Multiple Linear Regression

Model Assessment

Dr. Maria Tackett

09.30.19

STA 210

[Click for PDF of slides]

Announcements

- Lab 05 **due Tuesday at 11:59p**
- HW 03 **due Wednesday at 11:59p**
- Reading 06 for Wednesday

R packages

```
library(tidyverse)
library(knitr)
library(broom)
library(Sleuth3) # ex0824 data
library(cowplot) # use plot_grid function
```

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)

STA 210

Log transformation on y

```r
log_model <- lm(log_rate ~ Age, data = respiratory)
  kable(tidy(log_model, conf.int = TRUE), format = "markdown", dig
```

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

$$\widehat{\log \text{rate}} = 3.845 - 0.019 \times \text{Age}$$

- **Slope:** For every one month incraese in Age, we expect the median respiratory rate to be multiplied by a factor of $\exp\{-0.019\} = 1.019$ breaths per minute.

- **Intercept:** The expected respiratory rate for a child who is 0 months old (a newborn) is $\exp\{3.845\} = 46.76$ beats per minute.

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`

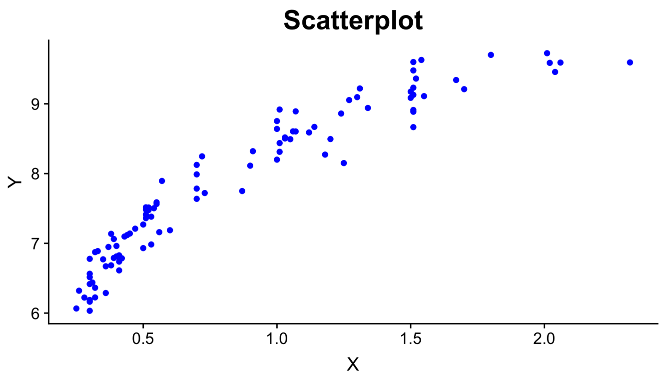
term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	3.845	0.013	304.500	0	3.82	3.870
Age	-0.019	0.001	-25.839	0	-0.02	-0.018

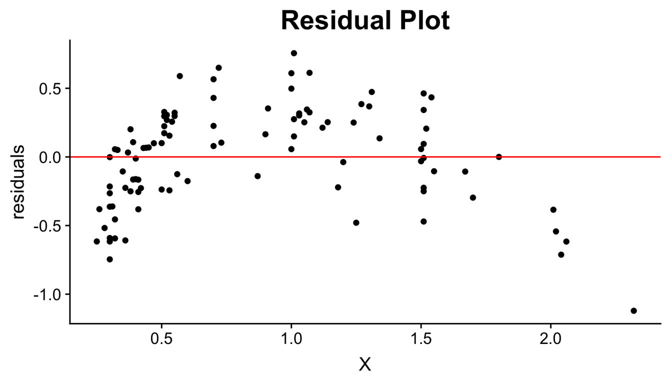
The 95% confidence interval for the coefficient of Age in terms of `Rate`:

$$[\exp\{-0.02\}, \exp\{-0.018\}] = [0.981, 0.982]$$

Interpretation: We are 95% confident that for each additional month in age, we can expect the median respiratory rate to be multiplied by a factor of 0.981 to 0.982.

Log Transformation on x





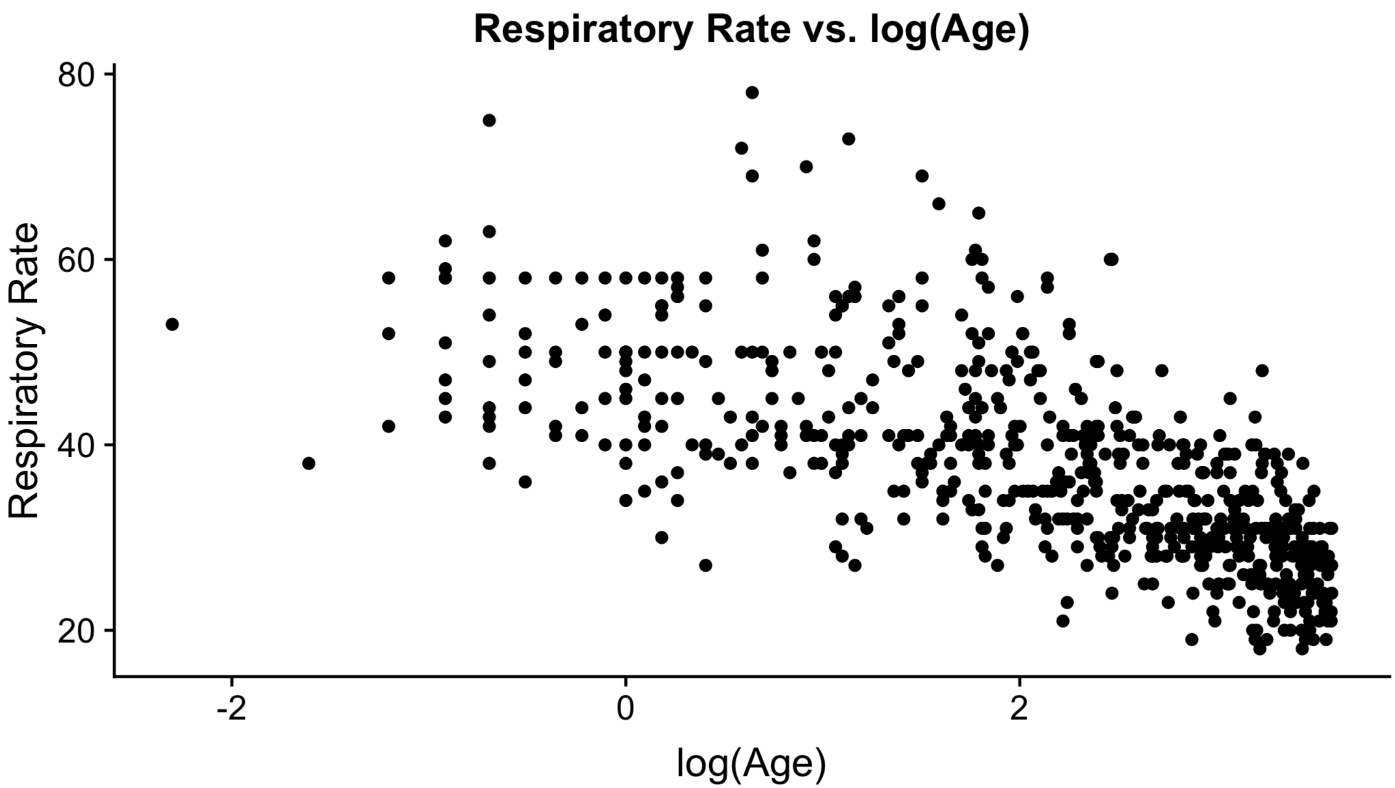
- Try a transformation on X if the scatterplot shows some curvature but the variance is constant for all values of X

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?

STA 210

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

Model Assessment & Selection

Restaurant tips

What affects the amount customers tip at a restaurant?

- **Response:**
  - **`Tip`**: amount of the tip
- **Predictors:**
  - **`Party`**: number of people in the party
  - **`Meal`**: time of day (Lunch, Dinner, Late Night)
  - **`Age`**: age category of person paying the bill (Yadult, Middle, SenCit)

```
tips <- read_csv("data/tip-data.csv") %>%
  filter(!is.na(Party))
```

STA 210

ANOVA table for regression

We can use the Analysis of Variance (ANOVA) table to decompose the variability in our response variable

	Sum of Squares	DF	Mean Square	F-Stat	p-value
Regression (Model)	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	p	$\frac{MSS}{p}$	$\frac{MMS}{RMS}$	$P(F > \text{F-Stat})$
Residual	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-p-1$	$\frac{RSS}{n-p-1}$		
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$	$\frac{TSS}{n-1}$		

The estimate of the regression variance, $\hat{\sigma}^2 = RMS$

STA 210

R^2

- **Recall**: R^2 is the proportion of the variation in the response variable explained by the regression model
- R^2 will always increase as we add more variables to the model
  - If we add enough variables, we can always achieve $R^2 = 100\%$
- If we only use R^2 to choose a best fit model, we will be prone to choose the model with the most predictor variables

Adjusted R^2

- **Adjusted R^2**: a version of R^2 that penalizes for unnecessary predictor variables
- Similar to R^2, it measures the proportion of variation in the response that is explained by the regression model
- Differs from R^2 by using the mean squares rather than sums of squares and therefore adjusting for the number of predictor variables

R^2 and Adjusted R^2

$$R^2 = \frac{\text{Total Sum of Squares} - \text{Residual Sum of Squares}}{\text{Total Sum of Squares}}$$

$$Adj.\,R^2 = \frac{\text{Total Mean Square} - \text{Residual Mean Square}}{\text{Total Mean Square}}$$

- $Adj.\,R^2$ can be used as a quick assessment to compare the fit of multiple models; however, it should not be the only assessment!
- Use R^2 when describing the relationship between the response and predictor variables

STA 210

Restaurant tips: model

```
model1 <- lm(Tip ~ Party + Meal + Age , data = tips)
kable(tidy(model1),format="html",digits=3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	1.254	0.394	3.182	0.002
Party	1.808	0.121	14.909	0.000
MealLate Night	-1.632	0.407	-4.013	0.000
MealLunch	-0.612	0.402	-1.523	0.130
AgeSenCit	0.390	0.394	0.990	0.324
AgeYadult	-0.505	0.412	-1.227	0.222

Restaurant tips: ANOVA

- R output

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Party	1	1188.636	1188.636	311.002	0.000
Meal	2	88.460	44.230	11.573	0.000
Age	2	13.032	6.516	1.705	0.185
Residuals	163	622.979	3.822	NA	NA

- ANOVA table

	Sum of Squares	DF	Mean Square	F-Stat	p-value
Regression (Model)	1290.12829	5	258.025658	67.5113618	0
Residual	622.97932	163	3.821959		
Total	1913.10761	168			

STA 210

Calculating R^2 and Adj R^2

	Sum of Squares	DF	Mean Square	F-Stat	p-value
Regression (Model)	1290.12829	5	258.025658	67.5113618	0
Residual	622.97932	163	3.821959		
Total	1913.10761	168			

```
#r-squared
tss <- 1188.63588 + 88.46005 + 13.03236 + 622.97932
rss <- 622.97932
(r_sq <- (tss - rss)/tss)
```

```
## [1] 0.6743626
```

```
#adj r-squared
tms <- tss/(nrow(tips)-1)
rms <- 3.821959
(adj_r_sq <- (tms - rms)/tms)
```

```
## [1] 0.6643738
```

Restaurant tips: R^2 and Adj. R^2

```
glance(model1)
```

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <d
## 1     0.674         0.664  1.95      67.5 6.14e-38     6  -350.  714.  7
## # … with 2 more variables: deviance <dbl>, df.residual <int>
```

- Close values of R^2 and Adjusted R^2 indicate that the variables in the model are significant in understanding variation in tips

STA 210

ANOVA F Test

- Using the ANOVA table, we can test whether any variable in the model is a significant predictor of the response. We conduct this test using the following hypotheses:

$$
\begin{aligned}
&H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\
&H_a : \text{at least one } \beta_j \text{ is not equal to } 0
\end{aligned}
$$

- The statistic for this test is the F test statistic in the ANOVA table
- We calculate the p-value using an F distribution with p and $(n-p-1)$ degrees of freedom

STA 210

ANOVA F Test in R

```
model0 <- lm(Tip ~ 1, data=tips)
```

```
model1 <- lm(Tip ~ Party + Meal + Age , data = tips)
```

```
kable(anova(model0,model1),format="html")
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
168	1913.1076	NA	NA	NA	NA
163	622.9793	5	1290.128	67.51136	0

At least one coefficient is non-zero, i.e. at least one predictor in the model is significant

Testing subset of coefficients

- Sometimes we want to test whether a subset of coefficients are all equal to 0

- This is often the case when we want test

  - whether a categorical variable with k levels is a significant predictor of the response

  - whether the interaction between a categorical and quantitative variable is significant

- To do so, we will use the **Nested F Test**

Nested F Test

- Suppose we have a full and reduced model:

$$
\begin{aligned}
\text{Full} &: y = \beta_0 + \beta_1 x_1 + \dots + \beta_q x_q + \beta_{q+1} x_{q+1} + \dots \beta_p x_p \\
\text{Red} &: y = \beta_0 + \beta_1 x_1 + \dots + \beta_q x_q
\end{aligned}
$$

- We want to test whether any of the variables $x_{q+1}, x_{q+2}, \dots, x_p$ are significant predictors. To do so, we will test the hypothesis:

$$
\begin{aligned}
H_0 &: \beta_{q+1} = \beta_{q+2} = \dots = \beta_p = 0 \\
H_a &: \text{at least one } \beta_j \text{ is not equal to } 0
\end{aligned}
$$

STA 210

Nested F Test

- The test statistic for this test is

$$F = \frac{(RSS_{reduced} - RSS_{full}) \big/ (p_{full} - p_{reduced})}{RSS_{full} \big/ (n - p_{full} - 1)}$$

- Calculate the p-value using the F distribution with $(p_{full} - p_{reduced})$ and $(n - p_{full} - 1)$ degrees of freedom

STA 210

Is `Meal` a significant predictor of tips?

term	estimate	std.error	statistic	p.value
(Intercept)	1.254	0.394	3.182	0.002
Party	1.808	0.121	14.909	0.000
AgeSenCit	0.390	0.394	0.990	0.324
AgeYadult	-0.505	0.412	-1.227	0.222
MealLate Night	-1.632	0.407	-4.013	0.000
MealLunch	-0.612	0.402	-1.523	0.130

STA 210

Tips data: Nested F Test

$$H_0 : \beta_{latenight} = \beta_{lunch} = 0$$
$$H_a : \text{ at least one } \beta_j \text{ is not equal to } 0$$

```
reduced <- lm(Tip ~ Party + Age, data = tips)
```

```
full <- lm(Tip ~ Party + Age + Meal, data = tips)
```

```
kable(anova(full,reduced),format="html")
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
163	622.9793	NA	NA	NA	NA
165	686.4439	-2	-63.46457	8.302623	0.0003684

At least one coefficient associated with `Meal` is not zero. Therefore, `Meal` is a significant predictor of `Tips`.

STA 210

Why is it not good practice to use the individual p-values to determine a categorical variable with $k > 2$ levels) is significant? *Hint*: What does it actually mean if none of the $k - 1$ p-values are significant?

Practice with Interactions

term	estimate	std.error	statistic	p.value
(Intercept)	1.2764989	0.4910882	2.5993270	0.0102086
Party	1.7947980	0.1715003	10.4652753	0.0000000
AgeSenCit	0.4007889	0.3969295	1.0097230	0.3141431
AgeYadult	-0.4701634	0.4197146	-1.1201978	0.2642977
MealLate Night	-1.8454674	0.7089728	-2.6030159	0.0101039
MealLunch	-0.4608832	0.8651044	-0.5327487	0.5949421
Party:MealLate Night	0.1108600	0.2846584	0.3894491	0.6974586
Party:MealLunch	-0.0500822	0.2825586	-0.1772455	0.8595384

1. Write the general form of the model.
2. Write the model for `Meal == "Late Night"`.
3. How does the mean change when `Meal == "Late Night"`?
4. How does the slope of `Party` change when `Meal == "Late Night"`?

STA 210

Nested F test for interactions

Is the interaction between `Party` and `Meal` significant?

```
reduced <- lm(Tip ~ Party + Age + Meal, data = tips)
```

```
full <- lm(Tip ~ Party + Age + Meal + Meal*Party, data = tips)
```

```
kable(anova(full,reduced),format="html")
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
161	621.9651	NA	NA	NA	NA
163	622.9793	-2	-1.014261	0.1312743	0.877071