Applied Linear Modeling

September 17, 2019

To-do

- Exercises:
  - *pick a to-do task*
  - *when finished, put your names on the to-do task and drop in Done jar*
- Create a week-2 folder on your laptop and put the following files in it (from GitHub):
  - *week-4-workshop.Rmd*
  - *dist_ssp_amfar_ch9.csv*
- No new R packages for today

All the things for today

- Discussion of exercises
- Workshop
  - *Simple linear regression*

Infectious disease outbreaks & needle exchanges

Health | Local News

North Seattle HIV cluster among drug users and homeless people worries health officials

April 19, 2019 at 5:07 pm | *Updated April 19, 2019 at 10:07 pm*

By Ryan Blethen
Seattle Times staff reporter

The opioid epidemic and a rising homeless population have helped fuel an HIV outbreak in North Seattle.

Outbreak of HIV found among Boston drug users

By Felice J. Freyer Globe Staff, January 29, 2019, 11:16 a.m.

- HIV outbreaks and rates of Hepatitic C and other infectious diseases have been increasing with the rise of opioid use since 2010
- One strategy for protecting people from infectious disease is providing clean needles to injection drug users via needle exchange programs
- The evidence is limited but promising on the effectiveness of needle programs
  - *Evidence shows needle programs reduce needle litter and often provide referrals for treatment and disease testing*
  - *Needle programs are illegal in some states with the rationale that providing needles promotes drug use*

Data source

- The Foundation for AIDS Research (amFAR) includes distance to syringe exchange programs as one variable in the **Opioid & Health Indicators Database** (https://opioid.amfar.org/indicator/)
- The `dist_ssp_amfar_ch9.csv` data set is county-level data that includes the distance to a syringe exchange program from counties in the US

```
# distance to syringe program data
dist.ssp <- read.csv(file = "dist_ssp_amfar_ch9.csv")

# summary
summary(object = dist.ssp)
```

```
##               county    STATEABBREVIATION    dist_SSP
##  jackson county   :  5   TX     : 50       Min.   :  0.00
##  jefferson county :  5   GA     : 30       1st Qu.: 35.12
##  lincoln county   :  5   KS     : 21       Median : 75.94
##  washington county:  5   NC     : 21       Mean   :107.74
##  benton county    :  4   TN     : 21       3rd Qu.:163.83
##  decatur county   :  4   KY     : 19       Max.   :510.00
##  (Other)          :472   (Other):338
##  HIVprevalence     opioid_RxRate       pctunins          metro
##  Min.   :  -1.00   Min.   :  0.20   Min.   : 3.00   metro    :226
##  1st Qu.:  52.98   1st Qu.: 45.12   1st Qu.: 8.60   non-metro:274
##  Median : 101.15   Median : 62.40   Median :11.70
##  Mean   : 165.75   Mean   : 68.33   Mean   :12.18
##  3rd Qu.: 210.35   3rd Qu.: 89.95   3rd Qu.:15.00
##  Max.   :2150.70   Max.   :345.10   Max.   :35.90
##
```


Codebook

Based on the amFAR website, the variables have the following meanings:

- county: the county name
- STATEABBREVIATION: the two-letter abbreviation for the state the county is in
- dist_SSP: distance in miles to the nearest syringe services program
- HIVprevalence: people age 13 and older living with diagnosed HIV per 100,000
- opioid_RxRate: number of opioid prescriptions per 100 people
- pctunins: percentage of the civilian noninstitutionalized population with no health insurance coverage
- metro: county is non-metro, which includes open countryside, rural towns, or smaller cities with up to 49,999 people, or metro

A research question

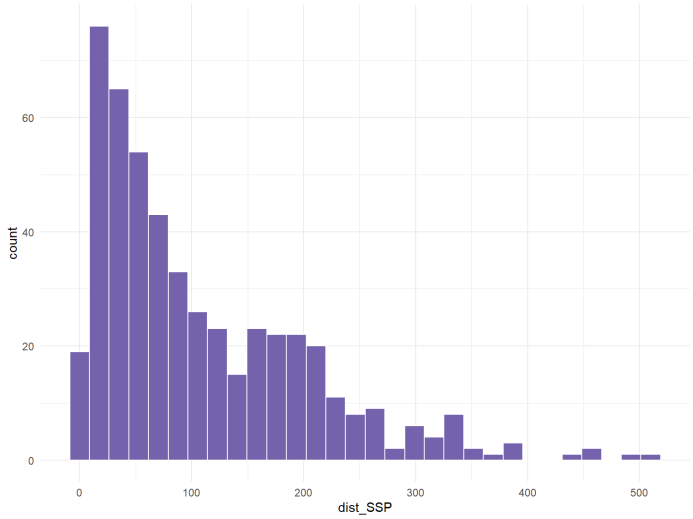
How can uninsurance, metro or non-metro status, HIV prevalence, and number of opioid prescriptions predict or explain distance to the nearest syringe program at the county level?

But first, descriptives

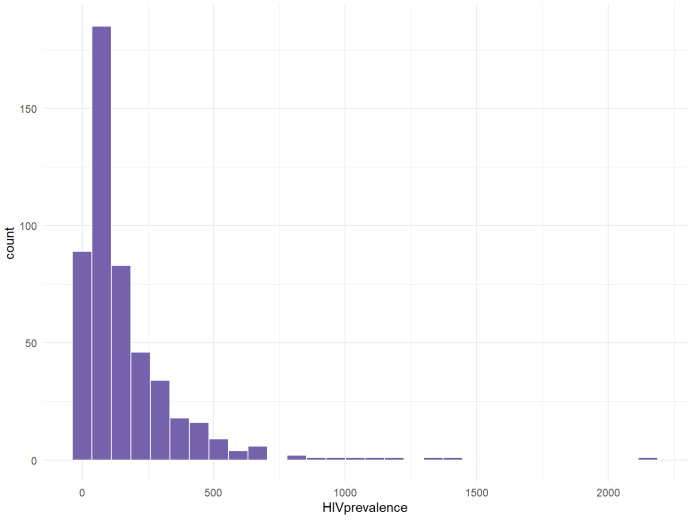
```
# open the tidyverse
library(package = "tidyverse")

# checking distributions for continuous
dist.ssp %>%
  ggplot(aes(x = dist_SSP)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  theme_minimal()
dist.ssp %>%
  ggplot(aes(x = HIVprevalence)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  theme_minimal()
dist.ssp %>%
  ggplot(aes(x = opioid_RxRate)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  theme_minimal()
dist.ssp %>%
  ggplot(aes(x = pctunins)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  theme_minimal()
```

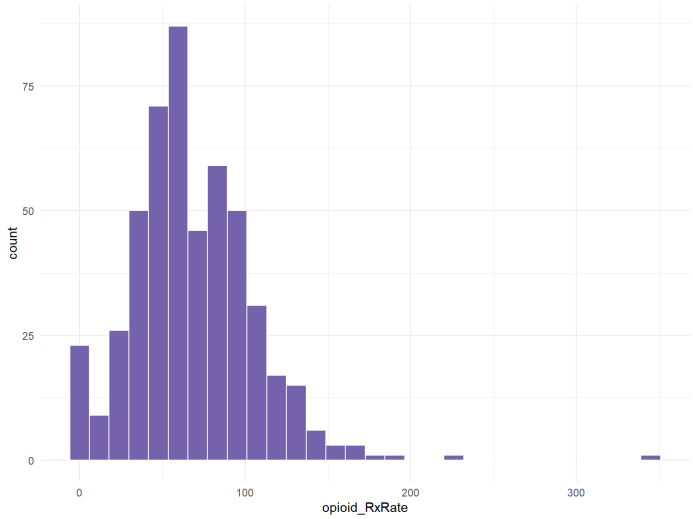

Distance to needle exchange



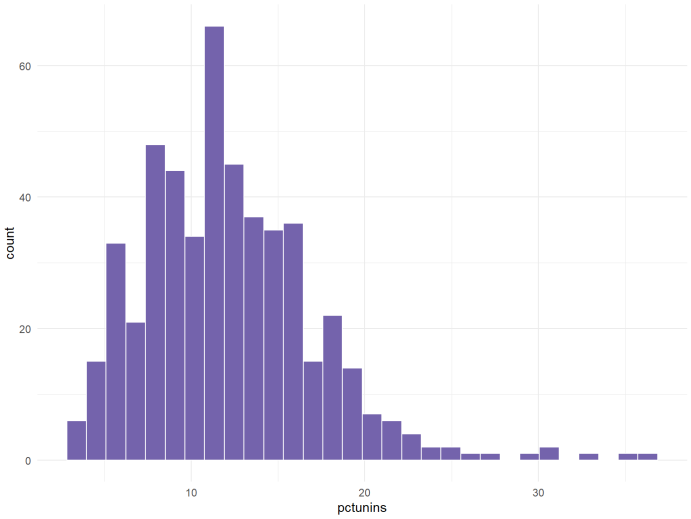
HIV prevalence



Opioid prescriptions per 100 people



Percent noninstitutionalized who are uninsured



Make the table

```
# make a table of descriptives
library(package = "tableone")

# dist_SSP and HIVprevalence are skewed
ssp.table <- CreateTableOne(data = dist.ssp,
                        vars = c('dist_SSP', 'HIVprevalence',
                                 'opioid_RxRate', 'pctunins',
                                 'metro'))
print(ssp.table, nonnormal = c("dist_SSP", "HIVprevalence"),
      showAllLevels = TRUE)
```

```
##
##                                level     Overall
##    n                                       500
##    dist_SSP (median [IQR])              75.94 [35.12, 163.83]
##    HIVprevalence (median [IQR])        101.15 [52.98, 210.35]
##    opioid_RxRate (mean (SD))            68.33 (36.81)
##    pctunins (mean (SD))                 12.18 (4.97)
##    metro (%)                 metro        226 (45.2)
##                              non-metro    274 (54.8)
```


The statistical model for regression

$$y = mx + b$$

Where:

- m is the slope of the line
- b is the y-intercept of the line, or the value of y when x = 0
- x and y are the coordinates of each point along the line

Other ways of writing the statistical model for regression

In statistics, the intercept is often represented by c or b_0 and the slope is often represented by b_1. Rewriting the equation for a line with these options would look like:

$$y = b_0 + b_1 x$$

$$y = c + b_1 x$$

Substituting values into the equation for a line

The equation for a regression line is useful in two ways:

- to *explain* values of the outcome (y)
- to *predict* values of the outcome (y)

For example, the equation for a line with a slope of 2 and a y-intercept of 3 would be written:

$$y = 3 + 2x$$

Using the equation for a line to predict

- Consider a scenario where this model predicts the number of gallons of clean drinking water per person per week needed to survive
- In this example, x would be the number of weeks and y would represent the gallons of water needed to survive

$$gallons = 3 + 2 \cdot weeks$$

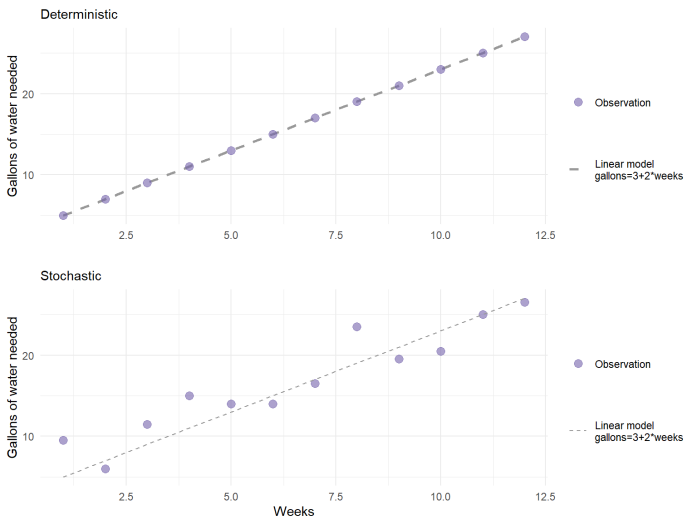
- A person hears from their landlord that there is a major repair needed with the pipes and the water supply in the building will not be safe for drinking for up to 4 weeks.
- Based on the equation, how many gallons of drinking water would be needed to survive?

$$gallons = 3 + 2 \cdot 4$$

$$gallons = 11$$

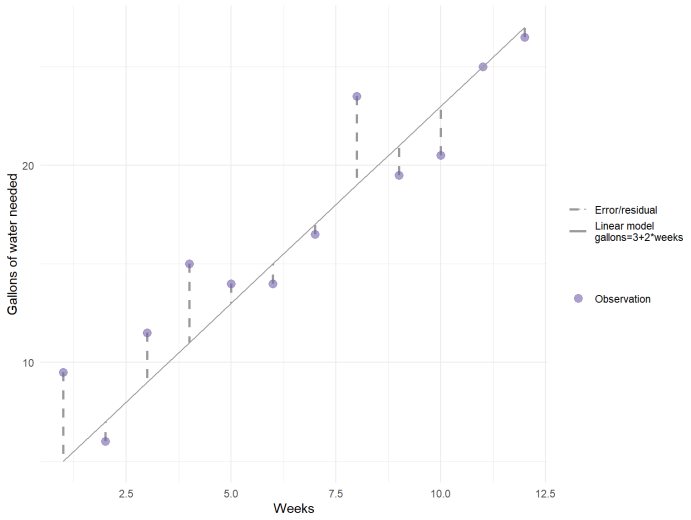
Deterministic vs. stochastic

- 11 gallons is a precise prediction and probably not correct for every human
- People (and other things) tend to vary, so regression captures *patterns* rather than precise predictions
  - *It is stochastic rather than deterministic*



Regression model with error term

$$outcome = b_0 + b_1 * predictor + error$$



Computing the slope and the intercept

Write out the model we are testing:

$$distance = b_0 + b_1 * uninsured + error$$

Compute the slope:

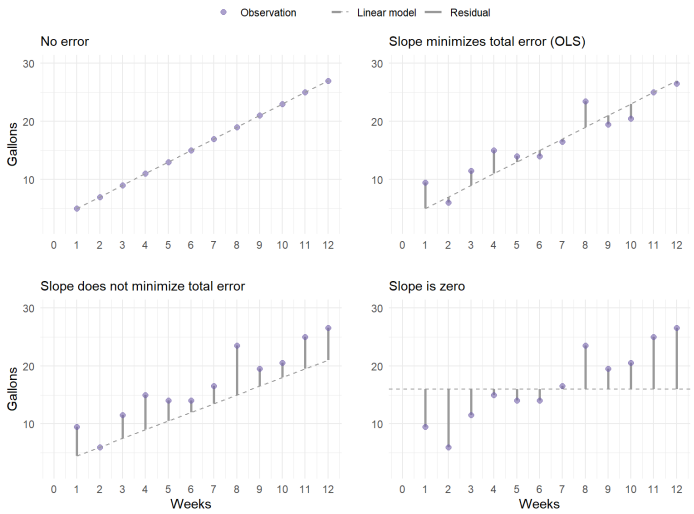
$$b = \frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{\sum_{i=1}^{n}(x_i - m_x)^2}$$

Where:

- i is an individual observation, in this case a county
- n is the sample size, in this case 500
- x_i is the value of `pctunins` for i
- m_x is the mean value of `pctunins` for the sample
- y_i is the value of `dist_SSP` for i
- m_y is the mean value of `dist_SSP` for the sample
- $\sum$ is the symbol for sum
- b is the slope

More about the slope

- The formula for the slope *minimizes the total error/residual*
- This method of regression is called ordinary least squares (OLS)



Using R to compute the slope

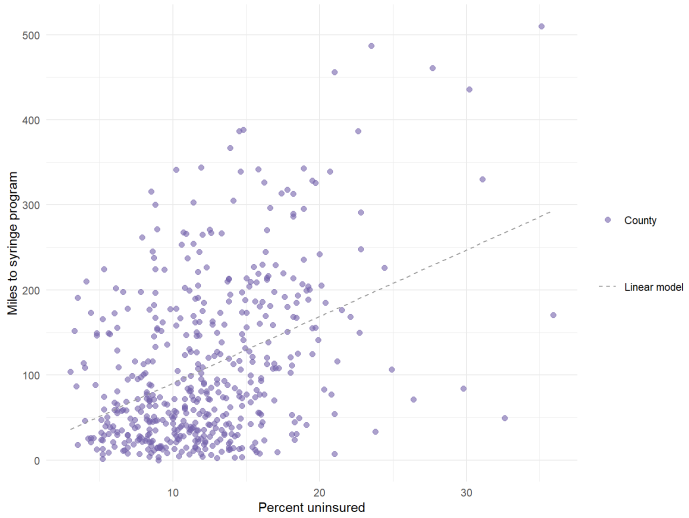
```
# linear regression of distance to syringe program by percent uninsured
dist.by.unins <- lm(formula = dist_SSP ~ pctunins,
                    data = dist.ssp, na.action = na.exclude)
summary(dist.by.unins)
```

```
##
## Call:
## lm(formula = dist_SSP ~ pctunins, data = dist.ssp, na.action = na.exclude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -217.71  -60.86  -21.61   47.73  290.77
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.4798    10.1757   1.226    0.221
## pctunins      7.8190     0.7734  10.110   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.91 on 498 degrees of freedom
## Multiple R-squared:  0.1703, Adjusted R-squared:  0.1686
## F-statistic: 102.2 on 1 and 498 DF,  p-value: < 2.2e-16
```

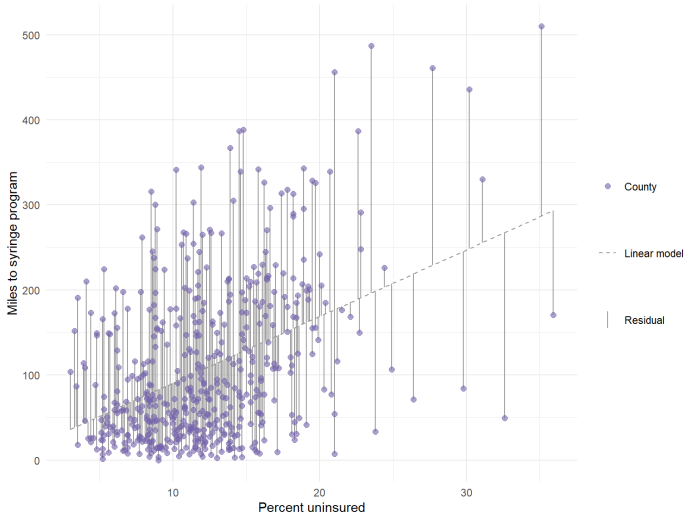

Navigating the linear regression output

$$distance = 12.4798 + 7.8190 \cdot pctunins$$

Visualizing the regression line with this slope and intercept



Visualizing the regression line with residuals



Slope value interpretation

- distance to syringe program = 12.48 + 7.82 * uninsured
- distance to syringe program = 12.48 + 7.82 * 10
- distance to syringe program = 90.67

Another county with 11% uninsured would have a predicted distance to syringe program of:

- distance to syringe program = 12.48 + 7.82 * uninsured
- distance to syringe program = 12.48 + 7.82 * 11
- distance to syringe program = 98.49

Because the slope is 7.82, the distance to the nearest syringe program increases by 7.82 miles for each 1% increase in people without insurance.

Slope significance testing: NHST Step 1 Write the null and alternate hypothesis

H0: The slope of the line is equal to zero

HA: The slope of the line is not equal to zero

NHST Step 2: Compute the test statistic

- The test statistic for the Wald test in OLS regression is the t-statistic
- The formula is the same as the formula for the one-sample t-test, but with the slope of the regression model in the numerator instead of the mean

$$t = \frac{b_1 - 0}{se_{b_1}}$$

* The formula can be used by substituting in the slope and standard error from the model output.

$$t = \frac{7.8190 - 0}{.7734} = 10.1099$$

NHST Step 3: Calculate the probability that your test statistic is at least as big as it is if there is no relationship (i.e., the null is true)

- The p-value of `<2e-16` for the slope is in the output
- There is a tiny chance that the t-statistic for the slope would be as big as it is (or bigger) if the null hypothesis were true

NHST Steps 4 & 5: Reject or retain the null hypothesis based on the p-value

- The null hypothesis is rejected in favor of the alternate hypothesis that the slope is not equal to zero

Interpretation:

The percentage of uninsured residents in a county is a statistically significant predictor of the distance to the nearest syringe program (b = 7.82; p < .05) in our sample. For every 1% increase in uninsured residents in a county, the predicted distance to the nearest syringe program increases by 7.82 miles.

Computing confidence intervals for the slope and intercept

- Compute confidence intervals using the standard error of the slope using `confint()` function with the `dist.by.unins` linear regression model object

```
# confidence interval for regression parameters
ci.dist.by.unins <- confint(dist.by.unins)
ci.dist.by.unins
```

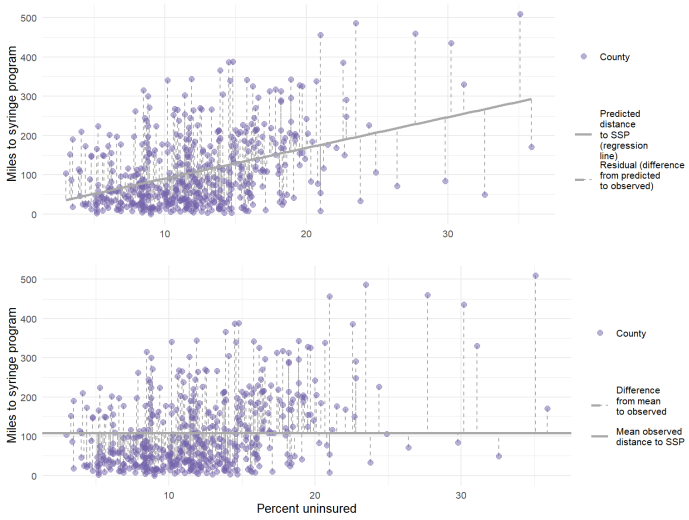
```
##                 2.5 %    97.5 %
## (Intercept) -7.512773 32.472391
## pctunins     6.299493  9.338435
```

Interpretation:

The percentage of uninsured residents in a county is a statistically significant predictor of the distance to the nearest syringe program (b = 7.82; p = 0). For every 1% increase in uninsured residents in a county, the predicted distance to the nearest syringe program increases by 7.82 miles. The value of the slope in the sample is 7.82 and the value of the slope is likely between 6.30 and 9.34 in the population that the sample came from (95% CI: 6.30-9.34). With every 1% increase in uninsured residents, the nearest syringe program is between 6.30 and 9.34 more miles away. These results suggest that communities with a larger percentage of uninsured are further from this resource, which may exacerbate existing health disparities.

Model significance and model fit

- Is the model better than the mean at explaining the data? (model significance)
- How much better is the model compared to the mean at explaining the data? (model fit)



Model significance NHST Step 1: Write the null and alternate hypotheses

H0: A model including percentage uninsured in a county is no better at explaining the distance to syringe programs than a baseline model of the mean value of distance

HA: A model including percentage uninsured in a county is no better at explaining the distance to syringe programs than a baseline model of the mean value of distance

NHST Step 2: Compute the test statistic

- The test statistic for this model is F
- F is a ratio of explained-to-unexplained variance
  - *How much more of the variation in the outcome (distance) did the model explain compared to the mean (numerator)*
  - *How much did it leave unexplained (denominator)*
- When the model explains more than it leaves unexplained, F gets larger and the model is considered *statistically significantly better than the mean at explaining the outcome*
- Here the value is F(1, 498) = 102.2

NHST Step 3: Calculate the probability that your test statistic is at least as big as it is if there is no relationship (i.e., the null is true)

There is a tiny probability ($p < .001$) of an F as big as 102.2 or bigger if the null hypothesis were true

NHST Steps 4 & 5: Reject or retain the null hypothesis

Given the tiny p-value, reject the null hypothesis in favor of the alternate hypothesis that percentage uninsured is helpful in explaining distance to syringe programs from a county

Add to the interpretation:

A simple linear regression analysis found that the percentage of uninsured residents in a county is a statistically significant predictor of the distance to the nearest syringe program (b = 7.82; p < .001). For every 1% increase in uninsured residents, the predicted distance to the nearest syringe program increases by 7.82 miles. The value of the slope is likely between 6.30 and 9.34 in the population that the sample came from (95% CI: 6.30-9.34). With every 1% increase in uninsured residents, there is likely a 6.30 to 9.34 increase in the miles to the nearest syringe program. The model was statistically significantly better than the mean of the distance to syringe program at explaining distance to syringe program [F(1, 498)=102.2; p < .001].

Last thing! Model fit

- R^2 is the measure of model fit for linear regression
- R^2 is the percentage of the total variability in the outcome that is explained by the model
  - *Adjusted R^2 reduces the R^2 a small amount for each predictor in the model (preferred)*
- In this case, R^2 is .169, which can be multiplied by 100 for interpretation as a percentage
  - *16.9% of the variation in distance to a syringe program can be explained by the uninsured percentage in a county*

All the things to report with regression

The things that should be reported following any simple linear regression analysis:

- an interpretation of the value of the slope (b)
- the significance of the slope (t and p; confidence intervals)
- the significance of the model (F and p)
- model fit (R^2 or R^2_{adj})

A simple linear regression analysis found that the percentage of uninsured residents in a county is a statistically significant predictor of the distance to the nearest syringe program (b = 7.82; p < .001). For every 1% increase in uninsured residents, the predicted distance to the nearest syringe program increases by 7.82 miles. The value of the slope is likely between 6.30 and 9.34 in the population that the sample came from (95% CI: 6.30-9.34). With every 1% increase in uninsured residents, there is likely a 6.30 to 9.34 increase in the miles to the nearest syringe program. The model was statistically significantly better the mean of the distance to syringe program at explaining distance to syringe program [F(1, 498)=102.2; p < .001] and explained 16.9% of the variance in the outcome. These results suggest that communities with lower insurance rates are further from this resource, which may exacerbate existing health disparities.

The End

- We will add more variables to the model to finish answering our research question next week!