Linear Regression

Model significance and model fit

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

Model significance and model fit

- The other p-value toward the bottom of the output for the linear regression is not for the intercept or the slope, it is from a test statistic that measures how much better the regression line is at getting close to the data points compared the mean value of the outcome.
- Essentially, are the predicted values better than the mean value of distance to the syringe program at capturing the relationship between `pctunins` and `dist_SSP`.

Model significance test statistic

- Like the t-statistic is the test statistic for a t-test comparing two means, the F-statistic is the test statistic for linear regression comparing the regression line to the mean.
- It is the same F with the same F-distribution as ANOVA; ANOVA is a special type of linear model where all the predictors are categorical.
- The F-statistic is used to determine whether the line showing the regression model is better overall at getting close to the data points than the line showing the mean of the outcome.

Understanding the F-statistic

- The F-statistic is a ratio of explained information (in the numerator) to unexplained information (in the denominator).
- If a model explains more than it leaves unexplained, the numerator is larger and the F-statistic is greater than 1.
- F-statistics that are much greater than 1 indicate a model is explaining much more of the variation in the outcome than it leaves unexplained.
- Large F-statistics are more likely to be statistically significant.

Computing the F statistic

- $F = \frac{\frac{\sum_{i=1}^{n} (\hat{y}_i - m_y)^2}{k-1}}{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-k}}$

- i is an individual observation, in this case a county

- n is the sample size, or total number of counties
- k is the number of parameters in the model; the slope and intercept are parameters
- y_i is the observed outcome of distance to syringe program for county i
- $\hat{y}_i$ is the predicted value of distance to syringe program for county i

- m_y is the mean of the observed outcomes of distance to syringe program

- The numerator of F was how much the predicted values differ from the mean observed value, *on average*.

- This is divided by how much the predicted values differ from the actual observed values, *on average*.

- The F-statistic is how much a predicted value differs from the mean value on average---which is explained variance, or how much better (or worse) the prediction is than the mean at explaining the outcome---divided by how much an observed value differs from the predicted value on average, which is the residual information or unexplained variance.

Values of the F-statistic

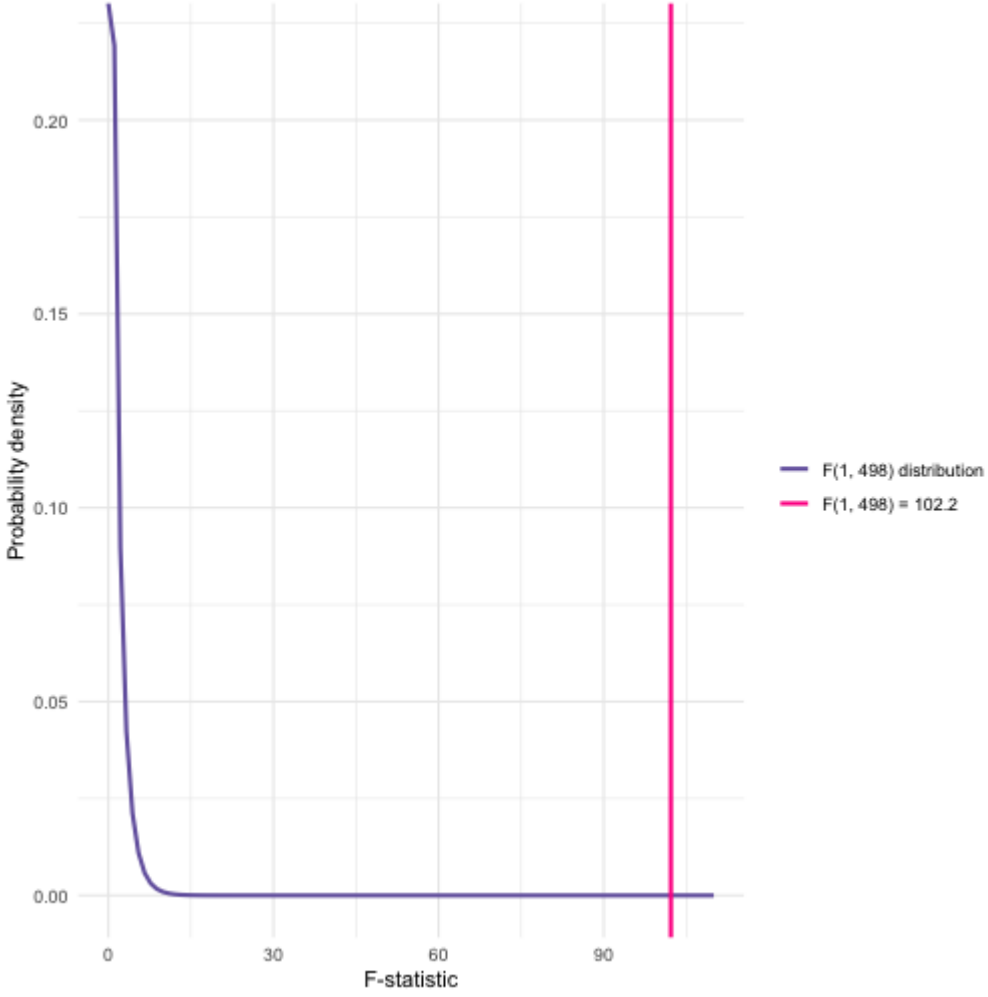
- The F-statistic is always positive due to the squaring of the terms in the numerator and denominator. As a result, the F-distribution starts at zero (when the regression line is exactly the same as the mean) and goes to the right.
- The shape of the F-distribution depends on the number of parameters in the statistical model and the sample size, two degrees of freedom values.

Interpreting the F-statistic

- The more the model explains the variation in the outcome, the larger the F-statistic gets.
- Like t-statistics and chi-squared statistics, larger values of F-statistics are less likely to occur when there is no relationship between the variables.

```
##
## 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
```

Visualizing the F-statistic



F-distribution with 2 and 498 degrees of freedom for model of distance to syringe program by uninsured

F-statistics and probability

- Just like with the t-statistic and chi-squared, the probability of an F-statistic this large or larger *if the null is true* is the area under the F-distribution curve starting at the vertical line (F = 102.2) and going right.
- There is really very little space under the curve from F = 102.2 to the right, which is consistent with the tiny p-value ($p < .001$).
- This is nearly a 0% chance that an F-statistic this large or larger would occur under the null hypothesis that there is no relationship between percentage uninsured and distance to syringe program.
- Essentially, the F-statistic and the associated p-value suggest a statistically significant relationship between percentage uninsured and distance to syringe program at the county level.

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 and its 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.

Understanding the R^2 measure of model fit

- The measure that tells how well the model fits is the R^2 or R-squared.
- The R^2 is computed by squaring the value of the correlation between the 500 observed distances to syringe programs in the 500 counties and the values of distance to syringe program predicted for the 500 counties by the model.
- When the model predicts values that are close to the observed values, the correlation is high and the R^2 is high.
- The R^2 is the amount of variation in the outcome that the model explains and is reported as a measure of **model fit**.
- For the relationship between uninsured percentage and distance to syringe program the R^2 is 0.17.
- To get the percentage of variance explained by the model, multiply by 100 for 17.03% of the variation in distance to syringe programs is explained by the percentage of uninsured people living in a county.
- Subtracting one from R^2 ($1 - R^2$) and again multiplying by 100 for a percent will give the percent of variance *not* explained by the model (here, 82.97).

Reporting linear regression results

What 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})

```
##
## 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
```