Multiple Linear Regression

Model Assessment & Selection

Prof. Maria Tackett

03.02.20

Click for PDF of slides

Announcements

- Project proposal **due Thursday, March 5 at 11:59p**
- Reading 07 for today & Wednesday
- Sign Up for DataFest!

  🗓️ April 3 - 5

  📍 Penn Pavilion

  🔗 stat.duke.edu/datafest

R packages

```
library(tidyverse)
library(knitr)
library(broom)
library(patchwork)
```

Today's Agenda

- ANOVA for Regression
- Nested F Test
- Model Selection
  - R^2 vs. Adj. R^2
  - AIC & BIC
  - Strategies

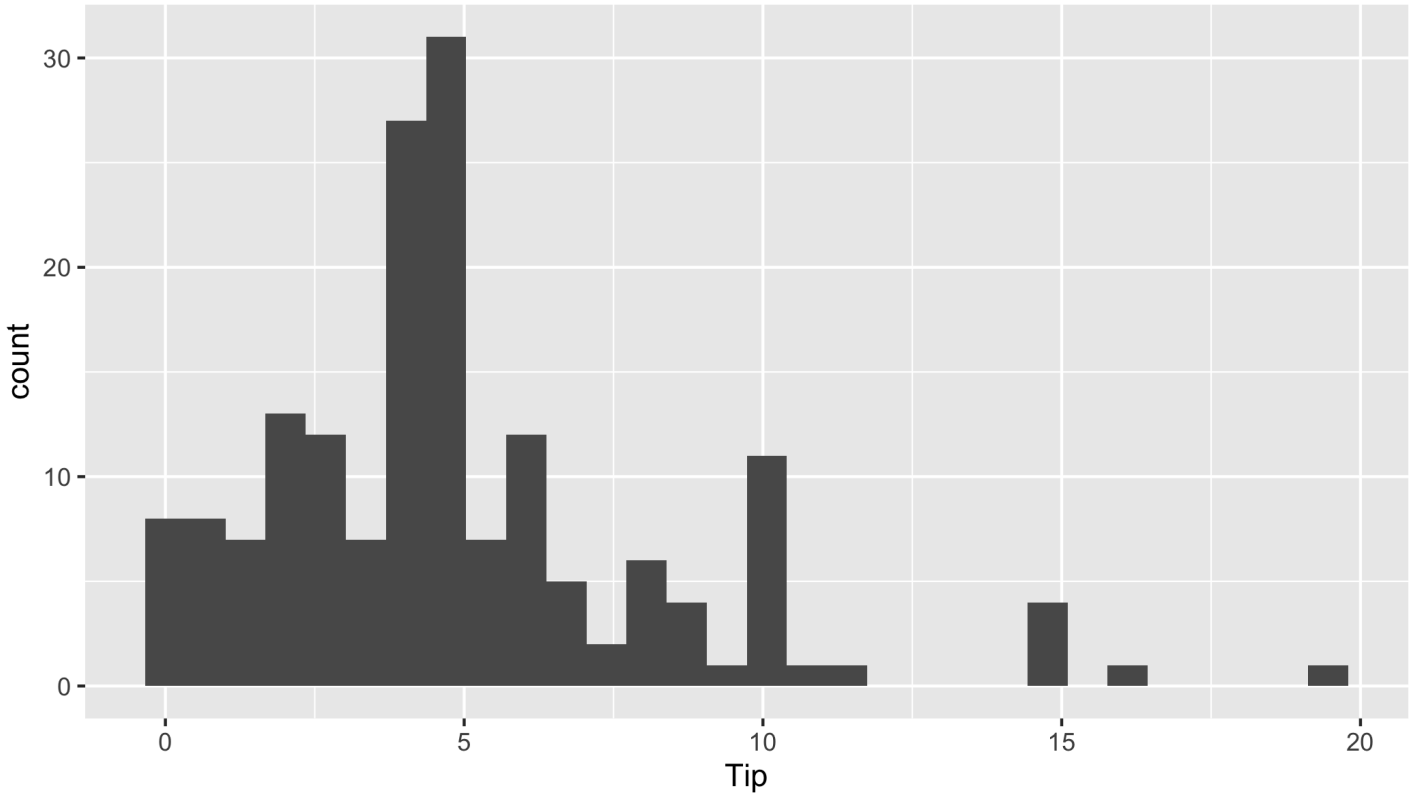
Model Assessment & Selection

Restaurant tips

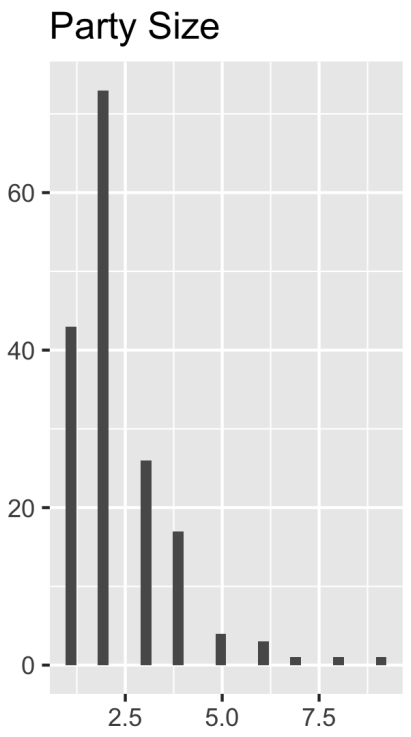
What affects the amount customers tip at a restaurant?

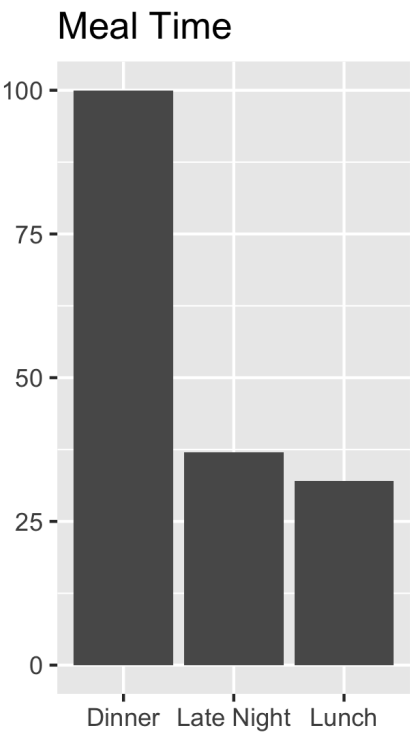
- **Response:**
  - `Tip`: amount of the tip
- **Predictors:**
  - `Party`: number of people in the party
  - `Meal`: time of day (Lunch, Dinner, Late Night)
  - `Age`: age category of person paying the bill (Yadult, Middle, SenCit)

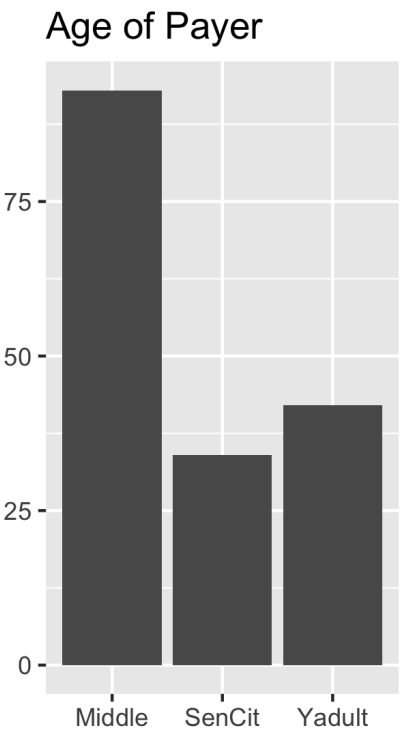
Response Variable



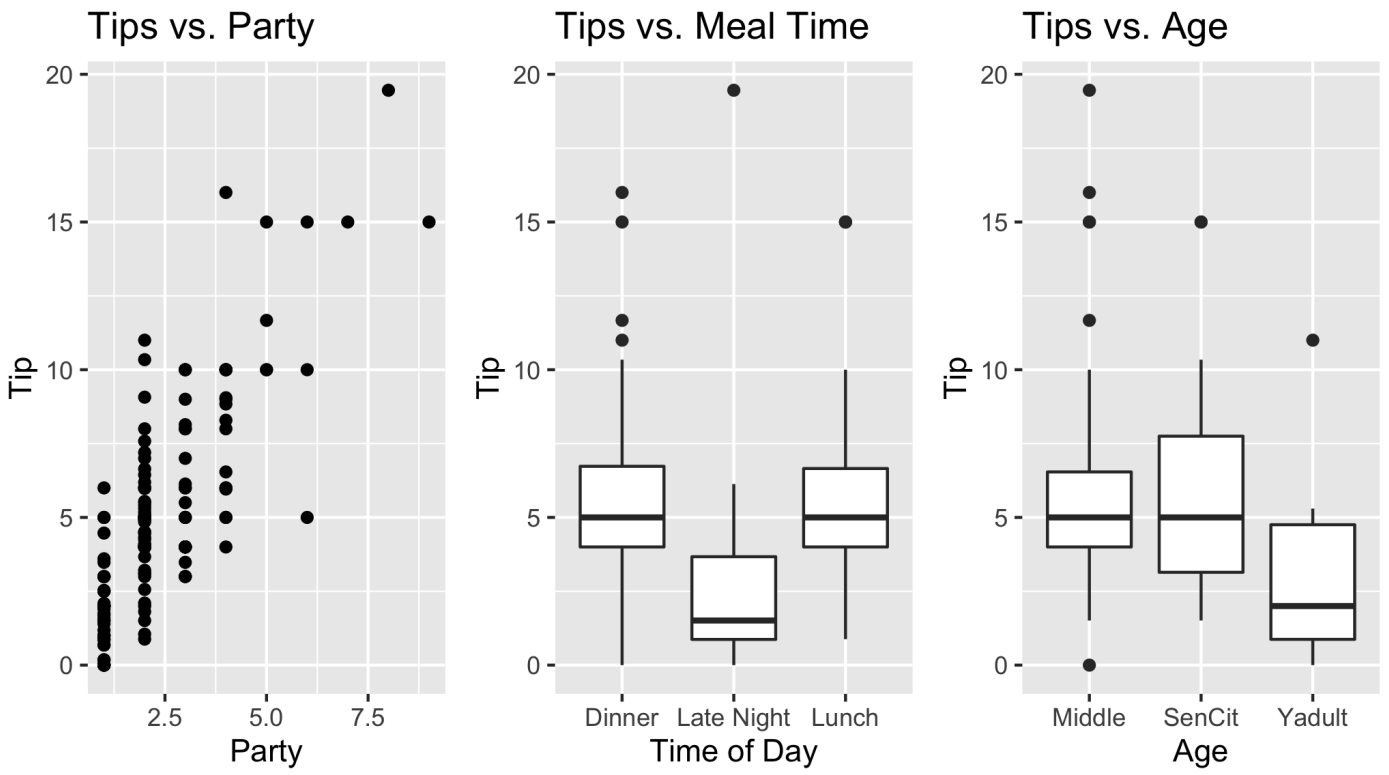
Predictor Variables







Response vs. Predictors



Restaurant tips: model

```
model1 <- lm(Tip ~ Party +  Age , data = tips)
tidy(model1, conf.int = TRUE) %>%
  kable(format = "markdown", digits=3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.838	0.397	2.112	0.036	0.055	1.622
Party	1.837	0.124	14.758	0.000	1.591	2.083
AgeSenCit	0.379	0.410	0.925	0.356	-0.430	1.189
AgeYadult	-1.009	0.408	-2.475	0.014	-1.813	-0.204

Is this the best model to explain variation in Tips?

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	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$	$\frac{TSS}{n-1}$		

In-class exercise

Use the ANOVA table to answer the question at

http://bit.ly/sta210-sp20-reg-anova.

Use **NetId@duke.edu** for your email address.

03:00

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:

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

- 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

ANOVA F Test in R

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

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

```
kable(anova(model0, model1), format="markdown", digits = 3)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
168	1913.108	NA	NA	NA	NA
165	686.444	3	1226.664	98.284	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 (Partial) F Test

Nested (Partial) F Test

- Suppose we have a full and reduced model:

$$\text{Full} : y = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \beta_{q+1} x_{q+1} + \ldots \beta_p x_p$$
$$\text{Red} : y = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$$

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

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

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

Is `Meal` a significant predictor of tips?

term	estimate
(Intercept)	1.254
Party	1.808
AgeSenCit	0.390
AgeYadult	-0.505
MealLate Night	-1.632
MealLunch	-0.612

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(reduced, full), format="markdown", digits = 3)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
165	686.444	NA	NA	NA	NA
163	622.979	2	63.465	8.303	0

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

Model with `Meal`

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.254	0.394	3.182	0.002	0.476	2.032
Party	1.808	0.121	14.909	0.000	1.568	2.047
AgeSenCit	0.390	0.394	0.990	0.324	-0.388	1.168
AgeYadult	-0.505	0.412	-1.227	0.222	-1.319	0.308
MealLate Night	-1.632	0.407	-4.013	0.000	-2.435	-0.829
MealLunch	-0.612	0.402	-1.523	0.130	-1.405	0.181

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?

Including interactions

Does the effect of `Party` differ based on the `Meal` time?

term	estimate
(Intercept)	1.276
Party	1.795
AgeSenCit	0.401
AgeYadult	-0.470
MealLate Night	-1.845
MealLunch	-0.461
Party:MealLate Night	0.111
Party:MealLunch	-0.050

Nested F test for interactions

Let's use a Nested F test to determine if `Party*Meal` is statistically significant.

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

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

```
kable(anova(reduced, full), format = "markdown", digits = 3)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
163	622.979	NA	NA	NA	NA
161	621.965	2	1.014	0.131	0.877

Final model for now

We conclude that the interaction between `Party` and `Meal` is not statistically significant.Therefore, we will use the original model that only included main effects.

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

Model Selection

Which variables should be in the model?

- This is a very hard question that is the subject of a lot of statistical research
- There are many different opinions about how to answer this question
- This lecture will mostly focus on how to approach variable selection
    - We will introduce some specific methods, but there are many others out there

Which variables should you include?

- It depends on the goal of your analysis
- Though a variable selection procedure will select one set of variables for the model, that set is usually one of several equally good sets
- It is best to start with a well-defined purpose and question to help guide the variable selection

Prediction

- **Goal:** to calculate the most precise prediction of the response variable
- Interpreting coefficients is **not** important
- Choose only the variables that are strong predictors of the response variable
    - Excluding irrelevant variables can help reduce widths of the prediction intervals

One variable's effect

- **Goal:** Understand one variable's effect on the response after adjusting for other factors
- Only interpret the coefficient of the variable that is the focus of the study
    - Interpreting the coefficients of the other variables is **not** important
- Any variables not selected for the final model have still been adjusted for, since they had a chance to be in the model

Explanation

- **Goal:** Identify variables that are important in explaining variation in the response
- Interpret any variables of interest
- Include all variables you think are related to the response, even if they are not statistically significant
    - This improves the interpretation of the coefficients of interest
- Interpret the coefficients with caution, especially if there are problems with multicollinearity in the model

Example: SAT Averages by State

- This data set contains the average SAT score (out of 1600) and other variables that may be associated with SAT performance for each of the 50 U.S. states. The data is based on test takers for the 1982 exam.
- **Data:** `case1201` data set in the `Sleuth3` package
- Response variable:
  - **SAT**: average total SAT score

SAT Averages: Explanatory Variables

- **`State`**: U.S. State
- **`Takers`**: percentage of high school seniors who took exam
- **`Income`**: median income of families of test-takers ($ hundreds)
- **`Years`**: average number of years test-takers had formal education in social sciences, natural sciences, and humanities
- **`Public`**: percentage of test-takers who attended public high schools
- **`Expend`**: total state expenditure on high schools ($ hundreds per student)
- **`Rank`**: median percentile rank of test-takers within their high school classes

In-Class Exercise:

Select the primary modeling objective for each scenario

http://bit.ly/sta210-sp20-selection

Use **NetId@duke.edu** for your email address.

If you finish early, discuss a modeling strategy for each scenario.

04:00

Model selection criterion

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 is a measure of the amount 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

SAT: ANOVA

```
sat_data <- Sleuth3::case1201 %>%
  select(-State)

sat_model <- lm(SAT ~ ., data = sat_data)
tidy(sat_model) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-94.659	211.510	-0.448	0.657
Takers	-0.480	0.694	-0.692	0.493
Income	-0.008	0.152	-0.054	0.957
Years	22.610	6.315	3.581	0.001
Public	-0.464	0.579	-0.802	0.427
Expend	2.212	0.846	2.615	0.012
Rank	8.476	2.108	4.021	0.000

SAT ANOVA

```
anova(sat_model) %>%
  kable(format = "markdown", digits = 3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Takers	1	181024.092	181024.092	260.838	0.000
Income	1	121.416	121.416	0.175	0.678
Years	1	14661.165	14661.165	21.125	0.000
Public	1	5154.528	5154.528	7.427	0.009
Expend	1	3984.227	3984.227	5.741	0.021
Rank	1	11222.976	11222.976	16.171	0.000
Residuals	43	29842.416	694.010	NA	NA

SAT Scores: R^2 and Adj. R^2

```
glance(sat_model)
```

```
## # 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     0.879         0.862  26.3      51.9 4.16e-18     7  -231.  477.  4
## # … 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 the response, i.e. that there aren't a lot of unnecessary variables in the model

Additional model selection criterion

- Akaike's Information Criterion (AIC):

$$AIC = n\log(RSS) - n\log(n) + 2(p+1)$$

- Schwarz's Bayesian Information Criterion (BIC):

$$BIC = n\log(RSS) - n\log(n) + log(n) \times (p+1)$$

See the supplemental note on AIC & BIC for derivations.

AIC & BIC

$$AIC = n\log(RSS) - n\log(n) + 2(p+1)$$
$$BIC = n\log(RSS) - n\log(n) + \log(n) \times (p+1)$$

- First Term: Decreases as p increases
- Second Term: Fixed for a given sample size n
- Third Term: Increases as p increases

Using AIC & BIC

$$AIC = n\log(RSS) - n\log(n) + 2(p+1)$$
$$BIC = n\log(RSS) - n\log(n) + \log(n) \times (p+1)$$

- Choose model with smallest AIC or BIC
- If $n \geq 8$, the penalty for BIC is larger than that of AIC, so BIC tends to favor *more parsimonious* models (i.e. models with fewer terms)

Backward Selection

- Start with model that includes all variables of interest
- Drop variables one at a time that are deemed irrelevant based on some criterion. Common criterion include
  - Drop variable with highest p-value over some threshold (e.g. 0.05, 0.1)
  - Drop variable that leads to smallest value of AIC or BIC
- Stop when no more variables can be removed from the model based on the criterion

Forward Selection

- Start with the intercept-only model
- Include variables one at a time based on some criterion. Common criterion include
  - Add variable with smallest p-value under some threshold (e.g. 0.05, 0.1)
  - Add variable that leads to the smallest value of AIC or BIC
- Stop when no more variables can be added to the model based on the criterion