Multiple Linear Regression

Interactions & Transformations

Prof. Maria Tackett

02.17.20

STA 210

[Click for PDF of slides]

Announcements

- Team Feedback #1 **due Wed, Feb 19 at 11:59p**
  - Check for email from Teammates
  - Please provide honest and constructive feedback. This team feedback will be graded for completion.
- HW 03 **due Mon, Feb 24 at 11:59p**

Today's Agenda

- Interactions
- Log Transformations

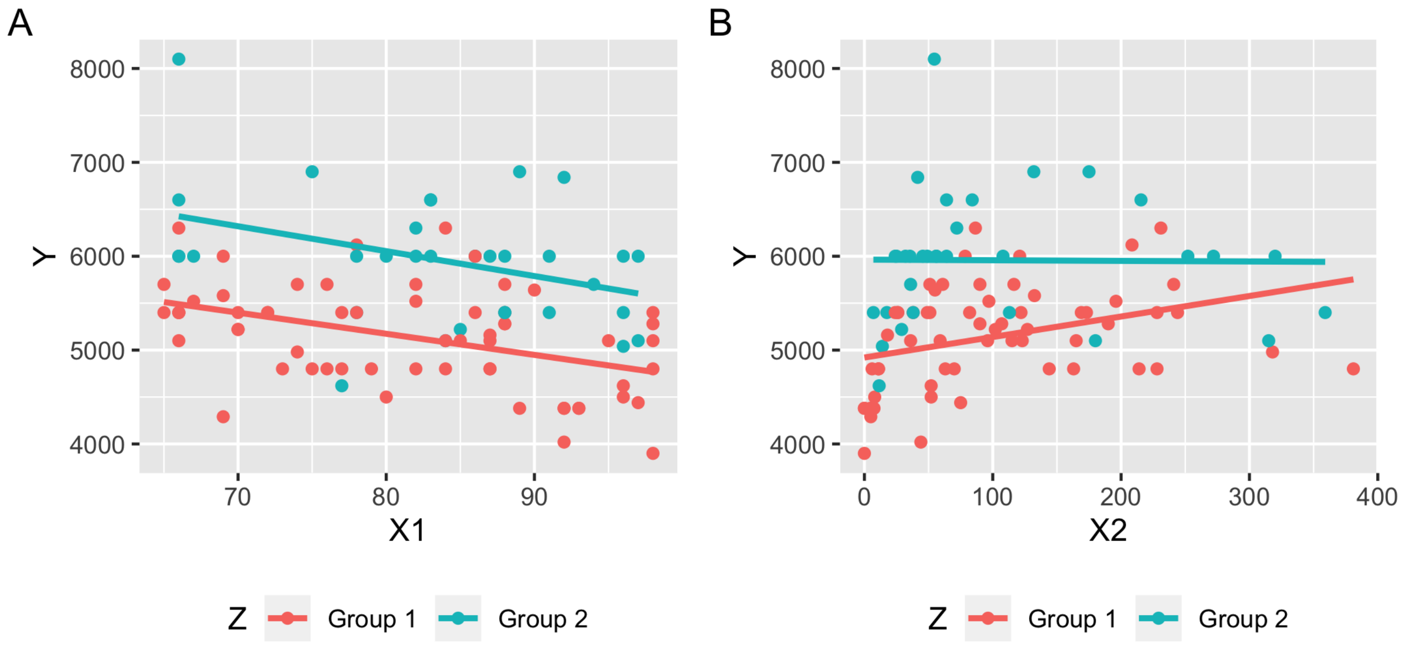
Interactions

Interaction Terms

- **Case:** Relationship of the predictor variable with the response depends on the value of another predictor variable
  - This is an **interaction effect**
- Create a new interaction variable that is one predictor variable times the other in the interaction
- **Good Practice:** When including an interaction term, also *include the associated <u>main effects</u>* (each predictor variable on its own) even if their coefficients are not statistically significant

STA 210

Checking for interactions in the EDA



The data

Predictors

- `verified_income`: Whether borrower's income source and amount have been verified (`Not Verified`, `Source Verified`, `Verified`)
- `debt_to_income`: Debt-to-income ratio, i.e. the percentage of a borrower's total debt divided by their total income
- `bankruptcy`: Indicator of whether borrower has had a bankruptcy in the past (`0`: No, `1`: Yes)
- `term`: Length of the loan in months
- `credit_util`: What fraction of total credit a borrower is utilizing, i.e. total credit utilizied divided by total credit limit

Response

- `interest_rate`: Interest rate for the loan

```
## Observations: 9,974
```

Add interaction term

```
model_w_int <- lm(interest_rate ~ verified_income + debt_inc_cent +
                    bankruptcy + term_cent + credit_util_cent +
                    debt_inc_cent * verified_income,
                  data = loans)
```

term	estimate	std.error	statistic	p.value
(Intercept)	11.298	0.074	151.764	0.000
verified_incomeSource Verified	1.094	0.100	10.940	0.000
verified_incomeVerified	2.704	0.119	22.730	0.000
debt_inc_cent	0.032	0.005	6.527	0.000
bankruptcy1	0.525	0.133	3.954	0.000
term_cent	0.154	0.004	38.764	0.000
credit_util_cent	4.841	0.163	29.689	0.000
verified_incomeSource Verified:debt_inc_cent	-0.009	0.007	-1.243	0.214
verified_incomeVerified:debt_inc_cent	-0.019	0.007	-2.699	0.007

Understanding interactions

- **Different intercept:** `verified_incomeVerified` = 2.704
- **Different slope** `verified_incomeVerified:debt_inc_cent` = -0.019

Log Transformations

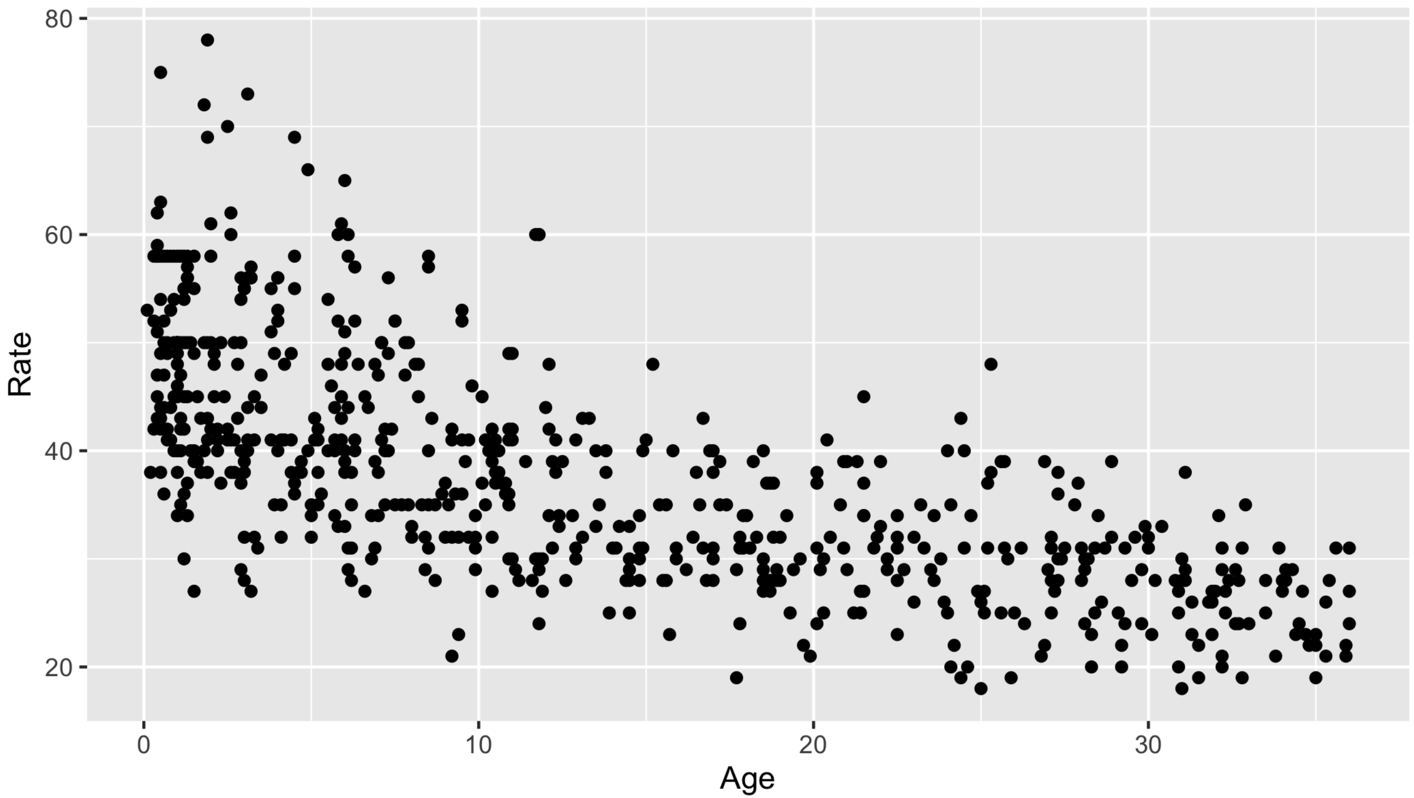
Respiratory Rate vs. Age

- 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)

STA 210

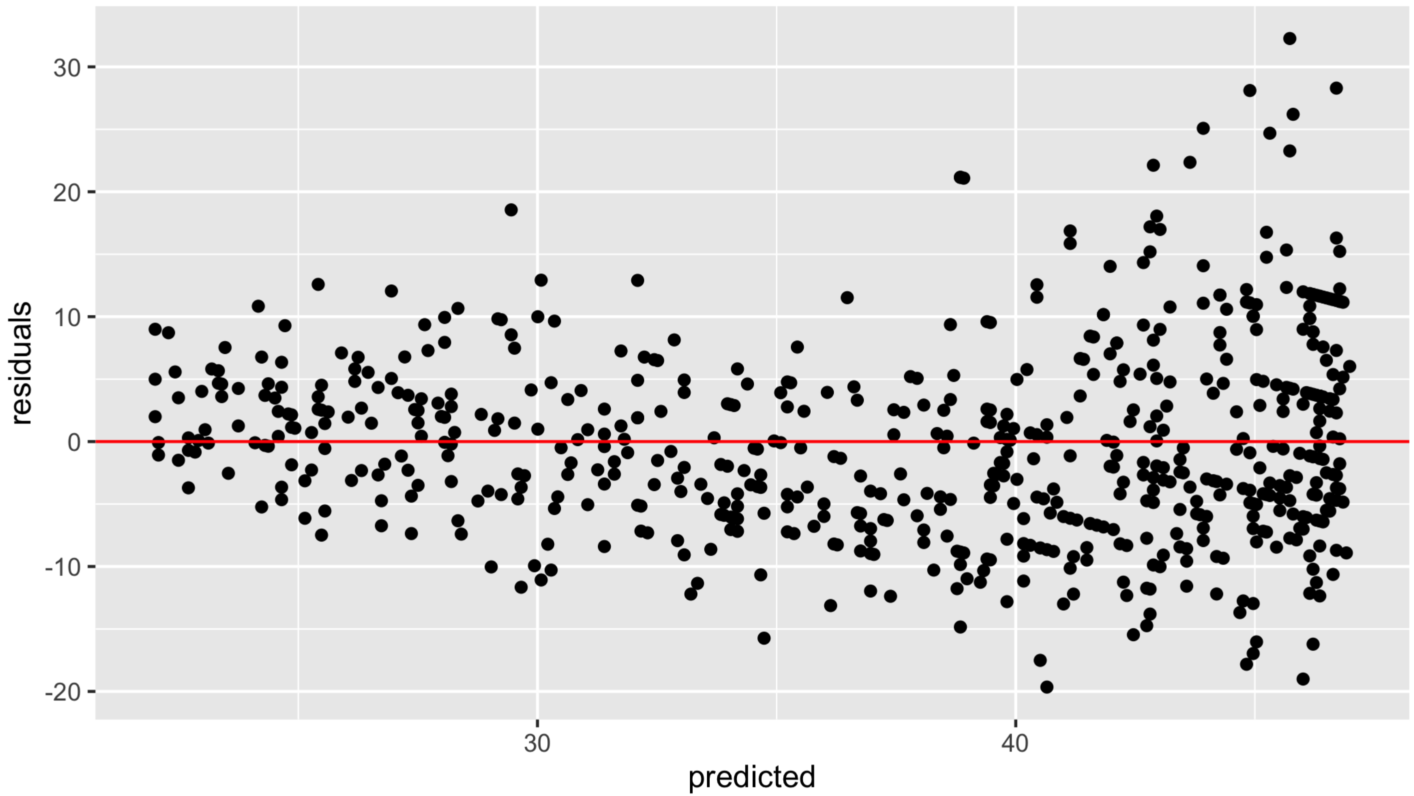
Rate vs. Age

```
respiratory <- ex0824
ggplot(data=respiratory, aes(x=Age, y=Rate)) +
  geom_point() +
  labs("Respiratory Rate vs. Age")
```



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 transformations

Need to transform y

- Typically, a "fan-shaped" residual plot indicates the need for a transformation of the response variable y
  - $\log(\mathbf{y})$: Easiest 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

- Use when the residual plot shows "fan-shaped" pattern
- If we apply a log transformation to the response variable, we want to estimate the parameters for the model...

$$\log(y) = \beta_0 + \beta_1 x$$

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

$$y = \exp\{\beta_0 + \beta_1 x\} = \exp\{\beta_0\}\exp\{\beta_1 x\}$$

Mean and logs

Suppose we have a set of values

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

Let's find the mean of the logged values of x, i.e. $\overline{\log(x)}$

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

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

Let's find mean of x and then log the mean value, i.e. $\log(\bar{x})$

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

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

STA 210

Median and logs

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

Let's find the median of the logged values of x, i.e. $\text{Median}(\log(x))$

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

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

Let's find median of x and then log the mean value, i.e. $\log(\text{Median}(x))$

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

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

Mean, Median, and log

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

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

```
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 y at the given value x_i. This doesn't hold when we log-transform y

- The mean of the logged values is **not** 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,

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

STA 210

Interpretation with log-transformed y

- Given the previous facts, if $\log(y) = \beta_0 + \beta_1 x$, then

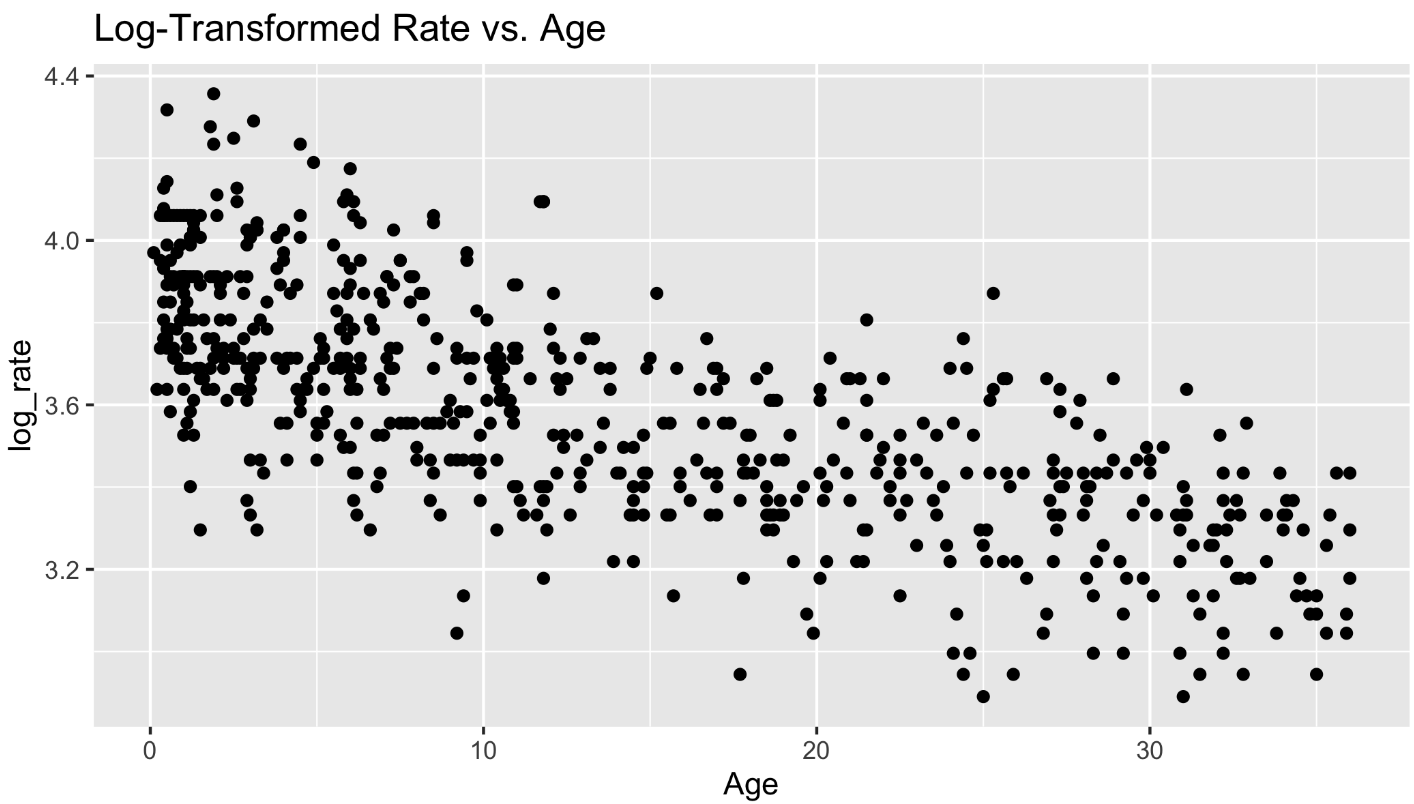
$$\text{Median}(y) = \exp\{\beta_0\}\exp\{\beta_1 x\}$$

- **Intercept:** When $x = 0$, the median of y is expected to be $\exp\{\beta_0\}$
- **Slope:** For every one unit increase in x, the median of y is expected to multiply by a factor of $\exp\{\beta_1\}$

STA 210

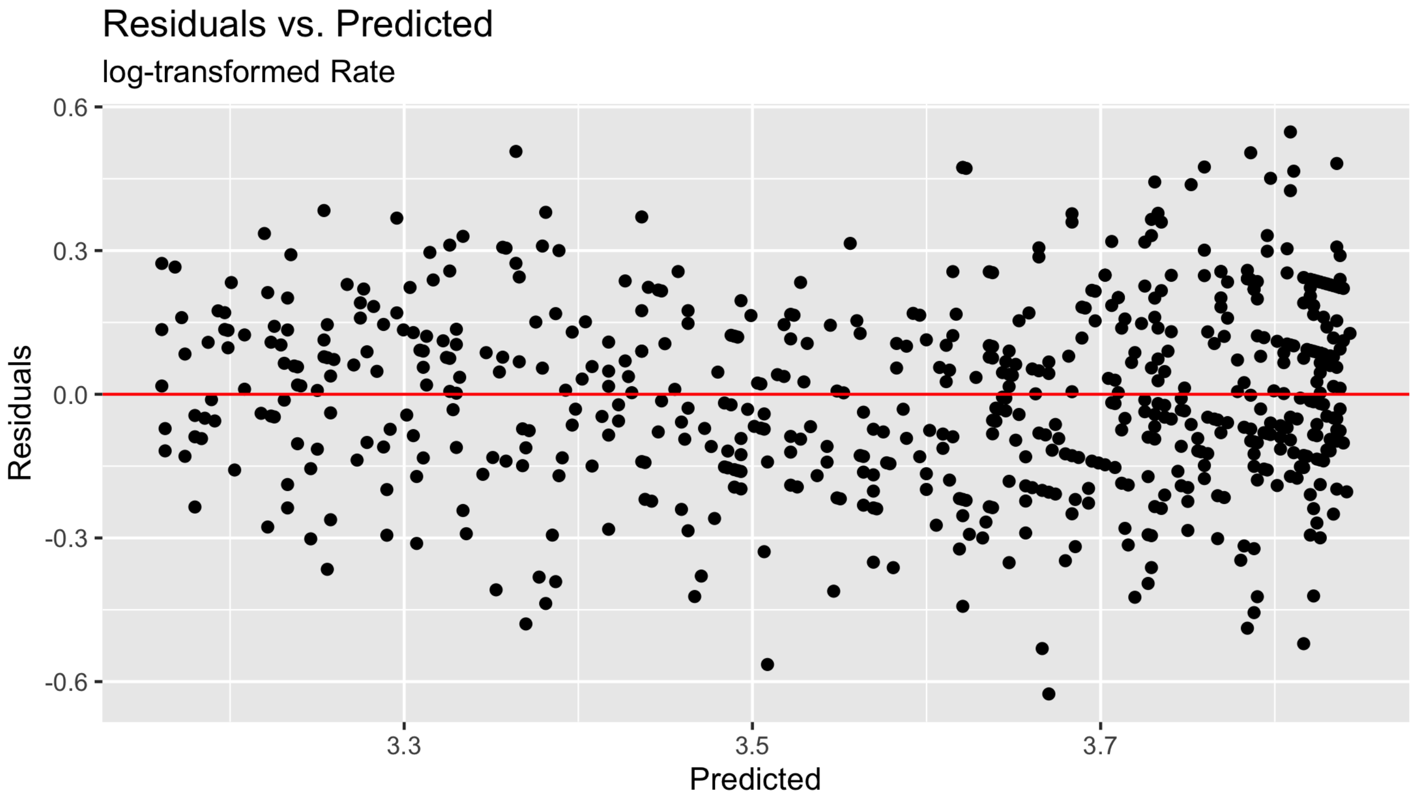
log(Rate) vs. Age

```
respiratory <- respiratory %>% mutate(log_rate = log(Rate))
```



log(Rate) vs. Age

```
log_model <- lm(log_rate ~ Age, data = respiratory)
```



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

- Go to http://bit.ly/sta210-sp20-logy and interpret the model.

04:00

STA 210

Confidence interval for β_j

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

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

- The confidence interval for the coefficient of x describing its relationship with y is

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

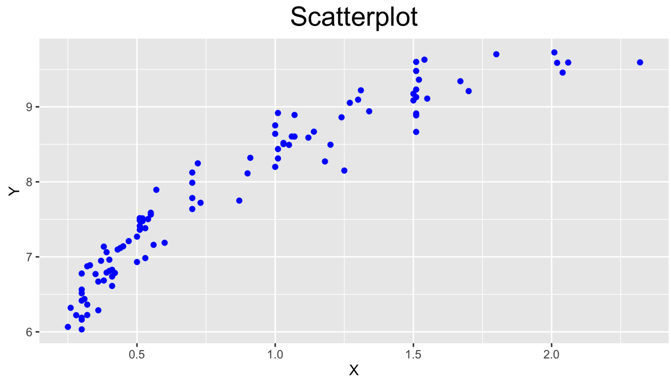
STA 210

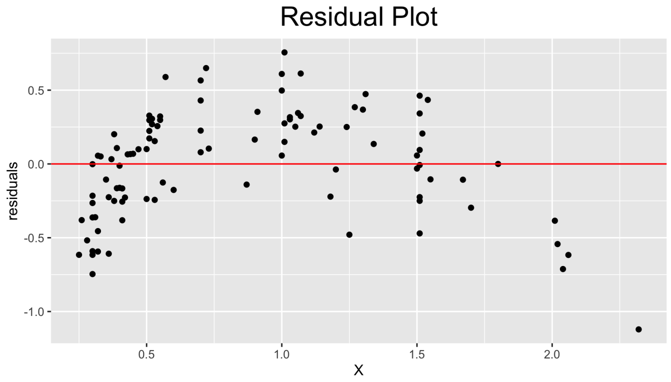
Coefficient of `Age`

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

Interpret the 95% confidence interval for the coefficient of `Age` in terms of *rate*.

Log Transformation on x





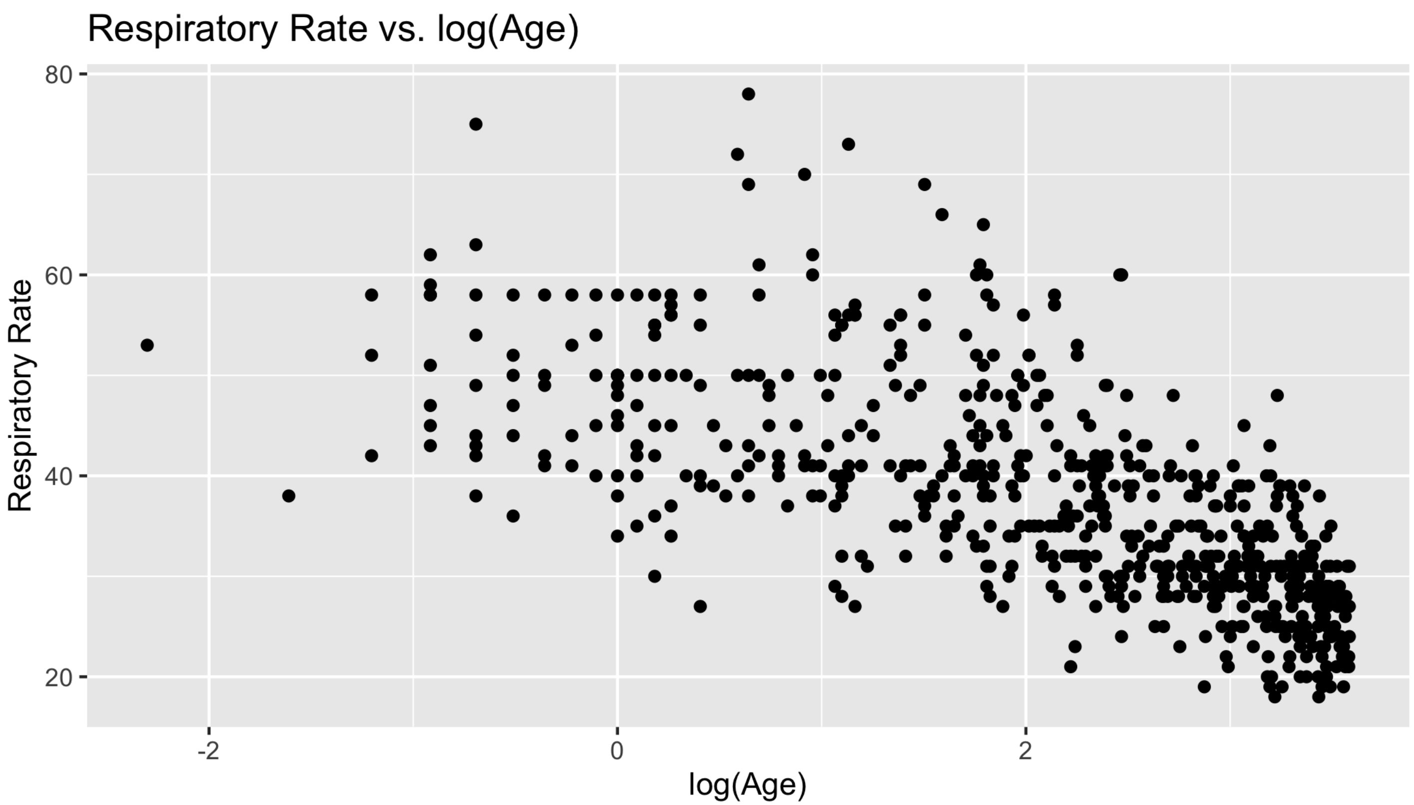
- Try a transformation on X if the scatterplot shows some curvature but the variance is constant for all values of X

Model with Transformation on x

$$y = \beta_0 + \beta_1 \log(x)$$

- **Intercept:** When $\log(x) = 0$, $(x = 1)$, y is expected to be β_0 (i.e. the mean of y is β_0)
- **Slope:** When x is multiplied by a factor of $\mathbf{C}$, y is expected to change by $\boldsymbol{\beta_1}\log(\mathbf{C})$ units, i.e. the mean of y changes by $\boldsymbol{\beta_1}\log(\mathbf{C})$
  - *Example*: when x is multiplied by a factor of 2, y is expected to change by $\boldsymbol{\beta_1}\log(\mathbf{2})$ units

Rate vs. log(Age)



Rate vs. 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

Go to http://bit.ly/sta210-sp20-logx and interpret the model.

04:00

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

STA 210