Multiple Linear Regression

Model Diagnostics

Dr. Maria Tackett

10.02.19

Click for PDF of slides

Announcements

- HW 03 **due TODAY at 11:59p**
- HW 04 **due Thursday, October 10 at 11:59p**
- Thursday's lab: Help Hours
- Looking ahead:
  - **Exam 1 on Mon, Oct 14 in class**
  - Practice exam on Sakai
  - Can bring 1 page of notes
  - Exam review on Oct 9
  - Lecture notes, past assignments, and textbook to study

R packages

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

Nested F Test

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)

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

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`**.

Why can we not rely on the individual p-values to determine if 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. What is the baseline level for `Meal`?
2. How do we expect the mean tips to change when `Meal == "Late Night"`, holding Age and Party constant?
3. How does the slope of `Party` change when `Meal == "Late Night"`, holding Age and Party constant?

Nested F test for interactions

Are there any significant interaction effects with Party in the model?

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

```
full <- lm(Tip ~ Party + Age + Meal + Age* Party + 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
159	615.380	4	7.6	0.491	0.742

Final model for now

We conclude that there are no significant interactions with `Party` in the model. Therefore, we will use the original model that only included main effects.

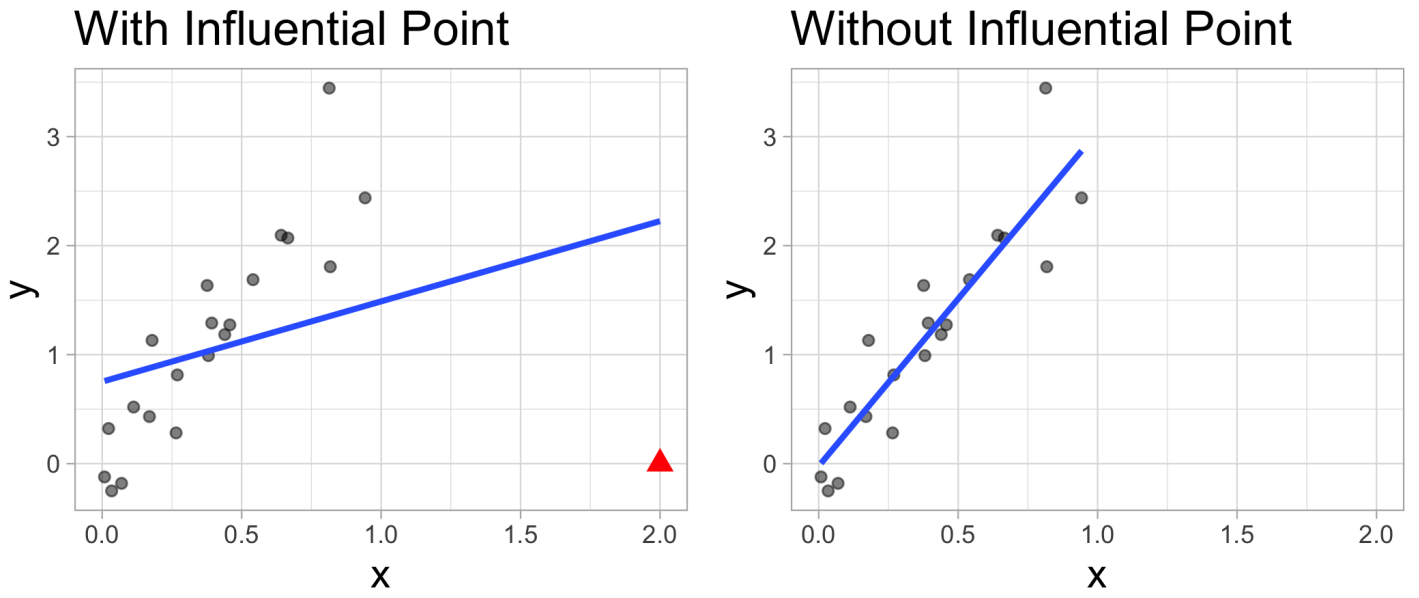
term	estimate	std.error	statistic	p.value
(Intercept)	1.254	0.394	3.182	0.002
Party	1.808	0.121	14.909	0.000
AgeSenCit	0.390	0.394	0.990	0.324
AgeYadult	-0.505	0.412	-1.227	0.222
MealLate Night	-1.632	0.407	-4.013	0.000
MealLunch	-0.612	0.402	-1.523	0.130

Model Diagnostics

Influential and Leverage Points

Influential Observations

An observation is **influential** if removing it substantially changes the coefficients of the regression model



Influential Observations

- In addition to the coefficients, influential observations can have a large impact on the standard errors
- Occasionally these observations can be identified in the scatterplot
    - This is often not the case - especially when dealing with multivariate data
- We will use measures to quantify an individual observation's influence on the regression model
    - **leverage**, **standardized residuals**, and **Cook's distance**

Leverage

- **Leverage:** measure of the distance between an observation's values of the predictor variables and the average values of the predictor variables for the whole data set
- An observation has high leverage if its combination of values for the predictor variables is very far from the typical combinations in the data
    - It is **potentially** an influential point, i.e. may have a large impact on the coefficient estimates and standard errors
- **Note:** Identifying points with high leverage has nothing to do with the values of the response variables

Calculating Leverage

- **Simple Regression:** leverage of the i^{th} observation is

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

- **Multiple Regression:** leverage of the i^{th} observation is the i^{th} diagonal of

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

High Leverage

- Values of leverage are between $\frac{1}{n}$ and 1 for each observation
- The average leverage for all observations in the data set is $\frac{(p+1)}{n}$
- There are different thresholds for determining when an observation has **high leverage**
  - We will use the threshold $h_i > \frac{2(p+1)}{n}$
- Observations with high leverage tend to have small residuals

High Leverage

- Questions to check if you identify points with high leverage:
  - Are they a result of data entry errors?
  - Are they in the scope for the individuals for which you want to make predictions?
  - Are they impacting the estimates of the model coefficients, especially for interactions?
- Just because a point has high leverage does not necessarily mean it will have a substantial impact on the regression. Therefore you should check other measures.

Standardized & Studentized Residuals

- What is the best way to identify outliers (points that don't fit the pattern from the regression line)?
- Look for points that have large residuals
- We want a common scale, so we can more easily identify "large" residuals
- We will look at each residual divided by its standard error

Standardized Residuals

$$std.res_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

- The standard error of a residual, $\hat{\sigma}\sqrt{1 - h_i}$ depends on the value of the predictor variables
- Residuals for observations that are high leverage have smaller variance than residuals for observations that are low leverage
    - This is because the regression line tries to fit high leverage observations as closely as possible

Standardized Residuals

- Values with very large standardized residuals are outliers, since they don't fit the pattern determined by the regression model
- Observations with standardized residuals of magnitude > 2 should be examined more closely
- Observations with large standardized residuals are outliers but may not have an impact on the regression line
- **Good Practice:** Make residual plots with standardized residuals
  - It is easier to identify outliers and check for constant variance assumption

Motivating Cook's Distance

- If a observation has a large impact on the estimated regression coefficients, when we drop that observation...
  - The estimated coefficients should change
  - The predicted Y value for that observation should change
- One way to determine each observation's impact could be to delete it, rerun the regression, compare the predicted Y values from the new and original models
  - This could be very time consuming
- Instead, we can use Cook's Distance which gives a measure of the change in the predicted Y value when an observation is dropped

Cook's Distance

- **Cook's Distance:** Measure of an observation's overall impact, i.e. the effect removing the observation has on the estimated coefficients
- For the i^{th} observation, we can calculate Cook's Distance as

$$D_i = \frac{1}{p}(std.res_i)^2\left(\frac{h_i}{1-h_i}\right)$$

- *Note:* Cook's distance, D_i, incorporates both the residual and the leverage for each observation
- An observation with large D_i is said to have a strong influence on the predicted values

Using these measures

- Standardized residuals, leverage, and Cook's Distance should all be examined together
- Examine plots of the measures to identify observations that may have an impact on your regression model
- Some thresholds for flagging potentially influential observations:
  - **Leverage**: $h_i > \frac{2(p+1)}{n}$ (some software uses $2p/n$)
  - **Standardized Residuals**: $|std.res_i| > 2$
  - **Cook's Distance**: $D_i > 1$

What to do with outliers/influential observations?

- It is OK to drop an observation based on the predictor variables if...
  - It is meaningful to drop the observation given the context of the problem
  - You intended to build a model on a smaller range of the predictor variables. Mention this in the write up of the results and be careful to avoid extrapolation when making predictions
- It is not OK to drop an observation based on the response variable
  - These are legitimate observations and should be in the model
- You can try transformations or increasing the sample size by collecting more data
- In either instance, you can try building the model with and without the outliers/influential observations

Model diagnostics in R

- Use the **`augment`** function in the broom package to output the model diagnostics (along with the predicted values and residuals)
- Output from `augment` :
  - response and predictor variables in the model
  - `.fitted`: predicted values
  - `.se.fit`: standard errors of predicted values
  - `.resid`: residuals
  - **`.hat`**: leverage
  - `.sigma`: estimate of residual standard deviation when corresponding observation is dropped from model
  - **`.cooksd`**: Cook's distance
  - **`.std.resid`**: standardized residuals

Example: 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))
```

Example: Tips

```
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

Using `augment` function

- Use the `augment` function to add predicted values and model diagnostics to data
  - Add the observation number for diagnostic plots

```
tips_output <- augment(model1) %>%
  mutate(obs_num = row_number())
```

Augmented data

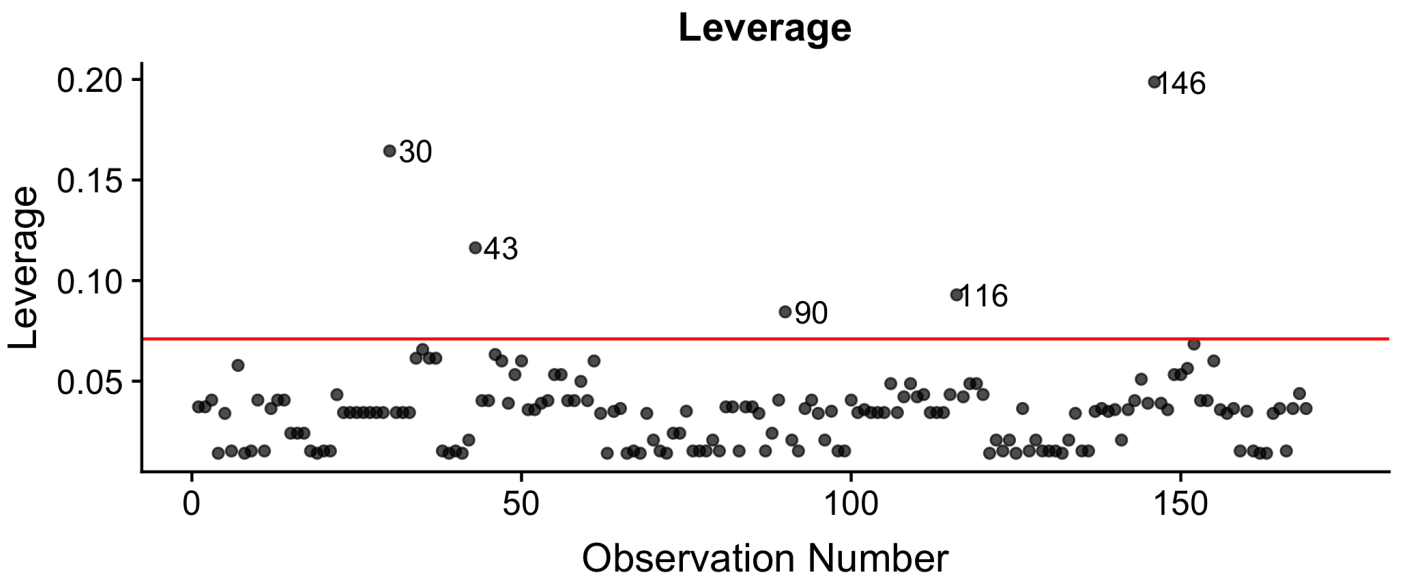
```
glimpse(tips_output)
```

```
## Observations: 169
## Variables: 12
## $ Tip        <dbl> 2.99, 2.00, 5.00, 4.00, 10.34, 4.85, 5.00, 4.00, 5.00
## $ Party      <dbl> 1, 1, 1, 3, 2, 2, 4, 3, 2, 1, 2, 2, 1, 1, 1, 1, 1, 2,
## $ Meal       <chr> "Dinner", "Dinner", "Dinner", "Dinner", "Dinner", "Di
## $ Age        <chr> "Yadult", "Yadult", "SenCit", "Middle", "SenCit", "Mi
## $ .fitted    <dbl> 2.5562830, 2.5562830, 3.4515838, 6.6766419, 5.2591209
## $ .se.fit    <dbl> 0.3771863, 0.3771863, 0.3939434, 0.2327069, 0.3604347
## $ .resid     <dbl> 0.43371698, -0.55628302, 1.54841620, -2.67664190, 5.0
## $ .hat       <dbl> 0.03722423, 0.03722423, 0.04060519, 0.01416878, 0.033
## $ .sigma     <dbl> 1.960700, 1.960502, 1.957071, 1.949536, 1.918486, 1.9
## $ .cooksd    <dbl> 3.294208e-04, 5.419132e-04, 4.612379e-03, 4.554814e-0
## $ .std.resid <dbl> 0.226100143, -0.289994804, 0.808622859, -1.378941928,
## $ obs_num    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
```

Leverage

```
leverage_threshold <- 2*(5+1)/nrow(tips)
```

```
ggplot(data = tips_output, aes(x = obs_num,y = .hat)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept = leverage_threshold,color = "red")+
  labs(x = "Observation Number",y = "Leverage",title = "Leverage")
  geom_text(aes(label=ifelse(.hat > leverage_threshold, as.charact
```



Points with high leverage

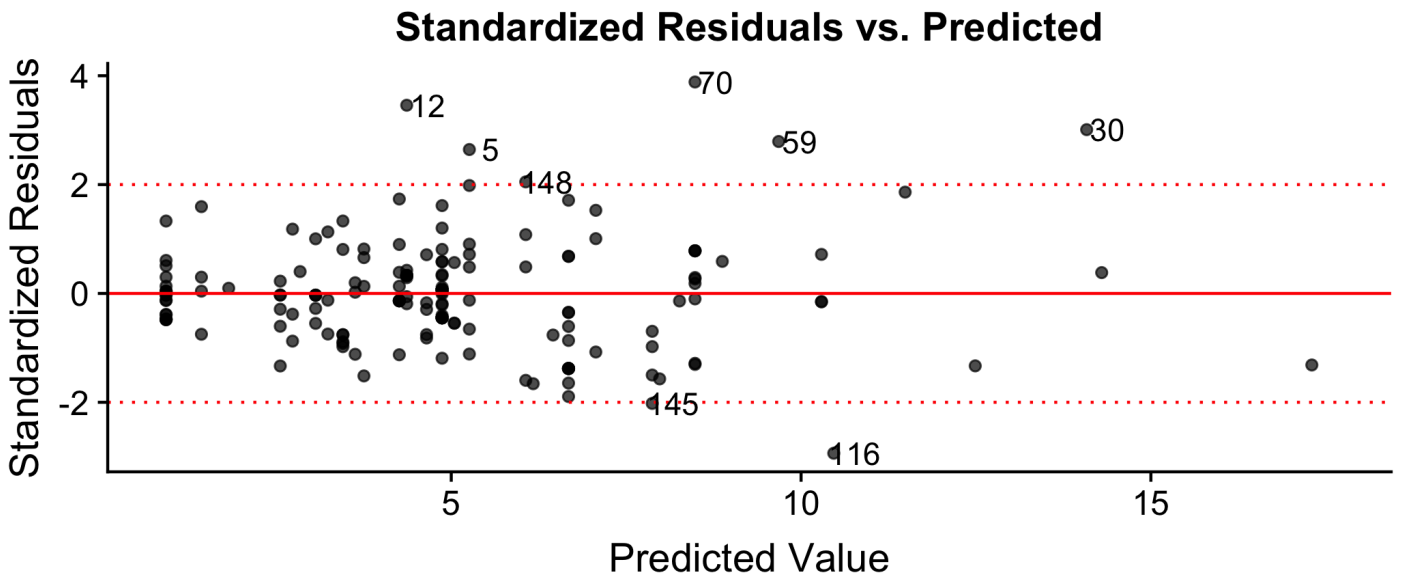
```
tips_output %>% filter(.hat > leverage_threshold) %>%
  select(Party, Meal, Age)
```

```
## # A tibble: 5 x 3
##   Party Meal       Age
##   <dbl> <chr>      <chr>
## 1     8 Late Night Middle
## 2     7 Dinner     SenCit
## 3     6 Dinner     SenCit
## 4     6 Late Night Middle
## 5     9 Lunch      SenCit
```

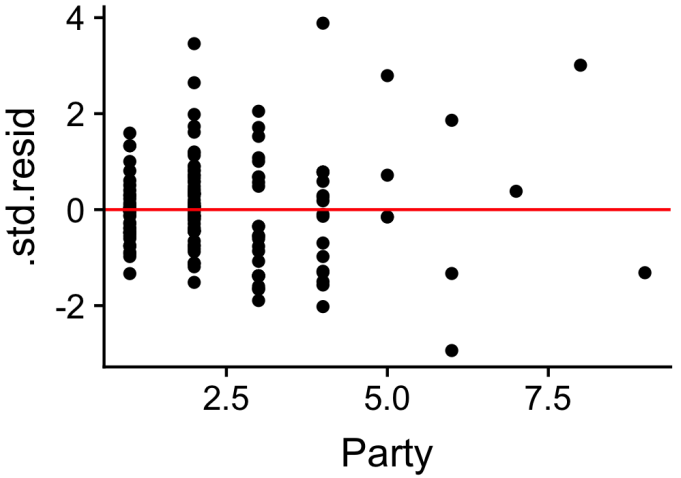
Why do you think these points have high leverage?

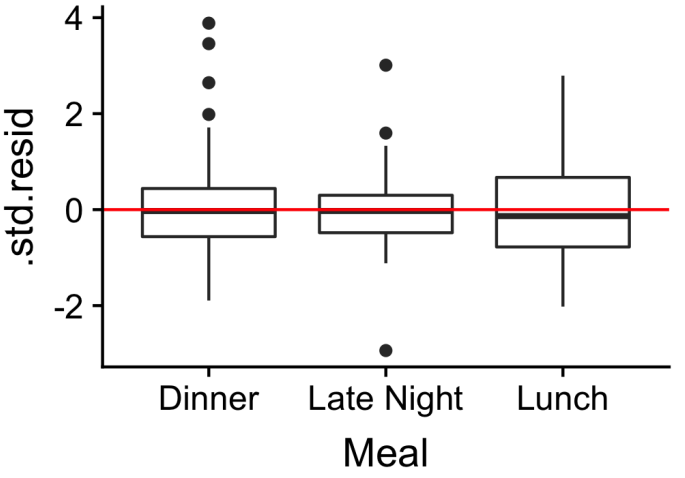
Standardized Residuals vs. Predicted

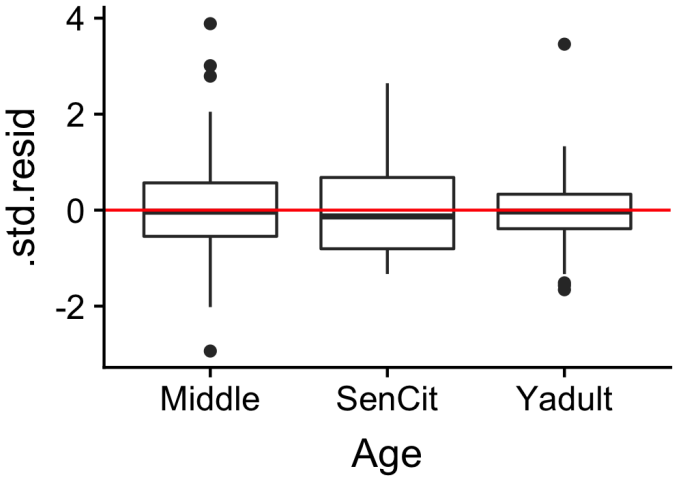
```
ggplot(data = tips_output, aes(x = .fitted,y = .std.resid)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept = 0,color = "red") +
  geom_hline(yintercept = -2,color = "red",linetype = "dotted") +
  geom_hline(yintercept = 2,color = "red",linetype = "dotted") +
  labs(x ="Predicted Value",y ="Standardized Residuals",title = "S
  geom_text(aes(label = ifelse(abs(.std.resid) >2,as.character(obs
```



Standardized residuals vs. predictors







Points with large magnitude std.res.

```
tips_output %>% filter(abs(.std.resid) > 2) %>%
  select(Party, Meal, Age, Tip)
```

```
## # A tibble: 8 x 4
##   Party Meal       Age      Tip
##   <dbl> <chr>      <chr>  <dbl>
## 1     2 Dinner     SenCit  10.3
## 2     2 Dinner     Yadult  11
## 3     8 Late Night Middle  19.5
## 4     5 Lunch      Middle  15
## 5     4 Dinner     Middle  16
## 6     6 Late Night Middle   5
## 7     4 Lunch      Middle   4
## 8     3 Lunch      Middle  10
```

- Why do you think these points have standardized residuals with large magnitude?
- What other variables could you examine?

Why we want to find outliers

Estimate of regression standard deviation, $\hat{\sigma}$, using all observations

```
glance(model1)$sigma
```

```
## [1] 1.954983
```

Estimate of $\hat{\sigma}$ without points with large magnitude standardized residuals

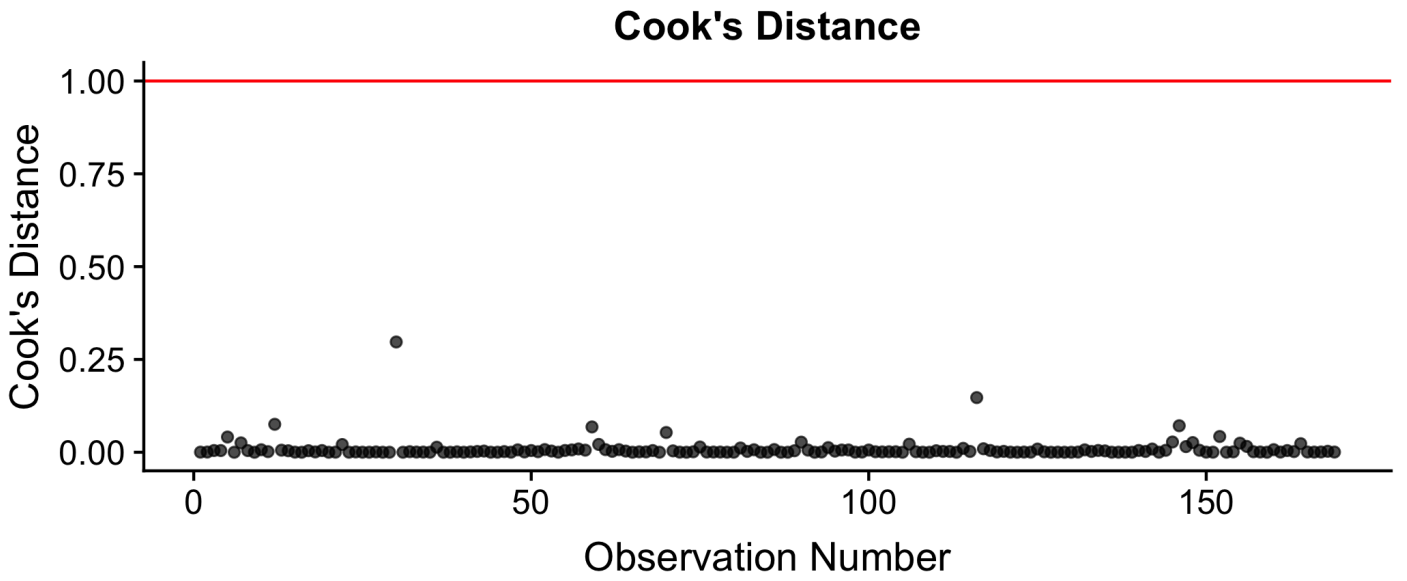
```
tips_output %>%
  filter(abs(.std.resid) <= 2) %>%
  summarise(sigma_est = sqrt(sum(.resid^2)/(n() - 5 - 1)))
```

```
## # A tibble: 1 x 1
##   sigma_est
##       <dbl>
## 1      1.56
```

Recall that we use $\hat{\sigma}$ to calculate the standard errors for all confidence intervals and p-values, so outliers can affect conclusions drawn from model

Cook's Distance

```
ggplot(data = tips_output, aes(x = obs_num, y = .cooksd)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept=1,color = "red")+
  labs(x= "Observation Number",y = "Cook's Distance",title = "Cook
  geom_text(aes(label = ifelse(.hat>1,as.character(obs_num),"")))
```



See the supplemental notes Details on Model Diagnostics for more details about standardized residuals, leverage points, and Cook's distance.

Multicollinearity

Why multicollinearity is a problem

- We can't include two variables that have a perfect linear association with each other
- If we did so, we could not pick a unique best fit model

Why multicollinearity is a problem

- Ex. Suppose the true population regression equation is $y = 3 + 4x$
- Suppose we try estimating that regression model using the variables x and $z = x/10$

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 \frac{x}{10} \\ &= \hat{\beta}_0 + \left(\hat{\beta}_1 + \frac{\hat{\beta}_2}{10}\right) x\end{aligned}$$

- We can set $\hat{\beta}_1$ and $\hat{\beta}_2$ to any two numbers such that $\hat{\beta}_1 + \frac{\hat{\beta}_2}{10} = 4$
  - We are unable then to choose the "best" combination of $\hat{\beta}_1$ and $\hat{\beta}_2$

Why multicollinearity is a problem

- When we have almost perfect collinearities (i.e. highly correlated explanatory variables), the standard errors for our regression coefficients inflate
- In other words, we lose precision in our estimates of the regression coefficients

Detecting Multicollinearity

Multicollinearity may occur when...

- There are very high correlations $(r > 0.9)$ among two or more explanatory variables, especially for smaller sample sizes
- One (or more) explanatory variables is an almost perfect linear combination of the others
- Include quadratic terms without first mean-centering the variables before squaring
- Including interactions with two or more continuous variables

Detecting Multicollinearity

- Look at a correlation matrix of the predictor variables, including all indicator variables
    - Look out for values close to 1 or -1
- If you think one predictor variable is an almost perfect linear combination of other predictor variables, you can run a regression of that predictor variable vs. the others and see if R^2 is close to 1

Detecting Multicollinearity (VIF)

- Variance Inflation Factor (VIF): Measure of multicollinearity

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

where $R^2_{X_j|X_{-j}}$ is the proportion of variation X that is explained by the linear combination of the other explanatory variables in the model.

- Typically $VIF > 10$ indicates concerning multicollinearity
- Use the `vif()` function in the `rms` package to calculate VIF

Tips VIF

- Calculate VIF using the **`vif`** function in the rms package

```
library(rms)
tidy(vif(model1))
```

```
## # A tibble: 5 x 2
##   names              x
##   <chr>          <dbl>
## 1 Party           1.19
## 2 MealLate Night  1.25
## 3 MealLunch       1.09
## 4 AgeSenCit       1.10
## 5 AgeYadult       1.40
```

Calculating VIF for `Party`

```
party_model <- lm(Party ~ Meal + Age, data=tips)
r.sq <- glance(party_model)$r.squared
(vif <- 1/(1-r.sq))
```

```
## [1] 1.193821
```

Calculating VIF for `MealLateNight`

```
# create indicator variables for Meal
tips <- tips %>%
  mutate(late_night = if_else(Meal=="Late Night",1,0),
         lunch = if_else(Meal=="Lunch",1,0))
```

```
late_night_model <- lm(late_night ~ lunch + Party + Age, data=tips
r.sq <- glance(late_night_model)$r.squared
(vif <- 1/(1-r.sq))
```

```
## [1] 1.250908
```