Applied Linear Modeling

September 24, 2019

To-do

- Exercises
  - *Communicate your results activity*
- Bonus warm-up
  - *Anscombe's quartet*
- We are using the same data from last week and a new R Markdown workshop file:
  - *week-5-workshop.Rmd*
  - *dist_ssp_amfar_ch9.csv*
- New R packages for today:
  - *foreign*
  - *broom*

All the things for today

- Discussion of exercises
- Assumption checking activity
- Workshop
  - *Multiple linear regression*

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

```
# 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 multiple regression

$$outcome = b_0 + b_1 \cdot predictor + b_2 \cdot predictor + \ldots + error$$

- Predictors can be continuous or categorical
- It is a good idea to have at least 10 observations for each predictor
  - *This is not a strict rule and opinions vary, but using many predictors with a small sample is usually considered poor statistical practice*

Revisiting our original model

```
# 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
```


Revisiting the confidence intervals

```
# 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
```

Interpreting it all:

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 percent of people insured are further from this resource, which may exacerbate existing health disparities.

Adding a binary variable to the model

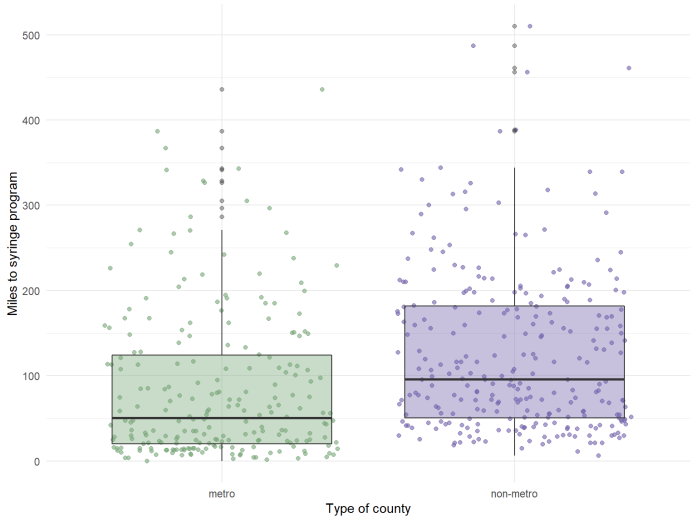
- Uninsured percentage accounted for 16.9% of the variation in distance to syringe program for counties, leaving about 83% still unexplained
- It would make sense that bigger cities are more likely to have syringe programs, so the `metro` variable seems like it might help to explain how far away a county is from a syringe program
- A boxplot might help better understand this relationship

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

# metro and distance to SSP
dist.ssp %>%
  ggplot(aes(x = metro, y = dist_SSP, fill = metro)) +
  geom_jitter(aes(color = metro), alpha = .6) +
  geom_boxplot(aes(fill = metro), alpha = .4) +
  labs(x = "Type of county",
       y = "Miles to syringe program") +
  scale_fill_manual(values = c("#78A678", "#7463AC"), guide = FALSE) +
  scale_color_manual(values = c("#78A678", "#7463AC"), guide = FALSE) +
  theme_minimal()
```


Distance to syringe program and metro status

- It looks like metro counties are closer to syringe programs, on average



Adding metro to the model

- Try naming the new regression object something different since it includes different variables now

```
# linear regression distance to syringe program by
# uninsured percent and metro status in 500 counties
dist.by.unins.metro <- lm(formula = dist_SSP ~ pctunins +
                      metro, data = dist.ssp,
                      na.action = na.exclude)
summary(object = dist.by.unins.metro)
```

```
##
## Call:
## lm(formula = dist_SSP ~ pctunins + metro, data = dist.ssp, na.action = na.exclude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -219.80  -60.07  -18.76   48.33  283.96
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.4240    10.3621   0.330 0.741212
## pctunins         7.3005     0.7775   9.389  < 2e-16 ***
## metronon-metro  28.0525     7.7615   3.614 0.000332 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.89 on 497 degrees of freedom
## Multiple R-squared:  0.1915, Adjusted R-squared:  0.1883
## F-statistic: 58.88 on 2 and 497 DF,  p-value: < 2.2e-16
```


Get some confidence intervals

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

```
##                    2.5 %    97.5 %
## (Intercept)    -16.934973 23.782947
## pctunins         5.772859  8.828114
## metronon-metro  12.803152 43.301754
```


Interpreting the multiple regression model results

- Interpreting the slopes
  - *Value and direction of the slopes*
  - *Interpretation of the value of the slopes*
  - *Interpretation of the direction of the slopes*
  - *Significance of the slope*
- Model fit
- Model significance

Interpreting the slope of percent uninsured

- The slope for `pctunins` was 7.30 (b = 7.30)
  - *For every one additional percent of uninsured people, the distance to a syringe program **increases** by 7.30 miles*
  - *The slope was statistically significantly different from 0 (t = 9.39; p < .001)*
  - *In the population, the slope was likely between 5.77 and 8.83*

There was a statistically significant positive relationship between percent uninsured in a county and the distance to a syringe program (b = 7.30; t = 9.39; p < .001); for every 1 percent increase in uninsured, the nearest syringe program is 7.30 more miles away. In the population, the distance to a syringe program likely increases by 5.77 to 8.83 miles for every additional 1 percent uninsured in the county (95% CI: 5.77-8.83).

Interpreting the slope of metro

- The slope of a categorical variable is interpreted with respect to its *reference group*
- Write out the regression model to see what this means…
  - *distance to syringe program = 3.42 + 7.3 * percent uninsured + 28.05 * non-metro*
- Substitute in values for an **example county** with 10% uninsured in a **non**-metro area:
  - *distance to syringe program = 3.42 + 7.3 * 10 + 28.05 * 1*
  - *distance to syringe program = 104.48*
- Substitute in the values for an **example county** with 10% uninsured in a **metro** area:
  - *distance to syringe program = 3.42 + 7.3 * 10 + 28.05 * 0*
  - *distance to syringe program = 76.43*

Interpreting with respect to reference group

- The reference group is the group not shown in the output
  - *In this case, non-metro is shown so **metro** is the reference group*
- The distance to a syringe program 28.05 miles further in **non-metro** counties compared to **metro** counties (b = 28.05).
- The slope was statistically significantly different from 0 (t = 3.61; p < .001)
- In the population, the slope was likely between 12.80 and 43.30

There was a statistically significant relationship metro status of a county and the distance to a syringe program (b = 28.05; t = 3.61; p < .001); non-metro counties are 28.05 miles further from the nearest syringe program compared to metro counties. In the population, the distance to a syringe program is 12.80 to 43.30 miles further for non-metro than for metro counties (95% CI: 12.80-43.30).

Model significance and model fit

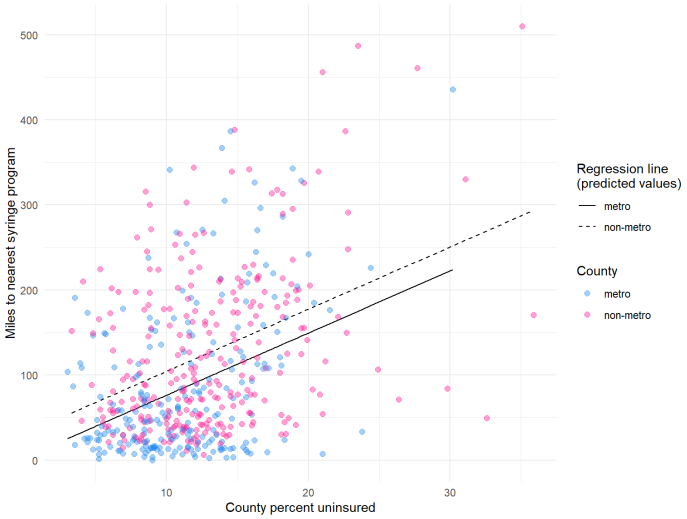
- The model was statistically significantly better than the mean distance at explaining the variation in distance to syringe program [F(2, 497) = 58.88; p < .001].
- The model explained 18.83% of the variation in distance (R^2_{adj} = .1883)

There was a statistically significant relationship metro status of a county and the distance to a syringe program (b = 28.05; t = 3.61; p < .001); non-metro counties are 28.05 miles further from the nearest syringe program compared to metro counties. In the population, the distance to a syringe program is 12.80 to 43.30 miles further for non-metro than for metro counties (95% CI: 12.80-43.30). The model was statistically significant [F(2, 497) = 58.88; p < .001] and explained 18.83% of the variation in distance.

Visualizing the model

```
# graphing the regression model with percent uninsured and metro
dist.ssp %>%
  ggplot(aes(x = pctunins, y = dist_SSP, group = metro)) +
  geom_line(data = broom::augment(dist.by.unins.metro),
            aes(y = .fitted, linetype = metro)) +
  geom_point(aes(color = metro), alpha = .4, size = 2) +
  theme_minimal() +
  scale_color_manual(values = c("dodgerblue2", "deeppink"), name = "County") +
  labs(y = "Miles to nearest syringe program", x = "County percent uninsured") +
  scale_linetype_manual(values = c(1,2), name = "Regression line\n(predicted values)")
```

Visualizing the model



Adding more variables to the model

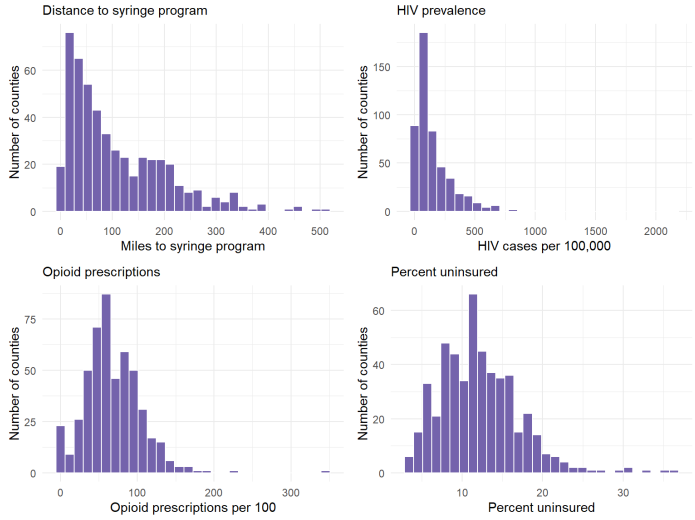
- HIV prevalence

A tiny taste of assumption checking

Before adding more variables to the model, let's look at a few assumptions:

- Observations are independent
- Outcome variable is continuous
- Continuous outcome and predictor variables are normally distributed
- The relationship between continuous predictors and the outcome is linear (linearity)
- The variance is constant with the points distributed equally around the line (homoscedasticity)
- The residuals are independent
- The residuals are normally distributed

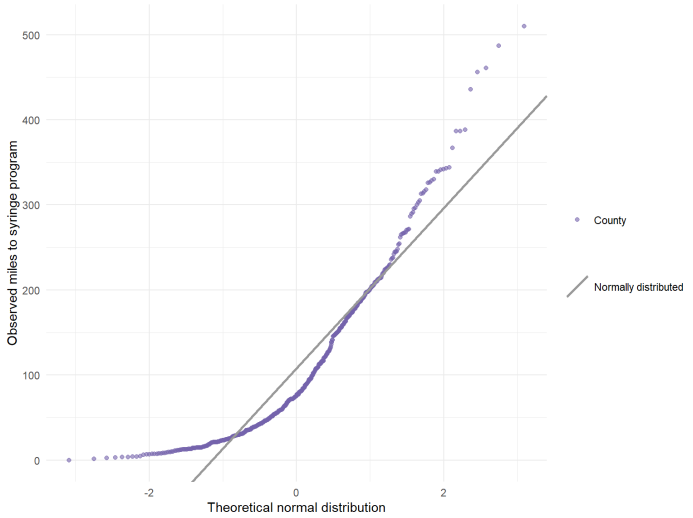
Reminder about normality



Another way to check normality: Q-Q plots

```
# Q-Q plot of distance variable to check normality
dist.ssp %>%
  ggplot(aes(sample = dist_SSP)) +
  stat_qq(aes(color = "County"), alpha = .6) +
  geom_abline(aes(intercept = mean(dist_SSP),
                  slope = sd(dist_SSP),
                  linetype = "Normally distributed"),
              color = "gray60", size = 1) +
  theme_minimal() +
  labs(x = "Theoretical normal distribution",
       y = "Observed miles to syringe program") +
  scale_color_manual(values = "#7463AC", name = "") +
  scale_linetype_manual(values = 1, name = "")
```

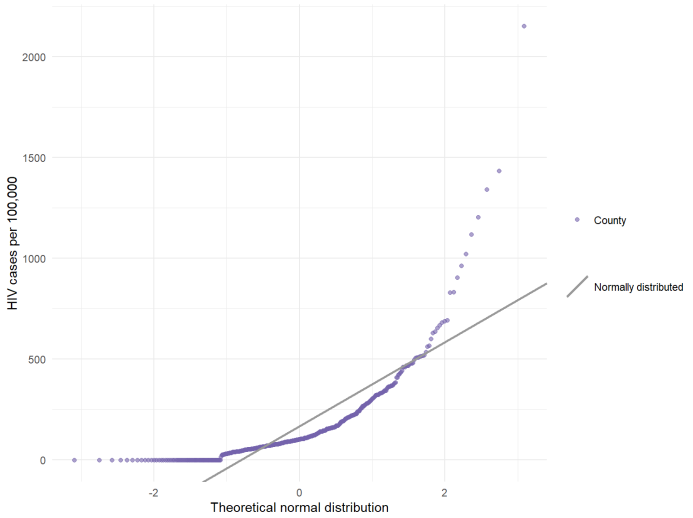

Another way to check normality: Q-Q plots



Another way to check normality: Q-Q plots

```
# Q-Q plot of HIV variable to check normality
dist.ssp %>%
  ggplot(aes(sample = HIVprevalence)) +
  stat_qq(aes(color = "County"), alpha = .6) +
  geom_abline(aes(intercept = mean(HIVprevalence),
                  slope = sd(HIVprevalence),
                  linetype = "Normally distributed"),
              color = "gray60", size = 1) +
  theme_minimal() +
  labs(x = "Theoretical normal distribution",
       y = "HIV cases per 100,000") +
  scale_color_manual(values = "#7463AC", name = "") +
  scale_linetype_manual(values = 1, name = "")
```


Another way to check normality: Q-Q plots

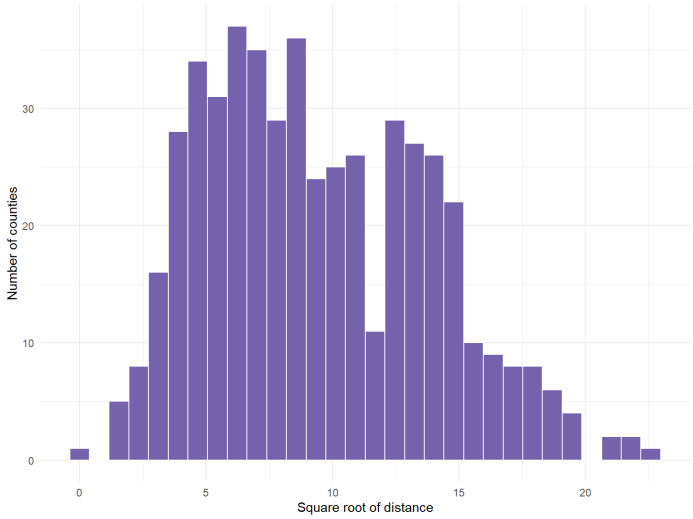


One strategy: Transforming non-normal variables

- Both of the non-normal variables were right-skewed
- Transformations that work best to make right-skewed data more normal are
  - *square root*
  - *cube root*
  - *reciprocal*
  - *log transformations*

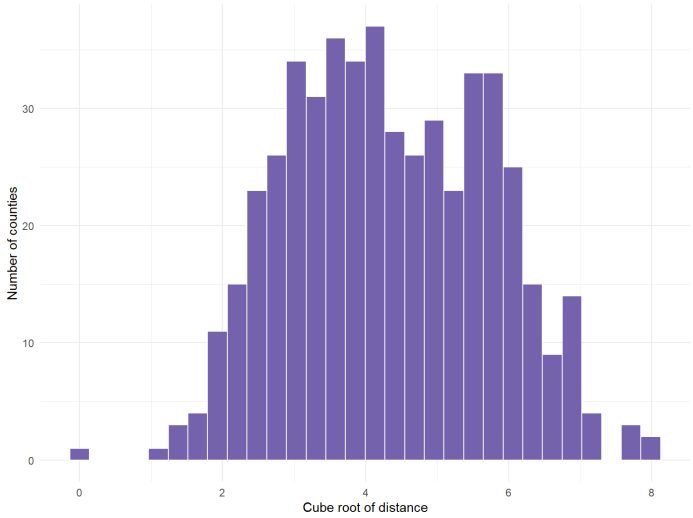
Square root transformation of distance

```
dist.ssp %>%
  ggplot(aes(x = sqrt(x = dist_SSP))) +
  geom_histogram(fill = "#7463AC", col = "white") +
  labs(x = "Square root of distance", y = "Number of counties")+
  theme_minimal()
```



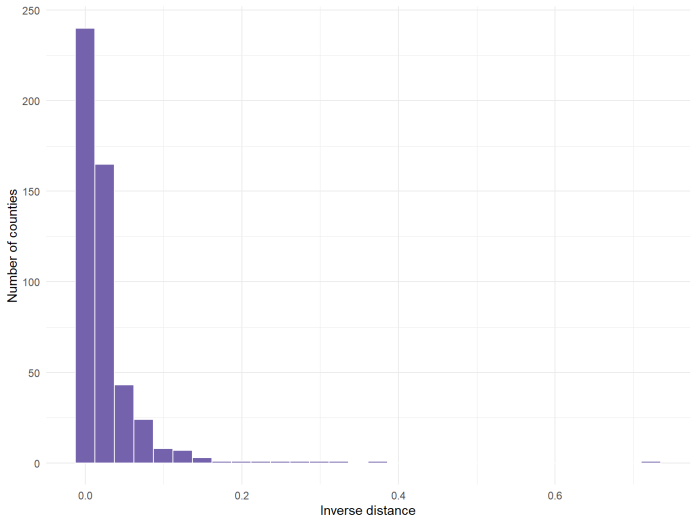
Cube root transformation of distance

```
dist.ssp %>%
  ggplot(aes(x = (dist_SSP)^(1/3))) +
  geom_histogram(fill = "#7463AC", col = "white") +
  labs(x = "Cube root of distance", y = "Number of counties") +
  theme_minimal()
```



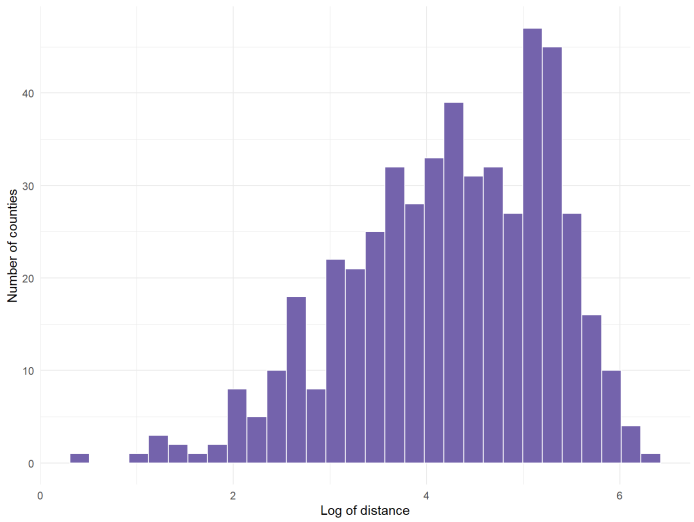
Inverse of distance

```
dist.ssp %>%
  ggplot(aes(x = 1/dist_SSP)) +
  geom_histogram(fill = "#7463AC", col = "white") +
  labs(x = "Inverse distance", y = "Number of counties") +
  theme_minimal()
```

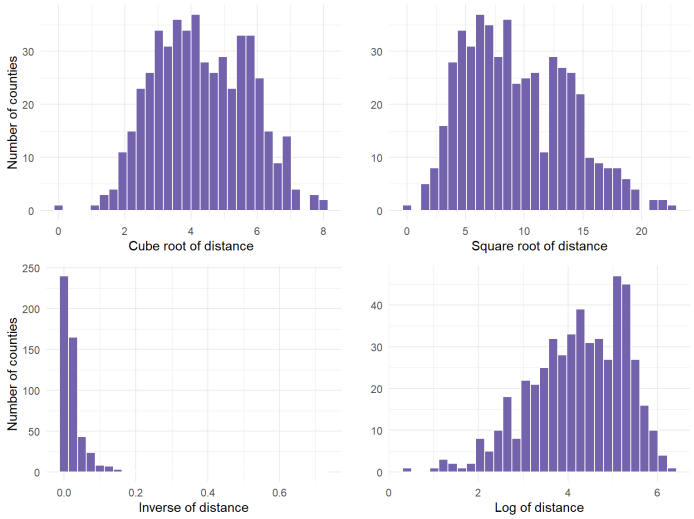


Log of distance

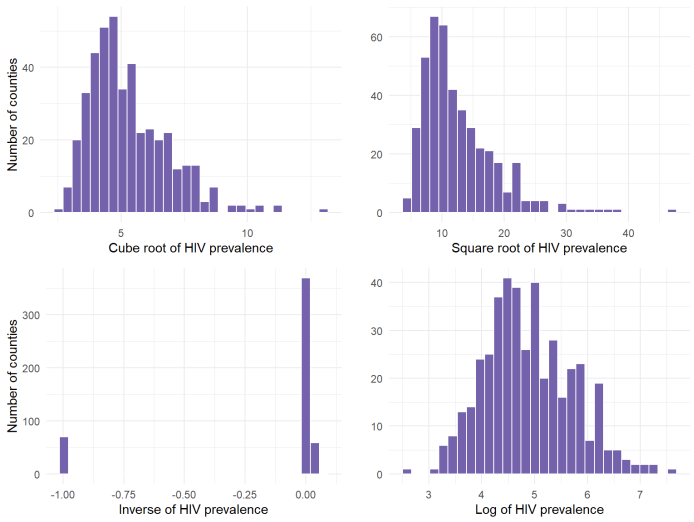
```
dist.ssp %>%
 ggplot(aes(x = log(x = dist_SSP))) +
 geom_histogram(fill = "#7463AC", col = "white") +
 labs(x = "Log of distance", y = "Number of counties") +
 theme_minimal()
```



All transformations for distance



All transformations for HIV prevalence



Model with transformed variables

- Cube root of distance
- Log of HIV prevalence

```
# linear regression of distance by percent uninsured, HIV prevalence,
# metro status
dist.full.model <- lm(formula = (dist_SSP)^(1/3) ~ pctunins +
                      log(x = HIVprevalence) + metro,
                      data = dist.ssp,
                      na.action = na.exclude)
summary(object = dist.full.model)
```

```
##
## Call:
## lm(formula = (dist_SSP)^(1/3) ~ pctunins + log(x = HIVprevalence) +
##     metro, data = dist.ssp, na.action = na.exclude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2624 -0.8959 -0.0745  0.8032  3.1967
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.03570    0.38448   7.896 2.48e-14 ***
## pctunins                  0.11269    0.01220   9.237  < 2e-16 ***
## log(x = HIVprevalence) -0.06529    0.07729  -0.845 0.398691
## metronon-metro            0.48808    0.12763   3.824 0.000151 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.214 on 426 degrees of freedom
##   (70 observations deleted due to missingness)
## Multiple R-squared:  0.2372, Adjusted R-squared:  0.2318
## F-statistic: 44.16 on 3 and 426 DF,  p-value: < 2.2e-16
```

Interpreting the model

- There is a positive and statistically significant relationship between percent uninsured and the transformed distance variable (b = .11; t = 9.24; p < .001)
  - *As uninsured percentage goes up, distance to a syringe program goes up*
- There is a negative and non-significant relationship between transformed HIV prevalence and transformed distance (b = -.07; t = -.85; p = .40)
- Non-metro counties are further from syringe programs compared to metro counties (b = .49; t = 3.82; p < .001)
- The model is statistically significantly better than the mean of the outcome at explaining variation in the outcome [F(3, 426) = 44.16; p < .001]
- The model explained 23.18% of the variation in the transformed distance to syringe program (R^2_{adj} = .2318)

Using the Partial-F test to choose a model

- Now that we have examined three models (`dist.by.unins`, `dist.by.unins.metro`, and `dist.full.model`), how do we choose which one to report?
  - *First, the model should address the research question of interest*
  - *Second, the model should include variables—if any—that have been demonstrated important in the past to help explain the outcome*
- After answering the research question and including available variables demonstrated important in the past, choosing a model can still be complicated
- One tool for choosing between two linear regression models is a statistical test called the Partial-F test
  - *The Partial-F test compares the fit of two **nested** models to determine if the additional variables in the larger model improved the model fit enough to warrant keeping them in the model*
    - **Nested models** must have the same observations and variables; the larger model can have more variables but must include ALL of the variables from the small model

Are the `dist.by.unins.metro` and `dist.full.model` models comparable using Partial-F?

- No
- Transforming the outcome and predictor variables for one model but not the other means the models are not *nested*
- Nested models must have the *same exact variables*, including transformed versions
  - *The larger model can have more variables, but must have all of the exact same variables that were used in the smaller model*
- Nested models must also have the *same observations*

Making the models comparable

- Use the transformed variables with the smaller model
- Remove observations with missing values for any variables in the larger model
  - *The log transformation results in NA for any values of 0 since the log of 0 is undefined*
- One strategy:

```
# drop observations with missing data
dist.ssp.small <- dist.ssp %>%
  select(HIVprevalence, dist_SSP, metro, pctunins) %>%
  mutate(log.HIV = log(x = HIVprevalence)) %>%
  drop_na()
```

Re-run models with the new data frame

```
# re-run regressions with the smaller data frame
dist.unins.metro.trans <- lm(formula = (dist_SSP)^(1/3) ~ pctunins + metro,
                     data = dist.ssp.small, na.action = na.exclude)

dist.full.model.small <- lm(formula = (dist_SSP)^(1/3) ~ pctunins +
                            metro + log(x = HIVprevalence),
                     data = dist.ssp.small, na.action = na.exclude)
```


Calculating the Partial-F statistic

$$F_{partial} = \frac{\frac{R^2_{full} - R^2_{reduced}}{q}}{\frac{1 - R^2_{full}}{n - p}}$$

Where:

- R^2_{full} is the R^2 for the larger model
- $R^2_{reduced}$ is the R^2 for the smaller nested model
- n is the sample size
- q is difference in the number of parameters for the two models
- p is the number of parameters in the larger model

The $F_{partial}$ statistic has q and n - p degrees of freedom.

Conducting a Partial-F test

- There is a way to do this using R code using the anova( ) function
- Enter the name of the smaller model first and then the larger model into anova( )
- The function will compare the two models using a Partial-F test

```
# partial F test for dist.by.unins.metro and dist.full.model
anova(object = dist.unins.metro.trans, dist.full.model.small)
```

```
## Analysis of Variance Table
##
## Model 1: (dist_SSP)^(1/3) ~ pctunins + metro
## Model 2: (dist_SSP)^(1/3) ~ pctunins + metro + log(x = HIVprevalence)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    427 629.14
## 2    426 628.09  1    1.0523 0.7137 0.3987
```

The Partial-F was .71 and the p-value was .40.

NHST Step 1: Write the null and alternate hypotheses

H0: The larger model is no better than the smaller model at explaining the outcome

HA: The larger model is better than the smaller model at explaining the outcome

NHST Step 2: Compute the test statistic

The Partial-F was .71

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 is large (p = .40)

NHST Step 4 & 5: Reject or retain the null hypothesis based on the probability

- The null hypothesis is retained
- This suggests that the larger model with uninsured percentage and metro status was not a better model for reporting than the full linear model
  - *When the larger one is not better, use the smaller one*
- Show the model results:

```
summary(dist.unins.metro.trans)
```

```
##
## Call:
## lm(formula = (dist_SSP)^(1/3) ~ pctunins + metro, data = dist.ssp.small,
##     na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.15317 -0.88038 -0.06389  0.81487  3.13344
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.73791    0.15349  17.838  < 2e-16 ***
## pctunins        0.10938    0.01155   9.470  < 2e-16 ***
## metronon-metro  0.52530    0.11974   4.387 1.45e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.214 on 427 degrees of freedom
## Multiple R-squared:  0.2359, Adjusted R-squared:  0.2323
## F-statistic: 65.92 on 2 and 427 DF,  p-value: < 2.2e-16
```

```
confint(dist.unins.metro.trans)
```

```
##                    2.5 %    97.5 %
## (Intercept)    2.43622367 3.0395891
```



```
## pctunins       0.08667989 0.1320849
## metronon-metro 0.28994762 0.7606574
```

Write the interpretation

A linear regression model including percentage uninsured and metro status of a county to explain the transformed distance in miles to the nearest syringe program was statistically significantly better than a baseline model at explaining the outcome [F(2, 427) = 65.92]. The model explained 23.23% of the variation in distance to syringe programs.

Percentage uninsured was a statistically significant predictor of distance to the nearest syringe program (b = .11; t = 9.47; p < .001). The 95% confidence interval for the coefficient suggested that a 1% increase in uninsured in a county was associated with a .09 to .13 increase in the transformed distance to the nearest syringe program.

Metro status was also a significant predictor of distance to a syringe program (b = .53; t = 4.39; p < .001). The 95% confidence interval for the coefficient suggested that non-metro counties have a .29 to .76 increase in transformed distance to a syringe program. Overall, the results suggest that more rural counties and counties with more uninsured are further from this health service, potentially exacerbating existing health disparities.

A note about adjusted R-squared

- Adjusted R^2 penalizes the model fit so that it does not pay to add extra variables unless they account for some variance in the outcome

$$R^2_{adj} = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$$

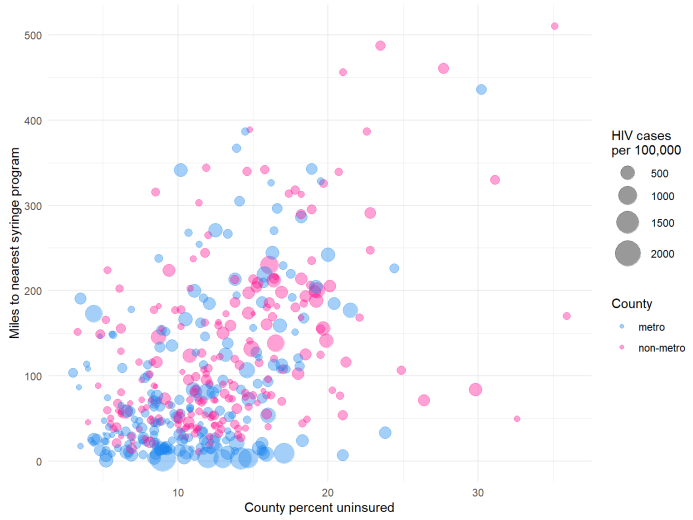
Where:

- n is sample size
- k is the number of parameters
- R^2 is the unadjusted model fit

A cool plot

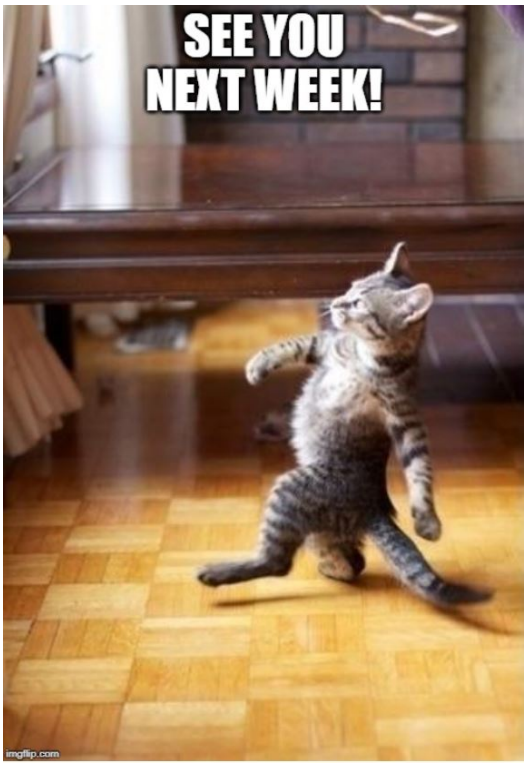
- Although we determined that we would report the model that only included percent uninsured and metro status, there may be situations when we would want to plot three variables together with a model
- We can use a bubble plot!

```
# graphing the regression model with three predictors
dist.ssp.small %>%
  ggplot(aes(x = pctunins, y = dist_SSP, group = metro, size = HIVprevalence)) +
  geom_point(aes(color = metro), alpha = .4) +
  theme_minimal() +
  scale_color_manual(values = c("dodgerblue2", "deeppink"), name = "County") +
  scale_size(range = c(1, 10), name="HIV cases\nper 100,000") +
  labs(y = "Miles to nearest syringe program", x = "County percent uninsured")
```

The End

- We will test all the assumptions and do more diagnostics next week!



9/24/2019 Applied Linear Modeling (1)