}} = \frac{(\overline{x'y'})^2}{\overline{x'^2}\,\overline{y'^2}}$$

r^2 = fraction of variance
- ☐ in data explained
- ☐ by the least squares fit

$$r = \frac{\overline{x'y'}}{\sigma_x\sigma_y}$$

r = correlation coefficient

solution

$$a_0 = \overline{y} - a_1\overline{x}$$

$$a_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\overline{x'y'}}{\overline{x'^2}}$$

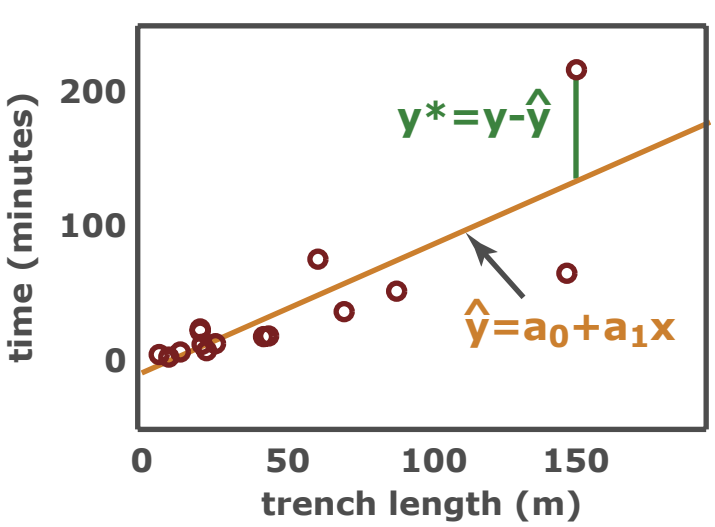
covariance of y with x

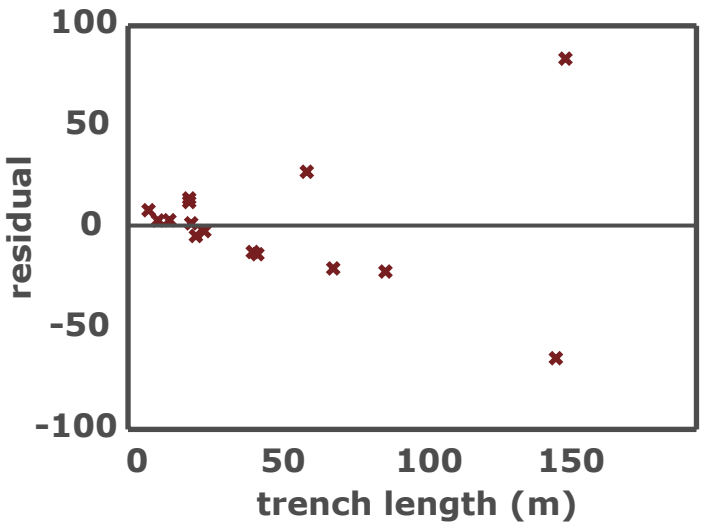
$$\overline{x'y'} = \mathrm{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

variance of x

$$\overline{x'^2} = \mathrm{var}(x) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Solution: a_0 = -9.6, a_1 = 0.93





regression
practical considerations

Normality Assumption

- **Check that the residuals are distributed normally.**
- **If the error is correlated with the indendent variable, the linear fit is not optimum.**

correlation

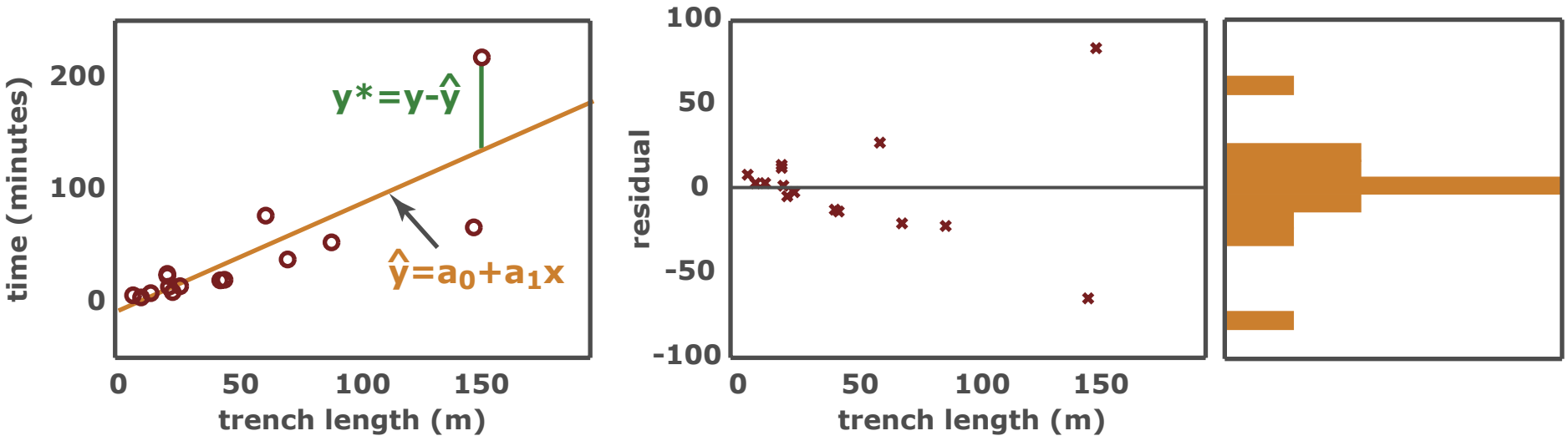
$$r^2 = a_1^2 \frac{\overline{x'^2}}{\overline{y'^2}} = \frac{(\overline{x'y'})^2}{\overline{x'^2}\,\overline{y'^2}}$$

covariance of y with x

$$\overline{x'y'} = \mathrm{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

variance of x

$$\overline{x'^2} = \mathrm{var}(x) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})^2$$



regression
practical considerations

RMS error = $(1-r^2)^{0.5}$

r	RMS error
0.98	0.20
0.90	0.43
0.50	0.87

Statistically significant correlations may not be useful for prediction.

correlation

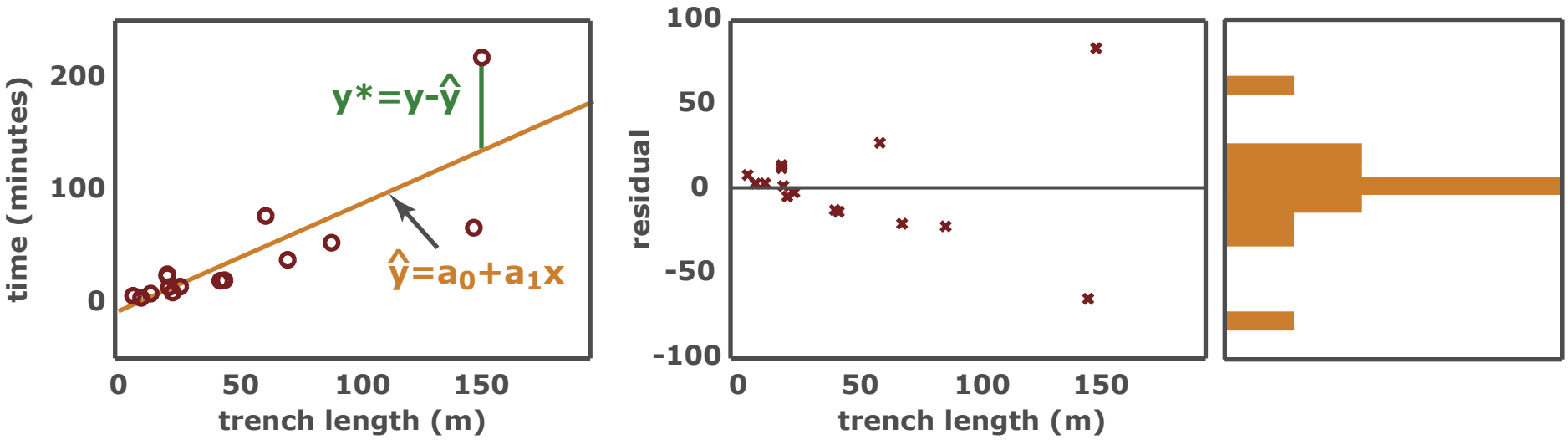
$$r^2 = a_1^2 \frac{\overline{x'^2}}{\overline{y'^2}} = \frac{(\overline{x'y'})^2}{\overline{x'^2}\,\overline{y'^2}}$$

covariance of y with x

$$\overline{x'y'} = \mathrm{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$$

variance of x

$$\overline{x'^2} = \mathrm{var}(x) = \frac{1}{N}\sum_{i=1}^{n}(x_i-\overline{x})^2$$



regression
practical considerations

Both variables uncertain:

1. **Define one variable as dependent.**

2. **Minimize the perpendicular distance to the regression line.**

- **(EOF analysis in two dimensions)**

correlation

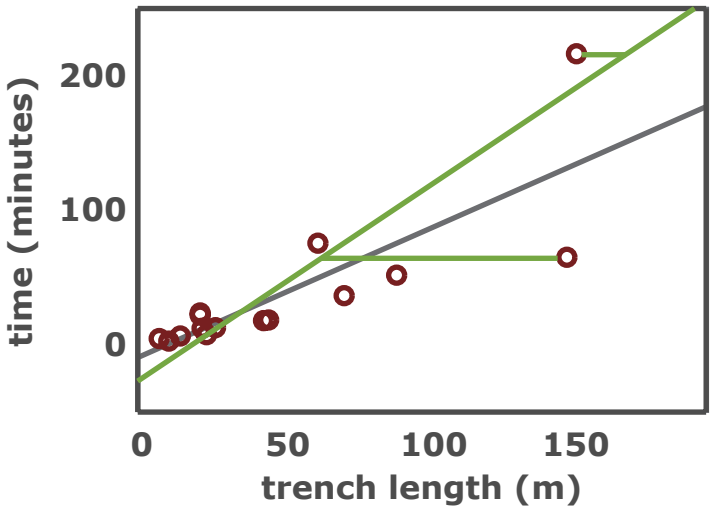
$$r^2 = a_1^2 \frac{\overline{x'^2}}{\overline{y'^2}} = \frac{(\overline{x'y'})^2}{\overline{x'^2}\,\overline{y'^2}}$$

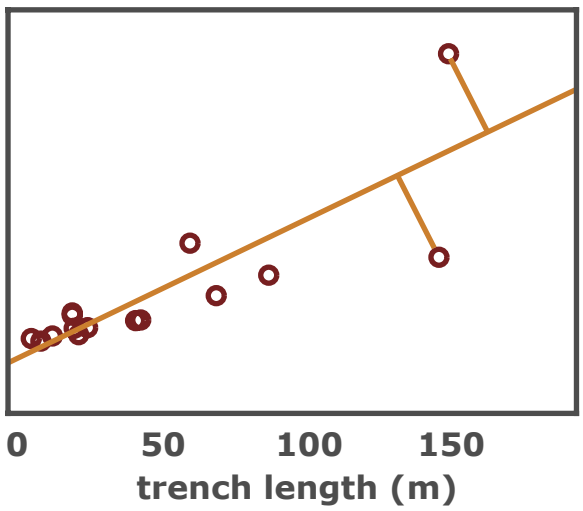
covariance of y with x

$$\overline{x'y'} = \mathrm{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

variance of x

$$\overline{x'^2} = \mathrm{var}(x) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

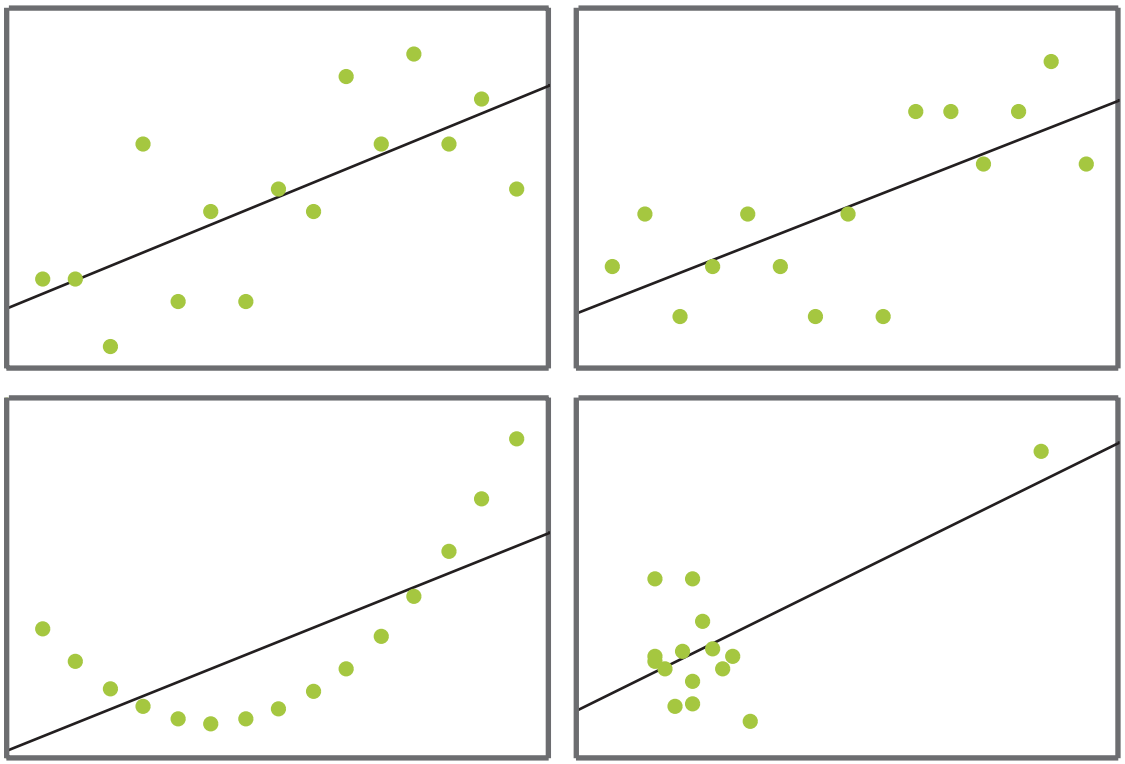




regression
practical considerations

r reveals information about linear relationships only

Example:
datasets fit with r=0.7 regression lines



correlation

$$r^2 = a_1^2 \frac{\overline{x'^2}}{\overline{y'^2}} = \frac{(\overline{x'y'})^2}{\overline{x'^2}\,\overline{y'^2}}$$

covariance of y with x

$$\overline{x'y'} = \mathrm{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

variance of x

$$\overline{x'^2} = \mathrm{var}(x) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Some final comments:

- cannot reveal non-linear or quadrature relationships
- when testing for statistical significance of correlations, check degrees of freedom
- be careful in interpreting correlations: relationships by chance or through a third variable?

regression
multiple regression

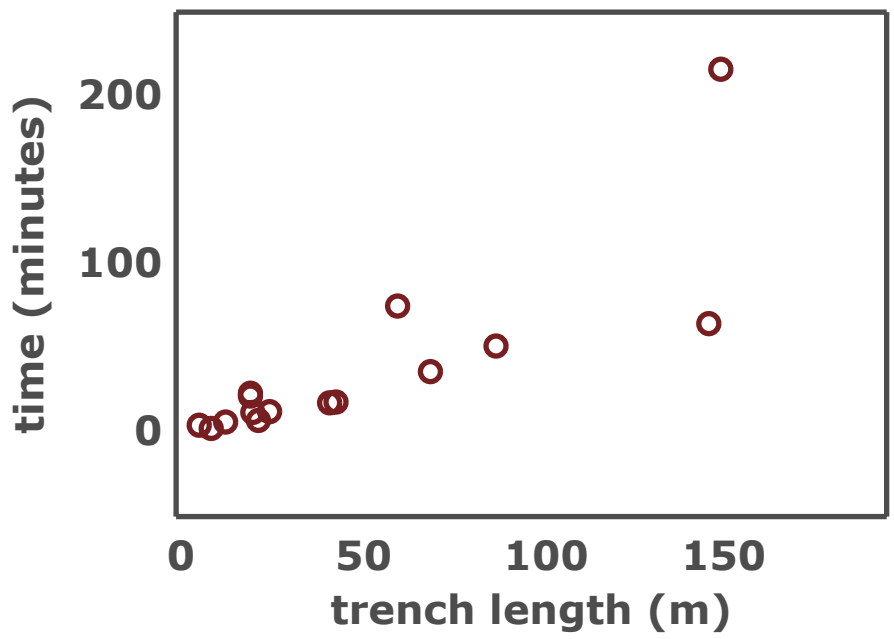
How long does it take to dig a trench?

Length is not the only factor to consider, so we generalize our model to include more predictors:

$$\hat{y} = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n$$

Miminimizing in a least squares sense:

$$\begin{bmatrix} \overline{x_1^2} & \overline{x_1x_2} & \overline{x_1x_3} & \cdots \\ \overline{x_2x_1} & \overline{x_2^2} & \overline{x_2x_3} & \cdots \\ \overline{x_3x_1} & \overline{x_3x_2} & \overline{x_3^2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix} = \begin{bmatrix} \overline{x_1y} \\ \overline{x_2y} \\ \overline{x_3y} \\ \vdots \end{bmatrix}$$



regression
multiple regression

How long does it take to dig a trench?

Length is not the only factor to consider, so we generalize our model to include more predictors:

$$\hat{y} = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n$$

Miminimizing in a least squares sense:

$$\begin{bmatrix} \overline{x_1^2} & \overline{x_1x_2} & \overline{x_1x_3} & \cdots \\ \overline{x_2x_1} & \overline{x_2^2} & \overline{x_2x_3} & \cdots \\ \overline{x_3x_1} & \overline{x_3x_2} & \overline{x_3^2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix} = \begin{bmatrix} \overline{x_1y} \\ \overline{x_2y} \\ \overline{x_3y} \\ \vdots \end{bmatrix}$$

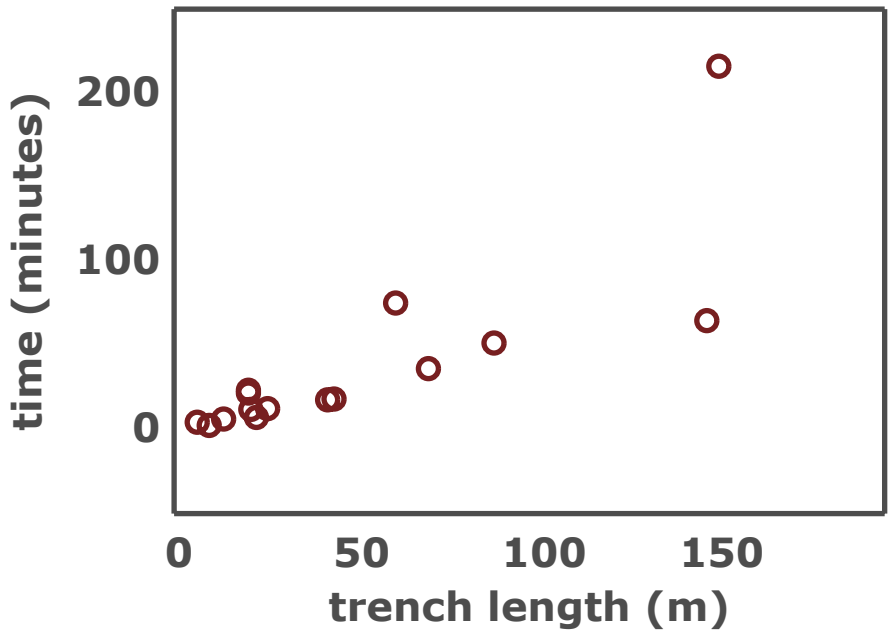
covariance matrix of predictors

coefficients

covariance of predictors with predictand

Or, in matrix form: $\overline{x_i x_j} a_j = \overline{x_i y}$

Solution of the multiple regression problem by matrix methods: see Strang 1988. The SVD places no restrictions on the covariance matrix and its factors provide important information about the matrix itself.



regression
multiple regression

How many predictors should be included?

$$\begin{bmatrix} \overline{x_1^2} & \overline{x_1x_2} & \overline{x_1x_3} & \dots \\ \overline{x_2x_1} & \overline{x_2^2} & \overline{x_2x_3} & \dots \\ \overline{x_3x_1} & \overline{x_3x_2} & \overline{x_3^2} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix} = \begin{bmatrix} \overline{x_1y} \\ \overline{x_2y} \\ \overline{x_3y} \\ \vdots \end{bmatrix} \quad \text{or} \quad \overline{x_i x_j} a_j = \overline{x_i y}$$

For two predictors, the fraction of explained variance is given by:

$$R^2 = \frac{{r_{x_1y}}^2 + {r_{x_2y}}^2 - 2r_{x_1y}r_{x_2y}r_{x_1x_2}}{1 - {r_{x_1x_2}}^2}$$

eg. trench length & soil type as predictors of digging time

regression
multiple regression

How many predictors should be included?

$$\square \begin{bmatrix} \overline{x_1^2} & \overline{x_1x_2} & \overline{x_1x_3} & \dots \\ \overline{x_2x_1} & \overline{x_2^2} & \overline{x_2x_3} & \dots \\ \overline{x_3x_1} & \overline{x_3x_2} & \overline{x_3^2} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix} = \begin{bmatrix} \overline{x_1y} \\ \overline{x_2y} \\ \overline{x_3y} \\ \vdots \end{bmatrix} \square \quad \square \quad \textbf{or}$$

$$\overline{x_i x_j} a_j = \overline{x_i y}$$

For two predictors, the fraction of explained variance is given by:

$$R^2 = \frac{r_{x_1y}{}^2 + r_{x_2y}{}^2 - 2r_{x_1y}r_{x_2y}r_{x_1x_2}}{1 - r_{x_1x_2}{}^2}$$

eg. trench length & soil type as predictors of digging time

r(x1,y)	r(x2,y)	r(x1,x2)	R2
0.5	--	--	0.25
--	0.5	--	0.25
0.5	0.5	0.99	0.251
0.5	0.5	0.5	0.33
0.5	0.5	0	0.5
0.5	0.25	0.5	0.25
0.5	0.25	0.25	0.27

regression
multiple regression

How many predictors should be included?

$$\square \begin{bmatrix} \overline{x_1^2} & \overline{x_1x_2} & \overline{x_1x_3} & \cdots \\ \overline{x_2x_1} & \overline{x_2^2} & \overline{x_2x_3} & \cdots \\ \overline{x_3x_1} & \overline{x_3x_2} & \overline{x_3^2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix} = \begin{bmatrix} \overline{x_1y} \\ \overline{x_2y} \\ \overline{x_3y} \\ \vdots \end{bmatrix} \square \quad \square \quad \textbf{or} \quad \overline{x_i x_j} a_j = \overline{x_i y}$$

For two predictors, the fraction of explained variance is given by:

$$R^2 = \frac{{r_{x_1y}}^2 + {r_{x_2y}}^2 - 2r_{x_1y}r_{x_2y}r_{x_1x_2}}{1 - {r_{x_1x_2}}^2}$$

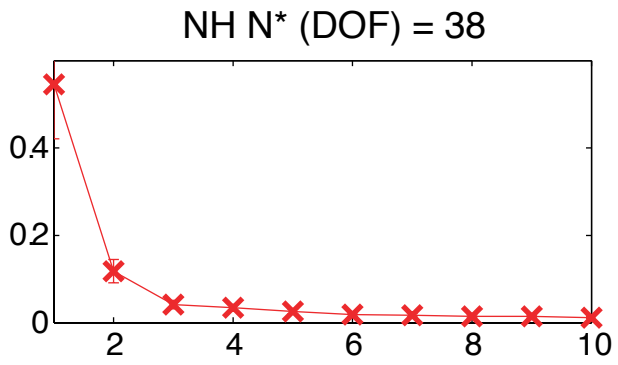
eg. trench length & soil type as predictors of digging time

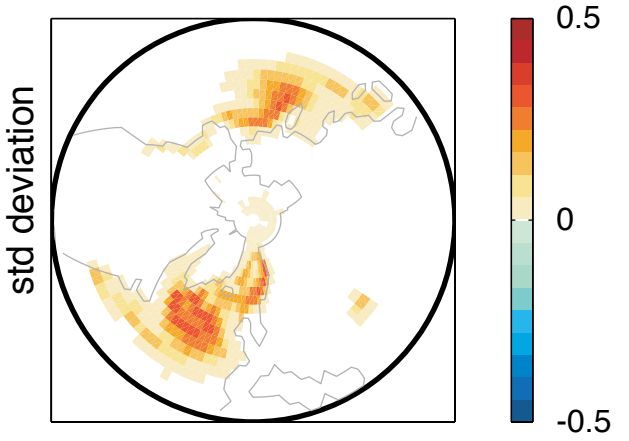
- the smaller the magnitude of $r(x_1,x_2)$, the more independent the predictors
- additional predictors must exceed the *minimum useful correlation*: $|r(x_2,y)|>|r(x_1,y)r(x_1,x_2)|$ for two predictors

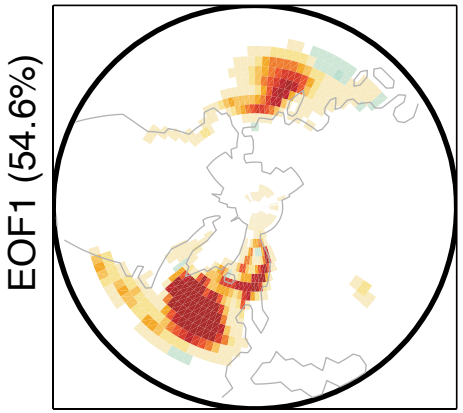
more predictors > fewer degrees of freedom in a_j coefficients > lower statistical significance of "fit" > poorer performance on independent data

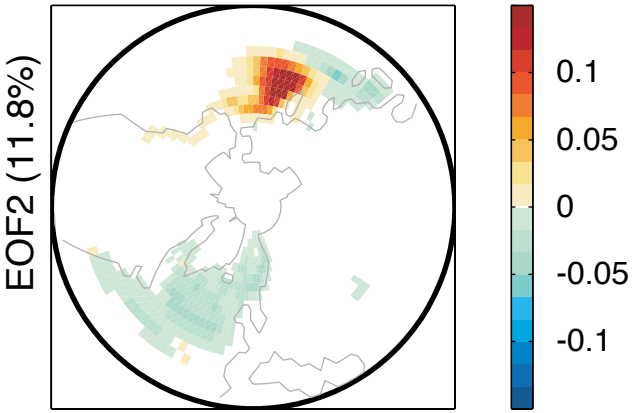
r(x1,y)	r(x2,y)	r(x1,x2)	R2
0.5	--	--	0.25
--	0.5	--	0.25
0.5	0.5	0.99	0.251
0.5	0.5	0.5	0.33
0.5	0.5	0	0.5
0.5	0.25	0.5	0.25
0.5	0.25	0.25	0.27

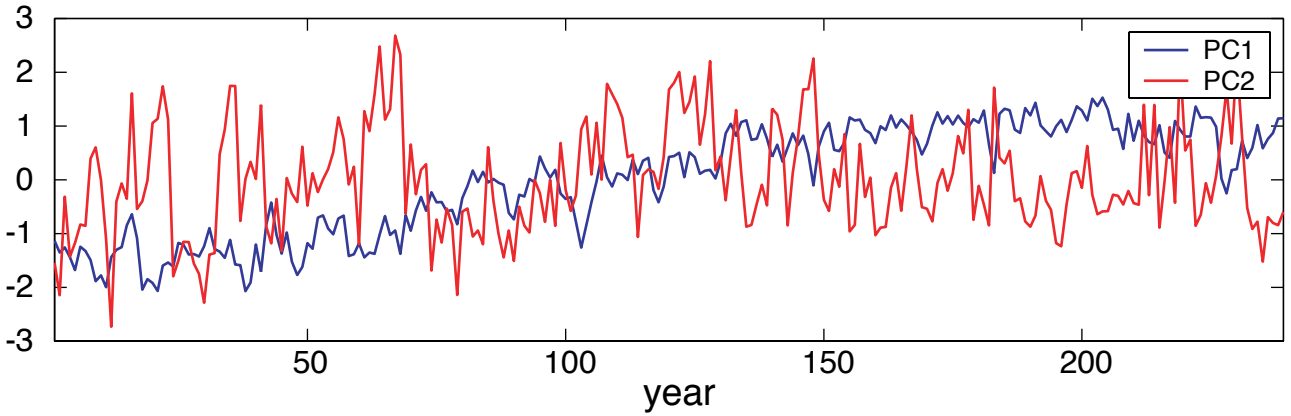
CCSM glacial run T42: 240 years JFMA ICEFRAC











CCSM glacial run: sea ice extent in millions of km^2