Model comparison

Prof. Maria Tackett

STA 210

[Click here for PDF of slides]

Topics

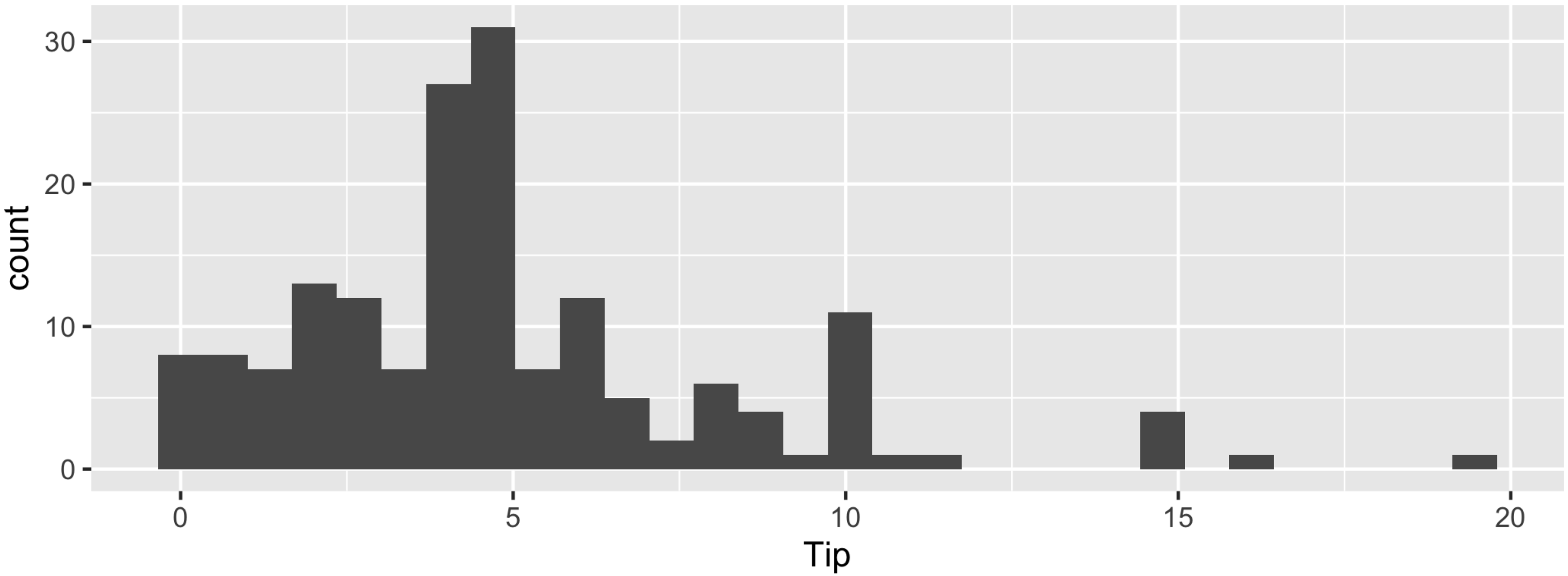
- ANOVA for Multiple Linear Regression
- Nested F Test
- R^2 vs. Adj. R^2
- AIC & BIC

Restaurant tips

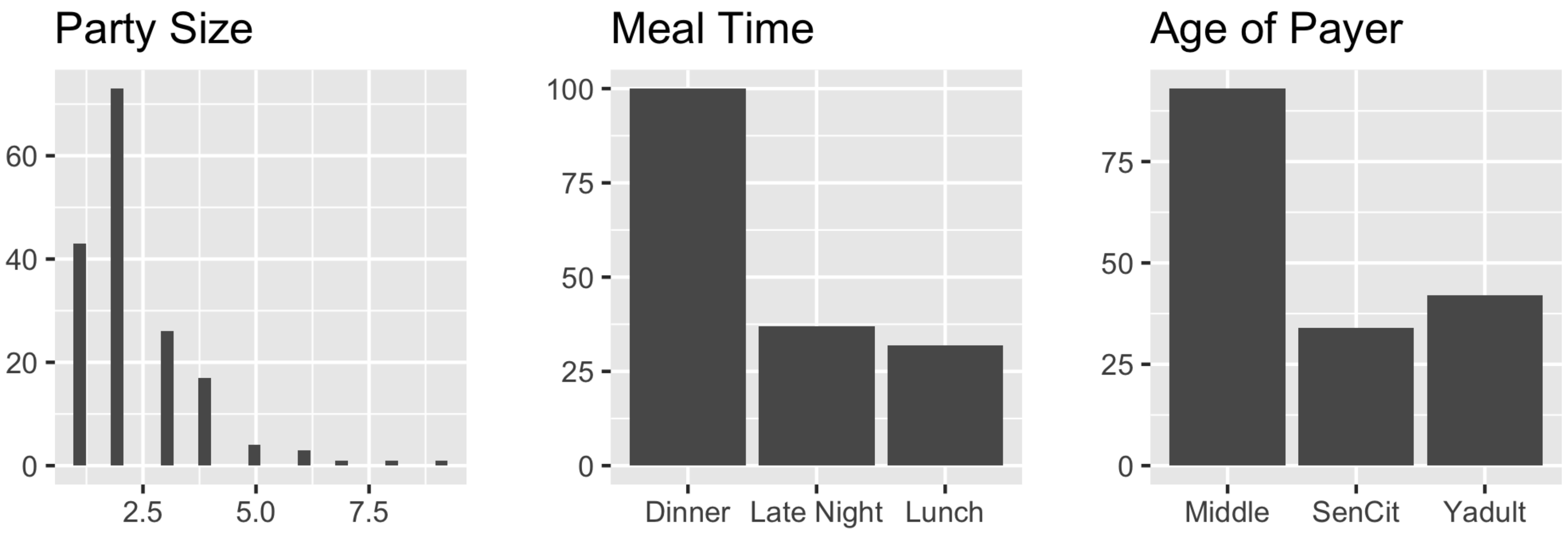
What affects the amount customers tip at a restaurant?

- **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)

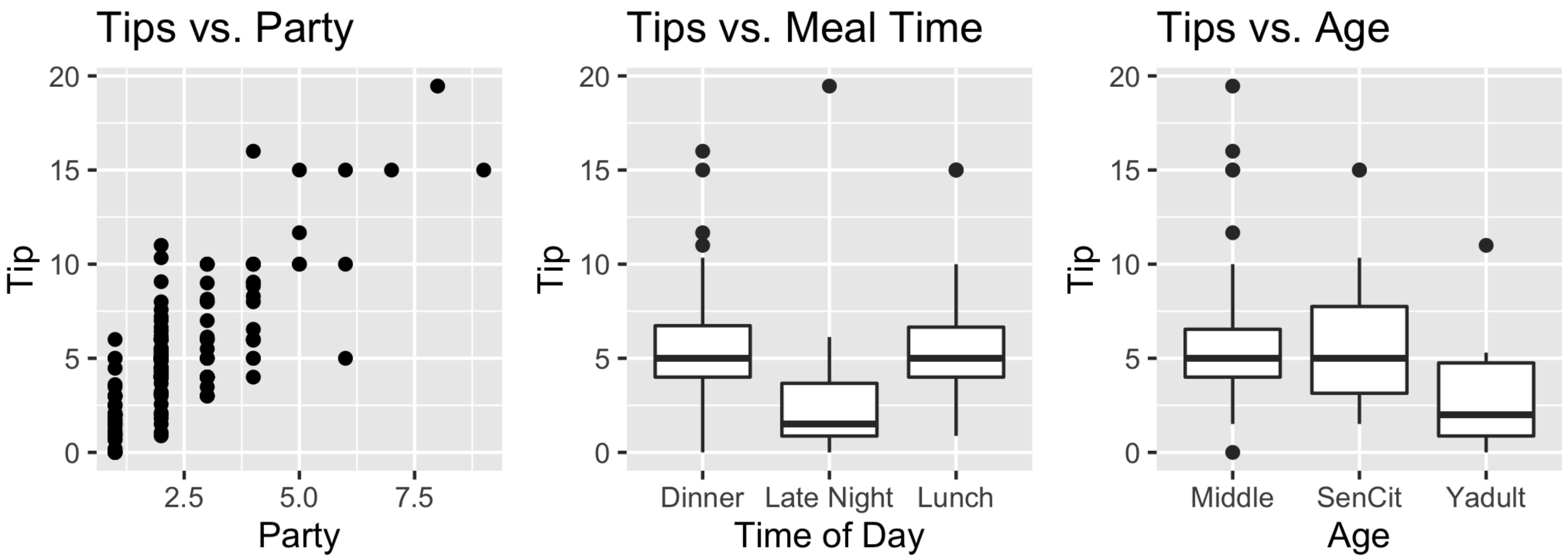
Response Variable



Predictor Variables



Response vs. Predictors



Restaurant tips: model

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 test for MLR

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

Tips: ANOVA Test

term	df	sumsq	meansq	statistic	p.value
Party	1	1188.636	1188.636	285.712	0.000
Age	2	38.028	19.014	4.570	0.012
Residuals	165	686.444	4.160		

Model df: 3

Model SS: 1188.636 + 38.028 = 1226.664

Model MS: 1226.664/ 3 = 408.888

FStat: 408.888 / 4.160 = 98.2903846

P-value: P(F > 98.2903846) ≈ 0

STA 210

Tips: ANOVA Test

term	df	sumsq	meansq	statistic	p.value
Party	1	1188.636	1188.636	285.712	0.000
Age	2	38.028	19.014	4.570	0.012
Residuals	165	686.444	4.160		

The data provide sufficient evidence to conclude that 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**

STA 210

Nested (Partial) F Test

- Suppose we have a full and reduced model:

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

- We want to test whether any of the variables $x_{q+1}, x_{q+2}, \ldots, x_p$ 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

$$F = \frac{\left(SSE_{reduced} - SSE_{full}\right) \Big/ \text{\# predictors tested}}{SSE_{full} \Big/ (n - p_{full} - 1)}$$

- Calculate the p-value using the F distribution with df1 = # predictors tested and df2 = $(n - p_{full} - 1)$

STA 210

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

STA 210

Tips: 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)
```

```
#Nested F test in R
anova(reduced, full)
```

Tips: Nested F test

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

F Stat: $\frac{(686.444-622.979)/2}{622.979/(169-5-1)} = 8.303$

P-value: P(F > 8.303) = 0.0003

- calculated using an F distribution with 2 and 163 degrees of freedom

The data provide sufficient evidence to conclude that at least one coefficient associated with `Meal` is not zero. Therefore, `Meal` is a

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

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

STA 210

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)
```

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

STA 210

Final model for now

We conclude that the effect of **`Party`** does not differ based **`Meal`**. 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 comparision

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: measure that includes a penalty 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

$$R^2 = \frac{SS_{Model}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

$$Adj.\,R^2 = 1 - \frac{SS_{Error}/(n-p-1)}{SS_{Total}/(n-1)}$$

STA 210

Using R^2 and $Adj. R^2$

$Adj. R^2$ 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

Tips: Comparing models

Let's compare two models:

```
model1 <- lm(Tip ~ Party + Age + Meal, data = tips)
glance(model1) %>% select(r.squared, adj.r.squared)
```

```
## # A tibble: 1 x 2
##   r.squared adj.r.squared
##       <dbl>         <dbl>
## 1     0.674         0.664
```

```
model2 <- lm(Tip ~ Party + Age + Meal + Day, data = tips)
glance(model2) %>% select(r.squared, adj.r.squared)
```

```
## # A tibble: 1 x 2
##   r.squared adj.r.squared
##       <dbl>         <dbl>
## 1     0.683         0.662
```

STA 210

AIC & BIC

Akaike's Information Criterion (AIC):

$$AIC = n\log(SS_\text{Error}) - n\log(n) + 2(p+1)$$

Schwarz's Bayesian Information Criterion (BIC)

$$BIC = n\log(SS_\text{Error}) - n\log(n) + log(n) \times (p+1)$$

See the supplemental note on AIC & BIC for derivations.

STA 210

AIC & BIC

$$AIC = n\log(SS_{\text{Error}}) - n\log(n) + 2(p+1)$$
$$BIC = n\log(SS_{\text{Error}}) - n\log(n) + \log(n) \times (p+1)$$

First Term: Decreases as *p* increases

STA 210

AIC & BIC

$$AIC = n\log(SS_{\text{Error}}) - n\log(n) + 2(p+1)$$
$$BIC = n\log(SS_{\text{Error}}) - n\log(n) + \log(n) \times (p+1)$$

Second Term: Fixed for a given sample size *n*

STA 210

AIC & BIC

$$AIC = n\log(SS_{\text{Error}}) - n\log(n) + 2(p+1)$$
$$BIC = n\log(SS_{\text{Error}}) - n\log(n) + \log(n)\times(p+1)$$

Third Term: Increases as *p* increases

Using AIC & BIC

$$AIC = n\log(SS_{Error}) - n\log(n) + 2(p+1)$$
$$BIC = n\log(SS_{Error}) - n\log(n) + \log(n) \times (p+1)$$

- Choose model with the smaller value of 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)

STA 210

Tips: AIC & BIC

```
model1 <- lm(Tip ~ Party + Age + Meal, data = tips)
glance(model1) %>% select(AIC, BIC)
```

```
## # A tibble: 1 x 2
##     AIC   BIC
##   <dbl> <dbl>
## 1  714.  736.
```

```
model2 <- lm(Tip ~ Party + Age + Meal + Day, data = tips)
glance(model2) %>% select(AIC, BIC)
```

```
## # A tibble: 1 x 2
##     AIC   BIC
##   <dbl> <dbl>
## 1  720.  757.
```

Recap

- ANOVA for Multiple Linear Regression
- Nested F Test
- R^2 vs. Adj. R^2
- AIC & BIC