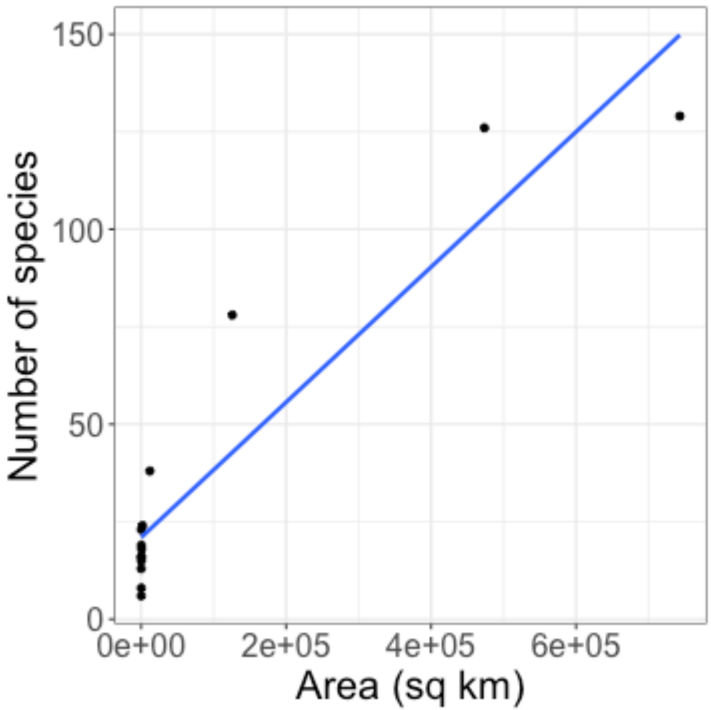
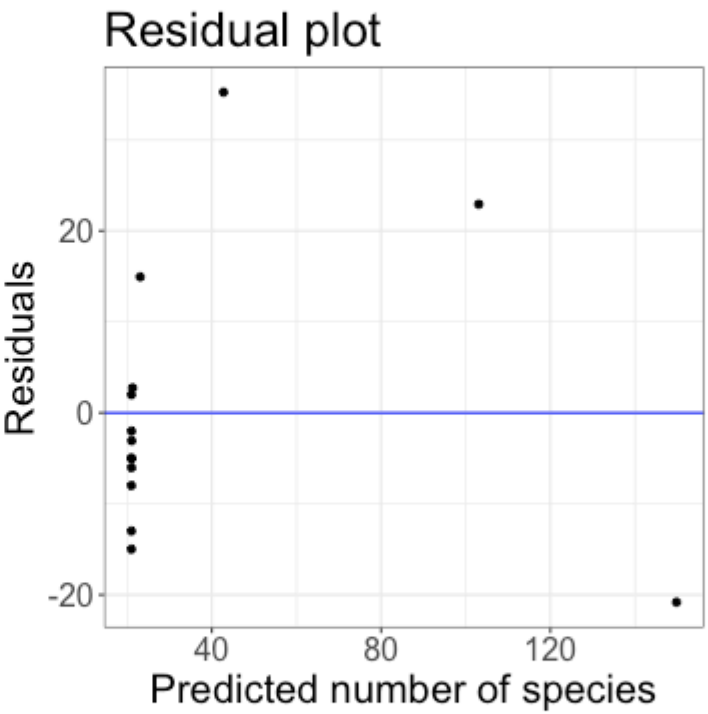
Nonlinear relationships and transformations

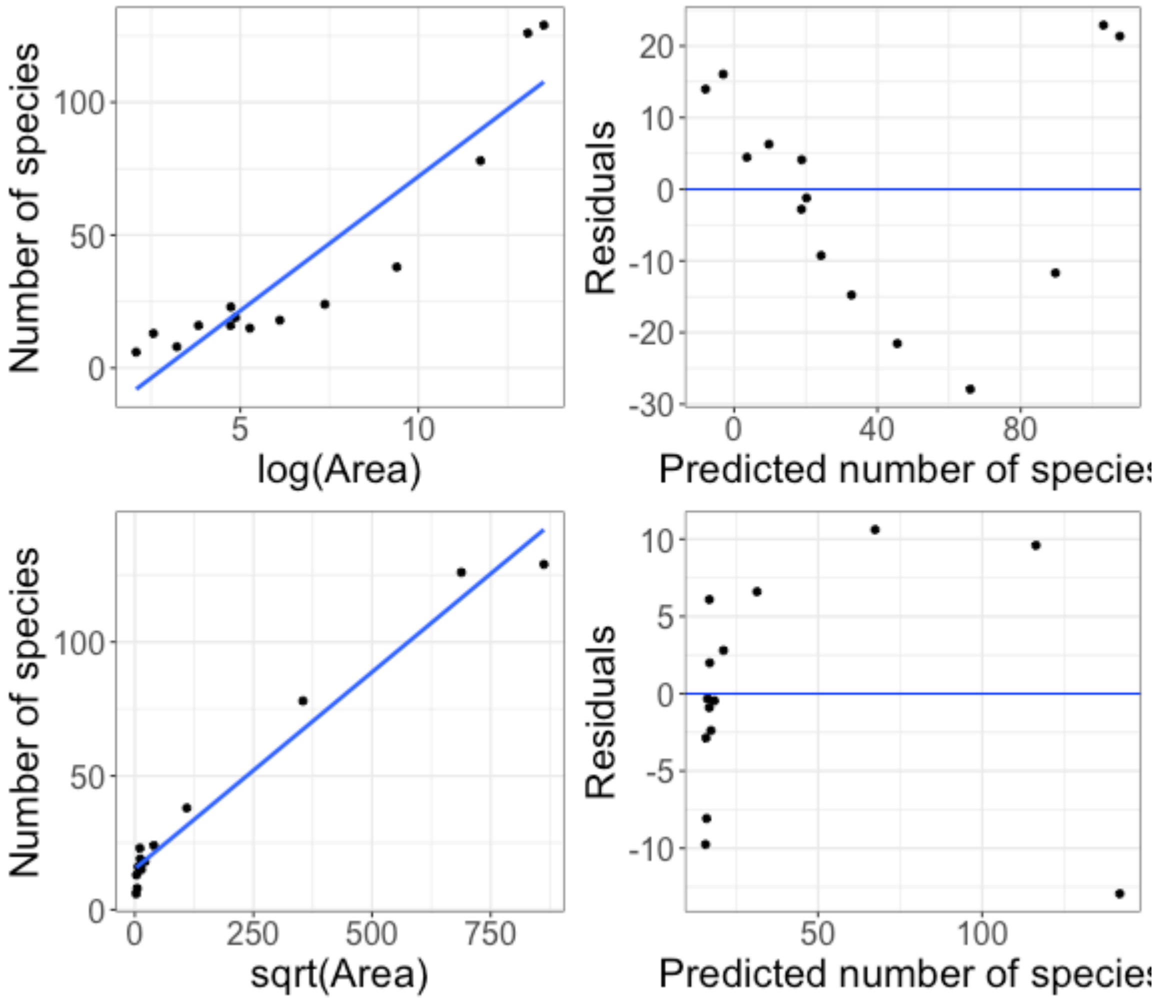
More transformations





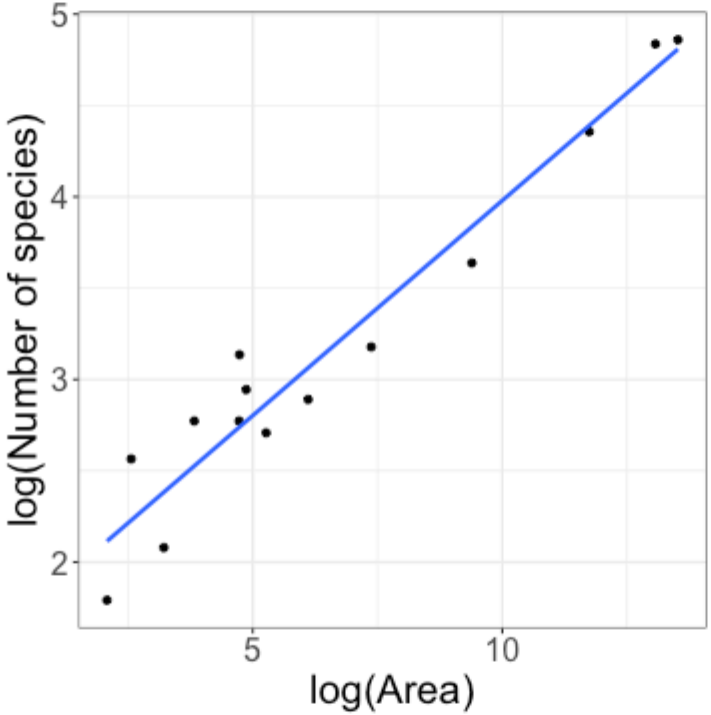
> What transformations could we try here?

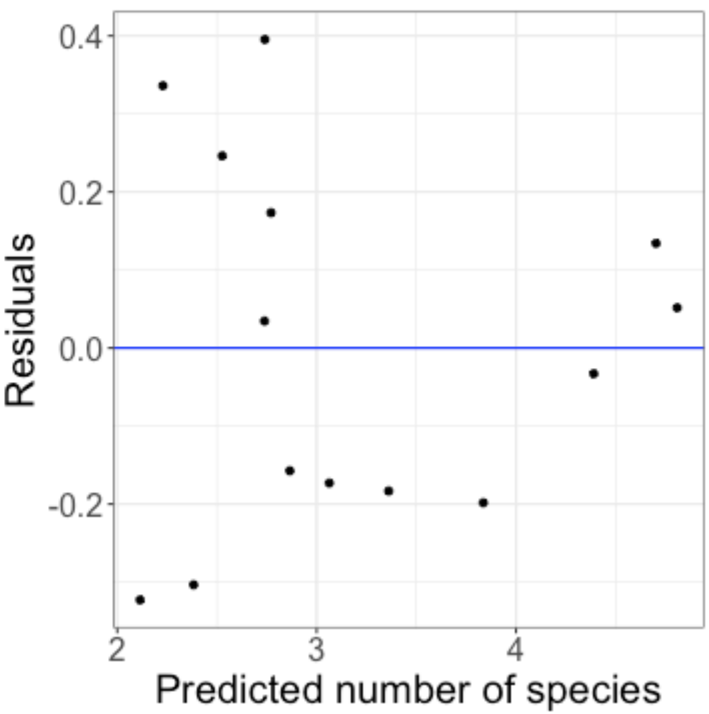
Trying some transformations



Transform both predictor and response?

$$\log(Species) = \beta_0 + \beta_1 \log(Area) + \varepsilon$$





$e^{a+b} = e^a e^b$ $\qquad$ $\log(b^a) = a\log(b)$

Interpreting the fitted model

$e^{\log(a)} = a$

$$\log(\widehat{Species}) = 1.625 + 0.246\log(Area)$$

$Area = x$

$\log(\widehat{species}) = 1.625 + 0.246\log(x)$

$\widehat{Species} = e^{1.625 + 0.246\log(x)}$

$= e^{1.625} e^{0.246\log(x)}$

$= e^{1.625} e^{\log(x^{0.246})}$

$= e^{1.625} (x)^{0.246}$

$Area = 1.01x$

(1% increase in Area)

$1.625 + 0.246\log(1.01x)$

$e^{1.625 + 0.246\log(1.01x)}$

$e^{1.625} e^{0.246\log(1.01x)}$

$e^{1.625} e^{\log((1.01x)^{0.246})}$

$e^{1.625} (1.01x)^{0.246}$

$$\frac{\widehat{Species} \text{ when } Area = 1.01x}{\widehat{Species} \text{ when } Area = x} = \frac{(1.01x)^{0.246}}{x^{0.246}} = (1.01)^{0.246}$$

estimated # species increases by a <u>factor</u> of $(1.01)^{0.246}$

Summary of model interpretation so far

$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

- A one-unit increase in x is associated with a change in y of $\widehat{\beta}_1$, on average

$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 log(x)$

- A 1% increase in x is associated with a change in y of $\widehat{\beta}_1 / 100$, on average

$\widehat{log(y)} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

- A one-unit increase in x is associated with a change in y by a **factor** of $e^{\widehat{\beta}_1}$, on average

$\widehat{log(y)} = \widehat{\beta}_0 + \widehat{\beta}_1 log(x)$

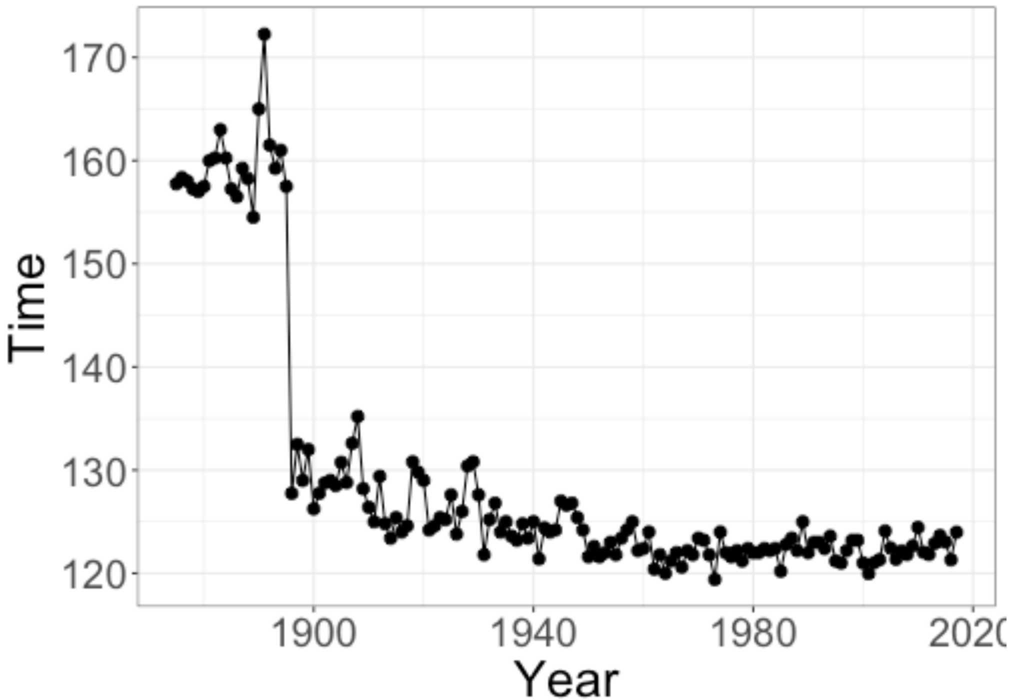
- A 1% increase in x is associated with a change in y by a **factor** of $(1.01)^{\widehat{\beta}_1}$, on average

Kentucky Derby data

Data on the Kentucky Derby winner from each year between 1875 and 2012. Variables include

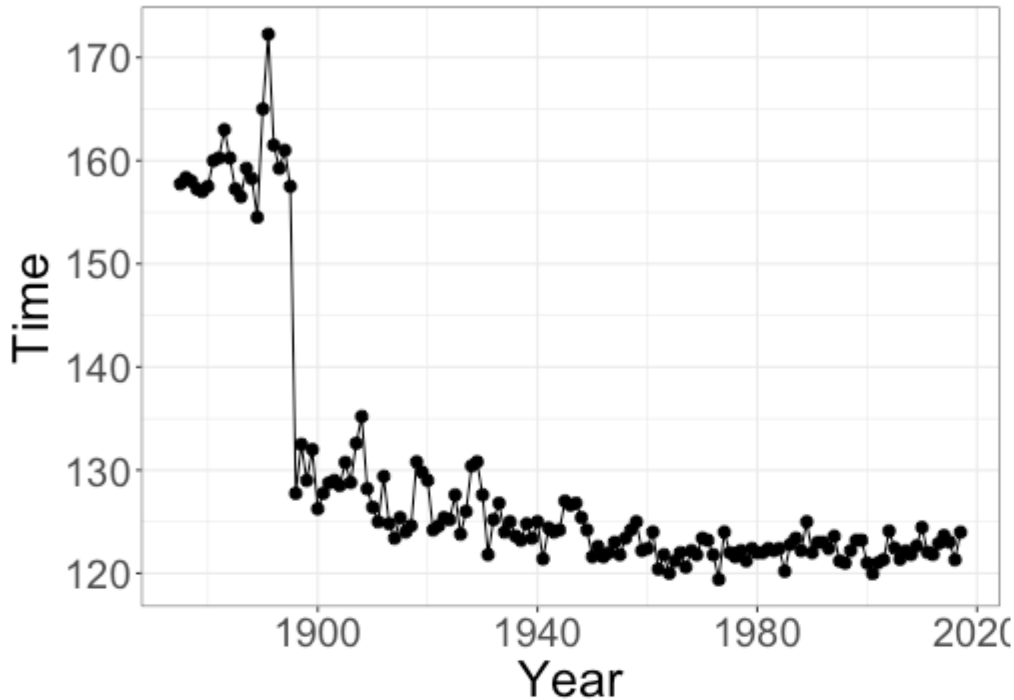
- `year`: year of race
- `speed`: speed of winning horse (mph)
- `time`: winning time (seconds)
- `condition`: condition of the track on the day of the race

EDA



What do you notice about the winning times?

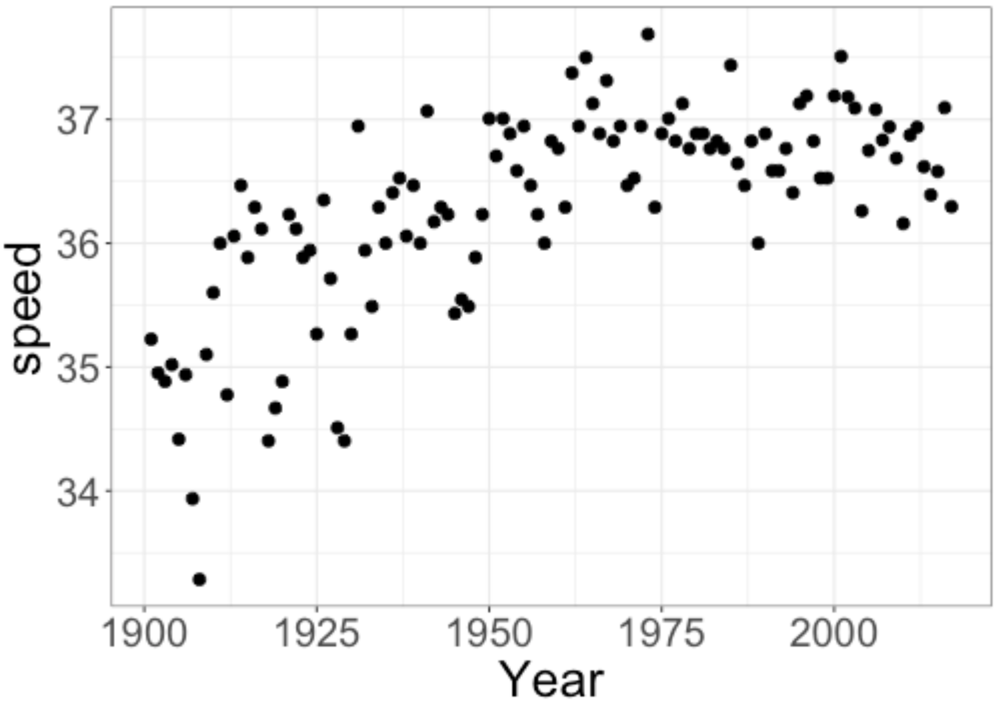
EDA



> What do you notice about the winning times?

There's a big drop around 1900, then they decrease gradually.

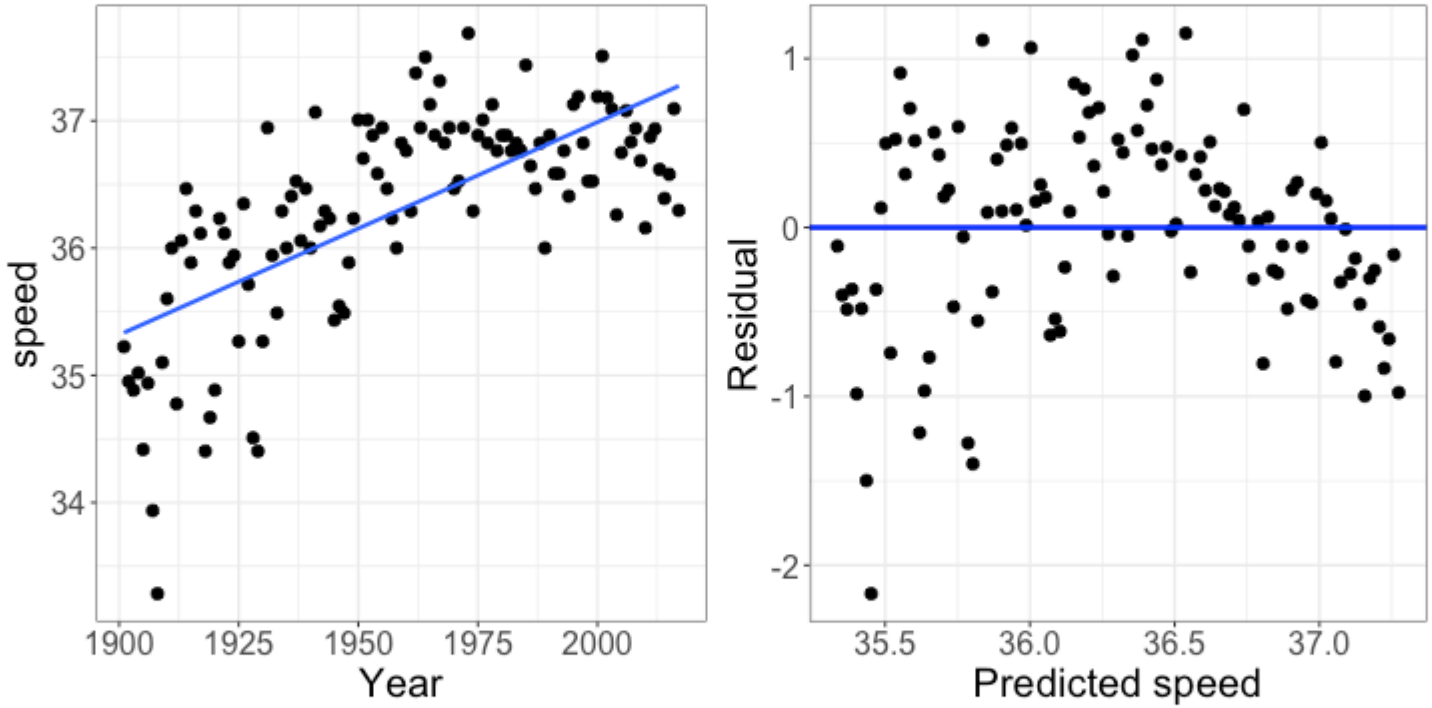
EDA



How would you describe the relationship?

Fitting a linear regression

$$\text{Speed} = \beta_0 + \beta_1 \ \text{Year} + \varepsilon$$



Polynomial transformations

Another possible transformation is a **polynomial**.

> **Definition:** A polynomial transformation of order k is a linear combination of the terms X, X^2, ..., X^k:
>
> $$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \varepsilon$$

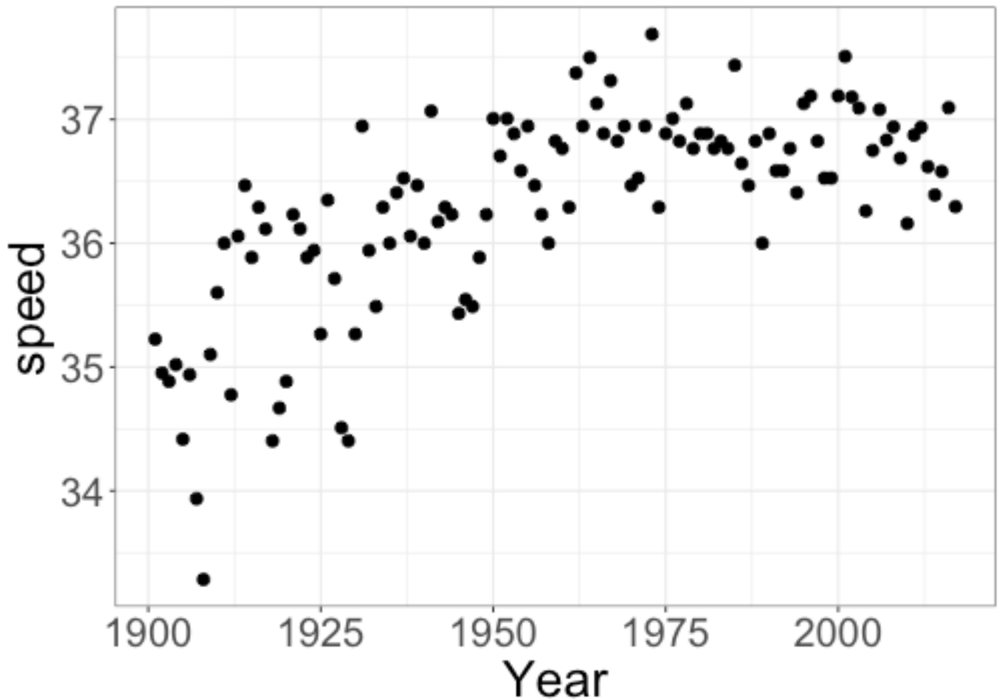
- Example: If we use a polynomial of order 3, our model is

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

2nd order polynomial:

3rd order polynomial:

Polynomial transformations

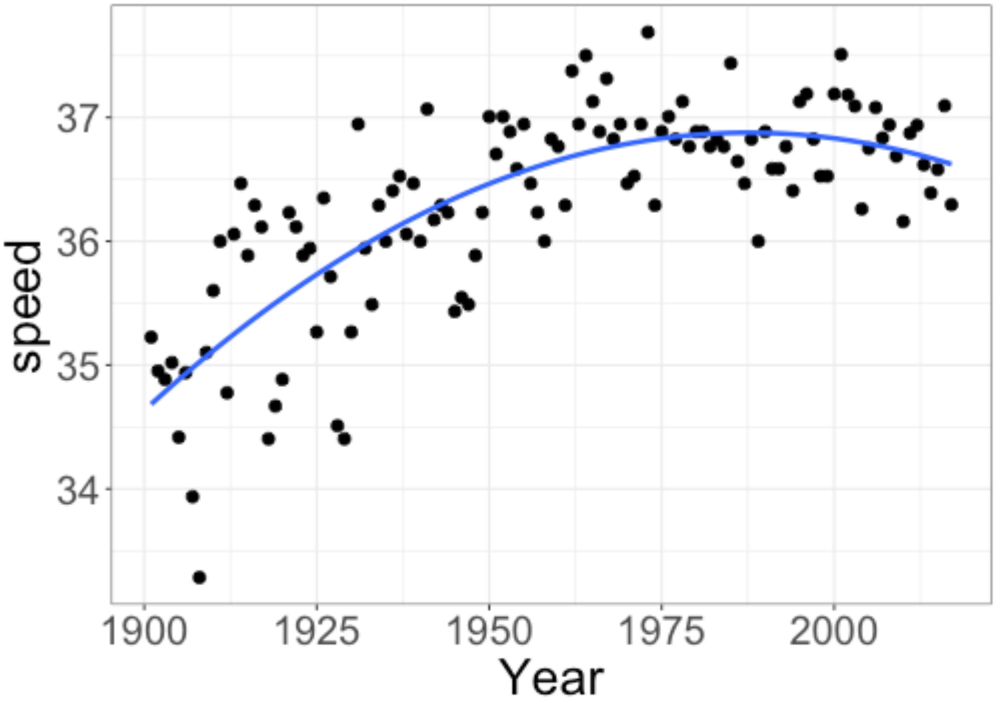


What order polynomial do you think we should use?

Polynomial transformation

Let's try a polynomial of order 2:

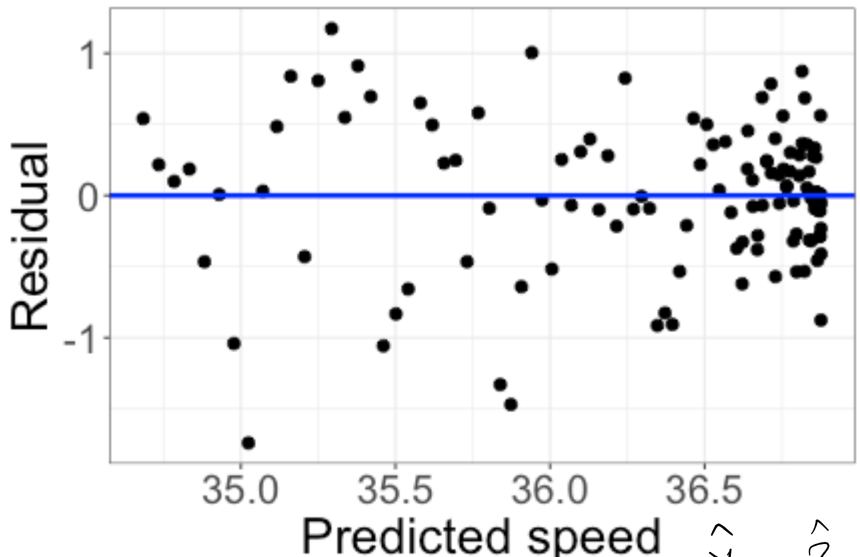
$$\text{Speed} = \beta_0 + \beta_1\ \text{Year} + \beta_2\ \text{Year}^2 + \varepsilon$$



Diagnostic plots

We use residual plots to check the model assumptions after transforming our predictor, just like with other transformations.

$Y - \hat{Y}$



$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$

How does the shape assumption look now?

Fitting the model in R

response

explanatory variable

```
derby_lm <- lm(speed ~ poly(Year, 2, raw=TRUE),
               data = derby)
```

order of polynomial

- `poly` means "fit a polynomial"
- `poly(..., 2, ...)` specifies a second-order polynomial (quadratic)

Fitting the model in R

```
derby_lm <- lm(speed ~ poly(Year, 2, raw=TRUE),
               data = derby)
derby_lm
```

```
##
## Call:
## lm(formula = speed ~ poly(Year, 2, raw = TRUE), data = d
##
## Coefficients:
##                (Intercept)  poly(Year, 2, raw = TRUE)1
##                 -1.120e+03                   1.164e+00
## poly(Year, 2, raw = TRUE)2
##                 -2.928e-04
```

$$\widehat{\text{Speed}} = -1120 + 1.16\ \text{Year} - 0.00029\ \text{Year}^2$$

Interpreting the fitted model

$$\widehat{\text{Speed}} = -1120 + 1.16 \text{ Year} - 0.00029 \text{ Year}^2$$

How would I interpret the intercept of the fitted polynomial regression model?

Interpreting the fitted model

$$\widehat{\text{Speed}} = -1120 + 1.16\ \text{Year} - 0.00029\ \text{Year}^2$$

> How would I interpret the intercept of the fitted polynomial regression model?

If Year = 0, we predict a speed of -1120 mph (doesn't make much sense, since year is nowhere near 0)

Interpreting the fitted model

$$\widehat{\text{Speed}} = -1120 + 1.16\ \text{Year} - 0.00029\ \text{Year}^2$$

How would I interpret the terms on Year and Year^2?

year 1900 $\rightarrow$ 1901

$-1120 + 1.16(1900) - 0.00029(1900)^2$

$\rightarrow -1120 + 1.16(1901) - 0.00029(1901)^2$

Interpreting the fitted model

$$\widehat{\text{Speed}} = -1120 + 1.16 \text{ Year} - 0.00029 \text{ Year}^2$$

> How would I interpret the terms on Year and Year^2?

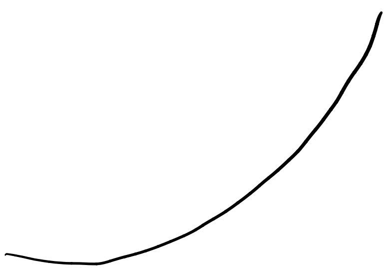
They're kind of difficult to interpret -- if I change Year, I also change Year^2

Interpreting the fitted model

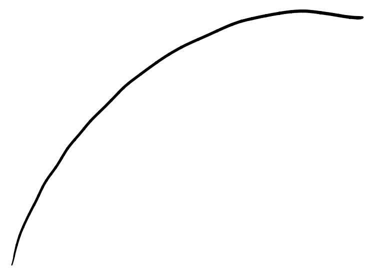
$$\widehat{\text{Speed}} = -1120 + 1.16\ \text{Year} - 0.00029\ \text{Year}^2$$

> Can I interpret the *sign* on Year^2?

x^2



$-x^2$



Interpreting the fitted model

$$\widehat{\text{Speed}} = -1120 + 1.16\ \text{Year} - 0.00029\ \text{Year}^2$$

> Can I interpret the *sign* on Year^2?

Yes (for a polynomial of order 2 specifically). The sign on Year^2 tells us about the concavity of the relationship.

Class activity

https://sta112-s26.github.io/class_activities/ca_10.html

- Experiment with polynomial regression
- Submit HTML at end of class