F-test in Multiple Regression

David Gerard

2018-12-07

Learning Objective

- Test for including multiple variables at the same time.
- Section 10.3 in the book

Case Study

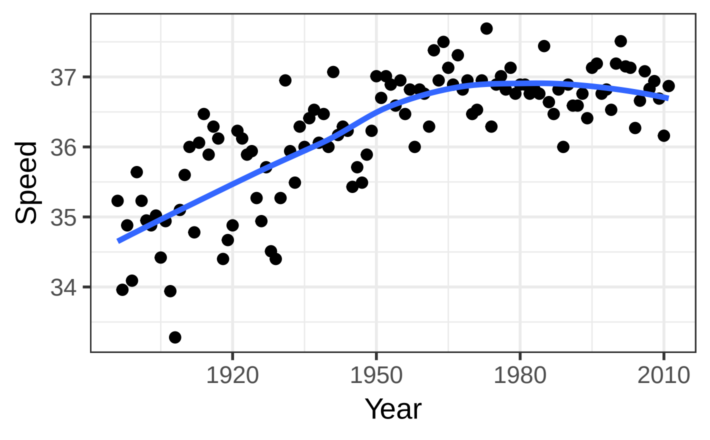
- Kentucky Derby
- Speed vs Year and Yeat^2.

```
library(Sleuth3)
data(ex0920)
head(ex0920)
```

```
##   Year         Winner Starters NetToWinner  Time Speed Tr
## 1 1896      Ben Brush        8        4850 127.8 35.23 Du
## 2 1897    Typhoon II         6        4850 132.5 33.96 He
## 3 1898        Plaudit        4        4850 129.0 34.88  C
## 4 1899         Manuel        5        4850 132.0 34.09  H
## 5 1900 Lieut. Gibson         7        4850 126.2 35.64  H
## 6 1901  His Eminence         5        4850 127.8 35.23  H
```

Year vs Speed

```
qplot(Year, Speed, data = ex0920) +
  geom_smooth(se = FALSE)
```



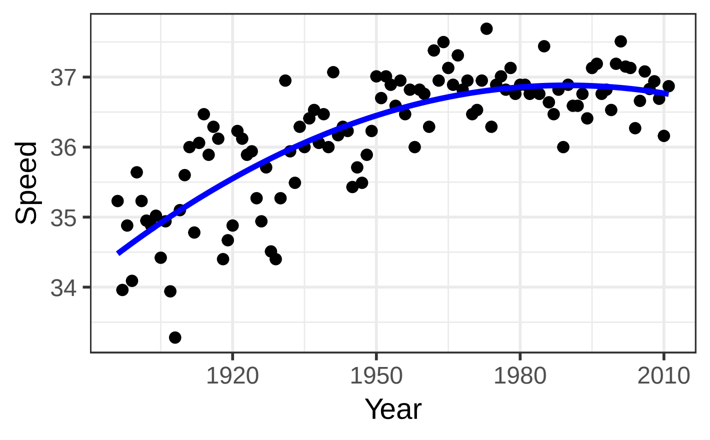
Goal

- Get a p-value for the association between year and Speed.
- It is clear that a quadratic model would be better than a linear model.
- $\mu(Speed|Year) = \beta_0 + \beta_1 Year + \beta_2 Year^2$
- So to see if year is important, we need to test:
  - $H_0 : \beta_1 = \beta_2 = 0$
  - $H_A :$ either $\beta_1 \neq 0$ or $\beta_2 \neq 0$

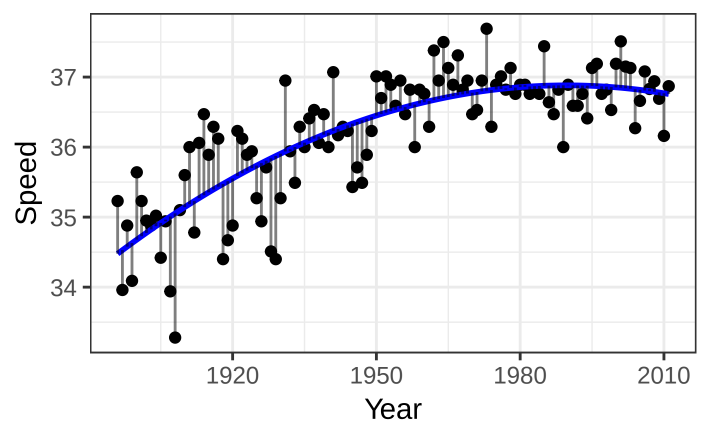
Full and Reduced Models:

- Full Model: $\mu(Speed|Year) = \beta_0 + \beta_1 Year + \beta_2 Year^2$
- Reduced Model: $\mu(Speed|Year) = \beta_0$
- Use F-test strategy to run this hypothesis test.
    1. Fit both full and reduced models.
    2. Calculate sum of squared residuals under both models and the corresponding degrees of freedom.
    3. Calculate the F-statistic.
    4. Compare to theoretical F-distribution under H_0

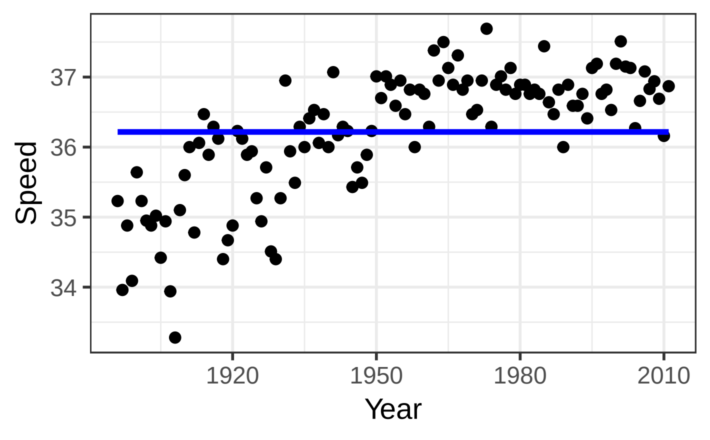
Fit Under Full



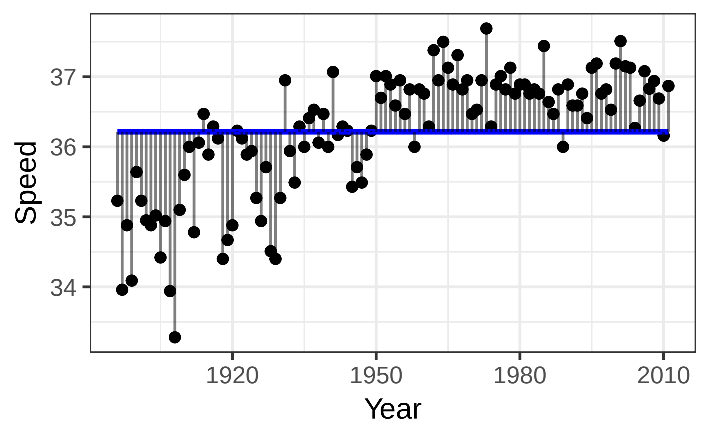
Residuals under Full



Fit under Reduced



Residuals under Reduced



In R

- First, fit both models

```r
ex0920$Year2 <- ex0920$Year ^ 2
lmfull <- lm(Speed ~ Year + Year2, data = ex0920)
lmreduced <- lm(Speed ~ 1, data = ex0920)
```

- Then use anova() with the reduced model as the first argument.

```r
anova(lmreduced, lmfull)
```

```
## Analysis of Variance Table
##
## Model 1: Speed ~ 1
## Model 2: Speed ~ Year + Year2
##   Res.Df  RSS Df Sum of Sq   F Pr(>F)
## 1    115 93.0
## 2    113 33.1  2      59.9 102 <2e-16
```

What is that Table?

```
## Analysis of Variance Table
##
## Model 1: Speed ~ 1
## Model 2: Speed ~ Year + Year2
##   Res.Df  RSS Df Sum of Sq   F Pr(>F)
## 1    115 93.0
## 2    113 33.1  2      59.9 102 <2e-16
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
$df_{reduced}$	$RSS_{reduced}$				
df_{full}	RSS_{full}	df_{extra}	ESS	F-stat	p-value

F-test

- We can use the F-test for any two **nested** models.
- **Nested**: The reduced model is a special case of the full model by setting constraints on some of the parameters of the full.

Another Example

- $\mu(Speed|Year, Starters) = \beta_0 + \beta_1 Year + \beta_2 Year^2 + \beta_3 Starters + \beta_4 Starters^2$
- $H_0 : \beta_3 = \beta_4 = 0$
- $H_A :$ either $\beta_3 \neq 0$ or $\beta_4 \neq 0$
- Full Model: $\mu(Speed|Year, Starters) = \beta_0 + \beta_1 Year + \beta_2 Year^2 + \beta_3 Starters + \beta_4 Starters^2$
- Reduced Model: $\mu(Speed|Year, Starters) = \beta_0 + \beta_1 Year + \beta_2 Year^2$

Another Example

```
ex0920$Starters2 <- ex0920$Starters ^ 2
lmfull <- lm(Speed ~ Year + Year2 + Starters +
               Starters2, data = ex0920)
lmreduced <- lm(Speed ~ Year + Year2, data = ex0920)
anova(lmreduced, lmfull)
```

```
## Analysis of Variance Table
##
## Model 1: Speed ~ Year + Year2
## Model 2: Speed ~ Year + Year2 + Starters + Starters2
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1    113 33.1
## 2    111 30.9  2      2.18 3.92  0.023
```

Example of a non-nested model

- Model 1: $\mu(Speed|Year, Starters) = \beta_0 + \beta_1 Year + \beta_2 Year^2$
- Model 2:
  $\mu(Speed|Year, Starters) = \beta_0 + \beta_1 Starters + \beta_2 Starters^2$
- **Cannot** use an F-test to compare these two models.
- Why? Mathematical theory only gaurantees the F-distribution when the models are nested.
- When models are not nested, use adjusted R^2, C_p, AIC, or BIC methods from section 12.4 (more on this later).