Multiple Linear Regression

Interactions & Transformations

Dr. Maria Tackett

09.25.19

[Click for PDF of slides]

Announcements

- HW 02 **due TODAY at 11:59p**
- Team Feedback #1 **due TODAY at 11:59p**
  - Please provide honest and constructive feedback. This team feedback will be graded for completion.
- Reading 05 for Monday
- HW 03 due Wednesday, 10/1 at 11:59p

Today's Agenda

- Categorical Predictors with $K > 2$ categories
- Interactions
- Log Transformations

R packages

```r
library(tidyverse)
library(knitr)
library(broom)
library(cowplot) # use plot_grid function
library(Sleuth3)
```

STA 210

Categorical Predictors

STA 210

Starting wages data

Explanatory

- **`Educ`:** years of Education
- **`Exper`:** months of previous work Experience (before hire at bank)
- **`Female`:** 1 if female, 0 if male
- **`Senior`:** months worked at bank since hire
- **`Age`:** Age in months

Response

- **`Bsal`:** annual salary at time of hire

Starting wages: Education categorical

term	estimate	std.error	statistic	p.value
(Intercept)	5637.224	183.730	30.682	0.000
SeniorCent	-21.710	5.320	-4.081	0.000
AgeCent	0.645	0.735	0.877	0.383
ExperCent	0.339	1.069	0.317	0.752
EducCat10	-665.340	535.844	-1.242	0.218
EducCat12	182.567	169.589	1.077	0.285
EducCat15	540.858	187.389	2.886	0.005
EducCat16	766.746	298.375	2.570	0.012
Female1	-756.105	129.586	-5.835	0.000

STA 210

EducCat Behind the scenes

- The categorical variable `EducCat` has 5 levels, so there are 4 indicator variables for Education in the model.
- For a given observation, a value is assigned for each of thr 4 indicator variables based on the following scheme:

	Indicator Variables			
Observations	**EducCat10**	**EducCat12**	**EducCat15**	**EducCat16**
Education = 8 (baseline)	0	0	0	0
Education = 10	1	0	0	0
Education = 12	0	1	0	0
Education = 15	0	0	1	0
Education = 16	0	0	0	1

Application Exercise

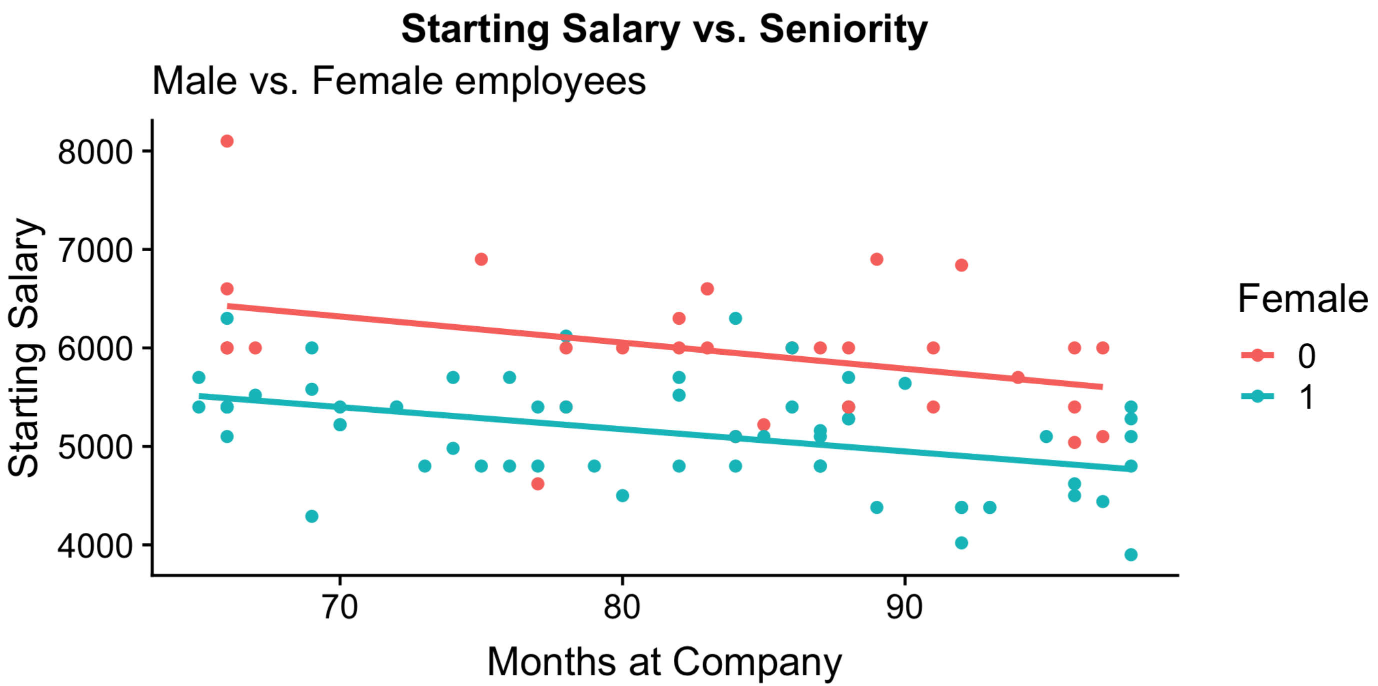
- Go to the **Wages** application exercise in RStudio Cloud.
- Fit a regression model with Education treated as a categorical variable.

- What is the baseline for Education?
- Interpret the coefficient for `EducCat16`.
- What is your conclusion from the p-value of `EducCat16`?
- Write the model equation for those with 8 years of education.
- Write the model equation for those with 16 years of education.

STA 210

Interactions

Checking for interactions



Do you think there is a significant interaction effect between `Female` and `Senior`? Why or why not?

STA 210

Checking for interactions

Do you think there is a significant interaction effect between `Female` and `Exper`? Why or why not?

Model with interactions

```
int_model <- lm(Bsal ~SeniorCent + AgeCent + ExperCent + EducCat +
kable(tidy(int_model), format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	5641.379	184.550	30.568	0.000
SeniorCent	-21.406	5.363	-3.991	0.000
AgeCent	0.583	0.745	0.783	0.436
ExperCent	-0.008	1.215	-0.007	0.995
EducCat10	-648.335	538.592	-1.204	0.232
EducCat12	180.877	170.251	1.062	0.291
EducCat15	531.351	188.744	2.815	0.006
EducCat16	738.594	303.058	2.437	0.017
Female1	-754.483	130.102	-5.799	0.000
ExperCent:Female1	0.741	1.219	0.608	0.545

STA 210

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.

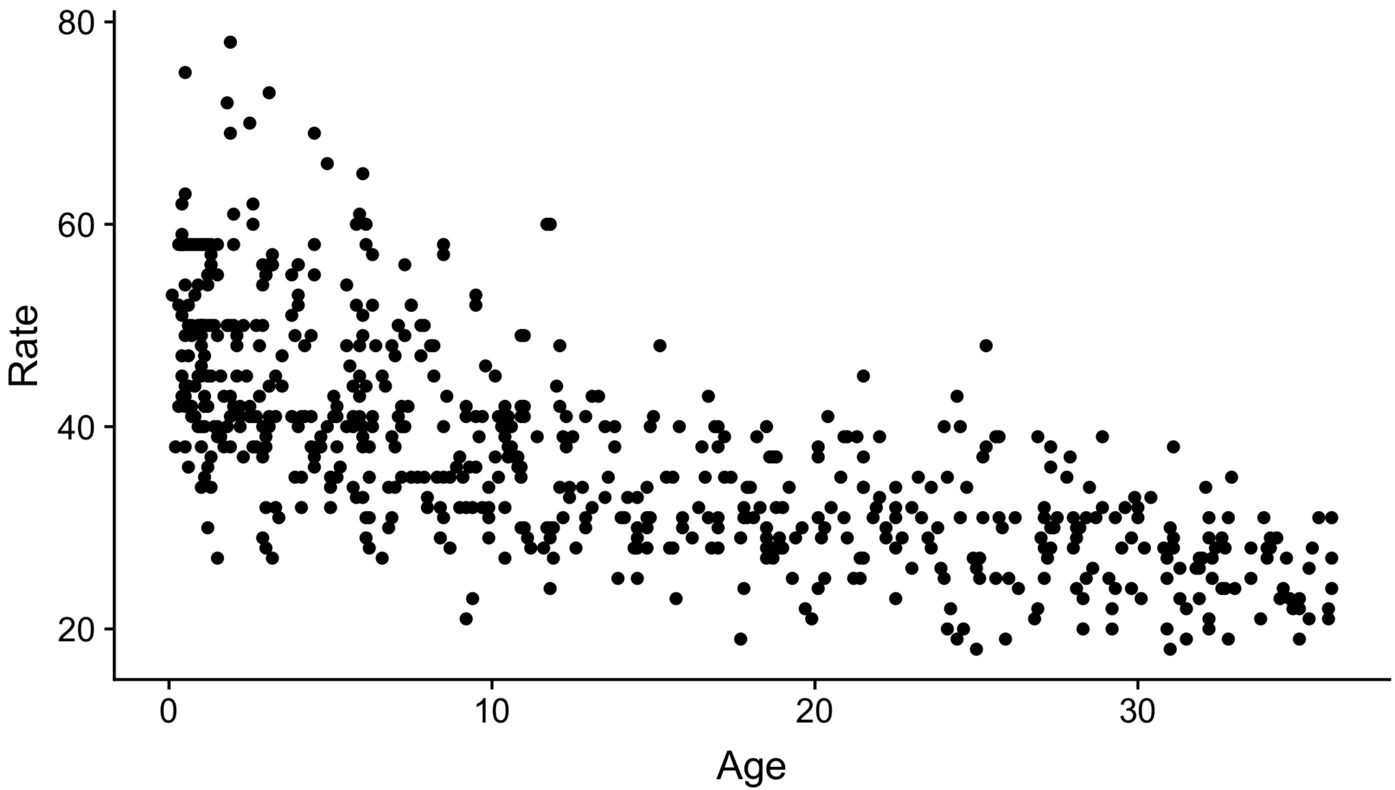
- The data contain the respiratory rate for 618 children ages 15 days to 3 years.

- **Variables**:

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

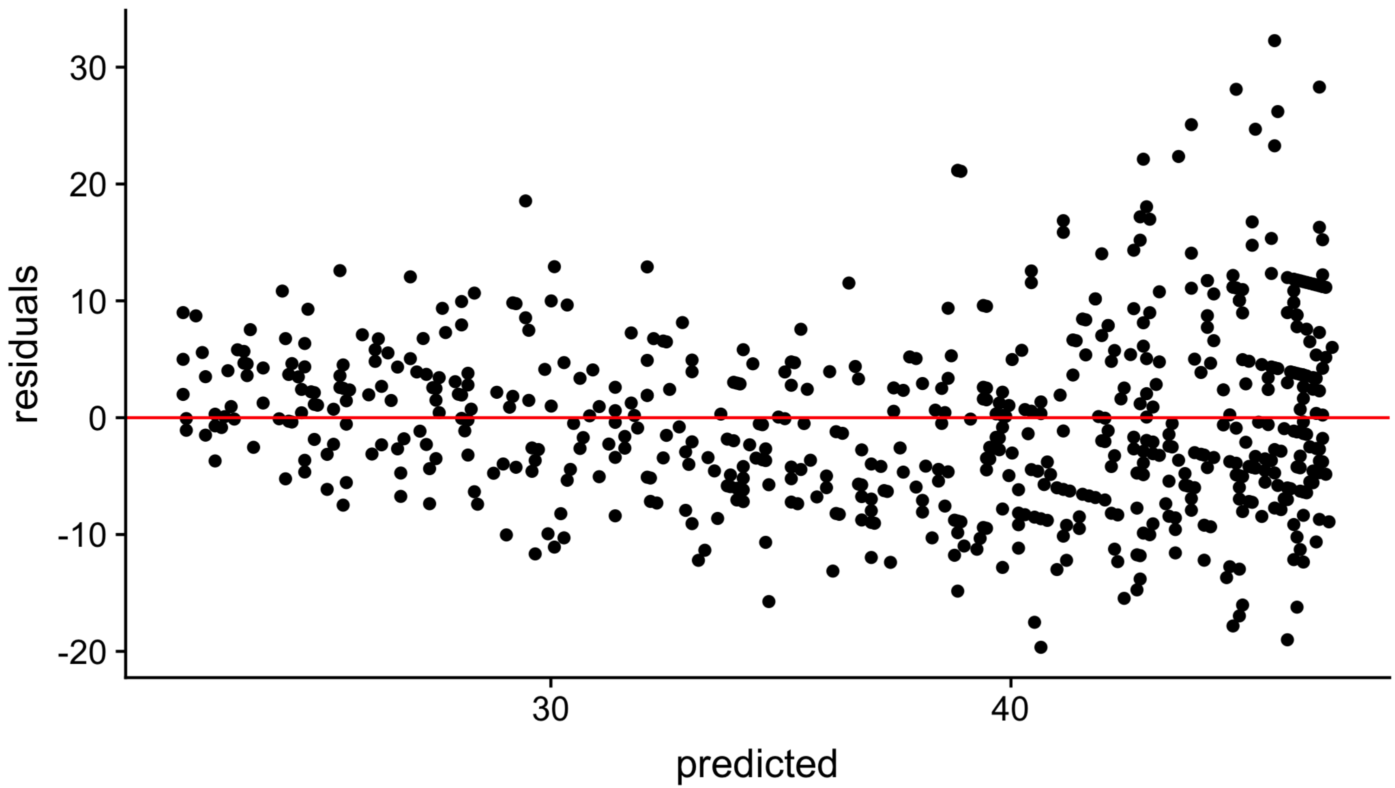
Rate vs. Age

```
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 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,

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

Interpretation with log-transformed y

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

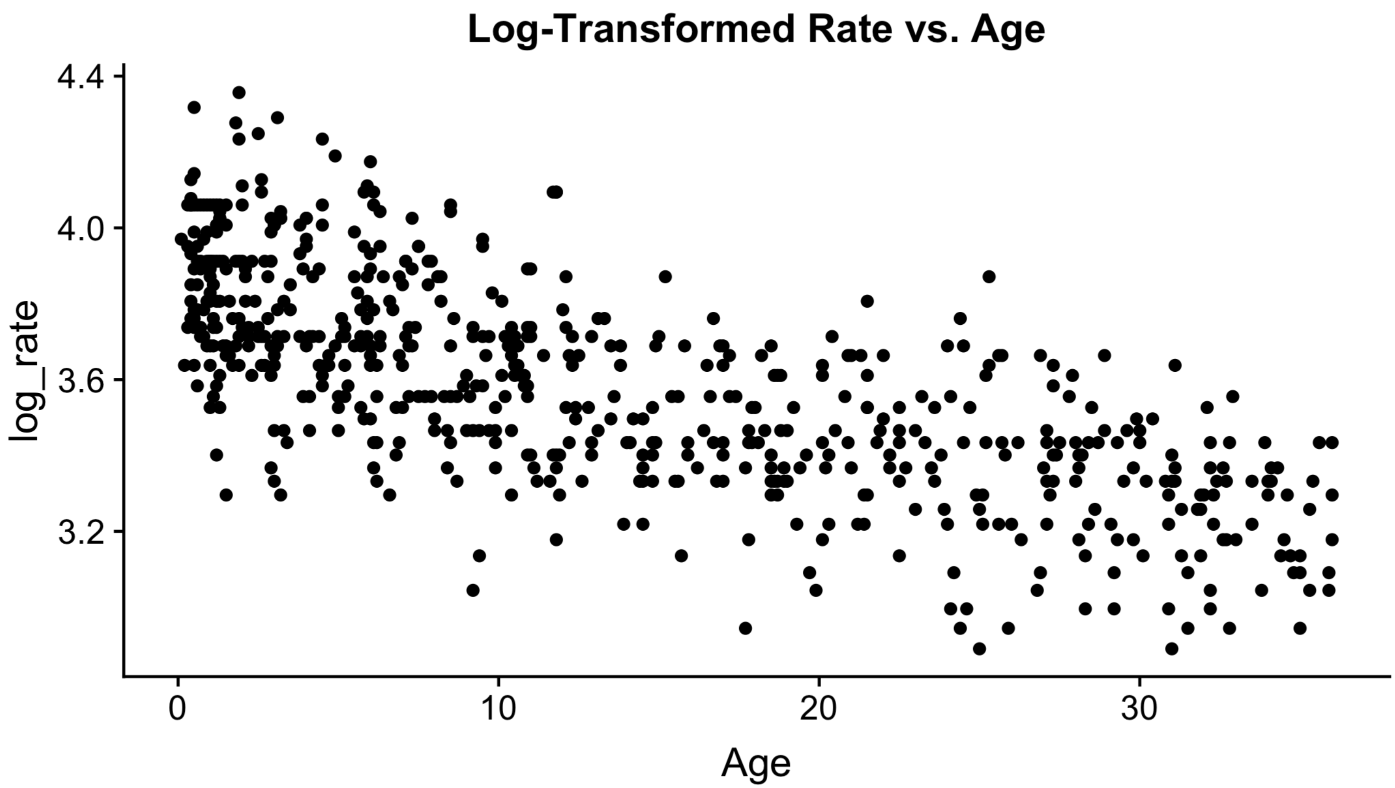
$$\text{Median}(y) = \exp\{\beta_0\}\exp\{\beta_1 x\}$$

- **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

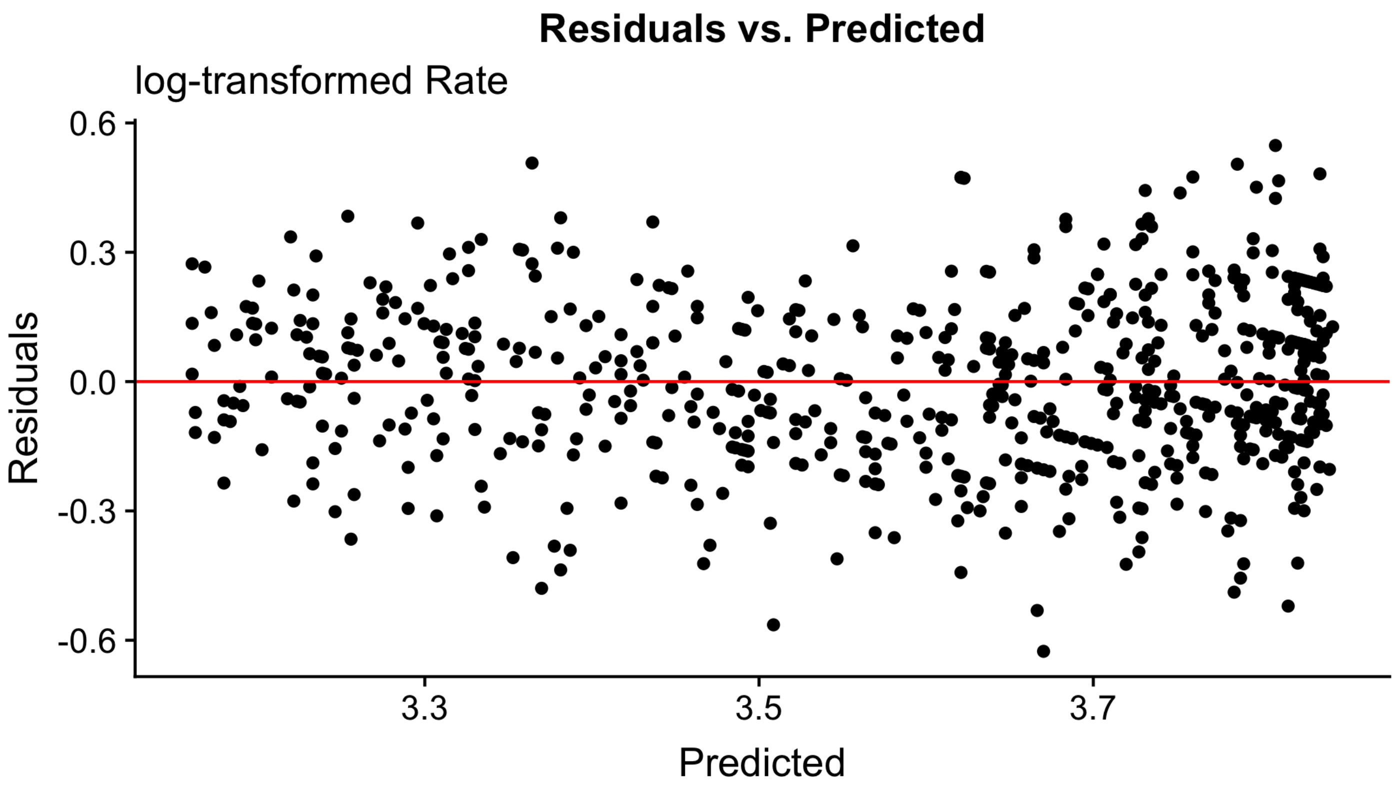
log(Rate) vs. Age

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



log(Rate) vs. Age

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



log(Rate) vs. Age

```
kable(tidy(log_model, conf.int=TRUE),format="html", digits=3)
```

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

1. Write the model in terms of $\log(rate)$.
2. Write the model in terms of $rate$. Interpret the slope and intercept.

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`

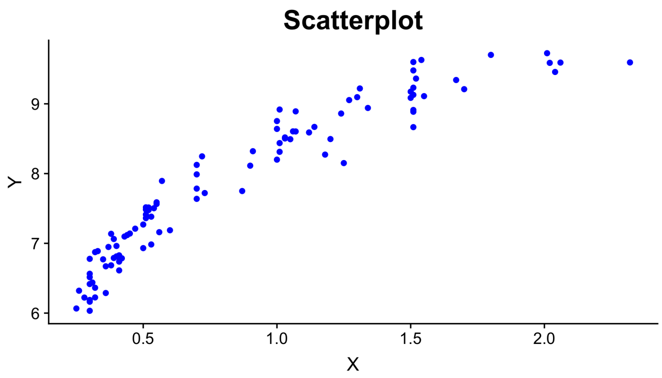
```
kable(tidy(log_model, conf.int=TRUE),format="html", digits=3)
```

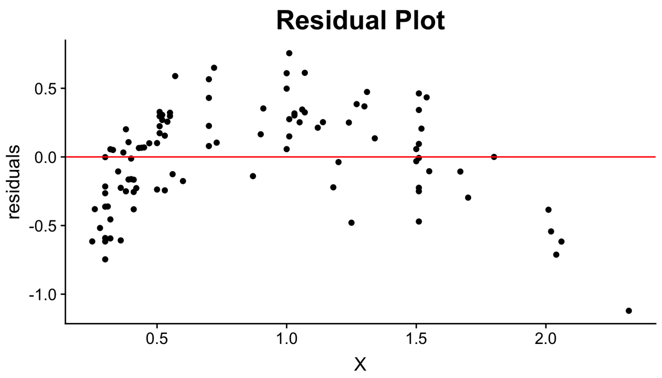
term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	3.845	0.013	304.500	0	3.82	3.870
Age	-0.019	0.001	-25.839	0	-0.02	-0.018

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

STA 210

Log Transformation on x





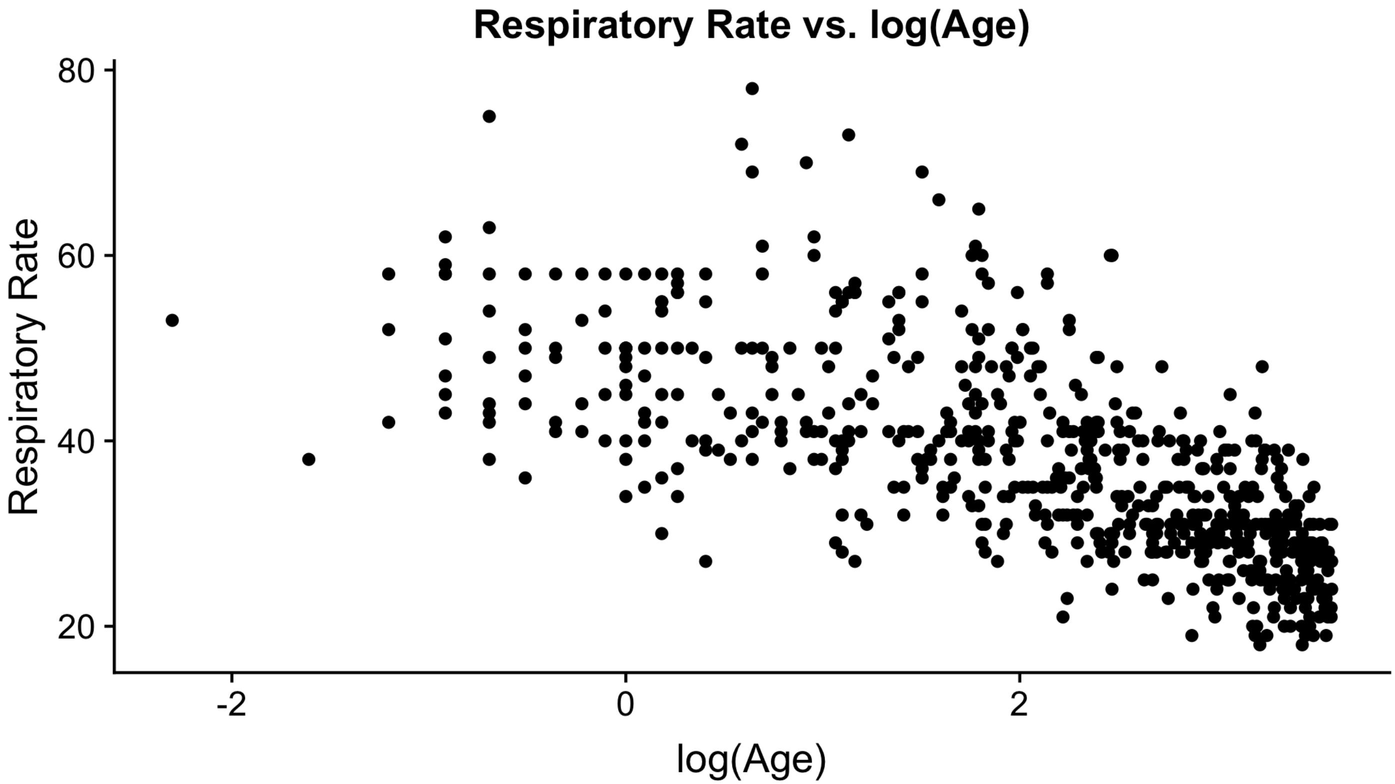
- Try a transformation on X if the scatterplot shows some curvature but the variance is constant for all values of X

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)	50.134533	0.6319775	79.32961	0	48.893441	51.375625
log.age	-5.982434	0.2626097	-22.78070	0	-6.498153	-5.466715

1. Write the equation for the model of y regressed on $\log(x)$.
2. Interpret the intercept in the context of the problem.
3. Interpret the slope in terms of how the mean respiratory rate changes when a child's age doubles.
4. Suppose a doctor has a patient who is currently 3 years old. Will this model provide a reliable prediction of the child's respiratory rate when her age doubles? Why or why not?

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

Before Next Class

- Reading 05 for Monday