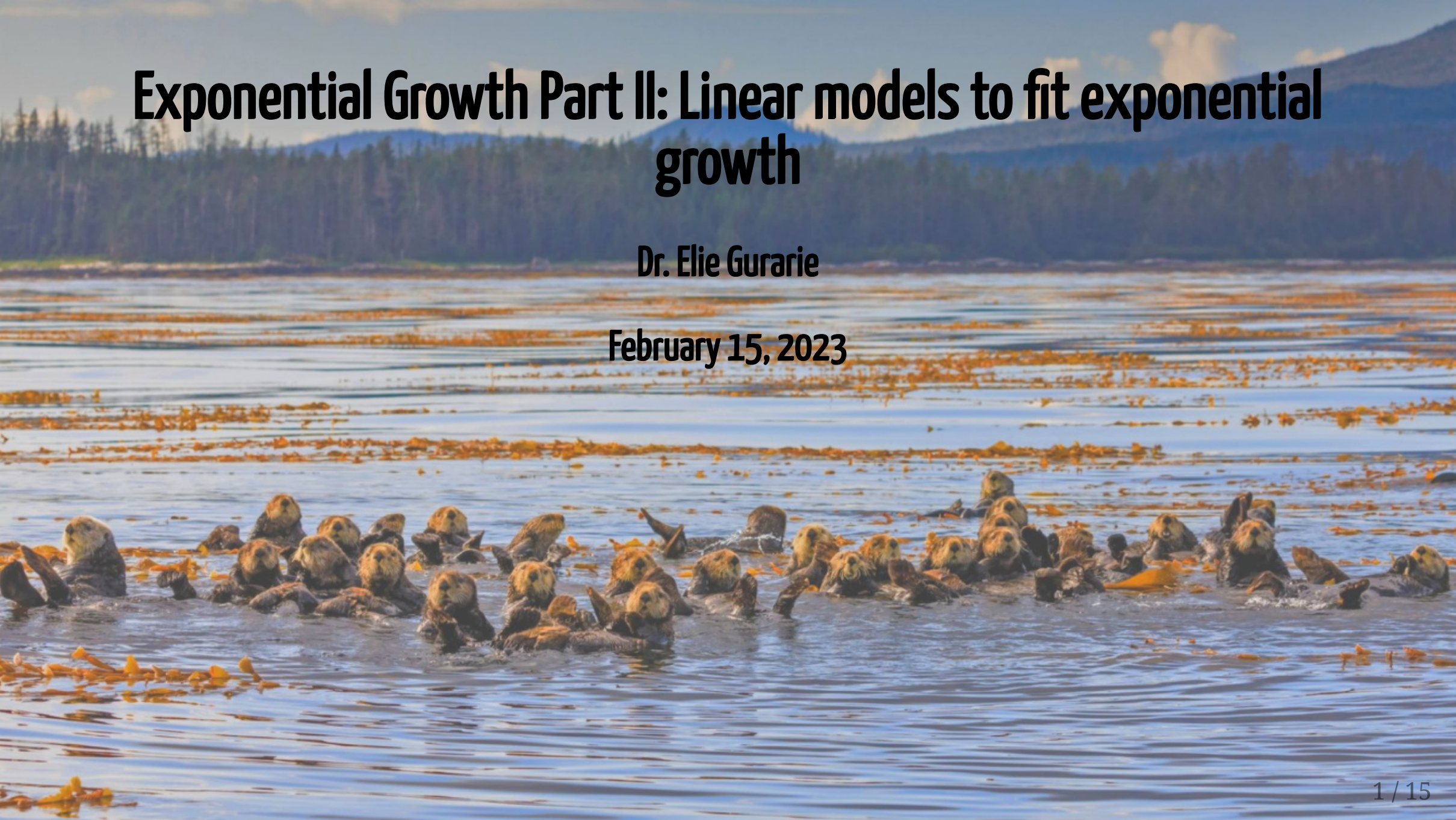


Exponential Growth Part II: Linear models to fit exponential growth

Dr. Elie Gurarie

February 15, 2023



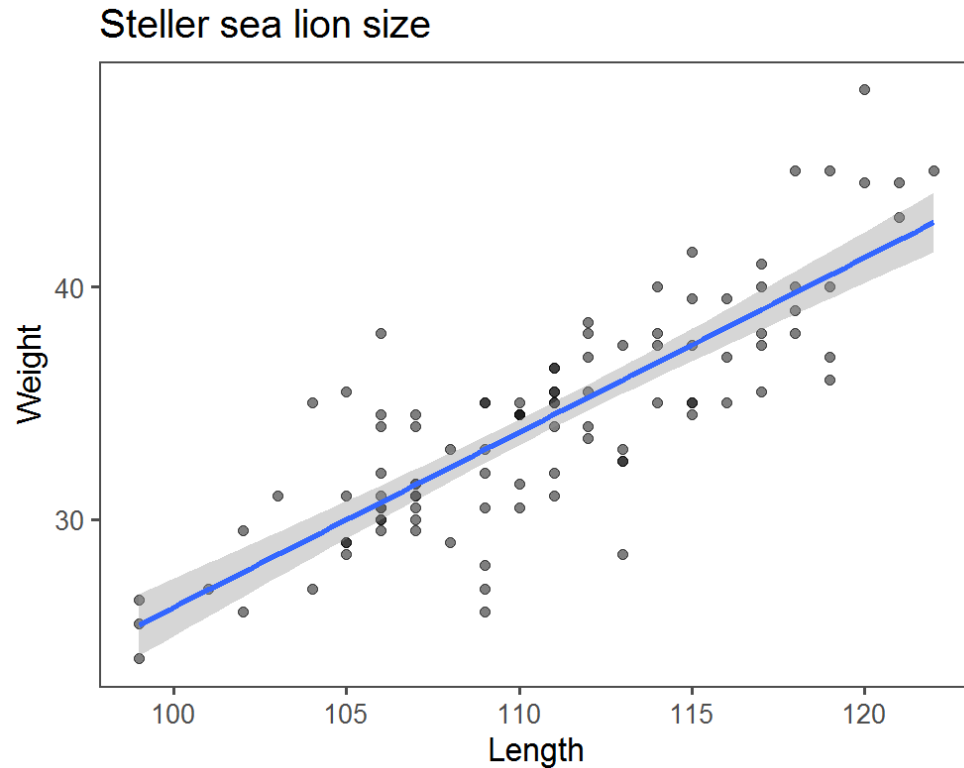
Steller sea lion (*Eumatopias jubatus*) - birth



Linear modeling

(aka **REGRESSION**, except I really don't like that term, for a variety of reasons to discuss in class.)

is a very general method to quantifying relationships among variables.



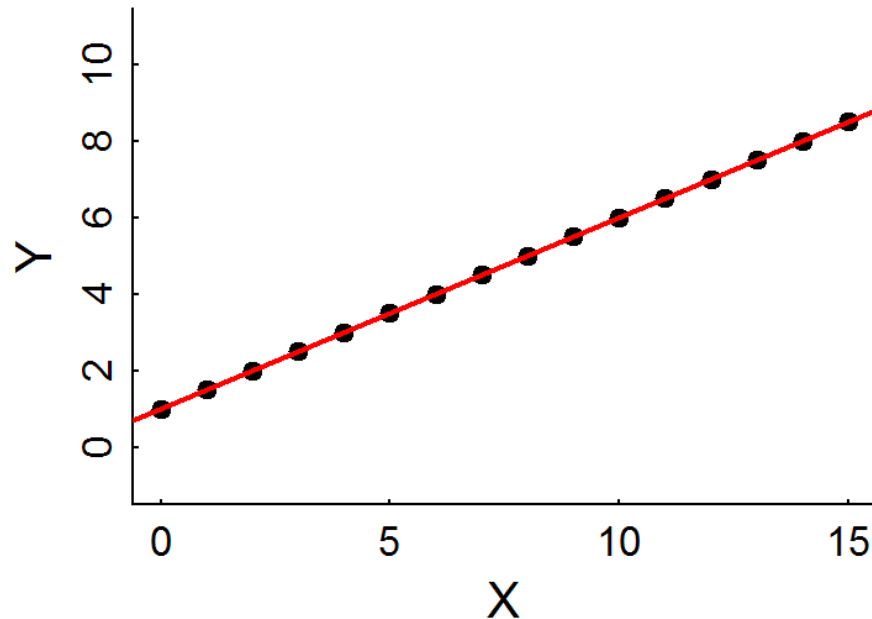
Linear Models

Deterministic:

$$Y_i = a + bX_i$$

a - intercept; b - slope

$a = 1$; $b = 1/2$



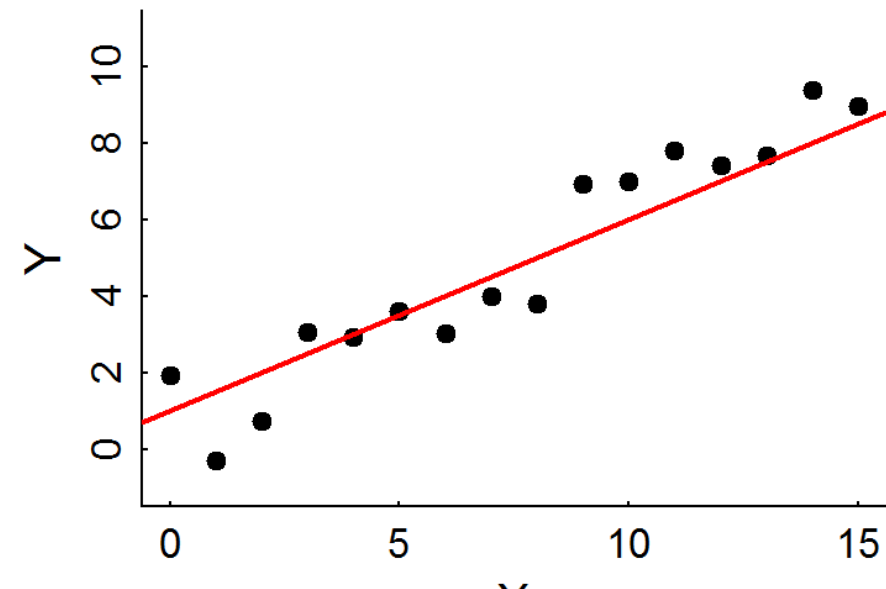
Statistical:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

α - intercept; β - slope; ϵ - **randomness!**

$$\epsilon_i \sim \mathcal{N}(0, \sigma)$$

$\alpha = 1$; $\beta = 1/2$; $\sigma = 1$



Fitting models is easy in



Point Estimate

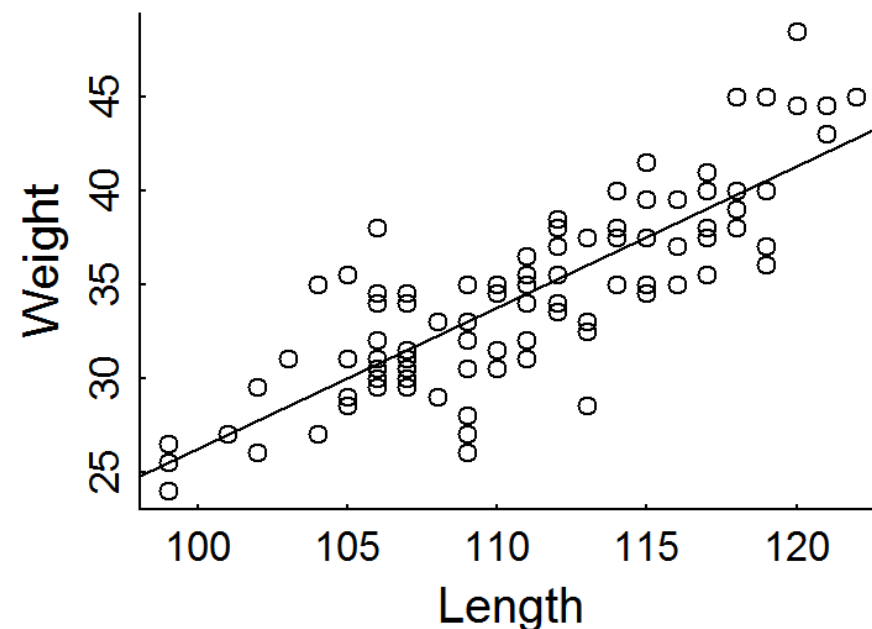
This command fits a model:

```
lm(Weight ~ Length, data = pups)
```

```
##  
## Call:  
## lm(formula = Weight ~ Length, data = pups)  
##  
## Coefficients:  
## (Intercept)      Length  
##    -49.1422      0.7535
```

So for **each 1 cm** of length, add another **754 grams**.

```
plot(Weight ~ Length, data = pups)  
abline(my_model)
```



The `abline` puts a line, with intercept a and slope b onto a figure.

Some comments on linear models

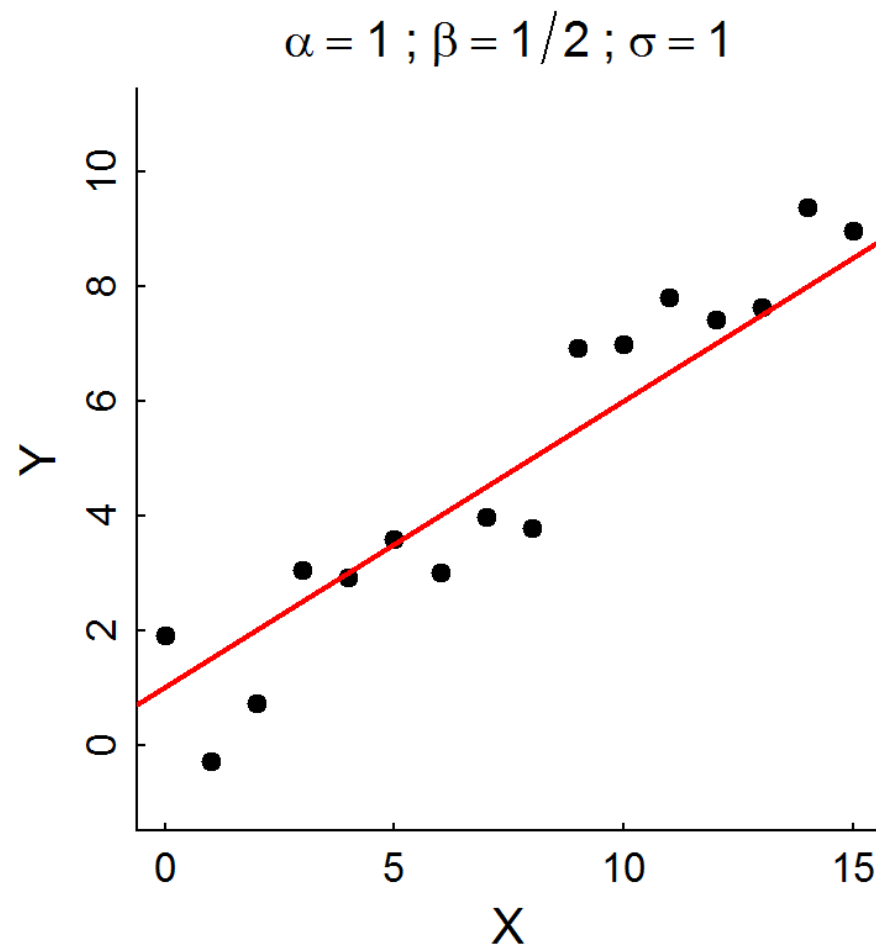
$$Y_i \sim \alpha + \beta X_i + \epsilon_i$$

1. ϵ_i is **unexplained variation** or **residual variance**. It is often (erroneously, IMO) referred to as "**error**". It is a **random variable**, NOT a **parameter** or **data**.

2. $\alpha + \beta X_i$ is the **predictor**, or the "**modeled**" portion. There can be any number of variables in the **predictor** and they can have different powers, so:

$$Y_i \sim \mathcal{N}(\alpha + \beta X_i + \gamma Z_i + \delta X_i^2 + \nu X_i Z_i, \sigma)$$

is also a **linear** model.



Statistical inference

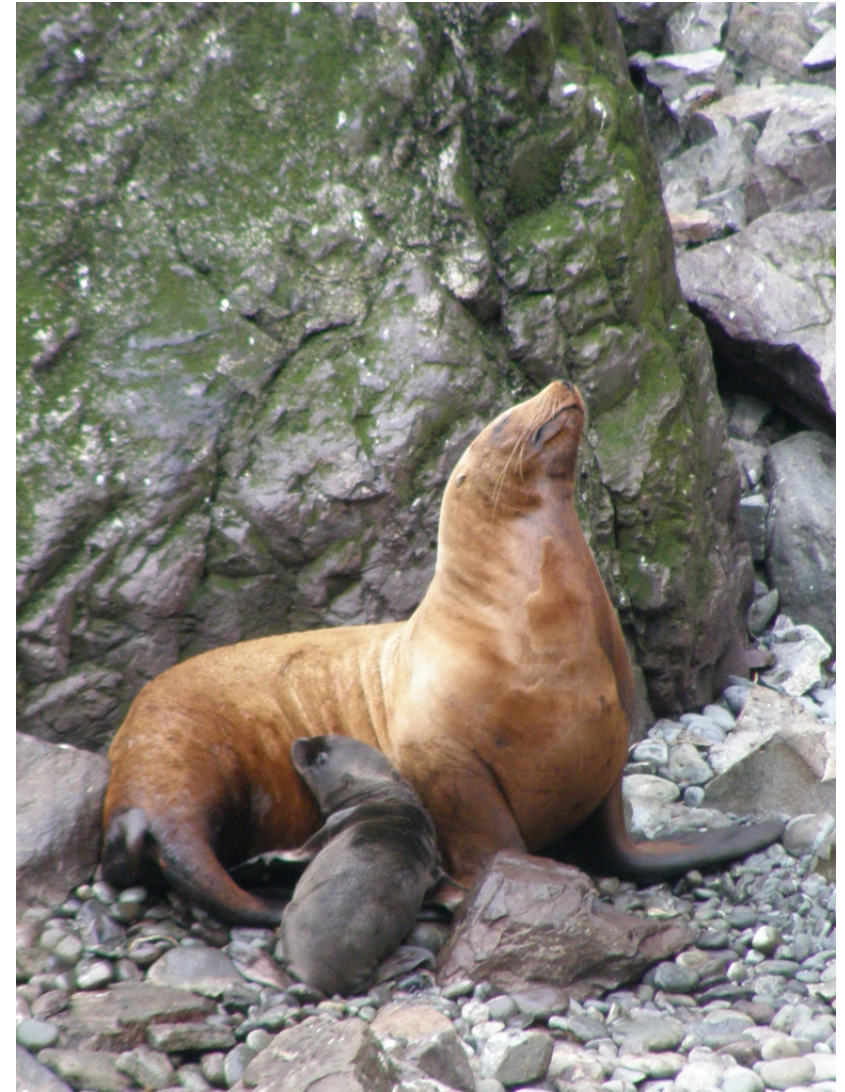
Statistical inference is the *science / art* of observing *something* from a **portion of a population** and making statements about the **entire population**.

In practice - this is done by taking **data** and **estimating parameters** of a **model**. (This is also called *fitting* a model).

Two related goals:

1. obtaining a **point estimate** and a **confidence interval** (precision) of the parameter estimate.
2. Assessing whether particular (combinations of) factors, i.e. **models**, provide any **explanatory power**.

This is (almost always) done using **Maximum Likelihood Estimation**, i.e. an algorithm searches through possible values of the parameters that make the model **MOST LIKELY** (have the highest probability) given the data.



Another gratuitous sea lion picture.

Statistical output

```
```\n##\n## Call:\n## lm(formula = Weight ~ Length, data = pups %>% subset(Island ==\n##      "Raykoke"))\n##\n## Residuals:\n##      Min       1Q   Median       3Q      Max \n## -7.498 -1.718  0.023  1.764  7.276 \n##\n## Coefficients:\n##              Estimate Std. Error t value Pr(>|t|)\n## (Intercept) -49.14222    5.75796  -8.535 1.81e-13 ***\n## Length      0.75345     0.05193  14.510 < 2e-16 ***\n## ---\n## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1\n##\n## Residual standard error: 2.761 on 98 degrees of freedom\n## Multiple R-squared:  0.6824,    Adjusted R-squared:  0.6791 \n## F-statistic: 210.5 on 1 and 98 DF,  p-value: < 2.2e-16\n```\n
```

## 1. Point estimates and confidence intervals

**Intercept ( $\alpha$ ):  $-49.14 \pm 11.5$**

**Slope ( $\beta$ ):  $0.75 \pm 0.104$**

## 2. Is the model a good one?

$p$ -values are very very small, in particular for **slope**

Proportion of variance explained is high:

$$R^2 = 0.68$$

# Statistical output

```
```\n##\n## Call:\n## lm(formula = Weight ~ Length, data = pups %>% subset(Island ==\n##       "Raykoke"))\n##\n## Residuals:\n##      Min       1Q   Median       3Q      Max \n## -7.498 -1.718  0.023  1.764  7.276 \n##\n## Coefficients:\n##              Estimate Std. Error t value Pr(>|t|)\n## (Intercept) -49.14222    5.75796  -8.535 1.81e-13 ***\n## Length      0.75345     0.05193  14.510 < 2e-16 ***\n## ---\n## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1\n##\n## Residual standard error: 2.761 on 98 degrees of freedom\n## Multiple R-squared:  0.6824,    Adjusted R-squared:  0.6791 \n## F-statistic: 210.5 on 1 and 98 DF,  p-value: < 2.2e-16\n```\n
```

Interpreting statistical results

The "standard error" around the **Length** factor is 0.05.

The "true value" lies within **TWO** standard errors of the **point estimate** with 95% probability.

So the estimate of the slope with **confidence interval** is (in g/cm):

$$\hat{\beta} = 754 \text{ g/cm} \pm 104$$

The p -value around the **Length** factor is $< 2 \times 10^{-16}$.. i.e. **0** This says that there is NO chance that you would get this steep a slope if there were NO relationship between Length and Weight (the null hypothesis).

So we've performed both **estimation** and **hypothesis testing** with this model.

Models and Hypotheses

Every p -value is a Hypothesis test.

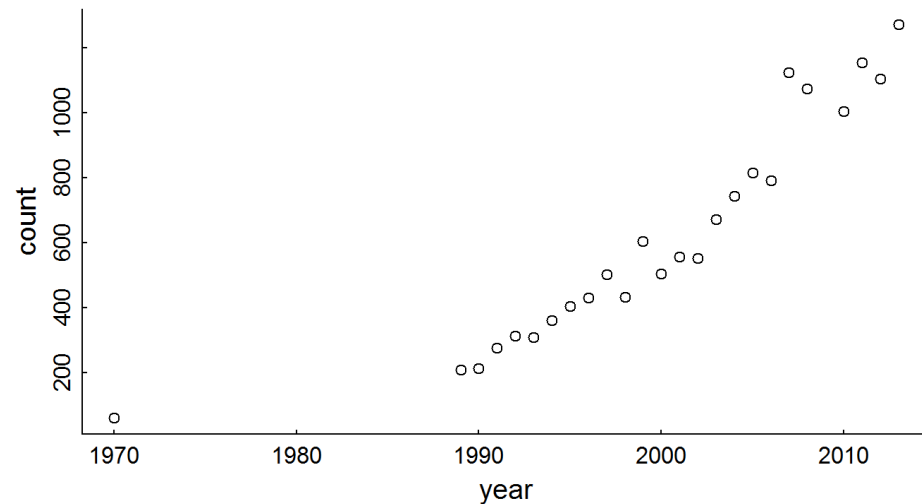
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-49.142	5.758	-8.535	0
Length	0.753	0.052	14.510	0

- First hypothesis test: H_0 **intercept** = 0
- Second hypothesis: H_0 **slope** = 0

Both null-hypotheses strongly rejected.

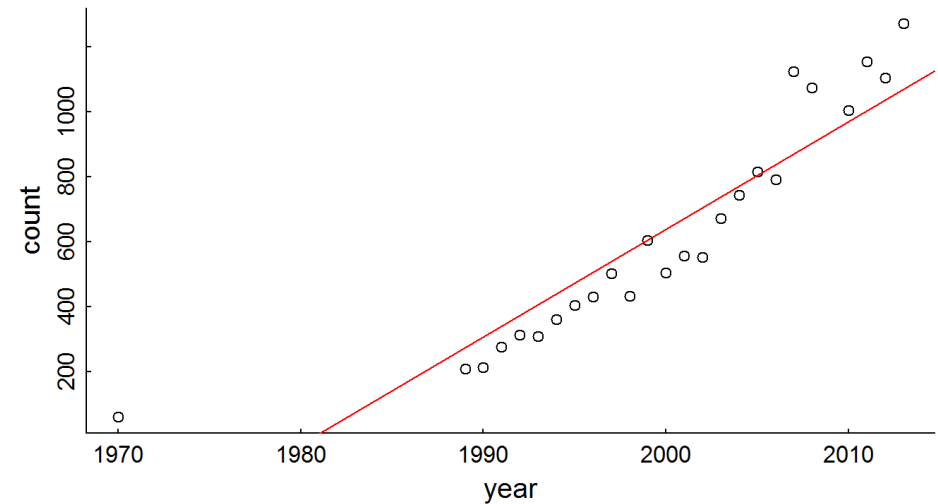
WA sea otter data:

```
WA <- read.csv("data/WA_SeaOtters_PopGrowth.c  
plot(WA)
```



Fit a linear model

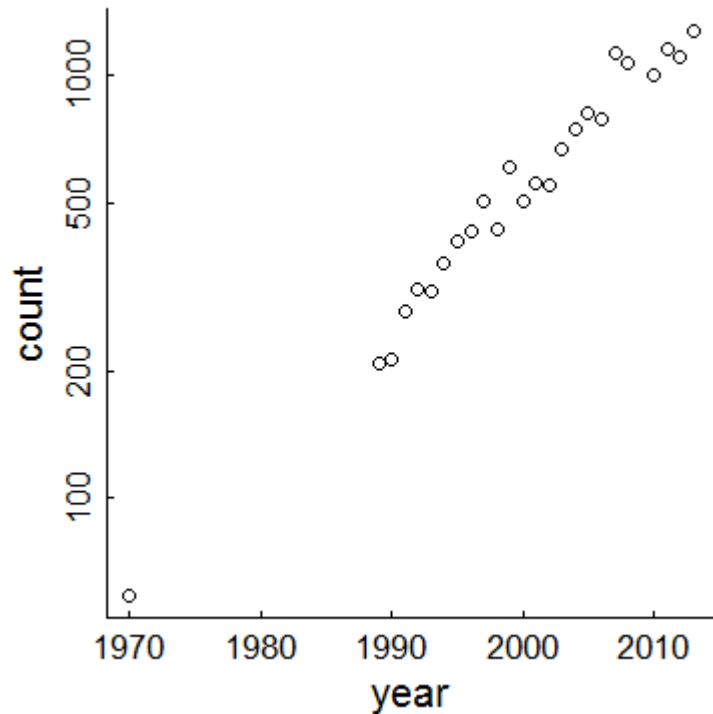
```
WA_lm <- lm(count ~ year, data = WA)  
plot(WA); abline(WA_lm, col = "red")
```



What are some problems with this model1?

Plot on Log scale: Much more linear looking!

```
plot(WA, log = "y")
```



Linear model of $\log(count)$

```
logWA_lm <- lm(log(count) ~ year, data = WA)  
logWA_lm
```

```
##  
## Call:  
## lm(formula = log(count) ~ year, data = WA)  
##  
## Coefficients:  
## (Intercept)          year  
## -140.22274         0.07325
```

Linear model of *log(count)*

```
logWA_lm <- lm(log(count) ~ year, data = WA)
logWA_lm

##
## Call:
## lm(formula = log(count) ~ year, data = WA)
##
## Coefficients:
## (Intercept)          year
## -140.22274         0.07325
```

A little math:

$$\log(N_i) = \alpha + \beta Y_i$$

$$N_i = \exp(\alpha) \times \exp(\beta Y_i)$$

$$N_i = e^\alpha e^{\beta Y_i}$$

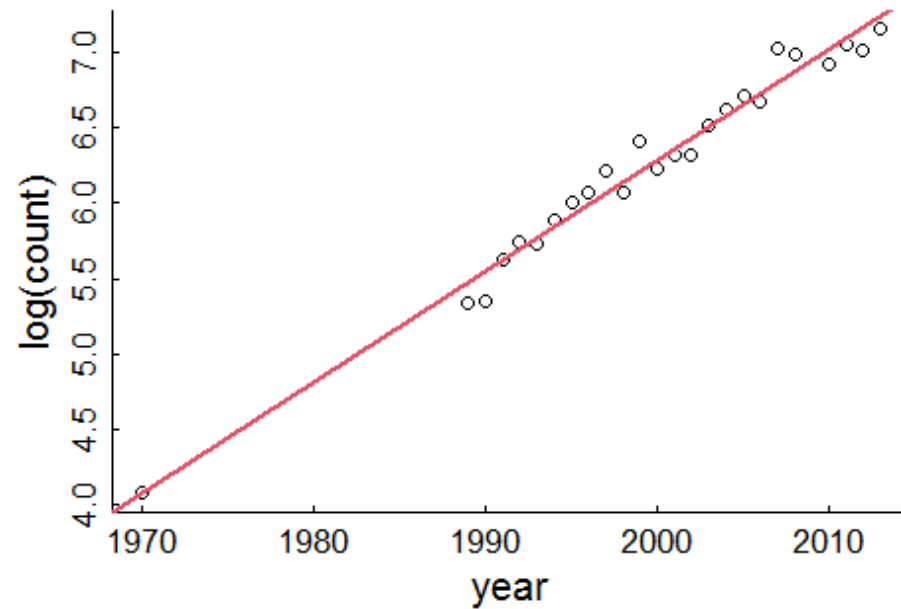
$$N_i = N_0 \lambda^{Y_i}$$

$$\lambda = e^\beta = e^{0.07325} = 1.076$$

SO ... percent rate of growth is about 7.6%.

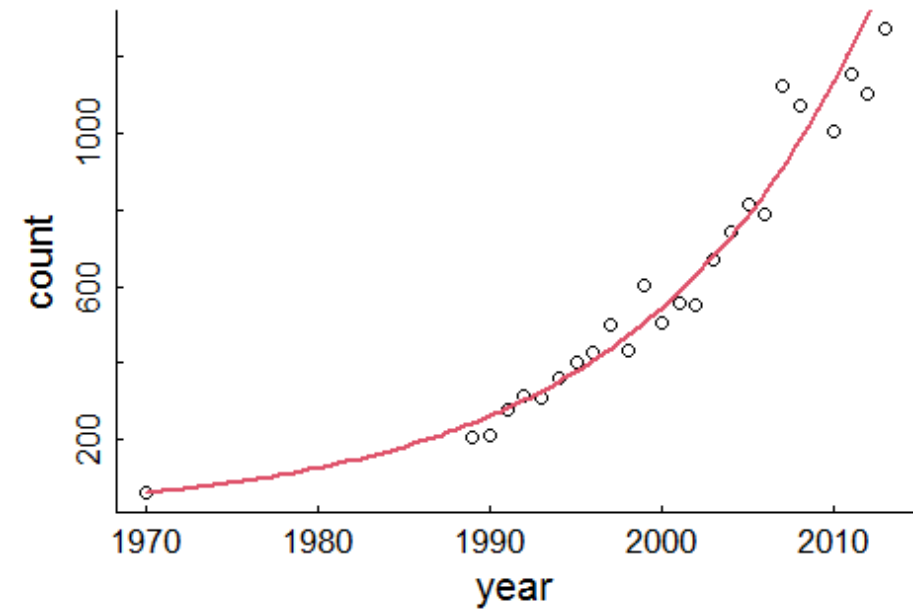
Plot linear model fit

```
plot(log(count)~year, data = WA)  
abline(lm(log(count)~year, data = WA), col =
```



Plot exponential growth

```
plot(count~year, data = WA)  
curve(exp(-140.2 + 0.07325 * x), add = TRUE,
```



Nice fit!

Summary stats and Confidence intervals

Summary stats

```
summary(logWA_lm)
```

##		Estimate	Std. Error	t value	Pr(> t)
##	(Intercept)	-140.2227	4.7318	-29.6344	0
##	year	0.0733	0.0024	30.9533	0

95% confidence intervals

$$\hat{\beta} = 0.073 \pm 2 \times 0.0024 = \{0.068, 0.078\}$$

$$\hat{\lambda} = \exp(0.073 \pm 2 \times 0.0024) = \{1.071, 1.081\}$$

So annual growth rate is $7.6\% \pm 0.5$, with 95% Confidence.

Key takeaway: With linear modeling we can use ALL the data to (a) get a great **point estimate** and (b) quantify **uncertainty** on that estimate.