Biost 517 / Biost 514
Applied Biostatistics I / Biostatistics I

Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 18:

Linear Regression: Geometric Mean and Log Transformation of Response and/or Predictor

Linear Regression

- For the simple linear regression, the mean value of a response variable is assumed to change linearly with the predictor:

$$E(Y \mid X) = \beta_0 + \beta_1 \times X$$

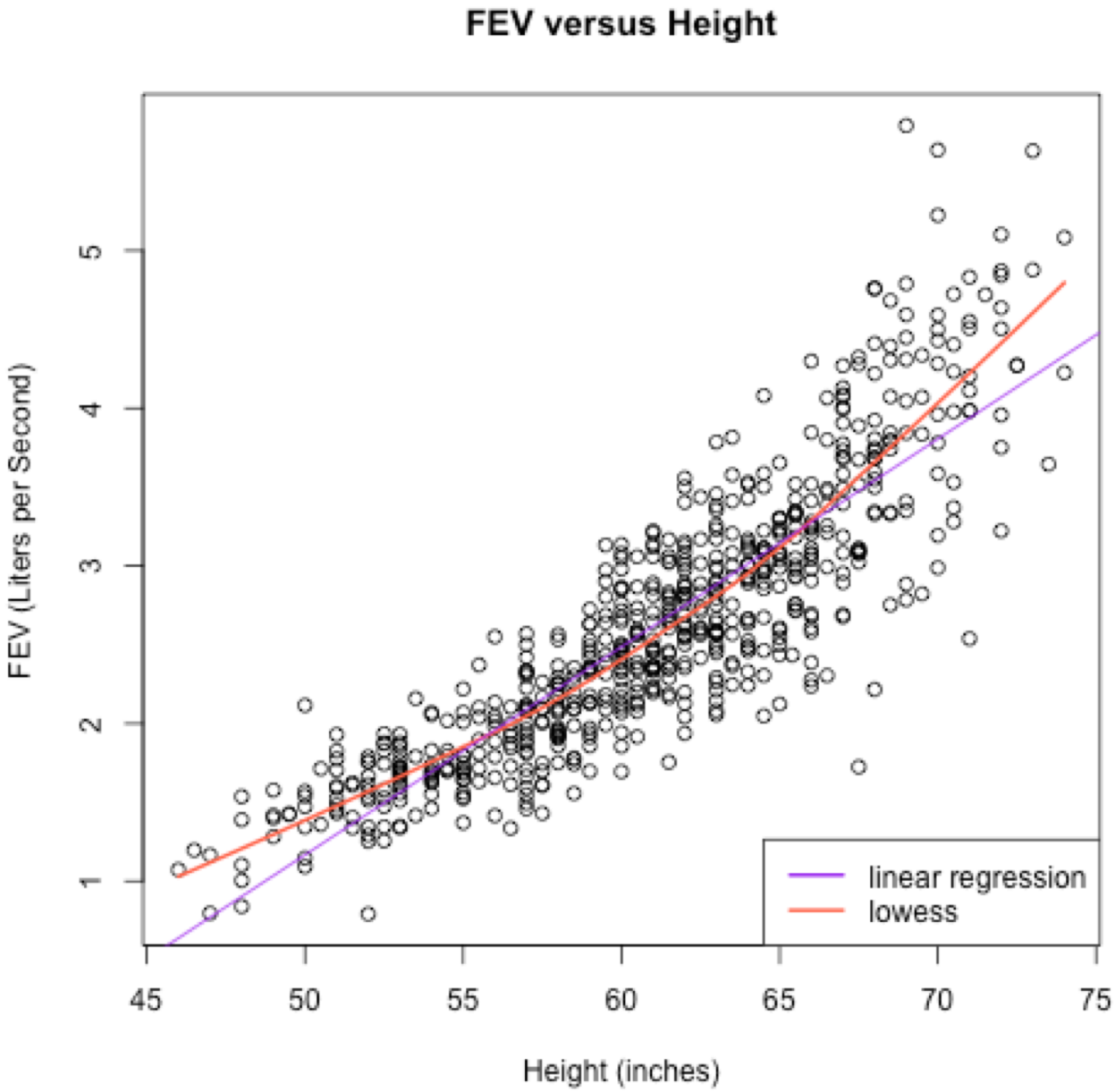
- A straight line, however, may not be an adequate representation of the true relationship between variables of interest.
- In addition, classical regression assumes constant variance for the response across the groupings defined by the predictor(s) of interest
- Transformation of response and/or predictors are often used with linear regression to:
  - make relationships between mean response and predictor linear
  - normalize residuals
  - stabilize non-constant variance

Example: Trends in FEV by Height

- FEV data set
  - A sample of 654 healthy children
-
- Lung function measured by forced expiratory volume (FEV)
  - maximal amount of air expired in 1 second (Liters per second)

- Scientific Question of interest: How does FEV differ across height groups?

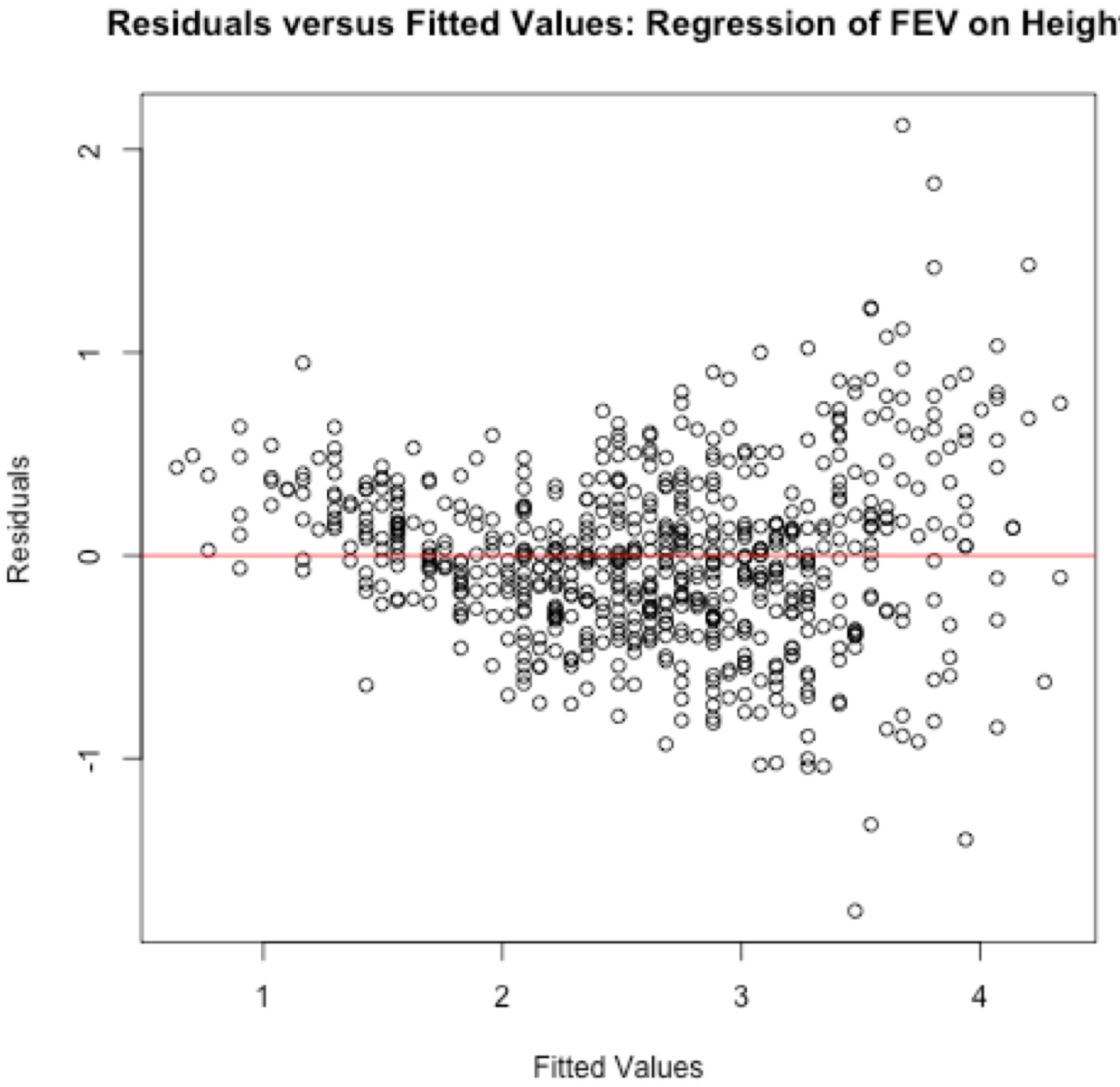
FEV versus Height



Characterization of Scatterplot

- Detection of outliers
  - None obvious
- Trends in FEV across groups
  - FEV tends to be larger for taller children
- Second order trends
  - Curvilinear increase in FEV with height
- Variation within height groups
  - “heteroscedastic”: unequal variance across groups
    - mean-variance relationship: higher variation in groups with higher FEV

FEV versus Height



Log Transformation of Response

- Log transformation of the response is the most common choice of transformation when a positive response variable is continuous
- Important to understand that linear regression on log transformed data changes the interpretation of the parameters
- Recall that for a linear regression with a predictor and outcome on their measured scale, the regression coefficient for the slope is interpreted as the change in the average value of the outcome for every unit increase in the predictor
- Linear regression with a log transformed response gives inference on the **geometric mean**, and mean value changes of the response for different values of the predictor relative are **percentage change** as opposed to absolute change.

Interpretation of Parameters with Log Transformed Response

- Linear regression on log transformed Y
  - Here, we "log" is referring to the natural log (ln)
  - Mean response has a linear relationship with the predictor on the log scale

Model $$E\left[\log Y_i \,\middle|\, X_i\right] = \beta_0 + \beta_1 \times X_i$$

$$X_i = 0 \qquad E\left[\log Y_i \,\middle|\, X_i = 0\right] = \beta_0$$

$$X_i = x \qquad E\left[\log Y_i \,\middle|\, X_i = x\right] = \beta_0 + \beta_1 \times x$$

$$X_i = x+1 \qquad E\left[\log Y_i \,\middle|\, X_i = x+1\right] = \beta_0 + \beta_1 \times x + \beta_1$$

Geometric Mean

For a sample of size n : Sample $GM_Y = \left(\prod_{i=1}^{n} Y_i \right)^{1/n}$

$$\log(\text{Sample } GM_Y) = \frac{1}{n} \sum_{i=1}^{n} \log(Y_i)$$

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \log(Y_i) = E[\log(Y)] = \mu_{log(Y)}$$

$$\text{Population } GM_Y = e^{\mu_{log(Y)}}$$

Interpretation of Parameters with Log Transformed Response

- Restated model as log link for geometric mean

Model $\log \mathrm{GM}\left[Y_i \mid X_i\right] = \beta_0 + \beta_1 \times X_i$

$X_i = 0$ $\log GM\left[Y_i \mid X_i = 0\right] = \beta_0$

$X_i = x$ $\log GM\left[Y_i \mid X_i = x\right] = \beta_0 + \beta_1 \times x$

$X_i = x+1$ $\log GM\left[Y_i \mid X_i = x+1\right] = \beta_0 + \beta_1 \times x + \beta_1$

Interpretation of Parameters with Log Transformed Response

- Interpretation of regression parameters by back-transforming model
  - Exponentiation is inverse of log

Model $$\mathrm{GM}\left[Y_i \mid X_i\right] = e^{\beta_0} \times e^{\beta_1 \times X_i}$$

$$X_i = 0 \qquad GM\left[Y_i \mid X_i = 0\right] = e^{\beta_0}$$

$$X_i = x \qquad GM\left[Y_i \mid X_i = x\right] = e^{\beta_0} \times e^{\beta_1 \times x}$$

$$X_i = x+1 \qquad GM\left[Y_i \mid X_i = x+1\right] = e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}$$

Interpretation of Parameters with Log Transformed Response

- Geometric mean when predictor is 0: **$\exp(\beta_0)$**
  - Found by exponentiation of the intercept from the linear regression on log transformed data
- Ratio of geometric means between two groups differing in the value of the predictor by 1 unit: **$\exp(\beta_1)$**
  - Found by exponentiation of the slope from the linear regression on log transformed data
  - Confidence intervals for geometric mean and ratios found by exponentiating the CI for regression parameters
  - What is $100(\exp(\beta_1)-1)$?

$$e^{\beta_1} - 1 = \frac{GM[Y_i \mid X_i = x+1]}{GM[Y_i \mid X_i = x]} - 1 = \frac{GM[Y_i \mid X_i = x+1] - GM[Y_i \mid X_i = x]}{GM[Y_i \mid X_i = x]}$$

  - So 100 times this value is the **percentage** increase (or decrease) in the geometric mean value of the outcome per unit increase in the predictor

Geometric Mean Model

- Science often used to dictate a model.
- Many variables of interest are measured on a log scale, so multiplicative models are often quite natural to use with regression.
- Statistical preference for transformation of response
  - In presence of heteroscedasticity “best linear unbiased estimator” requires weighting observations unequally
    - Okay if linear model truly holds
    - Scientifically unpleasing if linear model does not hold
  - Instead, we may be able to transform response to have equal variance across groups
    - “Homoscedasticity” tends toward easier and more precise inference when weighting all individuals equally
- Statistical preference for log transformation
  - Easier interpretation: multiplicative model
  - Compare groups using ratios

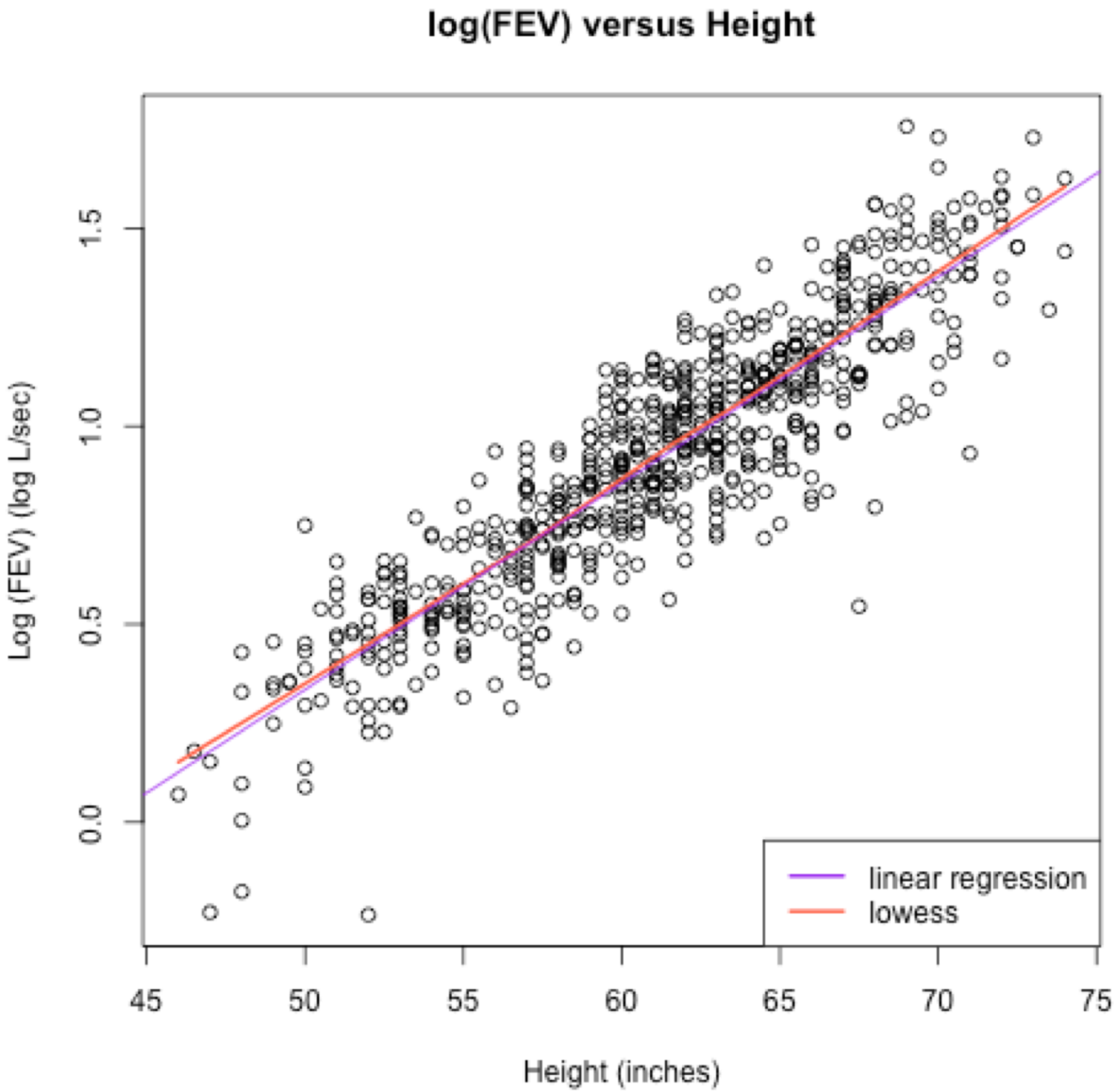
Alternative Representation of log Transformation of response

- Classical Linear Regression model with log transformed response:

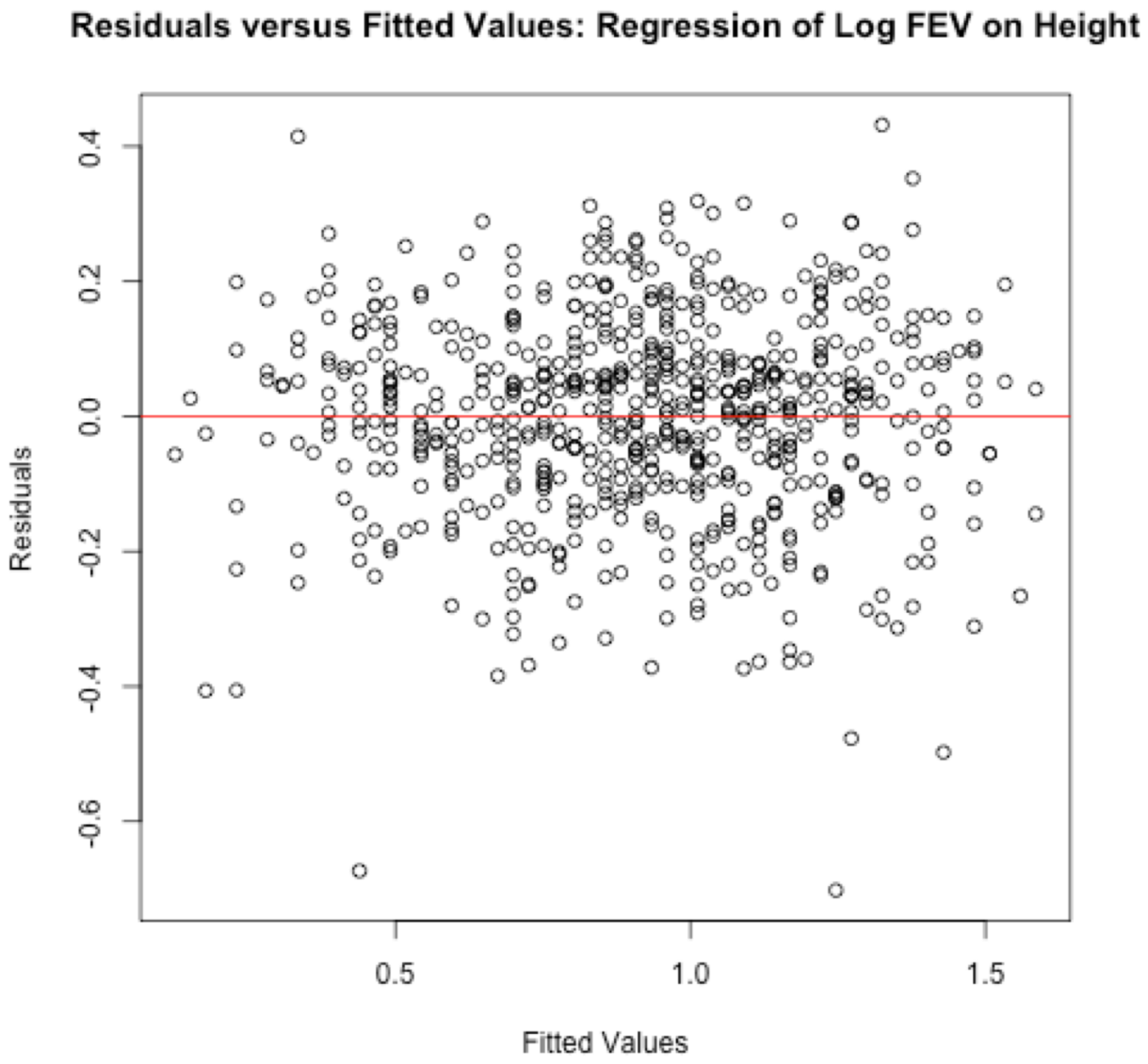
$$\log(Y_i) = \beta_0 + \beta_1 X_i + e_i$$

- Where it is assumed that $e_i \sim N(0, \sigma_e^2)$ and constant variance σ_e^2 for all groups defined by the predictor
- So the response Y is assumed to have a log-normal distribution.

Log(FEV) versus Height



Log(FEV) versus Height



Linear Regression with Log (FEV) on Height

```
> fevdata$logfev<-log(fevdata$fev)
> logreg1<-lm(logfev~height,data=fevdata)
> summary(logreg1)

Call:
lm(formula = logfev ~ height, data = fevdata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.70208 -0.08986  0.01190  0.09337  0.43174

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.271312   0.063531  -35.75   <2e-16 ***
height       0.052119   0.001035   50.38   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1508 on 652 degrees of freedom
Multiple R-squared:  0.7956,    Adjusted R-squared:  0.7953
F-statistic:  2538 on 1 and 652 DF,  p-value: < 2.2e-16
```

Linear Regression with Log (FEV) on Height with Robust SE

```
> library("sandwich")
> library("lmtest")
> coeftest(logreg1, vcov=vcovHC(logreg1, "HC1"))

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -2.2713118  0.0685518 -33.133 < 2.2e-16 ***
height       0.0521191  0.0011227  46.423 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Linear Regression with Log (FEV) on Height with Robust SE

```
> library(uwIntroStats)
> robustmodel1 <- regress ("mean", logfev~height,data=fevdata)
> robustmodel1

Call:
regress(fnctl = "mean", formula = logfev ~ height, data = fevdata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.70208 -0.08986  0.01190  0.09337  0.43174

Coefficients:
                Estimate  Naive SE   Robust SE     95%L      95%H        F stat    df Pr(>F)
[1] Intercept    -2.271     0.06353    0.06855     -2.406    -2.137        1097.78 1  < 0.00005
[2] height        0.05212   1.035e-03  1.123e-03    0.04991   0.05432      2155.08 1  < 0.00005

Residual standard error: 0.1508 on 652 degrees of freedom
Multiple R-squared:  0.7956,    Adjusted R-squared:  0.7953
F-statistic:  2155 on 1 and 652 DF,  p-value: < 2.2e-16
```

Regression Parameter Inference for Log FEV on Height: Robust SE

$$E[\log(FEV_i) \mid Height_i] = -2.271 + 0.0521 \times Height_i$$

- Ratio of geometric means of FEV between groups differing in Height by 1 inch is: $e^{\beta_1} = e^{.052} = 1.053$
- 95% confidence interval for this ratio is [1.051,1.056]:

$$(e^{.0499}, e^{.0543}) = (1.051, 1.056)$$

- P value < .0001 suggests geometric mean FEV differs across height
- The percentage increase in the geometric mean of FEV per 1 inch increase in height is estimated to be 5.3%:

$$100 * (e^{\beta_1} - 1) = 100 * (.053) = 5.3$$

Example: Interpretation with log transformed response

"From linear regression analysis on log transformed FEV using Huber-White estimates of the standard error, we estimate that for every 1 inch difference in height between two groups of children, the geometric mean FEV is 5.35% higher in the taller population. A 95% CI suggests that this observation is not unusual if the true relationship between geometric means were such that the taller group's geometric mean FEV were between 5.12% and 5.56% higher for each 1 inch difference in height. Because the two-sided P value is P < .0005, we reject the null hypothesis that there is no linear trend in the average log transformed FEV across height groups."

Log Transformed Predictor

- Log transformations of predictors are often used for linearity between the response and predictor in a linear regression analysis. As before, "log" is referring to the natural log (ln)

Model	$E[Y_i \mid log(X_i)] = \beta_0 + \beta_1 \times \log(X_i)$
$X_i = 1$	$E[Y_i \mid log(X_i = 1)] = \beta_0$
$X_i = x$	$E[Y_i \mid log(X_i = x)] = \beta_0 + \beta_1 \times \log(x)$
$X_i = kx$	$E[Y_i \mid log(X_i = kx)] = \beta_0 + \beta_1 \times \log(kx)$

- If $k = 1.01$, then $1.01x$ corresponds to a 1% increase in x.
- If $k = 1.1$, then $1.1x$ corresponds to a 10% increase in x.
- If $k = 2$, then $2x$ corresponds to a 100% increase in x.

Log Transformed Predictor

$$E[Y_i \mid log(X_i)] = \beta_0 + \beta_1 \times \log(X_i)$$

- Intercept β_0 is the mean of the response Y for subpopulation with $X = 1$. May not be of scientific interest. Maybe outside the range of data.
- With a linear regression model with a (natural) log transformed predictor what is the difference in mean response between two populations that differ in the predictor by 1%?

$$E[Y_i \mid log(X_i = 1.01x)] - E[Y_i \mid log(X_i = x)] =$$

$$\beta_1 \times \log(1.01x) - \beta_1 \times \log(x) = \beta_1 \log(1.01) = \beta_1 \log\left(1 + \frac{1}{100}\right) \approx \frac{\beta_1}{100}$$

Since

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - \cdots$$

$$= \sum_{n=1}^{\infty} (-1)^{(n-1)} \frac{x^n}{n} \text{ or } \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n}$$

Log Transformed Predictor

$$E[Y_i \mid log(X_i)] = \beta_0 + \beta_1 \times \log(X_i)$$

- So using a linear regression model with a (natural) log transformed predictor, two populations who differ by 1% for the predictor variable have a difference in mean response of

$$\beta_1 \log(1.01) \approx \frac{\beta_1}{100}$$

- What is the mean difference in response for two populations who differ by 5% for the predictor variable from a linear regression model with a natural log transformed predictor?

$$\beta_1 log(1.05) \approx \beta_1 \frac{5}{100} = \frac{\beta_1}{20}$$

Log Transformation of both Predictor and Response

- Interpretation of log transformed predictors with log link function (log transformed response)
  - Log link used to model the geometric mean
    - Exponentiated slope estimates ratio of geometric means across groups
  - Compare groups with a k-fold difference in their measured predictors is $\exp\left(\log(k)\times\beta_1\right)=k^{\beta_1}$
    - Estimated ratio of geometric means is :

$$\log[GM_{Y_i}\,|\log(X_i)]=\beta_0+\beta_1\log(X_i)$$

$$\log[GM_{Y_i}\,|\log(kX_i)]=\beta_0+\beta_1\log(kX_i)$$

$$\log[GM_{Y_i}\,|\log(kX_i)]-\log[GM_{Y_i}\,|\log(X_i)]$$

$$=\log\left(\frac{GM_{Y_i}\,|\log(kX_i)}{GM_{Y_i}\,|\log(X_i)}\right)=\beta_1\log(k)$$

$$\frac{GM_{Y_i}\,|\log(kX_i)}{GM_{Y_i}\,|\log(X_i)}=k^{\beta_1}$$

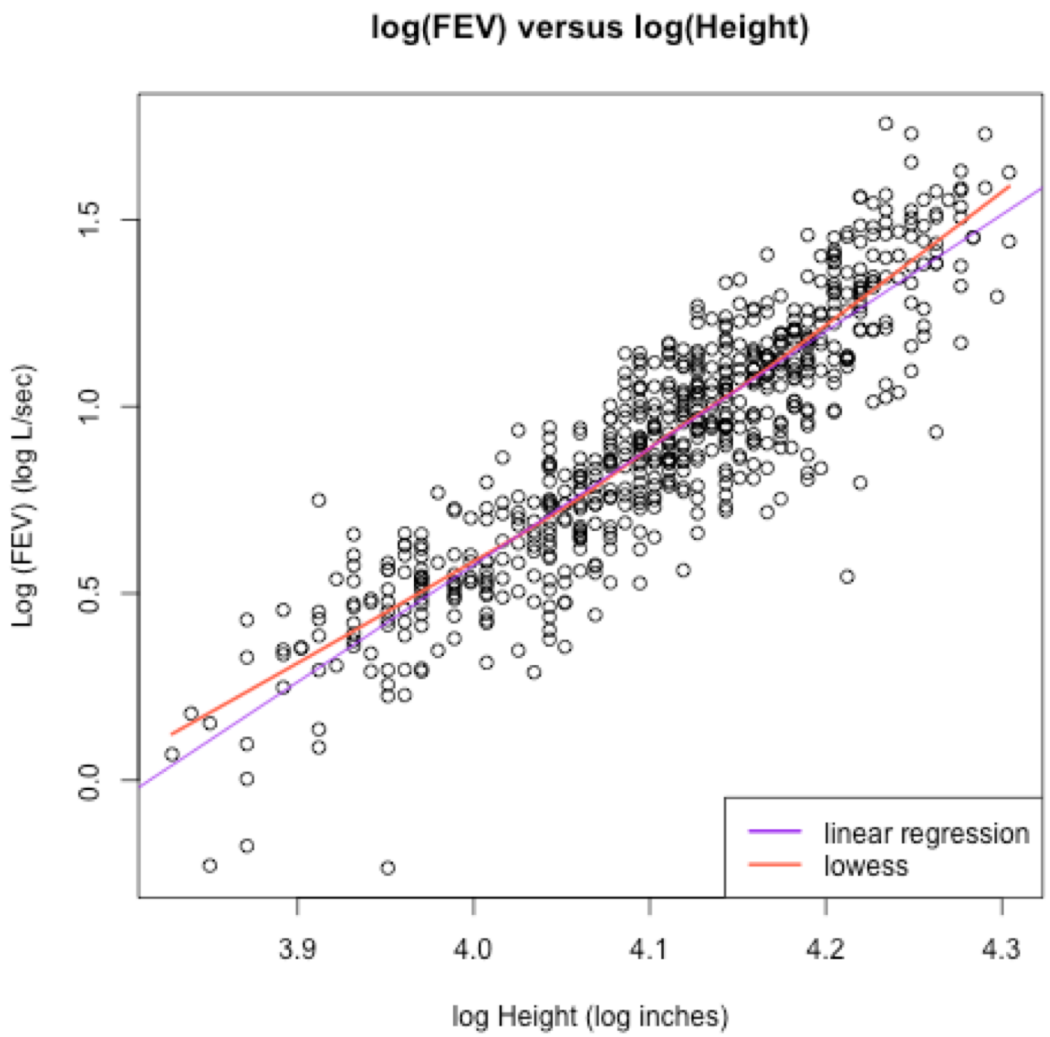
- Scientific justification for geometric mean for FEV and log of Height
  - FEV is a volume
  - Height is a linear dimension
    - Each dimension of lung size is proportional to height
  - Standard deviation likely proportional to height

Science $FEV \propto Height^3$

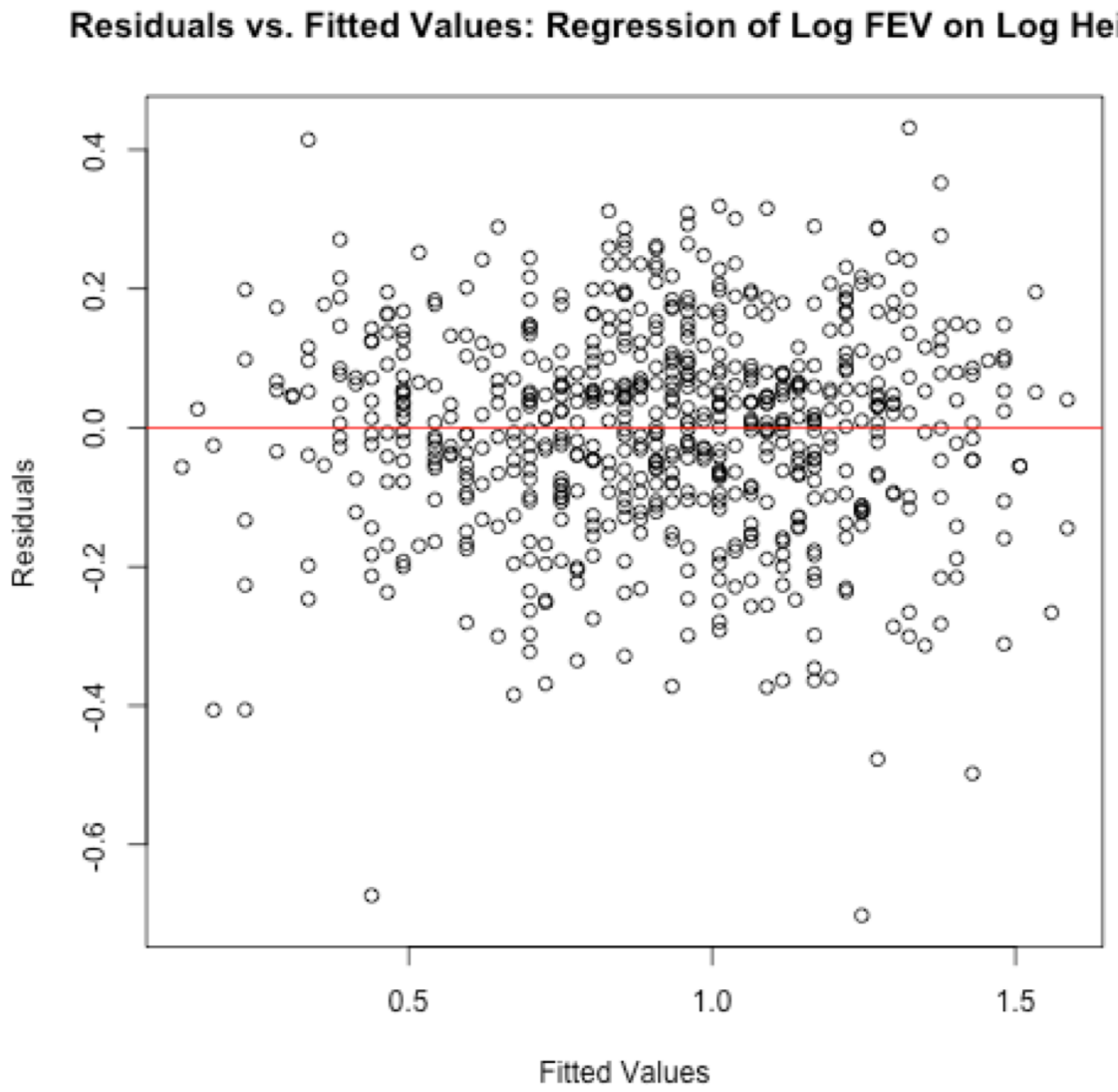
$\sqrt[3]{FEV} \propto Height$

Statistics $\log(FEV) \propto 3\log(Height)$

Log(FEV) versus log(Height)



Log(FEV) versus log(Height)



Regression: Log(FEV) versus log(Height)

```
> fevdata$loght<-log(fevdata$height)
> logreg2<-lm(logfev~loght,data=fevdata)
> summary(logreg2)

Call:
lm(formula = logfev ~ loght, data = fevdata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.69369 -0.09122  0.01145  0.09832  0.44965

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.92110    0.25577  -46.61   <2e-16 ***
loght         3.12418    0.06223   50.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1512 on 652 degrees of freedom
Multiple R-squared:  0.7945,    Adjusted R-squared:  0.7941
F-statistic:  2520 on 1 and 652 DF,  p-value: < 2.2e-16
```

Regression: Log(FEV) versus log(Height): Robust SE

```
> robustmodel2 <- regress ("mean", logfev~loght,data=fevdata)
> robustmodel2

Call:
regress(fnctl = "mean", formula = logfev ~ loght, data = fevdata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.69369 -0.09122  0.01145  0.09832  0.44965

Coefficients:
                Estimate  Naive SE  Robust SE    95%L      95%H         F stat    df Pr(>F)
[1] Intercept    -11.92     0.2558    0.2779     -12.47    -11.38         1840.19 1  < 0.00005
[2] loght          3.124    0.06223   0.06769      2.991     3.257        2130.18 1  < 0.00005

Residual standard error: 0.1512 on 652 degrees of freedom
Multiple R-squared:  0.7945,    Adjusted R-squared:  0.7941
F-statistic:  2130 on 1 and 652 DF,  p-value: < 2.2e-16
```

Regression Parameter Inference for Log FEV on Log Height: Robust SE

- Scientific interpretation of the slope

$$\log \mathrm{GM}\left[FEV_i \,|\, loght_i\right] = -11.9 + 3.12 \times loght_i$$

- Estimated ratio of geometric mean FEV for two groups differing by 10% in height (1.1-fold difference in height)
  - Exponentiate 1.1 to the slope: $1.1^{3.12} = 1.346$
    - Group that is 10% taller is estimated to have a geometric mean FEV that is 1.35 times higher (35% higher)
    - 95% confidence interval: $(1.1^{2.99}, 1.1^{3.26}) = (1.33, 1.36)$

Interpretation of Parameters

"From linear regression analysis on log transformed FEV using Huber-White estimates of the standard error, we estimate that when comparing two groups of children differing in height by 10%, the geometric mean FEV is 34.6% higher in the taller population. A 95% CI suggests that this observation is not unusual if the true relationship between geometric means were such that the taller group's geometric mean FEV were between 33.0% and 36.4% higher than that in the shorter group. Because the two sided P value is P < .0005, we reject the null hypothesis that there is no linear trend in average FEV across height groups."

More Interpretable Estimates by Rescaling

- rescale log height to base 1.1. Rescaled height values that differ by 1 unit on the new scale have a 10% difference in height values

$$\log_b x = \frac{\log_a x}{\log_a b}$$

- Conduct regression on the rescaled predictor and exponentiate coefficients
- (note that Robust SE have been transformed in complicated way, but CI computed on the log scale so are just transformations)

```
> fevdata$loght1.1<-fevdata$loght/log(1.1)
> robustmodel3 <- regress ("mean", logfev~loght1.1,data=fevdata)
> exp(robustmodel3$coefficients[,-c(6,7)])
                Estimate Naive SE Robust SE         95%L         95%H
(Intercept) 6.648610e-06 1.291453  1.320352 3.852509e-06 1.147408e-05
loght1.1    1.346847e+00 1.005949  1.006472 1.329892e+00 1.364018e+00
```

Why Transform Predictor?

- Typically chosen according to whether the data likely follow a straight line relationship

- Linearity ("model fit") necessary to predict the value of the parameter in individual groups
  - Linearity is not necessary to estimate existence of association
  - Linearity is not necessary to estimate a "first order trend" in the parameter across groups having the sampled distribution of the predictor
  - (Inference about these two questions will tend to be conservative if linearity does not hold)

Choice of Transformation

- Rarely do we know which transformation of the response or predictor variables provides best “linear” fit
- A Box-Cox transformation is often used to transform non-normal dependent variables into a normal shape. For some variable Y, the Box-Cox transformation of Y for some exponent λ is

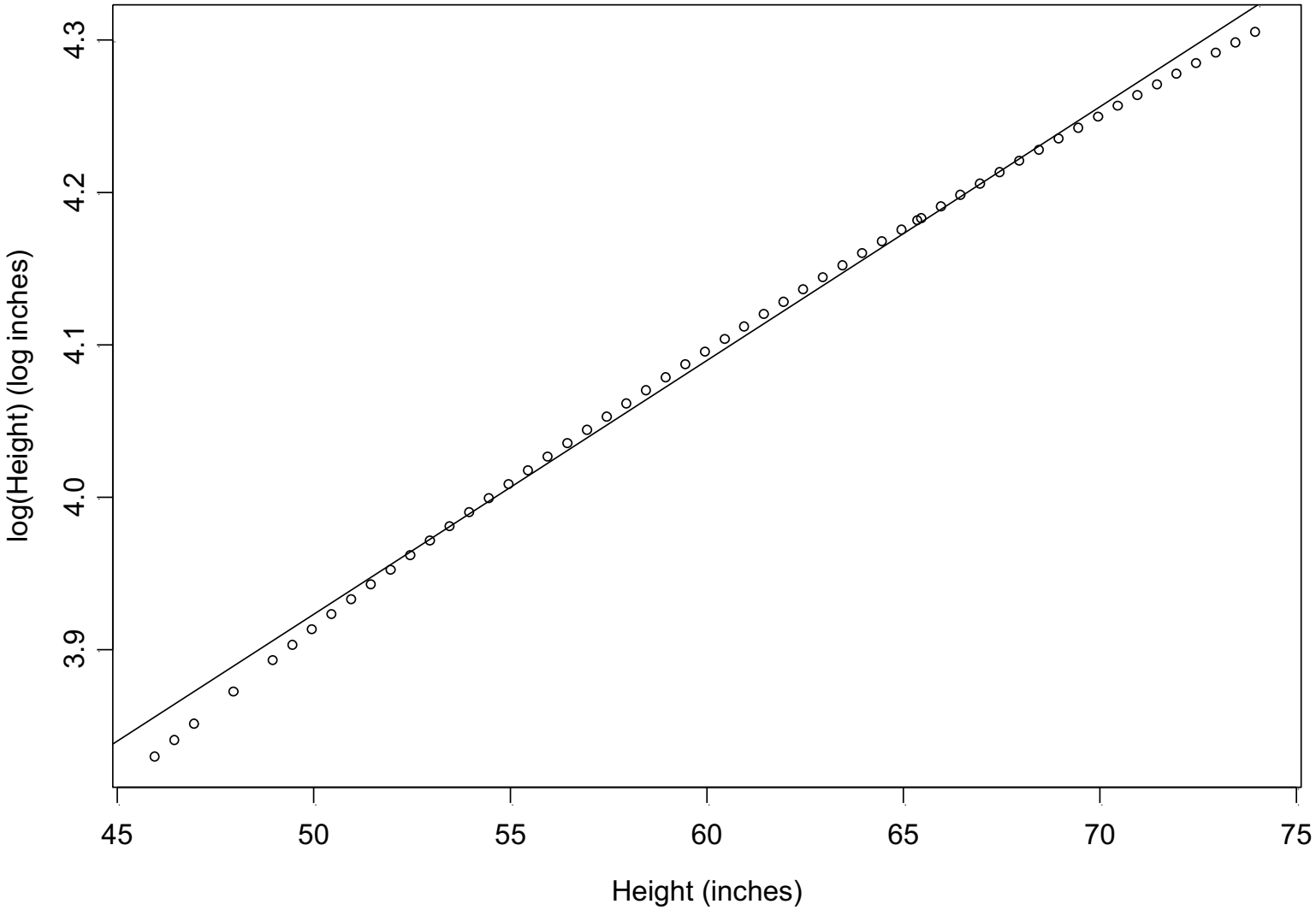
$$\frac{Y^{\lambda}-1}{\lambda} \text{ if } \lambda \neq 0; \text{ and } \log(Y) \text{ if } \lambda = 0$$

- As always, there is a danger in using the data to estimate the best transformation to use
  - If there is no association of any kind between the response and the predictor, a “linear” fit (with a zero slope) is the correct one
  - Trying to detect a transformation is thus an informal test for an association
    - Multiple testing procedures inflate the type I error

Sometimes Does Not Matter

- It is best to choose the transformation of the response and/or predictor on scientific grounds
- However, it is often the case that many functions are well approximated by a straight line over a small range of the data
  - Example: In the modeling of FEV as a function of height, the logarithm of height is approximately linear over the range of heights sampled

log(Height) versus Height



Untransformed Predictors

- It is thus often the case that we can choose to use an untransformed predictor even when science would suggest a nonlinear association

- This can have advantages when interpreting the results of the analysis
  - E.g., it is far more natural to compare heights by differences than by ratios
    - Chances are we would characterize two children as differing by 4 inches in height rather than as the 44 inch child as being 10% taller than the 40 inch child