Regression

A Practical Introduction

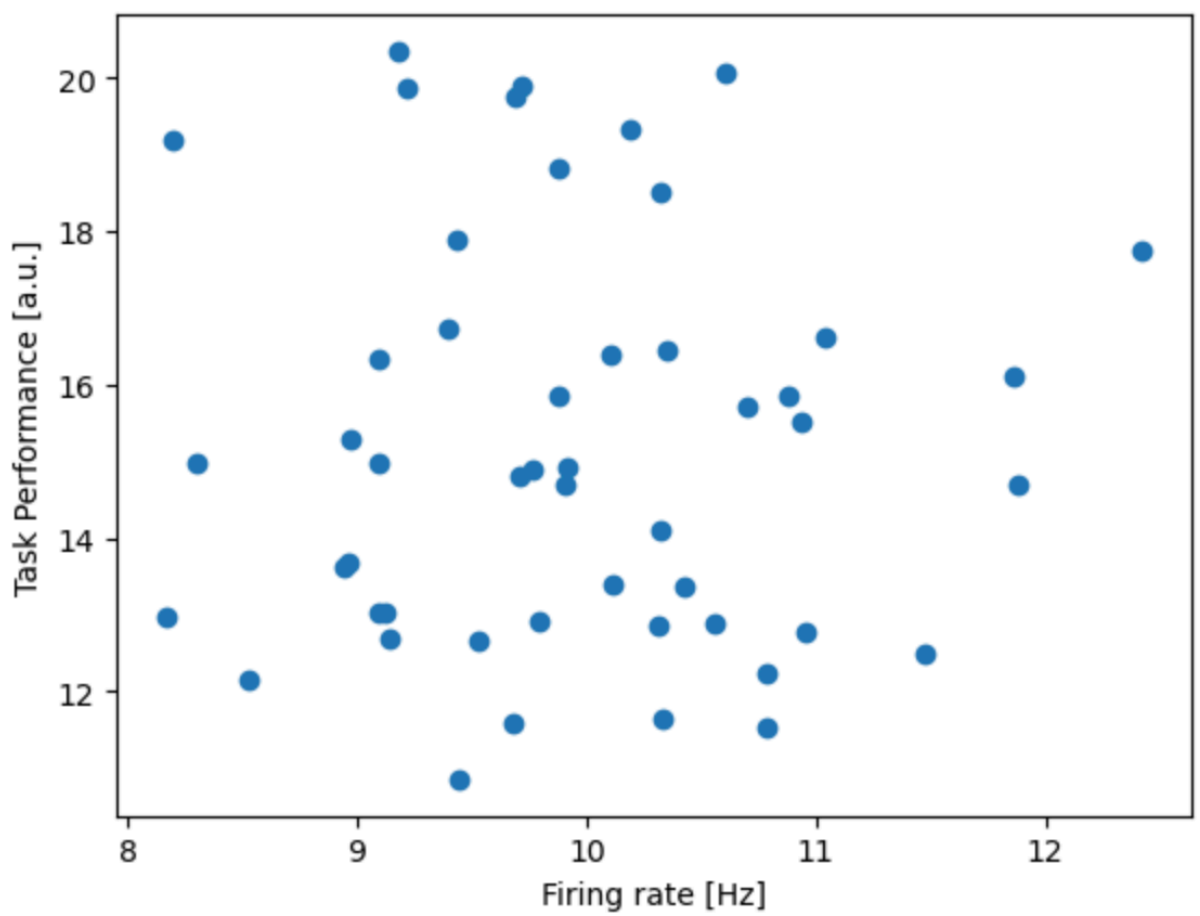
Instructor: Mark Kramer

Outline

A (very) practical introduction to linear regression

Main idea: model data as a line.

Here is my data

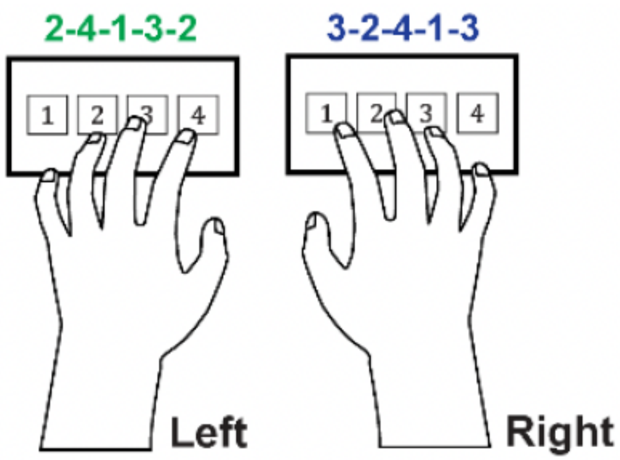


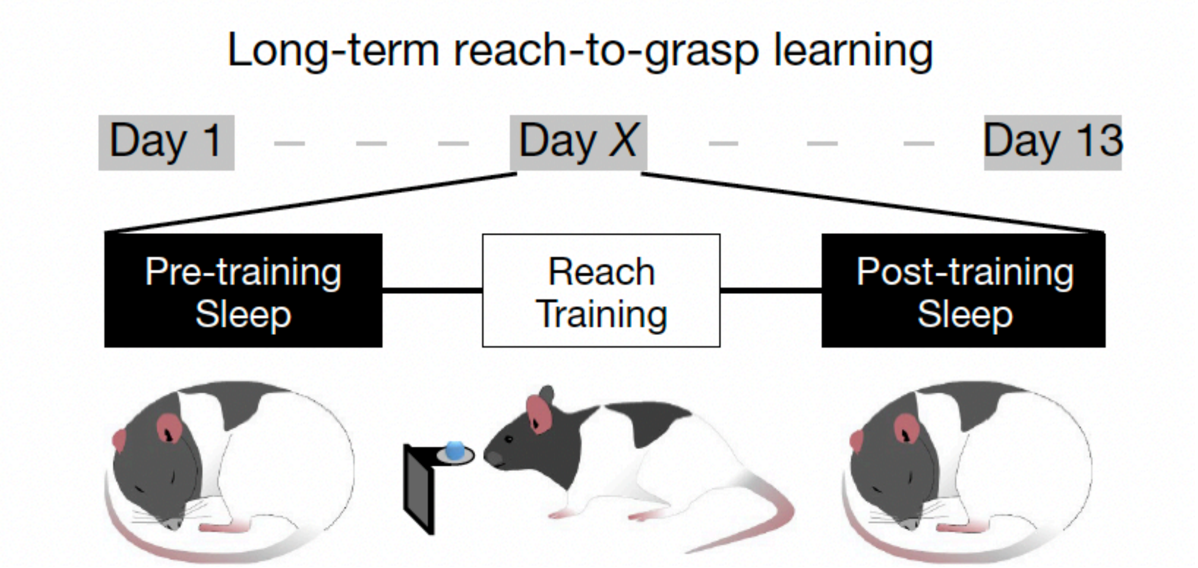
Here is my model

$$y = mx + b$$

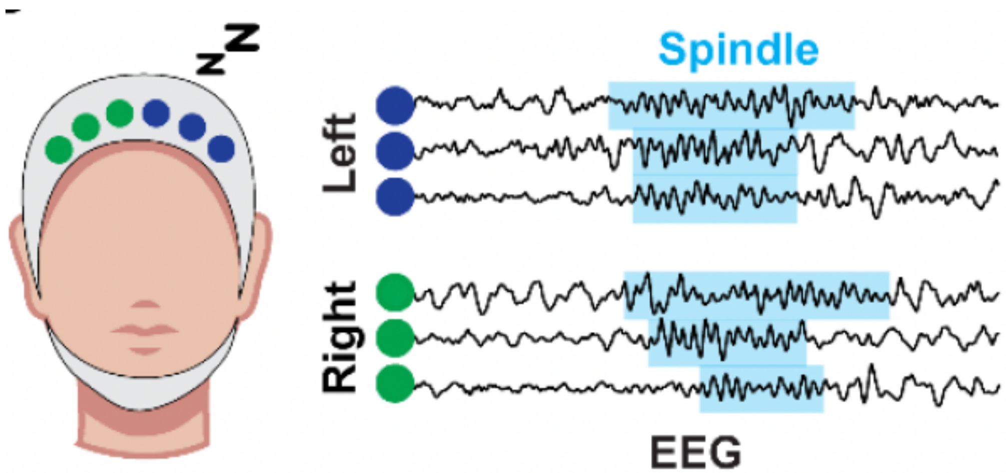
Data

Task performance (y)

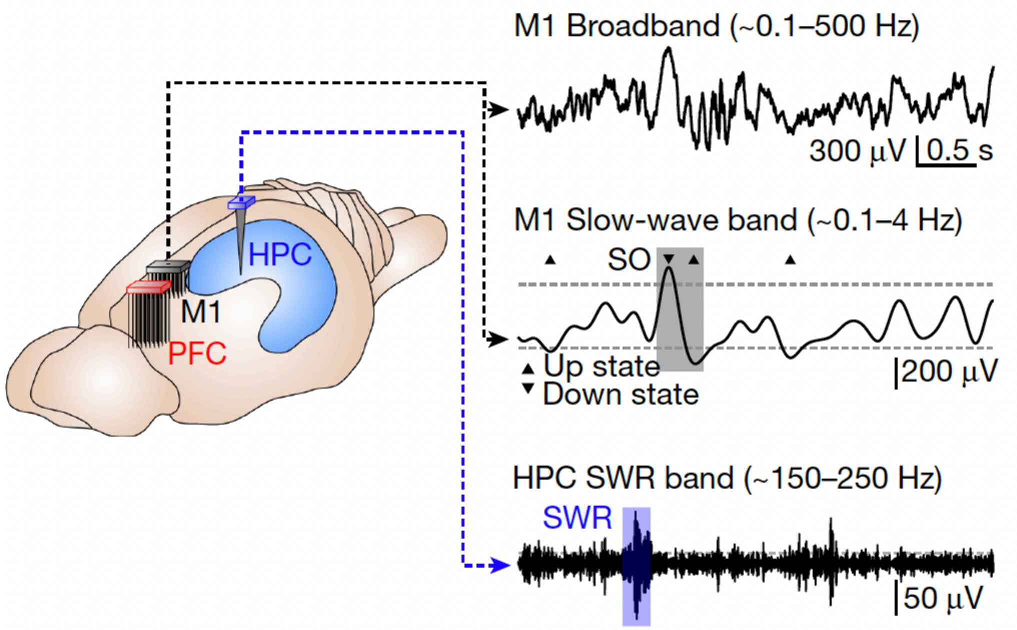




Brain activity (x)



[Kwon et al, bioRxiv, 2024]



[Kim et al, Nature, 2023]

Analyze the data (1)

Plot it …

Python

Visual inspection:

Analyze the data (2)

Compute a statistic? <u>Correlation</u> x_n and y_n: data at index n

$$C_{xy} = \frac{1}{N}\frac{1}{\sigma_x}\frac{1}{\sigma_y}\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})$$

number of data points

standard deviation of x

standard deviation of y

sum from indices 1 to N

mean of x

mean of y

mean of x $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n.$

sum the values of x for all n indices, then divide by the total number of points summed (N)

Analyze the data (2)

Compute a statistic? <u>Correlation</u> x_n and y_n: data at index n

$$C_{xy} = \frac{1}{N}\frac{1}{\sigma_x}\frac{1}{\sigma_y}\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})$$

number of data points

standard deviation of x

standard deviation of y

sum from indices 1 to N

mean of x

mean of y

variance of x $$\sigma_x^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{x})^2$$ characterizes the extent of fluctuations about the mean

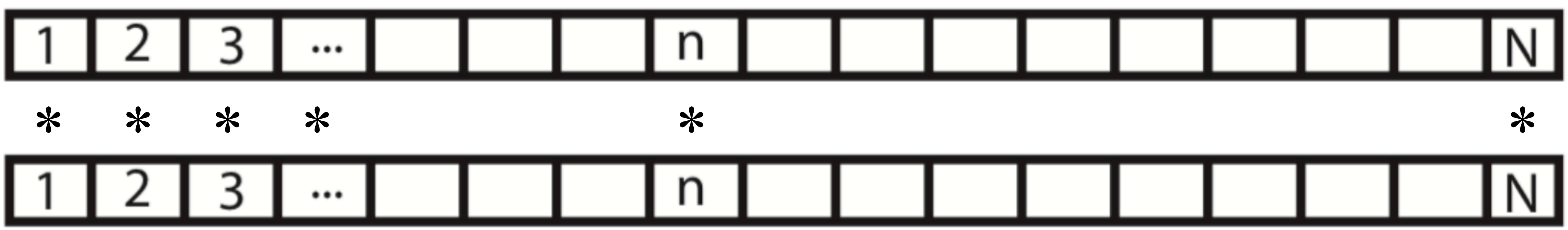
standard deviation of x $$\sigma_x = \sqrt{\sigma_x^2}$$

Analyze the data (2)

Compute a statistic? <u>Correlation</u>

$$C_{xy} = \frac{1}{N}\frac{1}{\sigma_x}\frac{1}{\sigma_y}\boxed{\sum_{n=1}^{N}}\boxed{(x_n - \bar{x})(y_n - \bar{y})}$$

sum from indices 1 to N



$x - \bar{x}$

$y - \bar{y}$

then sum & scale = C_{xy}

Analyze the data (2)

<u>Intuition</u>

<u>Correlation</u>

$$C_{xy} = \frac{1}{N}\frac{1}{\sigma_x}\frac{1}{\sigma_y}\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})$$

Assume $\bar{x} = \bar{y} = 0$

$$C_{xy} = \frac{1}{N}\frac{1}{\sigma_x}\frac{1}{\sigma_y}\sum_{n=1}^{N} x_n y_n$$

$$\boxed{\text{Reminder: } \sigma_x^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2}$$

What if x and y match? $C_{xy} = 1$

What if x equals $-y$? $C_{xy} = -1$

What if x and y are random? $C_{xy} \approx 0$

Analyze the data (2)

Compute a statistic? Correlation

$$C_{xy} = \frac{1}{N}\frac{1}{\sigma_x}\frac{1}{\sigma_y}\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})$$

Python

$$C_{xy} =$$

Conclusion:

Analyze the data (3): Regression

<u>Motivation</u>: Characterize relationships in the data.

To do so: build a *statistical* model containing

- **systematic effects**: things we know/observe that can explain the data

- **random effects**: unknown / haphazard variations that we make no attempt to model or predict

Regression

<u>Goal</u>: describe <u>succinctly</u> the systematic variations in the data, in a way that's <u>generalizable</u> to other related observations (e.g., by another experimenter, at another time, in another place).

random effects we don't model

Model $$y = \alpha + \beta x \quad \boxed{+ \text{ noise}}$$

y	outcome of measured system	(behavior)
x	predictor of measured system	(firing rate)
α, β	parameters	

<u>Note</u>: linear relationship

Regression

Note: we **cannot** observe y exactly … measurement error

We observe approximately linear relationship (corrupted by noise).

Challenge: Choose values (a, b) for parameter (α, β) in our model that "best describe" the data.

We observe $y_1, y_2, y_3, \ldots$ and $x_1, x_2, x_3, \ldots$ and fit our model

$$y = \alpha + \beta x$$

to choose the values (a, b) for parameter (α, β)

Regression

If we have (a, b), then we can compute <u>model predictions</u>:

$$\hat{y}_1 = a + bx_1$$

$$\hat{y}_2 = a + bx_2$$

$$\vdots$$

?

Choose (a, b) to make model predictions $\hat{y}_1, \hat{y}_2, \ldots$ $\boxed{\text{close}}$ to the observed outcomes $y_1, y_2, \ldots$

<u>Note</u>: Model predictions $\hat{y}_1, \hat{y}_2, \ldots$ do **not** reproduce exactly the observed outcomes $y_1, y_2, \ldots$

Regression

?

Choose (a, b) to make model predictions $\hat{y}_1, \hat{y}_2, \ldots$ $\boxed{\text{close}}$ to the observed outcomes $y_1, y_2, \ldots$

Q: "close" ?

A: A measure of discrepancy or distance

$$S_2(y, \hat{y}) = \sum_i (y_i - \hat{y}_i)^2$$ "least squares"

Choose (a, b) to <u>minimize</u> $S_2(y, \hat{y})$

to <u>minimize</u> the discrepancy between y and $\hat{y}$

Regression

Minimize $\quad S_2(y, \hat{y}) = \sum_i (y_i - \hat{y}_i)^2 \quad$ assumes

1. All observation on the same physical scale (e.g., # vs % correct)

2. Observations are independent or "exchangeable"

3. Deviations $(y_i - \hat{y}_i)$ similar for different values of y

(variability independent of mean)

Regression: estimate it

Estimate the model in Python

$$y = \alpha + \beta x$$

Task performance $= \alpha + \beta$ (firing rate)

intercept

slope

Python

Regression: estimate it

Estimate the model in Python

$$y = \alpha + \beta x$$

$$\text{Task performance} = \underset{\text{intercept}}{\alpha} + \underset{\text{slope}}{\beta}\,(\text{firing rate})$$

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:
Model:                            OLS   Adj. R-squared:
Method:                 Least Squares   F-statistic:
Date:                Mon, 07 Oct 2024   Prob (F-statistic):
Time:                        12:40:56   Log-Likelihood:
No. Observations:                  50   AIC:
Df Residuals:                      48   BIC:
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|
------------------------------------------------------------------------------
Intercept     15.0190      4.037      3.720      0.001
x              0.0158      0.404      0.039      0.969
==============================================================================
Omnibus:                        4.793   Durbin-Watson:
Prob(Omnibus):                  0.091   Jarque-Bera (JB):
Skew:                           0.459   Prob(JB):
Kurtosis:                       2.153   Cond. No.
==============================================================================
```

Interpret parameters …

Regression: plot it

Python

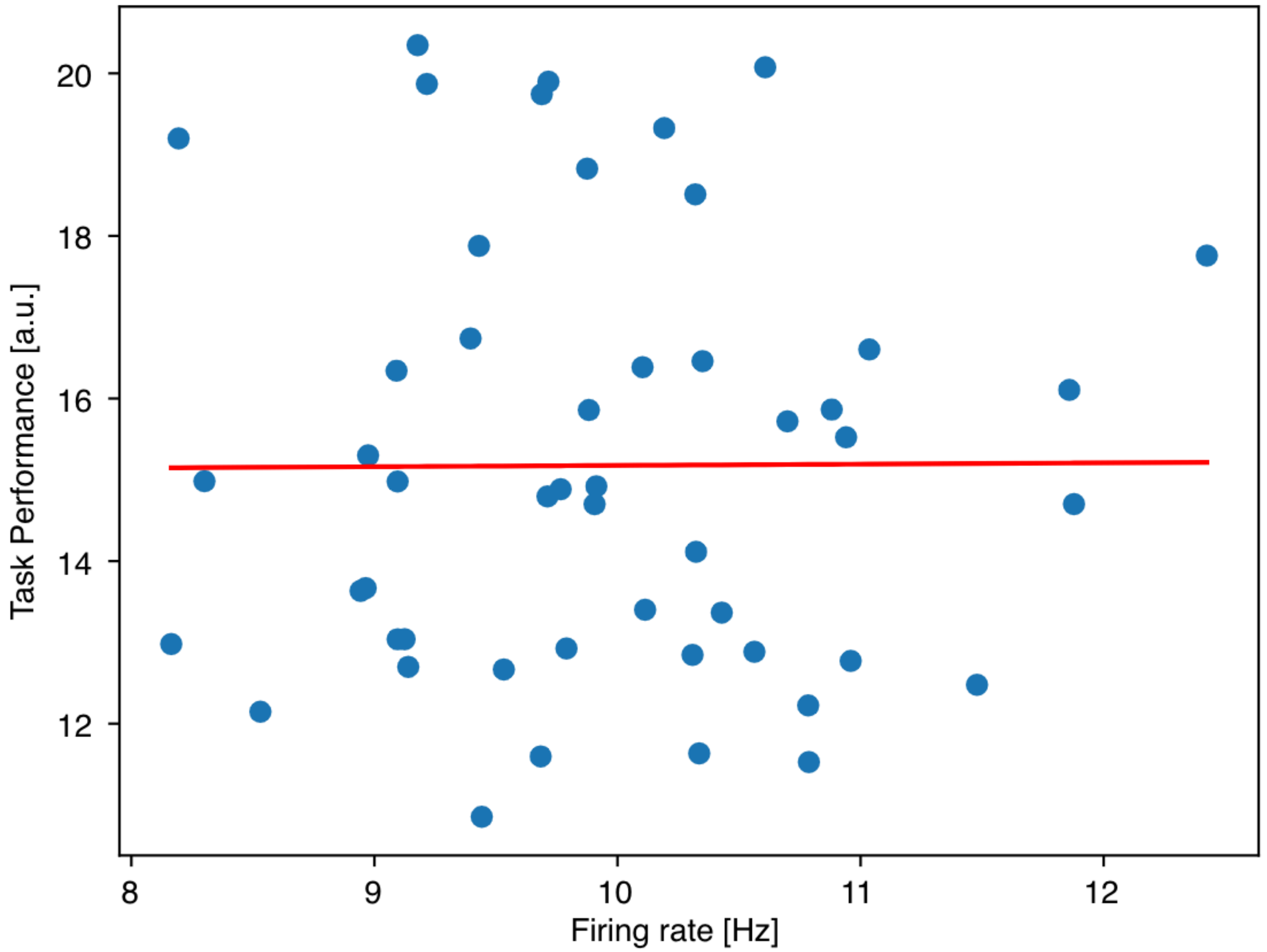
Regression: Interpret parameters

<u>Intercept</u>: $\alpha = 15.02$

- when firing rate (x) is 0, the task performance is ≈ 15

<u>Slope</u>: $\beta = 0.016$

- for each one-unit increase in firing rate, the task performance increases by 0.016.



Q: Evidence of a linear relationship between task performance and firing rate?

Regression: Interpret parameters

Q: Evidence of a linear relationship between task performance and firing rate?

A: Examine the p values

p-value: how much evidence we have to reject the null hypothesis (H_0)

Here, H_0 is that $\alpha = 0, \beta = 0$

Typically, we $\boxed{\text{reject } H_0 \text{ if } p < 0.05}$

The probability of observing the data, or something more extreme, under the null hypothesis is less than 5%.

The observed data is unlikely to have occurred by random chance alone, assuming the null hypothesis is true.

Regression: Interpret parameters

Q: Evidence of a linear relationship between task performance and firing rate?

A: Examine the <u>p values</u>

<u>Intercept</u>: $\alpha = 15.02, p = 0.001$

- Reject H_0 that intercept = 0

<u>Slope</u>: $\beta = 0.016, p = 0.969$

- No evidence to reject H_0 that slope = 0.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:
Model:                            OLS   Adj. R-squared:
Method:                 Least Squares   F-statistic:
Date:                Mon, 07 Oct 2024   Prob (F-statistic):
Time:                        12:40:56   Log-Likelihood:
No. Observations:                  50   AIC:
Df Residuals:                      48   BIC:
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|
------------------------------------------------------------------------------
Intercept     15.0190      4.037      3.720      0.001
x              0.0158      0.404      0.039      0.969
==============================================================================
```

<u>Note</u>: Never accept H_0. ~~We cannot conclude slope = 0~~

Instead: "*We fail to reject the null hypothesis that slope = 0.*"

CAS MA 665 A1 - Introduction to Modeling and Data Analysis in Neuroscience

Student

https://go.blueja.io/ie-TXlIb1kyOD50Y_F6mqg

Regression: conclusion (for now)

We considered this model:

Task performance $= \alpha + \beta$ (firing rate)

We found no evidence to reject the null hypothesis that $\beta = 0$.

We conclude that, in this model, we have no evidence of a relationship between task performance and firing rate.

Now what?

Regression: continued

Q: Now what?

A: Look for confounds.

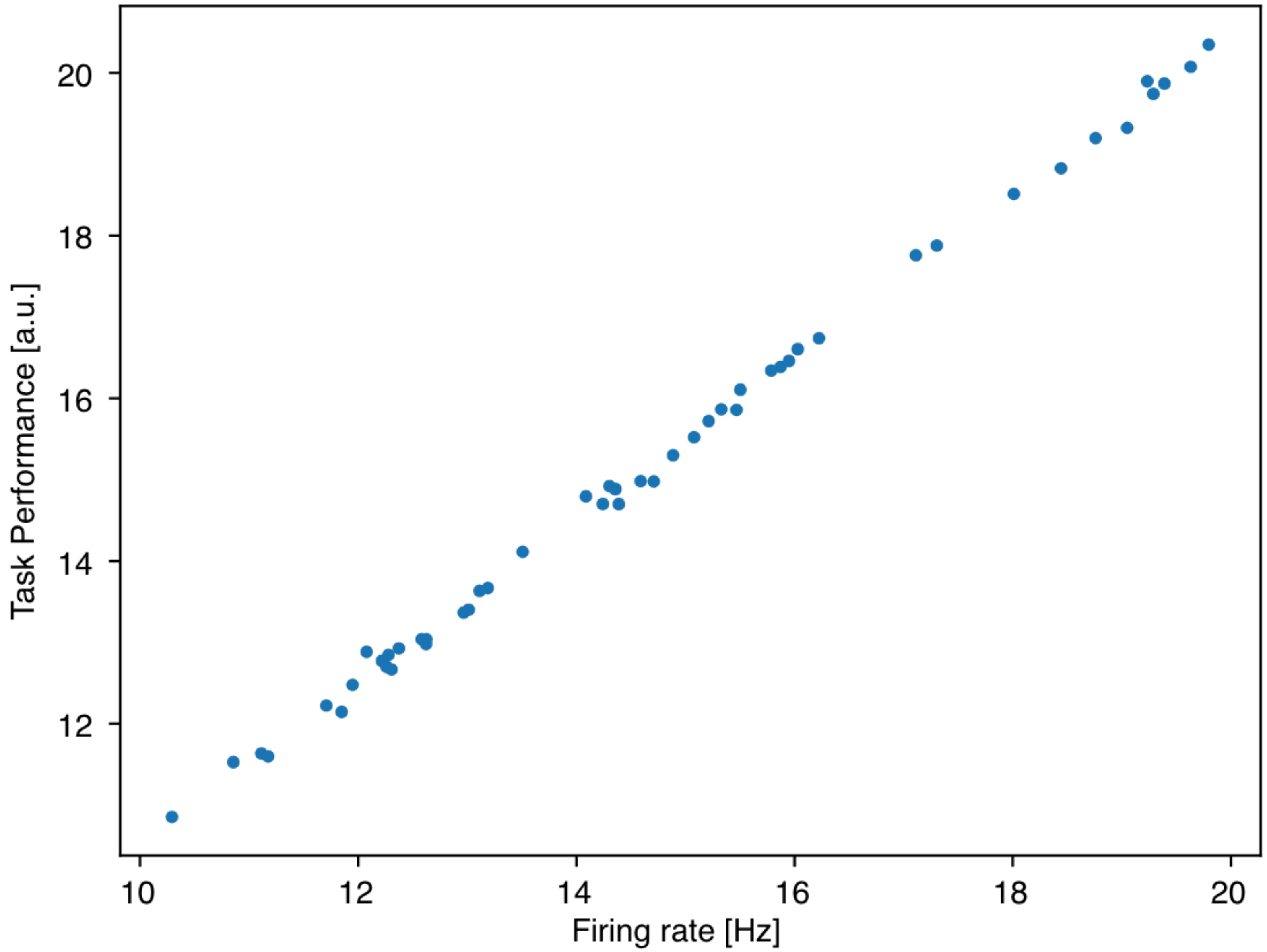
We learn that age impacts task performance

New variables:

y	task performance
x_1	firing rate
x_2	age

Analyze the data (1)

Plot it task performance <u>versus age</u>



Visual inspection:

Analyze the data (2)

Compute the correlation between task performance and age.

Python

$$C_{xy} =$$

Conclusion:

Analyze the data (3): Regression

Model $$y = \alpha + \beta_1 x_1 + \beta_2 x_2$$

$$\text{Task performance} = \alpha + \boxed{\beta_1}\,(\text{firing rate}) + \boxed{\beta_2\,(\text{age})}$$

parameter of interest

confound

Q: What is the relationship between task performance (y) and firing rate (x_1) after accounting for the confound of age (x_2)?

Analyze the data (3): Regression

Python

Regression: Interpret parameters

Intercept: $\alpha =$ $p =$

Slope (firing rate): $\beta_1 =$ $p =$

Slope (age): $\beta_2 =$ $p =$

	coef	std err	t	P>|t|
Intercept	0.0656	0.178	0.368	0.714
firing_rate	0.0466	0.016	2.961	0.005
age	0.9977	0.006	177.974	0.000

Regression: Plot the model

Python

Regression: conclusion (modified)

We considered the <u>updated model</u>:

$$\text{Task performance } = \alpha + \beta_1 \text{ (firing rate)} + \beta_2 \text{ (age)}$$

We found

We conclude that

What is a "good model" ?

A: A model that makes predictions $\hat{y}$ very close to y.

To do so, add more predictors (and parameters) to the model.

$$y = \alpha + \beta x_1 + \beta x_2 + \beta x_3 + \beta x_4 + \beta x_5 + \ldots$$

No reduction in complexity.

We want a simple theoretical pattern (e.g., line) for our ragged data

parsimony of parameters (only include what we need)

What is a "good model" ?

Parsimonious model

- easier to think about
- probably makes better prediction

Modeling is an art — no formal procedure, requires imagination

"*All models are wrong but some are useful.*" [George Box]

eternal truth not within our grasp

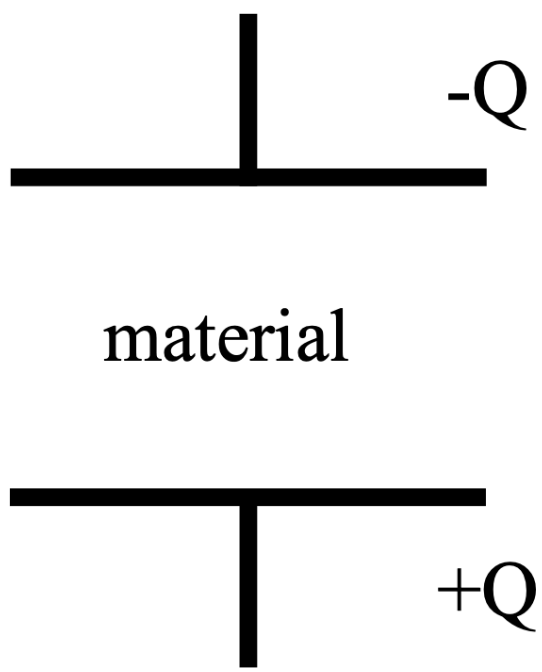
use those

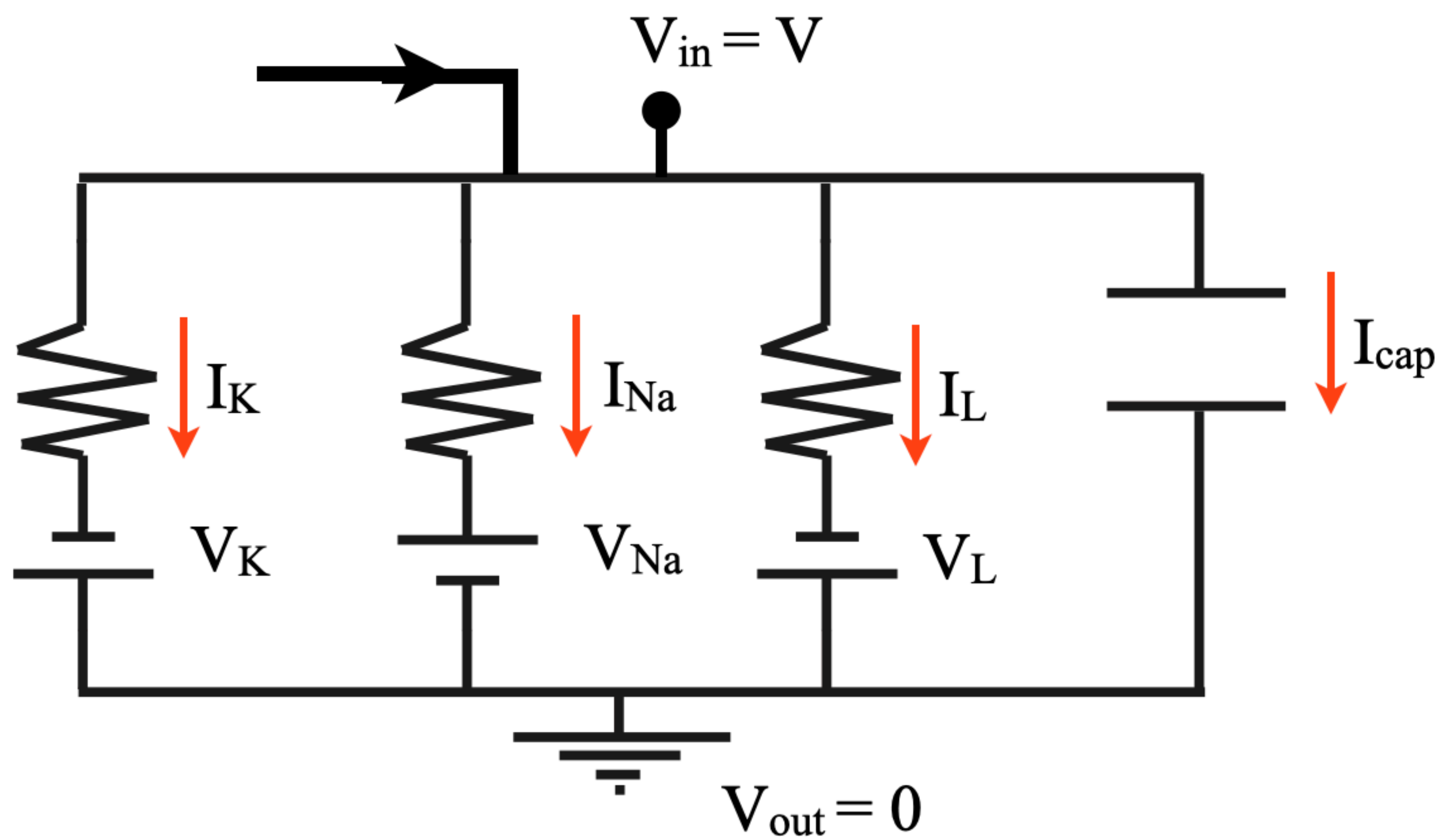
Check your model — look at errors or deviations $(y_i - \hat{y}_i)$

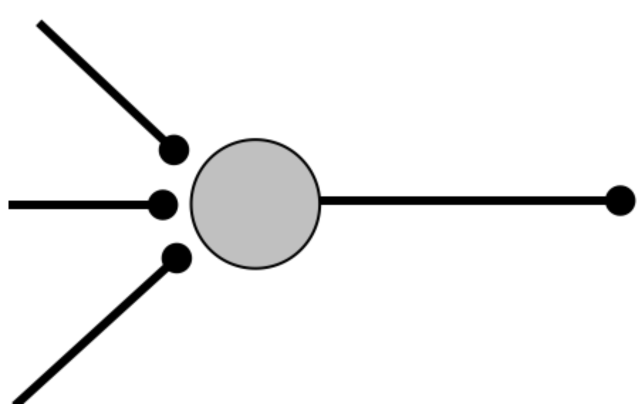
important but not covered here

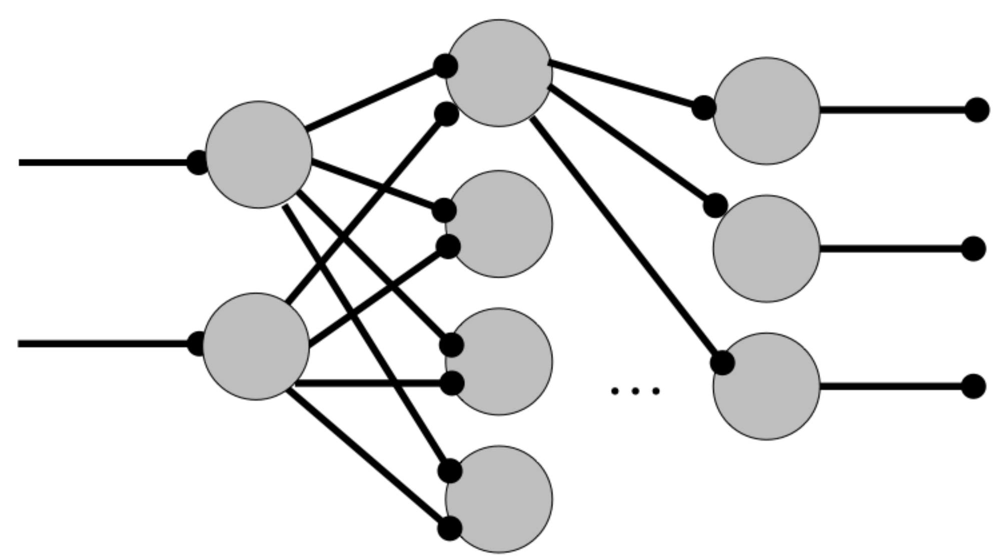
What is a model?

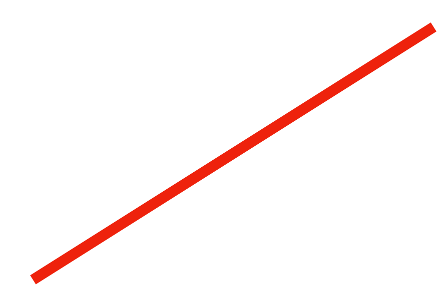
In MA665:











$$y = mx + b$$

What is computational neuroscience?

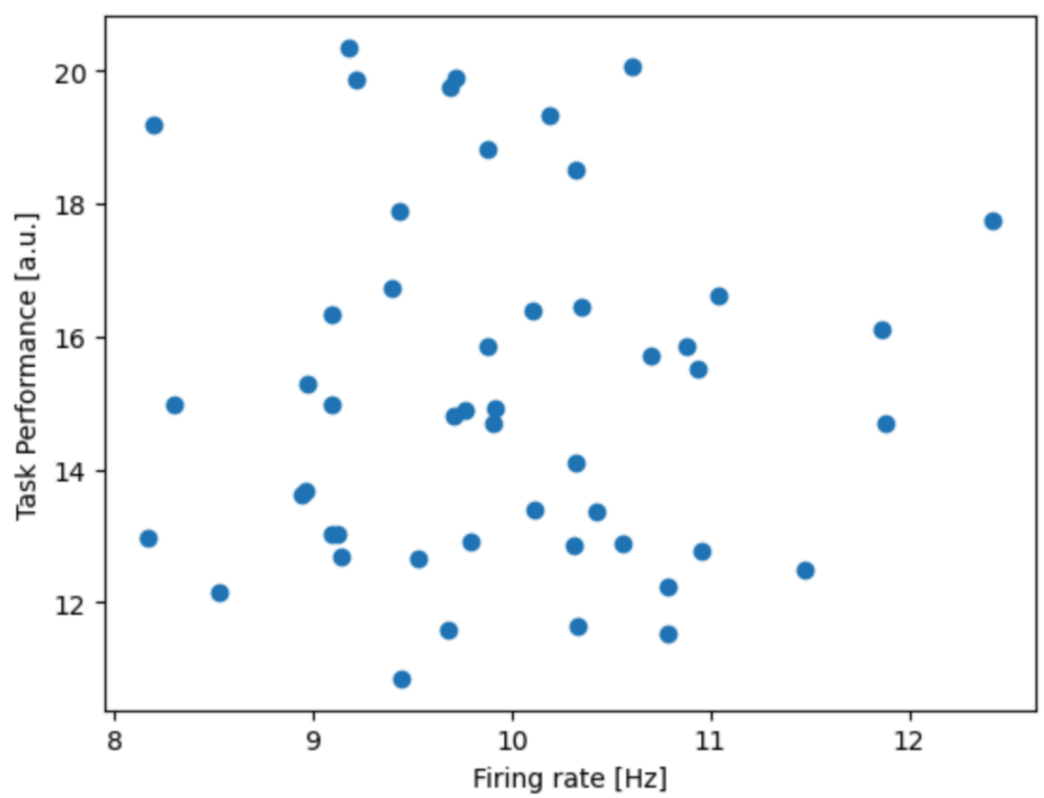
Mathematics:

$$
\begin{aligned}
C\,\frac{dV}{dt} &= I_{\text{input}}(t) - \bar{g}_{\text{K}}\, n^4 (V - V_{\text{K}}) - \bar{g}_{\text{Na}}\, m^3 h (V - V_{\text{Na}}) - \bar{g}_{\text{L}} (V - V_{\text{L}}) \\
\frac{dn}{dt} &= -\frac{n - n_\infty(V)}{\tau_n(V)} \\
\frac{dm}{dt} &= -\frac{m - m_\infty(V)}{\tau_m(V)} \\
\frac{dh}{dt} &= -\frac{h - h_\infty(V)}{\tau_h(V)}\,,
\end{aligned}
$$

Statistics:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.021
Method:                 Least Squares   F-statistic:                  0.001521
Date:                Mon, 07 Oct 2024   Prob (F-statistic):              0.969
Time:                        12:40:56   Log-Likelihood:                -119.04
No. Observations:                  50   AIC:                             242.1
Df Residuals:                      48   BIC:                             245.9
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
                           3.720      0.001      6.901      23.137
                           0.039      0.969     -0.797       0.829
==============================================================================
                              .793   Durbin-Watson:                   1.865
                              .091   Jarque-Bera (JB):                3.249
                              .459   Prob(JB):                        0.197
                              .153   Cond. No.                         108.
==============================================================================
```

Data:



Aside: C4R

NIH NINDS C4R

Home About Us Resources Blog Contact Join

Community for Rigor

Better Science Every Day

Welcome to the Community for Rigor! We are a free, open resource to help researchers of all kinds learn, practice, and promote scientific rigor.

Aside: Sample Size

https://tinyurl.com/2bm86mxn

Power/Sample Size Challenge Question

- Imagine that we have a scientific hypothesis based on previous work that suggests that substance x is a genetic biomarker for longevity (i.e., age at death).
- Before conducting an experiment to test the predictive power of this novel biomarker, we need to compute the **sample size** for our experiment.
- We will see that the sample size required to generate data that can support a scientific hypothesis depends directly on the prior beliefs and knowledge about that hypothesis.
- Here is the setup for the problem: suppose that we have only the following limited information about substance x and longevity:
    - *People have a normal distribution of expression of substance x.*
    - *Individuals at the high end of expression levels tend to live about 5 years longer than people at the low end.*

Aside: Sample Size

https://mark-kramer.github.io/METER-Units/

BU METER

Sample Size - How much data is enough for your experiment?

- Interactive notebook

Evaluate your evaluation methods! A key to meaningful inference.

- Interactive notebook

Putting the p-value in context: p<0.05, but what does it REALLY mean?

- Static notebook

Reproducible exploratory analysis: Mitigating multiplicity when mining data

- Static notebook

Aside: Sample Size

Q: Is there a relationship between x and lifespan?

A1: Do an experiment with sample size N.

A2: Fit a line…

$$lifespan = \beta_0 + \beta_1 x$$

$\beta_1 =$

p =

Conclusion:

Aside: Sample Size

Q: Now what?

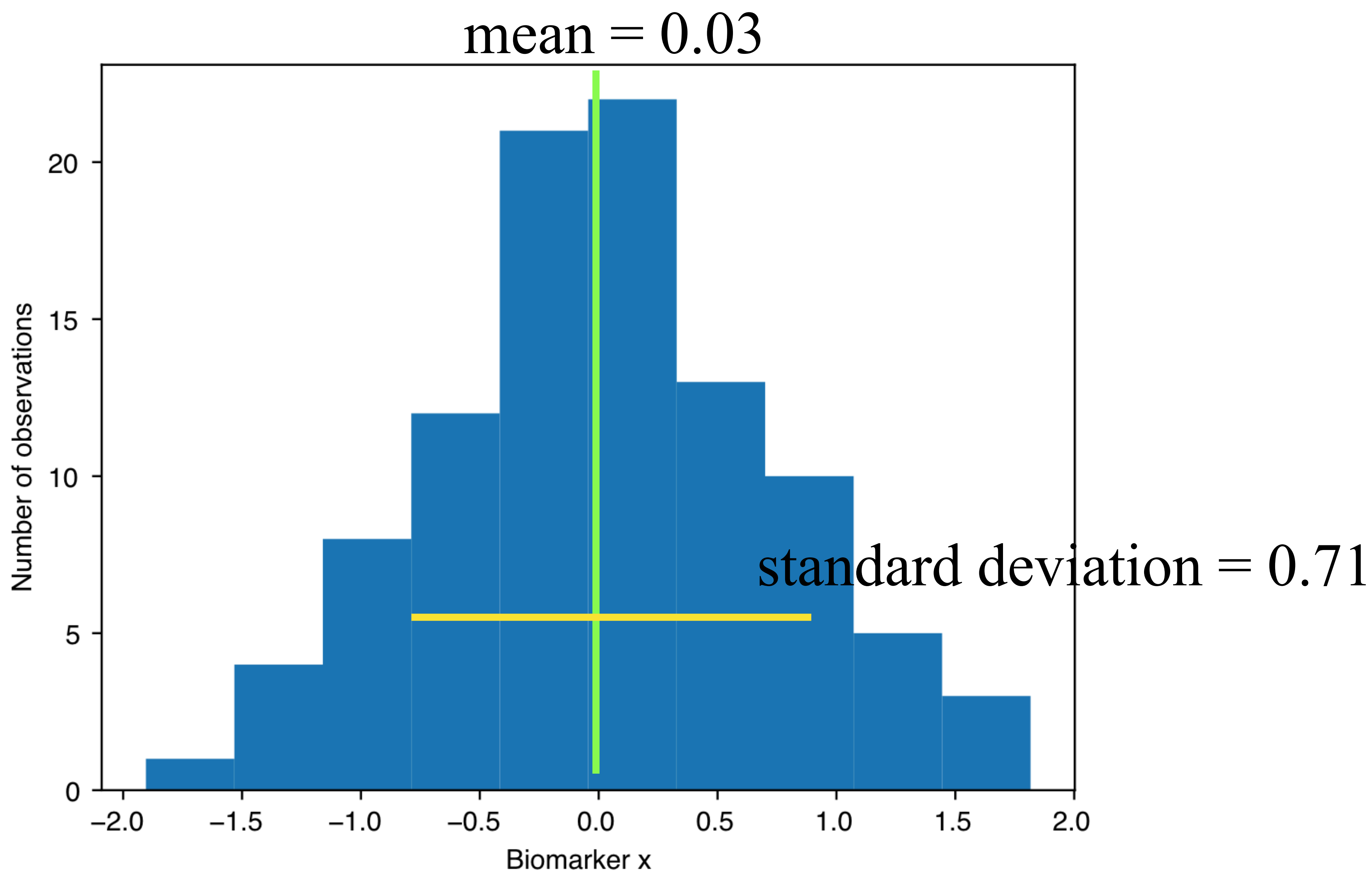
A: Maybe we failed to collect enough data to detect a relationship.

<u>Idea</u>:

- Reuse the data & model
- See how sample size (N) impacts conclusions.

Aside: Sample Size

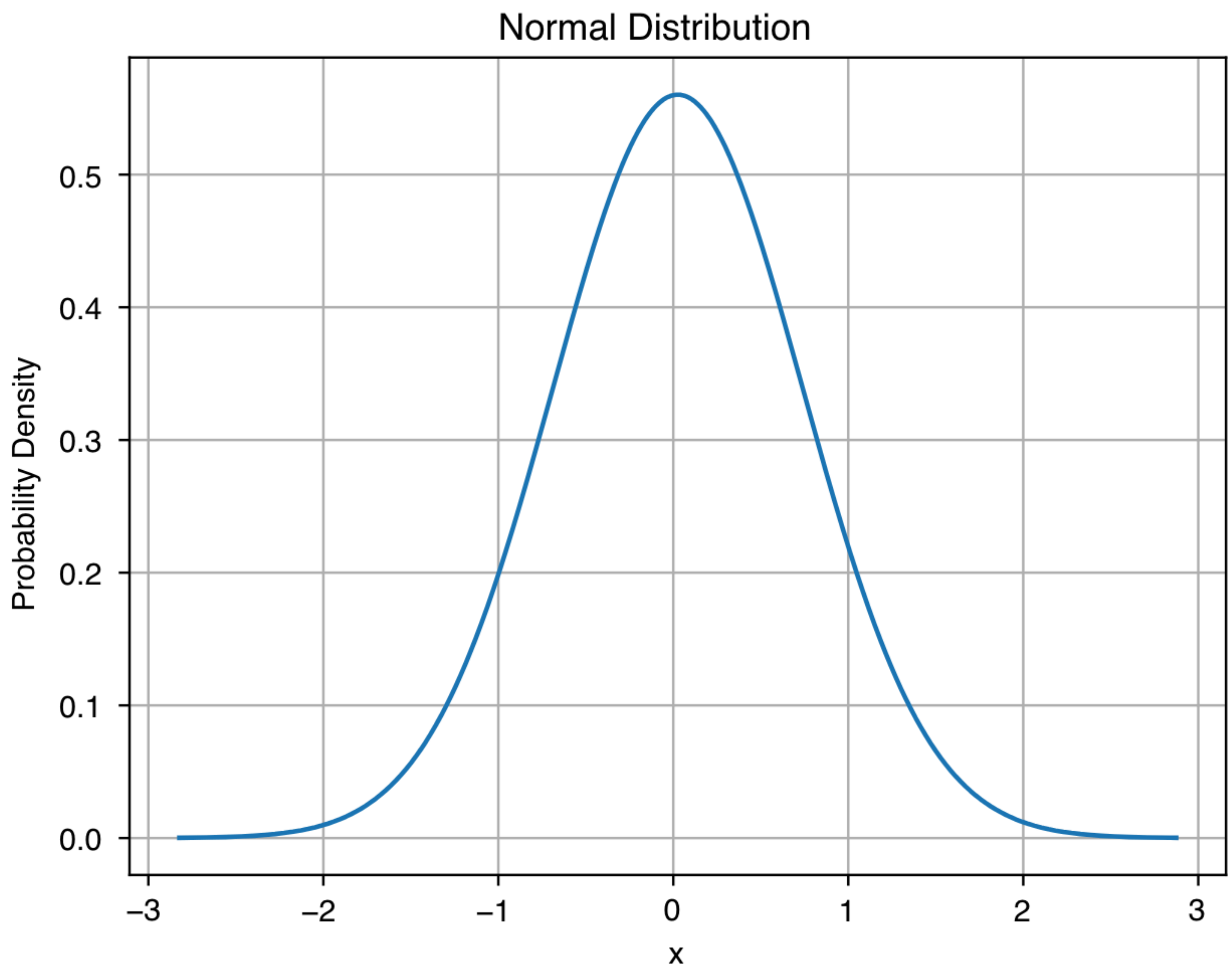
Consider biomarker x



Approximately normal

Aside: Sample Size

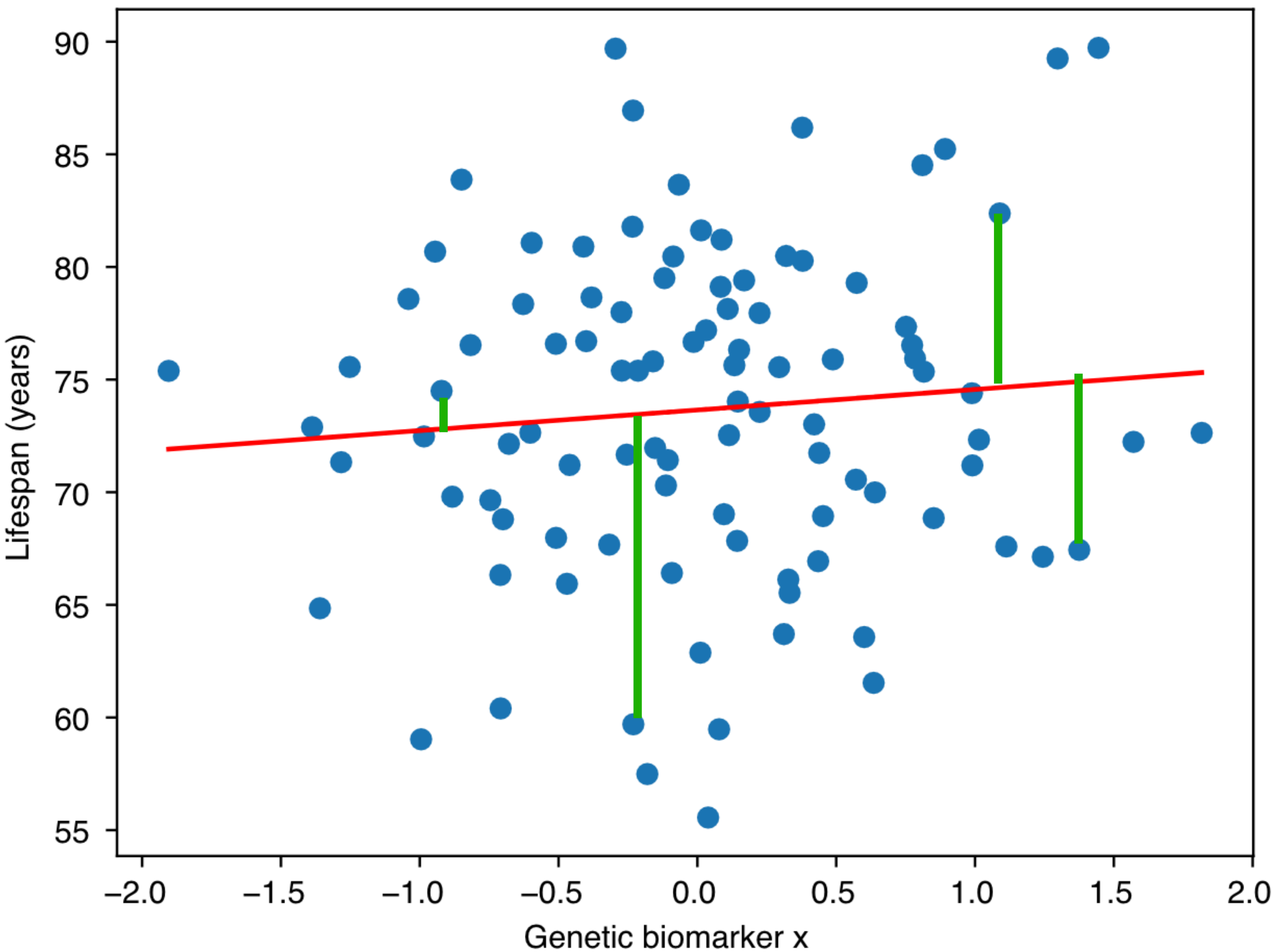
We can draw <u>random values of x</u> from this normal distribution



Draw 10 or 100 or 1000 or 10,000 values for x …

Aside: Sample Size

Consider model: $lifespan = \beta_0 + \beta_1 x$



$\beta_0 = 73.65$ (intercept)
$\beta_1 = 0.91$ (slope)

There's error in our model

Normally distributed:
mean ≈ 0
stand. dev. ≈ 7

To simulate new lifespans:

- Ask the model
- Include the error

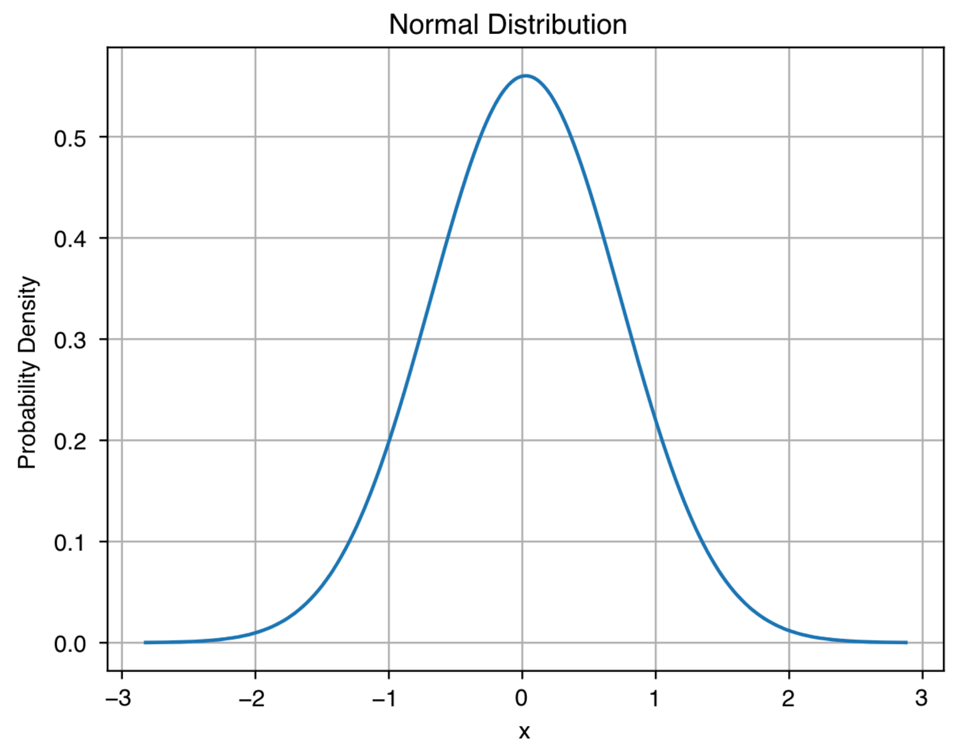
new lifespan = $\beta_0 + \beta_1 x$ + error

Aside: Sample Size

Create new data:

- Pick new sample size N*
- Draw new biomarkers x
- Draw new lifespans

  new lifespan = $\beta_0 + \beta_1 x$ + error



<u>Key insight</u>: Is there a relationship between x & lifespan <u>in the new data</u>?

Fit a (new) model: new lifespan = $\beta_0^* + \beta_1^*$ new x

Q: At what new sample size N* do you reliably detect a relationship?

… is $p < 0.05$ reliably.