Variable transformations

Prof. Maria Tackett

Click here for PDF of slides

Topics

- Log transformation on the response
- Log transformation on the predictor

Respiratory Rate vs. Age

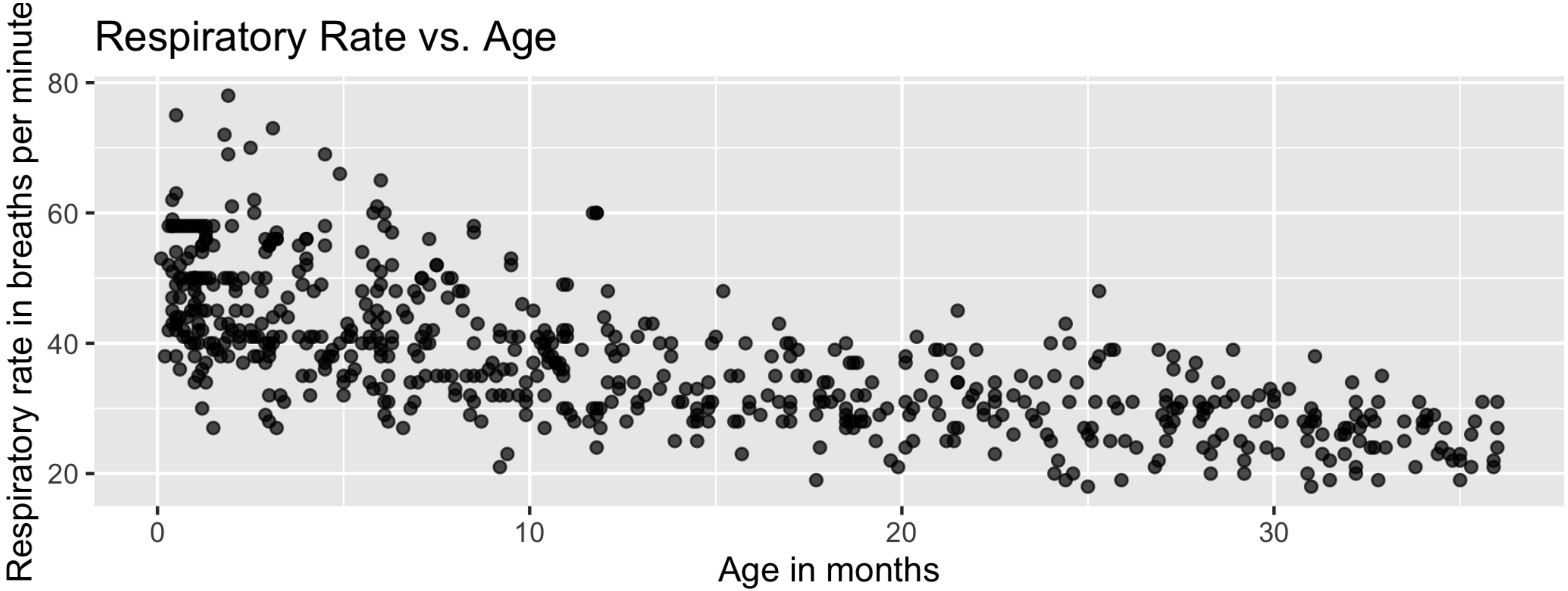
- A high respiratory rate can potentially indicate a respiratory infection in children. In order to determine what indicates a "high" rate, we first want to understand the relationship between a child's age and their respiratory rate.

- The data contain the respiratory rate for 618 children ages 15 days to 3 years.

- **Variables**:

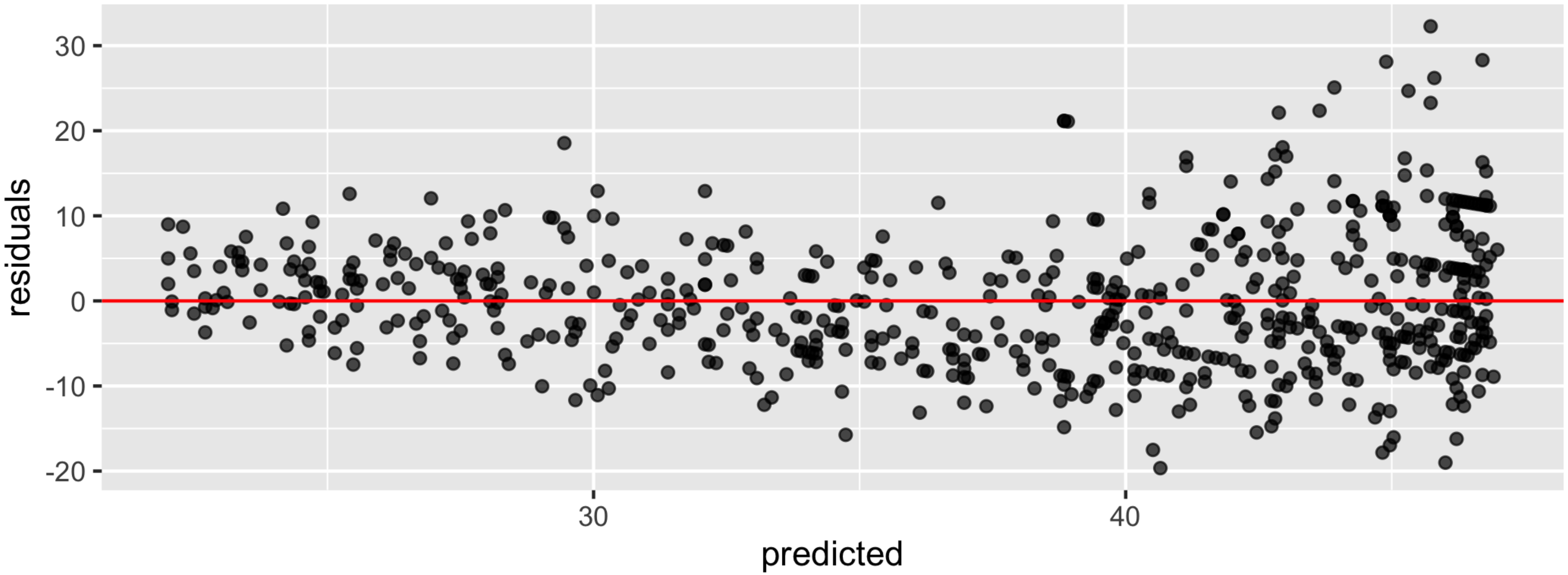
  - **`Age`**: age in months
  - **`Rate`**: respiratory rate (breaths per minute)

Rate vs. Age


Respiratory Rate vs. Age
Respiratory rate in breaths per minute
80
60
40
20
0
10
20
30
Age in months

Rate vs. Age

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	47.052	0.504	93.317	0	46.062	48.042
Age	-0.696	0.029	-23.684	0	-0.753	-0.638



Log transformation on the response

Need to transform Y

- Typically, a "fan-shaped" residual plot indicates the need for a transformation of the response variable y
  - log(Y) is the most straightforward to interpret
- When building a model:
  - Choose a transformation and build the model on the transformed data
  - Reassess the residual plots
  - If the residuals plots did not sufficiently improve, try a new transformation!

Log transformation on Y

- If we apply a log transformation to the response variable, we want to estimate the parameters for the statistical model

$$\log(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2_\epsilon)$$

- The regression equation is

$$\widehat{\log(Y)} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Log transformation on Y

- We want to interpret the model in terms of Y not $\log(Y)$, so we write all interpretations in terms of

$$\hat{Y} = \exp\{\hat{\beta}_0 + \hat{\beta}_1 X\} = \exp\{\hat{\beta}_0\}\exp\{\hat{\beta}_1 X\}$$

Mean and logs

Suppose we have a set of values

```
x <- c(3, 5, 6, 8, 10, 14, 19)
```

Let's calculate $\overline{\log(x)}$

```
log_x <- log(x)
mean(log_x)
```

```
## [1] 2.066476
```

Let's calculate $\log(\bar{x})$

```
xbar <- mean(x)
log(xbar)
```

```
## [1] 2.228477
```

Median and logs

```r
x <- c(3, 5, 6, 8, 10, 14, 19)
```

Let's calculate $\text{Median}(\log(x))$

```r
log_x <- log(x)
median(log_x)
```

```
## [1] 2.079442
```

Let's calculate $\log(\text{Median}(x))$

```r
median_x <- median(x)
log(median_x)
```

```
## [1] 2.079442
```

Mean, Median, and log

$$\overline{\log(x)} \neq \log(\bar{x})$$

```
mean(log_x) == log(xbar)
```

```
## [1] FALSE
```

$$\text{Median}(\log(x)) = \log(\text{Median}(x))$$

```
median(log_x) == log(median_x)
```

```
## [1] TRUE
```

Mean and median of $\log(Y)$

- Recall that $y = \beta_0 + \beta_1 X_i$ is the **mean** value of the response at the given value of the predictor x_i. This doesn't hold when we log-transform the response variable.

- Mathematically, the mean of the logged values is **not** necessarily equal to the log of the mean value. Therefore at a given value of x

$$\exp\{\text{Mean}(\log(y))\} \neq \text{Mean}(y)$$

$$\Rightarrow \exp\{\beta_0 + \beta_1 x\} \neq \text{Mean}(y)$$

STA 210

Mean and median of $\log(y)$

- However, the median of the logged values **is** equal to the log of the median value. Therefore,

$$\exp\{\text{Median}(\log(y))\} = \text{Median}(y)$$

- If the distribution of $\log(y)$ is symmetric about the regression line, for a given value x_i, we can expect $Mean(y)$ and $Median(y)$ to be approximately equal.

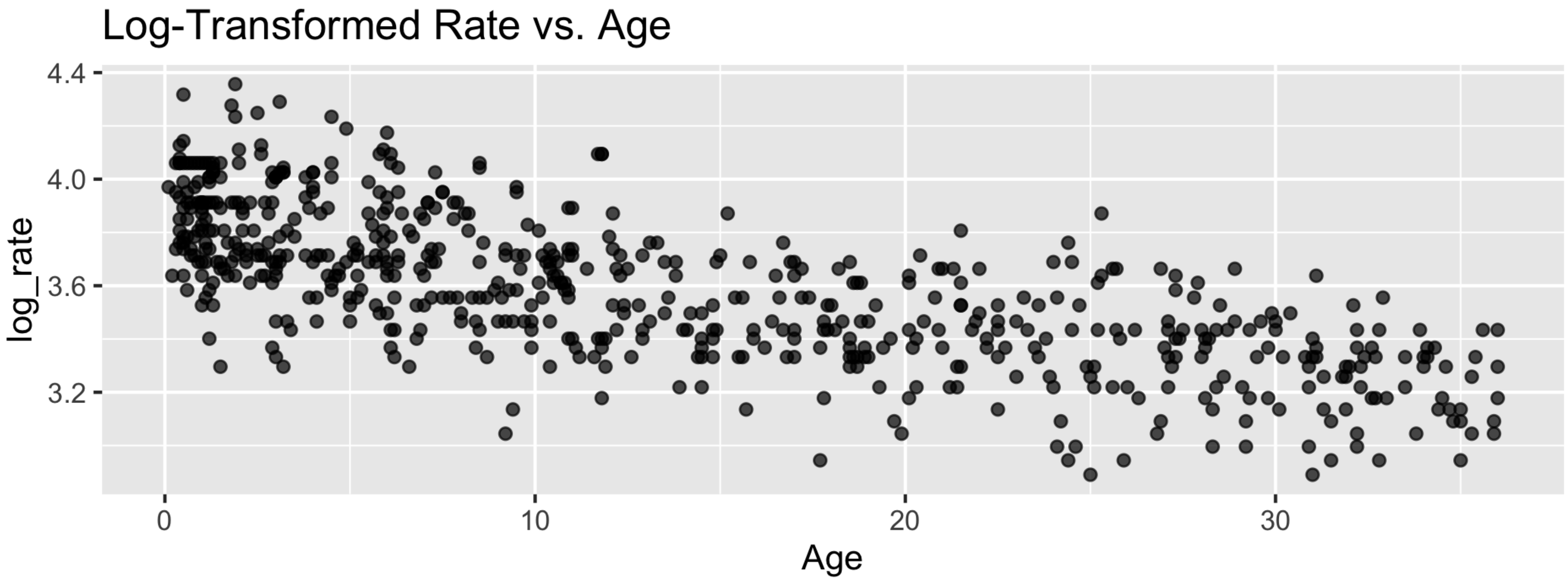
Interpretation with log-transformed y

- Given the previous facts, if $\widehat{\log(Y)} = \hat{\beta}_0 + \hat{\beta}_1 X$, then

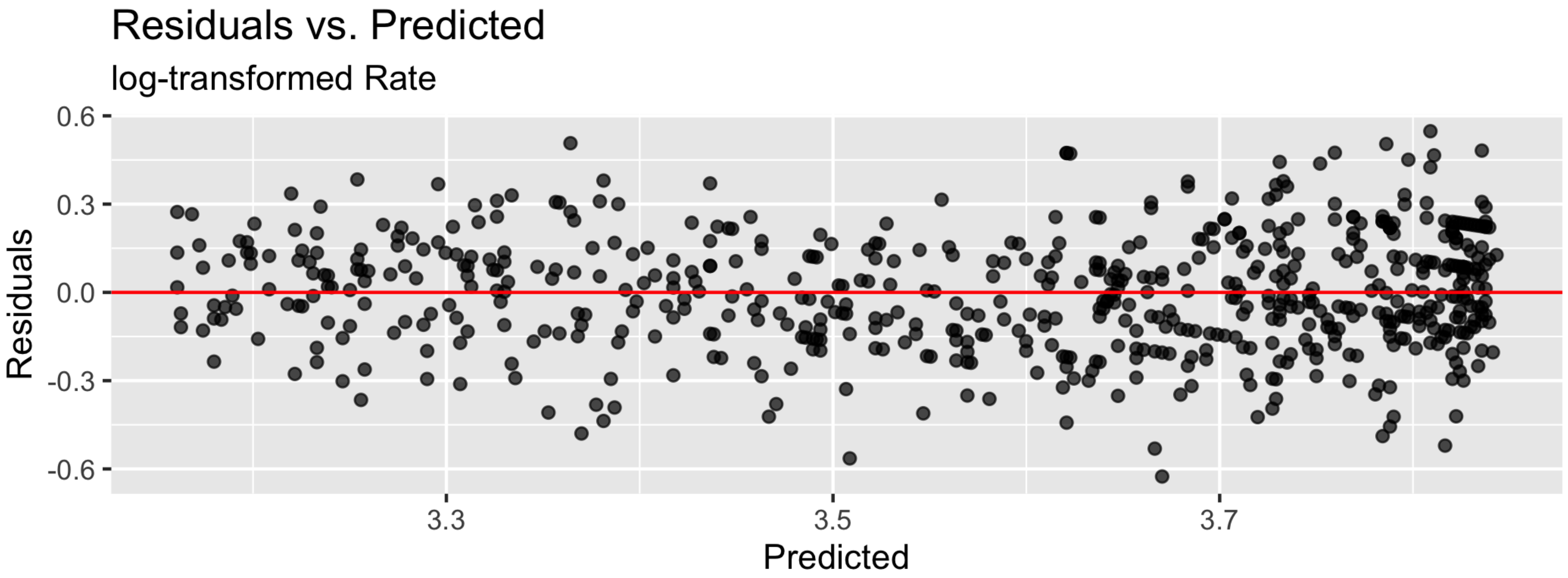
$$\text{Median}(\hat{Y}) = \exp\{\hat{\beta}_0\}\exp\{\hat{\beta}_1 X\}$$

- **Intercept:** When $X = 0$, the median of Y is expected to be $\exp\{\hat{\beta}_0\}$
- **Slope:** For every one unit increase in X, the median of Y is expected to multiply by a factor of $\exp\{\hat{\beta}_1\}$

log(Rate) vs. Age



log(Rate) vs. Age



log(Rate) vs. Age

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	3.845	0.013	304.500	0	3.82	3.870
Age	-0.019	0.001	-25.839	0	-0.02	-0.018

Intercept: The median respiratory rate for a new born child is expected to be 46.759 (exp{3.845}) breaths per minute.

Slope: For each additional month in a child's age, the respiratory rate is expected to multiply by a factor of 0.981 (exp{-0.019}).

Confidence interval for β_j

- The confidence interval for the coefficient of X describing its relationship with $\log(Y)$ is

$$\hat{\beta}_j \pm t^* SE(\widehat{\beta_j})$$

- The confidence interval for the coefficient of X describing its relationship with Y is

$$\exp\left\{\hat{\beta}_j \pm t^* SE(\widehat{\beta_j})\right\}$$

- Note: t^* is calculated from the t disribution with $n - p - 1$ df

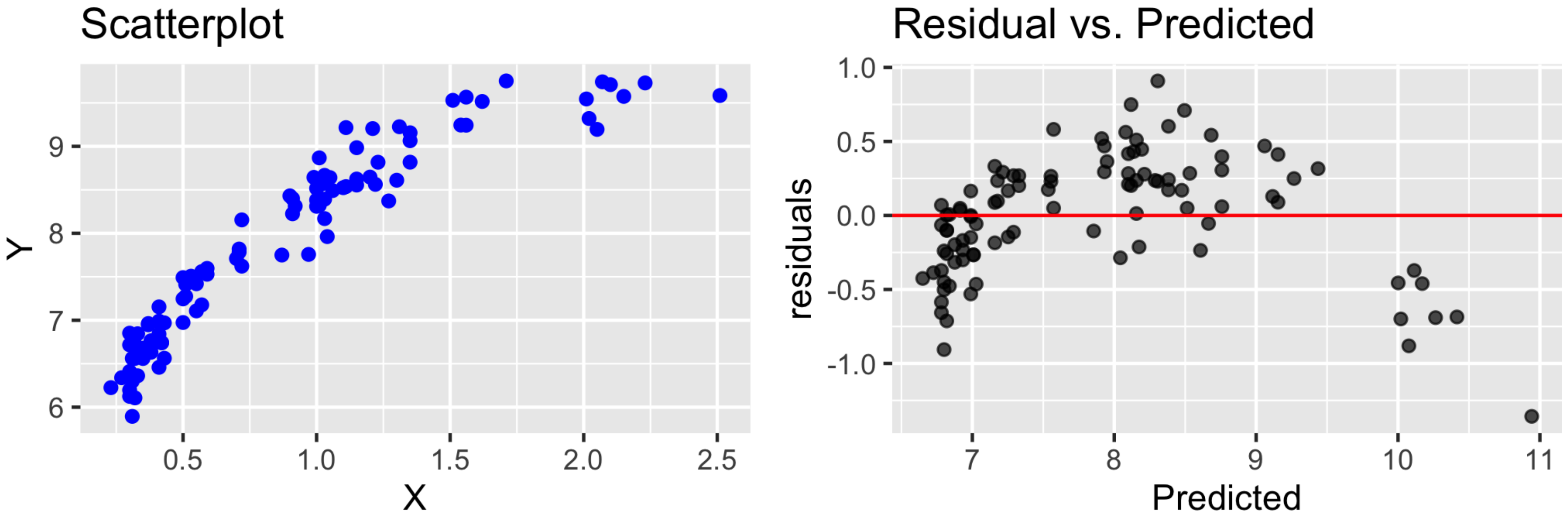
Coefficient of `Age`

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	3.845	0.013	304.500	0	3.82	3.870
Age	-0.019	0.001	-25.839	0	-0.02	-0.018

We are 95% confident that for each additional month in a child's age, the respiratory rate multiplies by a factor of 0.98 to 0.982 (exp{-0.02} to exp{-0.018}).

Log transformation on the predictor

Log Transformation on X



Try a transformation on X if the scatterplot shows some curvature but the variance is constant for all values of X

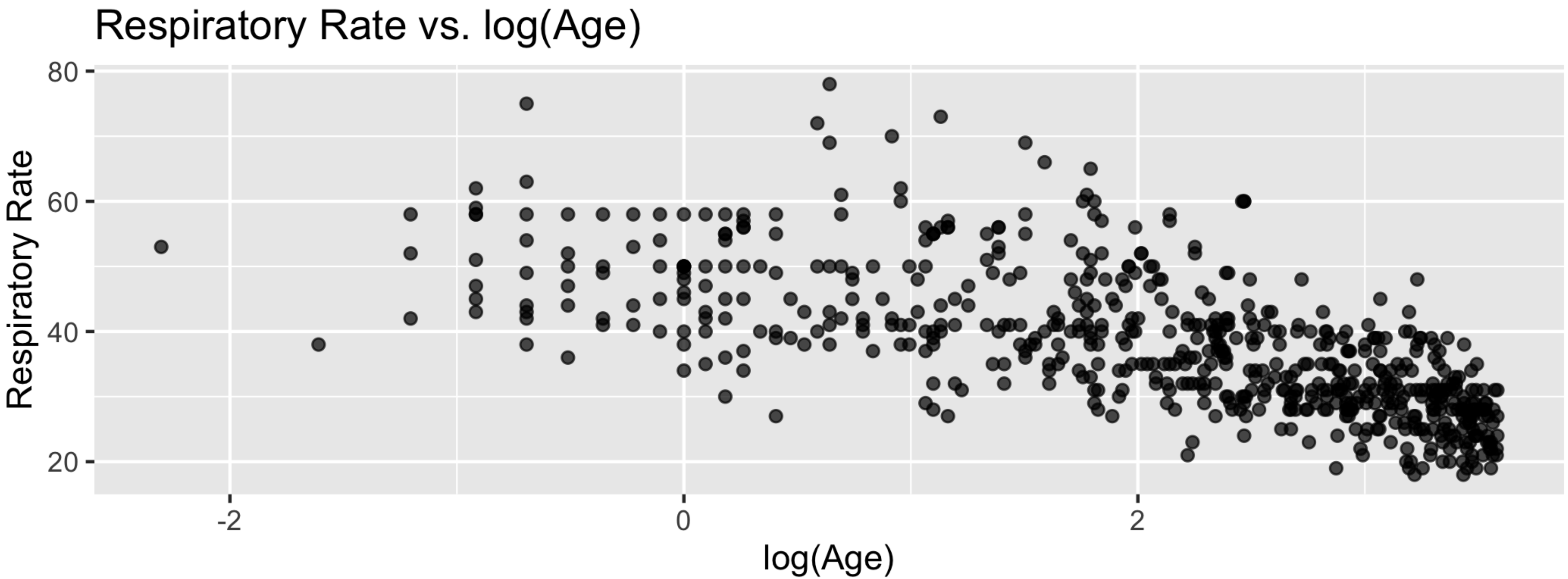
Model with Transformation on X

Suppose we have the following regression equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \log(X)$$

- **Intercept:** When $X = 1$ $(\log(X) = 0)$, Y is expected to be $\hat{\beta}_0$ (i.e. the mean of y is $\hat{\beta}_0$)
- **Slope:** When X is multiplied by a factor of $\mathbf{C}$, the mean of Y is expected to increase by $\boldsymbol{\hat{\beta}_1 \log(C)}$ units
  - *Example*: when X is multiplied by a factor of 2, Y is expected to increase by $\boldsymbol{\hat{\beta}_1 \log(2)}$ units

Rate vs. log(Age)



STA 210

Rate vs. log(Age)

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	50.135	0.632	79.330	0	48.893	51.376
log_age	-5.982	0.263	-22.781	0	-6.498	-5.467

Intercept: The expected (mean) respiratory rate for children who are 1 month old (log(1) = 0) is 50.135 breaths per minute.

Slope: If a child's age doubles, we expect their respiratory rate to decrease by 4.146 (5.982*log(2)) breaths per minute.

See Log Transformations in Linear Regression for more details about interpreting regression models with log-transformed variables.

Recap

- Log transformation on the response
- Log transformation on the predictor