Linear Regression

Checking assumptions and conducting diagnostics

Jenine Harris
Brown School

Importing and merging data sources

```
# distance to syringe program data
dist.ssp <- read.csv(file = "/Users/harrisj/Box/teaching/Teaching/Fall20

# regression
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
```

Codebook

Leslie looked through the variables and the codebook and determined that the variables had 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

Assumptions of simple linear regression

The calculations underlying the simple linear regression model of `dist_SSP ~ pctunins` are based on several assumptions about the data:

- Observations are independent
- The outcome is continuous
- The relationship between the two variables 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

Checking the independent observations assumption

- While the data had been collected in a reasonable way, one could argue that counties in the same state were not really independent because, like siblings in the same household, they may have the same characteristics and therefore may not vary as much from each other as truly independent areas would.
- This assumption is often seen as a possible problem for geographic data, but researchers continue to use linear regression and other similar models to work this type of data.
- This is a limitation to discuss in reporting the results.

Checking the continuous variables assumption

The outcome is continuous, so this assumption is met.

Checking the linearity assumption

- The *linearity* assumption is met if a scatterplot of the two variables shows a relationship that falls along a line.
- The earlier plot showing purple points and straight line drawn through them suggests that this assumption may be met.
- When graphed, the points fall generally the straight line.
- However, unlike the graph of water access and female education, the data points seem to not follow the line as well.
- Using a *Loess curve*, the relationship between the two variables is shown without constraining the line to be straight.
- In this case, a Loess curve shows deviation from a linear relationship, especially at lower values of the predictor, percent uninsured.

Checking the homoscedasticity assumption

- The assumption of homoscedasticity requires the data points to be evenly distributed around the regression line.
- For simple linear regression, this assumption is checked the same way as it was checked for the correlation analysis.
- Specifically, a scatterplot where all the points were relatively evenly spread out around the line would be one way to check the assumption.
- In the plots, the points seem closer to the line on the far left and then are more spread out around the line starting at the higher values of percentage uninsured.
    - It appears this assumption is *not met*.

Using Breusch-Pagan to test constant variance

- Breusch-Pagan can also be used to statistically test the null hypothesis that the variance is constant.

```
# testing for equal variance
const.var.test <- lmtest::bptest(formula = dist.by.unins)
const.var.test
```

```
##
##      studentized Breusch-Pagan test
##
## data:  dist.by.unins
## BP = 46.18, df = 1, p-value = 1.078e-11
```

- The Breusch-Pagan test statistic has a tiny p-value (BP = 46.18; $p < .001$), indicating that the null hypothesis of constant variance would be rejected.

- This is consistent with the scatterplot showing higher variance on the right hand side.

Plotting residuals to examine constant variance

- If there is a pattern in a scatterplot of residuals and predicted values, this indicates that the model was better at some kinds of predictions than others, suggesting bias.

- A dashed line is shown to indicate no relationship between the fitted (or predicted) values and the residuals, which would be the idea situation to meet the assumption.

- This line is a helpful reference point for looking at these types of graphs; for the homoscedasticity assumption to be met, the points should be roughly evenly distributed around the dashed line with no clear patterns.

Testing the independence of residuals assumption

*Both of the last two assumptions are about residuals, which are the distances between each data point and the regression line.

- Conceptually, residuals are the variation in the data that the regression line does not explain.
- The first assumption for residuals is that the residuals are independent or unrelated to each other.
- Residuals that are independent are residuals that do not follow a pattern.
- A pattern in the residuals suggests that the regression model is doing better for certain types of observations and worse for others.
- The **Durbin-Watson** test can be used to determine whether the model violates the assumption of independent residuals.
- The null hypothesis for the Durbin-Watson test is that *the residuals are independent*; the alternative hypothesis is that the residuals are not independent.
- A Durbin-Watson or D-W statistic with a value of two indicates perfectly independent residuals.

```r
# test independence of residuals
lmtest::dwtest(formula = dist.by.unins)
```

Testing the normality of residuals assumption

- The last assumption to check is normality of residuals; normally distributed residuals indicate that the regression line is far above a few points, far below a few others, and relatively near most of the points.

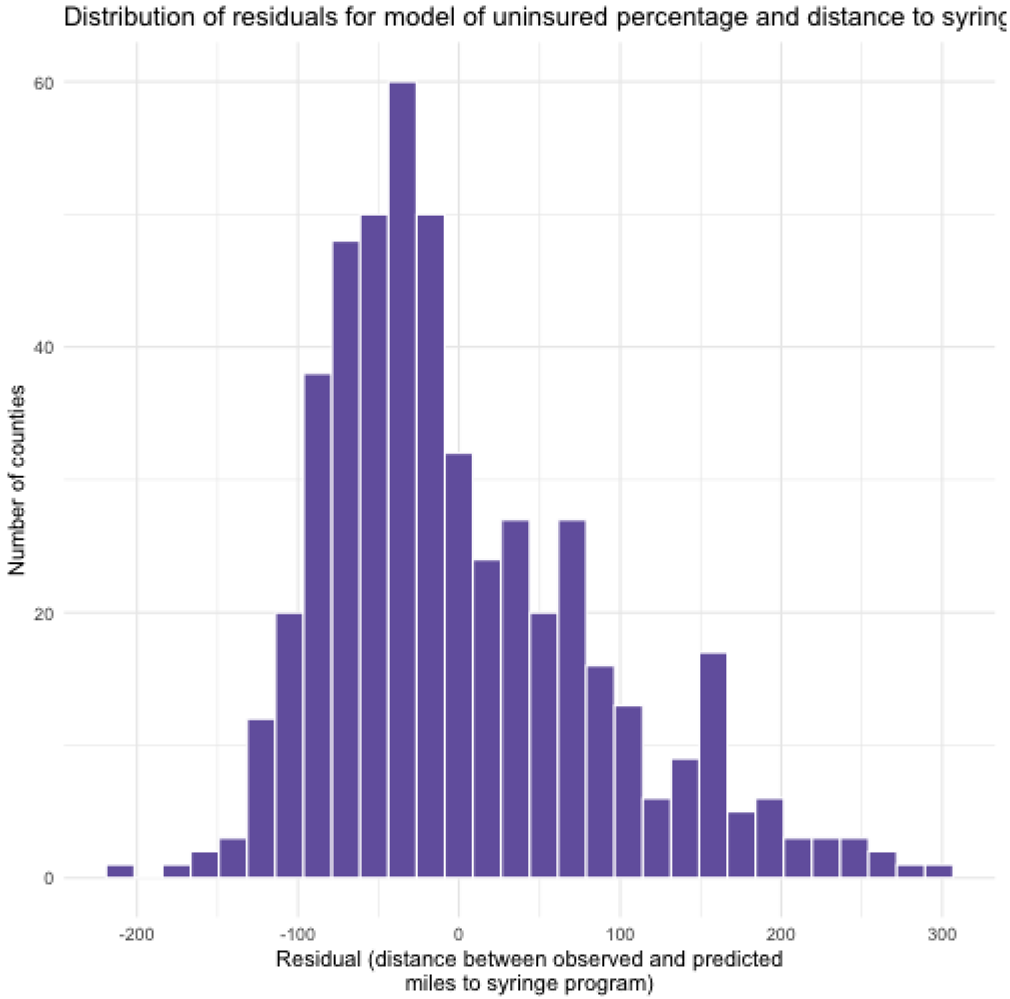
- If the residuals are skewed, that would mean that the regression line does a better job at explaining either the higher values of the outcome or the lower values of the outcome.

- Check normality using a histogram or Q-Q plot

- The `dist.by.unins` model object includes residuals to use in the plot and that they can be piped into `ggplot()` by first using the `data.frame()` function.

```
# check residual plot of uninsured percent and distance to syringe progr
data.frame(dist.by.unins$residuals) %>%
  ggplot(aes(x = dist.by.unins.residuals)) +
  geom_histogram(fill = "#7463AC", col = "white") +
  theme_minimal() +
  labs(x = "Residual (distance between observed and predicted\nmiles to
       y = "Number of counties",
       title = "Distribution of residuals for model of uninsured percent
```

Examine histogram of residuals



Interpreting the results of the assumption checking

- The assumptions it met were:
  - continuous variables,
  - and independence of residuals.
- The assumptions it failed were:
  - linearity,
  - homoscedasticity,
  - and normally distributed residuals.
- It may violate independence of observations, but this is not as clear.
- Because it does not meet all the assumptions, the model is considered biased and should be interpreted with caution.

Using model diagnostics to find outliers and influential values

- In addition to testing *assumptions*, model **diagnostics** are useful for determining whether there are any observations that are outliers or influential observations that may be having some impact on the model.

- An **outlier** is an observation with unusual values.

- A **regression outlier** has an unusual value of the outcome given its value(s) of predictor(s).

- An **influential observation** changes the slope of the regression line.

- There are several measures to help identify outliers and influential observations: **standardized residuals**, **df-betas**, **Cook's distance**, and **leverage.**

- One good strategy for identifying the truly problematic observations is to identify those that observations that are outliers or influential observations by on two or more of these four measures.

Using standardized residuals to find outliers

- *Standardized residuals* are z-scores for the residual values.

- The residuals are the distances between the observed and predicted values of the outcome and z-scores over 1.96 or below -1.96 (often rounded to 2) are 1.96 standard deviations or more away from the mean of the measure.

- Standardized residuals can be computed using the `rstandard()` function on the model object and then added to the data frame as a new variable.

- Create a new variable containing standardized residuals and name it `standardres`.

```
# add standardized residuals to data frame
dist.ssp.diag <- dist.ssp %>%
  mutate(standardres = rstandard(model = dist.by.unins))
```

Find high values of standardized residuals

- Once the standardized residuals were added to the data, the counties with high standardized residuals can be examined.

- After adding standardized residuals to the data frame, find the counties with large standardized residuals by examining the subset of counties where the absolute value (`abs()`) of the residuals is greater than 1.96.

- Use the `filter()` function and the absolute value function `abs()` to choose the observations with standard deviations above 1.96 or below -1.96.

- Print the information needed to examine the counties that were outliers and use `select()` to select the relevant information including the two variables in the model, the name of the county and the state it is in, and the predicted values and standardized residuals.

```
# get a subset of counties with standardized residuals > 2
dist.ssp.diag %>%
  filter(abs(standardres) > 1.96) %>%
  select(county, STATEABBREVIATION, dist_SSP, pctunins, predicted, stand
```

```
##              county STATEABBREVIATION dist_SSP pctunins predicted standardre
## 1       webb county                TX   436.00     30.2 248.61252    2.21263
## 2   garfield county                NE   300.00      8.8  81.28669    2.54945
## 3      starr county                TX   510.00     35.1 286.92545    2.65633
```

Using df-betas to find influential values

- The *df-beta* measure removes each observation from the data frame, conducts the analysis again, and compares the intercept and slope for the model with and without the observation.
- Observations with high df-beta values, usually with a cutoff of greater than 2, may be influencing the model.
- Using the same strategy as with the standardized residuals, identify counties with high df-betas.
- df-betas are different for slope and intercept, so she will have to use subsetting and choose the part of the `dist.by.unins` object with the intercept and slope separately.

```
# get dfbetas and add to data frame
# there will be one new variable per parameter
dist.ssp.diag <- dist.ssp %>%
  mutate(standardres = rstandard(model = dist.by.unins)) %>%
  mutate(dfbeta.intercept = dfbeta(dist.by.unins)[ , 1]) %>%
  mutate(dfbeta.slope = dfbeta(dist.by.unins)[ , 2])

# get subset of states with dfbetas > 2 for intercept and slope
dist.ssp.diag %>%
  filter(abs(dfbeta.intercept) > 2 | abs(dfbeta.slope) > 2) %>%
  select(county, STATEABBREVIATION, dist_SSP, pctunins, predicted,
         dfbeta.intercept, dfbeta.slope)
```

Using Cook's Distance to find influential values

- *Cook's Distance* is often shortened to Cook's D and is computed in a very similar way to the df-beta.
- That is, each observation is removed and the model is re-estimated without it.
- Cook's D then combines the differences between the models with and without an observation for *all the parameters* together instead of one at a time like the df-betas.
- The cutoff for a high Cook's D value is usually 4/n.

```
# cooks distance
# greater than 4/n is some influence
dist.ssp.diag <- dist.ssp %>%
  mutate(standardres = rstandard(model = dist.by.unins)) %>%
  mutate(dfbeta.intercept = dfbeta(dist.by.unins)[ , 1]) %>%
  mutate(dfbeta.slope = dfbeta(dist.by.unins)[ , 2]) %>%
  mutate(cooks.dist = cooks.distance(dist.by.unins))

# find counties with some influence
dist.ssp.diag %>%
  filter(cooks.dist > 4/n()) %>%
  select(county, STATEABBREVIATION, dist_SSP, pctunins, predicted,
         cooks.dist)
```

Using Leverage to find influential values

- *Leverage* is the influence that the observed value of the outcome has on the predicted value of the outcome.

- Specifically, leverage is the amount the predicted value of the outcome would change if the observed value of the outcome was changed by one unit. Leverage values range between 0 and 1.

- To determine which leverage values indicate influential observations, a cutoff of $\frac{2\cdot(k+1)}{n}$ is often used.

- The leverage values to find influential observations are computed by using the `hatvalues()` function.

- The predicted value of y is often depicted as $\hat{y}$, which looks like a y wearing a little hat.

```r
# leverage values
# identify those that are greater than 2(k+1)/n
dist.ssp.diag <- dist.ssp %>%
  mutate(standardres = rstandard(model = dist.by.unins)) %>%
  mutate(dfbeta.intercept = dfbeta(dist.by.unins)[ , 1]) %>%
  mutate(dfbeta.slope = dfbeta(dist.by.unins)[ , 2]) %>%
  mutate(cooks.dist = cooks.distance(dist.by.unins)) %>%
  mutate(lever = hatvalues(dist.by.unins))
```

Summarizing outliers and influential values

- It is often useful to have all the counties identified by these four measures in a single list or table to more easily see all the counties that seemed problematic.

```
# sum the number of times observations were outliers/influential
dist.ssp.diag <- dist.ssp %>%
  mutate(standardres = rstandard(model = dist.by.unins)) %>%
  mutate(dfbeta.intercept = dfbeta(dist.by.unins)[ , 1]) %>%
  mutate(dfbeta.slope = dfbeta(dist.by.unins)[ , 2]) %>%
  mutate(cooks.dist = cooks.distance(dist.by.unins)) %>%
  mutate(lever = hatvalues(dist.by.unins)) %>%
  mutate(outlier.infl = as.numeric(lever > 2*3/n()) +
           as.numeric(cooks.dist > 4/n()) +
           as.numeric(abs(dfbeta.intercept) > 2) +
           as.numeric(abs(dfbeta.slope) > 2) +
           as.numeric(abs(standardres) > 1.96))

# subset those with 2 or more measures indicating outlier/influential
dist.ssp.diag %>%
  filter(outlier.infl >= 2)
```

```
##               county STATEABBREVIATION dist_SSP HIVprevalence opioid_RxRate
## 1        webb county                TX   436.00         215.3          24.9
## 2   garfield county                NE   300.00          -1.0          75.8
```

Full 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 distance to the nearest syringe program is expected to increase 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 baseline 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 type of resource, which may exacerbate existing health disparities.

An examination of the underlying assumptions found that the data fail several of the assumptions for linear regression and so the model should be interpreted with caution; the results do not necessarily generalize to other counties beyond the sample. In addition, regression diagnostics found a number of counties that were problematic. Many of these counties were in Texas, which may suggest that counties in Texas are unlike the rest of the sample and might be considered separately.