Populations & Samples

PSY 3307

Jonathan A. Pedroza, PhD

Cal Poly Pomona

2022-02-15

Terms

- sample
  - Xbar is the sample mean
  - S or SD is the sample standard deviation
- population
  - mu is the population mean
  - sigma is the population standard deviation

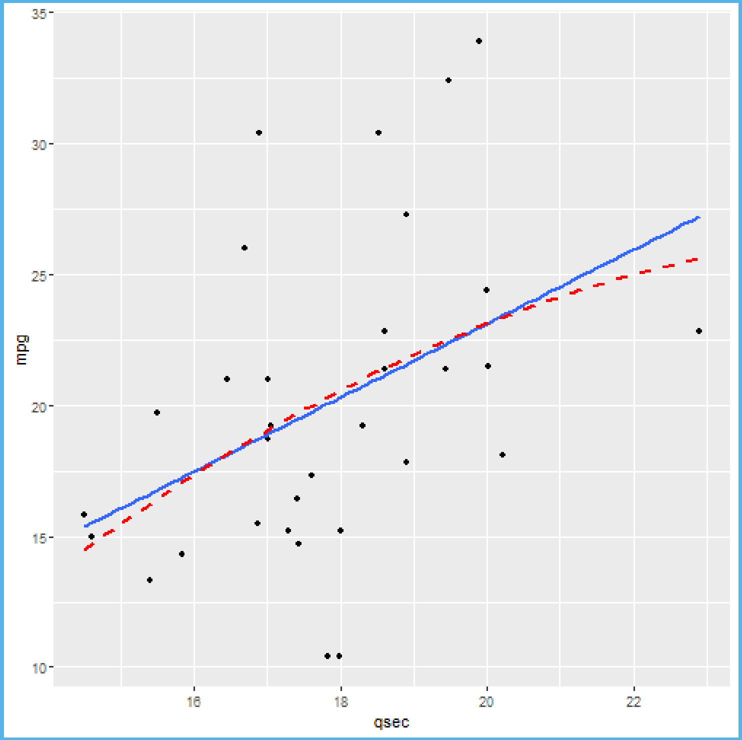
SPINE of Statistics

- **Standard Error**
- **Parameters**
- **Interval estimates**
- **Null hypothesis significance testing**
- **Estimation**

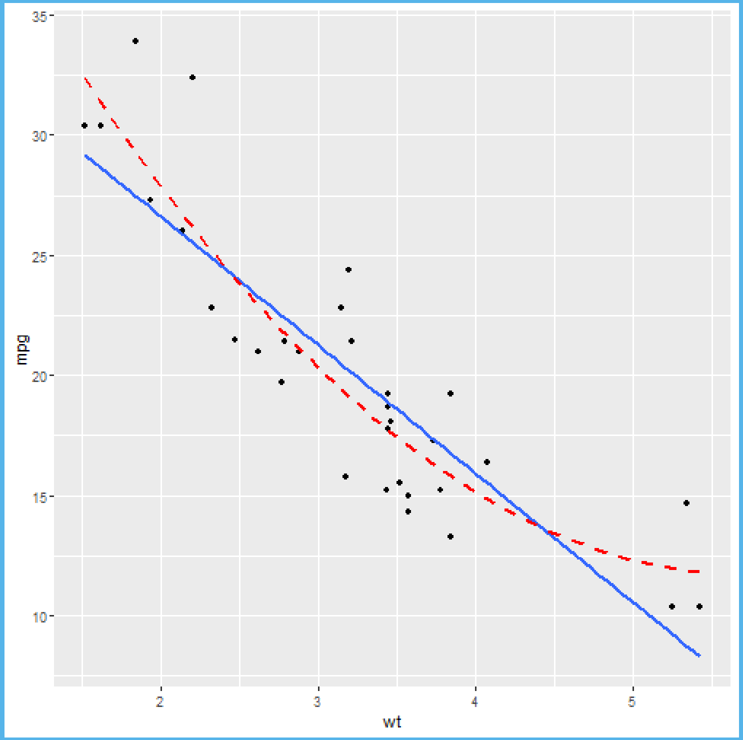
Statistical Models

- anything where you are testing predictions between IVs and DV
  - not all models are good models
- **fit** is the degree to which a statistical model represents the data
  - goodness of fit is a common measurement in statistical models (especially higher order models)
  - we'll talk aobut using the sum of squared errors as a measure of fit
- all models have some sort of fit

$$outcome_i = (model) + error_i$$



35
30
25
20
15
10
mpg
16
18
20
22
qsec



35
30
25
20
15
10
mpg
2
3
4
5
wt

Populations & Samples

- **population** is the entire number of entities
  - everyone you are interested in studying
- **sample** a smaller subset of the population that you infer things about
  - the bigger the sample the more it is like the population
  - the sample should be *representative* of the population

Populations & Samples

- CPP students
- CPP Psychology Students
- students from each CSU
- males from CPP
- PSY 3307 students

P is for Parameters

- **parameters** are a part of each statistical model
    - parameters are not necessarily measured and are usually constraints to represent some truth about the relationship between IV and DV
- parameters are symbols used
- JP Note: parameters can also mean values for any equations using the population only
    - statistics = sample
    - parameters = population

P is for Parameters

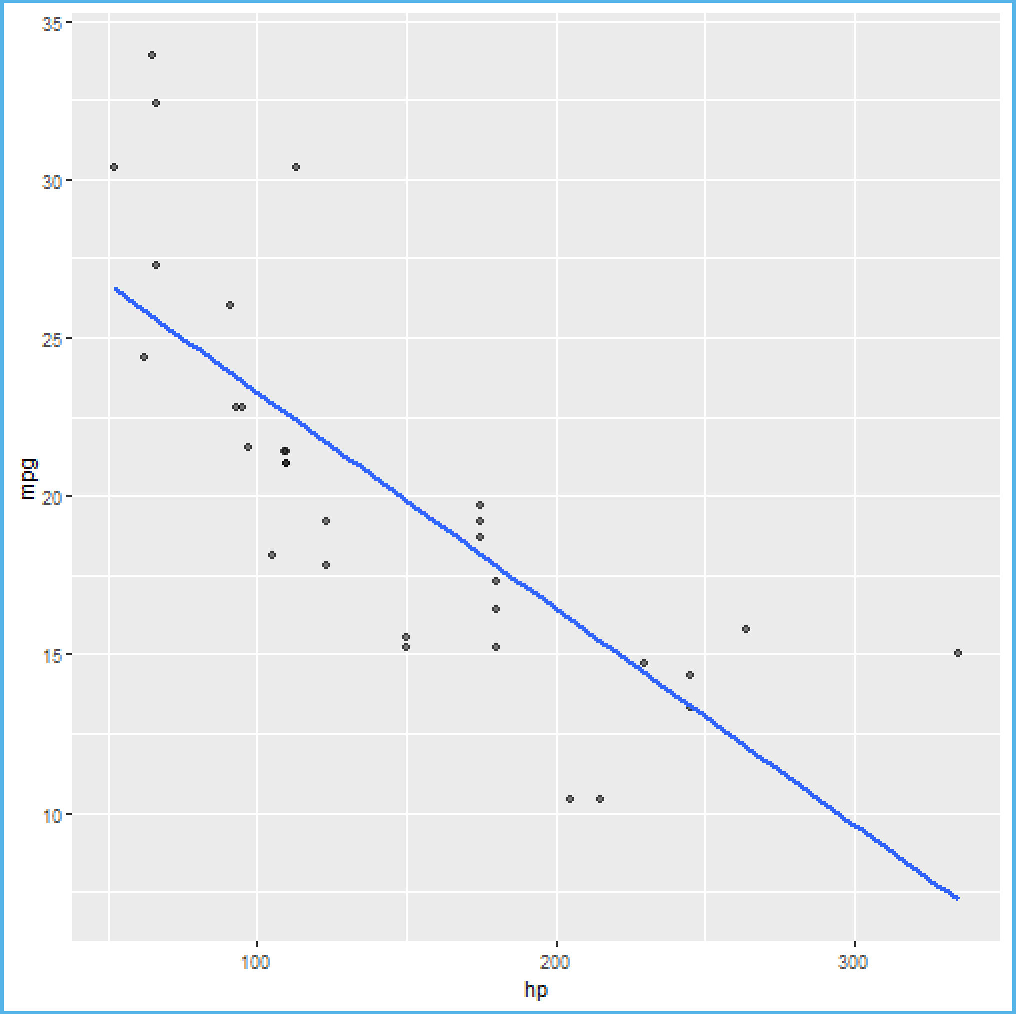
- when writing out models, we tend to use the first definition of parameters
- to work out the models below, we estimate the parameters (the values of the b)
    - we'll talk about this later when conducting statistics

$$outcome_i = (model) + error_i$$

$$outcome_i = (b_0) + error_i$$

$$outcome_i = (b_0 + b_1X_i) + error_i$$

$$outcome_i = (b_0 + b_1X_{1i} + +b_2X_{2i}) + error_i$$


35
30
25
20
15
10
mpg
100
200
300
hp

P is for Parameters

- the values that we are actually calculating are estimates
    - because we are using the sample to estimate what a relationship looks like in a population, we refer to them as **parameter estimates**
- this is in hopes that we are seeing that the sample is representative of the population

Mean as a Statistical Model

- b0 is commonly referred to as the intercept
    - it really is just the average value for the outcome

Mean as a Statistical Model

```
(10 + 8 + 4 + 7)/4
```

```
## [1] 7.25
```

$$outcome_i = (b_0) + error_i$$

- b0 is the mean of the outcome; here it refers to teacher eval scores
    - we give estimates of our data `hats` to show that we are *estimating* the data, and because they *could, in theory not be true*

$$outcome_i = (\hat{b}_0) + error_i$$

```r
data_df <- data.frame(eval_values = c(10, 8, 4, 7),
                      mean = c(7.25, 7.25, 7.25, 7.25),
                      deviance = c('', '', '', ''),
                      dev_squared = c('', '', '', ''))
data_df
```

```
##   eval_values mean deviance dev_squared
## 1          10 7.25
## 2           8 7.25
## 3           4 7.25
## 4           7 7.25
```

Assessing Model Fit - Sums of Squares & Variance

- let's use what we have learned with a different lens to see how it relates to modeling
  - **error** is another word for deviance

$$deviance = outcome_i - model_i$$

$$outcome_{lecture1} = (\hat{b}_0) + error_{lecture1}$$

- we can use the participant score = the mean estimate of the outcome and the error (deviance)

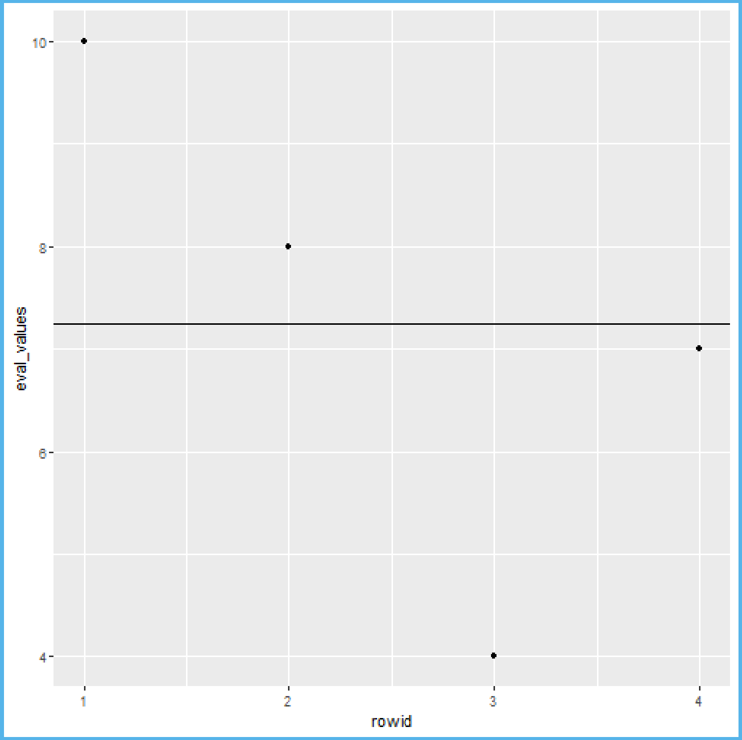
$$10 = 7.25 + error_{lecture1}$$

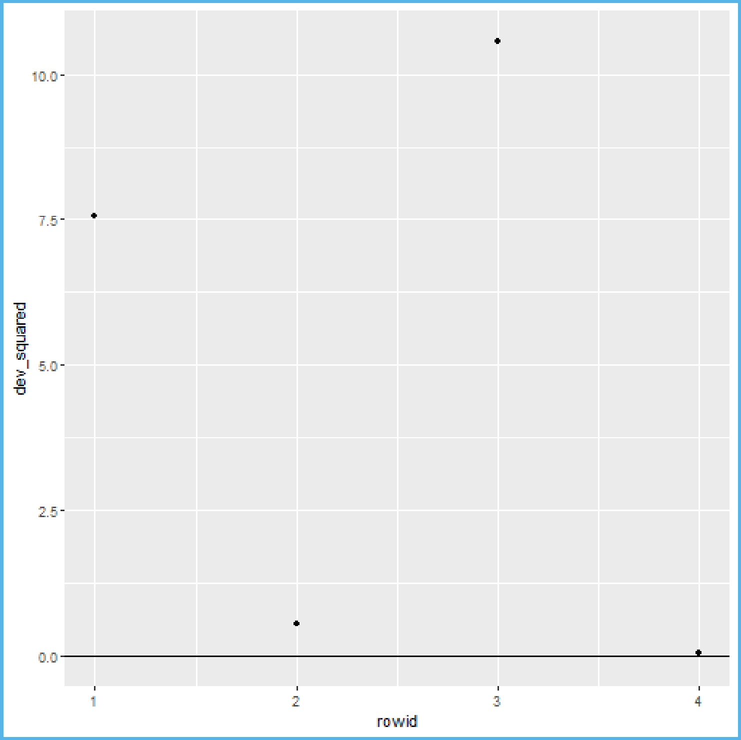
$$10 - 7.25 = error_{lecture1}$$

$$2.75 = error_{lecture1}$$

```r
data_df
```

```
##   eval_values mean deviance dev_squared
## 1          10 7.25     2.75      7.5625
## 2           8 7.25     0.75      0.5625
## 3           4 7.25    -3.25     10.5625
## 4           7 7.25    -0.25      0.0625
```


eval_values
10
8
6
4
1
2
3
4
rowid


dev_squared
10.0
7.5
5.0
2.5
0.0
1
2
3
4
rowid

Assessing Model Fit - Sums of Squares & Variance

- from the formulas above, we know the fit for the first evaluation
- now we can see the fit overall

$$total\ error = sum\ of\ errors = \sum_{i=1}^{n}(outcome_i - model_i)$$

$$\sum_{i=1}^{n}(outcome_i - model_i) = \sum_{i=1}^{n}(X_i - \overline{X})$$

$$2.75 + .75 + (-3.25) + (-.25) = 0$$

$$sum\ of\ squared\ errors(sum\ of\ squares) = \sum_{i=1}^{n}(outcome_i - model_i)^2$$

$$\sum_{i=1}^{n}(outcome_i - model_i)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$7.56 + .56 + 10.56 + .06 = 18.75$$

$$total\ error = \sum_{i=1}^{n}(observed_i - model_i)^2$$

$$mean\ squared\ error = \frac{SS}{df} = \frac{\sum_{i=1}^{n}(outcome_i - model_i)^2}{N-1}$$

$$df = N - 1$$

$$18.75/3 = 6.25$$

Assessing Model Fit - Sums of Squares & Variance

- to compute average error, we divide sum of squares by the number of values (N), expect we are focusing on the population and how to estimate it
    - **degrees of freedom** is the number of scores used to compute teh total adjusted for the fact that we're estimating the population value
- **mean squared error** is also known as the variance
    - variance is a special case that can be applied to more complex models
    - model fit can be assessed with sum of squared errors or mean squared errors
- mean squared error is often seen as how far predicted values (in models) are away from the participants' actual/raw values

E is for Estimating Parameters

- if we wanted to create predictions, we could then include scores to see how well they would fit with the data
  - put numbers in for the squared error that you'd like to test

$$outcome_i = (\hat{b}_0) + error_i$$

- we can rearrange this formula to get the error

$$error_i = outcome_i - (\hat{b}_0)$$

- from this we then test to see the difference between every participant's score and the new value you included (we'll cover this in the activity)
  - then you can add up the squared error values
  - can compare between predictions b1 to b2 values
  - we can then compare the b1, b2, and b0 error values to see which is the best fit

$$(x_i - b_1)^2$$

$$(x_i - b_2)^2$$

- the process of minimizing the sum of squared errors/sum of squares is know as the **method of least squares** or **ordinary least squares (OLS)**