PM 592
Regression Analysis for Public Health Data Science

Week 5

Multiple Regression

Multiple Regression

Shared Covariance

The Multiple Regression Model

Multiple Regression Assumptions

Collinearity

Regression Diagnostics

Lecture Objectives

- Explain the effect that shared covariance of independent variables has on predicting an outcome.
- Write the form of the multiple regression model.
- Interpret coefficients from a multiple regression model.
- Interpret the R-squared value from a multiple regression model.
- Diagnose collinearity effects.
- Perform a comprehensive analysis for diagnosing outliers and influential points in a linear regression model.

- ✓ How to assess the assumptions of linear regression
- ✓ Explain the concept behind the ANOVA table
- ✓ Advantages and disadvantages of transforming Y (and X)
- ✓ Correlation and its relation to regression
- ✓ Interpreting the slope for a binary X variable
- ✓ Dummy variable sets for categorical predictors

Example

Suppose researchers collected the following data from participants at a previous ToonCon:

`sw` – Liking of the Star Wars franchise (0-100 scale)

`st` – Liking of the Star Trek franchise (0-100 scale)

`sa` – Score on a “social adjustment” scale (0-100)

`male` – Gender of participant

If we want to know the relationship between the liking of these two franchises and social adjustment score, one possible approach would be to perform **several univariable linear regression models**.

```
> lm(sa ~ sw, data = data5) %>% summary()

Call:
lm(formula = sa ~ sw, data = data5)

Residuals:
    Min      1Q  Median      3Q     Max 
-30.095  -9.107   2.049   8.616  33.240 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 95.03918    4.69294   20.25   <2e-16 ***
sw          -0.92798    0.08812  -10.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.64 on 98 degrees of freedom
Multiple R-squared:  0.5309,	Adjusted R-squared:  0.5261 
F-statistic: 110.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

Each unit increase in liking Star Wars is associated with a 0.93-unit decrease in predicted Social Adjustment score.

If we want to know the relationship between the liking of these two franchises and social adjustment score, one possible approach would be to perform **several univariable linear regression models**.

```
> lm(sa ~ st, data = data5) %>% summary()

Call:
lm(formula = sa ~ st, data = data5)

Residuals:
    Min      1Q  Median      3Q     Max 
-38.410  -8.285   0.513  10.936  29.081 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 102.75775    5.74864   17.88  < 2e-16 ***
st           -0.96413    0.09769   -9.87 2.32e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.1 on 98 degrees of freedom
Multiple R-squared:  0.4985,	Adjusted R-squared:  0.4934 
F-statistic: 97.41 on 1 and 98 DF,  p-value: 2.321e-16
```

Each unit increase in liking Star Trek is associated with a 0.96-unit decrease in predicted Social Adjustment score.

If we want to know the relationship between the liking of these two franchises and social adjustment score, one possible approach would be to perform **several univariable linear regression models**.

```
> lm(sa ~ male, data = data5) %>% summary()

Call:
lm(formula = sa ~ male, data = data5)

Residuals:
    Min      1Q  Median      3Q     Max 
-40.804 -13.506  -0.517  13.332  39.082 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   53.381      2.868  18.610  < 2e-16 ***
male         -10.230      3.868  -2.645  0.00951 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.24 on 98 degrees of freedom
Multiple R-squared:  0.06663,	Adjusted R-squared:  0.05711 
F-statistic: 6.996 on 1 and 98 DF,  p-value: 0.009515
```

Being male is associated with a 10.2-unit decrease in predicted Social Adjustment score.

It appears that the effect of each of our variables is as follows:

Model	**Beta**	**P-Value**	**R^2**
1: SW	-0.93	<0.001	0.53
2: ST	-0.96	<0.001	0.50
3: Male	-10.2	0.01	0.07

So we conclude that liking Star Wars, Star Trek, and being male are all associated with lower predicted social adjustment scores.

Is there more to this story, though?

When we run several separate regressions, we inherently assume all the effects are independent of each other.

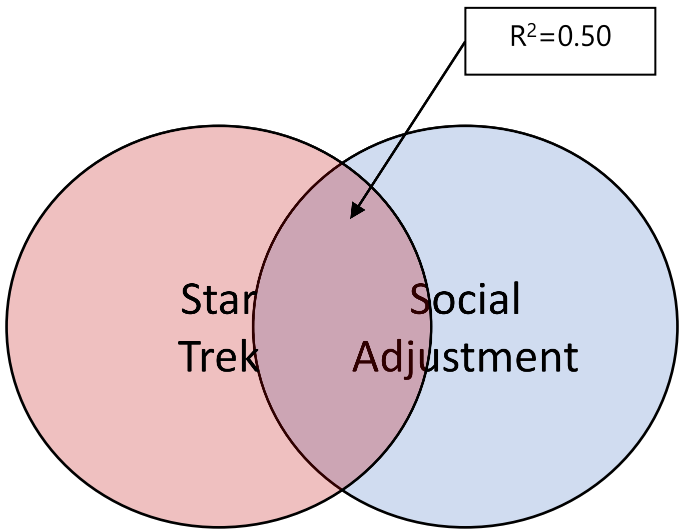
Model	Beta	P-Value	R^2
1: SW	-0.93	<0.001	0.53
2: ST	-0.96	<0.001	0.50
3: Male	-10.2	0.01	0.07

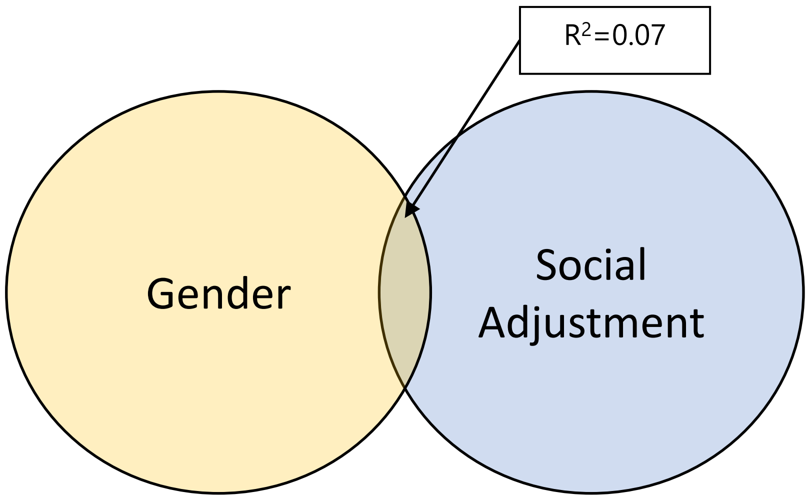
But if these effects were independent of each other (r=0 among all X), then the regression R^2 for all of them combined would be 0.53 + 0.50 + 0.07 = 110%. This clearly can't be the true!

In each model, the dependent variable explains a particular portion of the variance in the outcome.

For Liking Star Trek, this is 50%. For gender, it is 7%.

We can visualize this as follows:

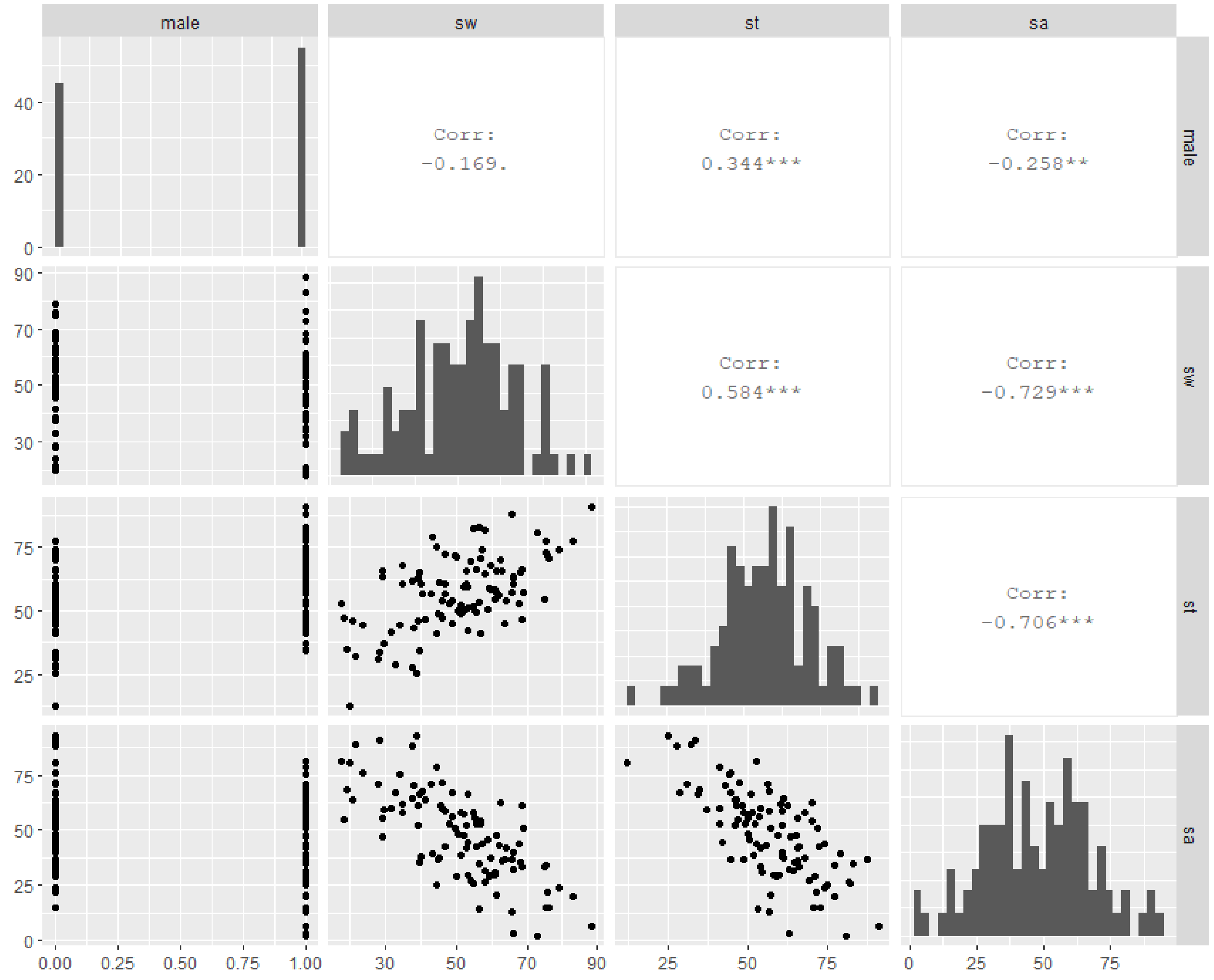




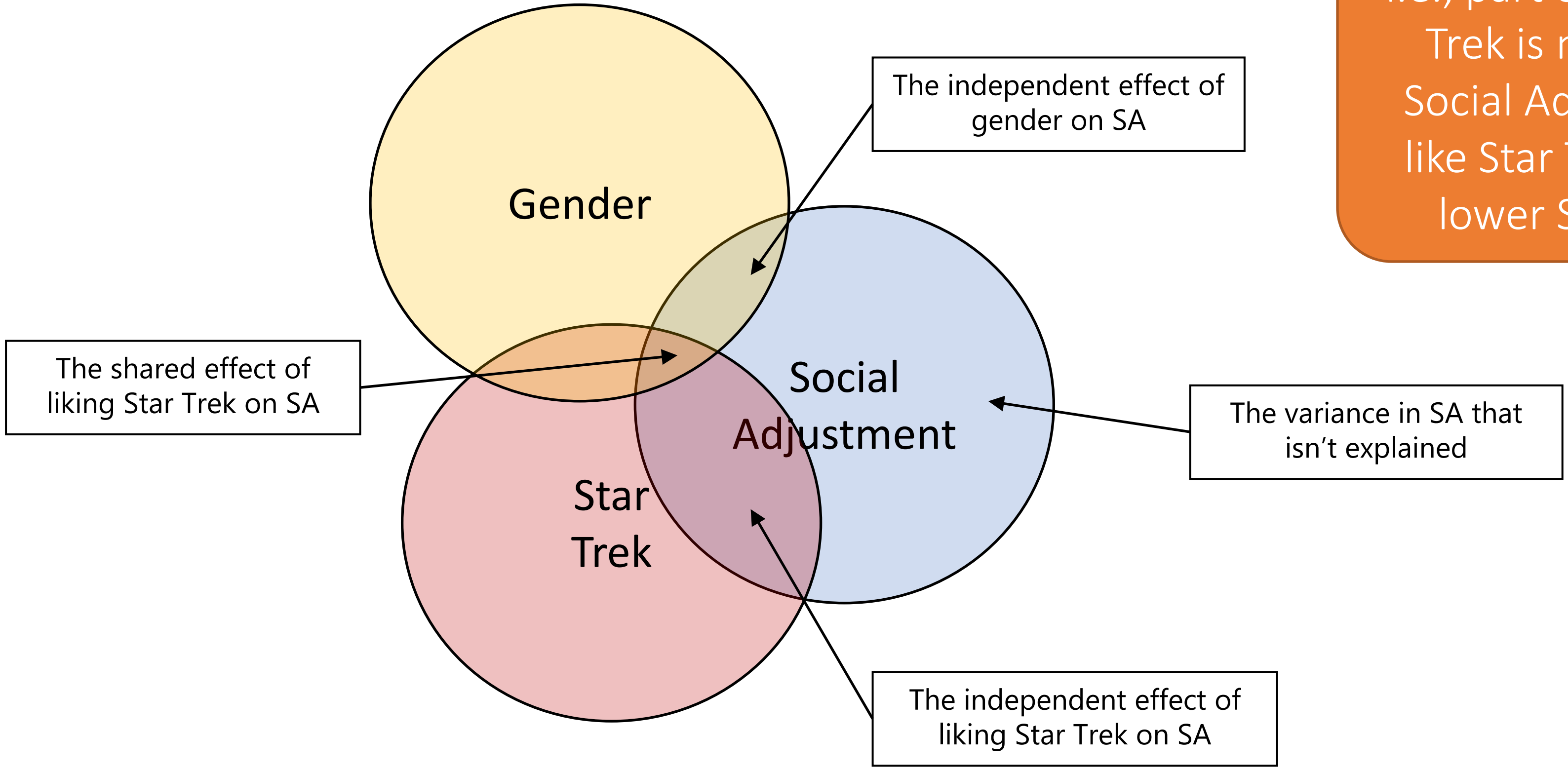
2. Shared Covariance

But here we see that all of our predictor variables are correlated to some extent!

Liking Star Trek is correlated with being male.



Because of the correlation between liking Star Trek and being male, they do not contribute <u>independent</u> effects on Social Adjustment.



i.e., part of the reason why liking Star Trek is negatively associated with Social Adjustment is because males like Star Trek more, and males have lower Social Adjustment scores.

We can combine these effects into a **multiple regression equation**.

```
> lm(sa ~ sw + st + male, data = data5) %>% summary()

Call:
lm(formula = sa ~ sw + st + male, data = data5)

Residuals:
    Min      1Q  Median      3Q     Max 
-24.760  -7.393  -1.535   9.472  25.327 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 113.77223    4.66003  24.414  < 2e-16 ***
sw           -0.81597    0.09935  -8.213 1.00e-12 ***
st           -0.31905    0.11180  -2.854  0.00529 ** 
male        -11.34294    2.67302  -4.243 5.08e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.92 on 96 degrees of freedom
Multiple R-squared:  0.7056,	Adjusted R-squared:  0.6964 
F-statistic:  76.7 on 3 and 96 DF,  p-value: < 2.2e-16
```

1) We see that 71% of the variance in SA is explained by these three predictors; lower than the sum of the individual single regressions.

2) We also see that the parameter estimates have changed. When the effects are taken *together*, the effects of SW and ST have been attenuated.

Recap

- Sometimes X variables are related to each other as well as to Y
- When X variables are related to each other, their effect on Y may be "shared" with other X variables

Recap

- Explain the concept of “shared covariance” among X variables and how it affects their relationship with Y.

What did we do here? We combined ST, SW, and male into a single **multiple regression model**.

- Let $\mu_{Y|X_1,X_2,\ldots,X_K}$ denote the mean value of Y for a given set of X variables.
- Let $\sigma^2_{Y|X_1,X_2,\ldots,X_K}$ denote the corresponding variance of Y for a given set of X variables.

We define the multiple regression equation as:

$$\mu_{Y|X_1,X_2,\ldots,X_k} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$$

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + e$$

e is a residual with mean 0 and variance σ^2

Parameter Interpretations

β_0: The estimated mean when <u>all</u> X variables equal 0.

β_k: The estimated difference in mean Y associated with a 1-unit change in X_k, **when all other X variables are held constant.**

The slope estimates represent partial regression coefficients. The following terminology is interchangeably used:

- Holding all other X constant
- Controlling for all other X
- Partialling out all other X
- Adjusting for all other X
- Taking all other X into account

What is the best-fit multiple regression model?

```
> lm(sa ~ sw + st + male, data = data5) %>% summary()

Call:
lm(formula = sa ~ sw + st + male, data = data5)

Residuals:
    Min      1Q  Median      3Q     Max 
-24.760  -7.393  -1.535   9.472  25.327 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 113.77223    4.66003  24.414  < 2e-16 ***
sw           -0.81597    0.09935  -8.213 1.00e-12 ***
st           -0.31905    0.11180  -2.854  0.00529 ** 
male        -11.34294    2.67302  -4.243 5.08e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.92 on 96 degrees of freedom
Multiple R-squared:  0.7056,	Adjusted R-squared:  0.6964 
F-statistic:  76.7 on 3 and 96 DF,  p-value: < 2.2e-16
```

$$\hat{Y} = 113.77 - 0.81X_{SW} - 0.32X_{ST} - 11.34X_{MALE}$$

What is the best-fit multiple regression model?

```
> lm(sa ~ sw + st + male, data = data5) %>% summary()

Call:
lm(formula = sa ~ sw + st + male, data = data5)

Residuals:
    Min      1Q  Median      3Q     Max 
-24.760  -7.393  -1.535   9.472  25.327 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 113.77223    4.66003  24.414  < 2e-16 ***
sw           -0.81597    0.09935  -8.213 1.00e-12 ***
st           -0.31905    0.11180  -2.854  0.00529 ** 
male        -11.34294    2.67302  -4.243 5.08e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.92 on 96 degrees of freedom
Multiple R-squared:  0.7056,	Adjusted R-squared:  0.6964 
F-statistic:  76.7 on 3 and 96 DF,  p-value: < 2.2e-16
```

"A 1-point increase in liking Star Wars is associated with a 0.81-point decrease in Social Adjustment score, adjusting for liking Star Trek and gender."

For people of the same gender and same ST score, a 1-point increase in liking Star Wars is associated with a 0.81-point decrease in Social Adjustment score.

$$\hat{Y} = 113.77 - 0.81X_{SW} - 0.32X_{ST} - 11.34X_{MALE}$$

What is the interpretation of R^2?

- The multiple R^2 now reflects the proportion of variation in Y that is explained by all the X variables in the model.
- In single linear regression $\sqrt{R^2} = R$ reflected the correlation between X and Y.
- In multiple linear regression, multiple R reflects the correlation between *an optimally-weighted linear combination of independent variables* (i.e., our predicted value of Y) and the actual value of Y.
- Unlike r, multiple R always ranges between 0 and 1.

Practice Questions

$$\hat{Y} = 113.77 - 0.81X_{SW} - 0.32X_{ST} - 11.34X_{MALE}$$

1) What is the estimated difference in Y for a 10-point increase in SW score, holding ST and MALE constant?

2) What is the estimated difference in Y for a male who scored 75 on both SW and ST scales, vs. a female who scored 50 on both SW and ST scales?

Practice Questions

$$\hat{Y} = 113.77 - 0.81X_{SW} - 0.32X_{ST} - 11.34X_{MALE}$$

1) What is the estimated difference in Y for a 10-point increase in SW score, holding ST and MALE constant?

$$\widehat{Y_1} = 113.77 - 0.81X_{SW} - 0.32X_{ST} - 11.34X_{MALE}$$

$$\widehat{Y_2} = 113.77 - 0.81(X_{SW} + 10) - 0.32X_{ST} - 11.34X_{MALE}$$

$$\hat{Y}_2 - \hat{Y}_1 = -0.81\,(X_{SW} + 10) - (-0.81)(X_{SW}) = -8.1$$

Alternately, we know $\beta_{SW} = -0.81$ is the effect on Y for a 1-unit increase in X_{SW}, so the effect for a 10-unit increase would be 10(-0.81) = -8.1

Practice Questions

$$\hat{Y} = 113.77 - 0.81X_{SW} - 0.32X_{ST} - 11.34X_{MALE}$$

2) What is the estimated difference in Y for a male who scored 75 on both SW and ST scales, vs. a female who scored 50 on both SW and ST scales?

$$\widehat{Y_1} = 113.77 - 0.81(50) - 0.32(50) - 11.34(0)$$

$$\widehat{Y_2} = 113.77 - 0.81(75) - 0.32(75) - 11.34(1)$$

$$\hat{Y}_2 - \hat{Y}_1 = (113.77 - 113.77) - 0.81X_{SW}(75 - 50) - 0.32(75 - 50) - 11.34(1 - 0)$$

$$= (113.77 - 113.77) - 0.81(25) - 0.32(25) - 11.34(1) = -39.59$$

A male who scored 75 on both SW and ST scales is predicted to have a 39.59-point lower SA score compared to a female who scored 50 on both SW and ST scales.

Recap

- The slope coefficients in multiple regression models must be interpreted *relative to the other variables in the model*
- Multiple R^2 reflects the variance in Y that is explained by all X variables collectively

Recap

- Interpret slope estimates from a multiple regression model
- Interpret intercepts from a multiple regression model
- Explain the concept of R^2
- Explain how R^2 can be interpreted in the context of a correlation coefficient

Test Yourself

Covid patients' blood lab values were taken, including CO2. Is age statistically associated with CO2 level?

a) Yes. Older paitents tend to have lower CO2 levels (p=0.004).

b) No. Age only explains 2.1% of the variance in CO2 levels.

c) No. The median of the residuals is 0.50.

```
> lm(co2 ~ age, data=dat14387A) %>% summary()

Call:
lm(formula = co2 ~ age, data = dat14387A)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.1219  -1.6192   0.5037   2.4885  11.3808 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.09683    0.73134  32.949  < 2e-16 ***
age         -0.03591    0.01236  -2.906  0.00387 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.985 on 394 degrees of freedom
  (47 observations deleted due to missingness)
Multiple R-squared:  0.02098,	Adjusted R-squared:  0.01849 
F-statistic: 8.443 on 1 and 394 DF,  p-value: 0.003872
```

Test Yourself

Covid patients' blood lab values were taken, including CO2. Is age statistically associated with CO2 level?

a) Yes. Older paitents tend to have lower CO2 levels (p=0.004).

b) No. Age only explains 2.1% of the variance in CO2 levels.

c) No. The median of the residuals is 0.50.

```
> lm(co2 ~ age, data=dat14387A) %>% summary()

Call:
lm(formula = co2 ~ age, data = dat14387A)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.1219  -1.6192   0.5037   2.4885  11.3808 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.09683    0.73134  32.949  < 2e-16 ***
age         -0.03591    0.01236  -2.906  0.00387 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.985 on 394 degrees of freedom
  (47 observations deleted due to missingness)
Multiple R-squared:  0.02098,	Adjusted R-squared:  0.01849 
F-statistic: 8.443 on 1 and 394 DF,  p-value: 0.003872
```

Test Yourself

Presence of congestion was added as a predictor in this model. What is the correct interpretation of the congestion beta coefficient?

a) Individuals with congestion are predicted to have 0.57ppm higher CO2 levels than those without congestion.

b) A 0.57-unit increase in congestion is associated with a 0.029-unit decrease in age.

c) Individuals with congestion are predicted to have higher CO2 than those without congestion, holding age constant.

```
> lm(co2 ~ age + congestion, data=dat14387A) %>% summary()

Call:
lm(formula = co2 ~ age + congestion, data = dat14387A)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.6132  -1.6135   0.3298   2.3868  11.2152 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 23.04056    0.84722  27.196   <2e-16 ***
age         -0.02855    0.01267  -2.253   0.0248 *  
congestion   0.57158    0.22322   2.561   0.0108 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.891 on 375 degrees of freedom
  (65 observations deleted due to missingness)
Multiple R-squared:  0.03807,	Adjusted R-squared:  0.03294 
F-statistic: 7.421 on 2 and 375 DF,  p-value: 0.0006906
```

Test Yourself

Presence of congestion was added as a predictor in this model. What is the correct interpretation of the congestion beta coefficient?

a) Individuals with congestion are predicted to have 0.57ppm higher CO2 levels than those without congestion.

b) A 0.57-unit increase in congestion is associated with a 0.029-unit decrease in age.

c) Individuals with congestion are predicted to have higher CO2 than those without congestion, holding age constant.

```
> lm(co2 ~ age + congestion, data=dat14387A) %>% summary()

Call:
lm(formula = co2 ~ age + congestion, data = dat14387A)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.6132  -1.6135   0.3298   2.3868  11.2152 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 23.04056    0.84722  27.196   <2e-16 ***
age         -0.02855    0.01267  -2.253   0.0248 *  
congestion   0.57158    0.22322   2.561   0.0108 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.891 on 375 degrees of freedom
  (65 observations deleted due to missingness)
Multiple R-squared:  0.03807,	Adjusted R-squared:  0.03294 
F-statistic: 7.421 on 2 and 375 DF,  p-value: 0.0006906
```

Test Yourself

Fatigue was added as a predictor to this model. What is the predicted difference in CO_2 for a 48-year-old patient with congestion and fatigue, compared to a same-age patient without either symptom?

```
> lm(co2 ~ age + congestion + fatigue, data=dat14387A) %>% summary()

Call:
lm(formula = co2 ~ age + congestion + fatigue, data = dat14387A)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.6518  -1.6245   0.3799   2.3382  11.2978 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 22.86329    0.90281  25.325   <2e-16 ***
age         -0.02736    0.01289  -2.123   0.0344 *  
congestion   0.43564    0.32214   1.352   0.1771    
fatigue      0.21089    0.35458   0.595   0.5524    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.9 on 373 degrees of freedom
  (66 observations deleted due to missingness)
Multiple R-squared:  0.03912,	Adjusted R-squared:  0.03139 
F-statistic: 5.062 on 3 and 373 DF,  p-value: 0.0019
```

Test Yourself

Fatigue was added as a predictor to this model. What is the predicted difference in CO2 for a 48-year-old patient with congestion and fatigue, compared to a same-age patient without either symptom?

$$\hat{Y}_B = 22.863 - 0.027(48) + 0.436(1) + 0.211(1) = 22.214$$

$$\hat{Y}_A = 22.863 - 0.027(48) + 0.436(0) + 0.211(0) = 21.567$$

$$\hat{Y}_B - \hat{Y}_A = 22.214 - 21.567 = 0.647$$

```
> lm(co2 ~ age + congestion + fatigue, data=dat14387A) %>% summary()

Call:
lm(formula = co2 ~ age + congestion + fatigue, data = dat14387A)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.6518  -1.6245   0.3799   2.3382  11.2978 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 22.86329    0.90281  25.325   <2e-16 ***
age         -0.02736    0.01289  -2.123   0.0344 *  
congestion   0.43564    0.32214   1.352   0.1771    
fatigue      0.21089    0.35458   0.595   0.5524    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.9 on 373 degrees of freedom
  (66 observations deleted due to missingness)
Multiple R-squared:  0.03912,	Adjusted R-squared:  0.03139 
F-statistic: 5.062 on 3 and 373 DF,  p-value: 0.0019
```

4. Multiple Regression Model Assumptions

Overall, the model assumptions for multiple regression are the same as for single regression.

We will go over a couple of nuances as we look at multiple independent variables.

Linearity

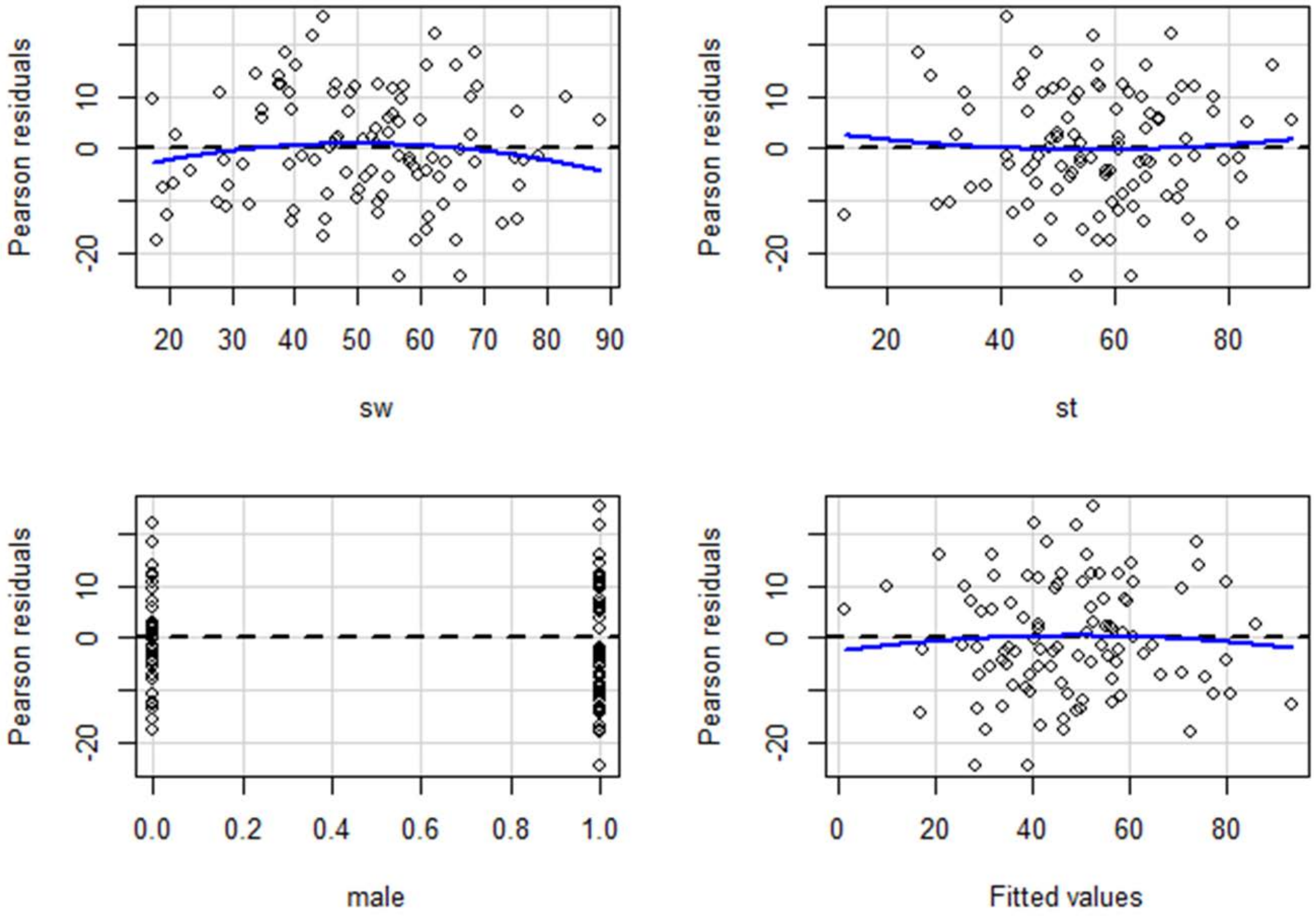
- The mean of Y is a linear function of the independent variables.
- i.e., $\mu_{Y|X_1,X_2,\ldots,X_k} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$
- Because all the X variables are included together, we can't just look at a plot of Y vs. any specific X; we'll have to examine the residuals for this.

4. Multiple Regression Model Assumptions

The CAR (Companion to Applied Regression) package gives us some more tools for evaluating regression models.

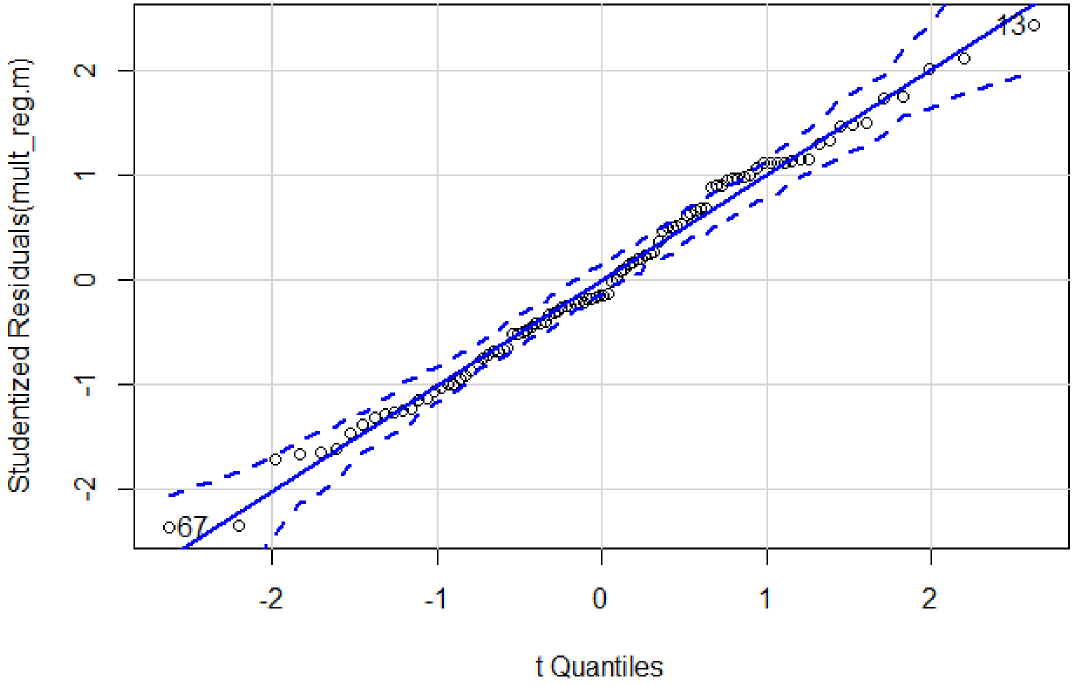
```
> car::residualPlots(mult_reg.m)
           Test stat Pr(>|Test stat|)
sw           -0.9792           0.3299
st            0.4186           0.6764
male         -1.2734           0.2060
Tukey test   -0.4648           0.6421
```

These t-tests are for curvature – a significant p-value indicates a change in the location of the residuals over X (non-significant = OK)



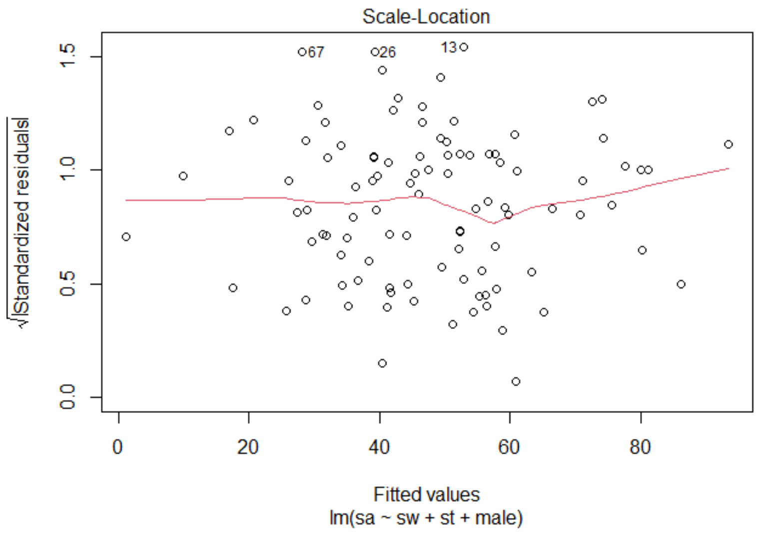
Normality

- For any fixed values of $X_1, X_2, \dots, X_k$, Y has a normal distribution.
- With large N, the central limit theorem makes inferences robust to deviations from this assumption.



Homoscedasticity

- For any fixed values of $X_1, X_2, \ldots, X_k$, the variance of Y is a constant.
- Technically, we want to see that the variance of Y is constant across the multivariate distribution of X values
- Since this is difficult to do, we frequently examine the variance of Y across the fitted values of Y.



Recap

- Multiple regression model assumptions are the same as for single regression
- To make examination of plots easier when there are many X variables, sometimes we examine the residuals as a function of $\hat{Y}$ instead of as a function of X.

Recap

- Assess the LINE assumptions from a multiple regression model

Example

100 participants were surveyed before a cruise that offered an all-you-can-eat food plan. Participants' weight (lbs.) was recorded before and after the cruise. Participants also filled out a survey asking about their liking of particular food items.

WGAIN: weight gain post- vs. pre-trip

FRIES: score of liking French fries (1-100 scale)

CHIPS: score of liking potato chips (1-100 scale)

Individually, these two variables appear to be strong predictors of weight gain during the trip.

```
> lm(wgain ~ fries, data = data6) %>% summary()

Call:
lm(formula = wgain ~ fries, data = data6)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.6333 -2.1853  0.0965  1.9174  4.9363 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.08230    0.85101  -0.097    0.923    
fries        0.09885    0.01622   6.095 2.16e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.486 on 98 degrees of freedom
Multiple R-squared:  0.2749,	Adjusted R-squared:  0.2675 
F-statistic: 37.15 on 1 and 98 DF,  p-value: 2.158e-08
```

```
> lm(wgain ~ chips, data = data6) %>% summary()

Call:
lm(formula = wgain ~ chips, data = data6)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.9340 -2.0774  0.1946  1.9435  4.6659 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.31171    0.78837   0.395    0.693    
chips        0.08967    0.01469   6.104 2.07e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.485 on 98 degrees of freedom
Multiple R-squared:  0.2755,	Adjusted R-squared:  0.2681 
F-statistic: 37.26 on 1 and 98 DF,  p-value: 2.075e-08
```

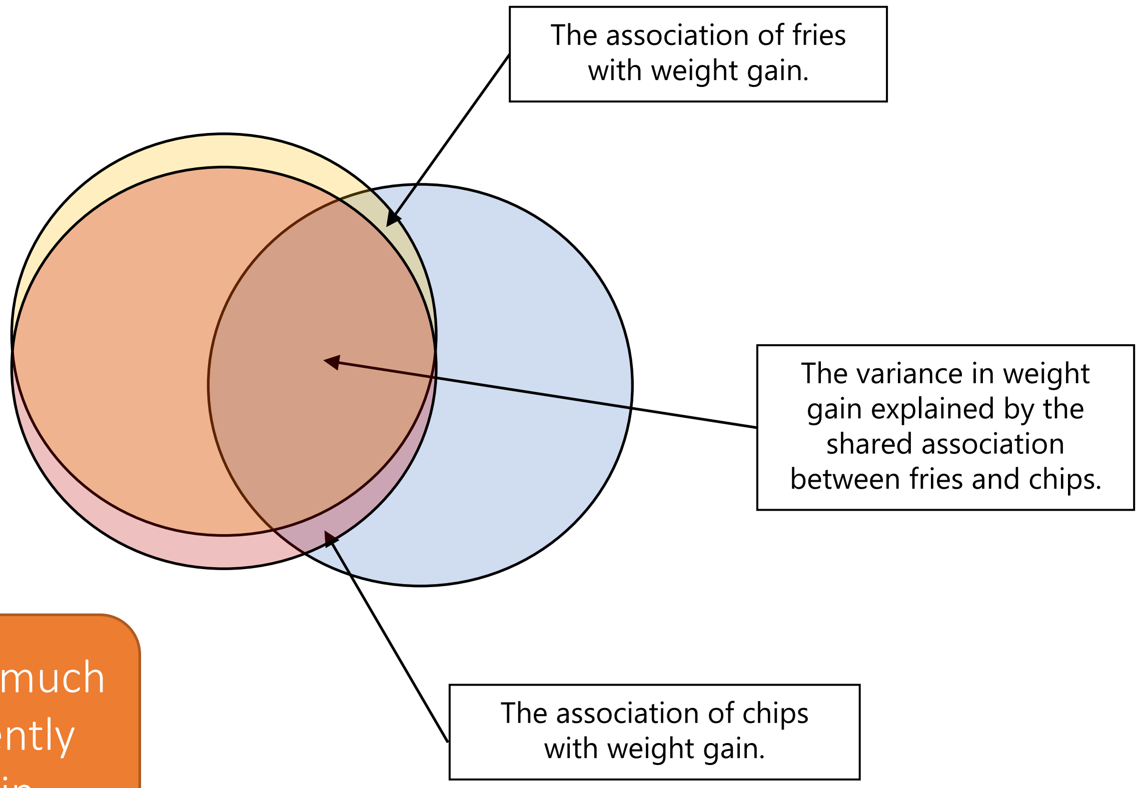
Yet, when we include them both in the same model, neither one appears to have an association with weight gain.

```
> lm(wgain ~ fries + chips, data = data6) %>% summary()

Call:
lm(formula = wgain ~ fries + chips, data = data6)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7589 -2.1982  0.1247  1.9397  4.8465 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.005777   0.856270   0.007    0.995
fries       0.049524   0.053859   0.920    0.360
chips       0.046878   0.048809   0.960    0.339

Residual standard error: 2.487 on 97 degrees of freedom
Multiple R-squared:  0.2817,	Adjusted R-squared:  0.2669 
F-statistic: 19.02 on 2 and 97 DF,  p-value: 1.071e-07
```

This is especially peculiar as the overall model test <u>is</u> significant.

This is a case of **collinearity**.

- The X variables are highly correlated (r = 0.95).
- The X variables share the same variance with the outcome.



Because our predictors share so much variance, they don't independently contribute much to weight gain.

This also causes problems with the ANOVA table.

```
> lm(wgain ~ chips + fries, data = data6) %>% anova()
Analysis of Variance Table

Response: wgain
          Df Sum Sq Mean Sq F value    Pr(>F)
chips      1 230.03 230.029 37.2025 2.174e-08 ***
fries      1   5.23   5.228  0.8455    0.3601
Residuals 97 599.77   6.183
```

```
> lm(wgain ~ fries + chips, data = data6) %>% anova()
Analysis of Variance Table

Response: wgain
          Df Sum Sq Mean Sq F value    Pr(>F)
fries      1 229.55 229.554 37.1256 2.237e-08 ***
chips      1   5.70   5.703  0.9224    0.3392
Residuals 97 599.77   6.183
```

The sum of squares explained by fries and chips differs depending on the order they are entered into the model.

The anova() function shows us the **Type I (sequential) Sums of Squares**.

```
> lm(wgain ~ chips + fries, data = data6) %>% anova()
Analysis of Variance Table

Response: wgain
          Df Sum Sq Mean Sq F value    Pr(>F)
chips      1 230.03 230.029 37.2025 2.174e-08 ***
fries      1   5.23   5.228  0.8455    0.3601
Residuals 97 599.77   6.183
```

Once we add chips into the model, fries does not contribute anything additional to explaining the variance in Y

```
> lm(wgain ~ fries + chips, data = data6) %>% anova()
Analysis of Variance Table

Response: wgain
          Df Sum Sq Mean Sq F value    Pr(>F)
fries      1 229.55 229.554 37.1256 2.237e-08 ***
chips      1   5.70   5.703  0.9224    0.3392
Residuals 97 599.77   6.183
```

Once we add fries into the model, chips does not contribute anything additional to explaining the variance in Y

We can use the CAR package to request **Type III (simultaneous) Sums of Squares**.

```
> lm(wgain ~ fries + chips, data = data6) %>% car::Anova(type = 3)
Anova Table (Type III tests)

Response: wgain
            Sum Sq Df F value Pr(>F)
(Intercept)   0.00  1  0.0000 0.9946
fries         5.23  1  0.8455 0.3601
chips         5.70  1  0.9224 0.3392
Residuals   599.77 97
```

Now the SS for fries and chips account for each other being in the model.

Note that in this approach the sum of all SS values does *not* equal the total SS in our Y variable.

When we have **collinear variables** in the model, we sometimes see the following consequences:

- Unexpected changes in the sign of the regression coefficients.
- Standardized coefficients greater than 1.
- Highly-inflated standard error estimates.

Example: Including mother's education, father's education, and parents' overall education as independent variables in a model.

Diagnosing Collinearity

- The **variance inflation factor (VIF)** is the amount the standard errors have been inflated by because of the inclusion of other correlated variables.
- The **tolerance** is 1/VIF.

The VIF will be 1 when the X variable is not correlated with any other independent variables in the model.

Rule of Thumb: A VIF value >10 indicates serious issues with collinearity.

The following collinearity diagnostic values show that fries and chips are highly collinear.

```
> lm(wgain ~ fries + chips, data = data6) %>% ols_vif_tol()
  Variables  Tolerance      VIF
1     fries 0.09073192 11.02148
2     chips 0.09073192 11.02148
```

The standard errors of the β estimates are inflated by a multiplicative factor of 11 when they are included in the same model.

Collinearity: What to Do?

- Sometimes centering the variables can help.
- When variables are highly collinear, they provide similar information about Y and little is lost by including only one.
- How to choose which variable to drop?
  - Whichever was least significant in a single linear regression model.
  - Whichever has the highest VIF value.
  - Whichever has less missing data.
  - Whichever is harder or more expensive to obtain.

Recap

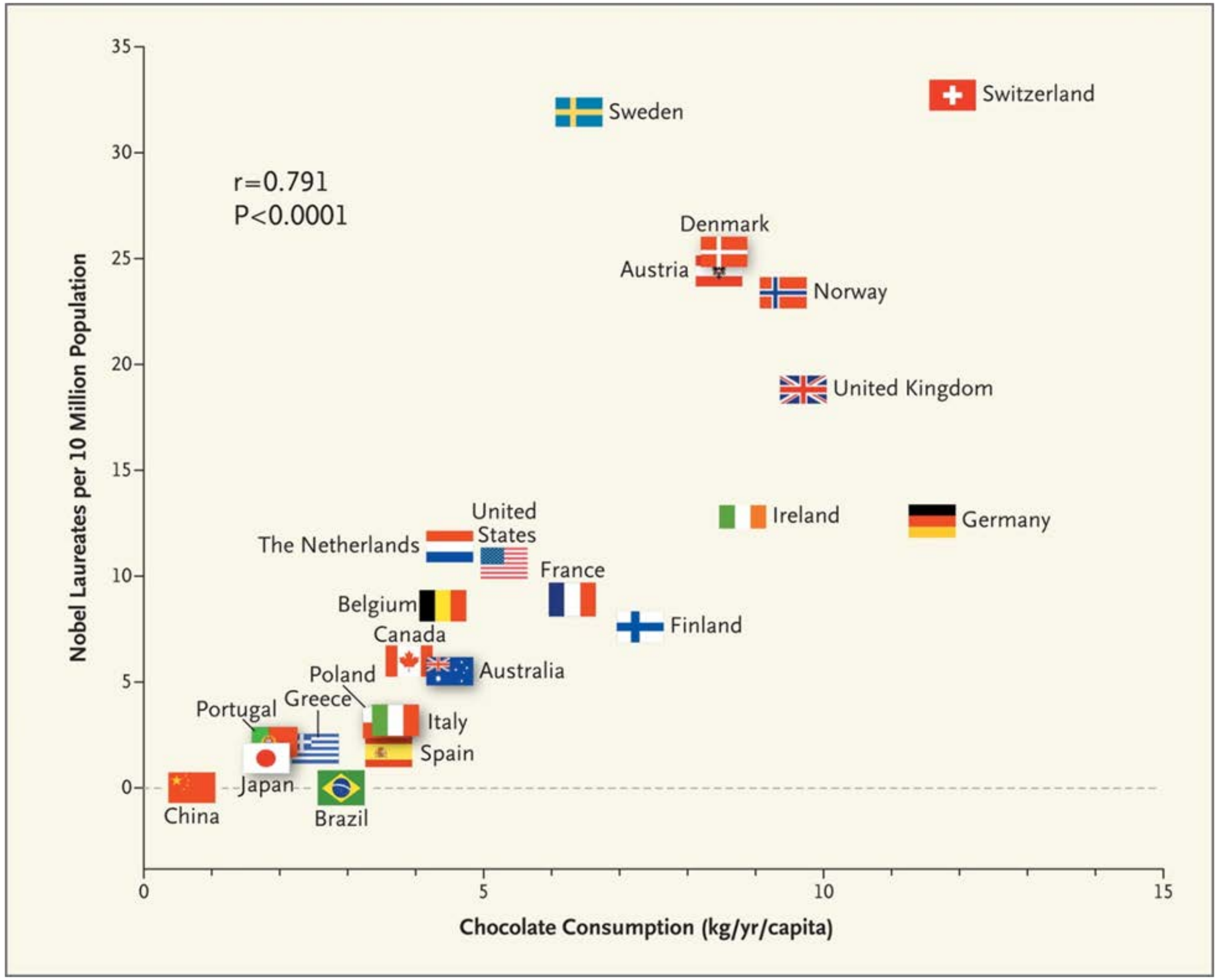
- Collinearity is a problem in multiple regression when two or more X variables are very highly related
- Collinearity can cause numerical problems in the regression output

Recap

- Explain how to diagnose collinearity in a multiple regression model
- Explain how to proceed with model building upon encountering collinearity
- Explain the difference between Type I and Type III Sums of Squares

Correlation between Countries' Annual Per Capita Chocolate Consumption & # of Nobel Laureates per 10 Million Population

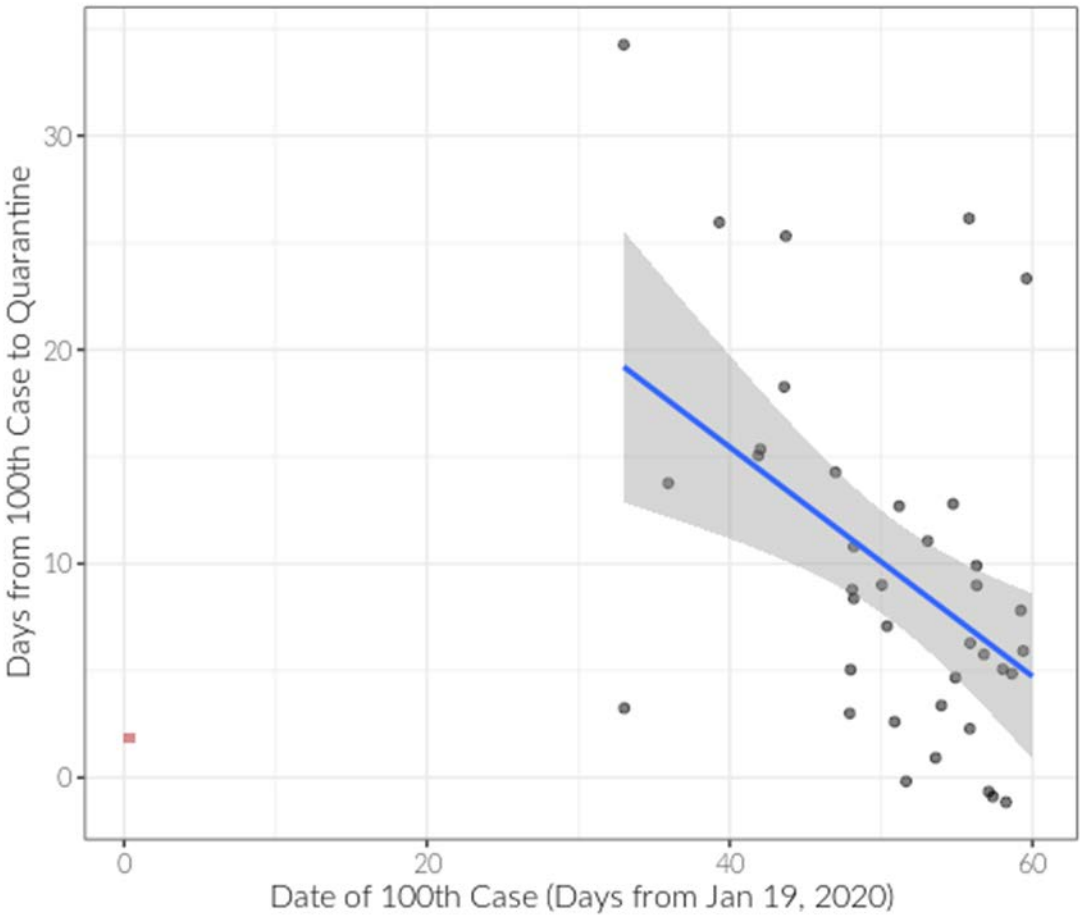
"There was a close, significant, linear correlation (r=0.791, p<.0001) between chocolate consumption per capita and the number of Nobel laureates per 10 million persons in a total of 23 countries… When recalculated with the exclusion of Sweden, the correlation coefficient increased to 0.862."



Messerli FH. N Engl J Med 2012;367:1562-1564.

Days from 100th COVID-19 case to quarantine plotted against date of 100th case (in days from Jan 19, 2020; China's 100th case).

The two variables are correlated ($r = -0.47$, $p=.003$) after removing China (red square).



Ultimately in regression we want a “good” model in that it:

1. **Satisfies the assumptions** of linear regression
   - Basic satisfaction of the LINE assumptions, and more to be covered
2. **Generalizes** to some population of interest
   - This is typically addressed through study design, but also by detection of outliers
3. **Accurately estimates** the relationship between Y and each X
   - Adequate sample size, no collinearity, account for confounding and effect modification (next week)
4. Has some **basis in reality**
   - Up to you and your colleagues to decide
   - More complicated models are harder to interpret; the best models are simple yet good (“parsimonious”)

The best way to meet these goals is to **know your data**

- Print your data in the results window or open it in the data viewer
- Exploratory data analysis for frequencies, means, maximum and minimum values
- Plotting the univariate distributions (histograms) and multivariate distributions (scatterplots) to spot potential data errors
- Correlation matrices to understand pairwise relationships among variables

Diagnostics Definitions

Outlier. A rare or unusual observation that appears at one of the extremes of the univariate or multivariate distributions.

Leverage (h_i). The extremeness of an observation with respect to the independent variables. The leverage of a data point depends on the distance of its X-value from the corresponding mean of all X values.

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{(n-1)S_X^{\;2}}$$

Influence. A data point is influential if, by itself, it has a substantial impact on the parameter estimates (intercepts, slopes) in a model. This typically happens when an observation is an outlier with high leverage (extreme in X and an unusual pattern of Y|X).

Advanced Residual Analysis

There are subtle differences among the different types of residuals we can examine.

The true residual. This is a bit of an abstract concept. The errors e_i are assumed to have mean 0 and variance σ^2. We can never really tell what the true residual is, as we can only measure them empirically based on the model we choose.

The estimated residual. What we are accustomed to calculating. This is given as: $\hat{e}_i = Y_i - \hat{Y}_i$. The mean of all estimated residuals is 0.

The standardized residual. $z_i = \frac{\hat{e}_i}{s}$. Essentially, each residual is converted to a z-score by dividing by the standard deviation of all residuals (mean 0 and variance 1). Because they're standardized, we can immediately determine how extreme the residual value is.

The studentized residual. $r_i = \frac{z_i}{\sqrt{1-h_i}}$. These values follow a T distribution with n-k-1 d.f. if regression assumptions are met.

The jackknife residual. $r_{(-i)} = \frac{\hat{e}_i}{S_{(-i)}\sqrt{1-h_i}}$, where $S_{(-i)}$ is the standard deviation of the residuals computed from a model where the i^{th} observation is deleted.

This type of residual is especially useful for identifying influential points.

These values follow a T-distribution with n-k-2 degrees of freedom if regression assumptions are met.

Jackknife residuals are large when the studentized residual is large.

Measures of Influence

There are 3 statistics that quantify the amount of influence an observation has on the estimated regression slope(s) or predicted value of Y.

1. Cook's Distance

A measure of how much all the fitted values change with the deletion of each observation.

$$d_i = \frac{e_i^2 h_i}{(k+1)S^2(1-h_i)^2} = \frac{r_i^2 h_i}{(k+1)(1-h_i)}$$

Observations with $d_i > 0.5$ may be worth investigating.
Observations with $d_i > 1$ are likely worth investigating.
Any observation with a Cook's distance that "sticks out" should be investigated.

2. DFBETAS

A measure of how much the regression coefficients change with the exclusion of the i^{th} observation.

$$\Delta\beta = \frac{\hat{\beta} - \hat{\beta}_{(-i)}}{S_{(-i)}\sqrt{\Sigma X_i^2}}$$

Observations with $\Delta\beta > 2/\sqrt{n}$ are influential.

3. DFFITS

A measure of how much the predicted value for the i^{th} observation changes when the i^{th} observation is deleted.

$$\Delta\hat{Y}_i = \frac{\hat{Y}_i - \hat{Y}_{i(-i)}}{S_{(-i)}\sqrt{\Sigma h_i}}$$

Observations with $\Delta\hat{Y}_i > 2/\sqrt{k/n}$ are influential.

Summary of Rules of Thumb

Leverage	Cook's D	DFBETAS	DFFIT
$h_i > 2(k+1)/n$	$d_i > 1$	$\lvert\Delta\beta\rvert > 2/\sqrt{n}$	$\lvert\Delta\hat{Y}_i\rvert > 2/\sqrt{k/n}$

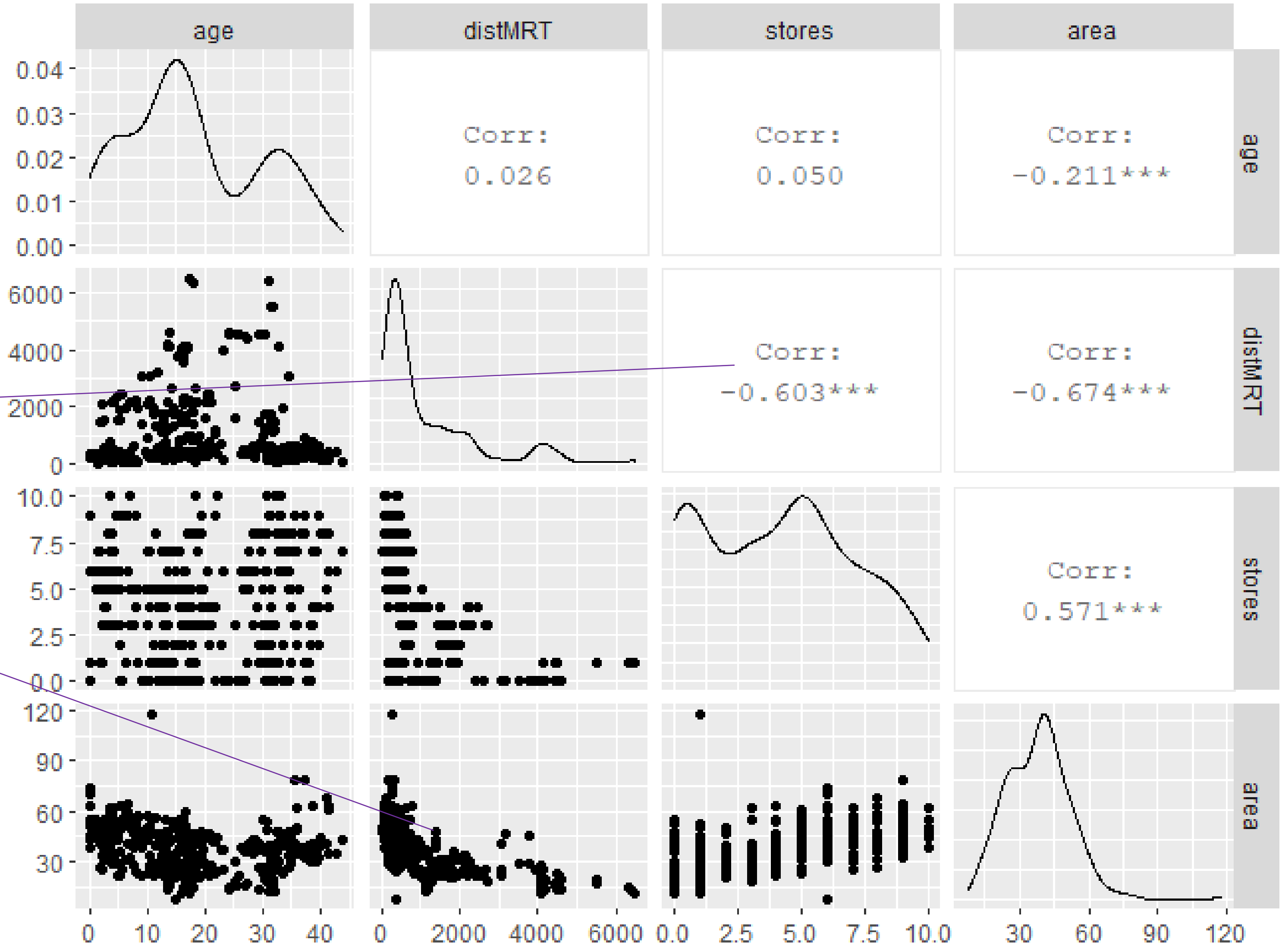
Example

What is the combined effect of house age, number of stores nearby, and distance to the nearest MRT station on price per unit area?

All of the independent variables are correlated with price per unit area.

⚠ Possible collinearity as DISTMRT and STORES are correlated.

⚠ Possible nonlinear relationship between DISTMRT & AREA.



```
> ols_vif_tol(re.m)
  Variables Tolerance      VIF
1       age 0.9927044 1.007349
2   distMRT 0.6338826 1.577579
3    stores 0.6327390 1.580431.
```

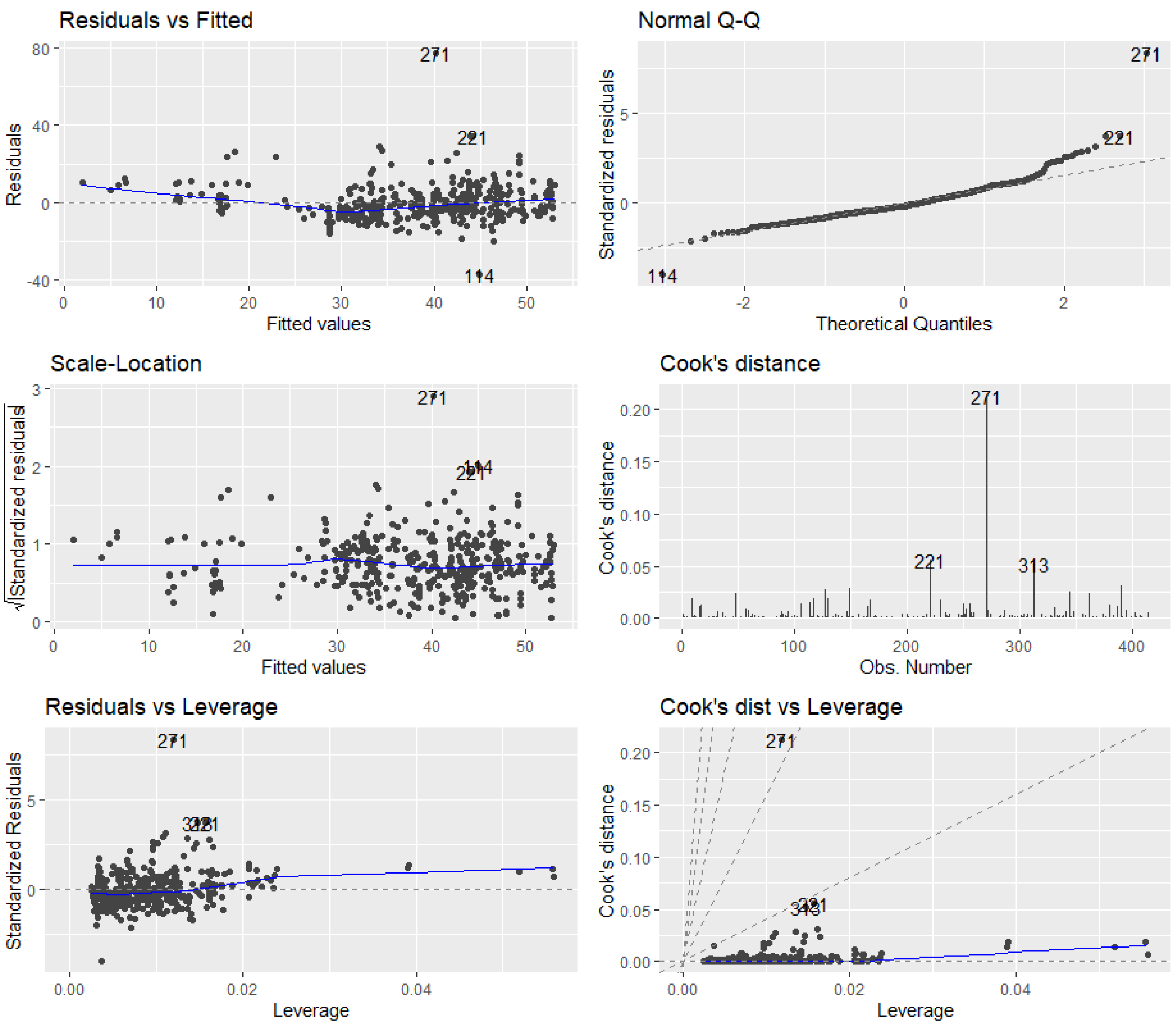
✓ No evidence of collinearity.

```
> summary(re.m)

Call:
lm(formula = area ~ age + distMRT + stores, data = re)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.304  -5.430  -1.738   4.325  77.315 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 42.977286   1.384542  31.041  < 2e-16 ***
age         -0.252856   0.040105  -6.305 7.47e-10 ***
distMRT     -0.005379   0.000453 -11.874  < 2e-16 ***
stores       1.297443   0.194290   6.678 7.91e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.251 on 410 degrees of freedom
Multiple R-squared:  0.5411,	Adjusted R-squared:  0.5377 
F-statistic: 161.1 on 3 and 410 DF,  p-value: < 2.2e-16
```

This is the full complement of plots available through plot() or autoplot()

Check for high Cook's Distance, especially toward the ends of the fitted distribution.

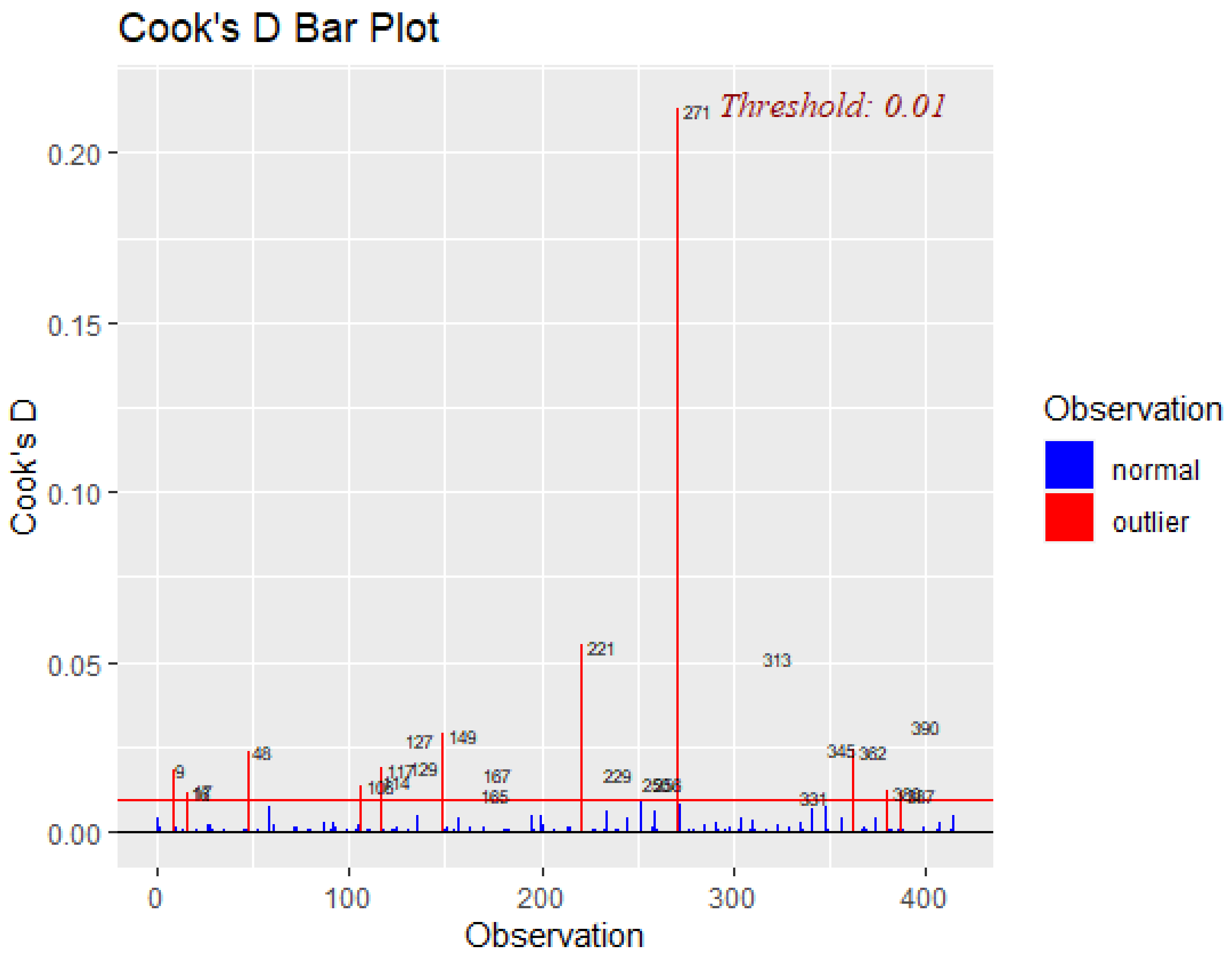
Check for high residuals ($|z_i|>2$) that also have high leverage.



Let's investigate in more detail

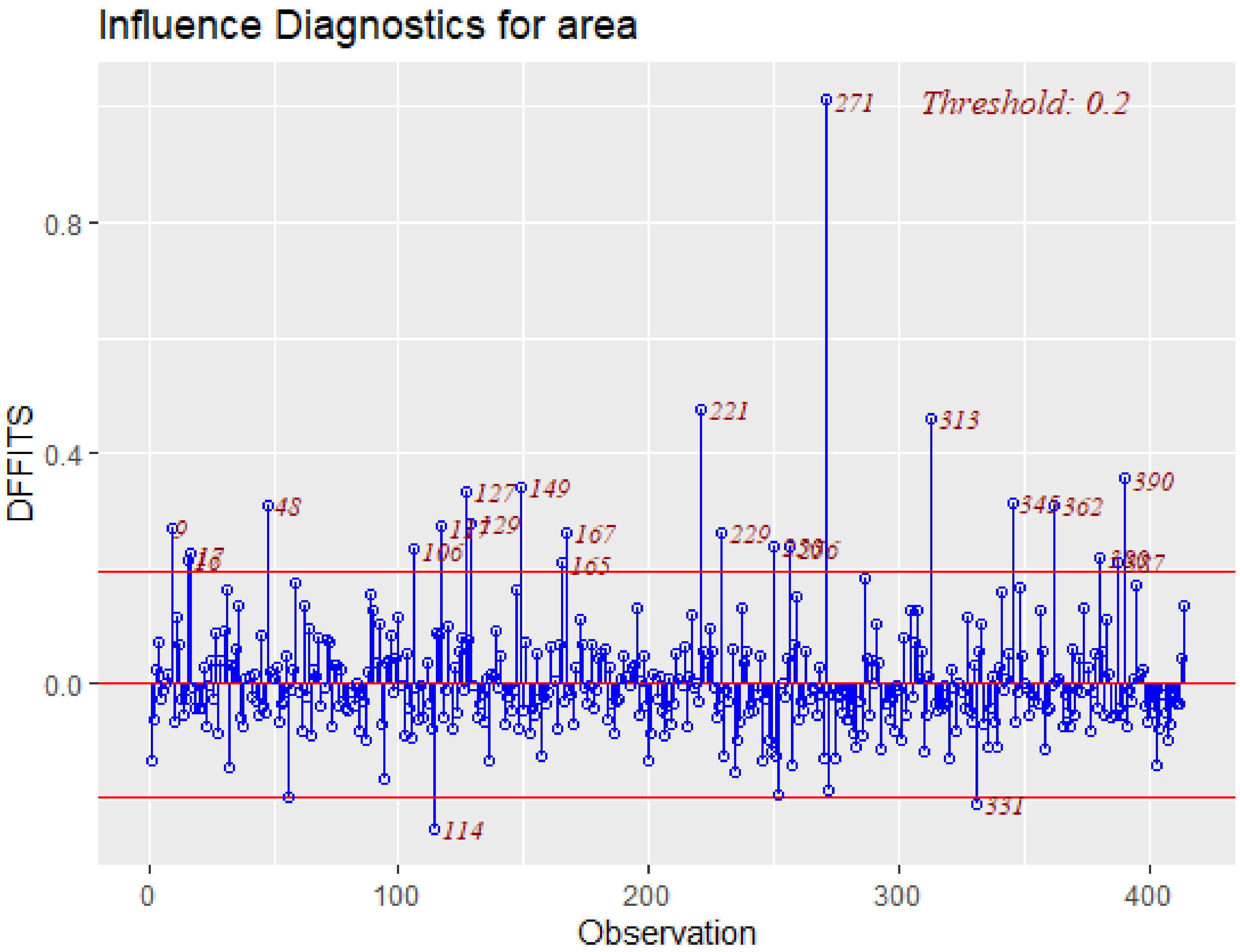
```
ols_plot_cooksd_bar(re.m)
```

⚠ Observation 271 definitely seems influential – it stands out like a sore thumb.



```
ols_plot_dffits(re.m)
```

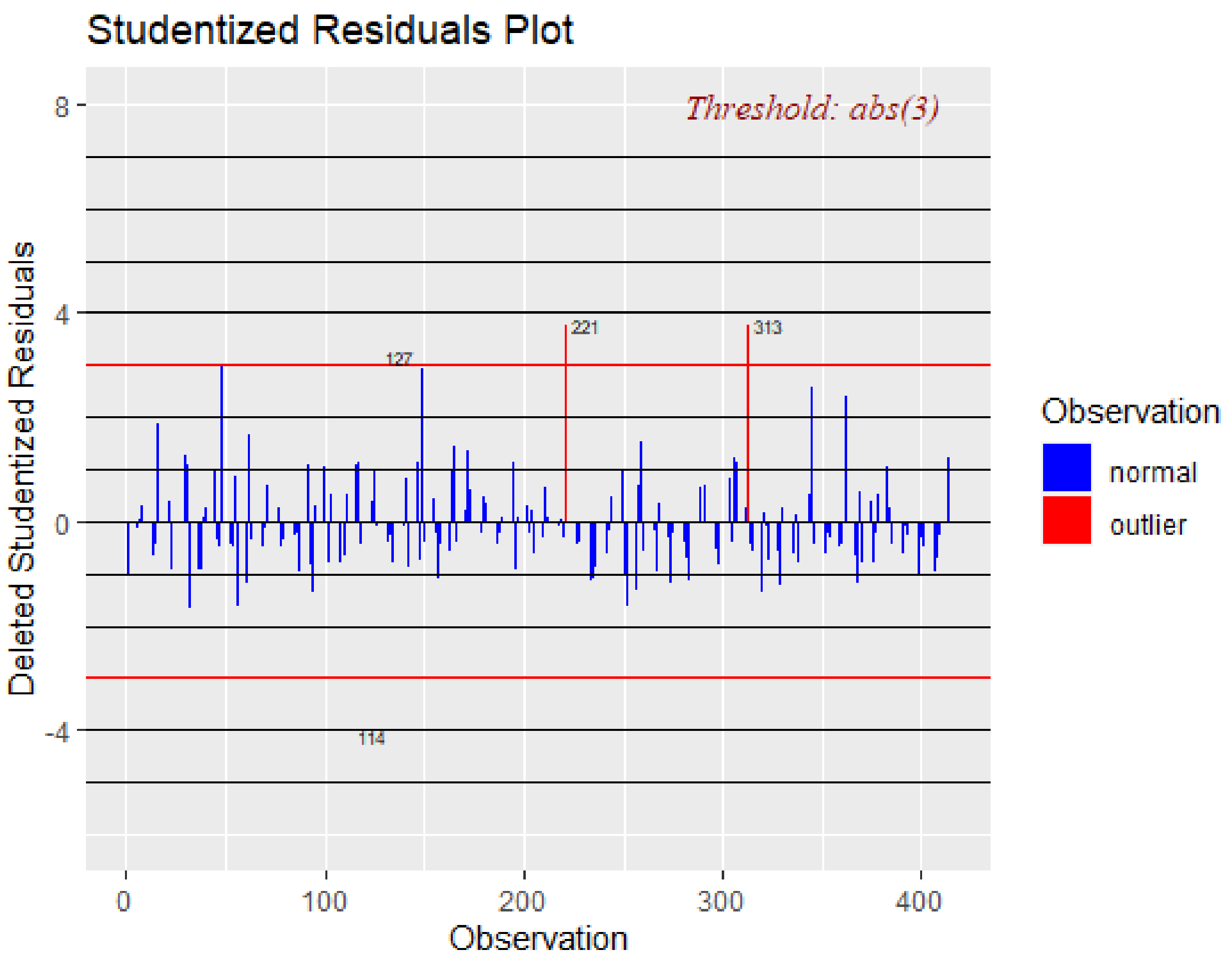
⚠ Again, the removal of observation 271 when model fitting will drastically change its predicted value.



Jackknife (externally studentized) residuals

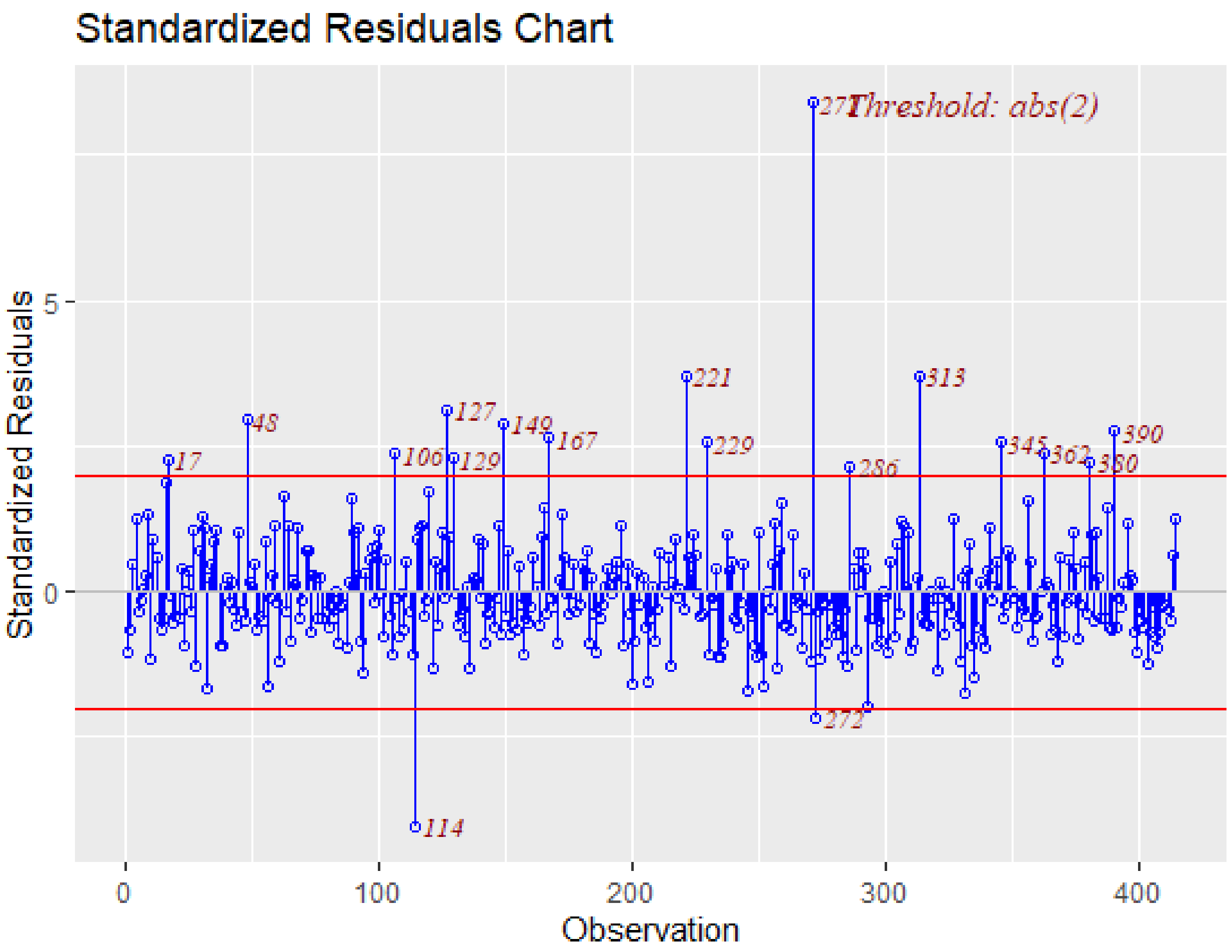
```
ols_plot_resid_stud(re.m)
```

Observations 127, 221, and 313 have high jackknife residuals, but they likely aren’t as influential as observation 271.



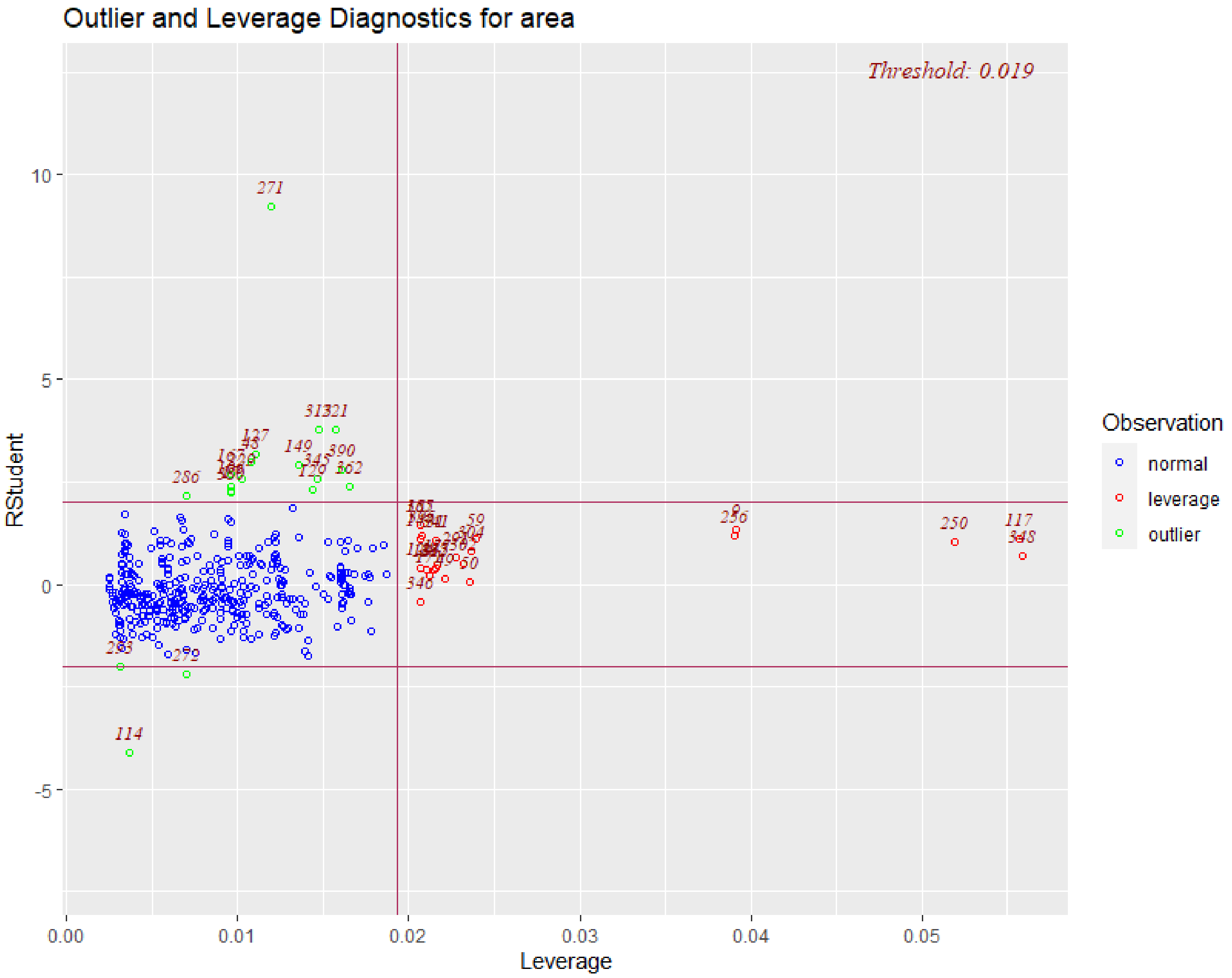
Studentized (internally studentized or “standardized”) residuals

```
ols_plot_resid_stand(re.m)
```



Leverage vs. Studentized Residual

```
ols_plot_resid_lev(re.m)
```



Here we see that, while a large outlier, observation 271 isn't classified as having "high" leverage. I would still examine that observation further, though.

To get influence statistics directly, we can use the influence.measures() function.

```
> influence.measures(re.m)$infmat %>% as_tibble()
# A tibble: 414 x 8
     dfb.1_   dfb.age  dfb.dMRT  dfb.strs    dffit cov.r     cook.d      hat
      <dbl>     <dbl>     <dbl>     <dbl>    <dbl> <dbl>      <dbl>    <dbl>
 1  0.0744  -0.0577   -0.0227   -0.0933   -0.131   1.02  0.00432    0.0159
 2  0.0249  -0.00128  -0.0152   -0.0509   -0.0631  1.02  0.000998   0.00971
 3  0.0125  -0.00864  -0.00582   0.00234   0.0259  1.01  0.000168   0.00320
 4  0.0349  -0.0240   -0.0162    0.00652   0.0720  0.997 0.00129    0.00320
 5 -0.0159   0.0177    0.00606  -0.000972 -0.0254  1.01  0.000162   0.00610
 6 -0.00250  0.00663  -0.00592  -0.00181  -0.0113  1.02  0.0000317  0.00660
 7 -0.00161  0.00240   0.000322  0.00144   0.00343 1.02  0.00000295 0.00978
 8  0.00206  0.00303  -0.00442   0.00441   0.0174  1.01  0.0000759  0.00376
 9 -0.146    0.0698    0.238     0.0840    0.271   1.03  0.0184     0.0391
10 -0.0117  -0.000348 -0.0234    0.00270  -0.0650  1.00  0.00106    0.00316
# ... with 404 more rows
```

Let's figure out why Observation 271 was such an anomaly.

```
> re %>%
+   bind_cols(
+     tibble(
+       pred = predict(re.m)
+       )
+     ) %>%
+   .[271,]
# A tibble: 1 x 11
     No  date   age distMRT stores   lat  long  area expense expense_ln  pred
  <dbl> <dbl> <dbl>   <dbl>  <dbl> <dbl> <dbl> <dbl>   <dbl>      <dbl> <dbl>
1  271 2013.  10.8    253.      1  25.0  122.  118.    3.16       1.15  40.2
```

This observation was of average age, didn't have many stores near it, and wasn't that close to the MRT. It was predicted to sell for $40.2 per unit area but it actually sold for $118 (almost 3x what was expected)!

Summary of Residuals

Standardized

- Z-score of residuals
- Identifies outliers

Studentized (internally)

- Standardizes to a t-distribution
- Accounts for leverage
- Identifies outliers

Jackknife (externally studentized)

- Studentized residual from fitting a model that deletes the observation in question
- Identifies influential points
- Better at identifying outliers on Y

What to do when you encounter problematic data points?

- Check for obvious data errors
  - Check for the accuracy of the data point
  - Delete the point if it is not representative of your intended population
- Consider the formulation of your model
  - Did you leave out important predictors?
  - Is there nonlinearity, or are there interactions you have to consider?
- Always justify why you alter data
  - Have a good objective reason for deleting data points
  - You can always perform “sensitivity” analysis; report the results with the data point included and excluded

Recap

- While not part of the LINE assumptions, models should be checked to see if any data point has undue influence on the regression model
- Problematic observations are those that 1) have high leverage (potential to be influential) and 2) have high residuals
- There are no firm rules about what constitutes an influential value; you'll need to make a decision based on the evidence you find

Recap

- Explain the different ways of calculating a model residual
- Explain the different regression diagnostic metrics and what each measures
- Explain how to spot an influential observation
- Describe the steps that should be taken upon finding influential observations

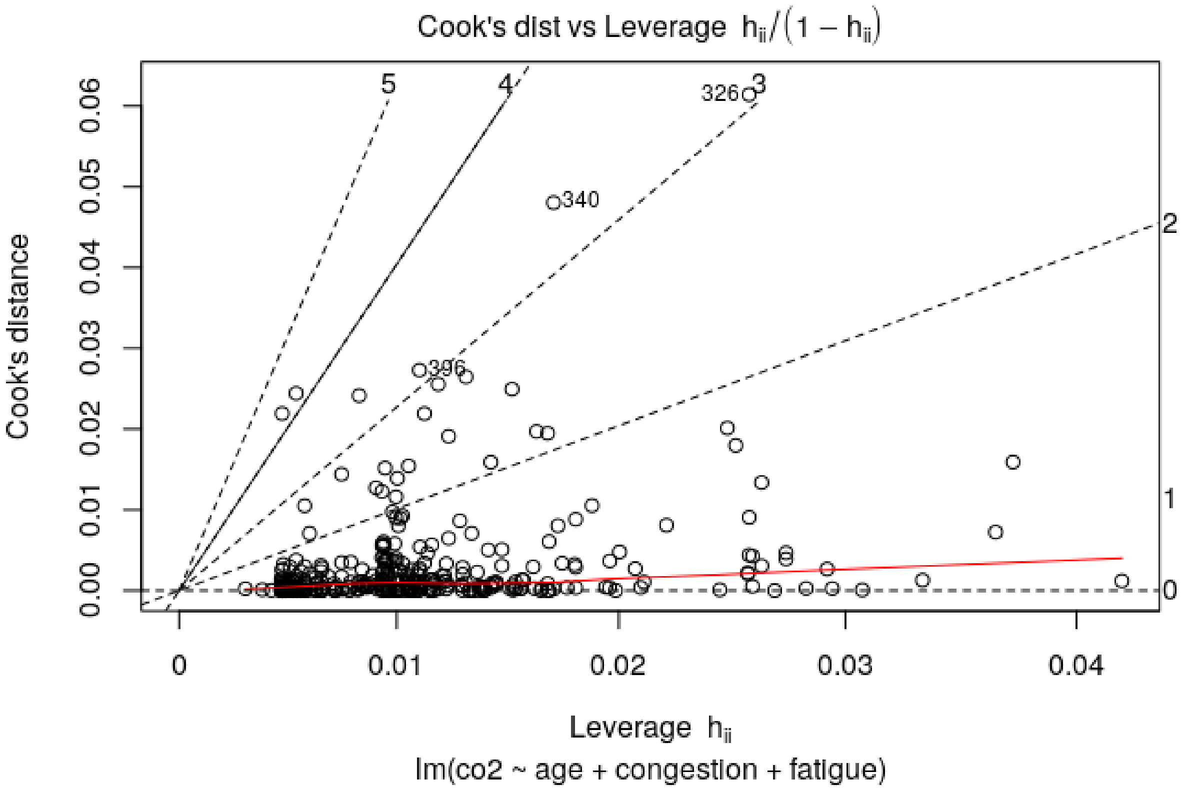
Test Yourself

This diagnostic plot was obtained from the model we previously assessed. What can we say about Observation 326?

a) This observation is the farthest from the mean of all X variables.

b) This observation has the highest residual.

c) This observation changes the regression fitted values the most.



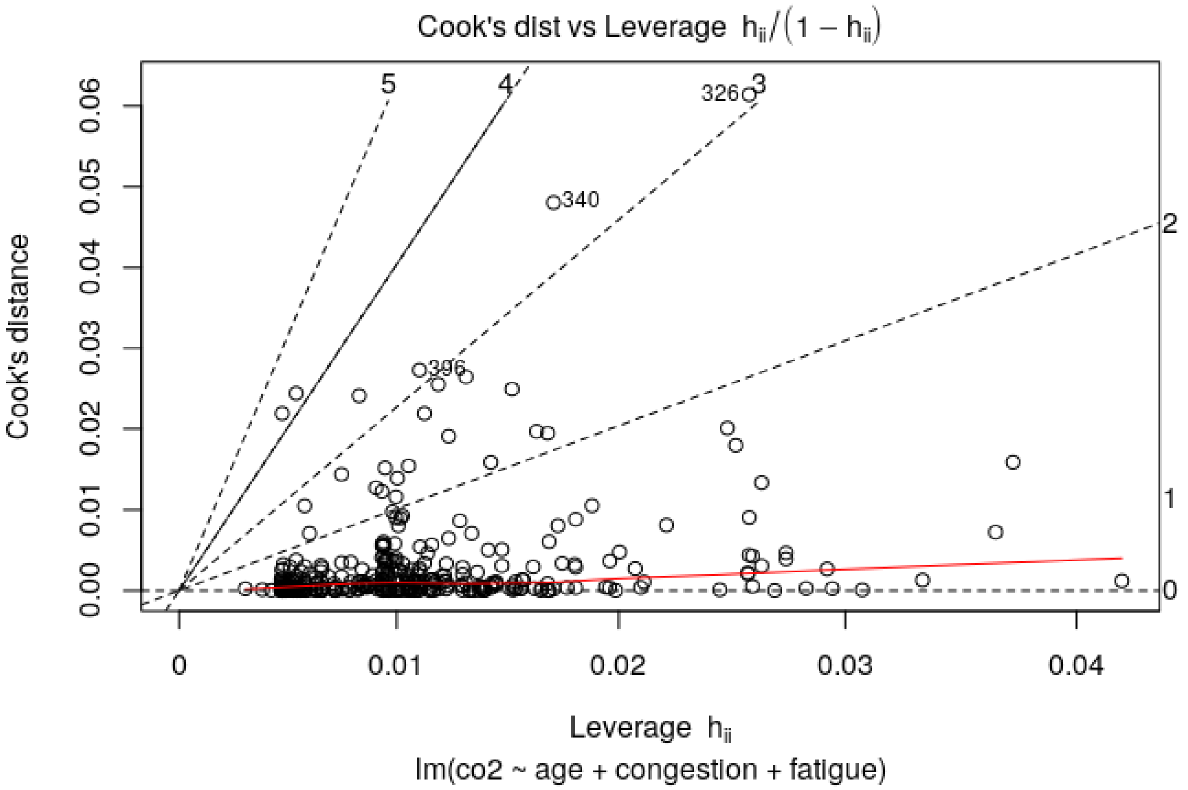
Test Yourself

This diagnostic plot was obtained from the model we previously assessed. What can we say about Observation 326?

a) This observation is the farthest from the mean of all X variables.

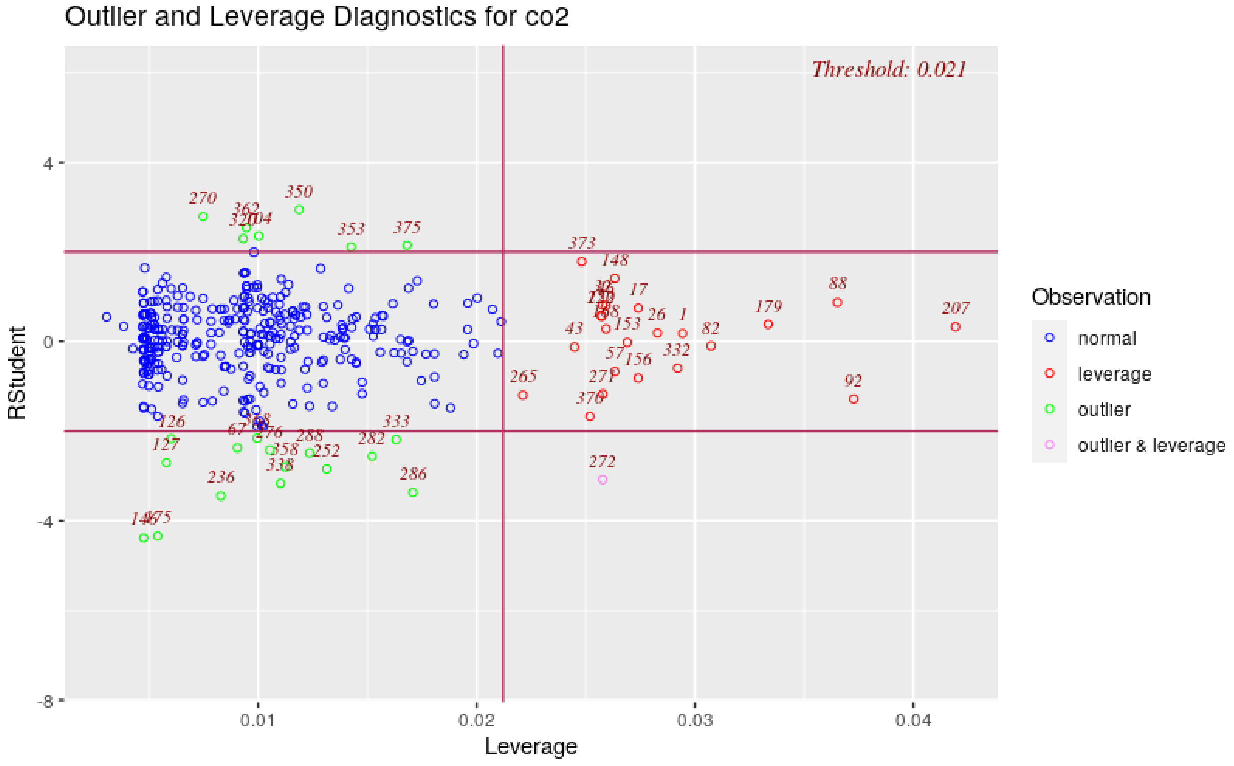
b) This observation has the highest residual.

c) This observation changes the regression fitted values the most.



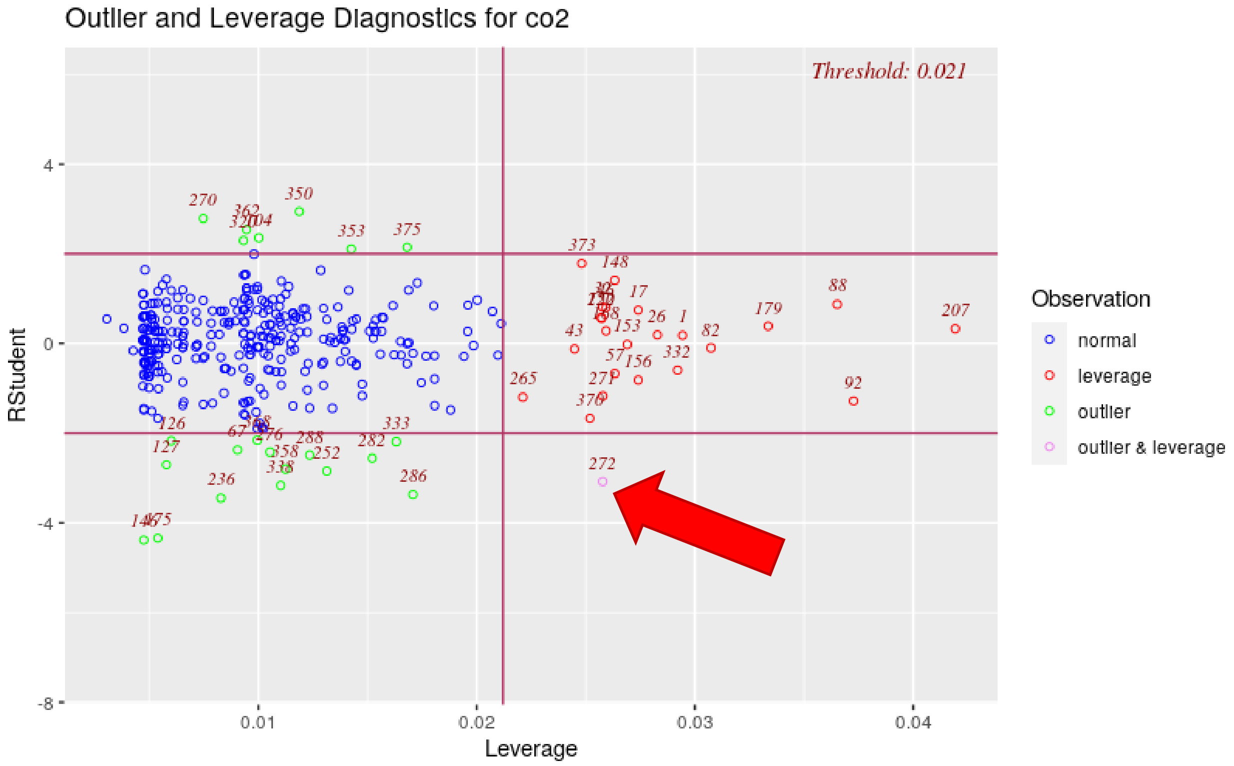
Test Yourself

This plot was obtained from the OLSRR package. Which point is most likely to be influential, according to this plot?



Test Yourself

This plot was obtained from the OLSRR package. Which point is most likely to be influential, according to this plot?



Additional Reading

Type I and Type III Sum of Squares

https://www.youtube.com/watch?v=mNzljQBKu5I

Partial and Semipartial Correlations

https://www.youtube.com/watch?v=yb0a4wPERZc

Packages and Functions

- `plot(lm_object)`
- `car::residualPlots()`
- `car::Anova()`
- `GGally::ggpairs()`
- `olsrr::ols_vif_tol()`
- `olsrr::ols_plot_cooksd_bar()`
- `olsrr::ols_plot_diffits()`
- `olsrr::ols_plot_resid_stud()`
- `olsrr::ols_plot_resid_stand()`
- `olsrr::ols_plot_resid_lev()`
- `influence.measures()`