Continuous data – regression diagnostics and transformations

Vocabulary and Notation

- Each subject has an observed value, a fitted value, and a residual

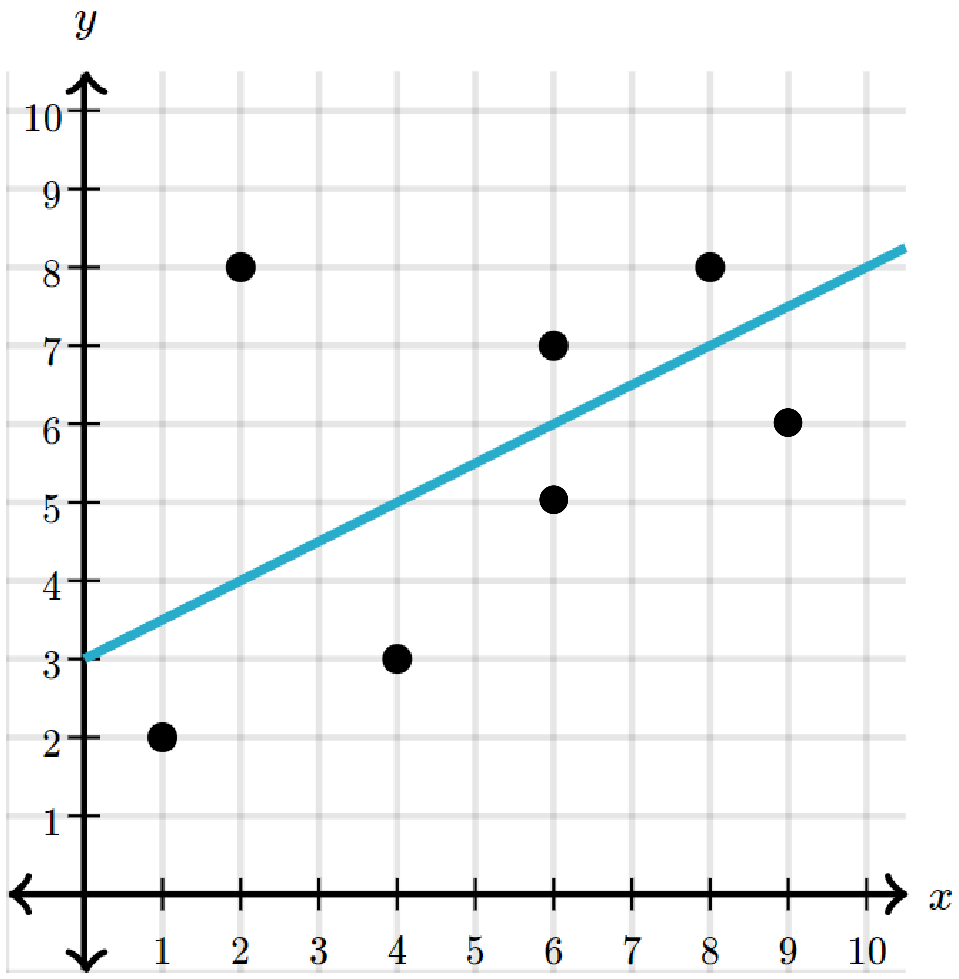
- Observed value = y
  - The actual observed value of the outcome variable in the dataset
- Fitted value = $\hat{y}$
  - The predicted value of the outcome variable (on the line)
- Residual = e
  - The vertical distance between the point and the regression line
  - Difference between y and $\hat{y}$

Vocabulary and Notation

- Proper notation for regression line:

$$\hat{y} = 3 + 0.5x \quad \text{or} \quad y = 3 + 0.5x + e$$

x	y	$\hat{y}$	e
1	2	3.5	-1.5
2	8	4	4
4	3	5	-2
6	7	6	1
6	5	6	-1
8	8	7	1
9	6	7.5	-1.5



Assumptions of Linear Regression

- **Independence** of the observations
- **Linearity** of the relationship between the two variables
- **Constant variance** of the residuals
- **Normality** of the residuals

Independence

What it means

- The observed outcomes in the dataset are independent
- One subject's outcome doesn't impact any other subjects' outcomes

How to assess it

- Knowledge of the study design and data collection
- Subject matter knowledge

How to fix it

- You can't! Avoid this problem at the study design phase
- Advanced statistical methodologies exist to handle correlated outcomes

Assumptions of Linear Regression

- **Independence** of the observations
- **Linearity** of the relationship between the two variables
- **Constant variance** of the residuals
- **Normality** of the residuals

Linearity

What it means

- Relationship between x and y is linear
- Linear regression won't pick up on non-linear associations

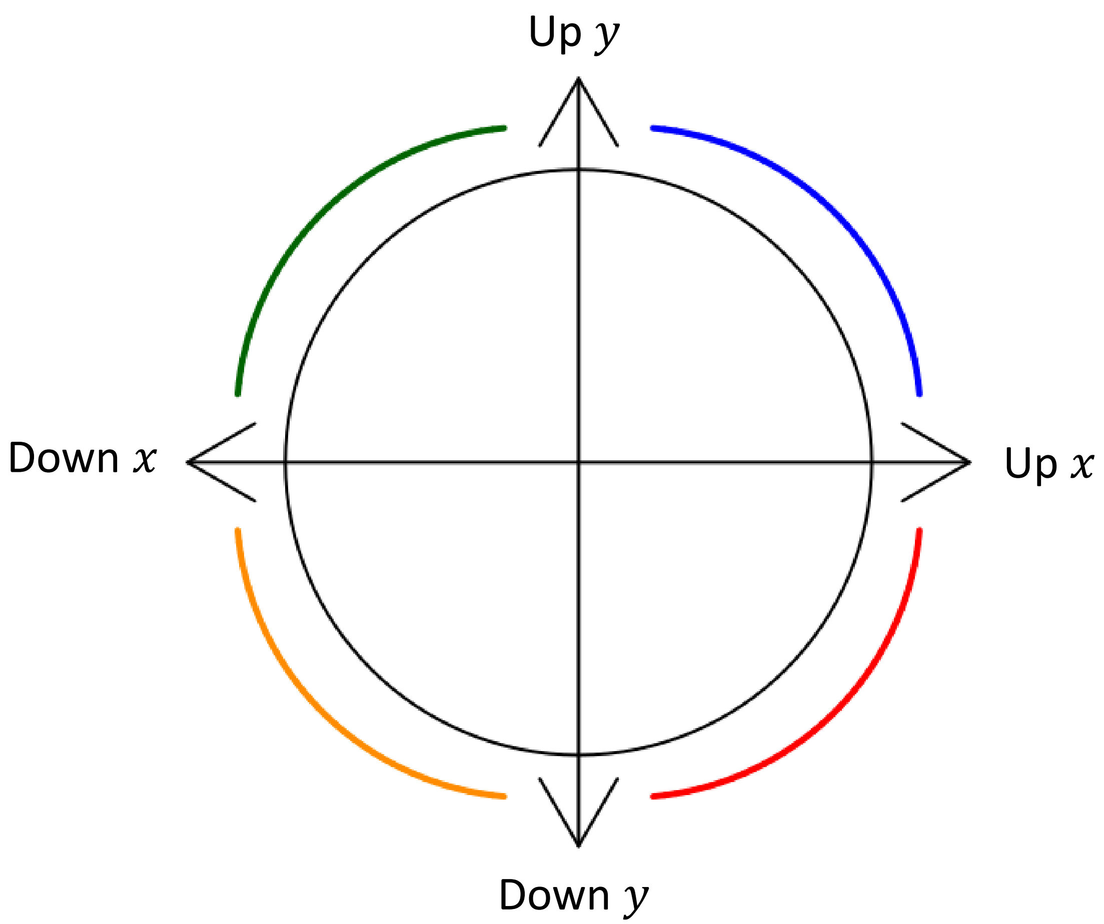
How to assess it

- Make a scatterplot of x and y, visual check for linear trend

How to fix it

- Transform x or y to make the relationship linear
- Use circle/ladder of powers to determine appropriate transformation

Circle/Ladder of Powers



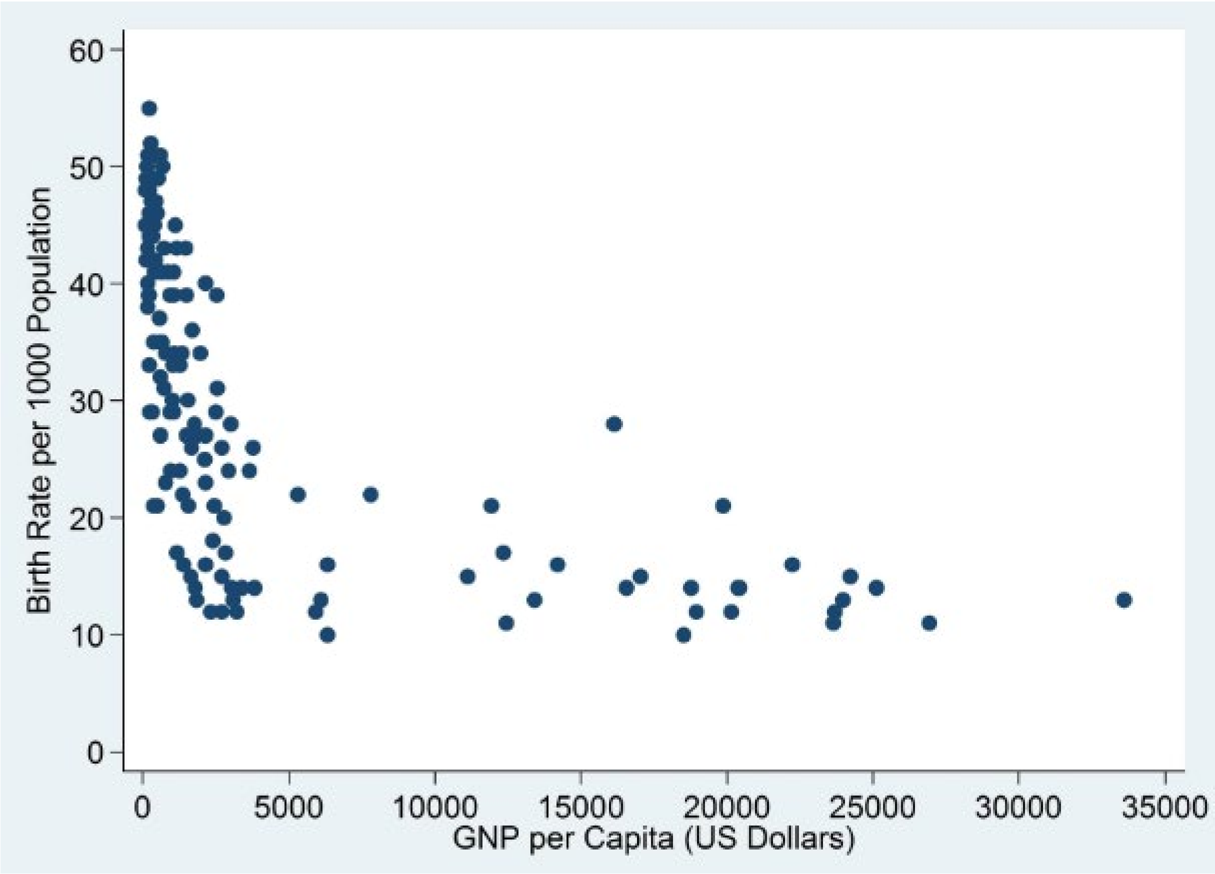
Power	Transformation
3	x^3
2	x^2
1	x
½	$\sqrt{x}$
0	$\log(x)$
-½	$\frac{1}{\sqrt{x}}$
-1	$\frac{1}{x}$
-2	$\frac{1}{x^2}$

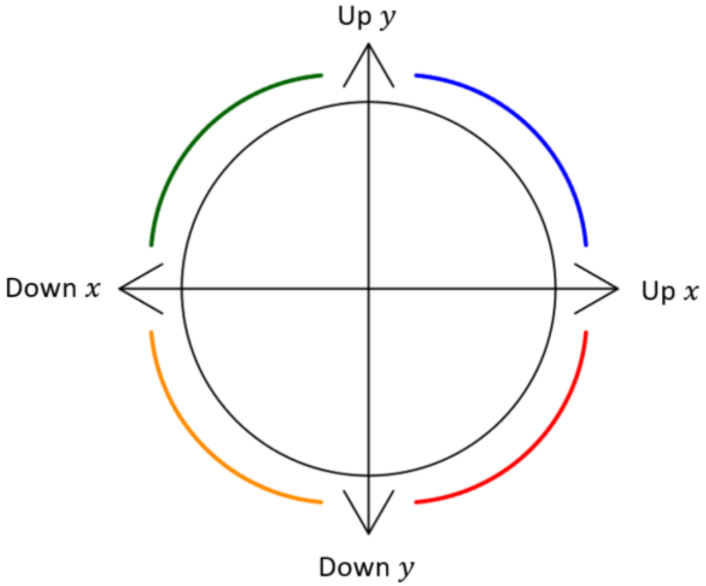
up the ladder

start here

down the ladder

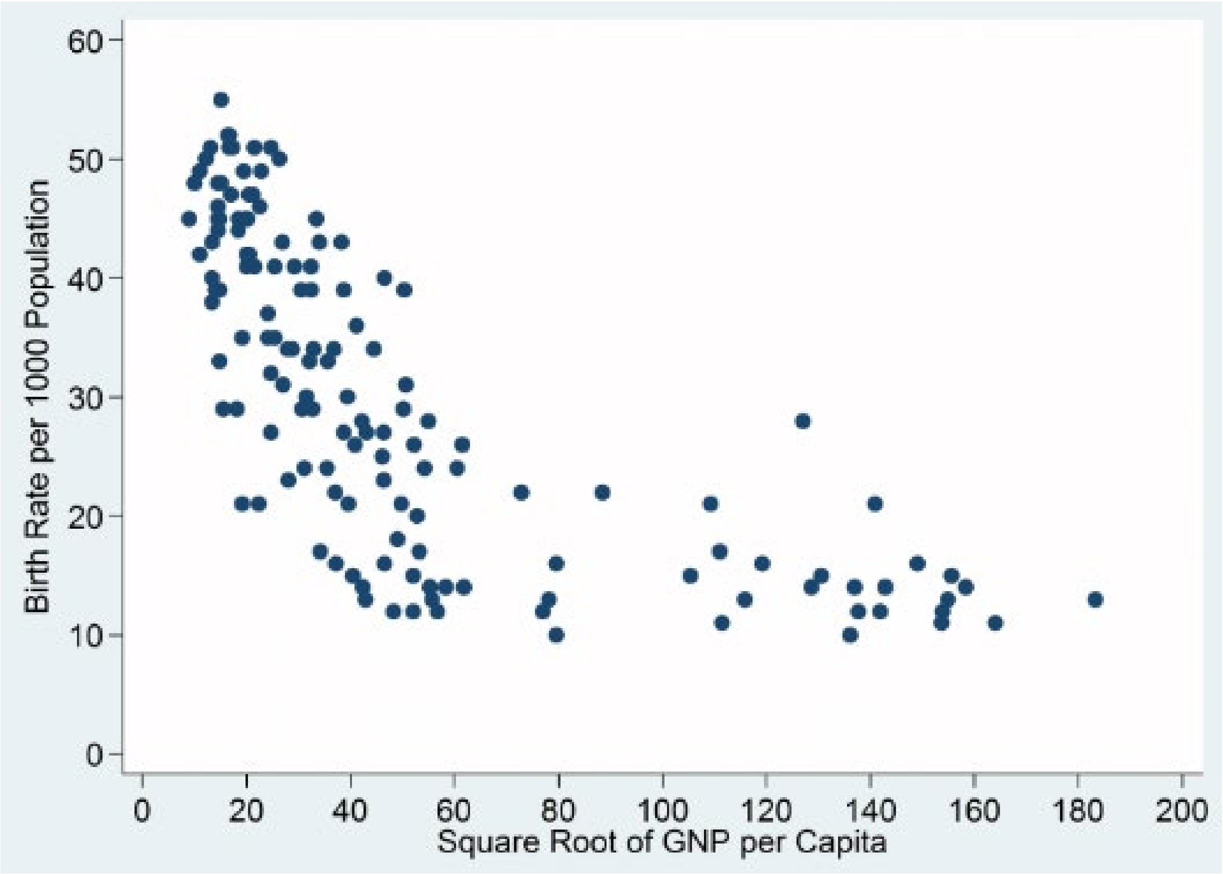
Example: GNP vs. Birth Rate

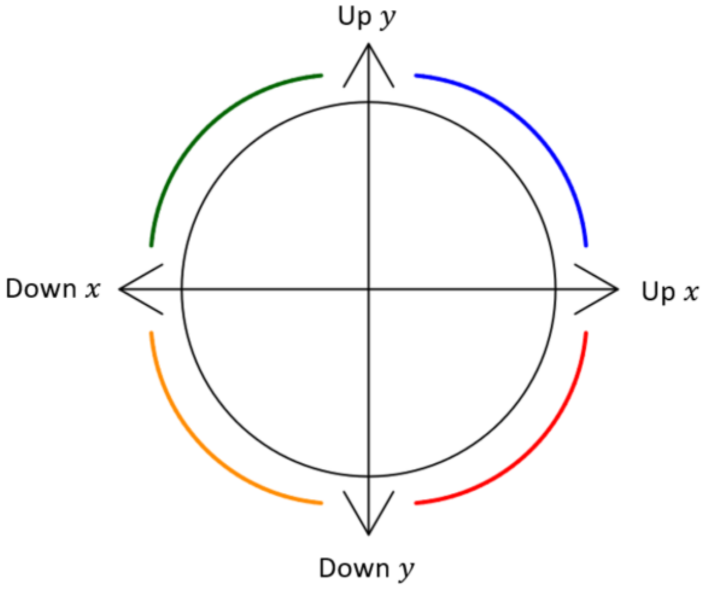




Power	Transformation
2	x^2
1	x
½	$\sqrt{x}$
0	$\log(x)$
-½	$\frac{1}{\sqrt{x}}$
-1	$\frac{1}{x}$

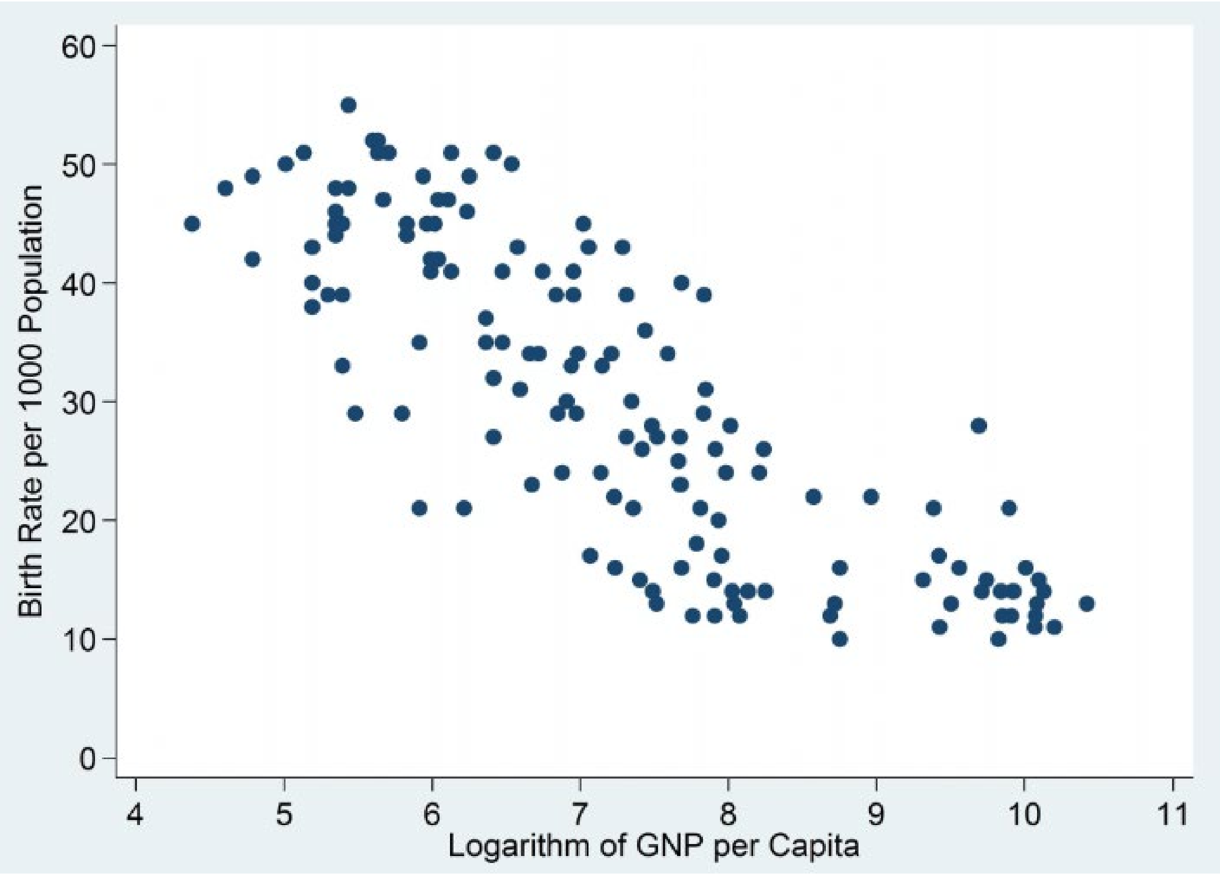
Example: GNP vs. Birth Rate

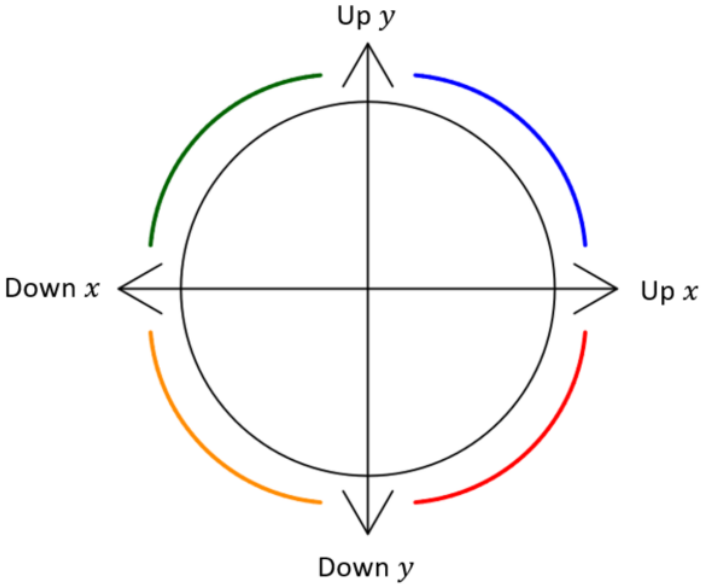




Power	Transformation
2	x^2
1	x
½	$\sqrt{x}$
0	$\log(x)$
-½	$\frac{1}{\sqrt{x}}$
-1	$\frac{1}{x}$

Example: GNP vs. Birth Rate





Power	Transformation
2	x^2
1	x
½	$\sqrt{x}$
0	$\log(x)$
-½	$\frac{1}{\sqrt{x}}$
-1	$\frac{1}{x}$

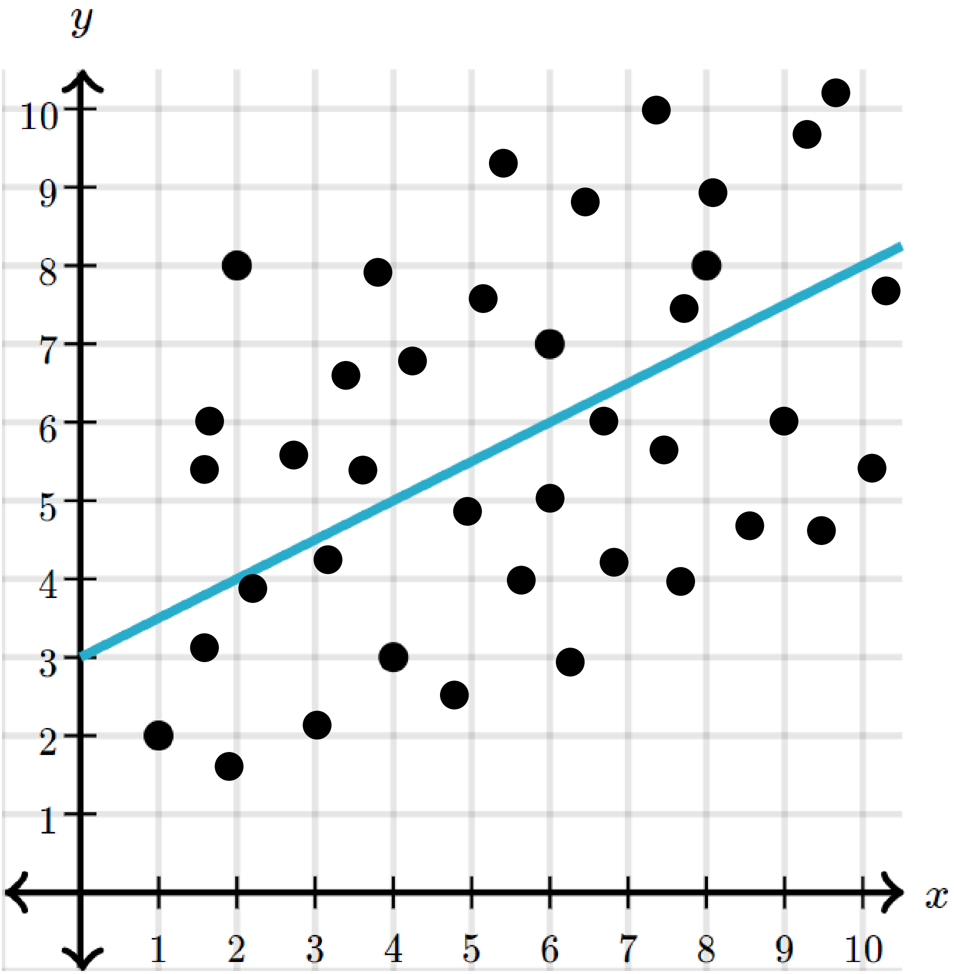
Assumptions of Linear Regression

- **Independence** of the observations
- **Linearity** of the relationship between the two variables
- **Constant variance** of the residuals
- **Normality** of the residuals

Constant Variance of the Residuals

What it means

- For a small range of x values, consider the residuals for all subjects whose observed x falls in that range. We want the variance of those residuals to be the same for all ranges of x.
- Also called homoscedasticity



Constant Variance of the Residuals

What it means

- For a small range of x values, consider the residuals for all subjects whose observed x falls in that range. We want the variance of those residuals to be the same for all ranges of x.
- Also called homoscedasticity

How to assess it

- Make a residual plot, looking for a random cloud of points
- Trends in the residual plot (especially a cone shape) reveal heteroscedasticity

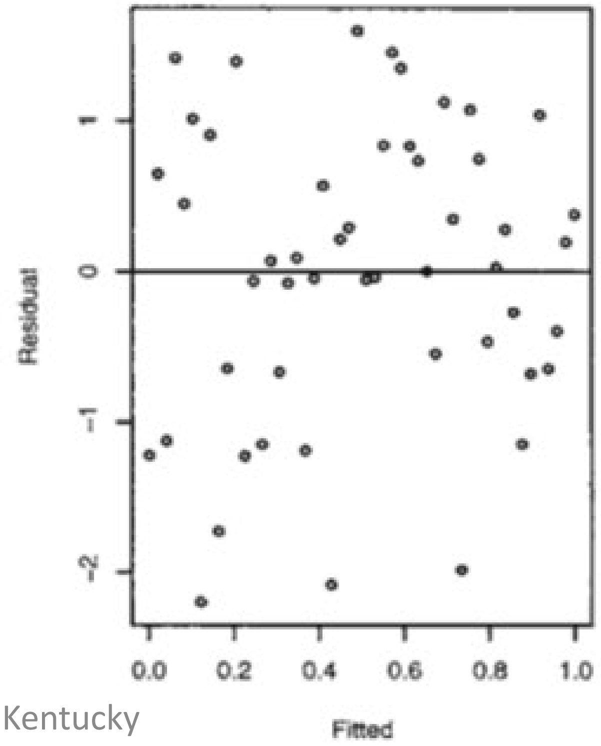
How to fix it

- Transformations of x or y can sometimes help
  - log transformation of y is particularly helpful if you see a cone shape

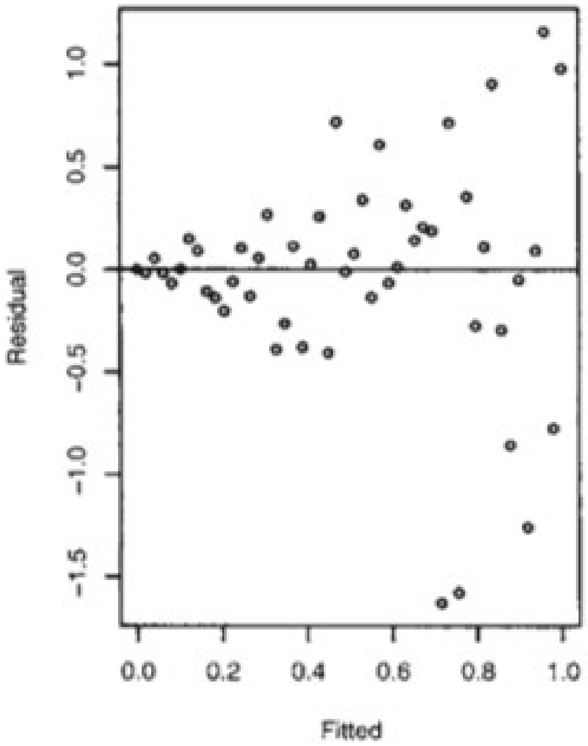
Residual Plot

- Plot fitted values on the x-axis and residuals on the y-axis.
- Want to see a random cloud of points with no trends

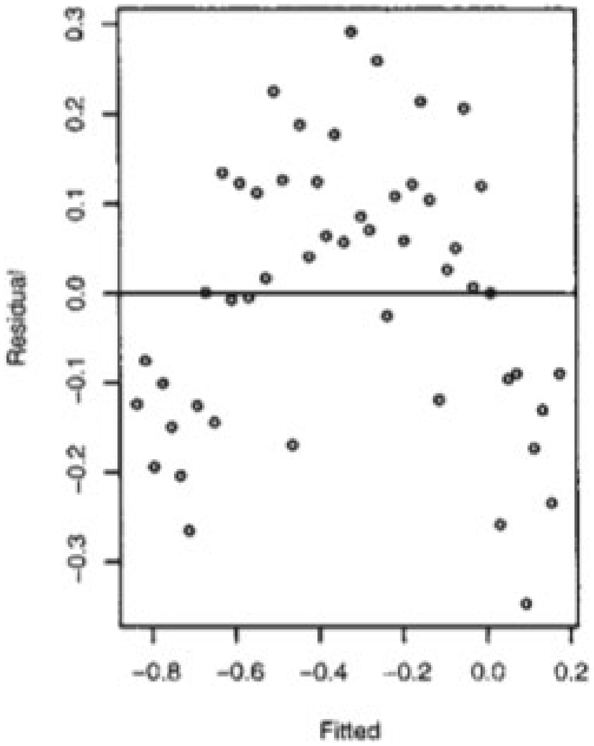
Good



Evidence of heteroscedasticity



Evidence of non-linearity



Assumptions of Linear Regression

- **Independence** of the observations
- **Linearity** of the relationship between the two variables
- **Constant variance** of the residuals
- **Normality** of the residuals

Normality of the Residuals

What it means

- The distribution of residuals follows a bell-shaped curve

How to assess it

- Make a normal QQ plot, visually check that the points fall approximately on a straight line
- If the points deviate a lot from the line, there is evidence that the residuals are not normally distributed

How to fix it

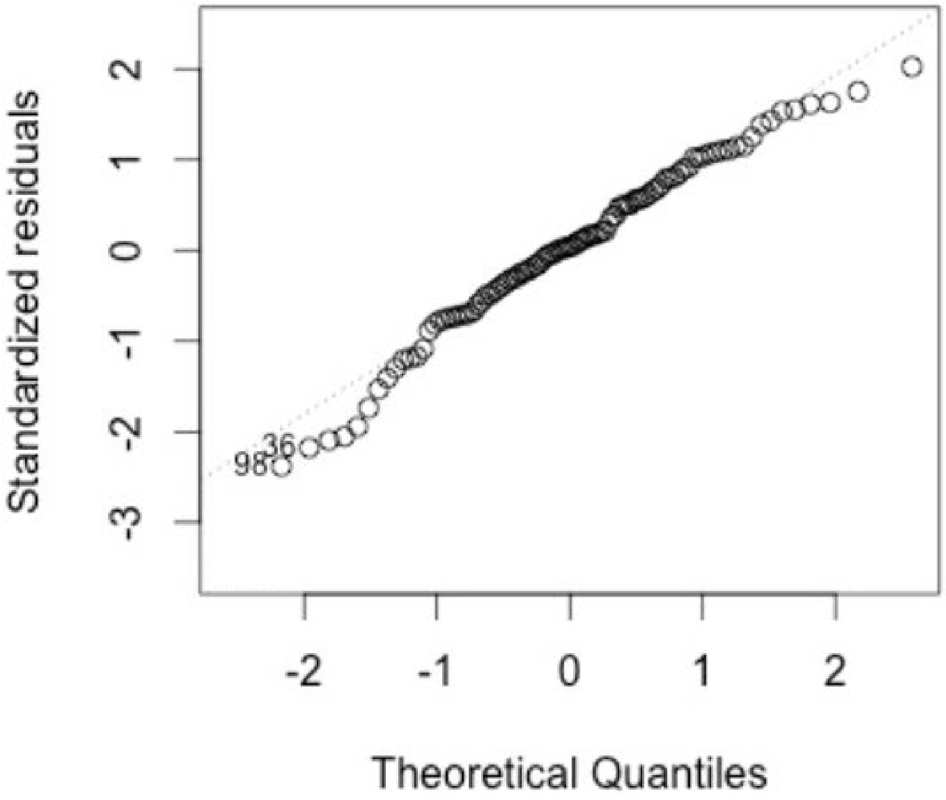
- Usually fixing non-linearity and/or heteroscedasticity will fix any issues with non-normality of residuals
- Small departures from normality actually aren't that big of a deal

Normal QQ Plot

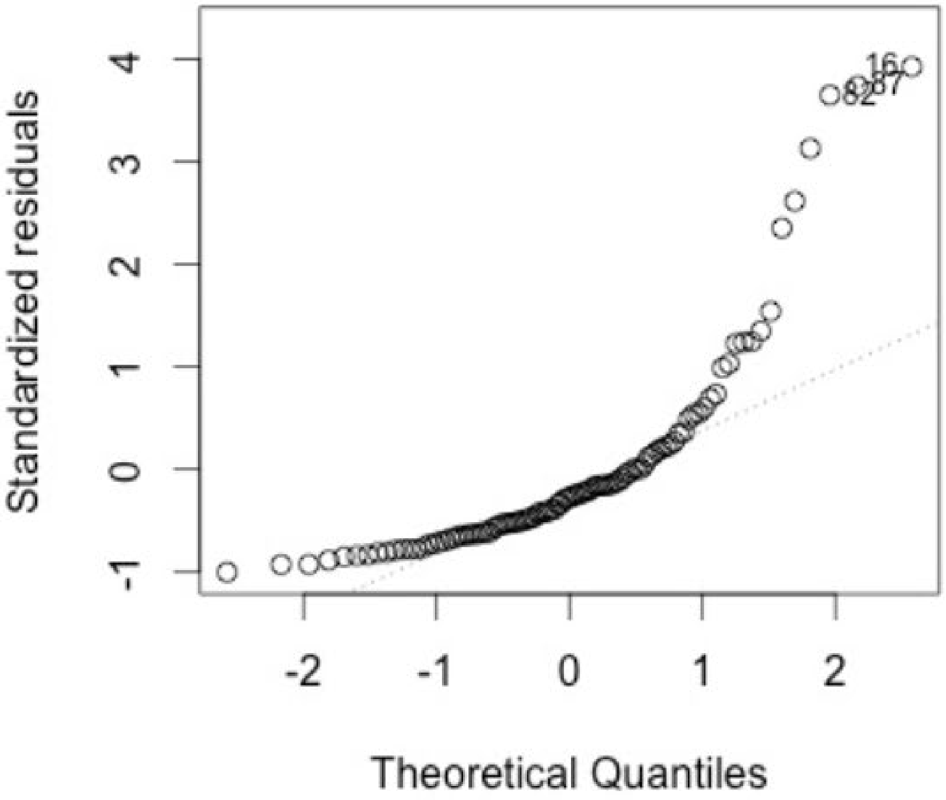
- Plot residuals on the y-axis vs. “theoretical quantiles” on the x-axis
- Remember: In a normal distribution, we expect a certain percentage of our data to fall within 1, 2, 3,… standard deviations of the mean (68-95-99.7 rule)
- Think about making a histogram of the residuals. In essence, QQ plot checks if the right percentage of residuals fall into each bucket of a normal distribution
- Want points on QQ plot to fall on a straight line

Normal QQ Plot

Good



Evidence of non-normal distribution of residuals



Assumptions of Linear Regression

- **Independence** of the observations
- **Linearity** of the relationship between the two variables
- **Constant variance** of the residuals
- **Normality** of the residuals

Assessing Model Fit: R^2

- Output for linear regression model gives us a number called R^2
  - Falls between 0 and 1
  - R^2 = correlation^2
- R^2 = "percent of variation in the outcome that can be explained by the predictor"
- Basically tells you how well the line is capturing the trend in the points
- Can use it to compare two models – higher R^2 indicates better fit

Transformations

- Pros: Can satisfy regression assumptions, improve model fit, make better predictions
- Cons: Interpretability of the model becomes more difficult

Example: Suppose we performed a log transformation of x. Regression model then relates $\log(x)$ to y. Interpretation of slope is then "on average, for every 1 unit increase in the log of (predictor variable), we expect a (slope) increase in (outcome variable)".

Example: Suppose we transform y to y^2. Regression model then relates x to y^2. Interpretation of slope is then "on average, for every 1 unit increase in (predictor variable), we expect a (slope) increase in (outcome variable) squared".

Model Building

- Building a regression model is just as much an art as it is a science
- There's no specific recipe for the perfect model
- Weigh interpretability vs. fit for the goals of your project
- Aim for a model that is **parsimonious** (as simple/interpretable as possible) while still being valid (satisfying assumptions)

FEV Data

- Information on respiratory health and exposure to secondhand smoke in 654 children

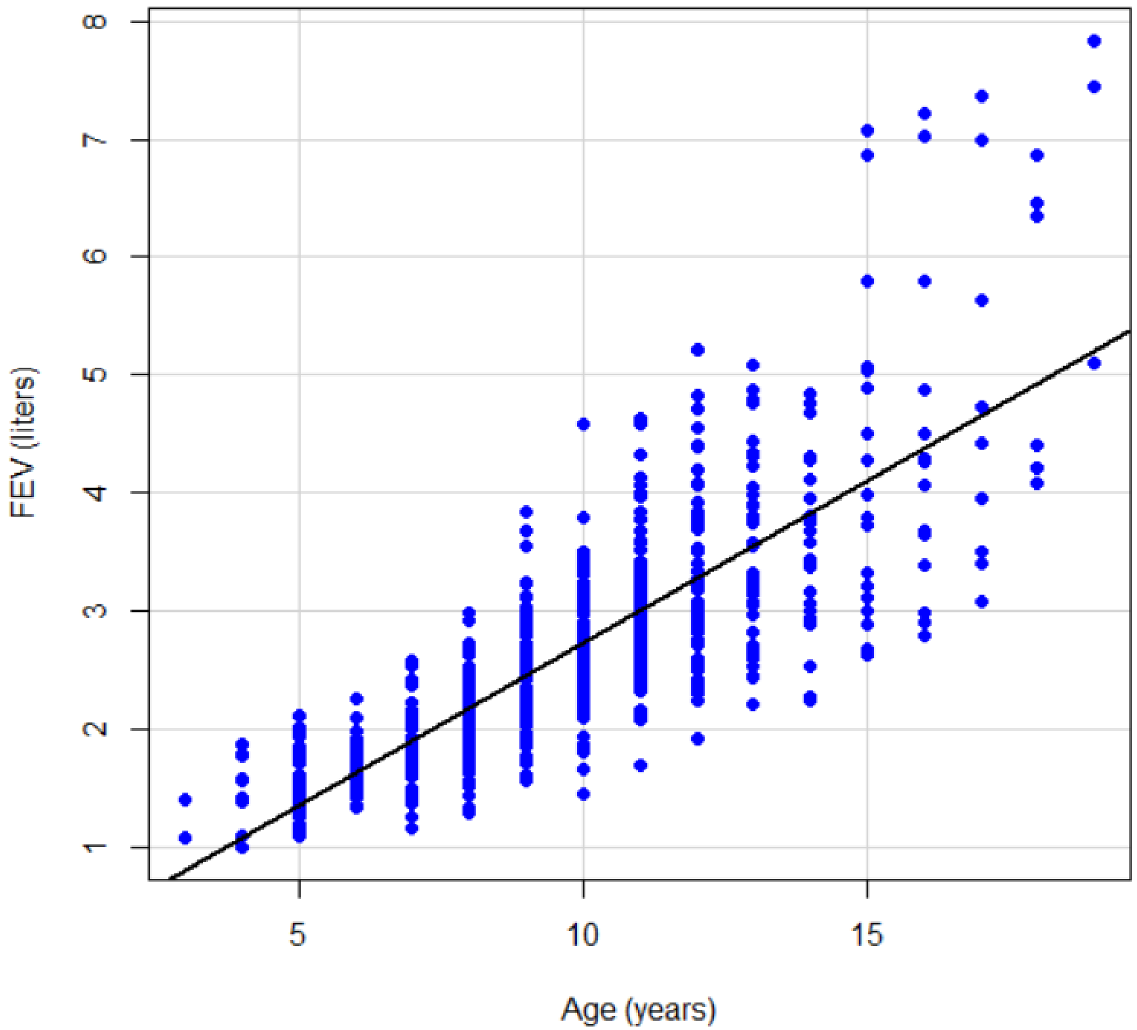
Variable	Description
Age	Age (years)
FEV	Forced expiratory volume (liters). FEV is the amount of air a person can exhale in the first second of a forceful breath.
Hgt	Height (inches)
Sex	Sex (Male, Female)
Smoke	Exposure to second-hand smoke (No, Yes)

Find the dataset (fev.xlsx) and the full data dictionary (fev Data Dictionary.pdf) in the Data Module on the Canvas site

Example: FEV vs. Age

- Fit a linear regression model relating age to forced expiratory volume (FEV). Assess whether the assumptions of the model are satisfied, and make any necessary transformations.

Example: FEV vs. Age



Example: FEV vs. Age

Linear regression model relating age and FEV:

```
Call:
lm(formula = FEV ~ Age, data = fev)

Residuals:
     Min       1Q   Median       3Q      Max
-1.59219 -0.39643 -0.03783  0.32006  2.97459

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.024841   0.086523  -0.287    0.774
Age          0.275217   0.008303  33.146   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6386 on 652 degrees of freedom
Multiple R-squared:  0.6276, Adjusted R-squared:  0.627
F-statistic:  1099 on 1 and 652 DF,  p-value: < 2.2e-16
```

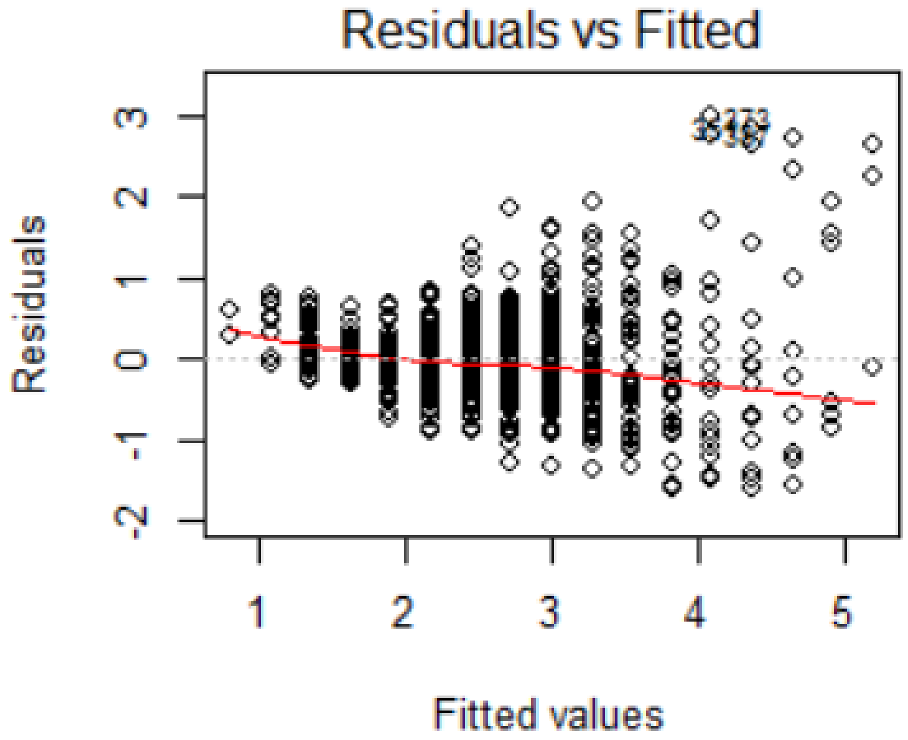
$\widehat{\text{FEV}} = -0.025 + 0.275 * \text{age}$

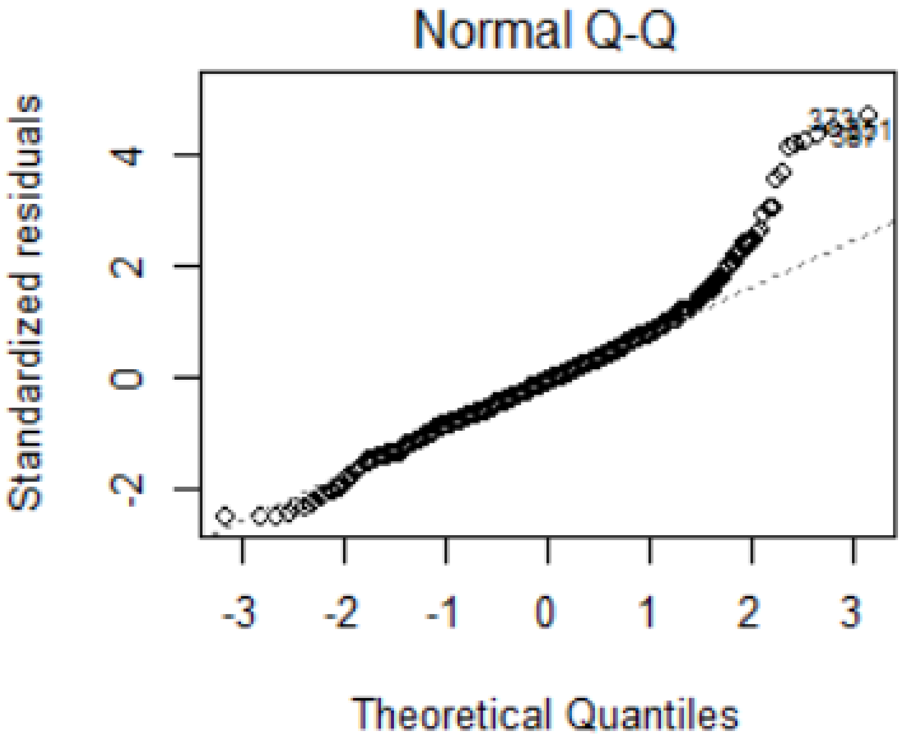
On average, every 1 year increase in age is associated with a 0.275 liter increase in FEV.

62.8% of the variation in FEV is explained by age.

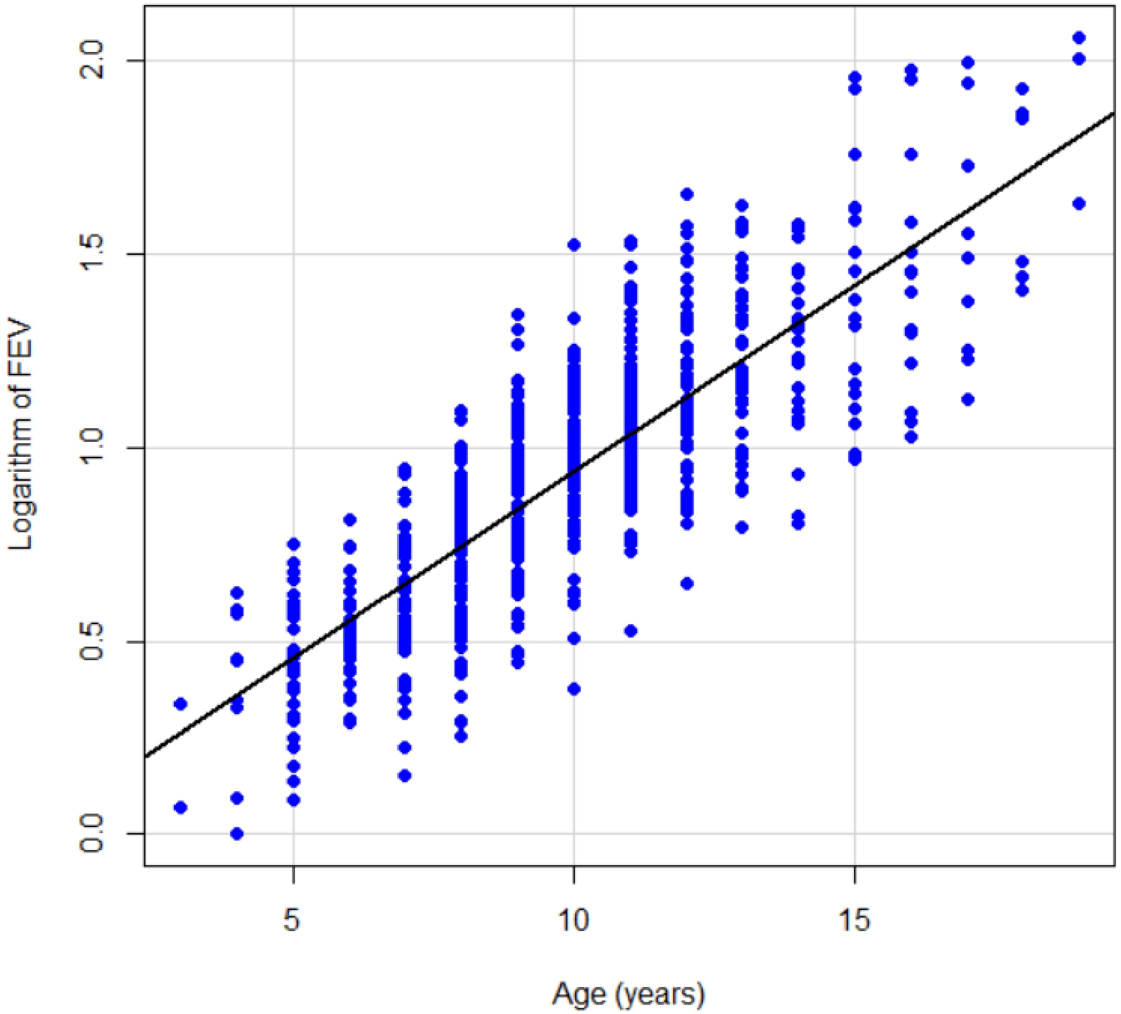
Example: FEV vs. Age

Cone shape in residuals plot usually means log transformation in outcome variable





Example: log(FEV) vs. Age



Example: log(FEV) vs. Age

Linear regression model relating age and log(FEV):

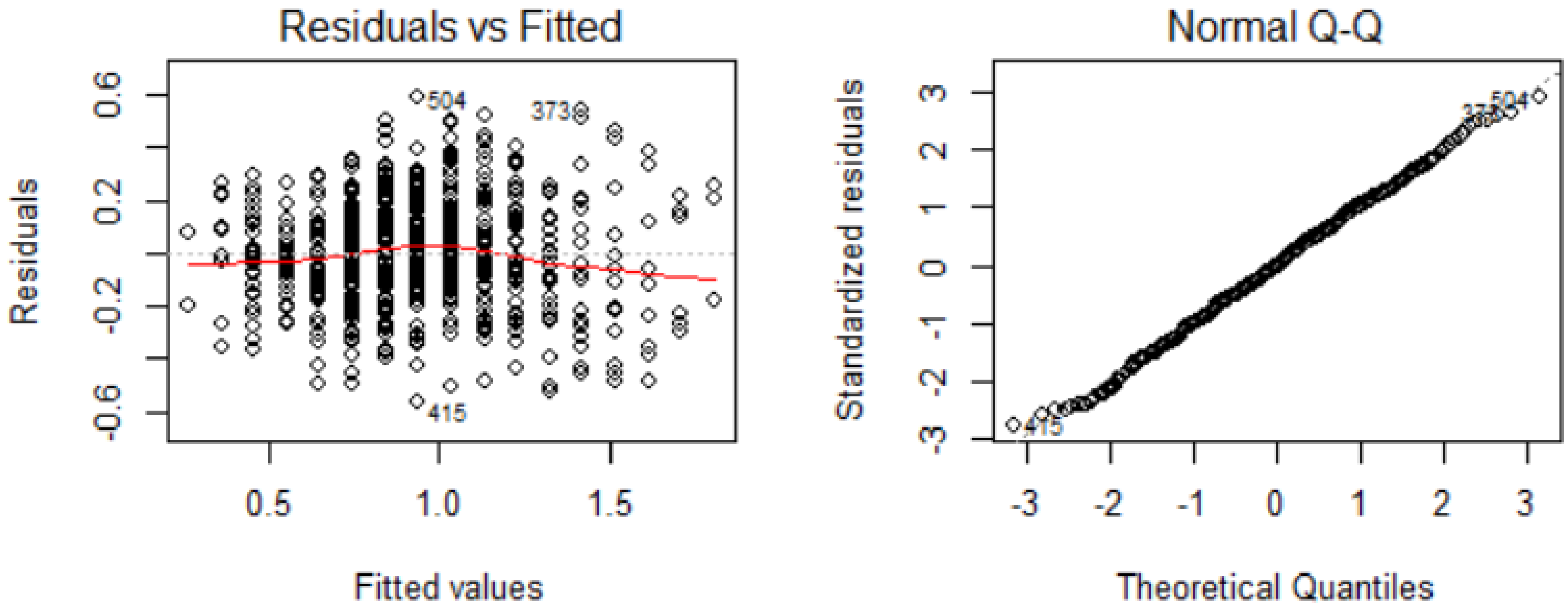
```
Call:
lm(formula = logFEV ~ Age, data = fev)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.56168 -0.12843 -0.00435  0.13679  0.58535 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.023022   0.027560  -0.835    0.404    
Age          0.096177   0.002645  36.364   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2034 on 652 degrees of freedom
Multiple R-squared:  0.6698,	Adjusted R-squared:  0.6693 
F-statistic:  1322 on 1 and 652 DF,  p-value: < 2.2e-16
```

$$\widehat{\log(\text{FEV})} = -0.023 + 0.096 * \text{age}$$

On average, every 1 year increase in age is associated with a 0.096 liter increase in log(FEV).

67.0% of the variation in log(FEV) is explained by age.

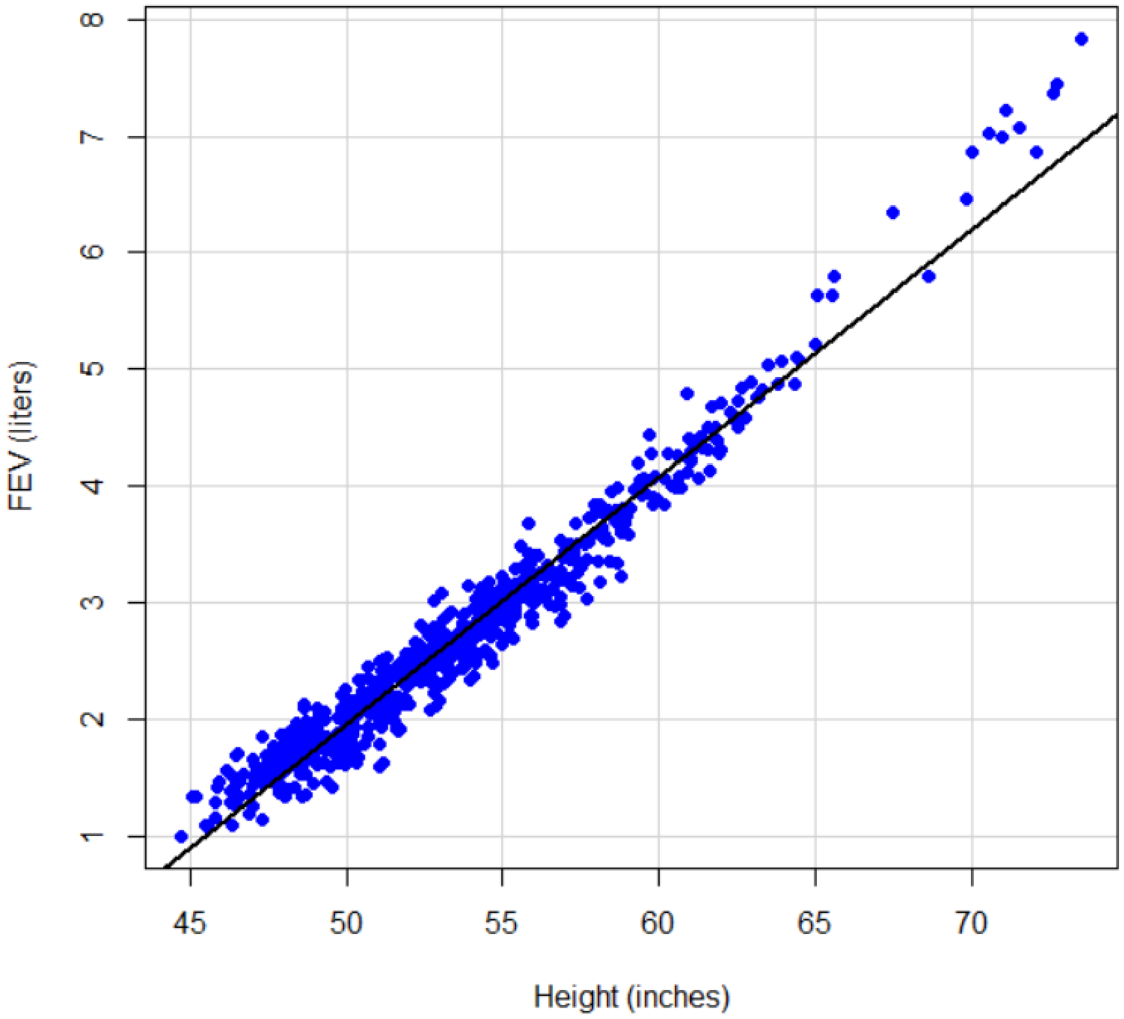
Example: log(FEV) vs. Age



Example: FEV vs. Height

- Fit a linear regression model relating height to forced expiratory volume (FEV). Assess whether the assumptions of the model are satisfied, and make any necessary transformations.

Example: FEV vs. Height



Example: FEV vs. Height

Linear regression model relating height and FEV:

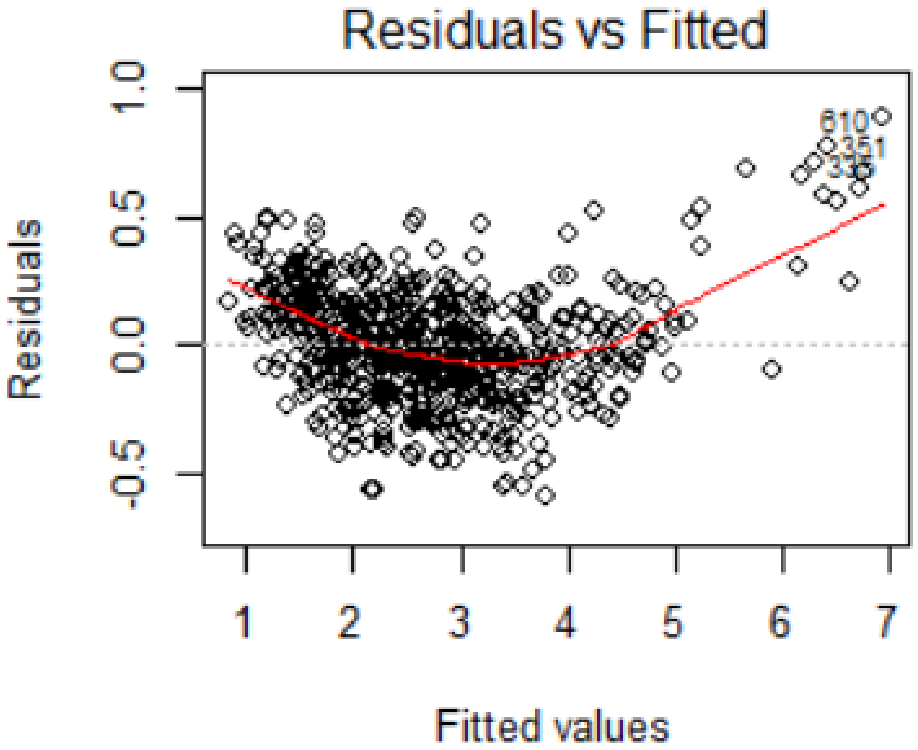
```
Call:
lm(formula = FEV ~ Hgt, data = fev)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59184 -0.13483 -0.00818  0.13203  0.88730

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.643255   0.094356   -91.6   <2e-16 ***
Hgt          0.211977   0.001753   120.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2162 on 652 degrees of freedom
Multiple R-squared:  0.9573, Adjusted R-squared:  0.9573
F-statistic: 1.462e+04 on 1 and 652 DF,  p-value: < 2.2e-16
```

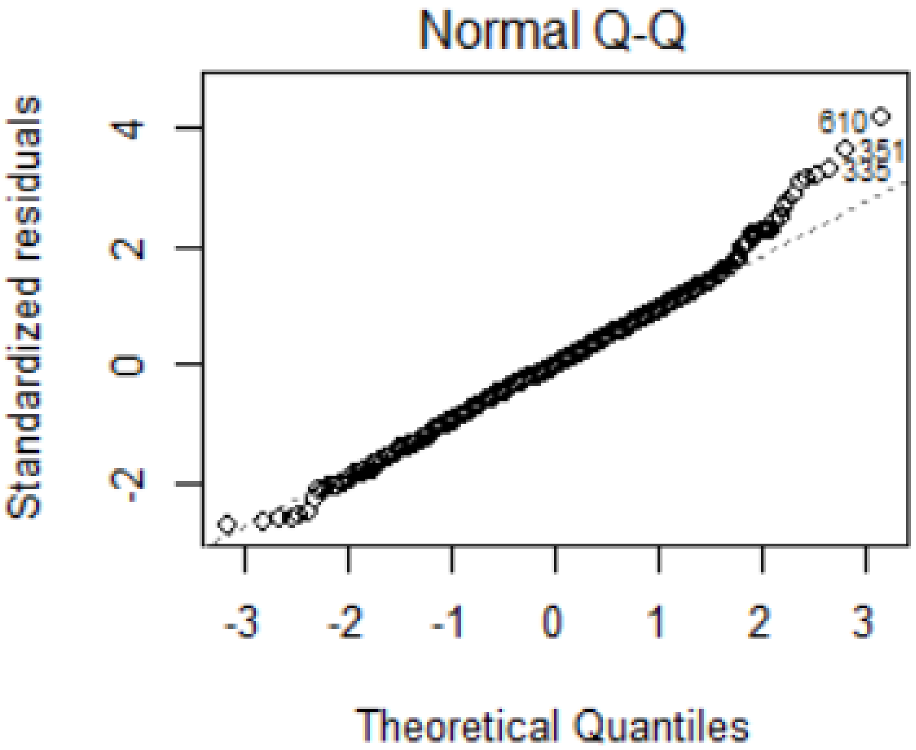
$$\widehat{\text{FEV}} = -8.643 + 0.212 * \text{height}$$

On average, every 1 inch increase in height is associated with a 0.212 liter increase in FEV.

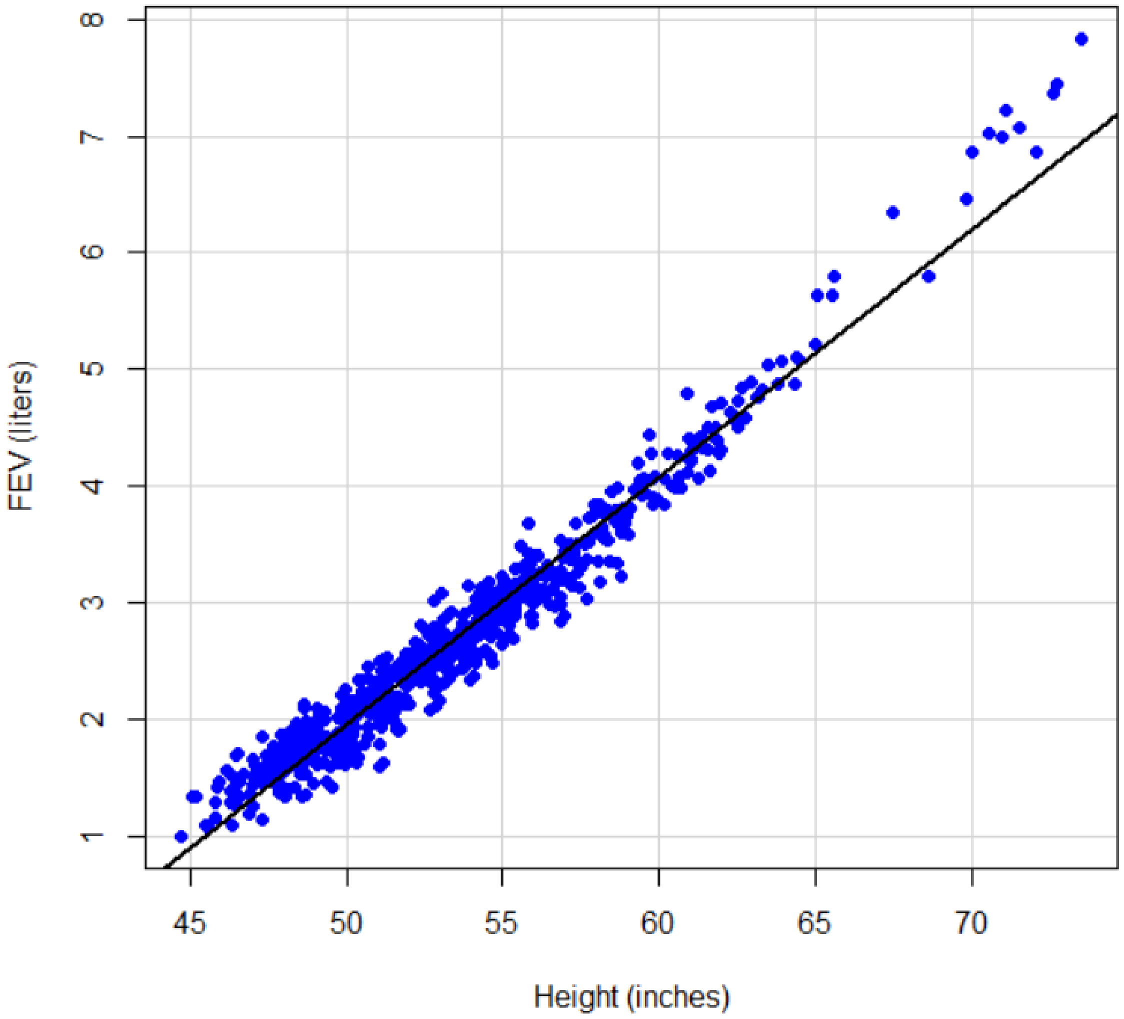
95.7% of the variation in FEV is explained by height.

Example: FEV vs. Height

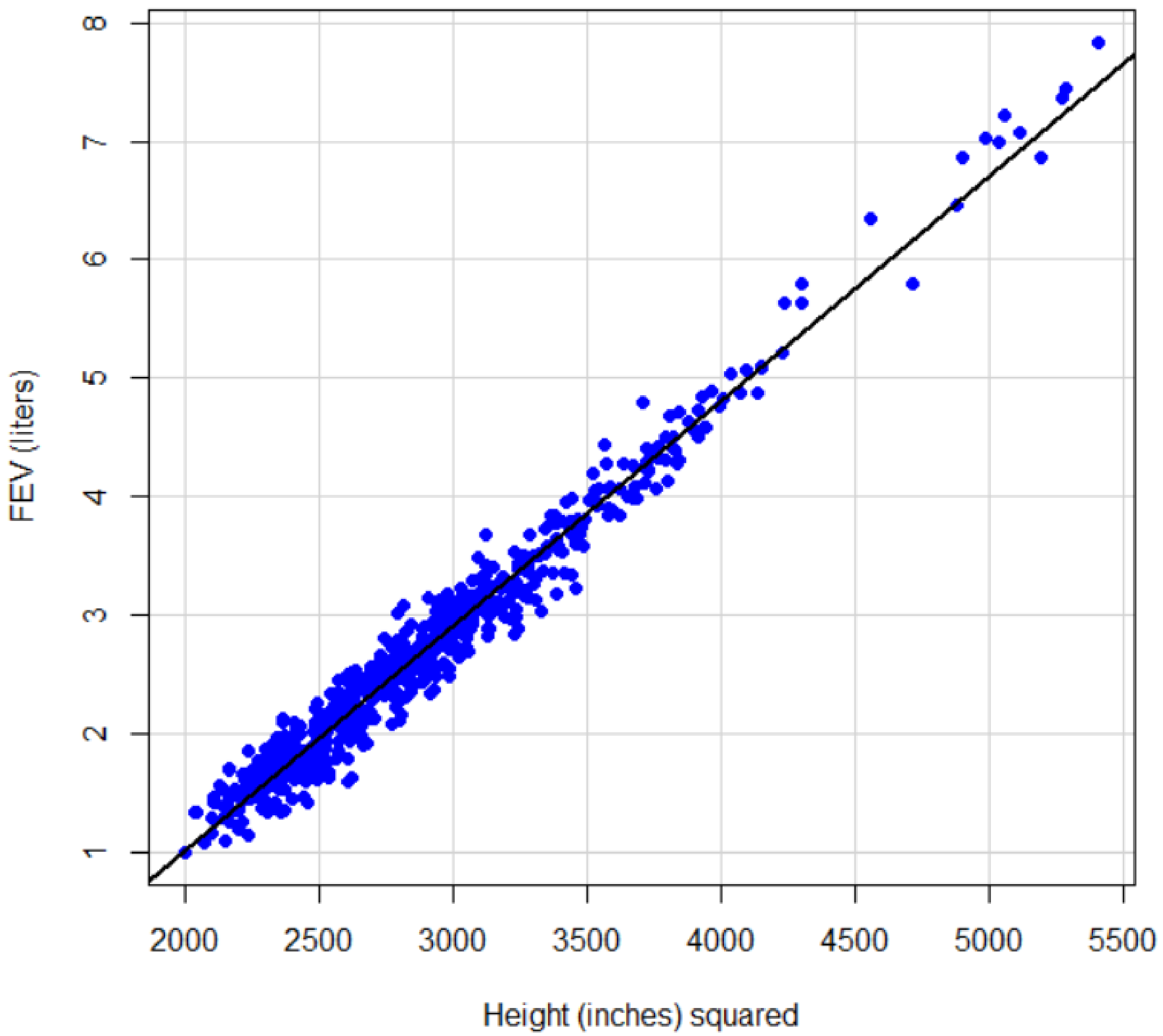




Example: FEV vs. Height



Example: FEV vs. Height^2



Example: FEV vs. Height^2

Linear regression model relating age and log(FEV):

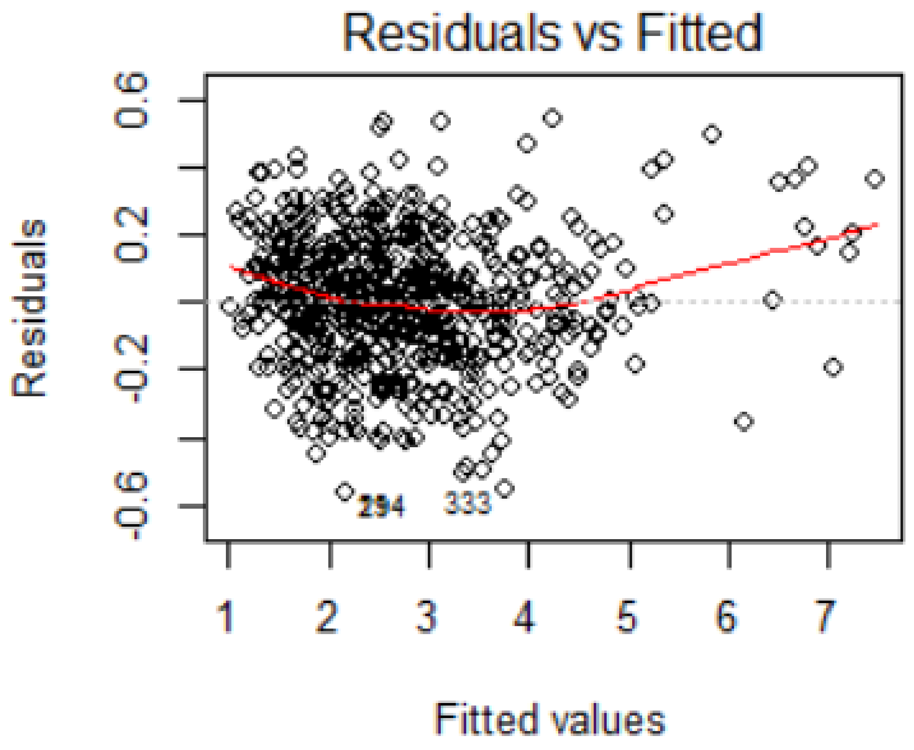
```
Call:
lm(formula = FEV ~ Hgt2, data = fev)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56360 -0.12192  0.00504  0.12499  0.53622

Coefficients:
               Estimate  Std. Error t value Pr(>|t|)
(Intercept) -2.76441875  0.04023454  -68.71   <2e-16 ***
Hgt2         0.00189324  0.00001365  138.71   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1894 on 652 degrees of freedom
Multiple R-squared:  0.9672, Adjusted R-squared:  0.9672
F-statistic: 1.924e+04 on 1 and 652 DF,  p-value: < 2.2e-16
```

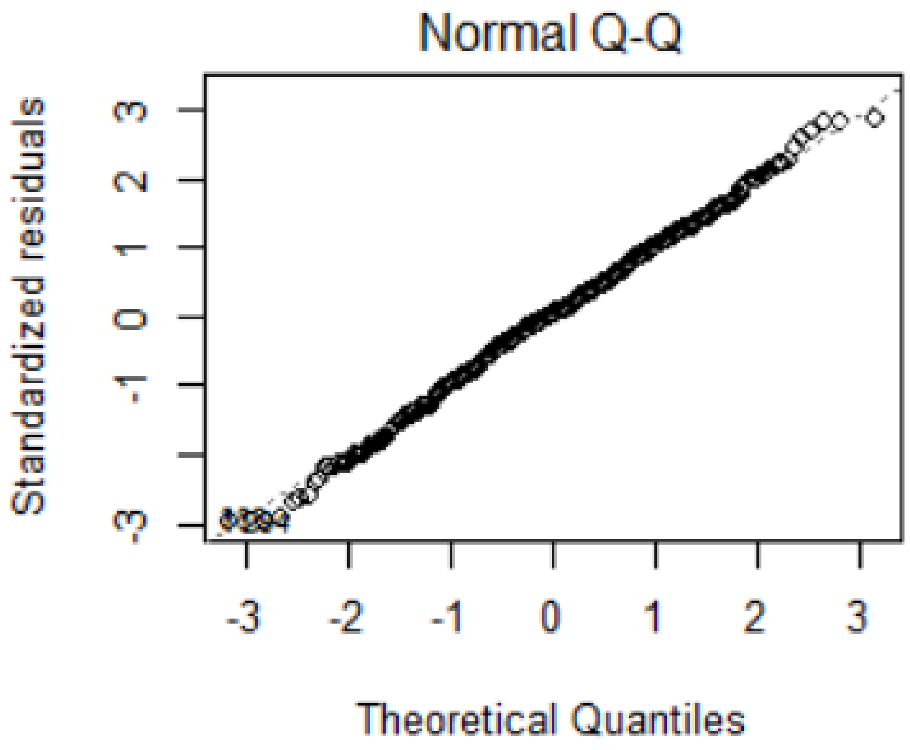
$\widehat{\text{FEV}} = -2.764 + 0.002 * \text{height}^2$

On average, every 1 inch^2 increase in height^2 is associated with a 0.002 liter increase in FEV.

96.7% of the variation in FEV is explained by height^2.

Example: FEV vs. Height^2





Important Points

- Four assumptions of linear regression
- How to check each assumption
- Transforming data to meet assumptions
  - Interpretation of model output
- Trade-off between model fit and interpretability