

STAT 224 Lecture 10

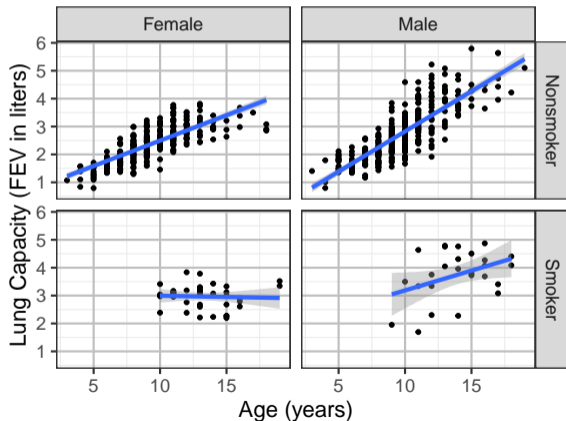
Polynomial Models

Yibi Huang
Department of Statistics
University of Chicago

Does FEV Grow Linearly in Age?

Children stop growing after they turn adults.

FEV might not grow linearly with age, at least for **female nonsmokers** in the plots below.



Test of Non-Linearity (Female Nonsmokers)

Let's focus on female nonsmokers first.

```
f.nonsmokers = subset(fevdata, sex == "Female" & smoke == "Nonsmoker")
```

To test non-linearity, one can add some nonlinear function of age, e.g. age^2 and see if the nonlinear term is significant.

```
summary(lm(fev ~ age + I(age^2), data=f.nonsmokers))$coef
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.50746	0.21039	-2.412	1.652e-02
age	0.43979	0.04297	10.235	4.715e-21
I(age^2)	-0.01298	0.00212	-6.121	3.176e-09

► The tiny p-value for the age^2 is strong evidence of non-linearity.

Polynomial Models

- ▶ Fitting the polynomial model

$$\text{fev} = \beta_0 + \beta_1 \text{age} + \beta_2 (\text{age})^2 + \varepsilon$$

doesn't mean we believe it's correct. It might just be a decent approximation to the true underlying nonlinear model

$$\text{fev} = f(\text{age}) + \varepsilon$$

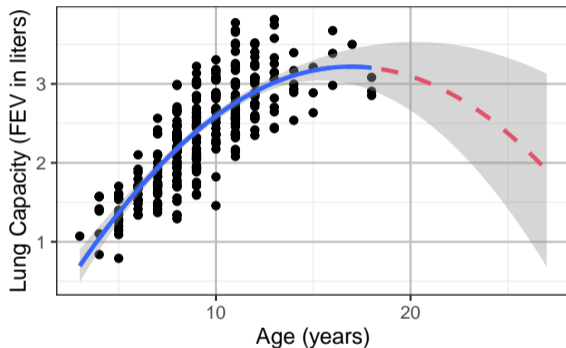
- ▶ One can try higher-order polynomials

$$\text{fev} = \beta_0 + \beta_1 \text{age} + \beta_2 (\text{age})^2 + \dots + \beta_k (\text{age})^k + \varepsilon$$

if lower-order ones don't capture the nonlinear pattern well.

Extrapolation is Dangerous!

Lung capacity decreases after children turn adults?



We are not sure whether the nonlinear relations is a polynomial (it's just an approximation!). Extrapolating the model beyond the range of data is dangerous.

Test of Non-linearity (Male Nonsmokers)

```
m.nonsmokers = subset(fevdata, sex == "Male" & smoke == "Nonsmoker")
summary(lm(fev ~ age + I(age^2), data=m.nonsmokers))$coef
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.143622	0.298683	0.4809	0.63096439
age	0.245875	0.059137	4.1577	0.00004175
I(age^2)	0.002058	0.002821	0.7296	0.46619995

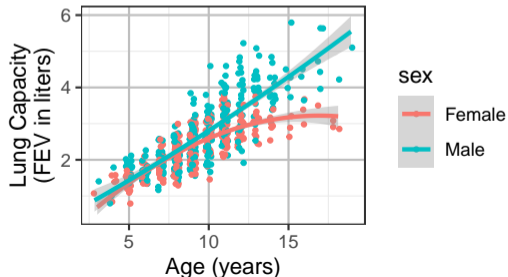
- ▶ The large P -value 0.466 for the age^2 means little evidence non-linearity
- ▶ This just means fev is approx. linear in age in the range of data for male smokers. Extrapolating the line beyond of the range of data remain dangerous
- ▶ The discrepancy in the significance of age^2 between boys and girls is an evidence of $\text{sex}:\text{age}^2$ interaction — lung capacities of girls stop growing earlier than boys.

Age:Sex Interactions — Nonsmokers Only

```
nonsmokers = subset(fevdata, smoke == "Nonsmoker")
summary(lm(fev ~ (age + I(age^2))*sex, data=nonsmokers))$coef
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.50746	0.263274	-1.927	5.440e-02
age	0.43979	0.053770	8.179	1.796e-15
I(age^2)	-0.01298	0.002653	-4.892	1.295e-06
sexMale	0.65108	0.369659	1.761	7.871e-02
age:sexMale	-0.19392	0.074369	-2.607	9.355e-03
I(age^2):sexMale	0.01504	0.003612	4.163	3.611e-05

Not surprisingly, the tiny P -value 3.61×10^{-5} for $I(\text{age}^2):\text{sexMale}$ indicates that boys and girls differ significantly in the curvatures of the growth curves of lung capacities.



```
lm(fev ~ (age + I(age^2))*sex, data=nonsmokers)$coef
```

(Intercept)	age	I(age^2)	sexMale
-0.50746	0.43979	-0.01298	0.65108
age:sexMale	I(age^2):sexMale		
-0.19392	0.01504		

$$\widehat{\text{fev}} \approx -0.507 + 0.44\text{Age} - 0.013(\text{Age})^2 + 0.651\text{Sex}_M \\ - 0.194(\text{Sex}_M \cdot \text{Age}) + 0.015(\text{Sex}_M \cdot \text{Age}^2)$$

The estimated growth curve for girls ($\text{Sex}_M = 0$) is

$$\widehat{\text{fev}} \approx -0.507 + 0.44\text{Age} - 0.013(\text{Age})^2,$$

The estimated growth curve for boys ($\text{Sex}_M = 1$) is

$$\widehat{\text{fev}} \approx (-0.507 + 0.651) + (0.44 - 0.194)\text{Age} + (-0.013 + 0.015)(\text{Age})^2 \\ = 0.144 + 0.246\text{Age} + 0.002(\text{Age})^2$$

Observe the coefficients are identical to the coefficients for the model including female nonsmokers only and the one for male nonsmokers only.

Interpretation of Coefficients in a Polynomial Model

Recall in Lecture 3 we said β_j = the regression coefficient for X_j , is the mean change in the response Y when X_j is increased by one unit **holding other X_i 's constant**.

However in a model that involves polynomial terms like

$$Y = \beta_0 + \underbrace{\beta_1 X_1 + \beta_2 X_1^2}_{\text{a polynomial of } X_1} + \beta_3 X_3 + \dots + \beta_p X_p + \varepsilon$$

it makes no sense to interpret a single coefficient for a polynomial like β_1 or β_2 since it's impossible to change X_1 while holding X_1^2 constant. We should interpret the entire polynomial altogether like,

*the mean of Y change with X_1 following the curve $\beta_1 X_1 + \beta_2 X_1^2$ **holding other X_i 's constant**.*

Test of Non-linearity (Smokers)

```
m.smokers = subset(fevdata, sex == "Male" & smoke == "Smoker")
```

```
summary(lm(fev ~ age + I(age^2), data=m.smokers))$coef
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-6.05764	4.7793	-1.267	0.21767
age	1.31395	0.7056	1.862	0.07539
I(age^2)	-0.04253	0.0255	-1.668	0.10889

```
f.smokers = subset(fevdata, sex == "Female" & smoke == "Smoker")
```

```
summary(lm(fev ~ age + I(age^2), data=f.smokers))$coef
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.98970	1.95968	3.056	0.004204
age	-0.43746	0.28431	-1.539	0.132627
I(age^2)	0.01536	0.01013	1.516	0.138246

- ▶ age^2 is insignificant for male smokers or female smokers, might be just due to the small sample size that makes it difficult to detect the non-linearity.

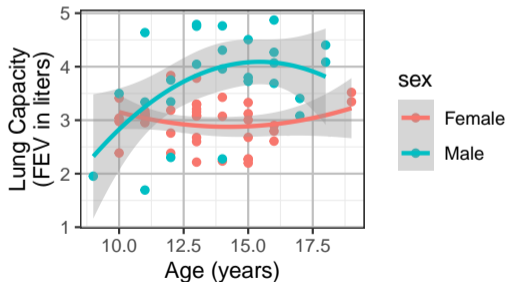
```

smokers = subset(fevdata, smoke == "Smoker")
summary(lm(fev ~ (age + I(age^2))*sex, data=smokers))$coef

```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.98970	2.79719	2.141	0.03639
age	-0.43746	0.40582	-1.078	0.28543
I(age^2)	0.01536	0.01447	1.062	0.29252
sexMale	-12.04734	4.53162	-2.659	0.01009
age:sexMale	1.75141	0.66463	2.635	0.01073
I(age^2):sexMale	-0.05790	0.02390	-2.423	0.01850

Male smokers still have significantly larger lung capacities than female nonsmokers, though neither show significant non-linearity



Can we remove the square term age^2 for smokers?

```
anova(lm(fev ~ (age + I(age^2))*sex, data=smokers),  
      lm(fev ~ age*sex, data=smokers))
```

Analysis of Variance Table

Model 1: fev ~ (age + I(age²)) * sex

Model 2: fev ~ age * sex

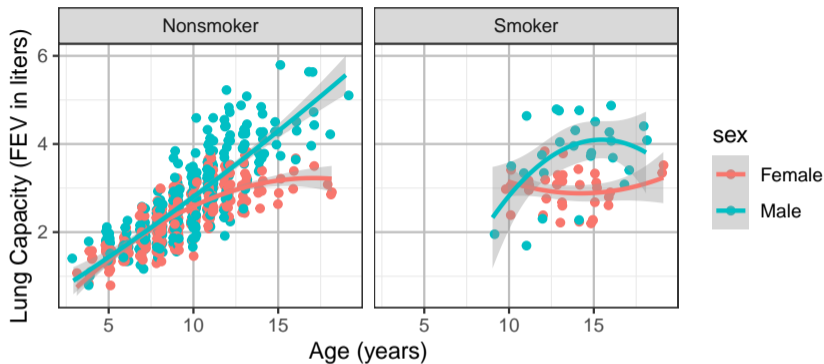
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	59	21.3				
2	61	23.5	-2	-2.21	3.06	0.054

The P-value 0.054 is at the borderline of significance, some moderate but not compelling evidence of non-linearity.

Nonlinear 3-Way Interaction Model

```
summary(lm(log(fev) ~ (age + I(age^2))*sex*smoke, data=fevdata))$coef
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.665172	0.101526	-6.552	1.167e-10
age	0.238379	0.020735	11.496	6.048e-28
I(age^2)	-0.007802	0.001023	-7.625	8.822e-14
sexMale	0.308447	0.142551	2.164	3.085e-02
smokeSmoker	2.831422	0.873835	3.240	1.256e-03
age:sexMale	-0.073525	0.028679	-2.564	1.058e-02
I(age^2):sexMale	0.004883	0.001393	3.506	4.864e-04
age:smokeSmoker	-0.395307	0.127614	-3.098	2.036e-03
I(age^2):smokeSmoker	0.013287	0.004604	2.886	4.031e-03
sexMale:smokeSmoker	-4.350760	1.413287	-3.078	2.169e-03
age:sexMale:smokeSmoker	0.651102	0.208206	3.127	1.845e-03
I(age^2):sexMale:smokeSmoker	-0.023863	0.007545	-3.163	1.637e-03



Can you explain from the plot why the 3-way interaction term $I(\text{age}^2) : \text{sexMale} : \text{smokeSmoker}$ is significant?