Linear Regression

Slope interpretation and significance

Jenine Harris
Brown School

foundations in public health
biostatistics

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

Slope interpretation and significance (b, p-value, CI)

- In addition to writing the regression model, there are three things to interpret for reporting the results of a linear regression analysis: model fit, model significance, and slope value and significance.

Interpreting the value of the slope

Using the regression model, the distance to syringe program would be predicted to be:

- 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.

Interpreting the statistical significance of the slope

- The output for the linear model included a p-value for the slope and a p-value for the intercept.
- The statistical significance of the slope in linear regression is tested using a **Wald test**, which is like a one-sample t-test where the hypothesized value of the slope is 0.
- To get the p-value from the regression model of distance to syringe program, the slope of 7.82 was compared to a hypothesized value of 0 using the Wald test.
- The null hypothesis was: *The slope of the line is equal to 0*.

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 to get t 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:
  - $t = \frac{7.8190 - 0}{.7734} = 10.1099$
- The t-statistic can also be found in the model output.

```
## 
## 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|)
```

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

- 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.
- The null hypothesis is rejected in favor of the alternate hypothesis that the slope is not equal to zero.
- This is often reported as the slope being statistically significantly different from 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 for slopes using the standard error of the slope from the regression output and a z-score of 1.96.

- Or, use the `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
```

- The output for `confint()` gives the confidence interval for the intercept and the slope; the intercept is typically ignored :-(

- 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

Using the model to make predictions

- **Predicted values** are the values of y predicted by the model for a given value of x.
- The regression line is essentially the line made of the predicted values based on the regression model.
- Predicted values of y are called y-hat and denoted $\hat{y}$.
- The R function, `predict()`, can be used to find the predicted values for all observations, or for a specific value of the independent variable.
- For example, for a county with 10% uninsured, `predict()` can be used to get the predicted value and the confidence interval around it for distance to nearest syringe program.
- The `predict()` function takes three arguments.
  - First, the `object =` argument takes the linear regression model object, which the team had named `dist.by.unins`.
  - The second argument is the name of a data frame with the observed value(s) of x in it.
  - Finally, the `interval =` argument would use "confidence" to get the confidence interval around each prediction. Nancy wrote the code and ran it for the team.

Interpreting predicted values

- The predicted distance to a syringe program from a county with with 10% of people uninsured is 90.67 miles with a confidence interval for the prediction (sometimes called a prediction interval) of 82.42 to 98.92 miles.

- The confidence interval shows where the population value of the statistic likely lies.

- In this case, the likely true distance to a syringe program from a county where 10% of people are uninsured is between 82.42 and 98.92 miles.

More predicted values

- The `predict()` function can predict $\hat{y}$ for all the observed values of x in the data.

- These predicted values can help determine how well the model fits the data.

```r
# use predict to find predicted value for all observed x
pred.dist.ssp.all <- predict(object = dist.by.unins,
                             interval="confidence")

# print out the first six predicted values and CI
head(x = pred.dist.ssp.all)
```

```
##         fit       lwr       upr
## 1  52.35653  39.20985  65.50321
## 2 157.13064 144.92001 169.34128
## 3 136.80134 127.37402 146.22865
## 4 109.43496 101.87889 116.99103
## 5 157.91254 145.58211 170.24297
## 6 128.20048 119.66871 136.73224
```