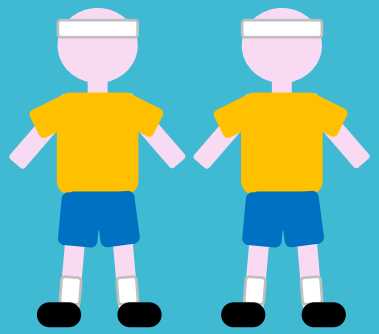
Linear regression: Part 2



Lecture Outline

A bit more on fitting

Adding uncertainty

Interpretation of results

How do the results fit in the scientific process?

Lecture Outline

A bit more on fitting

- EX1: Fit regression for 100m times

Adding uncertainty

- EX2: Calculate confidence intervals

Interpretation of results

- EX3: Interpret the results
- EX4: Prediction

How do the results fit in the scientific process?

- EX5: Discuss further steps/good models

A bit more on fitting

Recap of yesterday

Fit a regression line by minimizing sum of squares

This is also the likelihood

But it requires some assumptions:

- Normally distributed error (residuals)

Recap of yesterday

Fit a regression line by minimizing sum of squares

This is also the likelihood

But it requires some assumptions:

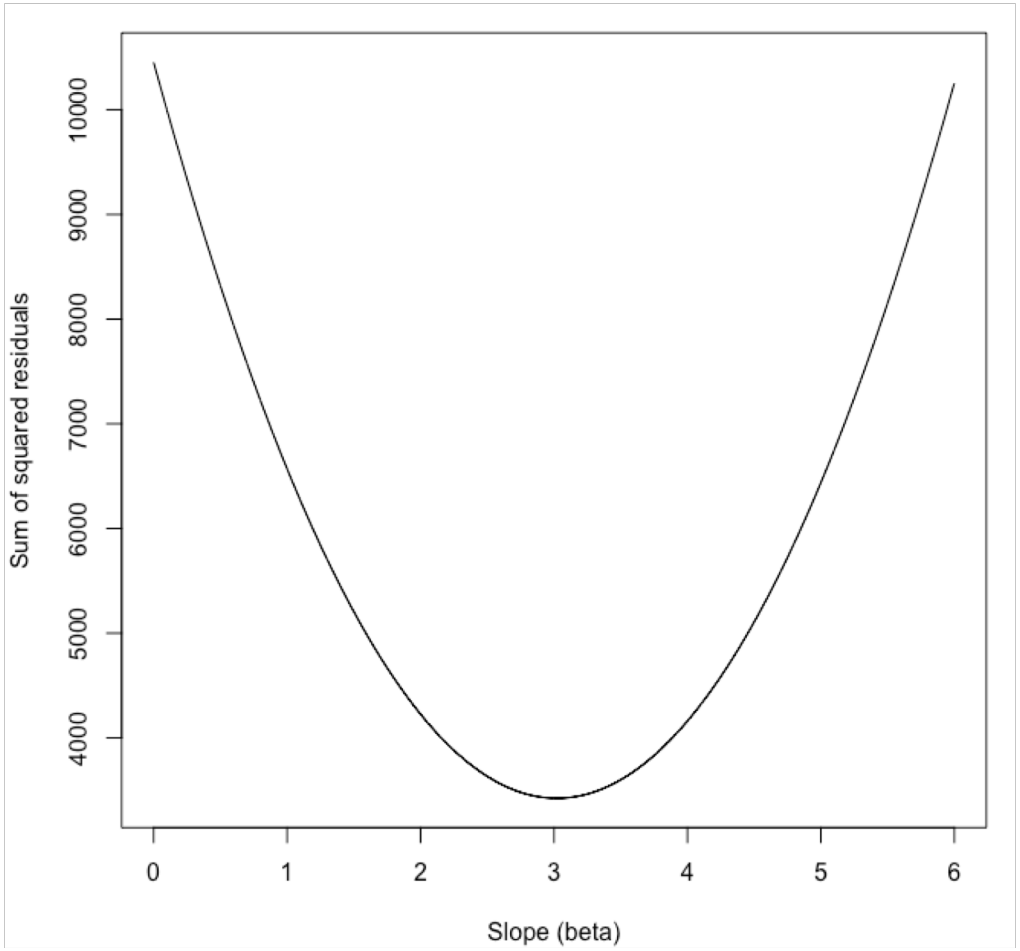
- **Normally distributed error (residuals)**
- **Error has equal variance along line**
- **Mean of the residuals = 0**
- **The relationship is linear**

Recap of yesterday

Approximated likelihood for β yesterday using sum of squared residuals

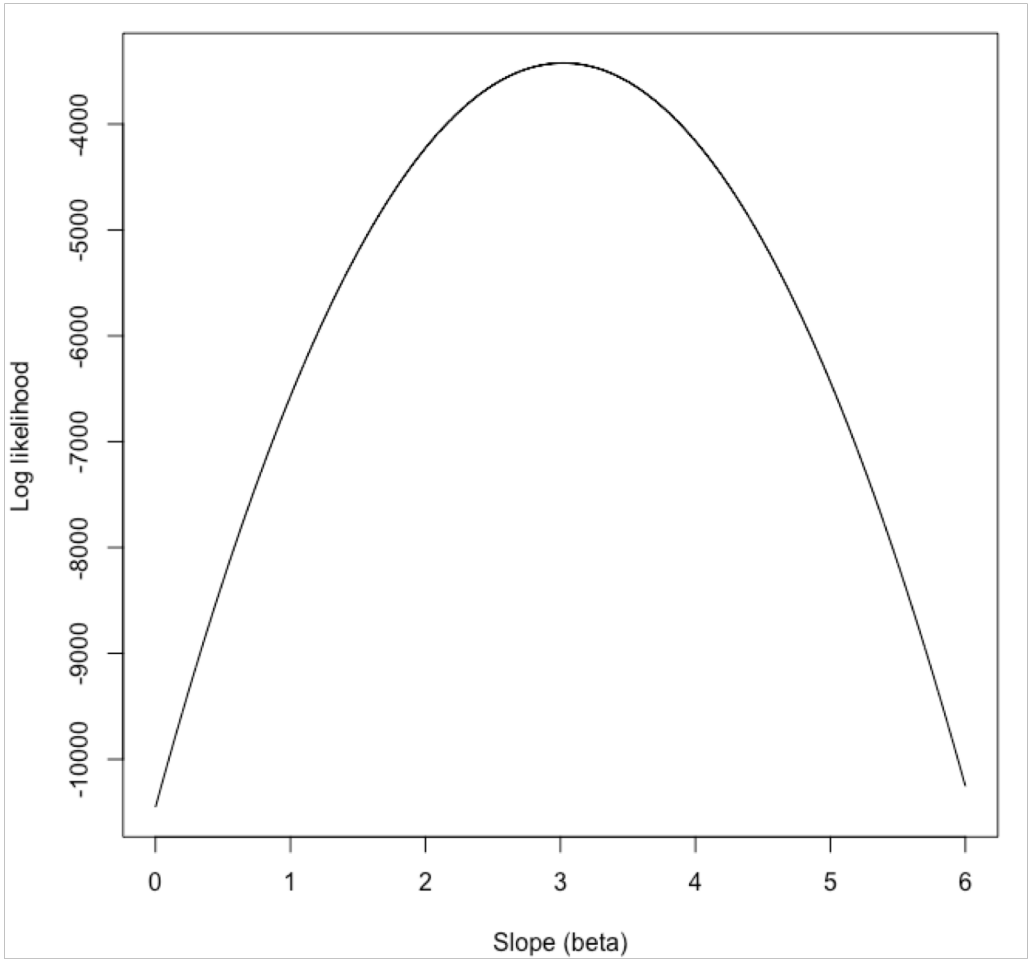
Recap of yesterday

Approximated likelihood for β yesterday using sum of squared residuals



Recap of yesterday

Approximated likelihood for β yesterday using sum of squared residuals



Maximum likelihood

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

Maximum likelihood

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

Three components:

α

β

Minimise sum of squared residuals

ε_i

Maximum likelihood

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

Three components:

α

β

Minimise sum of squared residuals

ε_i → $N(0, \sigma^2)$ Variance of residuals

Maximum likelihood

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

Three parameters that need to be estimated:

$$\alpha \qquad \beta \qquad \sigma^2$$

Variance is important too, even if we don't always interpret it

Back to fitting in R

Reminder! Fitting a linear regression in R

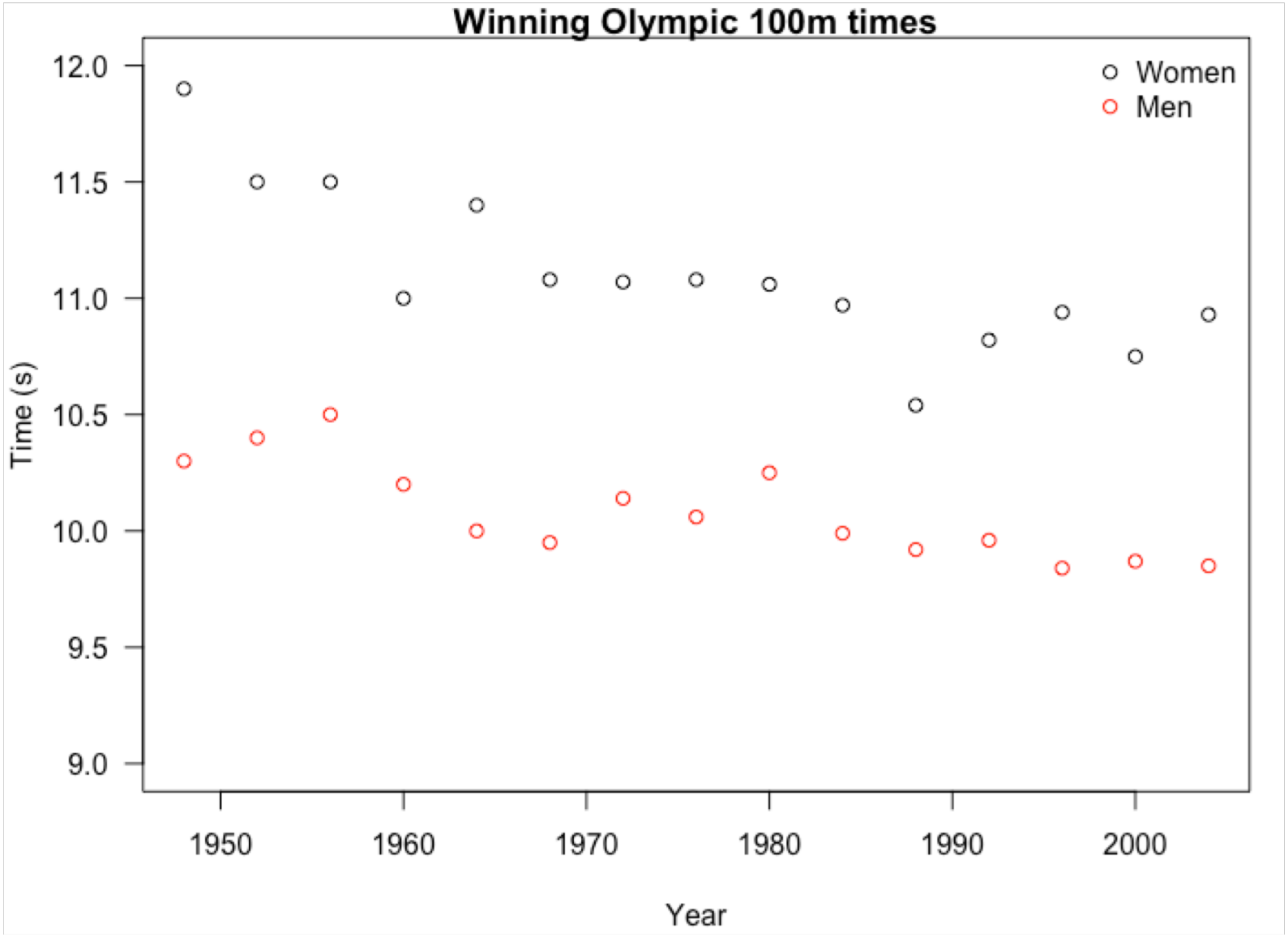
Arguments of lm():

lm(formula, data)

formula = Y ~ X
data = your data

Y is the response variable
X is the explanatory variable

Data from week 1



Exercise 1: Fit regression to 100m times

- Data from week 1 – now will fit a linear regression
- Some groups will run a regression on the women's times, the others will do one on the men's times (ONLY DO ONE)
- Data can be found at: https://www.math.ntnu.no/emner/ST2304/2019v/Week5/Times.csv
- It is a .csv file with headers (import using link above)
- **Fit the regression using lm(), make sure to assign the output as an object. Then look at the results.**

Hints and tips:

You will need to look at the column names to find the X and Y variables – use **head(), str(), or colnames()**

Remember to use **?FunctionName** for help on arguments

Exercise 1: Cntd

- Plot your data and add the regression line (code on next slide)
- Make sure to label your axes (arguments = xlab and ylab)
- Also choose the colour of your line

Plot line

We have estimates of the intercept and the slope of our line

So we can now plot the line regression line using:

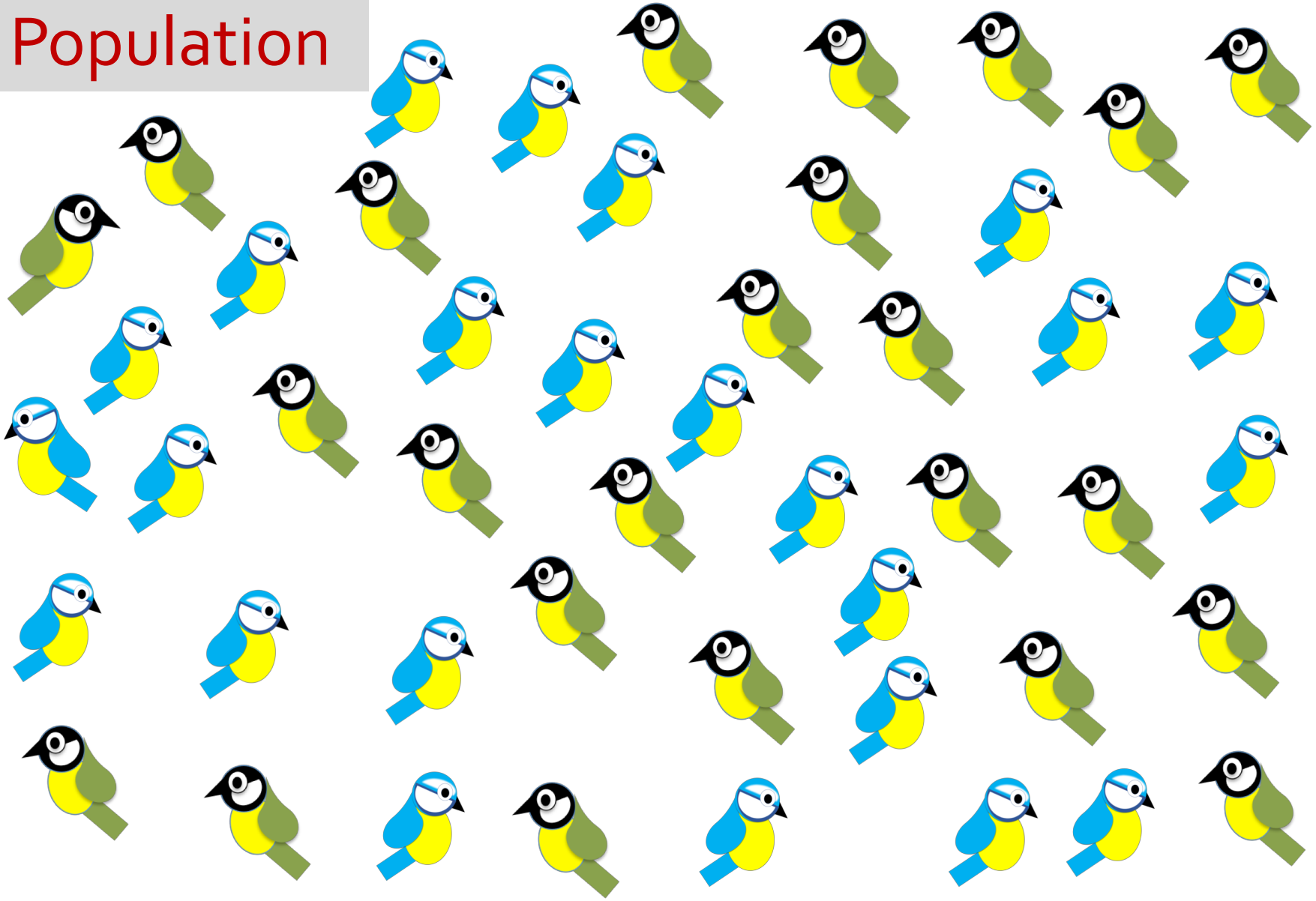
```
plot(X, Y, data = YourData)

abline(a = YourIntercept, b = YourSlope,
col = 2)
```

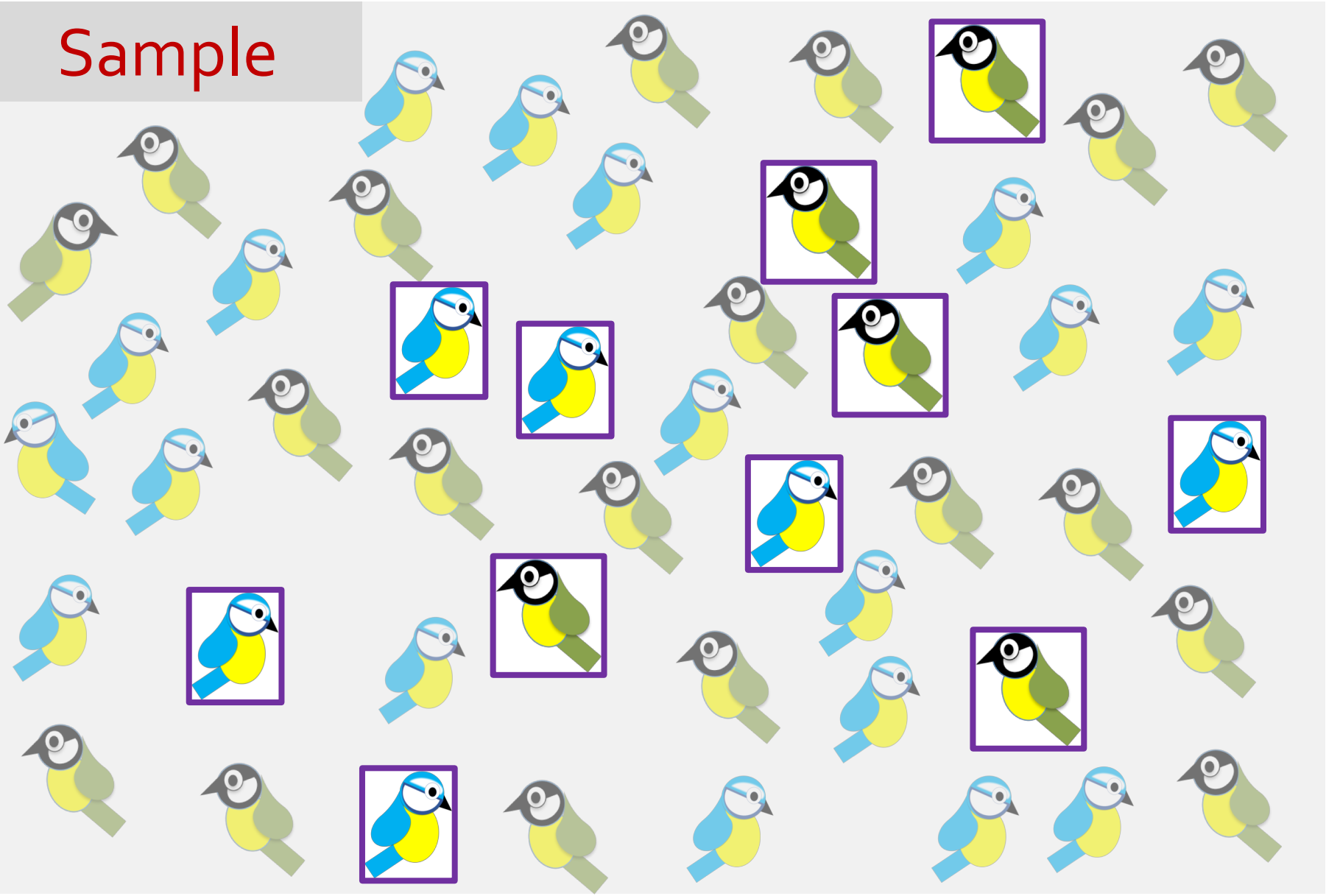
Adding uncertainty/ confidence

Why do we need to consider uncertainty?

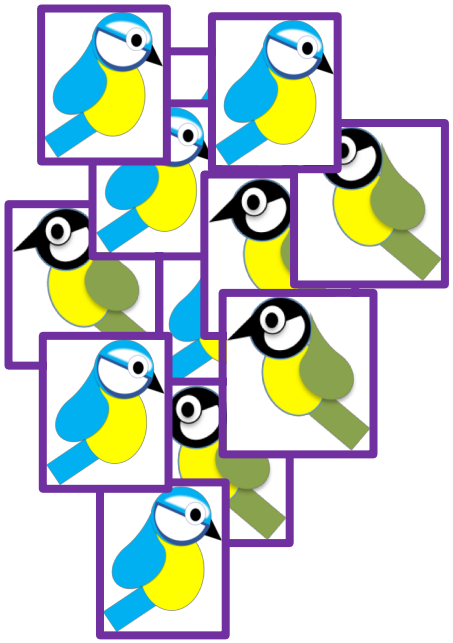
Population



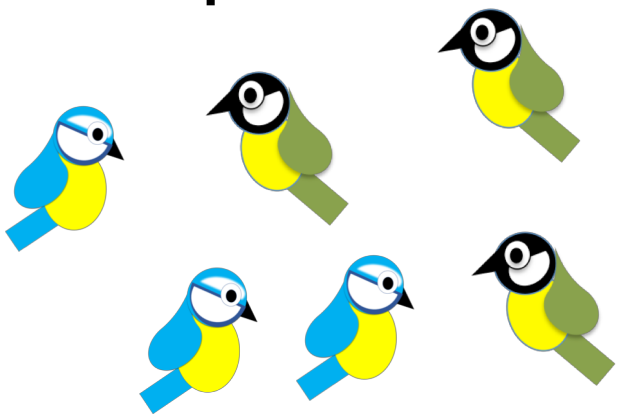
Sample

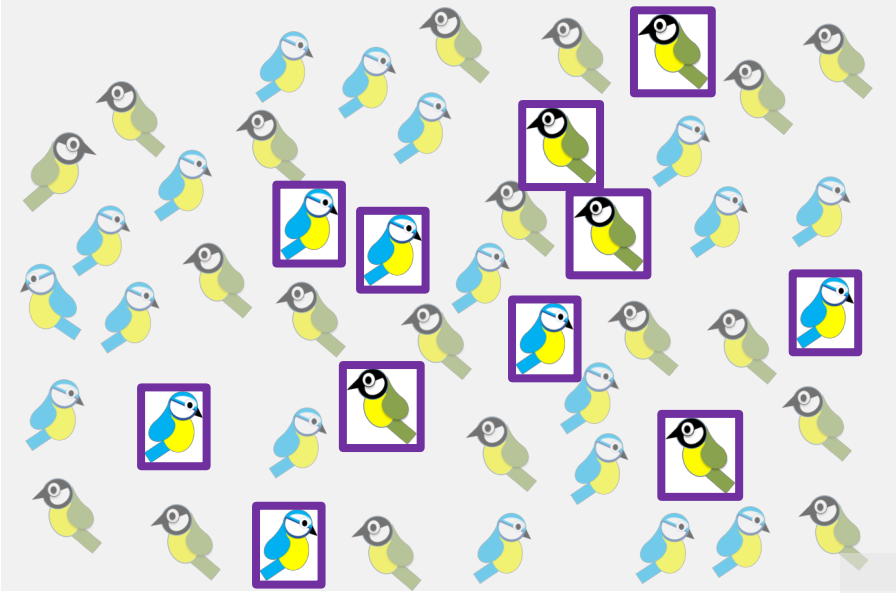


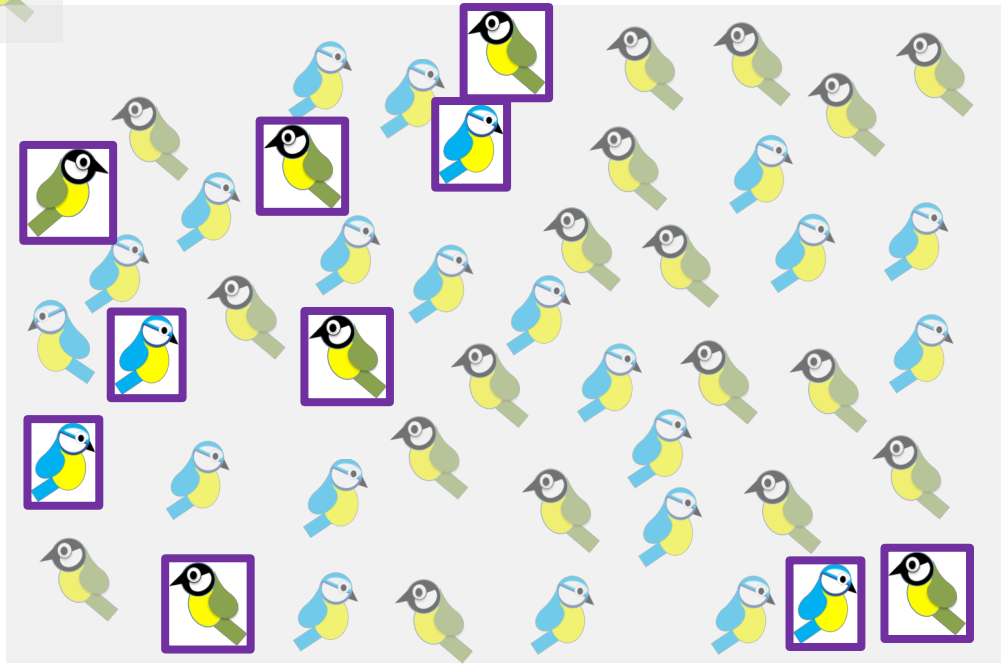
Sample



Population







Why do we need to consider uncertainty?

We want to take account of variability that would occur if we repeated our sampling

Let's us make general statements about the parameters at population level

How do we do this?

Why do we need to consider uncertainty?

We want to take account of variability that would occur if we repeated our sampling

Let's us make general statements about the parameters at population level

How do we do this? **Confidence intervals and standard errors**

P72-77 in The New Statistics with R

Confidence intervals

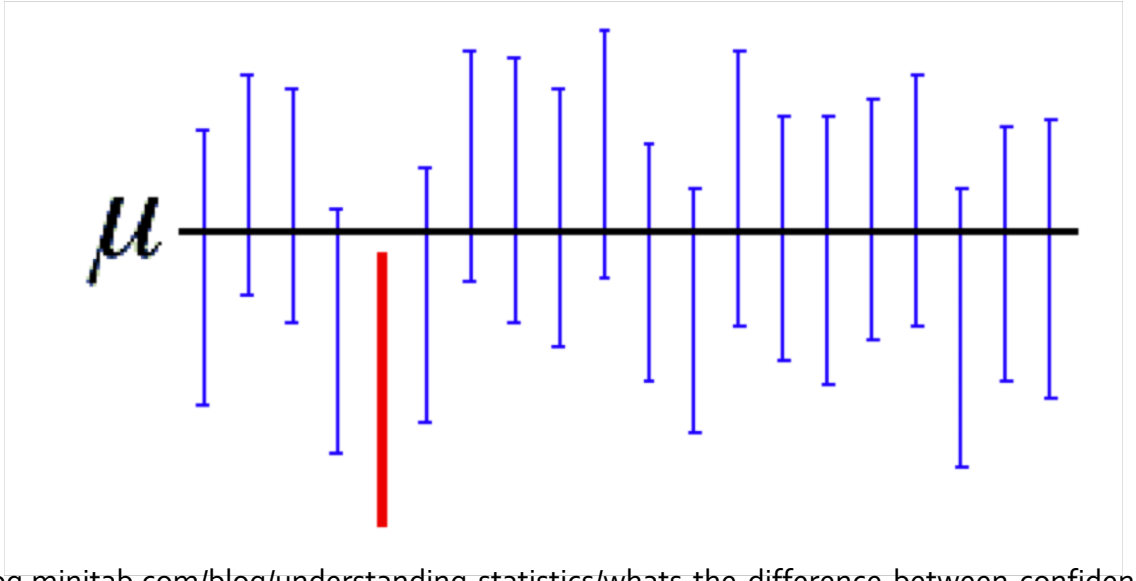
Represent what would happen if we repeat our sample many times

If we do this, 95% of the 95% confidence intervals drawn would contain the true population value

Confidence intervals

Represent what would happen if we repeat our sample many times

If we do this, 95% of the 95% confidence intervals drawn would contain the true population value



http://blog.minitab.com/blog/understanding-statistics/whats-the-difference-between-confidence-prediction-and-tolerance-intervals

Exercise 2: Adding confidence

- It is very easy to calculate the confidence intervals from an lm()

Use function:

```
confint(YourModelObject)
```

- Calculate the confidence intervals for your model
- Based on what you know already about the coefficients, work out what the confidence intervals are telling you

Exercise 2: Plotting confidence

- You know you can plot your regression line using abline()
- To plot your confidence intervals you need these two bits of code:

```
# generate some predictions based on the confidence
# interval
predictions <- predict(YourModelObject, interval =
“confidence”)

# plot the lower bound, column 2 in predictions
lines(YourXValues, predictions[,2], lty=2)

# plot the upper bound, column 3 in predictions
lines(YourXValues, predictions[,3], lty=2)
```

- lty=2 gives a dashed line

confint()

```
> confint(RegressionModel)
                    2.5 %       97.5 %
(Intercept) 45.40271555 66.39426309
Year        -0.02855252 -0.01792446
```

α intercept

β slope

Lower bound

Upper bound

Interpretation of results

Exercise 3: Interpret your results.

- You now have estimates of α and β and the associated confidence intervals for your regression model
- But what do these mean in terms of the relationship between 100m time and year?
- Pick out the key results from your model
- Prepare to present one key result per group

Exercise 4: Prediction

- Regression can also be used for prediction
- We can predict within our dataset and outside
- Why might we want to predict values of Y?
- In what ways could prediction be good or bad? (write some down on your whiteboard)

Uncertainty in prediction

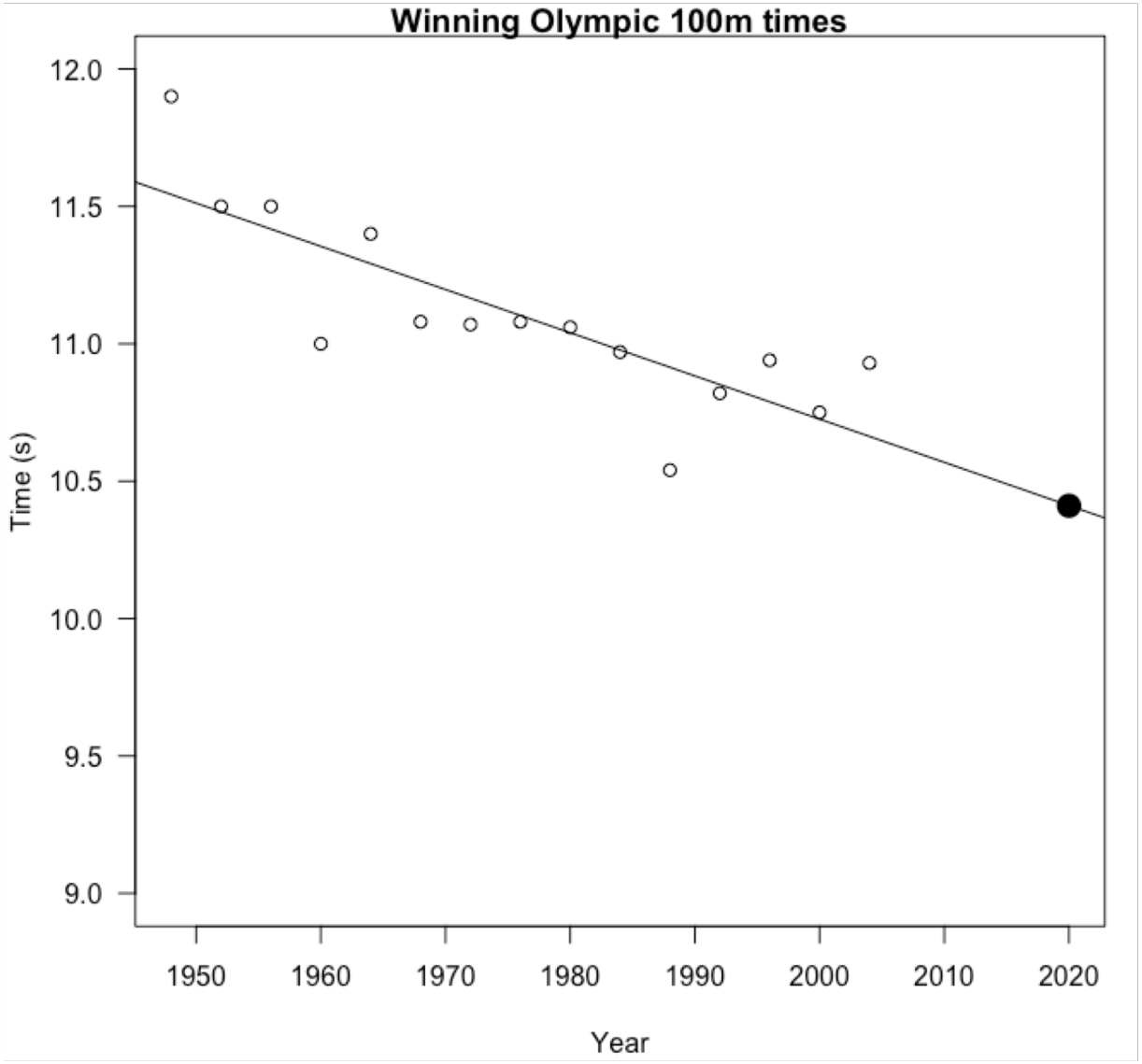
- When we use a regression for prediction it has uncertainty
- Uncertainty in relationship is captured by confidence intervals
- Prediction uncertainty captured by **prediction intervals**

Uncertainty in prediction

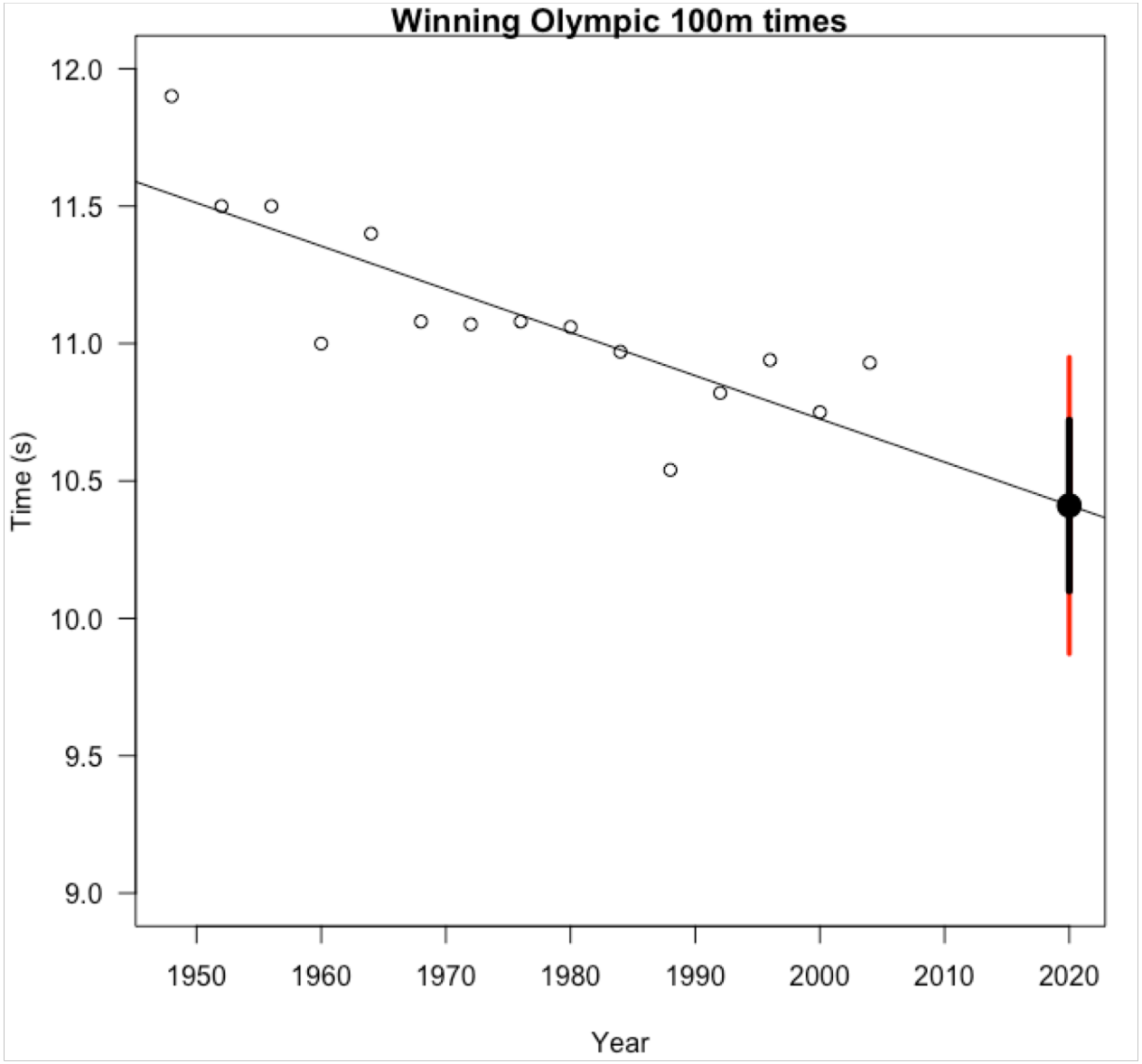
- For prediction we need to convey uncertainty in estimated relationship **and scatter around the line (error)**

- So prediction error includes the variance of the residuals too (σ^2 is finally useful!)

Uncertainty in prediction



Uncertainty in prediction



95% prediction interval for women in 2020 is between 9.87 and 10.94 seconds

Exercise 5: Further directions

- Think about the results we have had today for men's and women's times for 100m
- How well do you think your line captures the data?
- How could you work out how good the fit is? What would you want to look at?
- What did you find out about the times? Is there anything more might you want to know? (E.g. does temperature impact running speed?)

Summary of today's results

- Both men's and women's 100m winning Olympic times are decreasing over time
- Women by 0.016 seconds/year
- Men by 0.01 seconds/year
- We are unlikely to have seen the results if there was no trend (0 not in CIs)
- **Other questions:** How will times change in the future? Does this pattern happen outside of the Olympics? Are all humans getting faster? Is speed increase influenced by temperature?

Lecture Summary

A bit more on fitting

Adding uncertainty

Interpretation of results

How do the results fit in the scientific process?

Lecture Summary

A bit more on fitting

3 parameters estimated for maximum likelihood

Adding uncertainty

Interpretation of results

How do the results fit in the scientific process?

Lecture Summary

A bit more on fitting

3 parameters estimated for maximum likelihood

Adding uncertainty

We add uncertainty to represent taking a sample many times

Interpretation of results

How do the results fit in the scientific process?

Lecture Summary

A bit more on fitting

3 parameters estimated for maximum likelihood

Adding uncertainty

We add uncertainty to represent taking a sample many times

Interpretation of results

We can translate α β into change in Y with X (back into biological units) – make conclusion about relationship

How do the results fit in the scientific process?

Lecture Summary

A bit more on fitting

3 parameters estimated for maximum likelihood

Adding uncertainty

We add uncertainty to represent taking a sample many times

Interpretation of results

We can translate α β into change in Y with X (back into biological units) – make conclusion about relationship

How do the results fit in the scientific process?

We can use our results to tell us about the population.

We can ask new questions inspired by results of our models