Modeling Nonlinear Associations

Kevin Cummiskey

November 6, 2019

Review

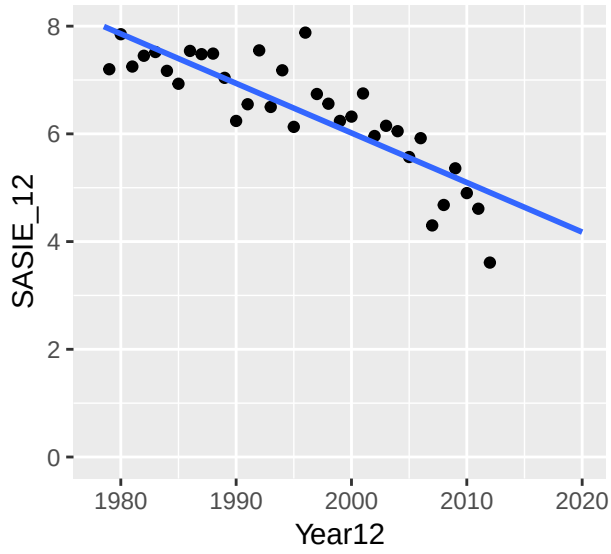
Explain a situation in which you would use the partial F-test. As part of your answer, describe your research question, the variables you would collect, the models you would fit, and hypotheses you would test.

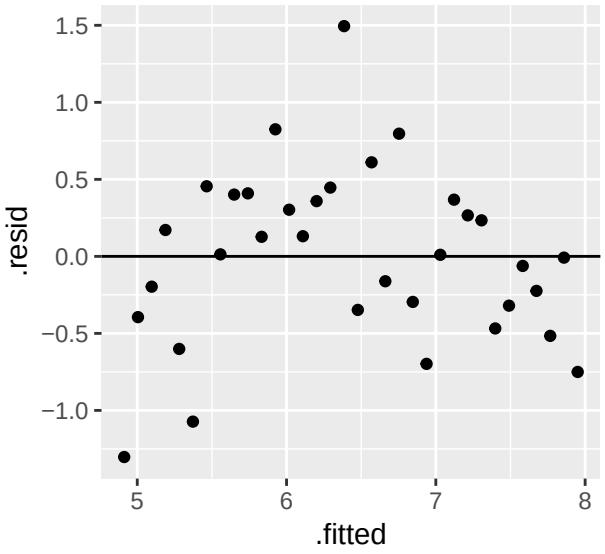
Example 5.3 Arctic Sea Ice

The purpose of this analysis is to predict when there will be no sea ice in Artic Ocean.

```
library(gridExtra)
ice = read.table(file = "http://www.isi-stats.com/isi2/data/ArticSeaIce.txt",
                 header = T)
model_linear = lm(SASIE_12 ~ Year12, data = ice)

p1 = ice %>% ggplot(aes(x = Year12, y = SASIE_12)) +
  geom_point() +
  geom_smooth(method = "lm", se = F, fullrange = T) +
  ylim(0,8) +
  xlim(1978, 2020)
p2 = model_linear %>% fortify() %>%
  ggplot(aes(x = .fitted, y = .resid)) + geom_point() +
  geom_hline(yintercept = 0)
grid.arrange(p1,p2, ncol = 2)
```





Is a linear model appropriate? What are the implications of choosing the wrong model?

Let's fit the following model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

How do we interpret the coefficients?

```
# Let's try to fit the model
model_quadratic = lm(SASIE_12 ~ Year12 + I(Year12^2), data = ice)
summary(model_quadratic)
```

```
##
## Call:
## lm(formula = SASIE_12 ~ Year12 + I(Year12^2), data = ice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9513 -0.1959  0.0493  0.2474  1.1259
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.510e+04  3.726e+03  -4.053 0.000315 ***
## Year12       1.524e+01  3.735e+00   4.080 0.000293 ***
## I(Year12^2) -3.841e-03  9.357e-04  -4.104 0.000273 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4691 on 31 degrees of freedom
## Multiple R-squared:  0.8217, Adjusted R-squared:  0.8102
## F-statistic: 71.43 on 2 and 31 DF,  p-value: 2.472e-12
```

What's unusual in the output?

```
# Let's standardize year first
ice = ice %>% mutate(std.year = (Year12-mean(Year12))/sd(Year12))
```

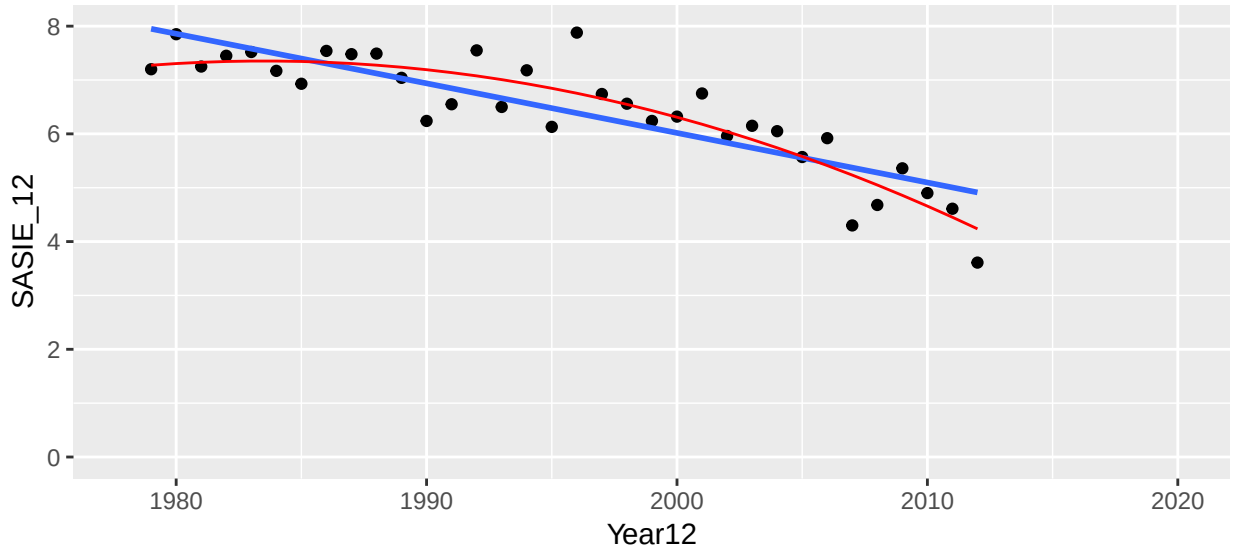
```
## Warning: The `printer` argument is deprecated as of rlang 0.3.0.
## This warning is displayed once per session.
```

```
model_std.quadratic = lm(SASIE_12 ~ std.year + I(std.year^2), data = ice)
summary(model_std.quadratic)
```

```
##
## Call:
## lm(formula = SASIE_12 ~ std.year + I(std.year^2), data = ice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9513 -0.1959  0.0493  0.2474  1.1259
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.80113    0.12076  56.317  < 2e-16 ***
## std.year      -0.91669    0.08166 -11.226 1.89e-12 ***
## I(std.year^2) -0.38086    0.09279  -4.104 0.000273 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4691 on 31 degrees of freedom
## Multiple R-squared:  0.8217, Adjusted R-squared:  0.8102
## F-statistic: 71.43 on 2 and 31 DF,  p-value: 2.472e-12
```

Why do we need to standardize year first?

```
ice = ice %>% left_join(model_std.quadratic %>% fortify())
ice %>% ggplot(aes(x = Year12, y = SASIE_12)) +
  geom_point() + geom_smooth(method = "lm", se = F) +
  geom_line(aes(y = .fitted), col = "red") +
  ylim(0,8) +
  xlim(1978,2020)
```



Let's look at the second-order term.

```
anova(model_std.quadratic, model_linear)
```

```
## Analysis of Variance Table
##
## Model 1: SASIE_12 ~ std.year + I(std.year^2)
## Model 2: SASIE_12 ~ Year12
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     31  6.822
## 2     32 10.529 -1   -3.7072 16.846 0.0002731 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What is the name of the test above?

What hypotheses are being tested?

What do you conclude from this test?

What year does each model predict there will be no more sea ice in the Arctic Ocean?