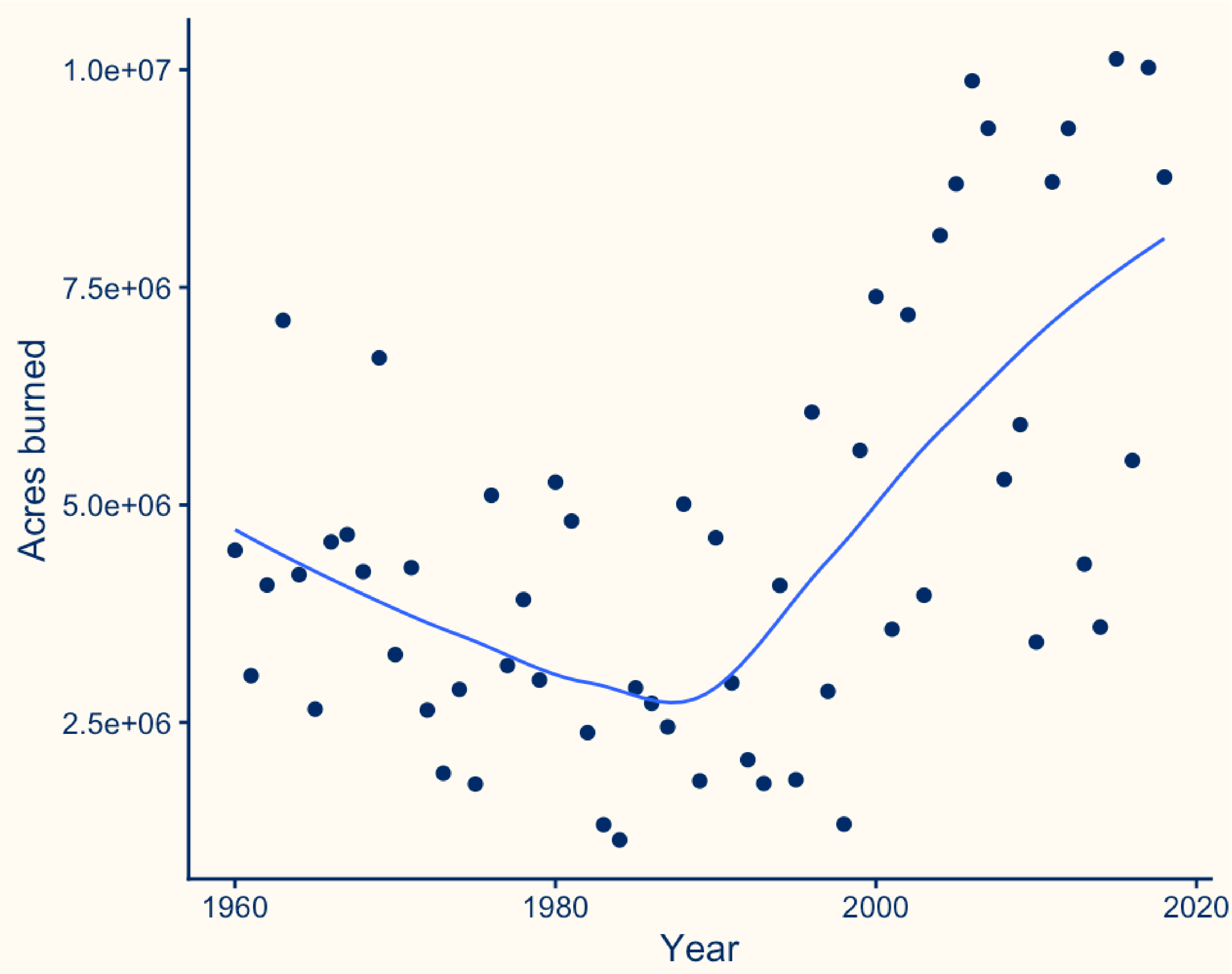
Polynomial Regression

Stat 230: Applied Regression Analysis

PDF version of slides

Wildfires

- The National Interagency Coordination Center at the National Interagency Coordination Center compiles annual wildland fire statistics for federal and state agencies.
- This information is provided through Situation Reports, which have been in use for several decades.
- Our goal is to model the number of acres burned over the years

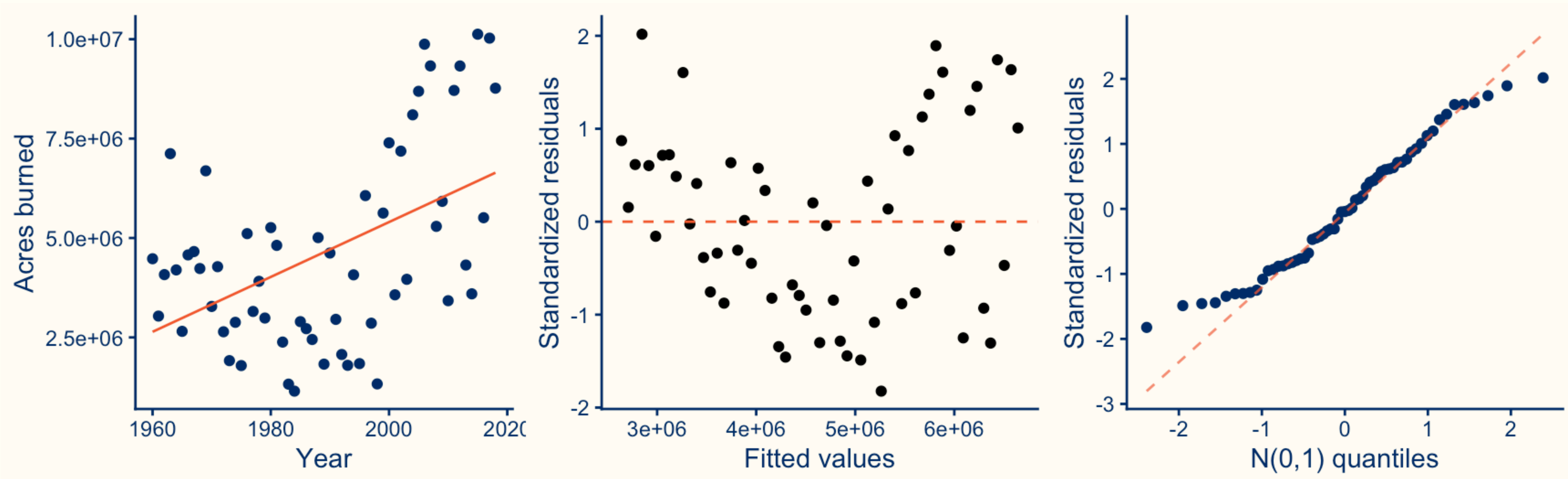


1.0e+07
7.5e+06
5.0e+06
2.5e+06
Acres burned
1960
1980
2000
2020
Year

Option 1: SLR model

$\mu\{y|x\} = \beta_0 + \beta_1 x$

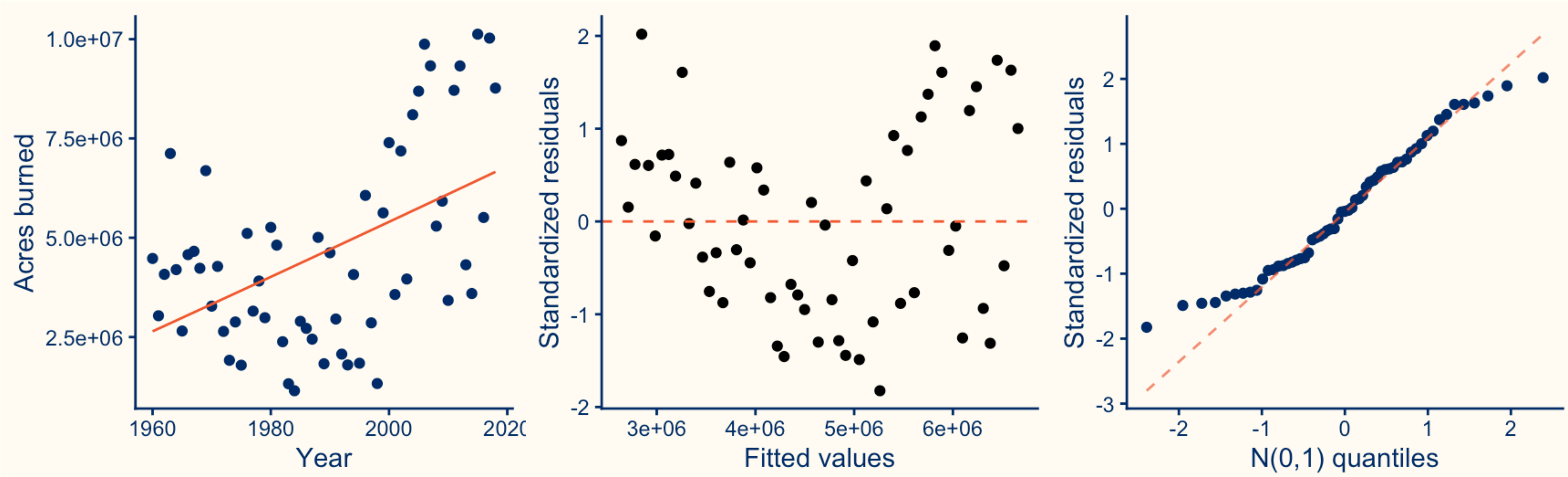
Is the fit reasonable?



Option 2: Transform X

$\mu\{y|x\} = \beta_0 + \beta_1 x^2$

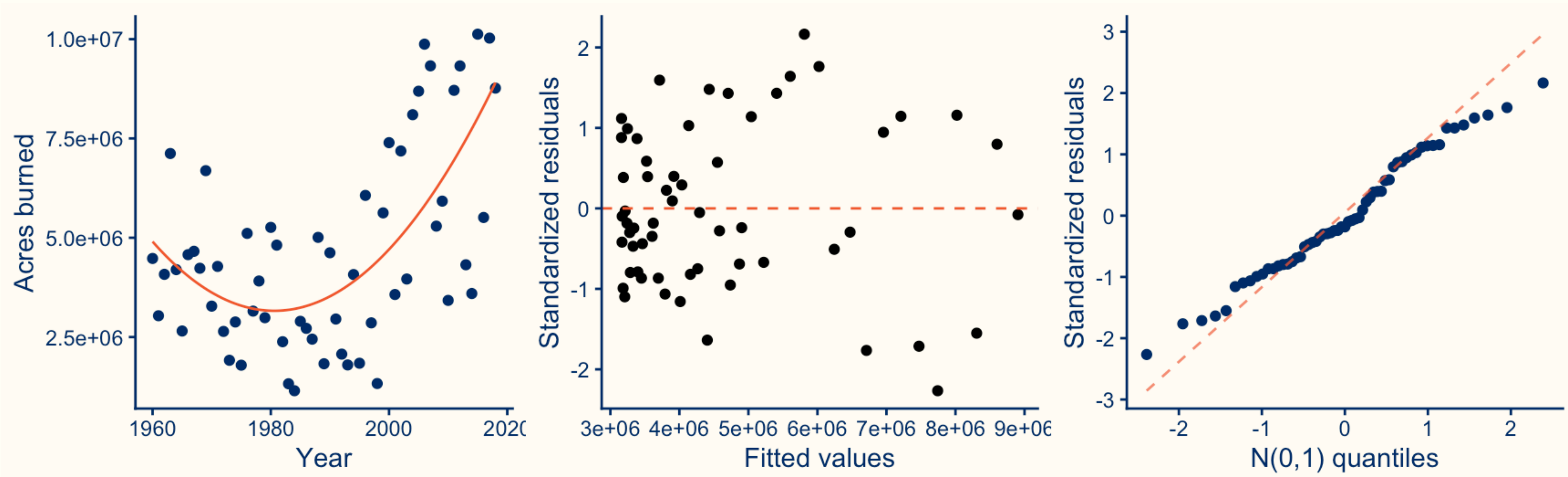
Is the fit reasonable?



Option 3: Polynomial model

$\mu\{y|x\} = \beta_0 + \beta_1 x + \beta_2 x^2$

Is the fit reasonable?



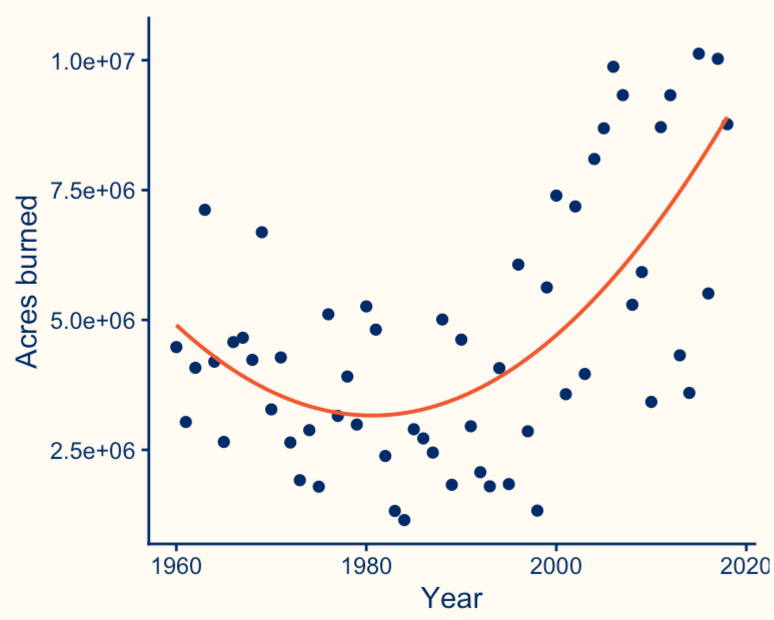
The polynomial regression model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma)$$

Assumptions — same as in SLR

1. $\mu\{Y|x_i\}$, is a linear function
2. For each x_i, the sub-population of responses is normally distributed
3. The standard deviation for each sub-population is σ
4. Independent observations

Interpreting the model



Focus on the expected change in y for a specific one-unit increase in x

- e.g. change in acres burned from 1985 to 1990
- e.g. change in acres burned from 2005 to 2010

$$\begin{aligned}\mu\{y|x+1\} - \mu\{y|x\} &= \left[\beta_0 + \beta_1(x+1) + \beta_2(x+1)^2\right] - \left[\beta_0 + \beta_1 x + \beta_2 x^2\right] \\ &= \beta_1 + \beta_2\,(2x+1)\end{aligned}$$

Inferences about coefficients

Inference uses the same t-based tools as SLR, but with

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n - (k+1)}}$$

i.e. the degrees of freedom of the t-distribution change

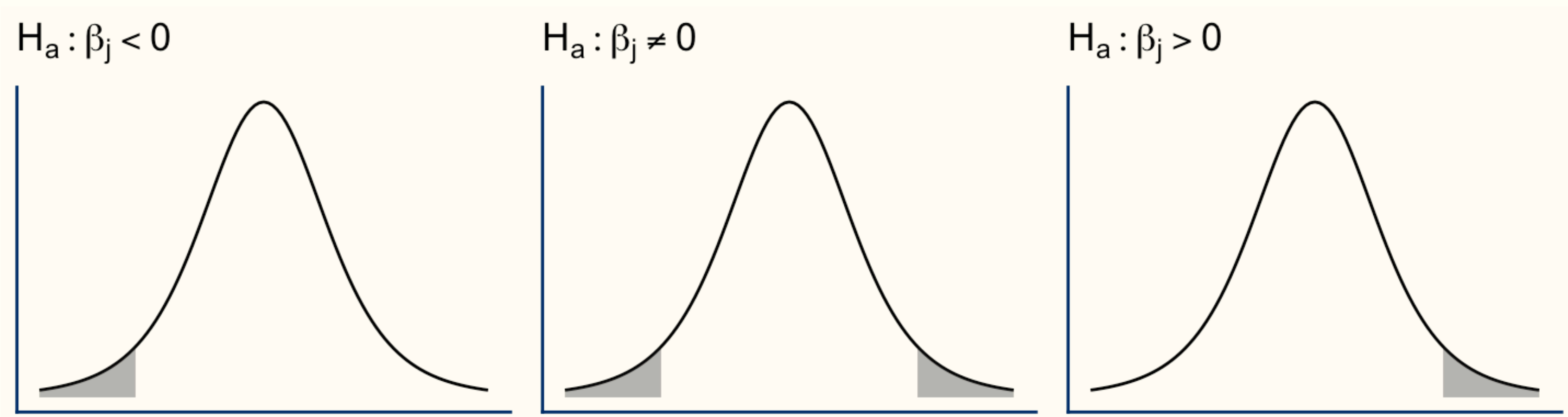
Testing a single coefficient

Hypotheses: $H_0 : \beta_j = \#$ vs. $H_a : \beta_j \begin{smallmatrix} < \\ \neq \\ > \end{smallmatrix} \#$

Test statistic: $t = \dfrac{\hat{\beta_j} - \#}{SE(\beta_j)}$

Reference distribution: t distribution with d.f. $= n - (k + 1)$

p-value: Area in the tail(s) specified by H_a



CIs for a single coefficient

$$\widehat{\beta}_j \pm t^*_{n-(k+1)} \cdot SE(\widehat{\beta}_j)$$

Wildfires example

Term	Estimate	SE	Lower	Upper
(Intercept)	16109806404	3793719340	8510073345	23709539462
Year	-16264428	3814896	-23906583	-8622273
I(Year^2)	4106	959	2185	6027

Your turn

Would a higher-order polynomial (e.g. cubic, quartic, quintic) provide a better fit to the wildfire data?

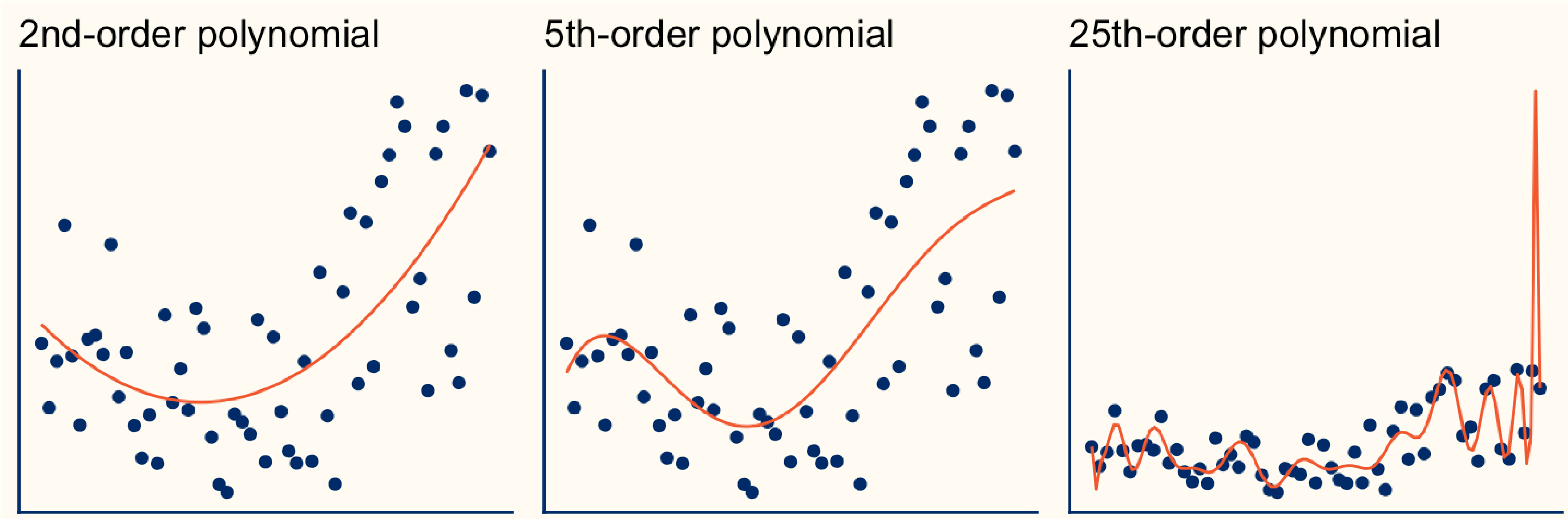
Work through that example on the handout with your neighbors

A warning about this analysis

> *Prior to 1983, sources of these figures are not known, or cannot be confirmed, and were not derived from the current situation reporting process. As a result the figures prior to 1983 should not be compared to later data.*
>
> *— NIFC*

A warning about polynomials

High-order polynomial regression models will over-fit your data (i.e. pick up on peculiarities specific to your one sample from the population)



Multiple linear regression

Polynomial regression is one example of the multiple regression model, but there are numerous ways to incorporate multiple predictors into a model

$$\mu\{Y|X\} = \beta_0 + \beta_1 x + \beta_2 x^2$$

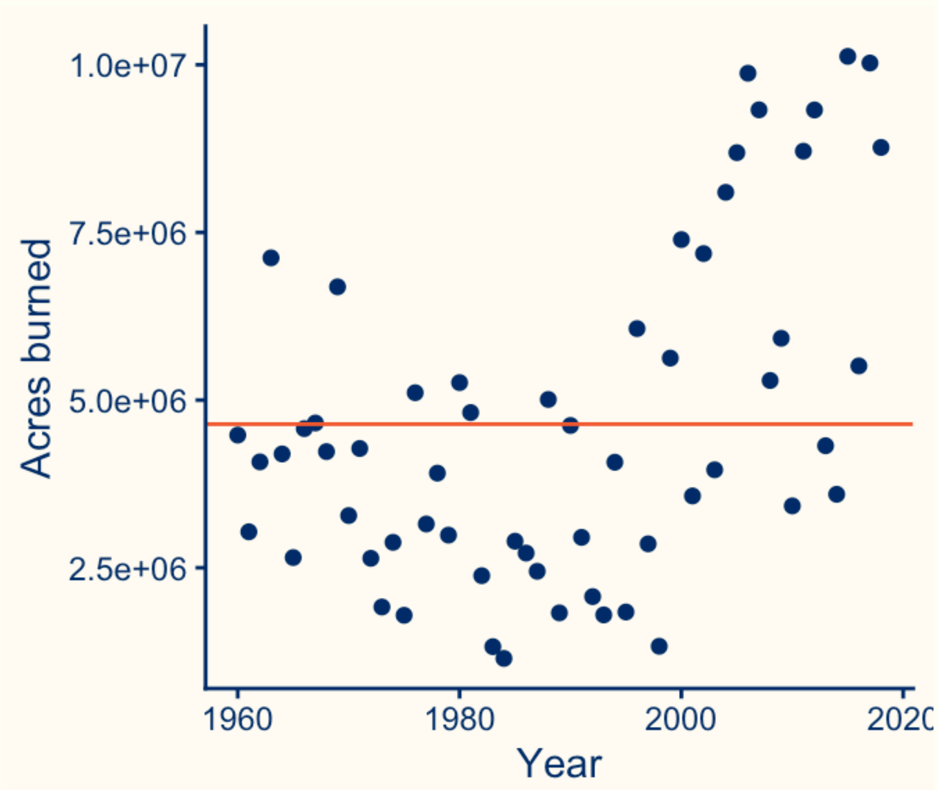
$$\mu\{Y|X\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\mu\{Y|X\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_1 x_2$$

Explained variance

The “null” model

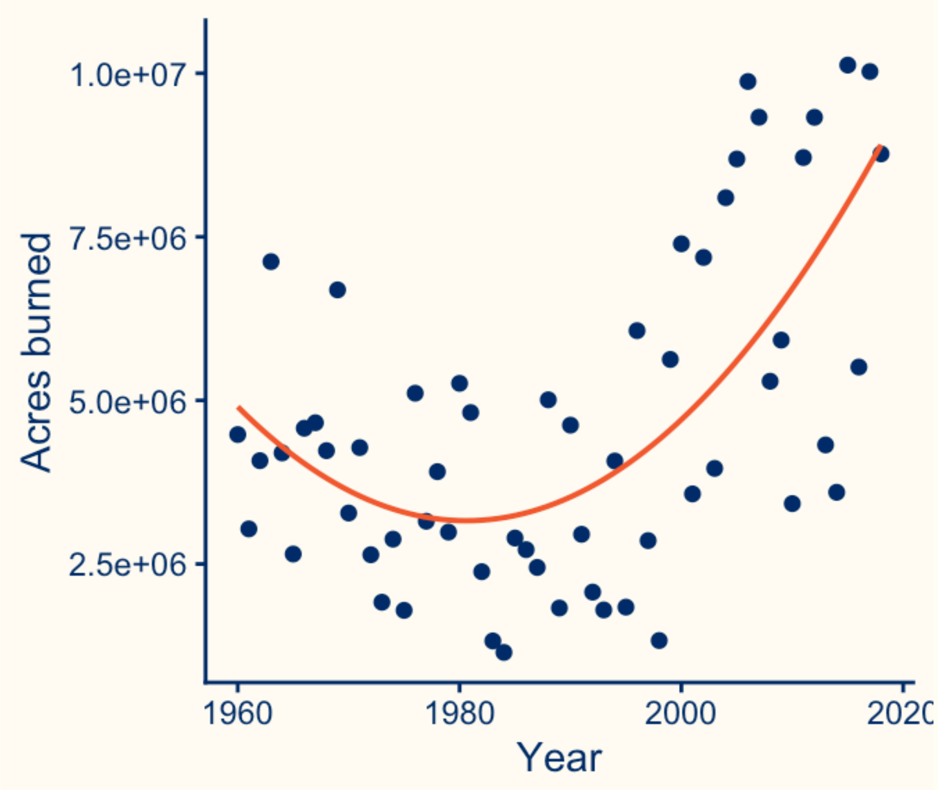
- Use the mean of Y as the prediction for all observations
- $\mu\{Y|X\} = \beta_0$
- Leaves a lot of variability unexplained



$$SD(Y) = 2.465548 \times 10^6 \text{ cm}$$

Polynomial model

- $\mu\{Y|X\} = \beta_0 + \beta_1 x + \beta_2 x^2$
- Using a predictor explains more variability

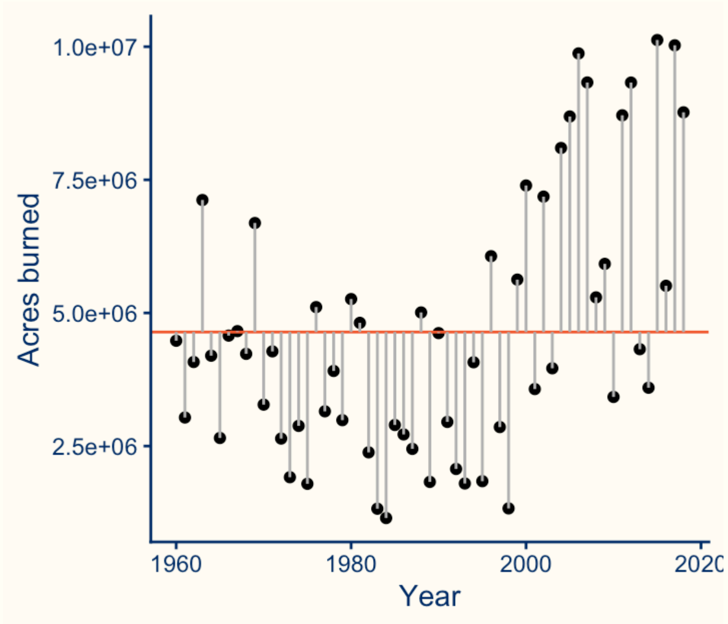


$SD(Y) = 1.9098377 \times 10^6$ cm

SSTotal

$$SSTotal = \sum(Y_i - \bar{Y})^2$$

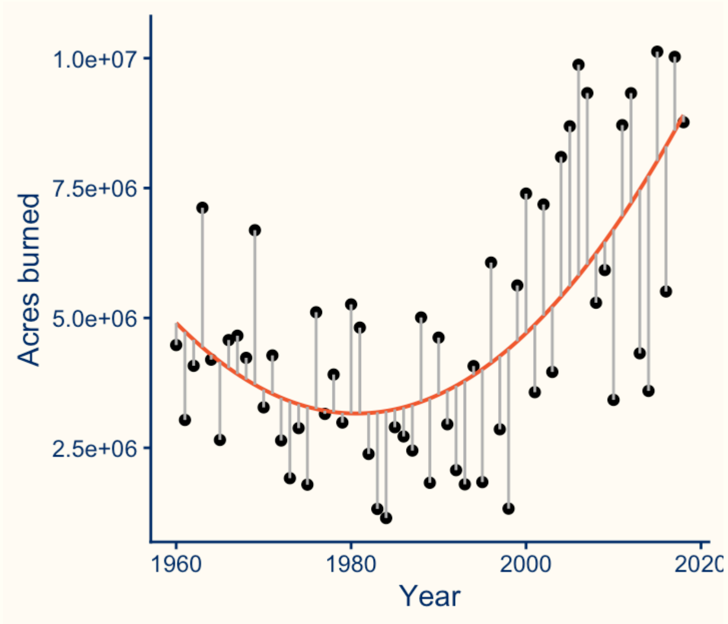
Measures the overall variability in Y



SSResidual (aka SSError)

$$\text{SSResidual} = \sum(Y_i - \widehat{Y}_i)^2$$

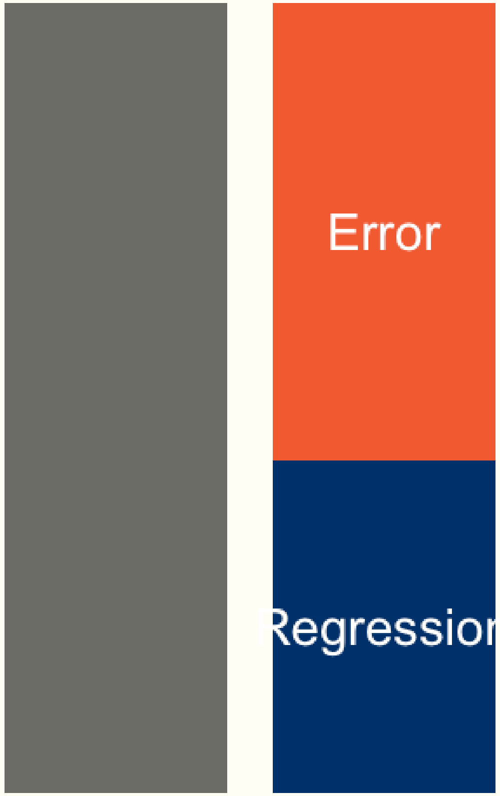
Measures the variability unexplained by the model



SSRegression

Measures the variability explained by the model

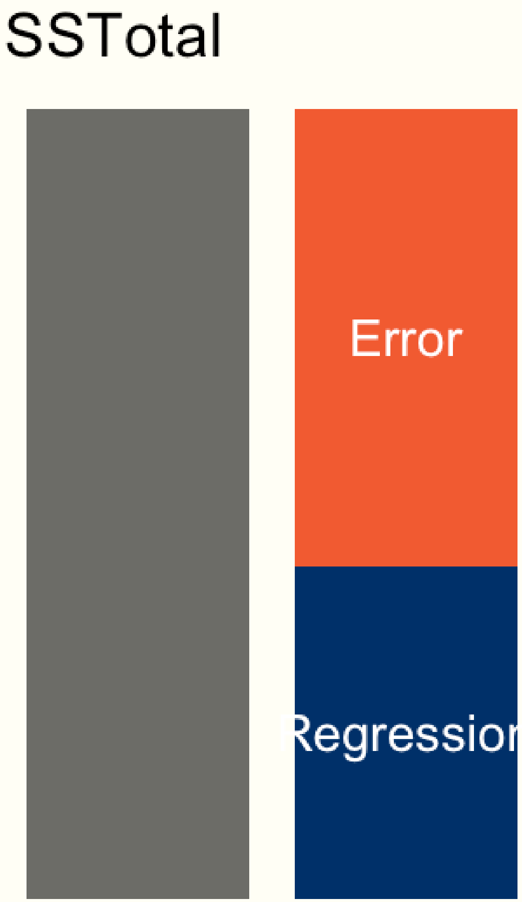
SSTotal



$$\text{SSRegression} = SSTotal - SSError$$

Coefficient of Determination: R^2

Proportion of the total variation in y explained by the linear regression model



$$R^2 = 0.421 = \frac{SSRegr}{SST} = 1 - \frac{SSE}{SST}$$

⚠️ Caution

R^2 only addresses how close the fitted values are to the data, on average. It says nothing about the validity of the model.

Comparing models

Wildfires example

Suppose we wish to compare a quadratic and quartic model for the wildfires data set:

- Quadratic: $\mu\{Y|X\} = \beta_0 + \beta_1 x + \beta_2 x^2$
- Quartic: $\mu\{Y|X\} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$

How do we decide which model is preferred?

Comparing models

Can we compare R^2 values?

- No! More complex models will always have a higher R^2 value, even if the additional predictors are not useful.

Can we run individual t-tests?

- No! The tests are not necessarily independent, and the Type I error rate will be inflated.

Comparing models with an F-test

Full model $\mu\{Y|X\} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$

Reduced model $\mu\{Y|X\} = \beta_0 + \beta_1 x + \beta_2 x^2$

Hypotheses $H_0: \ \beta_3 = \beta_4 = 0$

H_a : at least one $\beta_j \neq 0$, $j = 3,4$

Comparing models with an F-test

Test statistic

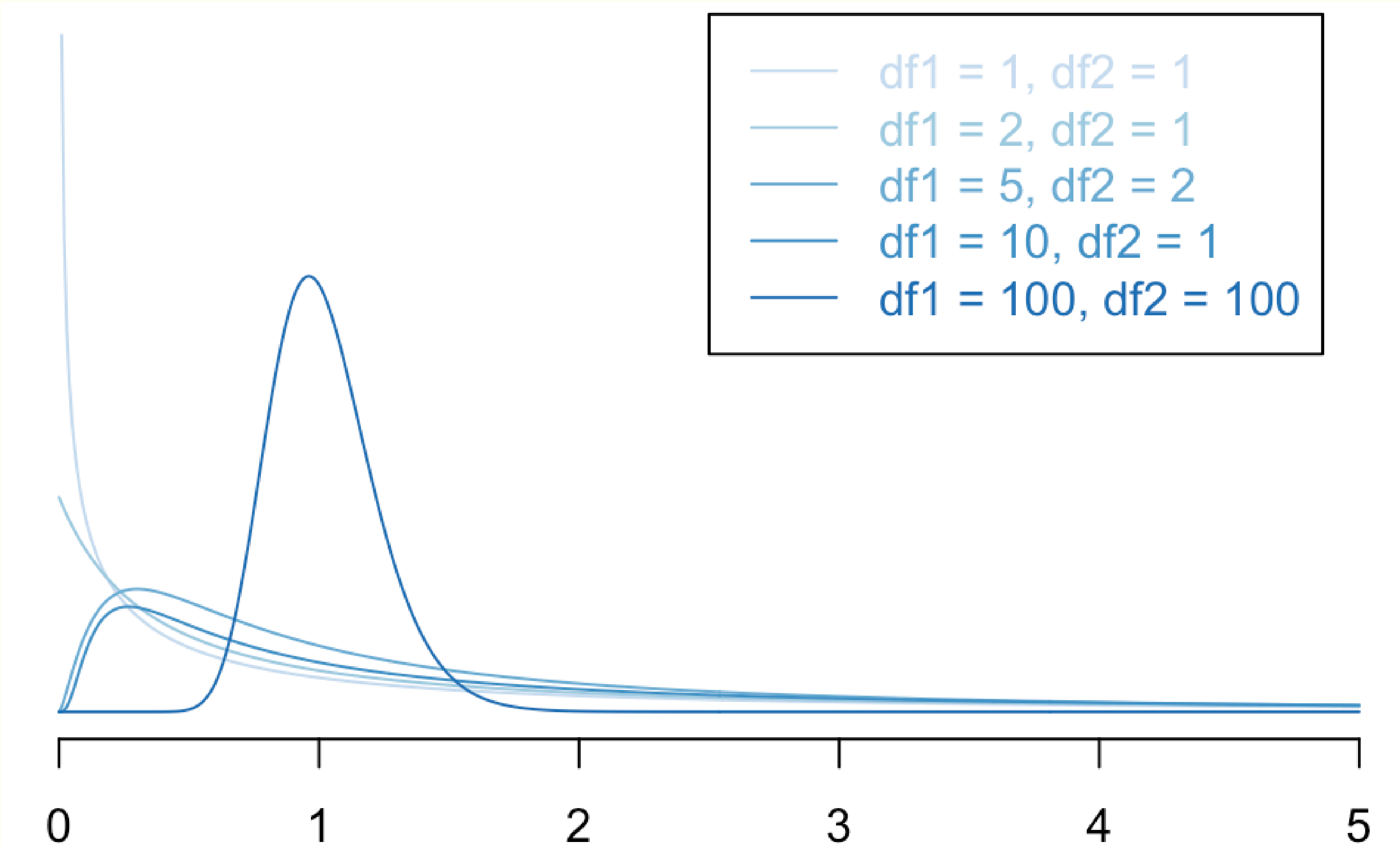
$$
\begin{aligned}
F &= \frac{(R^2_{\text{full}} - R^2_{\text{reduced}})/d}{(1 - R^2_{\text{full}})/df_{\text{full}}} \\
&= \frac{(\text{SSR}_{\text{full}} - \text{SSR}_{\text{reduced}})/d}{\text{MSE}_{\text{full}}} \\
&= \frac{(\text{SSE}_{\text{reduced}} - \text{SSE}_{\text{full}})/d}{\text{MSE}_{\text{full}}}
\end{aligned}
$$

where

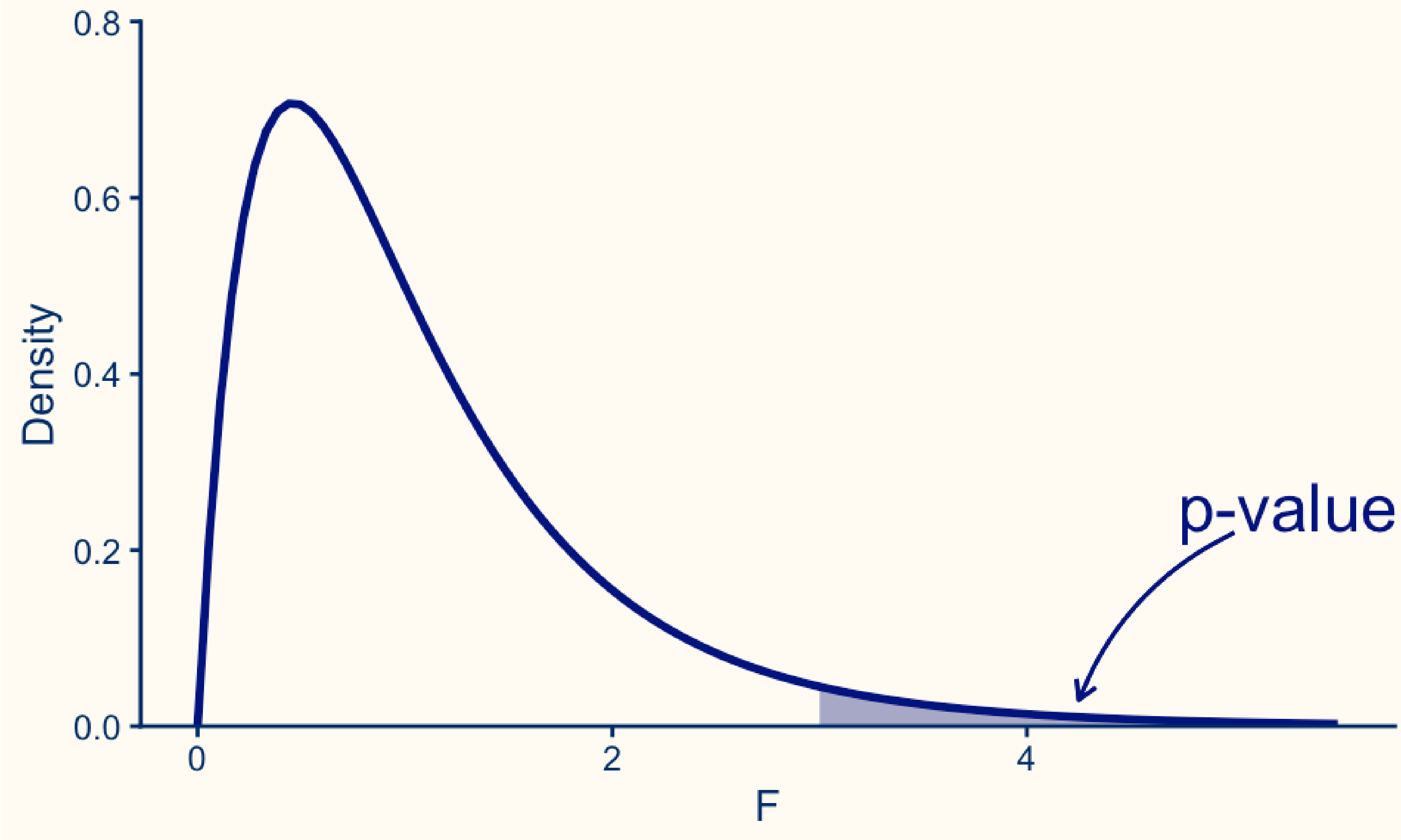
- $d = \mathrm{df}_{full} - \mathrm{df}_{reduced}$ = # betas being tested
- $\mathrm{df}_i = n - (p + 1)$ = error d.f. for model i
- $\mathrm{MSE}_{\mathrm{full}} = \dfrac{\mathrm{SSE}_{\mathrm{full}}}{\mathrm{df}_{\mathrm{full}}}$

F distribution

The F-statistics follows an F distribution with $\mathrm{df}_{full} - \mathrm{df}_{reduced}$ and $n - p - 1$ d.f.



Upper-tail p-values



To obtain upper-tail areas: `1 - pf(stat, df1, df2)`

Putting it all together

model	r.squared	df	df.residual	nobs
Full	0.453	4	54	59
Reduced	0.421	2	56	59

$$F = \frac{(R^2_{\text{full}} - R^2_{\text{reduced}})/d}{(1 - R^2_{\text{full}})/df_{\text{full}}} = \frac{(0.453 - 0.421)/2}{(1 - 0.453)/54} \approx 1.58$$

```r
1 - pf(1.58, 2, 54)
```

```
[1] 0.2153484
```

There is no evidence that the quartic model is an improvement over the quadratic model ($F = 4.467$,

Reading R output

```
Call:
lm(formula = Acres ~ poly(Year, 4), data = wildfires)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     4641410     245955  18.871  < 2e-16 ***
poly(Year, 4)1  9025270    1889218   4.777 1.40e-05 ***
poly(Year, 4)2  8177004    1889218   4.328 6.55e-05 ***
poly(Year, 4)3 -1194695    1889218  -0.632   0.5298    
poly(Year, 4)4 -3177711    1889218  -1.682   0.0983 .  

Residual standard error: 1889000 on 54 degrees of freedom
Multiple R-squared:  0.4534,	Adjusted R-squared:  0.4129 
F-statistic:  11.2 on 4 and 54 DF,  p-value: 1.096e-06
```

Extra sums of squares F-test in R

```r
full <- lm(Acres ~ poly(Year, 4), data = wildfires)
reduced <- lm(Acres ~ poly(Year, 2), data = wildfires)
anova(reduced, full)
```

```
Analysis of Variance Table

Model 1: Acres ~ poly(Year, 2)
Model 2: Acres ~ poly(Year, 4)
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1     56 2.0426e+14
2     54 1.9273e+14  2 1.1525e+13 1.6146 0.2084
```