BIOS 6312: Modern Biostatistics Methodology II

Andrew J. Spieker, Ph.D.

Associate Professor of Biostatistics
Vanderbilt University

Set 5: Transformations and splines

Version: 04/26/2025

Transformations and basis expansions

Types of transformations:

- Until now, we have largely been concerned with models that assume linearity between covariates and (mean) outcomes.
  - Note: Even models with interaction terms have stratum-specific linear models embedded within them, as we have seen.
- For reasons we will discuss, it is sometimes of interest to first transform the exposure and/or outcome before fitting a model. We will mostly focus on the following methods:
  - Shifting/scaling.
  - Log-transformation.
  - Basis expansions ($\mathrm{E}[Y|X = x] = f(x)$ for "more complicated" $f(\cdot)$).
- In our initial examples, we will focus on the case of a single exposure.

Table of Contents

Shifting and scaling

Shifting: Re-centering X

- You can center the exposure to have mean $\overline{x} - x_0$ for a known x_0.

$$\mathrm{E}[Y|X = x] = \beta_0 + \beta_1(x - x_0).$$

- How would we interpret coefficients?
  - $\beta_0 = \mathrm{E}[Y|X = x_0]$.
  - $\beta_1 = \mathrm{E}[Y|X = x + 1] - \mathrm{E}[Y|X = x]$ (no change from unshifted model).

Shifting and scaling

Shifting: Re-centering Y

- You can center the outcome to have mean $\overline{y} - y_0$ for a known y_0.

$$\mathrm{E}[Y - y_0 | X = x] = \beta_0 + \beta_1 x.$$

- How would we interpret coefficients?
  - $\beta_0 = \mathrm{E}[Y - y_0 | X = 0] = \mathrm{E}[Y | X = 0] - y_0$.
  - $\beta_1 = \mathrm{E}[Y | X = x + 1] - \mathrm{E}[Y | X = x]$ (no change from unshifted model).

Shifting and scaling

Shifting: Re-centering X and Y

- You can center both the exposure and outcome:

$$\mathrm{E}[Y - y_0 | X = x] = \beta_0 + \beta_1(x - x_0).$$

- How would we interpret coefficients?
  - $\beta_0 = \mathrm{E}[Y - y_0 | X = x_0] = \mathrm{E}[Y | X = x_0] - y_0$.
  - $\beta_1 = \mathrm{E}[Y | X = x + 1] - \mathrm{E}[Y | X = x]$ (no change from unshifted model).
- What happens when you choose $x_0 = \overline{x}$ and $y_0 = \overline{y}$? If you need a hint, look at the next slide. :)

Known intercepts:

- If X^* and Y^* denote the exposure and outcome after having been centered about their means, you can fit the following reduced model:

$$\mathrm{E}[Y^*|X^* = x] = \beta x.$$

- R formula: `y ~ -1 + x` removes the intercept.
- Be careful, though, as there is almost never adequate justification to force an intercept through the origin. Good reasons to remove the intercept include:
  - All variables have been centered about their respective means.
  - A particular basis expansion calls for it (will discuss later).
  - Very strong theoretical knowledge (like, we're talking "$F = ma$" level theoretical knowledge).

Scaling: Re-scaling X

- You can scale the exposure ($X^* = cX$ for a known c):

$$\mathrm{E}[Y|X^* = x^*] \quad = \quad \beta_0 + \beta_1 x^*$$

- How would we interpret coefficients?
  - $\beta_0 = \mathrm{E}[Y|X^* = 0] = \mathrm{E}[Y|X = 0]$ (no change from unscaled model).
  - $\beta_1 = \mathrm{E}[Y|X^* = x^* + 1] - \mathrm{E}[Y|X^* = x^*]$
    $= \mathrm{E}[Y|cX = cx + 1] - \mathrm{E}[Y|cX = cx]$
    $= \mathrm{E}[Y|X = x + 1/c] - \mathrm{E}[Y|X = x]$.
- Useful for converting between units.
- By linearity, $c\beta_1 = \mathrm{E}[Y|X = x + 1] - \mathrm{E}[Y|X = x]$.

Scaling: Re-scaling Y

- You can scale the outcome ($Y^* = cY$ for a known c):

$$\mathrm{E}[Y^*|X = x] \quad = \quad \beta_0 + \beta_1 x$$

- How would we interpret coefficients?
  - $\beta_0 = \mathrm{E}[Y^*|X = 0] = \mathrm{E}[cY|X = 0] = c \times \mathrm{E}[Y|X = 0]$.
  - $\beta_1 = \mathrm{E}[Y^*|X = x + 1] - \mathrm{E}[Y^*|X = x]$
    $= \mathrm{E}[cY|X = x + 1] - \mathrm{E}[cY|X = x]$
    $= c(\mathrm{E}[Y|X = x + 1] - \mathrm{E}[Y|X = x])$.

Shifting and scaling

Variable shifting and scaling:

- Shifting makes it easier to uncover stratum-specific means and stratum-specific associations.
- Scaling is useful for making it easy to interpret coefficients that would be too small or too large on the original scale (i.e., converting age from months to years or vice versa).
- There are instances in which shifting and scaling are actually required for a regression method to work properly (e.g., penalized regression).

Table of Contents

Logarithmic transformations

Log-transformations: Transforming the predictor

- If the exposure is positive-valued, it is possible to log-transform it.

$$\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 \log(x).$$

- $\beta_0 = \mathrm{E}[Y|\log(X) = 0] = \mathrm{E}[Y|X = 1]$.
- $\beta_1 \log(1 + q) = \mathrm{E}[Y|X = (1 + q)x] - \mathrm{E}[Y|X = x]$; $q > 0$.
  - Difference in mean Y between strata differing in X by $(100 \times q)\%$.
  - For instance, $\beta_1 \log(1.5)$ compares the mean Y between one stratum of X and the stratum with a value of X that is 50% higher.

Logarithmic transformations

Log-transformations: Transforming the predictor

- Useful for settings in which you believe the mean Y to possess a linear relationship with multiples of X.
- In my experience, a commonly encountered example is when X is a biological concentration/volume of some sort.
  - Prostate-specific antigen (PSA).
  - Lymphocyte count.
- Such quantities are often right-skewed, though contrary to common practice, skewness alone is *not* a sufficient reason to log-transform a variable.
- Avoid blanket statements!! Nuance is your friend.
  - ~~Log-transform all biological quantities/concentrations.~~
  - ~~It is only useful to log-transform a biological quantity/concentration.~~
  - ~~Log-transform all right-skewed variables.~~

Logarithmic transformations

Log-transformations: Transforming the outcome

- If the outcome is positive-valued, it is possible to log-transform it.

$$\mathrm{E}[\log(Y)|X = x] = \beta_0 + \beta_1 x.$$

- To interpret coefficients, you must understand the geometric mean.

The (sample) geometric mean:

$$\begin{aligned}
\exp\left(\overline{\log(Y)}\right) &= \exp\left(\frac{1}{n_0}\sum_{j=1}^{n_0}\log(Y_j)\right) \\
&= \exp\left(\frac{1}{N_0}\sum_{j=1}^{n_0}\log(Y_j)\right) \\
&= \sqrt[n_0]{\prod_{j=1}^{n_0} Y_j} \\
&= \text{GeoMean}(Y).
\end{aligned}$$

Log-transformations: Transforming the outcome

- If the outcome is positive-valued, it is possible to log-transform it.

$$\mathrm{E}[\log(Y)|X = x] = \beta_0 + \beta_1 x.$$

- $\exp(\beta_0) = \exp(\mathrm{E}[\log(Y)|X = 0])$.
  - Geometric mean Y among stratum $X = 0$.
- $\exp(\beta_1) = \exp(\mathrm{E}[\log(Y)|X = x + 1] - \mathrm{E}[\log(Y)|X = x])$
  $= \exp(\mathrm{E}[\log(Y)|X = x + 1])/\exp(\mathrm{E}[\log(Y)|X = x])$.
  - Geometric mean ratio, comparing strata differing in X by one unit.

Logarithmic transformations

Log-transforming: Transforming the outcome

- Useful for settings in which you are interested in the geometric mean Y and believe it to possess a multiplicative relationship with the exposure, X.
- In my experience, a commonly encountered example is when X is a biological concentration/volume of some sort.
  - Hemagglutination inhibition titer.
- The same qualifications discussed for log-transformation of the predictor apply.
- Right-skewness is again *not* a justifiable reason on its own to log-transform an outcome.

Additional notes:

- Having a model that explains your data well is important.
- At the same time, it is important to be able to articulate the scientific question that you're answering.
  - This is linked to coefficient interpretation.
- Outside of this class, you will need to rely more on intuition and the wisdom of the field you're working in.
  - Developed and acquired over time with patience, practice, and experience.

Exercise: Putting concepts together

- Consider the following model:

$$\mathrm{E}[\log(Y)|X = x] = \beta_0 + \beta_1(\log(x) - 5).$$

- Interpret the following quantities:
  - $\exp(\beta_0)$.
  - 1.6^{β_1}.

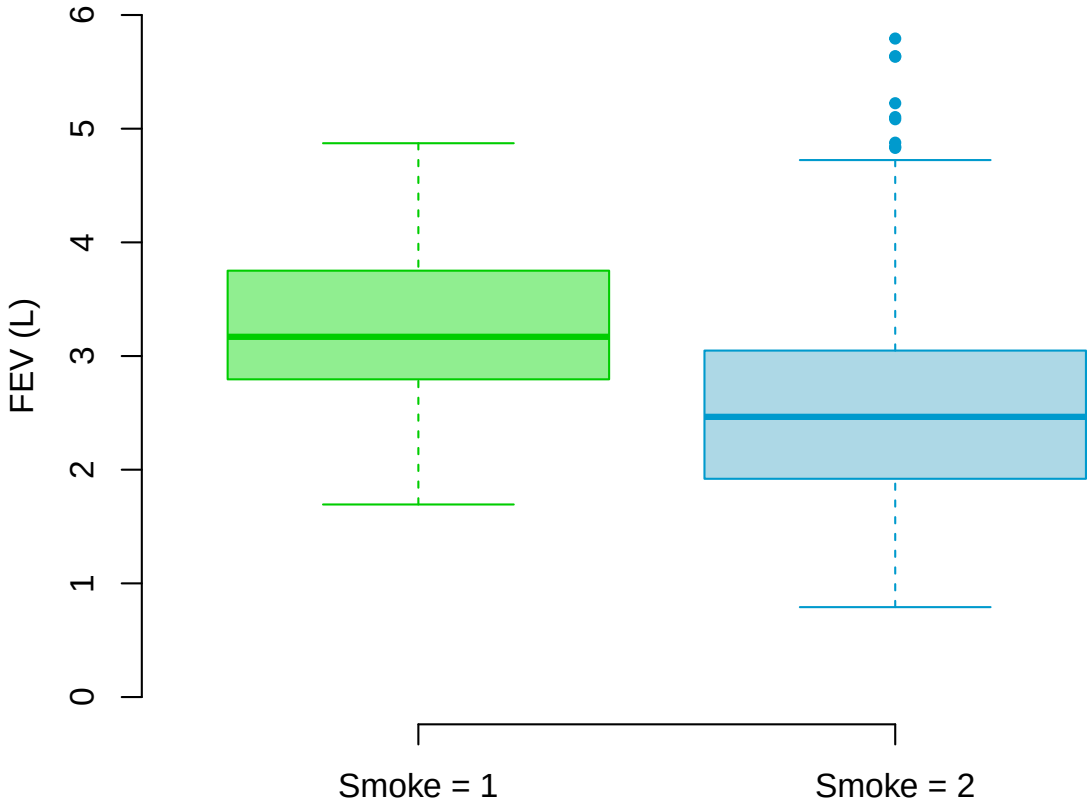
Multiple regression: Analogous results

- Model: $\mathrm{E}[Y|\mathbf{X}=\mathbf{x}] = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \cdots + \beta_K x_K$.
- Coefficient interpretation:
  - β_0: mean Y among $X_1 = 1, X_2 = \cdots = X_K = 0$.
  - $\beta_1 \log(1+q)$: mean difference comparing strata differing in X_1 by $100q\%$ but of the same value for other covariates.
  - β_2: mean difference comparing strata differing in X_2 by one unit but of the same value for other covariates.
- Model: $\mathrm{E}[\log(Y)|\mathbf{X}=\mathbf{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K$.
- Coefficient interpretation:
  - $\exp(\beta_0)$: geometric mean Y among $X_1 = \cdots = X_K = 0$.
  - $\exp(\beta_k)$: geometric mean ratio comparing strata differing in X_k by one unit but of the same value for other covariates.

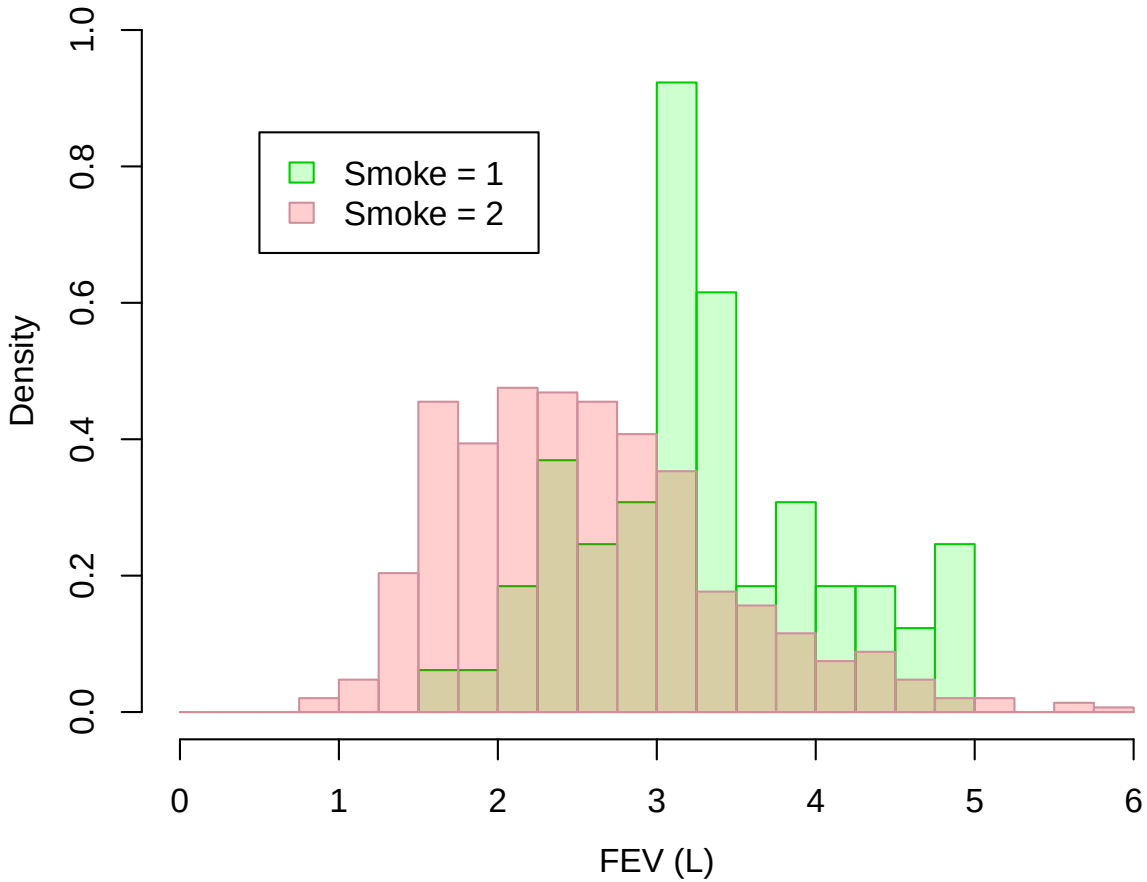
Example: FEV study

- These data came from a cohort study of 654 children that sought to evaluate the association between smoking and lung function in children.
- Lung function was measured by FEV (forced expiratory volume).
  - Amount of volume (L) blown out of lungs in one second.
  - For this reason, you will often see units of L and L/sec used interchangeably for FEV.
- Variables: Age (years), height (in), sex, smoking status, and FEV.
- This data set has a lot of really neat examples; many illustrations shown in this slide set are based on this data set.

Example: Smoking and FEV



Example: Smoking and FEV



Example: Smoking and FEV

- Re-code your smoking variable!
- Variables:
  - X: smoking (0 = No; 1 = Yes)
  - Y: FEV (L)
- Model: $\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x$.
- Coefficient interpretation:
  - β_0: mean FEV among non-smokers.
  - β_1: difference in mean FEV between smokers and non-smokers.

R: Smoking and FEV

```
## Re-code smoking variable properly!
dat$smoke01 <- 2 - dat$smoke

## Fit model
model <- regress("mean", fev ~ smoke01, data = dat)

> model

Call:
  regress(fnctl = "mean", formula = fev ~ smoke01, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-1.775 -0.634 -0.102  0.480  3.227

Coefficients:
               Estimate   Naive SE  Robust SE    95%L    95%H    F stat  df      Pr(>F)
[1] Intercept      2.57     0.0347     0.0351    2.50    2.64   5354.44   1  < 0.00005
[2] smoke01       0.711      0.110     0.0989   0.517   0.905     51.67   1  < 0.00005

Residual standard error: 0.841 on 652 degrees of freedom
Multiple R-squared:  0.0602,  Adjusted R-squared:  0.0588
F-statistic: 51.7 on 1 and 652 DF,  p-value: 1.81e-12
```

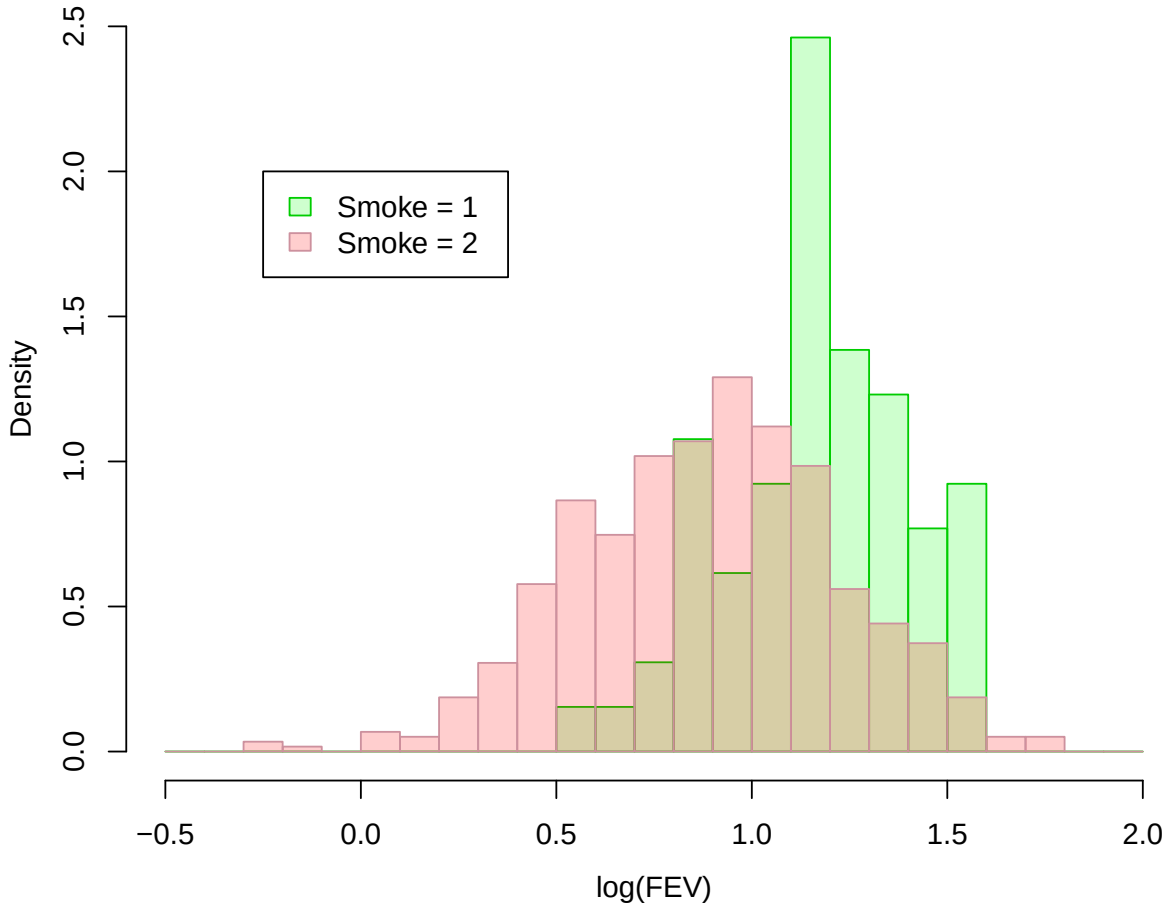
Example: Smoking and FEV

- We estimate the mean FEV among smokers to be 0.711 L higher than that of non-smokers.
  - Of course, this only applies to the study population from which these data were sampled.
- Precision of estimation is quite high.
  - 95% CI: [0.517, 0.905]; $p < 0.001$.
- Can someone give the one-sentence interpretation of the confidence interval?

Example: Smoking and log-transformed FEV

- Suppose, as an example, that we seek to understand the association between smoking and *geometric mean* FEV.
- This can be addressed by log-transforming the outcome.
- Variables:
  - X: smoking (0 = No; 1 = Yes)
  - Y: FEV (L)
- Model: $\mathrm{E}[\log(Y)|X = x] = \beta_0 + \beta_1 x$.
- Coefficient interpretation:
  - $\exp(\beta_0)$: geometric mean FEV among non-smokers.
  - $\exp(\beta_1)$: ratio of geometric mean FEV comparing smokers and non-smokers.

Example: Smoking and log-transformed FEV



R: Smoking and log-transformed FEV

```
## Define log-FEV variable
dat$logfev <- log(dat$fev)

## Fit model
model <- regress("mean", logfev ~ smoke01, data = dat)

> model

Call:
  regress(fnctl = "mean", formula = logfev ~ smoke01, data = dat)

Residuals:
   Min      1Q  Median      3Q     Max
-1.1229 -0.2280  0.0124  0.2178  0.8683

Coefficients:
               Estimate   Naive SE  Robust SE    95%L   95%H.  F stat  df      Pr(>F)
[1] Intercept     0.888     0.0133     0.0137   0.862   0.915  4220.18   1  < 0.00005
[2] smoke01       0.272     0.0423     0.0319   0.209   0.335    72.54   1  < 0.00005

Residual standard error: 0.323 on 652 degrees of freedom
Multiple R-squared:  0.0598,  Adjusted R-squared:  0.0583
F-statistic: 72.5 on 1 and 652 DF,  p-value: <2e-16
```

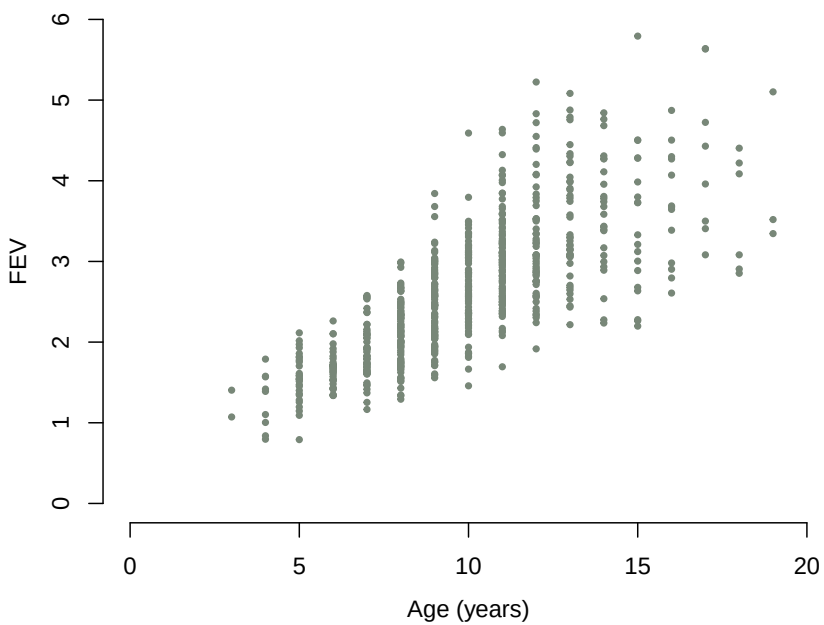
Example: Smoking and log-transformed FEV

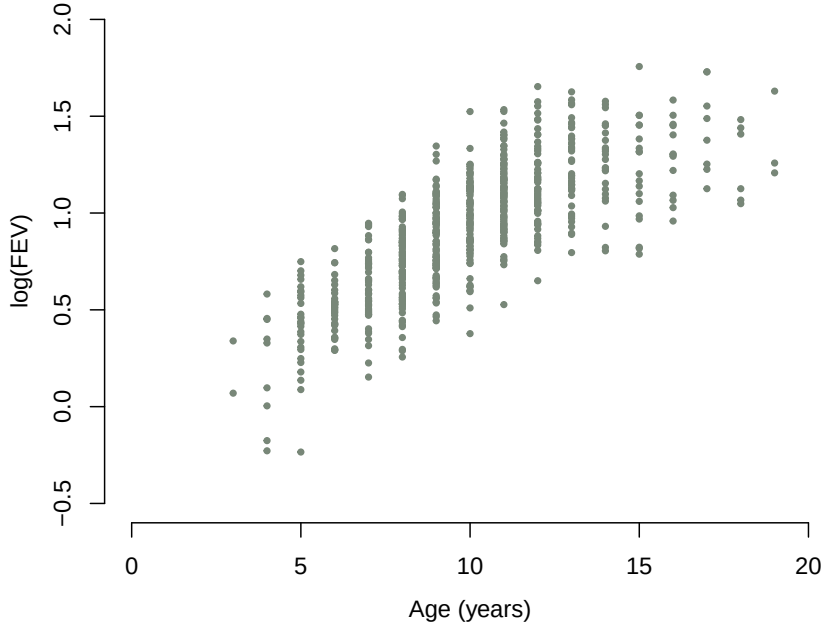
- We estimate the geometric mean FEV among smokers to be 31.3% higher than that of non-smokers.
  - Note that geometric mean ratios are unitless.
  - I obtain this number by exponentiating the coefficient (and then rounding suitably).
  - $\exp(0.2720807) = 1.3127$.
- You can obtain a confidence interval for the geometric mean ratio by expoentiating the endpoints of the confidence interval.
  - $\exp(0.2093539) = 1.2329$.
  - $\exp(0.3348074) = 1.3977$.
  - Based on a 95% confidence interval, our estimate of the geometric mean ratio would not be considered unusual if in truth the geometric mean were between 23.3% and 39.8% higher among smokers.

Example: Smoking and log-transformed FEV

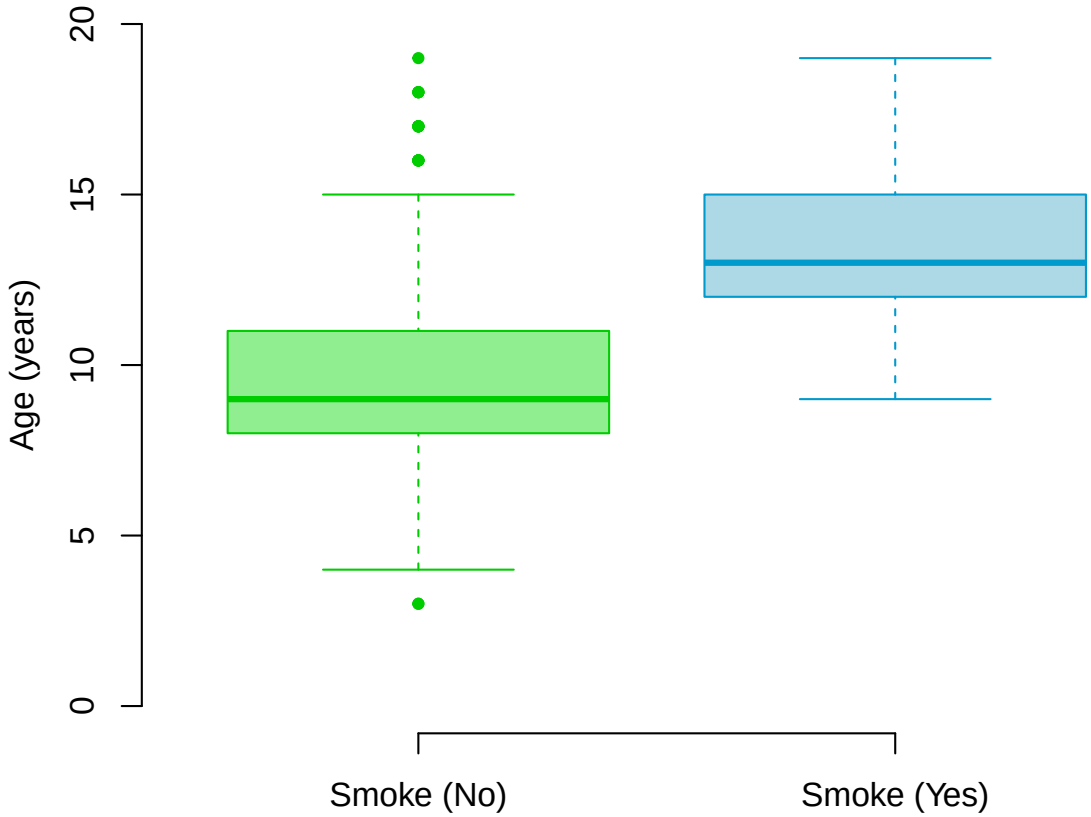
- What might explain the counter-intuitive finding that smoking is associated with greater lung function?
- Age may be serving as a possible confounder.
- This can be evidenced descriptively (although statistical/visual evidence is neither necessary nor sufficient to support a decision to adjust for a potential confounder).

Investigation: Age and FEV





Investigation: Age and smoking



Example: Smoking and log-transformed FEV

- Sticking with a geometric mean ratio as a target endpoint, let us run a model that adjusts for age.
- Variables:
  - X: smoking (0 = No; 1 = Yes)
  - Z: age (years)
  - Y: FEV (L)
- Model: $\mathrm{E}[\log(Y)|X = x, Z = z] = \beta_0 + \beta_1 x + \beta_2 z$.
- Let us interpret the coefficients as a group (for practice):
  - $\exp(\beta_0)$
  - $\exp(\beta_1)$

R: Smoking and log-transformed FEV (shortcut)

```
model <- regress("geometric mean", fev ~ smoke01 + age, data = dat)

> model

Call:
  regress(fnctl = "geometric mean", formula = fev ~ smoke01 + age,
          data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-0.711 -0.135  0.001  0.149  0.603

Coefficients:

  Raw Model:
                 Estimate  Naive SE  Robust SE    F stat  df      Pr(>F)
[1] Intercept      0.0229    0.0304     0.0339      0.46   1      0.4989
[2] smoke01       -0.0899    0.0301     0.0375      5.77   1      0.0166
[3] age            0.0908   0.00305    0.00353    662.75   1  < 0.00005

  Transformed Model:
                  e(Est)    e(95%L)   e(95%H)      F stat  df      Pr(>F)
[1] Intercept       1.02      0.957      1.09        0.46   1      0.4989
[2] smoke01        0.914      0.849     0.984        5.77   1      0.0166
[3] age             1.10       1.09      1.10      662.75   1  < 0.00005

Residual standard error: 0.211 on 651 degrees of freedom
Multiple R-squared:  0.601, Adjusted R-squared:   0.6
F-statistic:  361 on 2 and 651 DF,  p-value: <2e-16
```

Example: Smoking and log-transformed FEV

- We estimate the geometric mean FEV among smokers to be 8.60% lower than that of non-smokers of the same age.
  - I obtain this number by subtracting the GMR from one.
  - $1 - 0.9139978 = 0.08600$.
- Don't forget to *reverse* the order of the confidence interval endpoints.
  - $1 - 0.8491955 = 0.150805$
  - $1 - 0.9837452 = 0.016255$
  - Based on a 95% confidence interval, this estimate would not be considered unusual if in truth the geometric mean were between 1.63% and 15.1% lower.

Further thoughts:

- Recurring pattern in interpretation of regression parameters. Proper interpretation of a coefficient (particularly a non-intercept) should ordinarily check the following boxes...
  1. Make clear which strata are being compared.
     - ⋆ Strata differing in X by one unit?
     - ⋆ Strata differing in X by 5%?
  2. Make clear what is being compared between them.
     - ⋆ The (arithmetic) mean?
     - ⋆ The geometric mean?
  3. Make clear how they're being compared.
     - ⋆ A difference?
     - ⋆ A ratio?

Table of Contents

Handling nonlinearity: Motivation

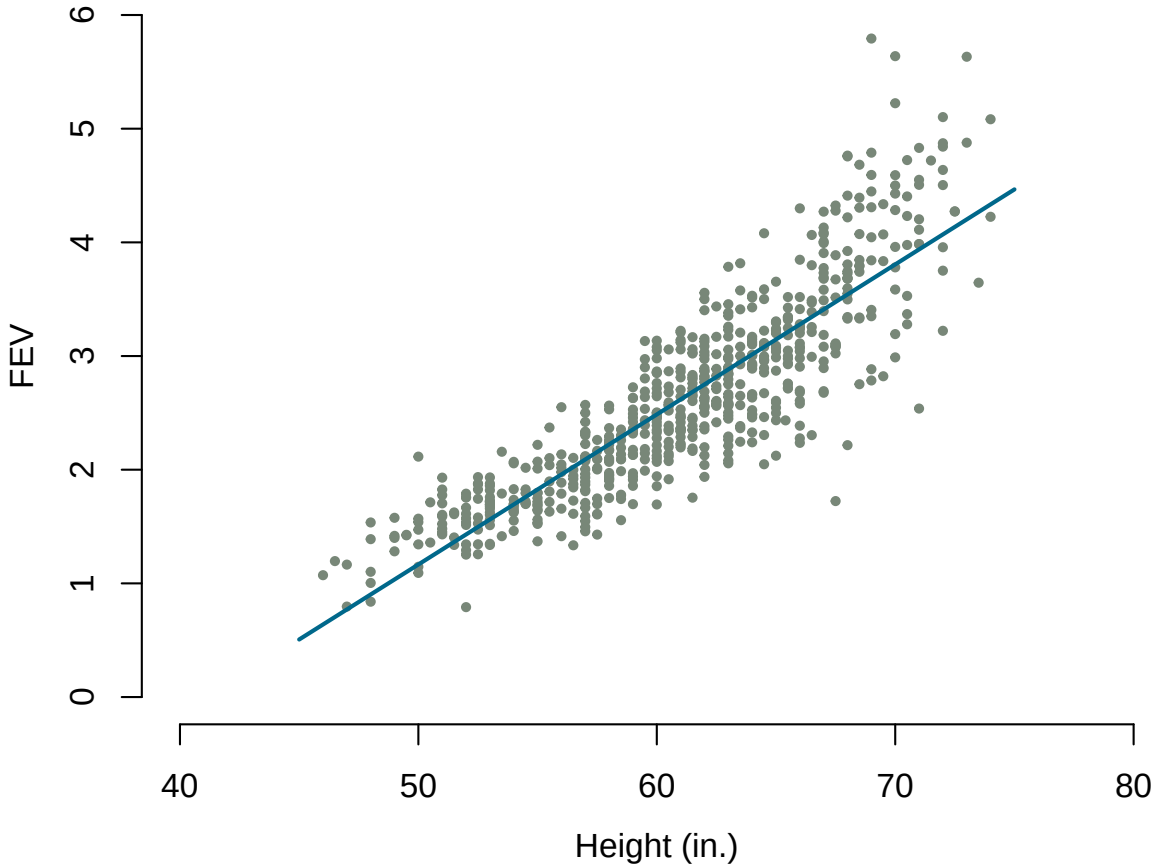
- When you're dealing a continuous predictor, it's almost certain that a linear model is not going to perfectly describe the relationship between X and (mean) Y.
  - Log-transformation of exposures and/or outcomes can be thought of as a specific way to address nonlinearity.
- Discussion point: The degree to which we care about specifying a model that captures the relationship very well depends to some extent upon the purpose for which the model is being used.

Nonlinearity

Example: FEV study

- Variables:
  - X: height (inches)
  - Y: FEV (L)
- Model: $\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x$.

FEV: Association between height and FEV



Nonlinearity

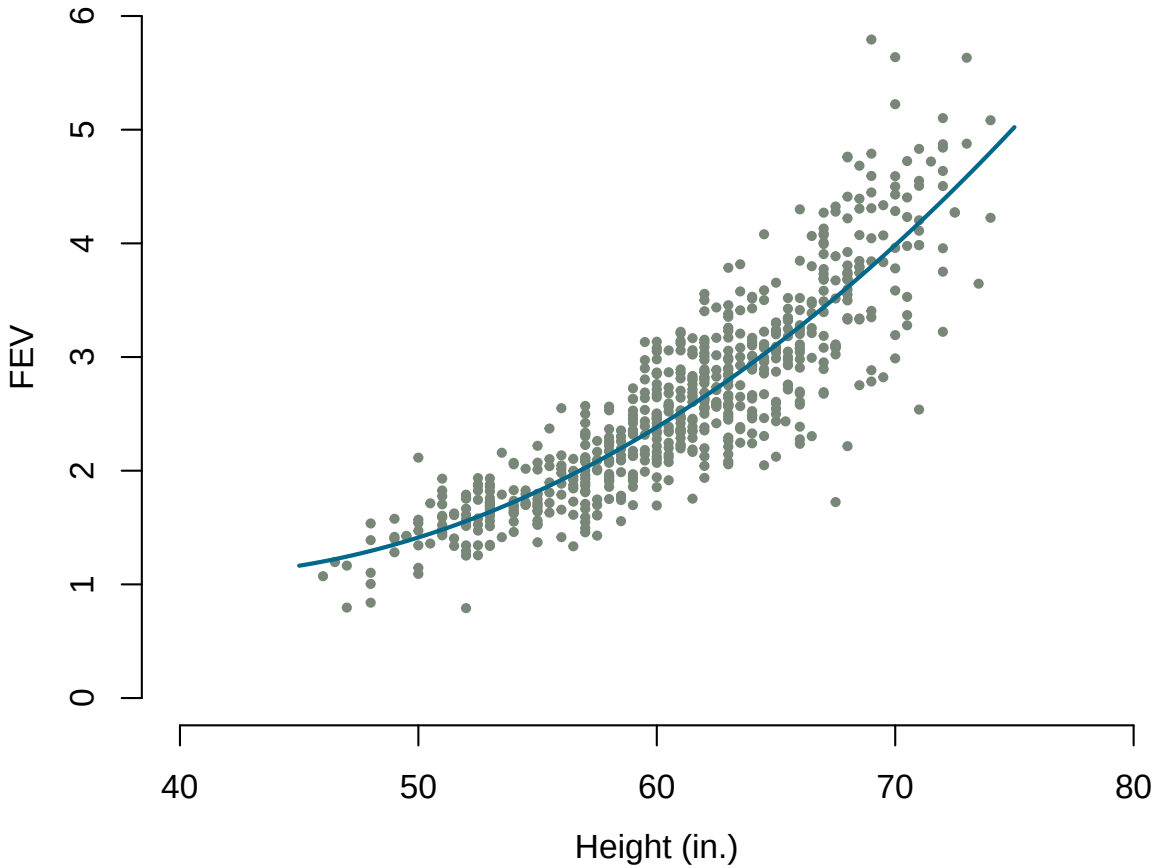
FEV: Nonlinearity

- There seems to be evidence of a curvilinear relationship. However, if all you're looking to do is assess whether there is an overall/first-order association between height an mean FEV, I'll grant you that this model pretty much has you covered.
- Therefore, if you're only thinking about first-order trends, the utility of a model that better describes the relationship between variables is not immediately obvious in this simple example.
- Thinking beyond this simple example, I want to make the broader benefit of accounting for nonlinearity clear.
  - A linear model applied to a curvilinear relationship will not properly estimate stratum-specific means.
  - Suppose Z is a confounder for the relationship between X and Y. Linear adjustment for Z is insufficient if the relationship between Z and Y (given X) is nonlinear.

FEV: Quadratic model

- One solution is to use a quadratic model.
- Variables:
  - X: height (inches)
  - Y: FEV (L)
- Model: $\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x + \beta_2 x^2$.
- Without even looking at the scatter plot, I know this model will provide a "better fit" to the data than a simple linear model.
- I'm appealing to a general principle that a model in which additional parameters are included from a starting model (i.e., nesting) cannot provide a worse fit (measured by, e.g., R^2).

Example: Height and FEV



Example: Height and FEV

- Variables:
  - X: height (inches)
  - Y: FEV (L)
- Model: $\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x + \beta_2 x^2$.
- The purpose of the quadratic model is to provide a better fit; it doesn't necessarily have scientific justification.
  - $\beta_0 = \mathrm{E}[Y|X = 0]$ (meaningless).
  - $\beta_1 = \ldots$? Here, the difference in mean FEV comparing strata differing in height by one unit is *not constant*. Also, it doesn't make sense to speak about the difference in mean FEV comparing strata differing in height but of the same squared height (in fact, this would only occur when comparing $X = -0.5$ to $X = 0.5$).

Nonlinearity

Note: Interpreting β_1 from a quadratic model

- Model: $\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x + \beta_2 x^2$.
- Interpretation of β_1 is *possible* but a little tough.

$$\frac{\partial \mathrm{E}[Y|X = x]}{\partial x} = \beta_1 + 2\beta_2 x;$$

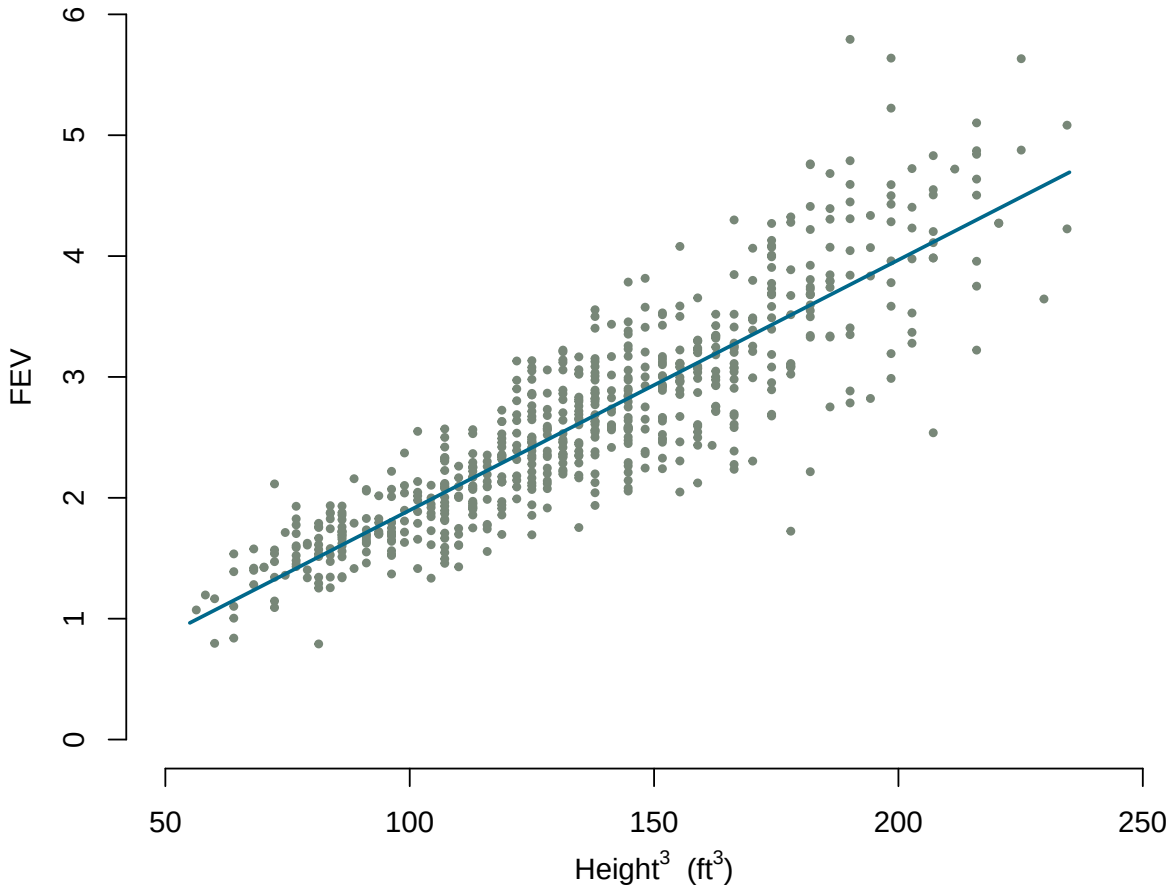
$$\left.\frac{\partial \mathrm{E}[Y|X = x]}{\partial x}\right|_{x=0} = \beta_1.$$

- β_1: rate of change of $\mathrm{E}[Y|X = x]$ with respect to x at $x = 0$.
- Under linearity, $\beta_2 = 0$ and β_1 denotes rate of change across all x.
- To test whether X is associated with mean Y, test $H_0 : \beta_1 = \beta_2 = 0$.
- Other principles surrounding adjustment, shifting, scaling, and log-transformations generalize. For instance, if X were centered to have mean 5, β_1 would denote the rate of change at $x = 5$.

FEV: Transforming height

- The same general principles hold for higher order polynomials. However, there is one particular transformation that may have some scientific justification.
- Consider cubing height (first scaling to *feet* in order provide more manageable numbers).
- Variables:
  - X: height (inches)
  - Y: FEV (L)
- Model: $\mathrm{E}[Y|X = x] = \beta_0 + \beta_1(x/12)^3$.
- Other than simply "to provide a better fit," why might this particular transformation have some theoretical justification?

FEV: Association between height and FEV



Nonlinearity

Basic transformations: Additional thoughts

- There are all sorts of transformations that are theoretically possible.
  - Square-root transformations.
  - Power transforms.
  - Inverse transforms.
- When selecting a transformation, all of the questions must be critically weighed against each other in a given setting:
  - What is my scientific question?
  - How easy will it be to answer my scientific question with this model?
  - What are the assumptions required by this model?
  - Does my ability to answer the scientific question rest upon selecting the right transformation?
- Instead of belaboring the point, we're going to now move to a more general (modern) framework that accommodates various forms of nonlinearity.

Table of Contents

Basis expansions

Modeling a function:

- Goal: Estimate $\mathrm{E}[Y|X = x] = f(x)$.
- Invariably, we need to specify a class of functions to which f belongs.
  - Simple linear regression: $f(x) = \beta_0 + \beta_1 x$.
  - Log-transformation of the exposure: $f(x) = \beta_0 + \beta_1 \log(x)$.
  - Cubic-transformation of the exposure: $f(x) = \beta_0 + \beta_1 x^3$.
- Consider instead the following class of representations for $f(x)$:

$$f(x) = \sum_{p=0}^{P-1} \beta_p h_p(x) \quad \text{[additive model]}.$$

- This is called a *basis expansion*; we have examples of this already!
  - $(P-1)$-degree polynomial regression: $h_p(x) = x^{p-1}$.
  - Categorization of continuous x: $h_p(x) = 1(x \in C_p)$; $\{C_p\}$ denotes a set of mutually exclusive and exhaustive sets spanning the range of X.
- The goal is to propose smart choices for $h_p(x)$ for $P > 1$.

Smart choices:

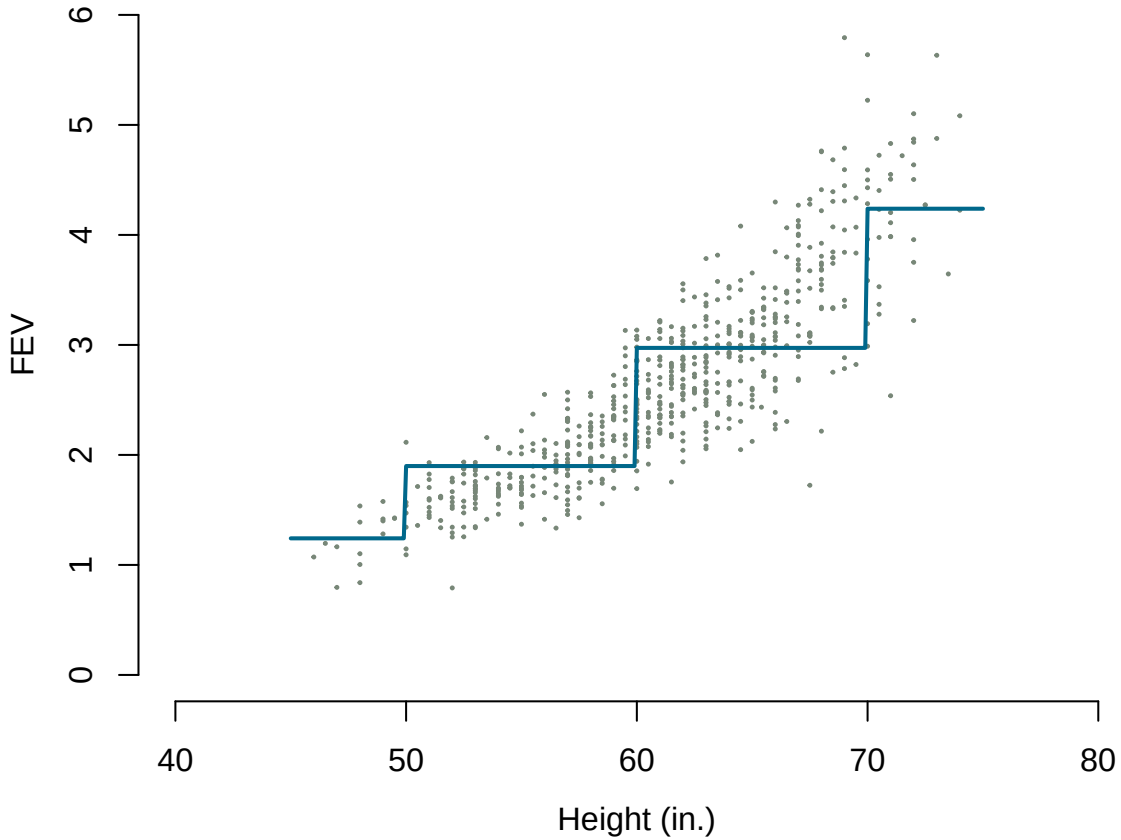
- We've spoken to some degree about the drawbacks of categorizing continuous variables (including predictors).
- Higher-degree polynomials are certainly more flexible. However, they are liable to "wobble around" more than necessary, particularly at the boundaries of the data where there is sparse information.
- The conceptual problem with polynomial interpolation is that it too much information is shared across the levels of X.
  - The local relationship between X and Y at low values of X is used to inform the local relationship between X and Y at high values of X.
- Note: A $(P-1)$-degree polynomial can fit P points perfectly.

Basis expansions

Revisiting categorization: Piecewise constant functions

- One option is to let $f(x)$ be a piecewise constant (steps at ζ_1, ζ_2, ..., which must be specified by the user).
  - User-specified "boundary" points are known as *knots*.
- Basis expansion for piecewise constant function with three knots:
  - $h_0(x) = 1(x < \zeta_1)$.
  - $h_1(x) = 1(\zeta_1 \leq x < \zeta_2)$.
  - $h_2(x) = 1(\zeta_2 \leq x < \zeta_3)$.
  - $h_3(x) = 1(x \geq \zeta_3)$.
- Model: four parameters/four degrees of freedom.
  - This is intuitive as this saturated model is estimating the mean value of Y separately within four distinct strata.
  - This is, in some sense, the "other extreme" of model flexibility in which *no* information is being shared across discrete segments.
  - This is a re-parameterization of the way we would have categorized X in the prior set of notes; they are equivalent models.

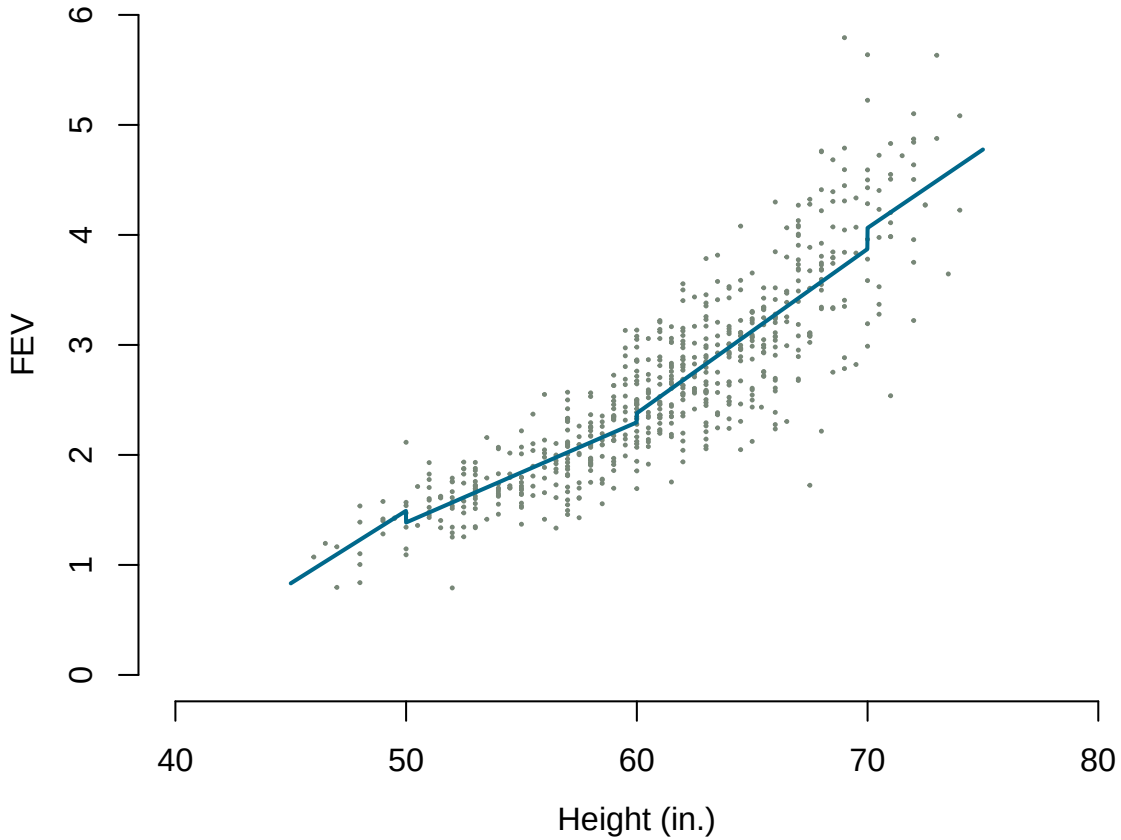
FEV: Piecewise constant function; knots at 50, 60, and 70 in.



Piecewise linear functions:

- Basis expansion for piecewise linear function with three knots:
  - $h_0(x) = 1(x < \zeta_1)$.
  - $h_1(x) = 1(\zeta_1 \leq x < \zeta_2)$.
  - $h_2(x) = 1(\zeta_2 \leq x < \zeta_3)$.
  - $h_3(x) = 1(x \geq \zeta_3)$.
  - $h_4(x) = x \times 1(x < \zeta_1)$.
  - $h_5(x) = x \times 1(\zeta_1 \leq x < \zeta_2)$.
  - $h_6(x) = x \times 1(\zeta_2 \leq x < \zeta_3)$.
  - $h_7(x) = x \times 1(x \geq \zeta_3)$.
- Model: eight parameters/eight degrees of freedom.
  - Why is this intuitive?

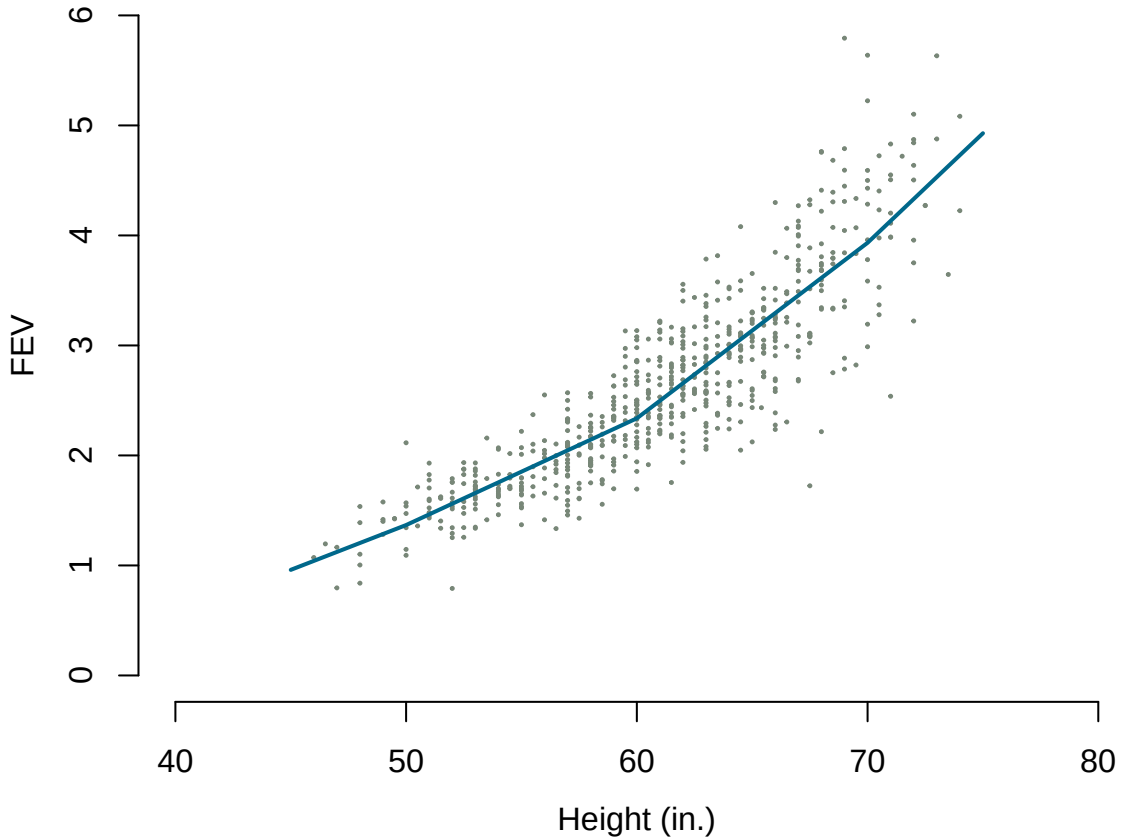
FEV: Piecewise linear; knots at 50, 60, and 70 in.



Continuous piecewise linear functions:

- Basis expansion for continuous piecewise linear function with three knots:
  - $h_0(x) = 1$.
  - $h_1(x) = x$.
  - $h_2(x) = (x - \zeta_1)_+ = \max(0, x - \zeta_1)$.
  - $h_3(x) = (x - \zeta_2)_+ = \max(0, x - \zeta_2)$.
  - $h_4(x) = (x - \zeta_3)_+ = \max(0, x - \zeta_3)$.
- Model: five parameters/five degrees of freedom.
  - Four linear functions (two parameters each); must join at three points.
  - $2 \times 4 - 3 = 5$.

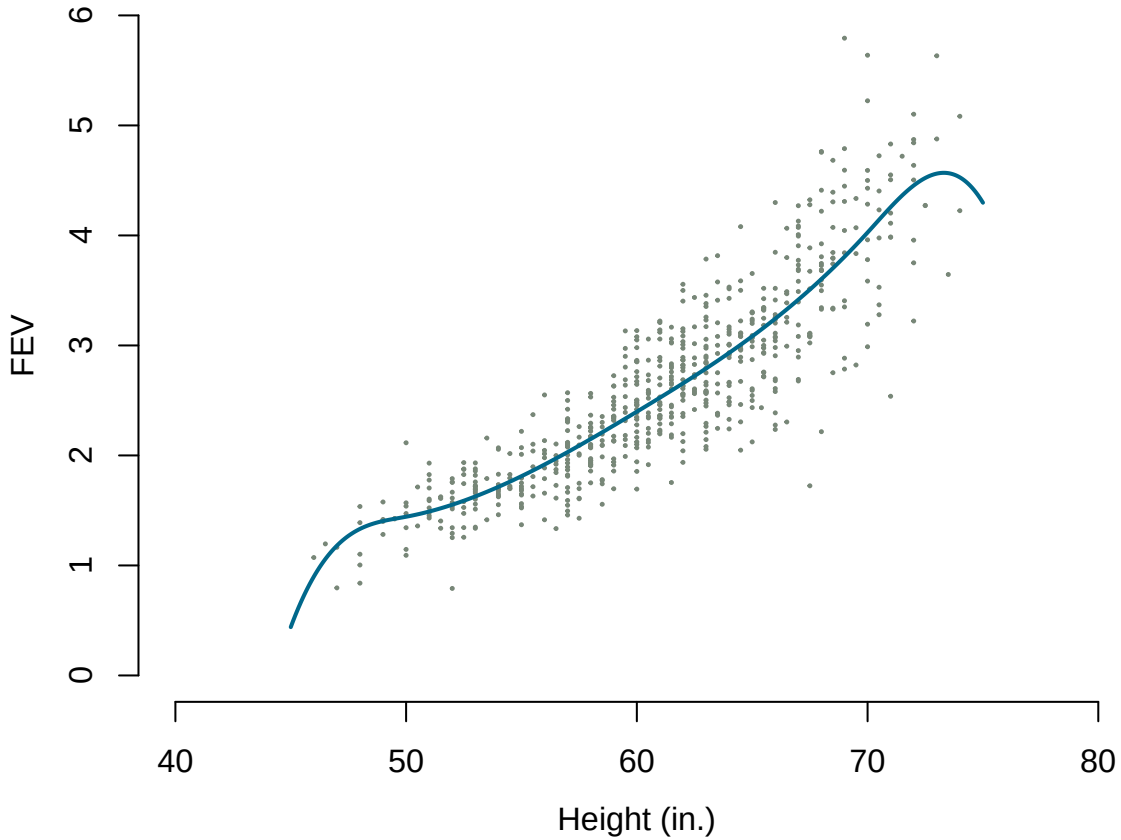
FEV: Continuous piecewise linear; knots at 50, 60, and 70 in.



Piecewise cubic w/ continuous first derivatives:

- Basis expansion for piecewise cubic function having continuous first derivatives function with three knots:
  - $h_0(x) = 1$.
  - $h_1(x) = x$.
  - $h_2(x) = x^2$.
  - $h_3(x) = x^3$.
  - $h_4(x) = (x - \zeta_1)^3_+$.
  - $h_5(x) = (x - \zeta_2)^3_+$.
  - $h_6(x) = (x - \zeta_3)^3_+$.
- Model: seven parameters/seven degrees of freedom.
  - Four cubic functions (four parameters each); three restrictions at three knots (specifically, must be continuous, differentiable, and have continuous first derivative at each knot).
  - $4 \times 4 - 3 \times 3 = 7$.
- This is referred to as a cubic spline.

FEV: Cubic spline; knots at 50, 60, and 70 in.



Natural cubic splines:

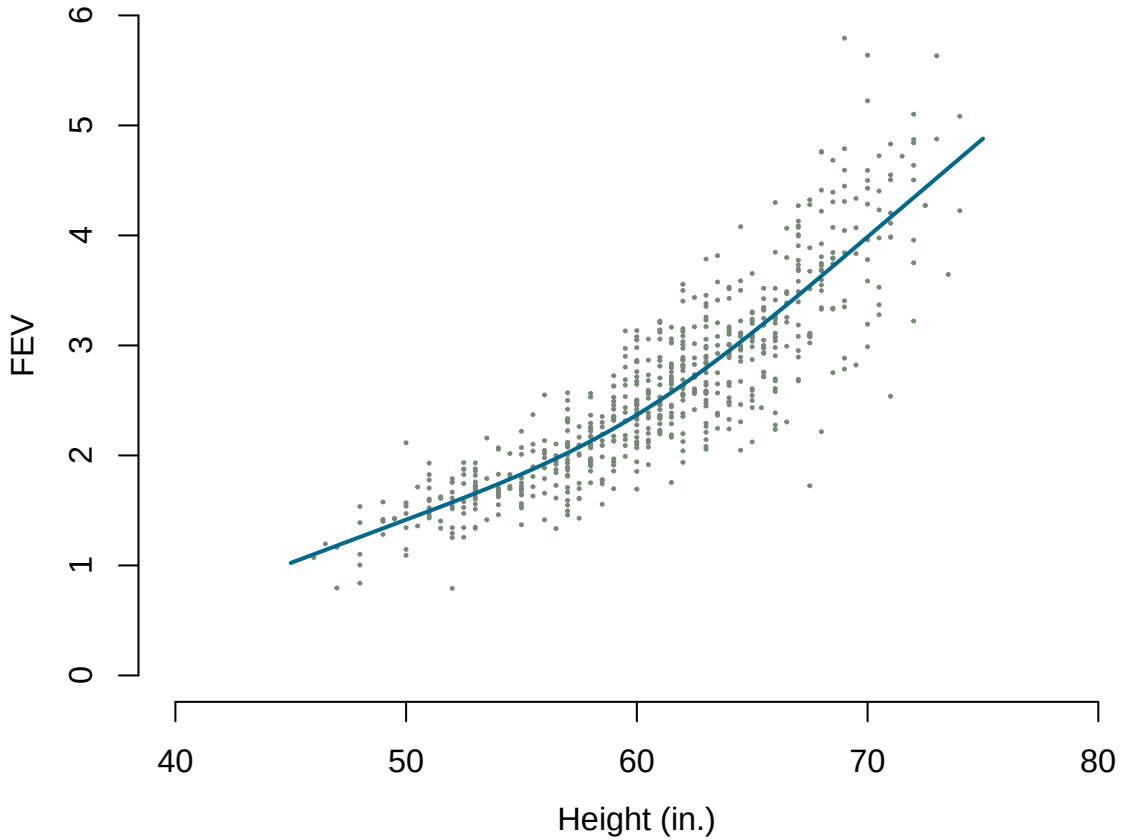
- Polynomial behavior is often erratic near boundaries, as we have seen.
- A *natural* or *restricted* cubic spline is linear beyond the boundary knots. This frees up four degrees of freedom (two per boundary).
- Hence, a natural cubic spline with P knots uses P basis functions.
- Basis expansion for natural cubic spline with P knots:
  - $h_0(x) = 1$.
  - $h_1(x) = x$.
  - $h_{p+2}(x) = d_{p+1}(x) - d_{P-1}(x)$, for $p = 0, \ldots, P-2$

  where

$$d_p(x) = \frac{(x-\zeta_p)_+^3 - (x-\zeta_P)_+^3}{\zeta_P - \zeta_p}.$$

- (This is a mess. I won't ask you to do a natural cubic spline by hand).

FEV: Natural cubic spline; knots at 50, 60, and 70 in.



More specifics:

- Discussion point: where do we put the knots?
  - Visual/as needed.
  - Equally spaced over range.
  - Quantiles (recommended: Harrell (2001)).

Example: FEV study

- Let's return to the example of smoking and FEV, this time adjusting for age and height using natural cubic splines.
- Variables (some re-coded):
  - X_1: smoke (0 = No; 1 = Yes).
  - X_2: age (years)
  - X_3: height (inches)
  - X_4: sex (0 = Female; 1 = Male)
  - Y: FEV (L)
- Model description: Log-transform FEV; include all predictors with natural cubic splines on age and height (four knots at default percentiles).
- As a bit of review, let's also test the overall association between each of age/height and (geometric mean) FEV.

Example: FEV study

```
## Generate basis functions for natural cubic splines
ncs.R <- function(x, knots, stub = "n") {
  N <- length(x); P <- length(knots)
  zP <- knots[P]; zP.1 <- knots[P - 1]
  bmat <- matrix(0, nrow = N, ncol = P - 1)
  bmat[,1] <- as.numeric(x)
  nms <- c(paste(stub, 1, sep = ""))
  for (j in 1:(P - 2))
  {
    zp <- knots[j]
    dp.num <- pmax(0, (x - zp)^3) - pmax(0, (x - zP)^3)
    dp <- dp.num/(zP - zp)
    dP.1.num <- pmax(0, (x - zP.1)^3) - pmax(0, (x - zP)^3)
    dP.1 <- dP.1.num/(zP - zP.1)
    bmat[,j + 1] <- dp - dP.1
    nms <- c(nms, paste(stub, j + 1, sep = ""))
  }
  bmat <- data.frame(cbind(1, bmat))
  names(bmat) <- c(paste(stub, 0, sep = ""), nms)
  return(bmat)
}
```

Example: FEV study

```
## Attach basis functions for age
dat <- cbind(dat,
             ncs.R(dat$age,
                   knots = quantile(dat$age, c(10,35,65,90)/100),
                   stub = "a")[,2:4])

## Attach basis functions for height
dat <- cbind(dat,
             ncs.R(dat$height,
                   knots = quantile(dat$height, c(10,35,65,90)/100),
                   stub = "h")[,2:4])

## Re-code sex
dat$sex01 <- 2 - dat$sex
```

Example: FEV study

```
## Fit model
model <- regress("geometric mean", fev ~ smoke01 + a1 + a2 + a3 + h1 + h2 + h3 + sex01,
                 data = dat)

## Abridged output
> model

Call:
  regress(fnctl = "geometric mean", formula = fev ~ smoke01 + a1 +
            a2 + a3 + h1 + h2 + h3 + sex01, data = dat)

Transformed Model:
                  e(Est)    e(95%L)   e(95%H)   F stat  df      Pr(>F)
[1] Intercept      0.115     0.0754     0.175   102.64   1  < 0.00005
[2] smoke01        0.956      0.911      1.00     3.46   1     0.0633
[3] a1              1.01      0.985      1.03     0.24   1     0.6237
[4] a2              1.01       1.00      1.01     5.11   1     0.0241
[5] a3             0.984      0.970     0.998     5.16   1     0.0234
[6] h1              1.05       1.04      1.06   122.78   1  < 0.00005
[7] h2             0.999      0.998      1.00     1.12   1     0.2906
[8] h3              1.00      0.998      1.00     0.48   1     0.4874
[9] sex01           1.04       1.01      1.06     7.09   1     0.0079

Residual standard error: 0.145 on 645 degrees of freedom
Multiple R-squared:  0.812, Adjusted R-squared:  0.81
F-statistic:  318 on 8 and 645 DF,  p-value: <2e-16
```

R: Testing age association

```
## Constraint matrix
R <- matrix(0, nrow = 3, ncol = 9)
R[1,3] <- R[2,4] <- R[3,5] <- 1

## Test of interest
> lincom(model, R, joint.test = TRUE, useFdstn = TRUE)

      F stat num df den df p value
[1,]   18.5    3.0    645 1.6e-11 ***
```

R: Testing age association

```
## Constraint matrix
R <- matrix(0, nrow = 3, ncol = 9)
R[1,6] <- R[2,7] <- R[3,8] <- 1

## Test of interest
> lincom(model, R, joint.test = TRUE, useFdstn = TRUE)

      F stat num df den df p value
[1,]    155      3    645  <2e-16 ***
```

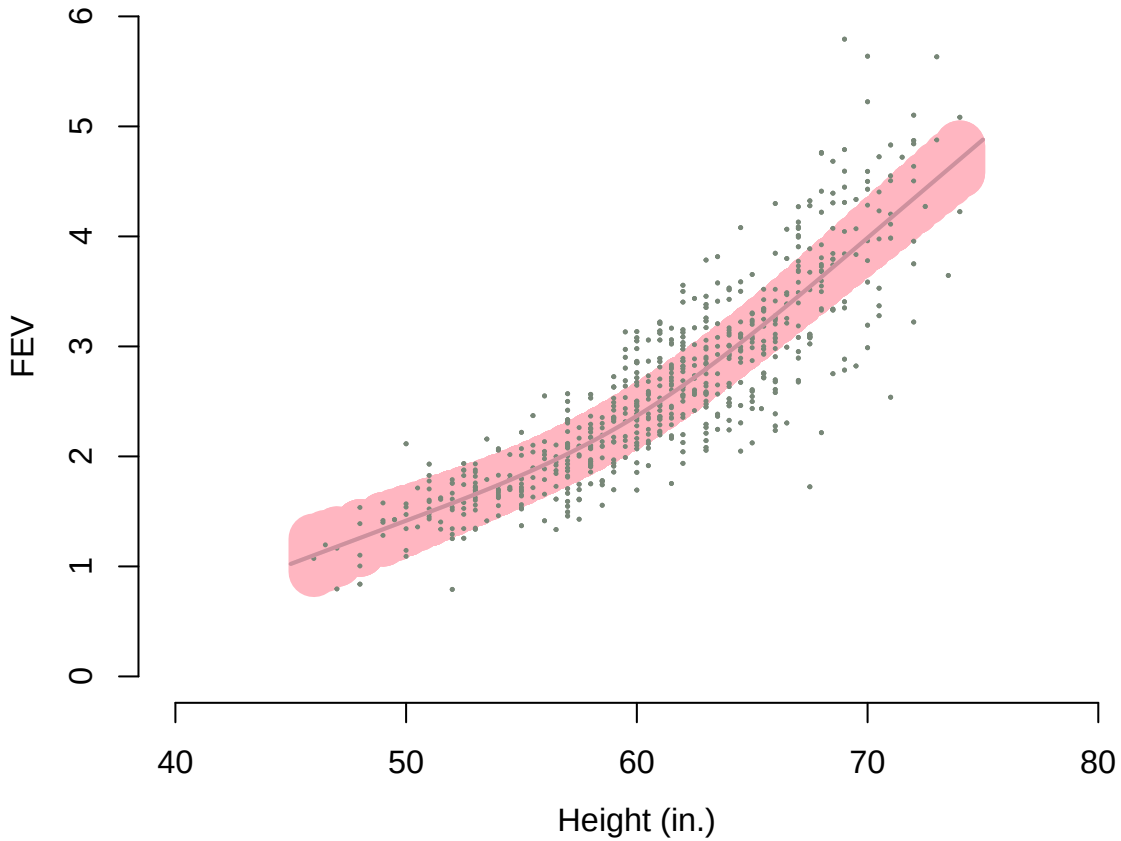
More thoughts:

- Flexible way to reduce bias (even if we take a hit in variability).
- Coefficients possess no straightforward interpretation.
- When conducting hypothesis tests to evaluate association between X and (mean) Y, don't forget joint test!

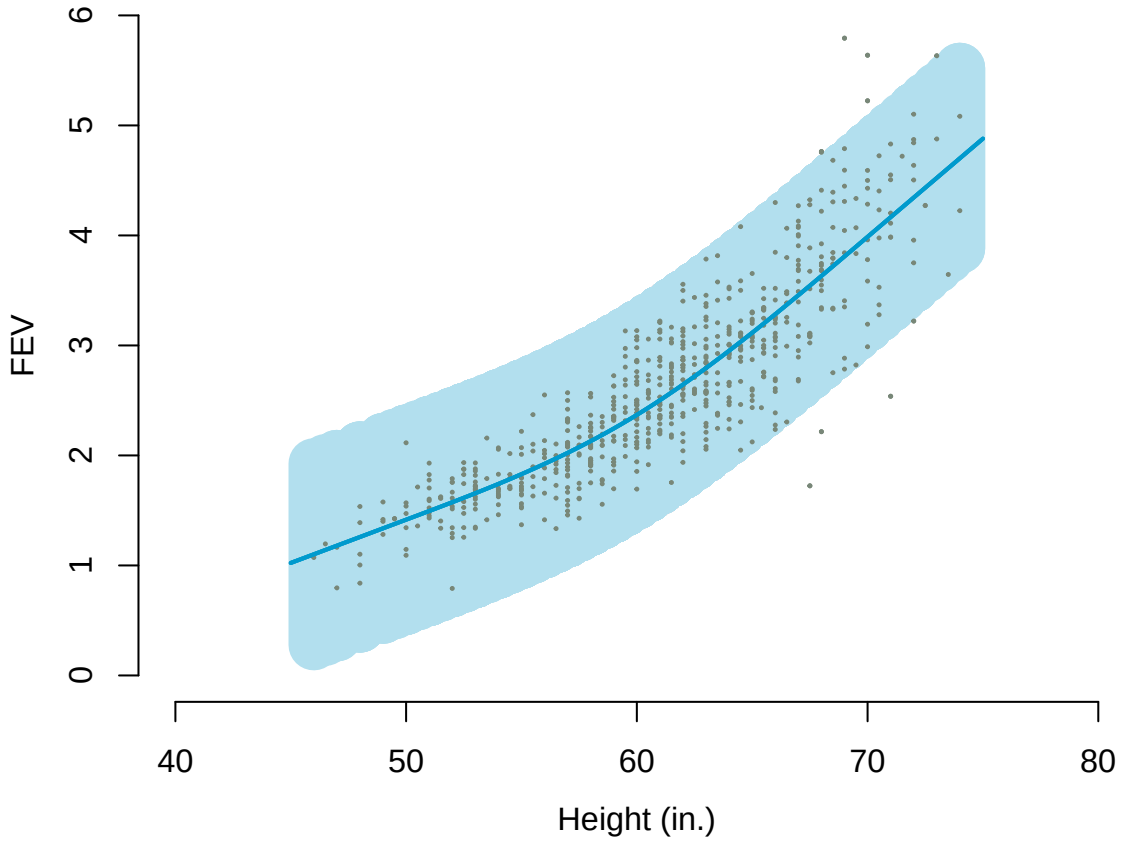
Notes on diagnostics, stratum-specific analysis:

- Diagnostics, stratum-specific confidence intervals, and stratum-specific prediction intervals all generalize in the way you would expect.
- As an example, return to the height model with a natural cubic spline with knots at 50, 60, and 70 in.

FEV: Point-wise confidence band



FEV: Point-wise prediction band for future observations



Summary

This unit:

- Shifting and scaling (convenience).
- Log-transformation (typically science-driven).
- Basis expansions, including natural cubic splines.

Summary

So far:

- Review.
- Simple linear regression.
- Multiple linear regression (foundations).
- Multiple linear regression (interactions and strata).
- Transformations and basis expansions.

Summary

Coming up:

- Regression with binary outcomes.
- Regression with nominal, ordinal, and count outcomes.
- Introduction to clustered data.
- Methods for time-to-event outcomes.
- Predictive capacity of regression models.