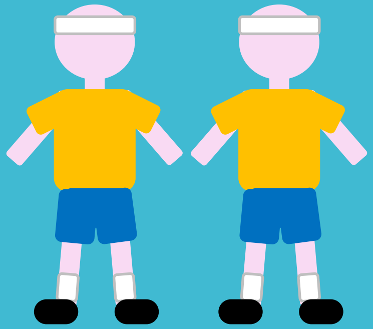
Linear regression: Part 2



Recap

What are linear models and linear regression?

How do we fit these models?

Using lm() in R

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

Prediction

- EX4: Prediction
- EX5: Discuss further steps/good models

A bit more on fitting

What the likelihood looks like

This is the log-likelihood for a linear regression:

$$l(y|x,\alpha,\beta,\sigma^2) = -\frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{(y_i-(\alpha+\beta x_i))^2}{2\sigma^2}$$

What the likelihood looks like

This is the log-likelihood for a linear regression:

$$l(y|x,\alpha,\beta,\sigma^2) = -\frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{(y_i - (\alpha + \beta x_i))^2}{2\sigma^2}$$

Our parameters

What the likelihood looks like

This is the log-likelihood for a linear regression:

$$l(y|x,\alpha,\beta,\sigma^2) = -\frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{(y_i-(\alpha+\beta x_i))^2}{2\sigma^2}$$

Our parameters

The explanatory variable

What the likelihood looks like

This is the log-likelihood for a linear regression:

$$l(y|x,\alpha,\beta,\sigma^2) = -\frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{(y_i-(\alpha+\beta x_i))^2}{2\sigma^2}$$

Our parameters

The explanatory variable

The response variable (our observed data)

What the likelihood looks like

This is the log-likelihood for a linear regression:

$$l(y|x,\alpha,\beta,\sigma^2) = -\frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{(y_i-(\alpha+\beta x_i))^2}{2\sigma^2}$$

Our parameters

The explanatory variable

The response variable (our observed data)

The sample size

What the likelihood looks like

This is the log-likelihood for a linear regression:

$$l(y|x, \alpha, \beta, \sigma^2) = -\frac{n}{2}\log \sigma^2 - \sum_{i=1}^{n} \frac{\left(y_i - (\alpha + \beta x_i)\right)^2}{2\sigma^2}$$

This is the log-likelihood for a normal distribution:

$$l(y|\mu, \sigma^2) = -\frac{n}{2}\log \sigma^2 - \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{2\sigma^2}$$

What the likelihood looks like

This is the log-likelihood for a linear regression:

$$l(y|x,\alpha,\beta,\sigma^2) = -\frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{\left(y_i - (\alpha + \beta x_i)\right)^2}{2\sigma^2}$$

This is the log-likelihood for a normal distribution:

$$l(y|\mu,\sigma^2) = -\frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{2\sigma^2}$$

Identical except:

$$\mu_i = (\alpha + \beta x_i)$$

What the likelihood looks like

This is the log-likelihood for a linear regression:

$$l(y|x, \alpha, \beta, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{\left(y_i - (\alpha + \beta x_i)\right)^2}{2\sigma^2}$$

This is the log-likelihood for a normal distribution:

$$l(y|\mu, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{2\sigma^2}$$

Identical except:

$\mu_i = (\alpha + \beta x_i)$ to get the mean for the normal distribution we use the linear equation

What the likelihood looks like

This is the log-likelihood for a linear regression:

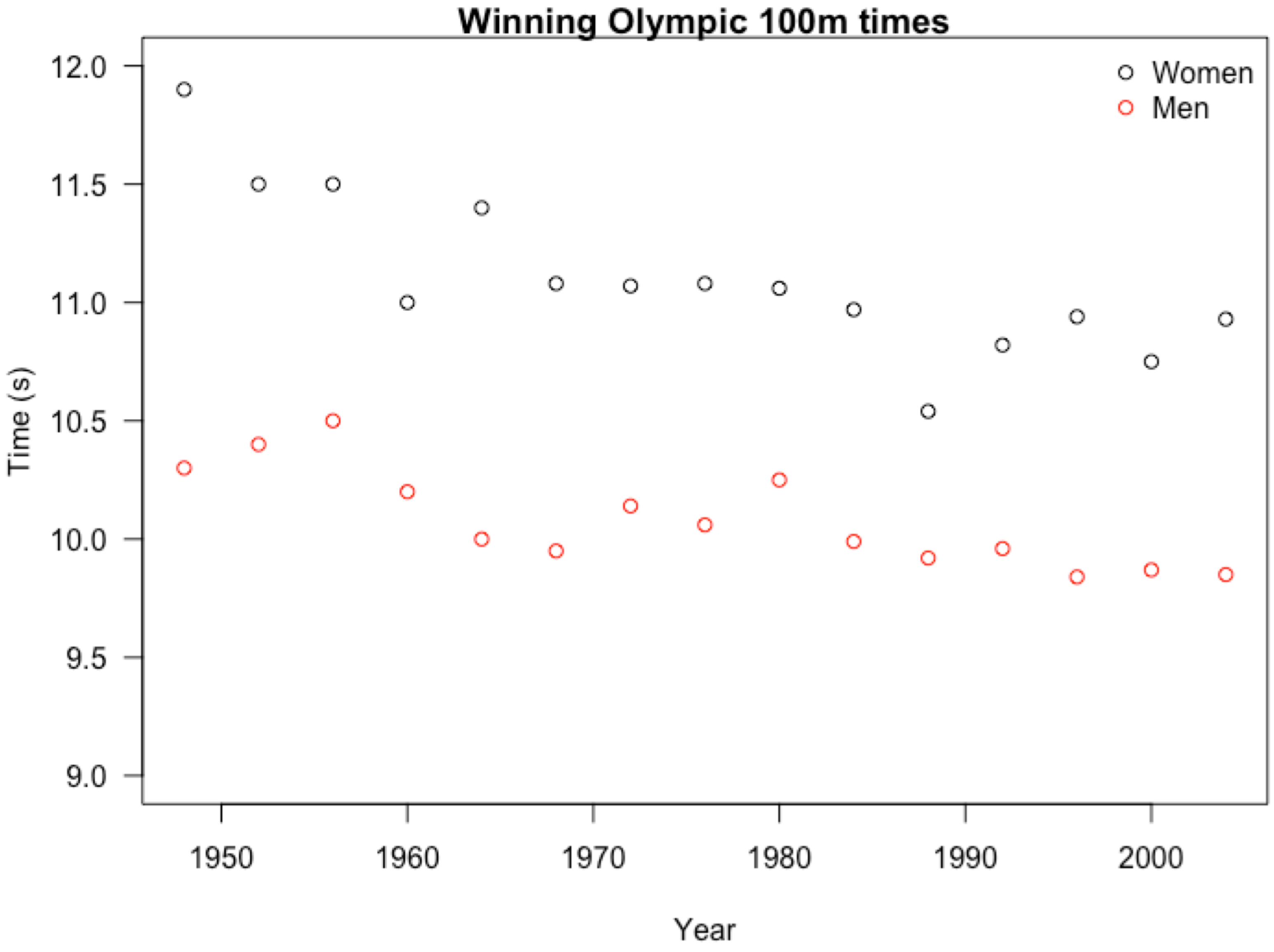
$$l(y|x,\alpha,\beta,\sigma^2) = -\frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{\boxed{\left(y_i - (\alpha + \beta x_i)\right)^2}}{2\sigma^2}$$

This is the log-likelihood for a normal distribution:

$$l(y|\mu,\sigma^2) = -\frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{2\sigma^2}$$

This part is the same as summing the squares (yesterday)

Data for today



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

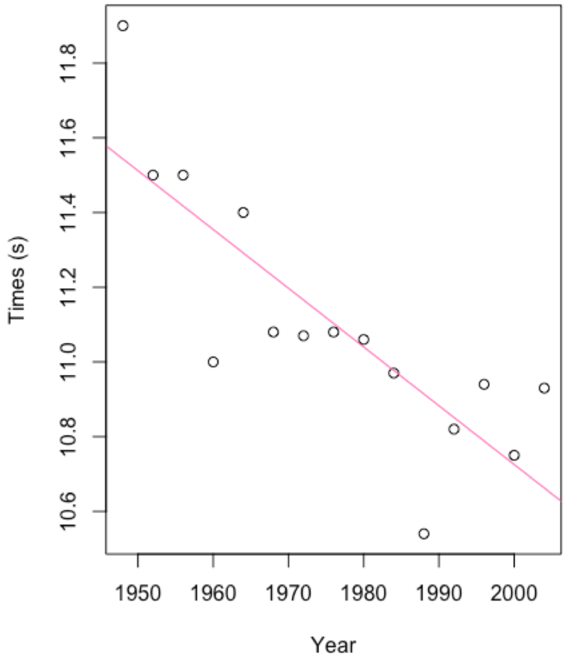
Exercise 1: Fit regression to 100m times

Part E of exercise module.

Some groups will run a regression on the women’s times, the others will do one on the men’s times (ONLY DO ONE)

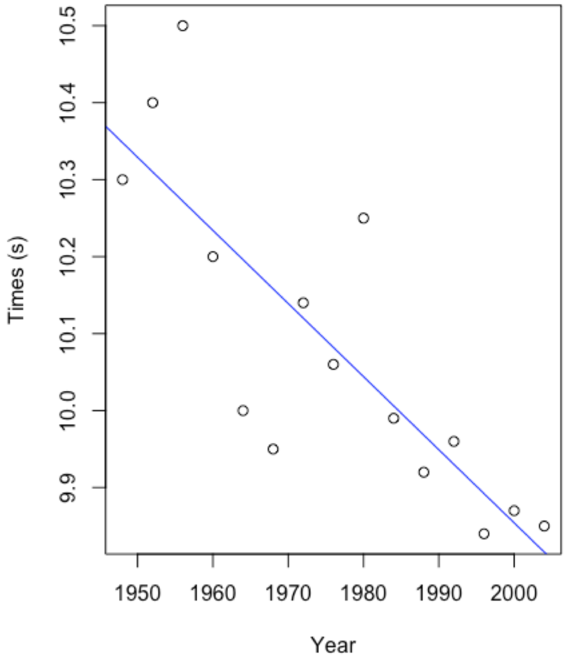
Summary Part E

Women



(Intercept) Year
42.19 -0.016

Men



(Intercept) Year
28.85 -0.0095

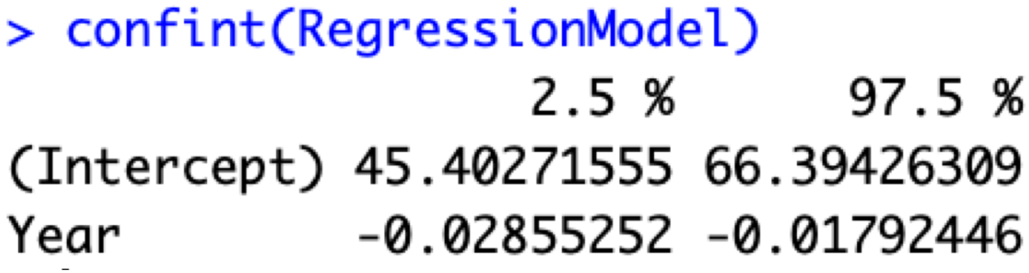
Adding uncertainty/ confidence

Exercise 2: Adding confidence

Part F

Some theory and practice

Summary Part F

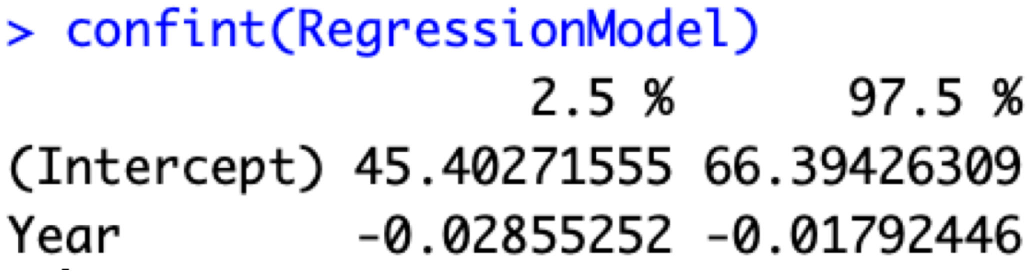


```
> confint(RegressionModel)
                  2.5 %      97.5 %
(Intercept) 45.40271555 66.39426309
Year        -0.02855252 -0.01792446
```

Lower bound

Upper bound

Summary Part F



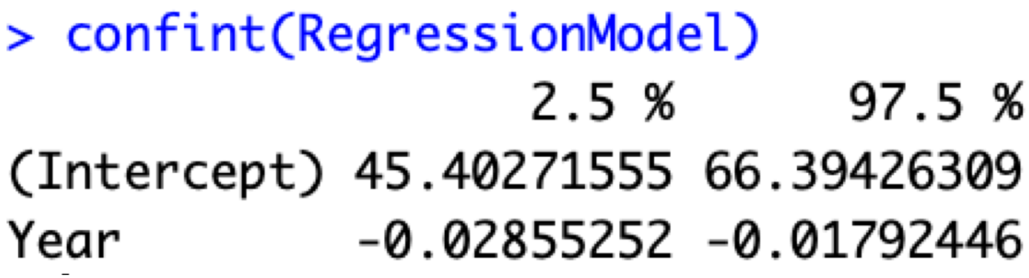
```
> confint(RegressionModel)
                  2.5 %      97.5 %
(Intercept) 45.40271555 66.39426309
Year        -0.02855252 -0.01792446
```

Lower bound

Upper bound

If you were to repeat this many many times, 95% of the time (on average) the confidence interval you draw would contain the true value.

Summary Part F



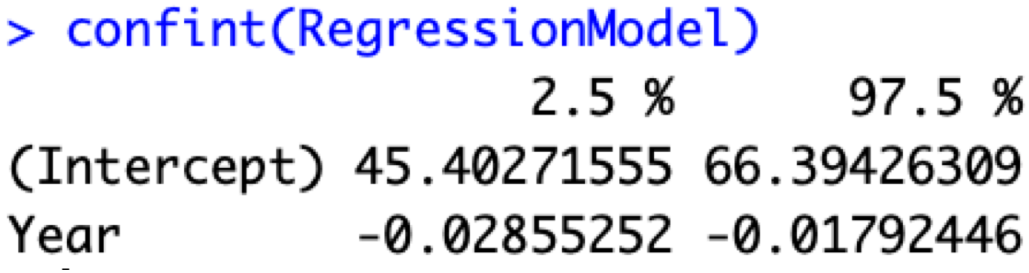
```
> confint(RegressionModel)
                  2.5 %      97.5 %
(Intercept) 45.40271555 66.39426309
Year        -0.02855252 -0.01792446
```

Lower bound

Upper bound

NOT: 95% probability that the true value is within the confidence interval

Summary Part F



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

Lower bound

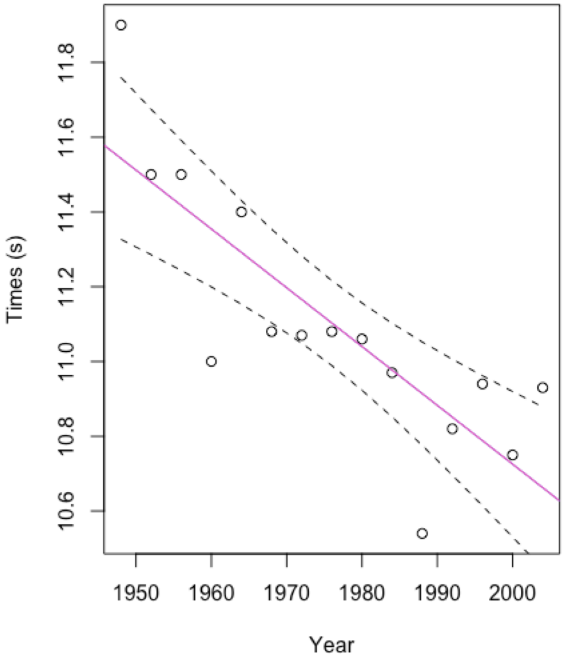
Upper bound

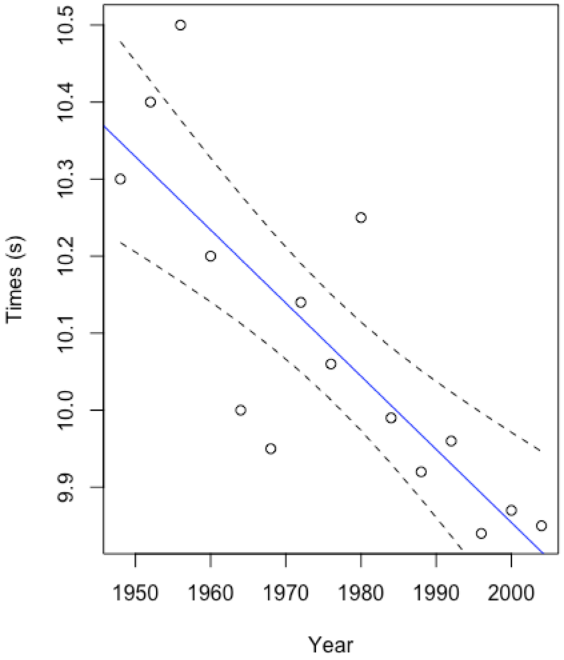
NOT: 95% probability that the true value is within the confidence interval

IS: the range of values that are more plausible to be the true value

IS: width says how uncertain we are (wider = less certain)

Summary Part F





Interpretation of results

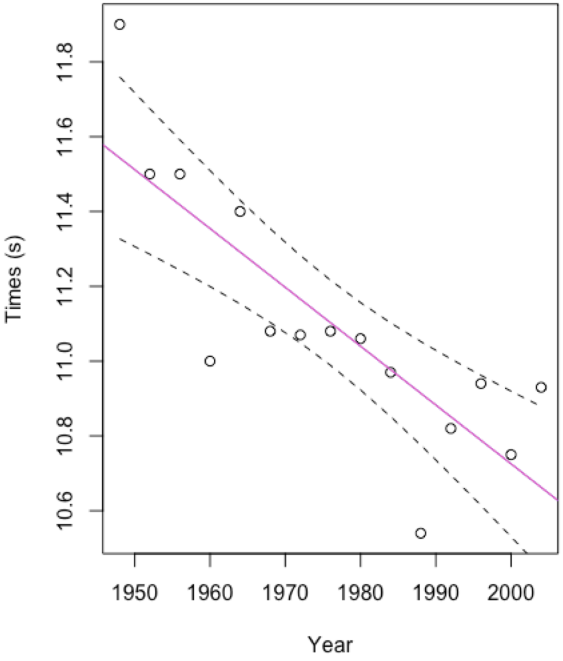
Exercise 3: Interpret your results.

Part G

Practice interpreting the results

Summary Part G

Which bit do we care about?



Maximum likelihood estimates:

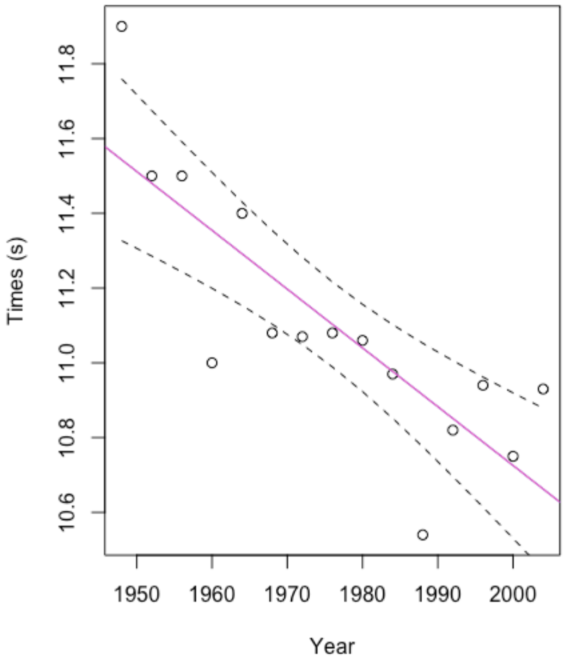
(Intercept)	Year
42.19	-0.016

Confidence intervals:

	2.5 %	97.5 %
(Intercept)	29.19	55.19
Year	-0.02	-0.009

Summary Part G

Which bit do we care about?



Maximum likelihood estimates:

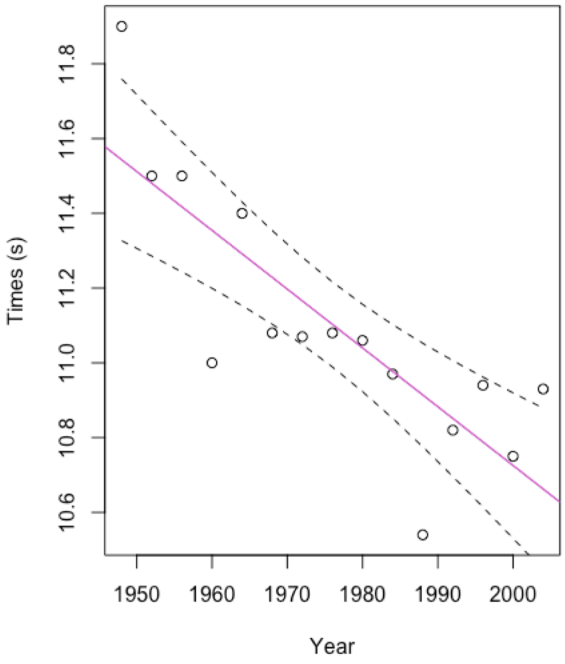
(Intercept)	Year
42.19	-0.016

Confidence intervals:

	2.5 %	97.5 %
(Intercept)	29.19	55.19
Year	-0.02	-0.009

Summary Part G

Which bit do we care about?



Maximum likelihood estimates:

(Intercept)	Year
42.19	-0.016

Confidence intervals:

	2.5 %	97.5 %
(Intercept)	29.19	55.19
Year	-0.02	-0.009

Exercise 3: Present results

5 minutes to update your results

Turn to same row on opposite side and tell them your result

Is it different for men and women?

Exercise 4: Prediction

Finish part G

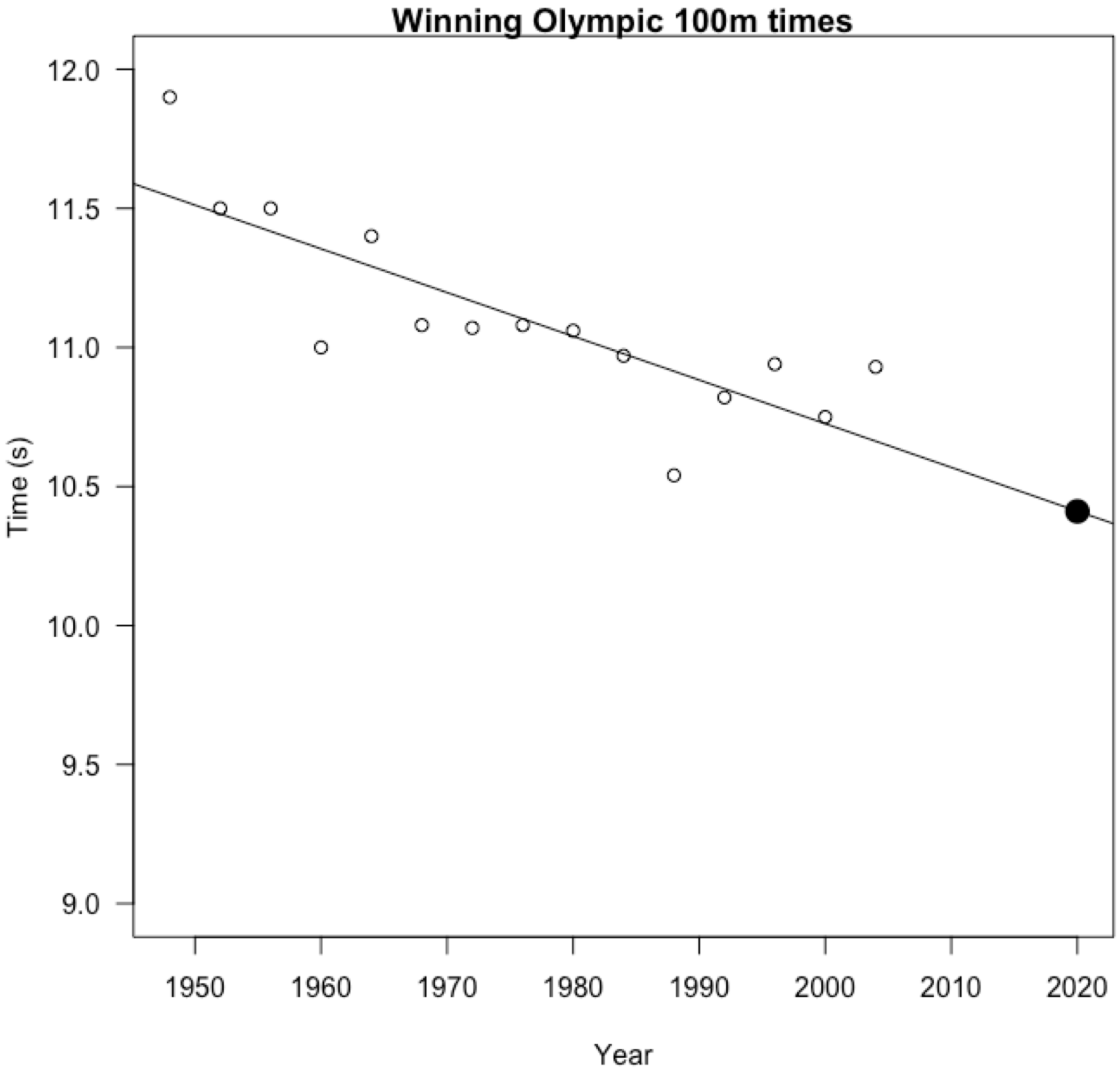
Summary Part G

Why predict?

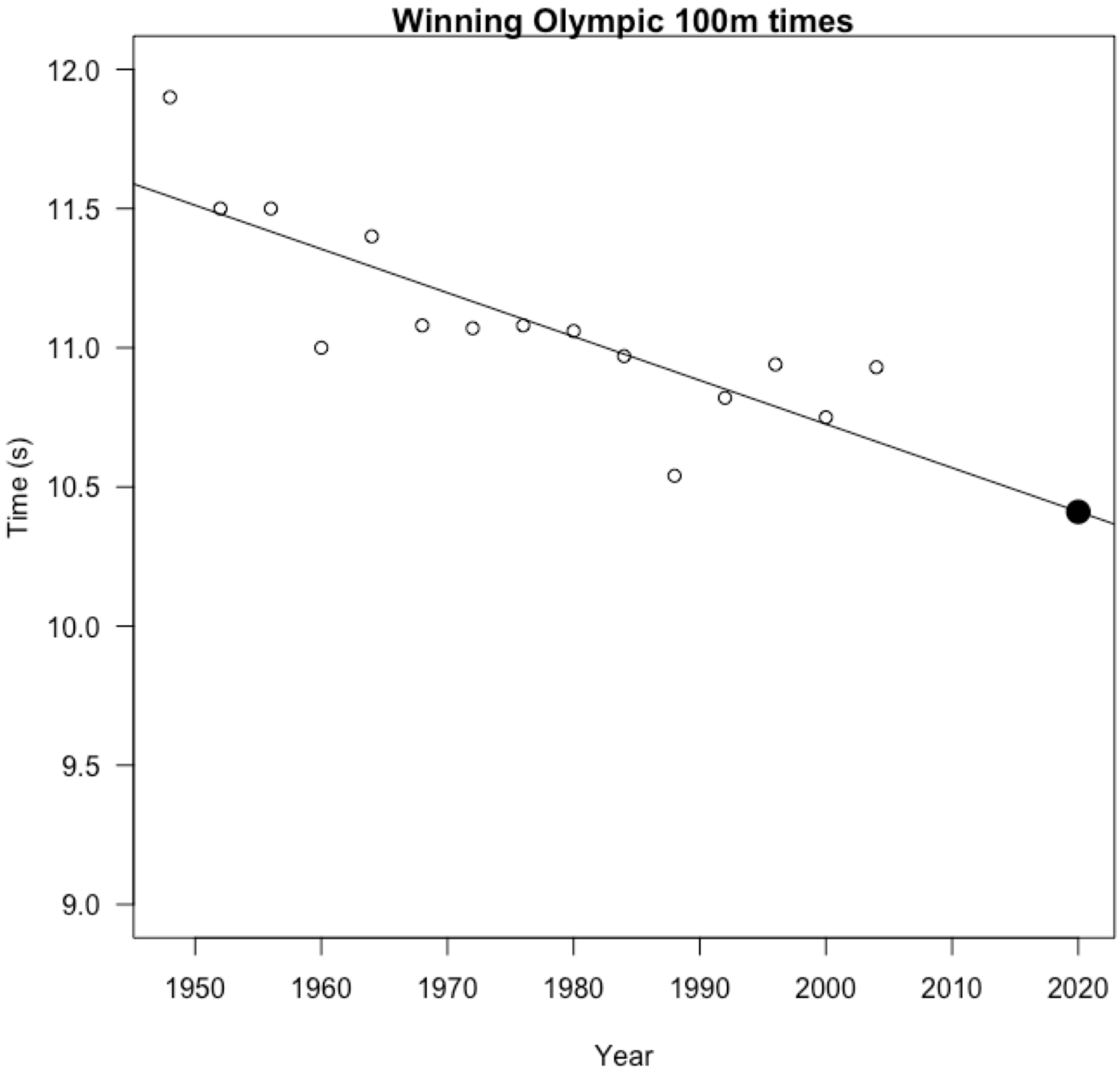
Fill in values within our data

Predict new values e.g. climate change

Uncertainty in prediction



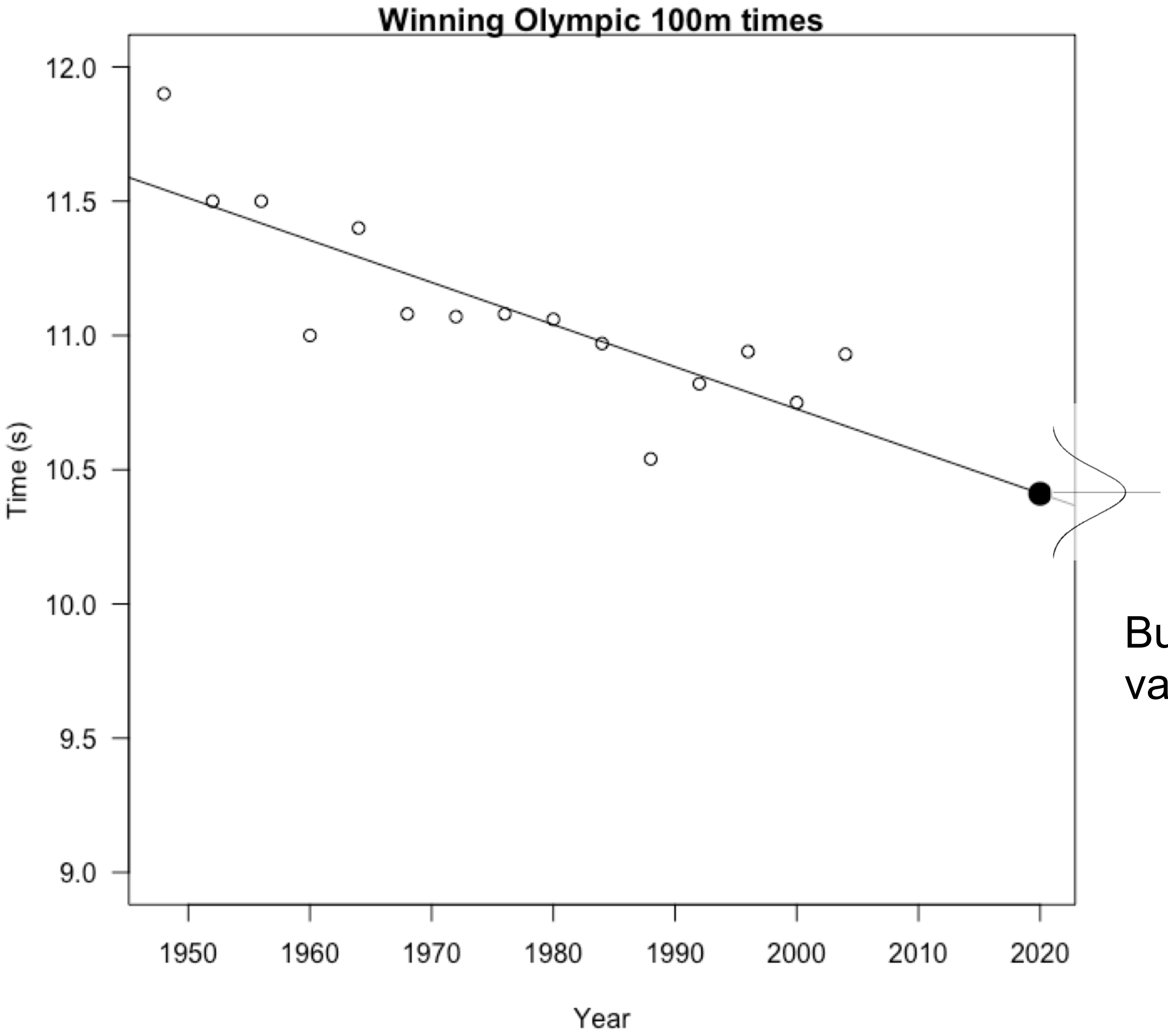
Uncertainty in prediction



X = 2020

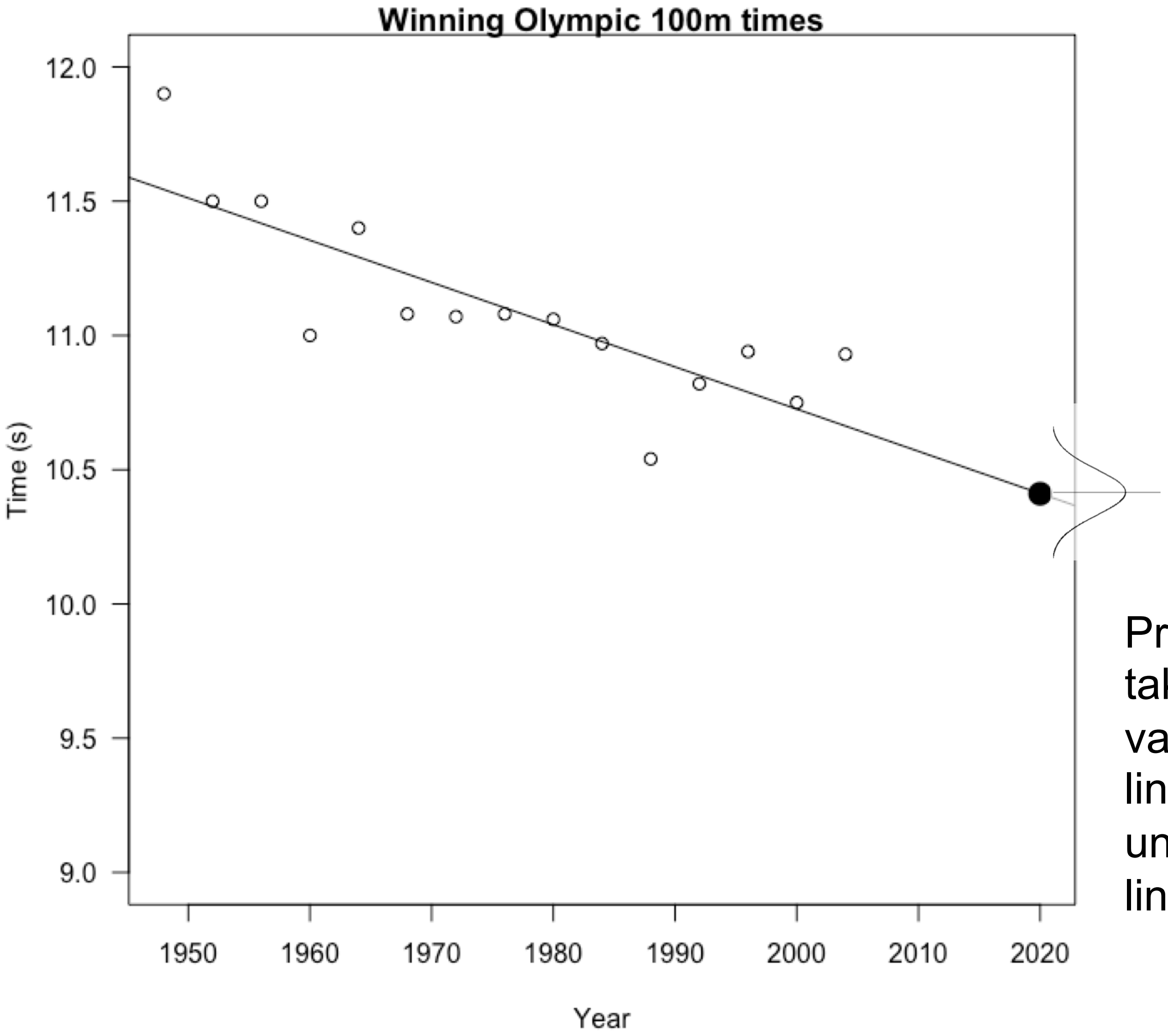
$\hat{Y}$= 10.41 seconds

Uncertainty in prediction



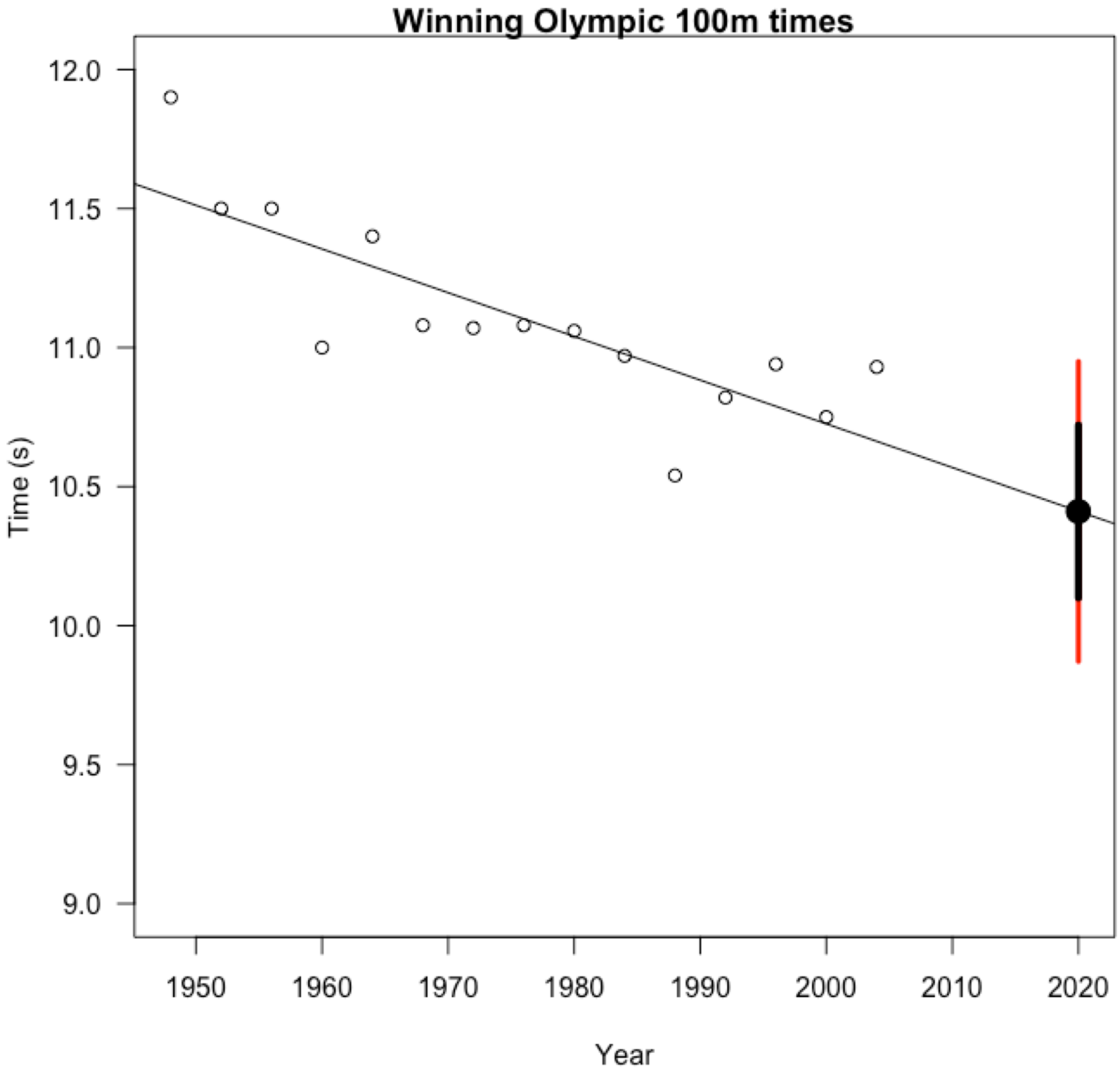
But what about variation???

Uncertainty in prediction

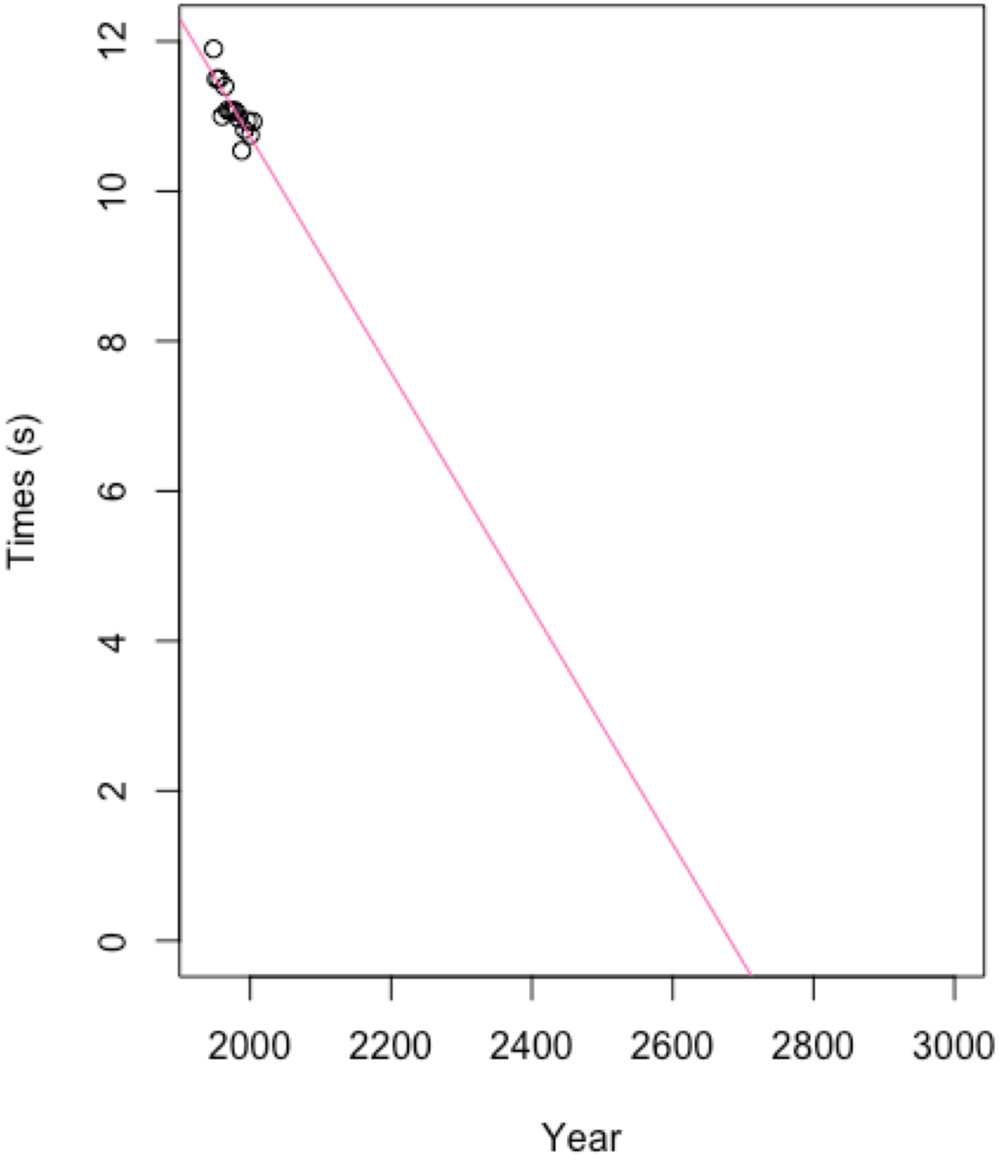


Prediction interval takes into account variation around the line as well as uncertainty in the line itself!

Uncertainty in prediction



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



Be careful with prediction

Exercise 5: Further directions

Part H

Exercise 5: Further directions

Feedback on further directions

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

Prediction

Lecture Summary

A bit more on fitting
Tried for a real example

Adding uncertainty

Interpretation of results

Prediction

Lecture Summary

A bit more on fitting

Tried for a real example

Adding uncertainty

We add uncertainty to represent taking a sample many times

Interpretation of results

Prediction

Lecture Summary

A bit more on fitting

Tried for a real example

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

Prediction

Lecture Summary

A bit more on fitting

Tried for a real example

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

Prediction

Can be useful but also need to be careful of going too far outside of your data

Give us feedback