Applications of Linear Regression

November 25th, 2019

UC DAVIS
UNIVERSITY OF CALIFORNIA

Outline

- When to use linear regression
- Assessing assumptions
- Interpretations
- Splines
- Polynomial regressions

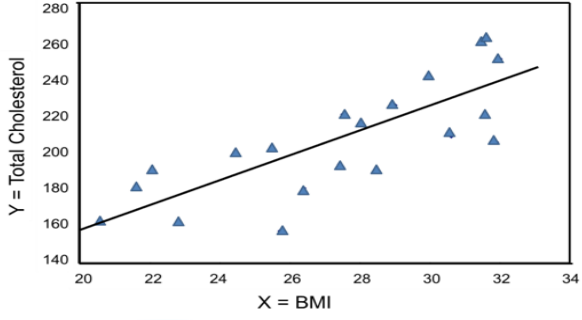
Multiple analyses for same data

Research question: Is body mass index (BMI) associated with high total cholesterol level?

Option 1:
- Convert both to categorical variables
- Contingency tables, chi-squared test and/or Fisher's Exact Test

Option 2:
- Keep both as continuous variables
- Correlation and/or regression analysis

		High total Cholesterol level (>300)		
		Yes	No	
High BMI (>30)	Yes	56	22	78
	No	12	65	77
		68	87	155



Contingency table chi-squared analysis for continuous variables

Advantages

- Ease of interpretation
- No distributional assumptions
- Can easily stratify by other variables
- Can calculate the odds ratio or the relative risk

Disadvantages

- Sometimes requires arbitrary grouping of continuous variables
- Loss of power and precision
- Cannot adjust for confounders easily

Regression analysis for continuous variables

Advantages

- Maintains continuity of data
- Can model one variable as a function of the other variable (regression analysis)
- More useful when both variables are continuous
- Can adjust for confounders

Disadvantages

- Primarily measures linear relationships
- For parametric methods, requires normality and linearity assumptions to be satisfied for testing hypotheses

Simple linear regression

- Purpose: to model the change in one variable (Y, dependent variable) as the other variable (X, independent variable) changes.

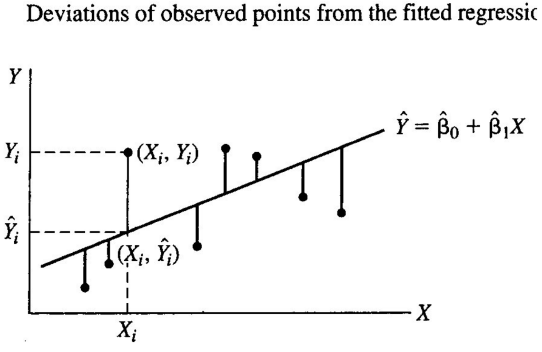
$$Y = \alpha + \beta X$$

- α = intercept (value of Y when X = 0)
- β = slope (change in Y for every 1 unit increase in X)

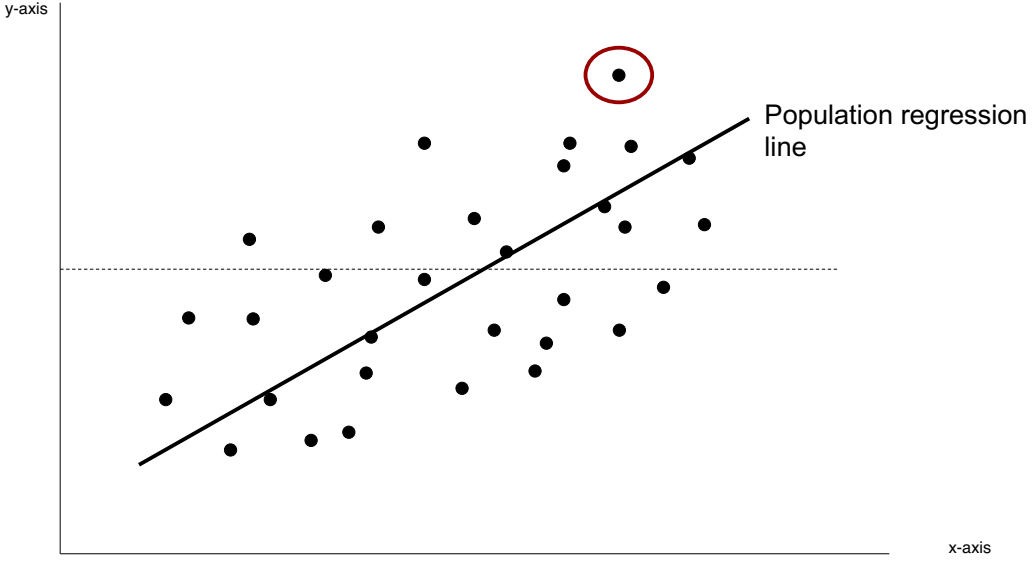
What are we actually doing?

Step 1: plot your data
Step 2: find the best fitting line
**Best fitting line = line that goes straight through your data, minimizing the distance between each point and your line

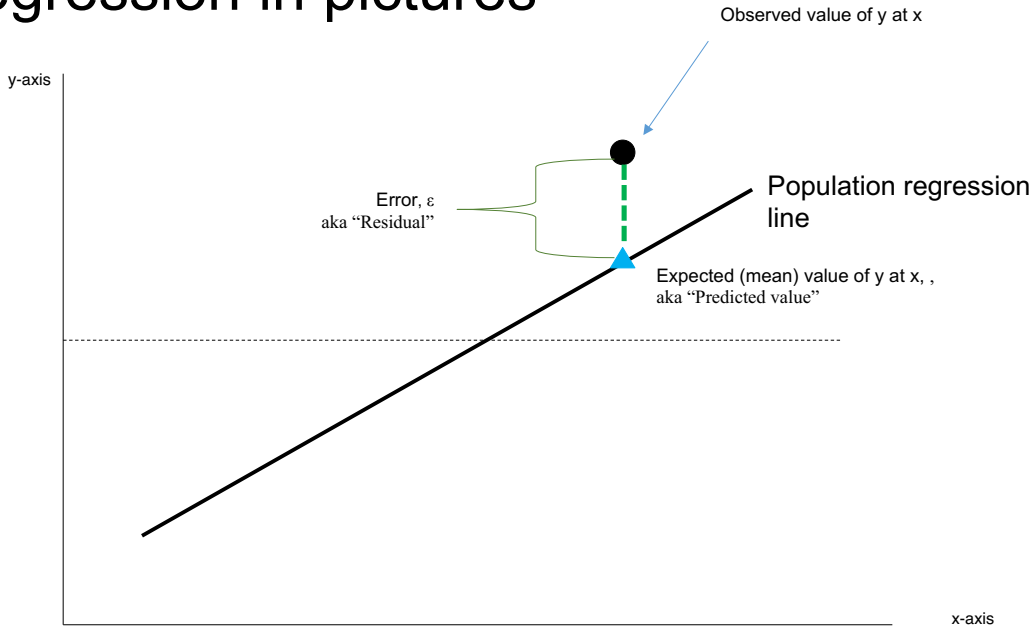
SAS will do this for you in a basic linear regression model



Linear regression in pictures



Linear regression in pictures



Example

```
PROC REG data=mph;
   model DIST=MPH;
run;
```

The REG Procedure
Dependent Variable: dist

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-122.34459	20.15624	-6.07	<.0001
mph	1	6.22708	0.47319	13.16	<.0001

What is the equation for the best fitting line?

A. Y = 6.22708 + 122.34459X
B. Y = 122.34459 + 6.22708X
C. Y = 122.34459 - 6.22708X
D. Y= -122.34459 + 6.22708X

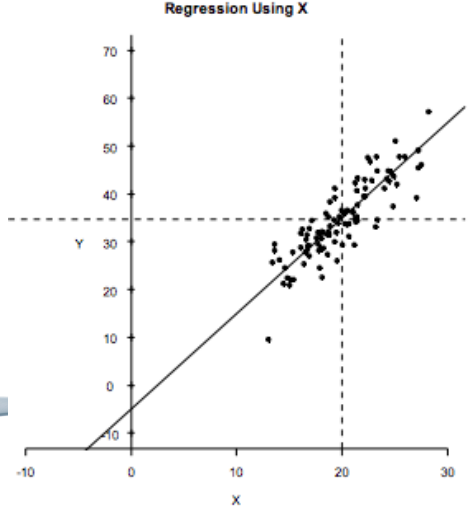
What does this mean?

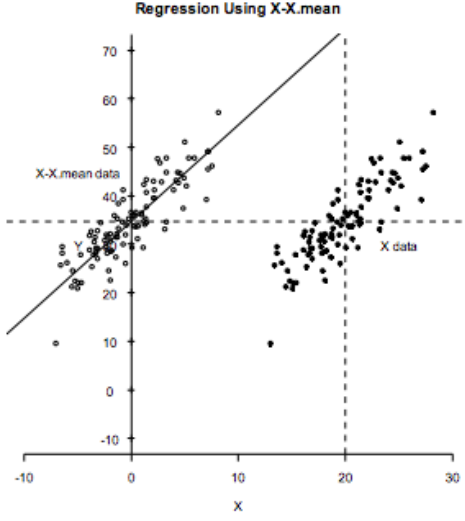
- Equation is Y = -122.34459 + 6.22708X
- Slope (B1): The estimated change in mean Y per increase of 1 unit in X
  - “Estimated mean stopping distance increases by 6.23 feet per 1 MPH increase in speed”
- Intercept (B0 or): The estimated mean when X=0
  - “The estimated mean stopping distance at 0 MPH is -122.3 feet”
  - Any problems with this interpretation?

doesn't change the slope, only changes y intercept and makes it more interpretable

Centering X variables

- All you are doing is “moving” the Y axis over a little, to give interpretation to your Y intercept. Nothing changes except the intercept





UCDAVIS

Centering X variables

If we center MPH on X = 39.37, what is the best interpretation of the new intercept (now 125.1)?

A. When one is traveling 125.1 MPH, the stopping distance is 39.37 feet
B. When one is traveling 0 MPH, the estimated mean stopping distance is 125.1 feet
C. When one is traveling 39.37 MPH, the stopping distance is 125.1 feet
D. When one is traveling 39.37 MPH, the estimated mean stopping distance is 125.1 feet

Centering X variables

Does everyone know how to center a variable on it's mean?

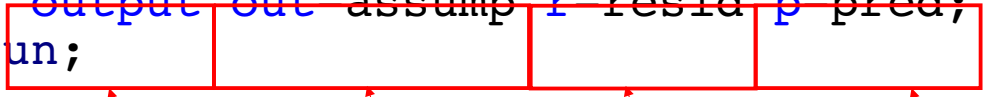
- PROC MEANS, to get the mean
- Create a new variable that subtracts the mean from the original variable
  - If the mean of MPH is 39.37, then:
    mph_cent = mph – 39.37

Assumptions

- "LINE"
  - Linearity: The mean of Y is a straight-line function of X
  - Independence: For any particular value of X, the Y-values are statistically independent of each other.
  - Normality: For any particular value of X, the Y-values have a normal distribution. of the residuals → not just the X or Y variable.
  - Homoscedasticity (aka Equal Variance): For any particular value of X, the Y-values have the same variance.

SAS Code

```
proc reg data=chs;
   model fev = height;
   output out=assump r=resid p=pred;
run;
```



Tells SAS to create a new dataset

Name given to the new dataset

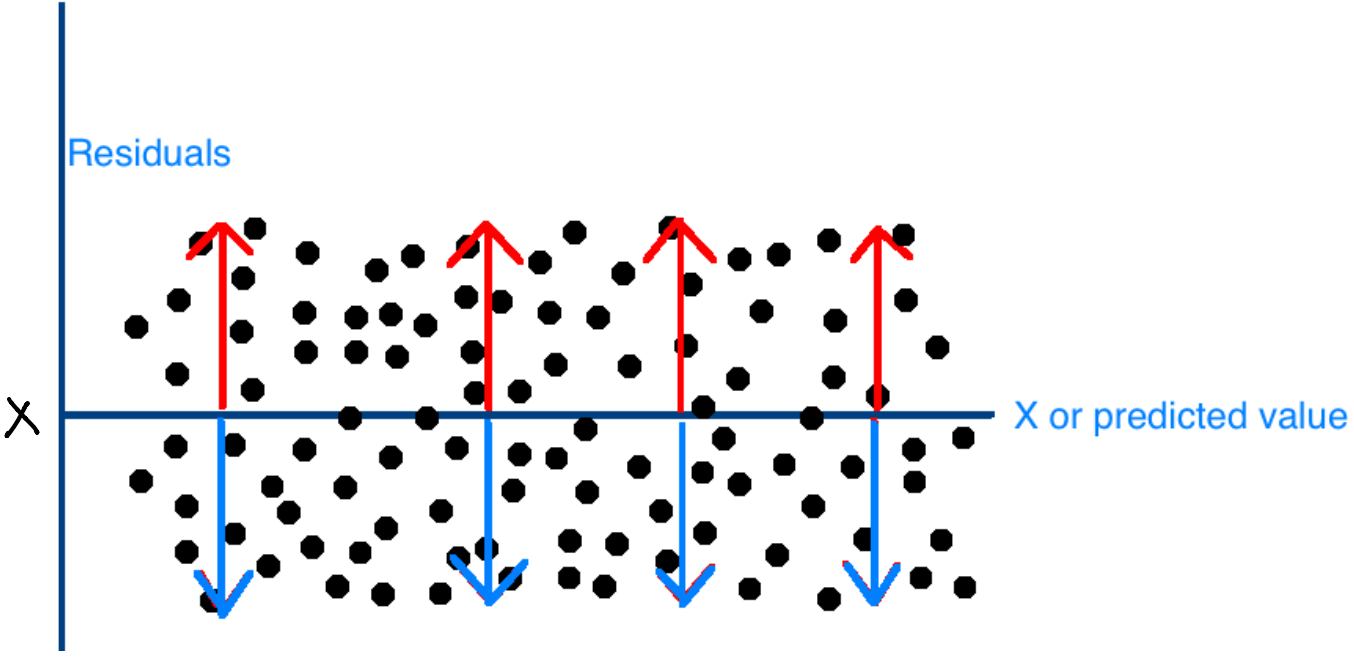
Tells SAS to output the residuals and call them “resid”

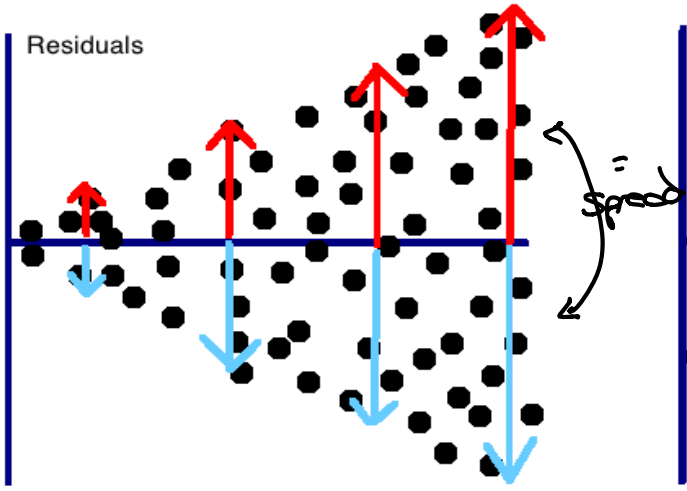
Tells SAS to output the predicted values and call them “pred”

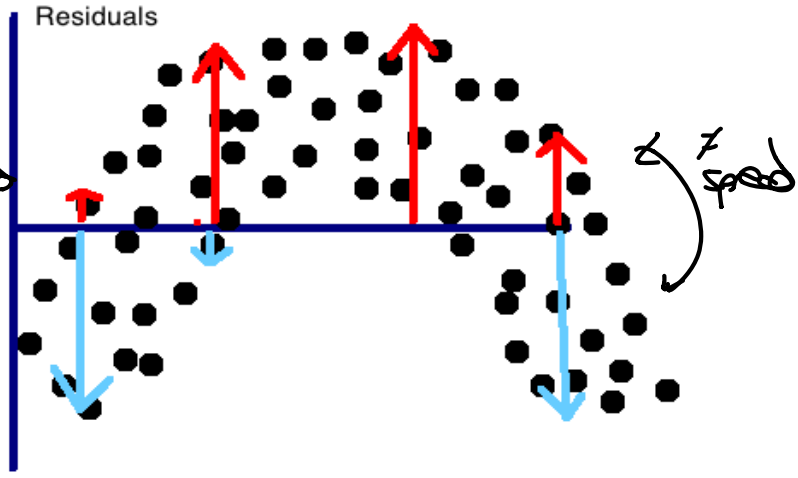
Linear Regression assumption: Linearity

- At any given value of X/predicted value (pick a low, medium, and high value), the spread of the positive residuals is approximately the same as the spread of the negative residuals.
- When plotted, do the points approximately follow a straight line?
  - Violated when data are U-shaped (quadratic) or curved in any way (e.g. exponential, cubic, etc)

Linearity







At a given X/predicted value, the red arrow is equal to the blue arrow.

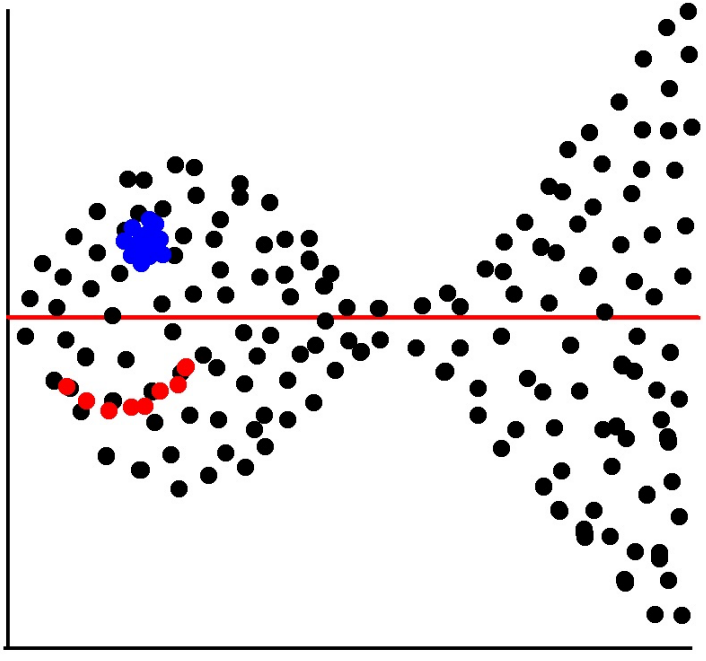
Linearity assumption is met

At a given X/predicted value, the red and blue arrows are different

Linearity assumption is NOT met

Is the linearity assumption met?

A. Yes
B. No
C. Can't tell from the data
D. I have no idea



Linear Regression assumption: Independence

- Typically satisfied when there is one measurement per study unit.
- We want to be sure that none of our observations are correlated with one another
- Violated in many situations, e.g.:
  - Y = Visual acuity (2 eyes per person) 2 obs / person
  - Y = Height, measured multiple times on each subject over time (longitudinally). Longitudinal studies, measures over time
  - Y = Cholesterol, measured on each study subject and their family members (familial clustering). genetic dependence

Linear Regression assumption: Normality

- For any fixed value of X, Y has a normal distribution.
  - As in other situations, normality is required for validity in hypothesis testing and confidence intervals.
  - With large n, the Central Limit Theorem makes inferences “robust” to deviations from this assumption.
    - E.g. If examining the relationship between Age and Height in a study with a sample size of 100,000. For a given age, say 25, height will follow a normal distribution – it will be bell-shaped.

Linear Regression assumption: Normality

- For any fixed value of X, Y has a normal distribution.
  - Run PROC UNIVARIATE on the residuals.
  - It's important to look at all the measures of normality and use them to make an overall decision, rather than just looking at 1 or 2 measures
  - Code:

```
proc univariate data=assump;
  var resid;
run;
```

Normality

20% of 10 = 2

ex. mean = 10 → median = 15 bad

mean = 10 → median = 11 good

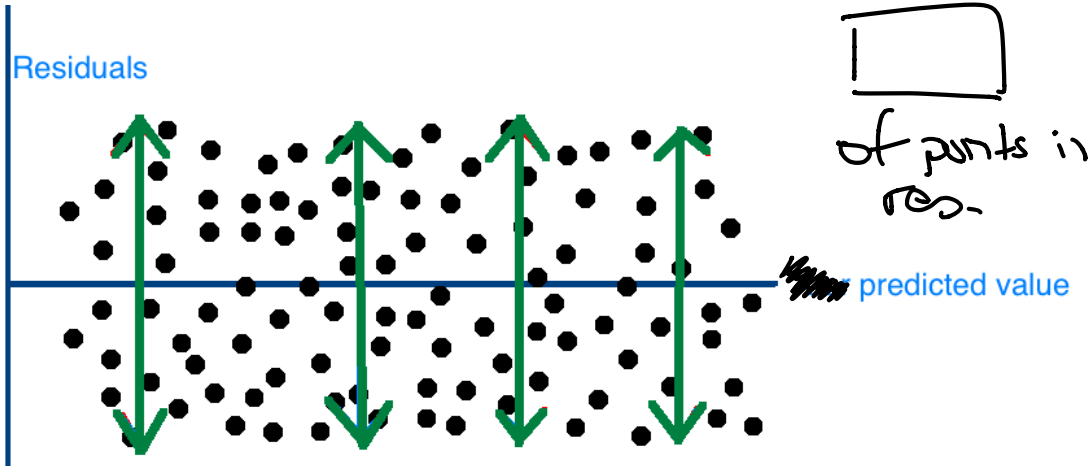
1. Mean ~ median (differ by < 20% of 1 SD)
2. Skewness/Kurtosis < |1|
3. Histogram is bell shaped
4. Normal probability plot (Q-Q plot) follows a straight line
5. Tests for normality are not statistically significant

   **NOTE: when you have a large sample size, the Shapiro-Wilk (and other statistical tests for normality) have a lot of power to detect even very small deviations from normality, so they often appear as statistically significant.

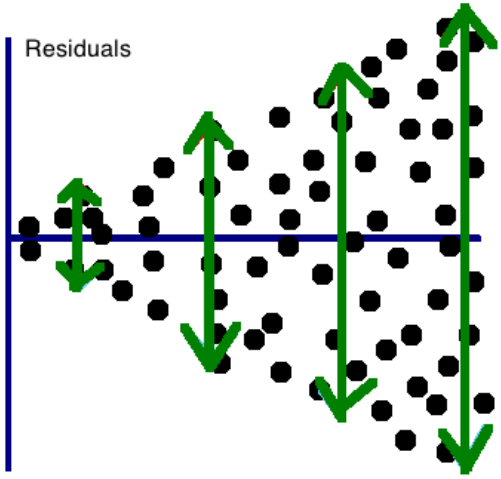
Linear Regression assumption: Equal Variance (Homoscedasticity)

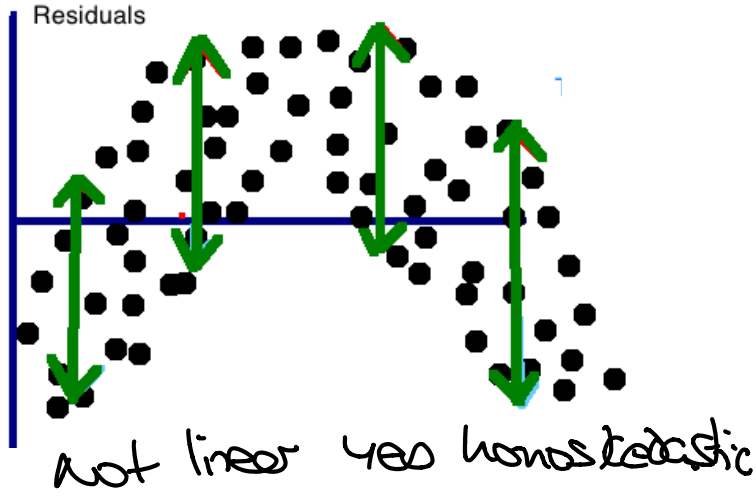
- For any fixed value of X, the variance of Y is a constant; i.e.,

Grey arrows should be the same size



# of points in res.





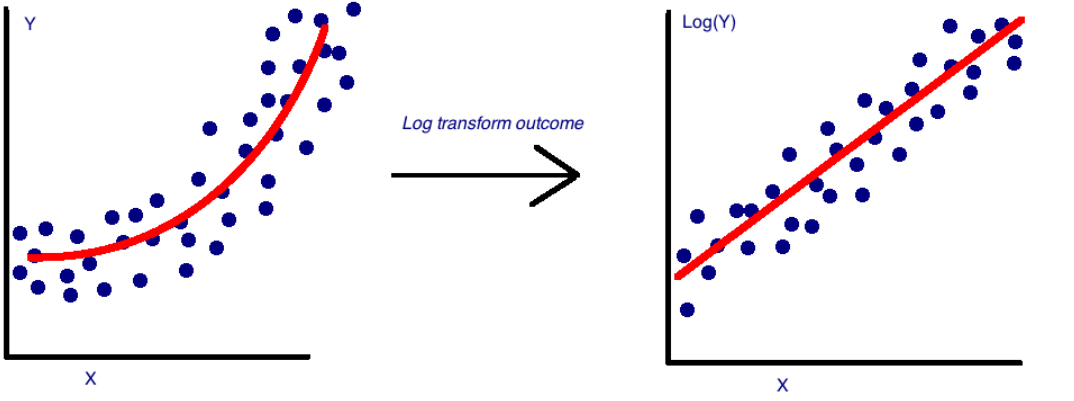
The size of the green arrows are different at different X/predicted values

Homoscedasticity assumption is NOT met

The size of the green arrows at different X/predicted values are approximately the same

Homoscedasticity assumption is met

Log Transformation



Can't run linear regression in this model

Can run linear regression, linear relationship observed

Writing Conclusions

- Things to include in your interpretation of an association:
  - Your X and Y variables (spelled out)
  - Significant/non-significant (include p-value!)
  - What you were measuring (linear association, correlation, etc)
  - Beta, R, or other measure of association
  - The direction of the association
  - Any other important or relevant statistics

Writing Conclusions

"There is [is not] a statistically significant, positive [negative], linear association between x-var and y-var (p = XXXX)." […]

Beta interpretation (continuous X variable):

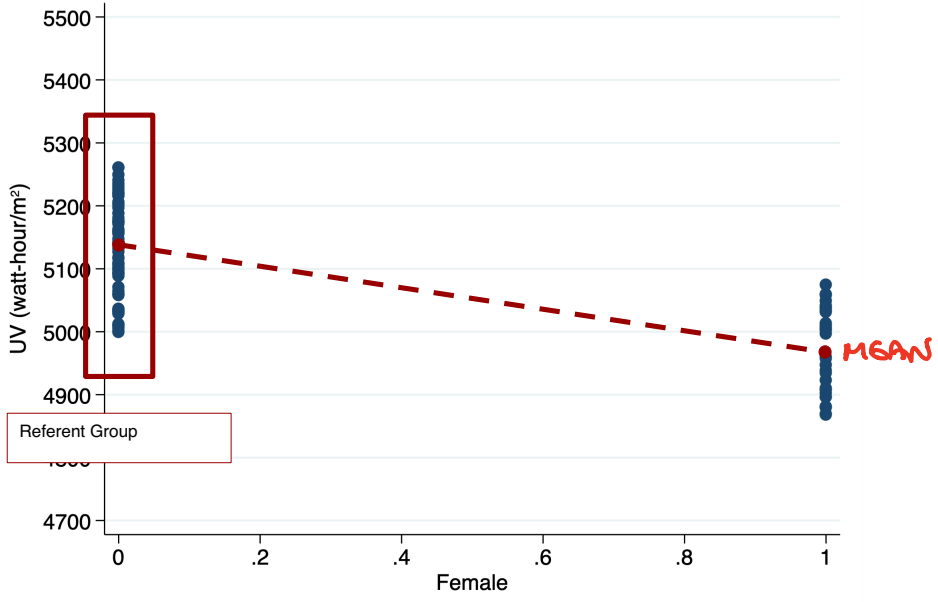
For every 1 unit increase in [x-var], mean y-var increases by XXX.

Beta interpretation (categorical X variable):

On average, the mean [y-var] is higher for [x-var cat1] than for [x-var cat2] (XXX v. ZZZ).

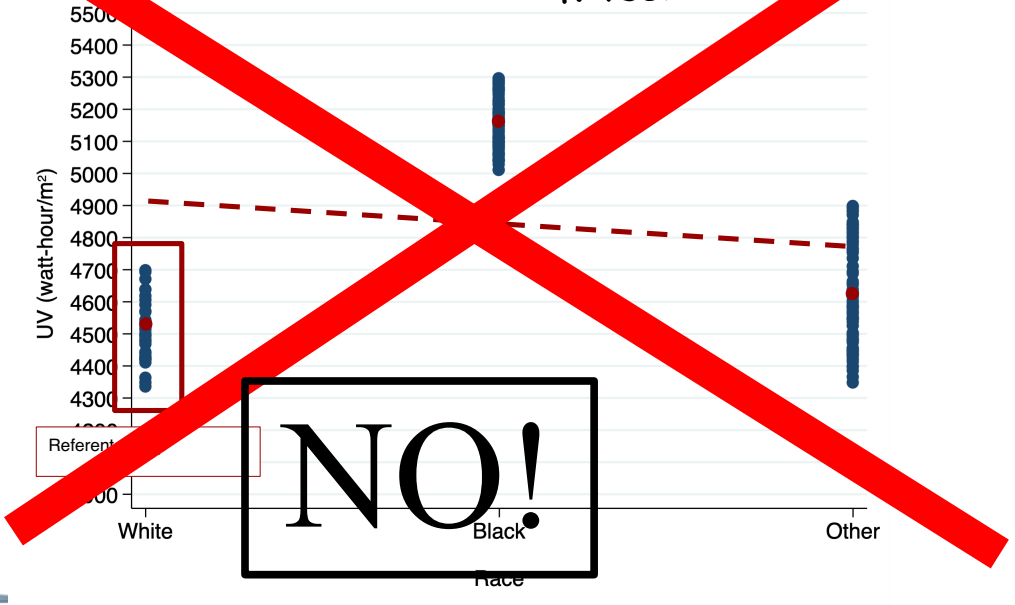
R2 interpretation: XX% of the variation in [y-var] can be explained by [x-var]

Categorical Variables

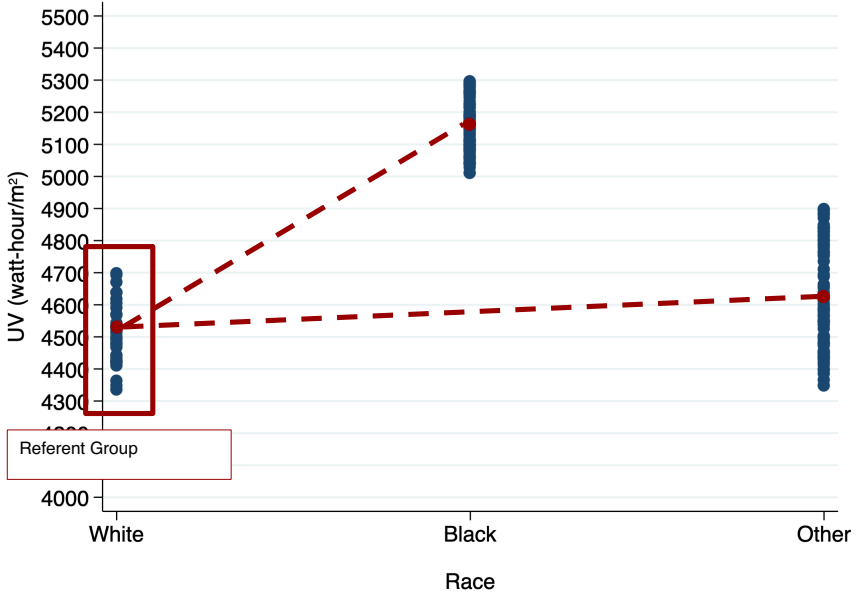


Categorical Variables

SAS looks up: if you don't make a dummy vble, it draws this weird line down the middle.



Categorical Variables



UCDAVIS

Manipulating X

- So far, we have seen that Y has a linear relationship with X
- In assessing residuals we may see a pattern that indicates LINE assumptions are violated
  - Fix linearity and/or homoscedasticity issue with a transformation on Y (typically log)
- Sometimes transforming Y doesn't work, sometimes we want another solution

Manipulating X

- Rather than transforming Y, we can do something to X
- Applying a transformation to X typically involves:
  - Polynomials, X2 or X3
  - Piecewise linear regressions (splines), splitting up X into sections
- Transforming X this way still results in a linear regression
  - Linear means linear in the parameters (βs)

Linear Splines

- Example: Out of pocket medical expenditures vs. length of stay
  - Outcome (Y) is “out of pocket” medical expenditures
  - A researcher tells you most Health Management Organizations (HMOs) will usually pay for the first week of a hospital stay only
  - She expects “out of pocket” expenditures to increase dramatically if length of stay (LOS) was longer than one week
  - What does the X-Y relationship look like?

Linear Splines

We use splines when the relation between X and Y does not fit one straight line

Linear Splines

- We define a new variable that is a function of X that will allow us to check whether the slope is indeed different if length of stay (LOS) is greater than 7 days.
- The idea is to create a variable which we will call (LOS-7)+ that equals

$$\begin{cases} (LOS - 7) & \text{if } LOS > 7 \\ 0 & \text{if } LOS \leq 7 \end{cases}$$

- When you fit the regression, you will include the two terms:
  - LOS
  - (LOS-7)+
- Regression model: $Y = \beta_0 + \beta_1 LOS + \beta_2 (LOS-7)+ + E$
  - Where E is the residual, assumed to be normally distributed with mean zero and variance s2.

Linear Splines

- When LOS ≤ 7: μ(expenditures| LOS ≤ 7) = $\beta_0 + \beta_1 LOS$
- When LOS > 7: μ(expenditures| LOS > 7) = $\beta_0 + \beta_1 LOS + \beta_2 (LOS-7)+$

$$= \beta_0 + \beta_1 LOS + \beta_2 (LOS) - \beta_2 (7)$$
$$= (\beta_0 - \beta_2 (7)) + (\beta_1 + \beta_2) LOS$$

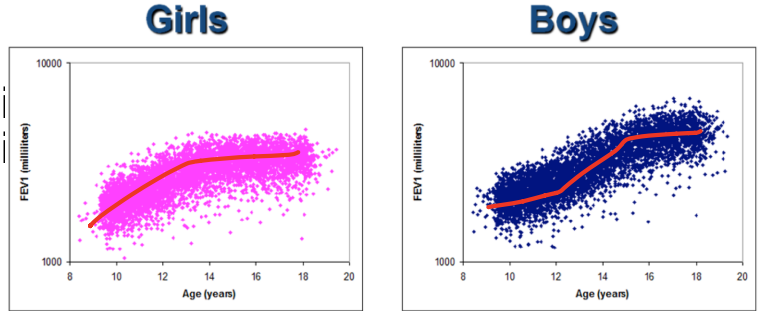
- Conclusion: by including the extra "spline term", we allowed for a different slope relating LOS to expenditures, for LOS values above 7 days. The regression coefficient on the spline term (β_2) quantifies the difference in slopes. We could look at the hypothesis test for H0: $\beta_2 = 0$ to assess whether there was a statistically significant change in slope after 7 days.

Linear Splines

- Example: Lung function vs. age
  - Outcome (Y) is FEV1 in children aged 10 to 18
  - Due to the growth spurt children experience during puberty and the reduction in growth children experience as they reach late teens, we expect a non-linear relationship between FEV1 and age during adolescence.
  - What does the X-Y relationship look like?

Linear Splines

- The relationship between Age (X) and FEV (Y) in girls and boys

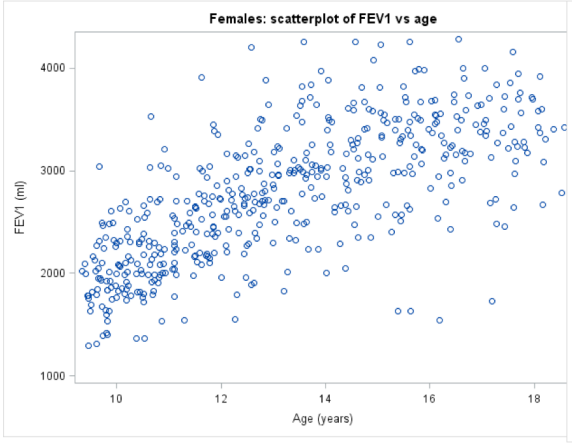


- The relationshi l bend at age 13
- The relationshi nall bends, at ages 12.5 and

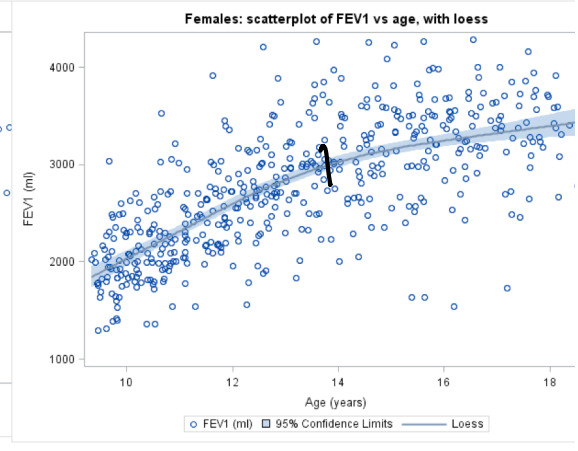
Linear Splines

We see from a loess smooth that the line fitting the girls data is in two pieces with a break at about age 13.

Scatterplot



Scatterplot with smooth



Slope changes

Linear Splines

If we fit a straight line, assessing the residuals shows we violate the assumption of linearity

R2=0.497

↓

not great R^2

Mean of residuals NOT = 0

Homoskedasticity kind of met.

Linearity not met.

Linear Splines

- Let’s fit two lines to these data, one for the relationship between FEV and age for girls younger than 13 and the relationship between FEV and age for girls older than 13
- We are breaking the data up at a knot or inflection point where we see the scatter changes
- We add both pieces (age ≤ 13 and >13)

Model: $\mu(FEV) = \beta_0 + \beta_1 AGEPFT + \beta_2 (AGEPFT-13)+$

So when:

AGEPFT ≤ 13: $\mu(FEV \mid AGEPFT \leq 13) = \beta_0 + \beta_1 AGEPFT$

AGEPFT > 13: $\mu(FEV \mid AGEPFT > 13) = (\beta_0 - 13\beta_2) + (\beta_1 + \beta_2)\, AGEPFT$

β_0: average FEV when AGEPFT is 0 (i.e., at birth)

β_1: average difference in FEV associated with a 1 year increase in age, for girls *younger* than age 13

$(\beta_1 + \beta_2)$: average difference in FEV associated with a 1 year increase in age, for girls *older* than age 13

β_2: *additional difference in average FEV* associated with a 1 year increase in age for girls older than age 13 as compared to girls less younger than age 13

How much it differs between the 2 age groups.

Linear Splines

In SAS, create the age<= 13 and age>13, then fit the model

```
data datf;
  set desktop.datf;
  if AGEPFT ne . then do;
    if AGEPFT > 13 then AGEPFTsp1 = AGEPFT-13;
   else AGEPFTsp1 = 0;
  end;
run;
proc reg data=datf;
  model FEV=AGEPFT AGEPFTsp1 /clb;
run;
```

Linear Splines

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-1005.95459	236.94546	-4.25	<.0001	-1471.49286	-540.41633
AGEPFT	1	303.08186	20.47060	14.81	<.0001	262.86228	343.30144
AGEPFTsp1	1	-202.62763	32.21673	-6.29	<.0001	-265.92541	-139.32986

yes!

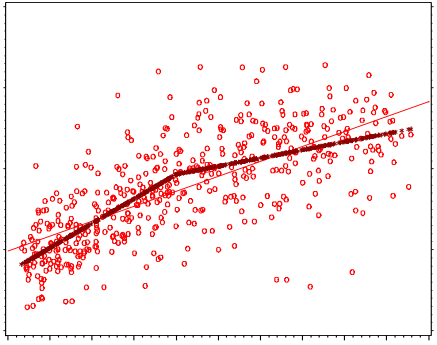
- The relationship between FEV and age for girls over the age of 13 is different from the relationship for girls under the age of 13 (P<.0001).
- Under the age of 13, average FEV increases by 303.1 ml per year (95% CI: 262.9, 343.3).
- Over the age of 13, average FEV increases by only 100 ml (303 – 203) per year.

303 + (-203)

Is the spline var stat. significant? → should we include splines in the model

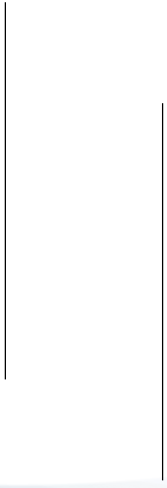
Linear Splines

- The visual results show the fit and residuals are better with spline model



Linear Splines

- What about for boys? We can add more than one break
- The age vs FEV for boys shows breaks in two places, at ages 12.5 and 15.5



Linear Splines

Model: μ(FEV) = β0 + β1AGEPFT+ β2(AGEPFT-12.5)+ + β3(AGEPFT-15.5)+

SAS code to create the two spline variables and fit the model:

```
data datm;
  set desktop.datm;
  if AGEPFT ne . then do;
    if AGEPFT > 12.5 then AGEPFTsp1 = AGEPFT-12.5;
    else AGEPFTsp1 = 0;
    if AGEPFT > 15.5 then AGEPFTsp2 = AGEPFT-15.5;
    else AGEPFTsp2 = 0;
  end;
run;
proc reg data=datm;
  model FEV=AGEPFT AGEPFTsp1 AGEPFTsp2 /clb;
run;
```

Linear Splines

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-96.63172	376.09702	-0.26	0.7973	-835.57144	642.30801
AGEPFT	1	213.22869	33.37585	6.39	<.0001	147.65322	278.80417
AGEPFTsp1	1	281.62524	57.20360	4.92	<.0001	169.23400	394.01648
AGEPFTsp2	1	-309.25985	64.27265	-4.81	<.0001	-435.54007	-182.97963

- For boys under the age of 12.5, average FEV increases by 213.2 ml per year (95% CI: 147.7, 278.8).
- The relationship between FEV and age for boys between the ages of 12.5 and 15.5 differs from the relationship for boys less than 12.5 years old (P<0.0001). For boys between the ages of 12.5 and 15.5 years old, a 1 year increase in age is associated with a 494.8 ml (213.2 + 281.6) increase in average FEV.
- The relationship between FEV and age for boys older than 15.5 years differs from the relationship for boys between the ages of 12.5 and 15.5 (P<0.0001). For boys ages 15.5 years and above, a 1 year increase in age is associated with a 185.5 ml (213.2 + 281.6 – 309.3) increase in average FEV.

Linear Splines

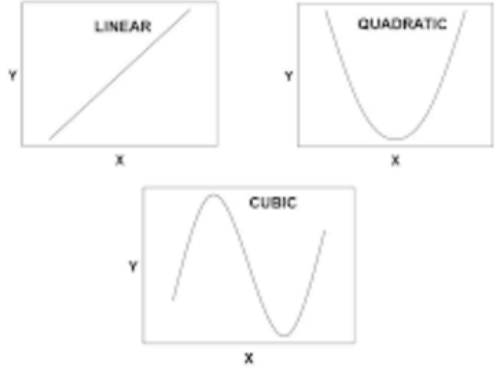
- This method only applies when X is continuous.
- Sometimes called “piecewise linear” because we are fitting separate linear fits and piecing them together.
- The breakpoints are often called “knots”.
  - More advanced methods use this terminology and often uses smooth curves rather than linear pieces. Cubic regression splines is probably the most used and versatile method.

Polynomial Regression

- The general polynomial regression model has the form:

$$Y = \beta 0 + \beta 1X + \beta 2X2 + \ldots + \beta k\text{-}1\ Xk\text{-}1 + \beta kXk + E$$

- The degree of the polynomial is k:
  - k=1 linear (no bends)
  - k=2 quadratic (one bed)
  - k=3 cubic (two bends)



Polynomial Regression

- Hard to interpret polynomials greater than order 3 (cubic)
- With N observations in a dataset, k can at most be N-1 (model identifiability)
- Let's fit a quadratic model on the CHS data for girls

$$FEV = \beta 0 + \beta 1Age + \beta 2Age2 + E$$

```
data datf;
  set datf;
  age2 = AGEPFT**2;
run;
PROC REG data=datf;
  model FEV=AGEPFT age2 /clb;
run;
```

Polynomial Regression

The REG Procedure
Model: MODEL1
Dependent Variable: FEV

Number of Observations Read	500
Number of Observations Used	500

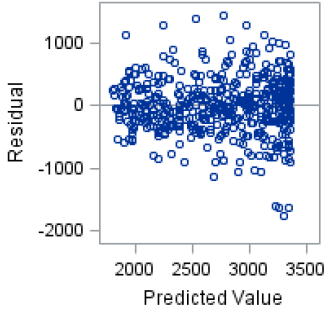
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	114615766	57307883	286.17	<.0001
Error	497	99529308	200260		
Corrected Total	499	214145073			

Root MSE	447.50439	R-Square	0.5352
Dependent Mean	2758.62677	Adj R-Sq	0.5334
Coeff Var	16.22200		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-3598.99146	619.87396	-5.81	<.0001	-4816.88795	-2381.09496
AGEPFT	1	783.59279	93.61581	8.37	<.0001	599.66126	967.52432
age2	1	-22.08649	3.43941	-6.42	<.0001	-28.84407	-15.32892



Conclusions:

- statistically significant quadratic trend in the relationship between FEV and Age in females (coefficient –22.1, p<0.0001).
- residuals from the quadratic model do not exhibit a pattern when plotted against predicted values, indicating that the non-linearity in the relationship between FEV and age has been well-modeled.
- residuals also appear approx. normally distributed (check with PROC UNIVARIATE).

UCDAVIS

Polynomial Regression

- Issues with polynomial regression
  - Correlation between X and X2, X3, etc.
  - This causes collinearity
- Solution: Center X on its mean, then square, cube, etc

Age and Age^2 are highly correlated.

Polynomial Regression

- Revisit the quadratic model on the CHS data for girls
  $FEV = \beta 0 + \beta 1 Age + \beta 2 Age2 + E$
- Mean age is 13.2, so take age-13.2 and then create a squared version of this
- Run the correlation and re-run the regression

Polynomial Regression

The CORR Procedure

2 Variables: AGEPFT age2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
AGEPFT	500	13.20179	2.50045	6601	9.33060	18.57632
age2	500	180.52688	68.05869	90263	87.06001	345.07958

Pearson Correlation Coefficients, N = 500
Prob > |r| under H0: Rho=0

	AGEPFT	age2
AGEPFT	1.00000	0.99633 <.0001
age2	0.99633 <.0001	1.00000

The CORR Procedure

2 Variables: cage cage2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
cage	500	-2.1283E-8	2.50045	-0.0000106	-3.87119	5.37453
cage2	500	6.23975	6.09287	3120	6.79361E-6	28.88560

Pearson Correlation Coefficients, N = 500
Prob > |r| under H0: Rho=0

	cage	cage2
cage	1.00000	0.29348 <.0001
cage2	0.29348 <.0001	1.00000

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-3598.99146	619.87396	-5.81	<.0001	-4816.88795	-2381.09496
AGEPFT	1	783.59279	93.61581	8.37	<.0001	599.66126	967.52432
age2	1	-22.08649	3.43941	-6.42	<.0001	-28.84407	-15.32892

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	2896.44100	29.34446	98.70	<.0001	2838.78650	2954.09549
cage	1	200.43057	8.38083	23.92	<.0001	183.96434	216.89680
cage2	1	-22.08649	3.43941	-6.42	<.0001	-28.84407	-15.32892

- Correlation is much smaller (0.996 vs 0.293) when you center
- Overall F and R2 are the same as non-centered model
- Squared term the same
- Intercept interpreted as before: estimated mean FEV for female aged 13.2 years is 2896 mL

To interpret: pick 3 points along your line and see the predicted value for those

Polynomial Shapes

- In general, the model $Y = \beta0 + \beta1X + \beta2X2+E$ defines:
  - A parabola facing upward if $\beta2 > 0$
  - A parabola facing downward if $\beta2 < 0$
  - The min Y ($\beta2 > 0$) or max Y ($\beta2 < 0$) occurs at $X = -\beta1 / (2\ \beta2)$
- For example, from our model in females, we have $\beta1 = 783.6$ $\beta2 = -22.1$
- Thus, the estimated age at which peak FEV occurs in females is $-783.6 /[2*(-22.1)] = 17.73$ years of age

Splines vs Polynomials

- Choose that which best satisfies LINE assumptions.
- Want best interpretability for your data (e.g. spline might be best for FEV-age example because there is a clear age when things change).
- Polynomial function might be best for parabolas (quadratic) or S -like functions (cubic).

QUESTIONS:

correlation vs linear regression. different? or parts of the same thing?

Normality: not just x and y but the relationship.