Bias

What is Bias?

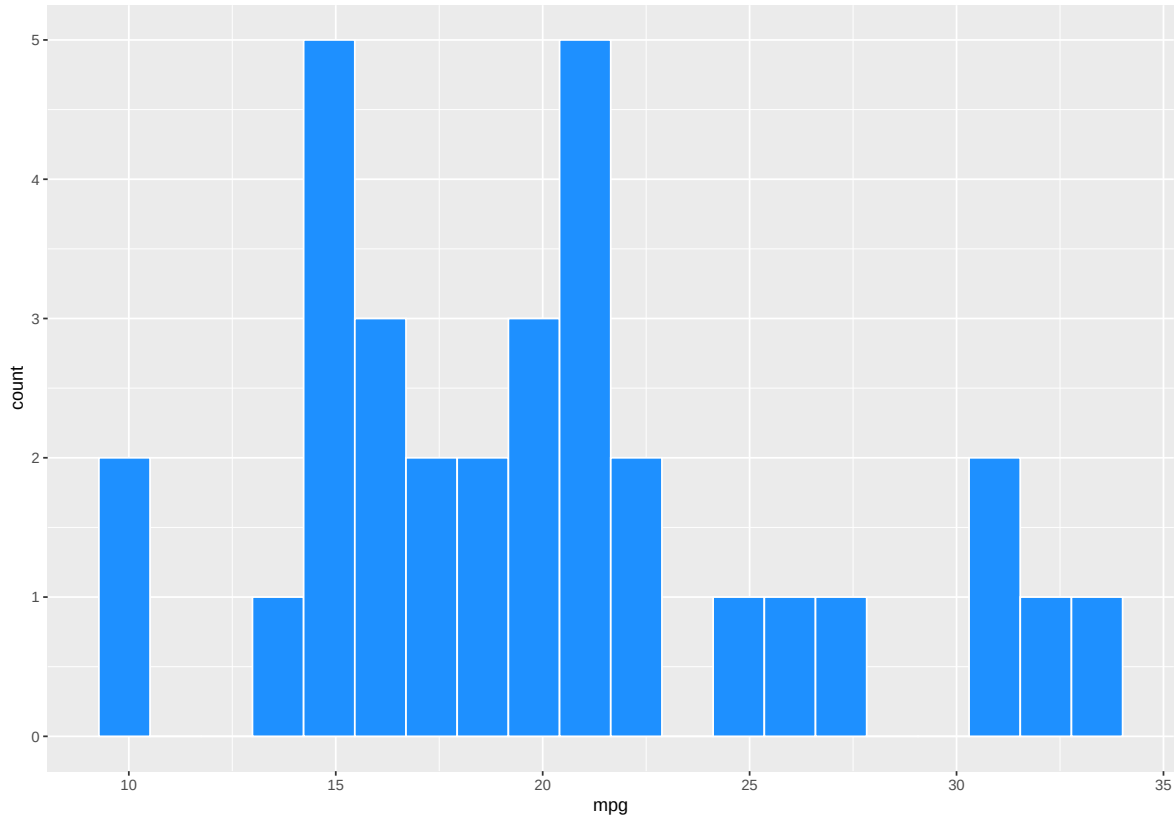
- first biased estimator we've focused on was using (N - 1) instead of N for variance and standard deviation
- three ways that statistical bias can enter the modeling process is from
  - things biasing the parameter estimates
  - things biasing SEs and CIs
  - things biasing test statistics and p-values

Outliers

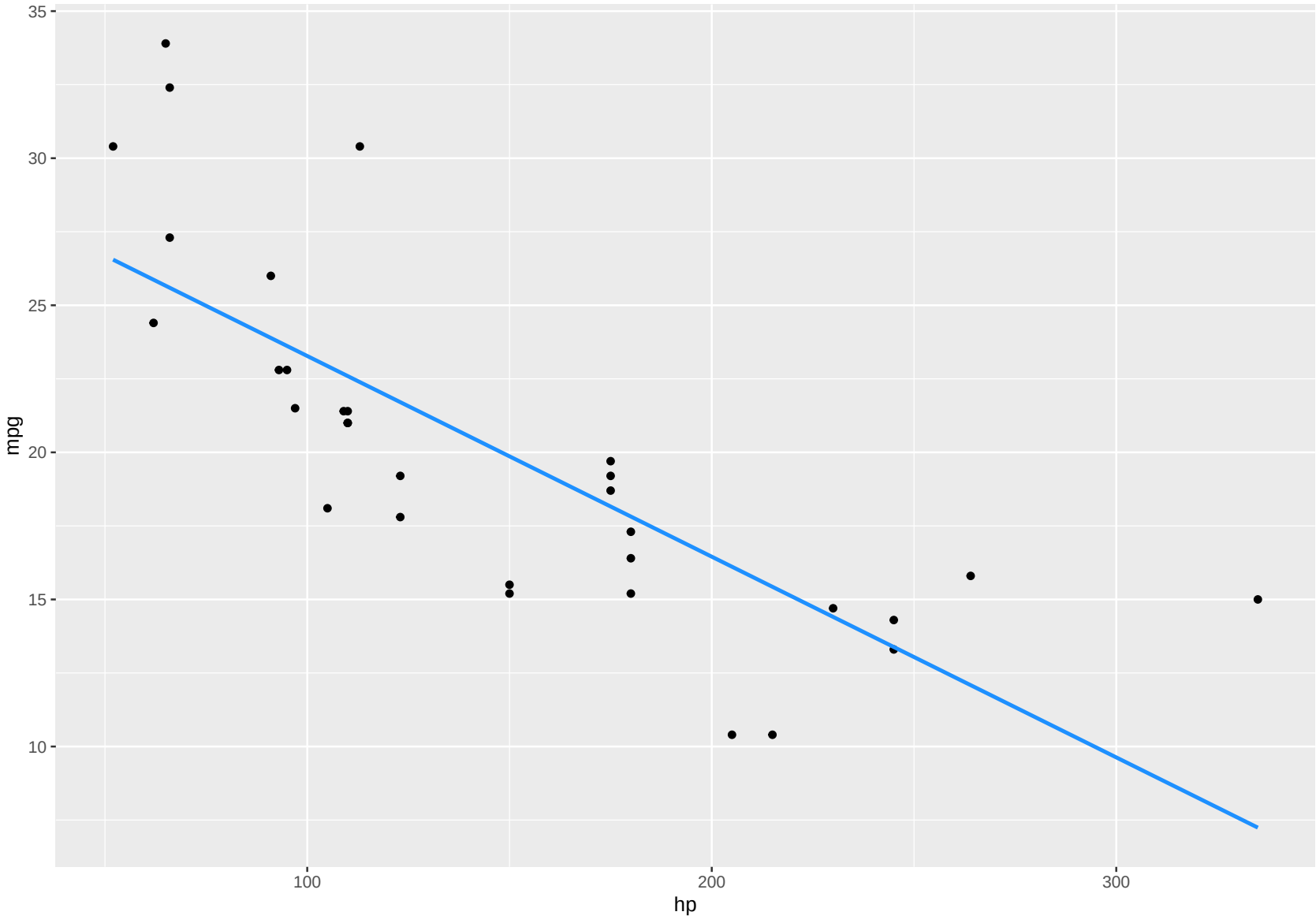
- an **outlier** is a score very different from the rest of the data
- when looking at bias in boxplots on SPSS, it will provide you with information on outliers and *influential* outliers
- researchers say that participants outside +-3SD are considered outliers
- outliers bias parameter estimates, but they have a greater impact on the error associated with teh estimate

Outliers

```
[1] 20.09062

[1] 6.026948

[1] 38.17147

[1] 2.009781
```



Outliers



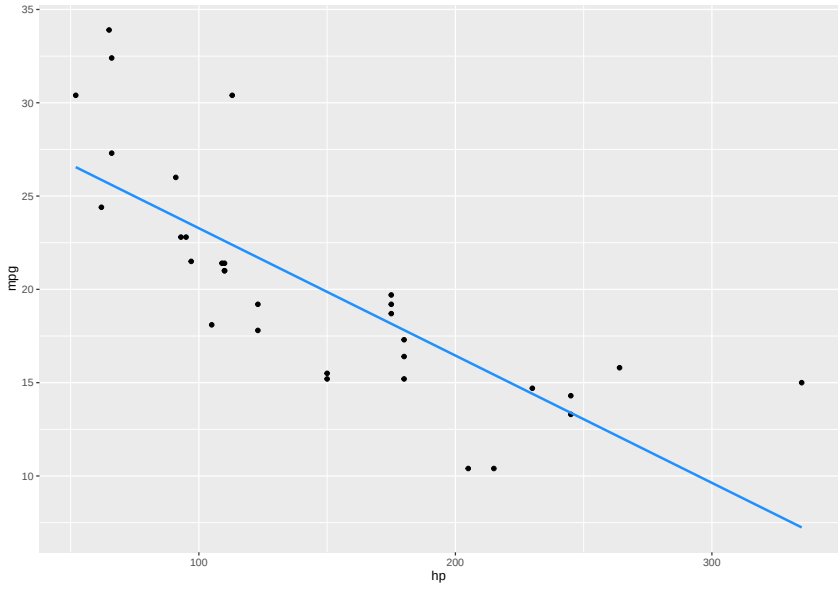
Assumptions

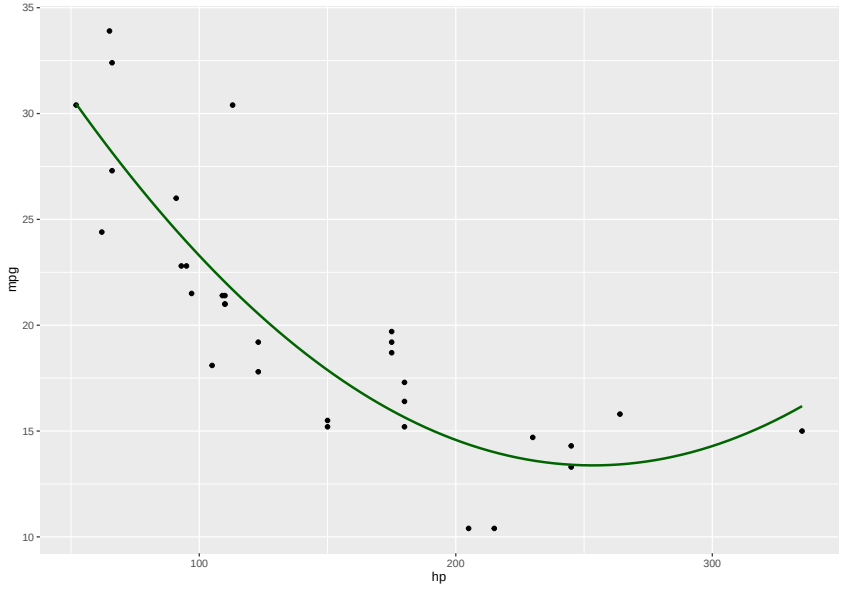
- Good researchers will always check the following assumptions before conducting inferential statistics
  - JP: I will warn you right now. **There are no 100% correct answers to these questions**.
    - Statistics is about to get ugly
- If any of the assumptions are violated, your test statistic and corresponding p-value may not be correct
- each model you run will have assumptions that need to be checked
  - we will talk about these every time we discuss a new test statistic
- The main assumptions are
  - additivity and linearity
  - normality
  - homoscedasticity/homogeneity of variance
  - independence

Additivity & Linearity

- many of the models we'll cover are based on linear relationships
  - even those that are comparing conditions/groups (e.g., t-tests)
- This assumption is simply based on whether there is a line that fits the data ***well***
  - very subjective
- Every other assumption falls on linearity because if you think you are looking at a linear relationship, you want to test it with a linear test

Additivity & Linearity





Normally Distributed

- many think that your data must be normally distributed
  - not the case
- parameter estimates
  - when we create a model, we have what we predicted in the model and the rest (error/deviance/residual)
  - what should be normally distributed are the residuals for the model
- confidence intervals
  - for confidence intervals to be accurate, your estimate (what you're predicting in your model) should have a normal sampling distribution
- null-hypothesis significance testing
  - if your parameter estimates are normal, then the test-statistic distribution should also be normally distributed, which would also make your alpha/p-value accurate

Central Limit Theorem

- populations should be normally distributed
  - if you have enough observations, your population's distribution will be normal

Assumption of Normality

- confidence intervals don't need to be worried about too much for normality because if our parameter estimate is normal then so are our confidence intervals if the sample is large enough
- significant tests of models will be accurate if the sample is large enough thanks to the central limit theorem
- in estimates of model parameters, residuals of the population must be normally distributed

Homoscedasticity/Homogeneity of Variance

- homoscedasticity affects parameters and NHST
  - **parameters** using Ordinary Least Squares (OLS) estimation, we get the best model fit when variance of the DV is equal across different values of the IV
    - variance is equal across our groups
- **null hypothesis significance testing** assume variance of the outcome if equal across different values of the IV
  - some tests (like t-tests) can account for the variance not being equal in groups

What is Homosceasticity/Homogeneity of Variance?

- **homoscedasticity** in correlational studies states that variance of the DV scores should be stable at all points of the IV
  - **homogeneity of variance** means that all the values should be fairly grouped together across each group/all points of the IV
- **heteroscedasticity** is when the variance of the DV scores is different at different points of the IV or different amount groups
  - if we were to use error bars (we'll talk about this later), then we can see that the points would be spread out more across each group, also called **heterogeneity of variance**

When does it matter?

- the best fitting model will have homoscedasticity
  - OLS models will still show model fit, but it won't be the best it could be
  - other options can fit the model better like **weighted least squares**, which is when each participant is weighted by a function of its variance (adjusts each participant)
- heteroscedasticity can result in biased standard errors, which then affects confidence intervals, p values, and parameter estimates

Independence

- simply put, this means that your errors in your model are not related to one another
- JP: I like to work with spatial data, and that tends to have problems with independence
  - Ex: if I am focusing on different counties, how different are LA county and Orange county

Spotting Outliers

- Visuals are the best method to spot outliers easily
  - histograms
  - boxplots
- boxplots are extremely helpful in SPSS because they show outliers and influential outliers
- JP: I tend to run my analyses with both outliers included and dropped from the model
  - just because they are an extreme case does not mean they are not a valid case

Spotting Normality

- **p-p plot** (probability-probability plot) and **q-q plot** (quantile-quantile plot)
  - p-p plots show the cumulative probability of a variable against the cumulative probability of a particular distribution
  - q-q plots show quantiles as dots rather than individual points/participants
    - essentially, what we want is for our dots to follow the line as close as possible for both

Using Numbers for Normality

- we can use measures of central tendency and measures of variability to get a better understanding of our data
  - we can look for minimum and maximum values (outliers)
  - we can see if the SD is large, then there is a lot of dispersion/spread/distance between points
- JP: the visuals are just much easier to gather this information from
  - afterward, you can check the numbers
- There are also tests to show normality in the data
  - **Kolmogorov-Smirnov** and **Shapiro-Wilk** tests compare the scores in a sample to a normally distributed set of scores with the same mean and SD
  - if the test is significant, your data is not normal
    - not really all that useful because large samples (N $\sim$ 150+) can make them be statistically significant
- I'll also cover how to get descriptive statistics for groups
  - not entirely relevant because our test statistics will give us that information anyway

Spotting Linearity

- in SPSS, spotting problems with linearity and homoscedasticity are related to using the residuals
  - in SPSS, its sometimes referred to as the *zpred vs zresid.* if you transform your variables into z-scores
- JP: I like to look at the raw values in a scatterplot as well as the residuals
- points should always look like no general pattern is forming
  - funnel/fanning= BAD
  - non-linear pattern = BAD
  - fanning & non linear = BAD

Spotting Heteroscedasticity

- **Levene's test** is a specific test to address whether the variance in your groups (categorical IVs) is equal (similar enough) to one another
  - Ex: Are hours of sleep similar in males and females
- if this test is statistically significant then the assumption can be made that the groups are not equal and you have heteroscedasticity
- **Hartley's Fmax** or **variance ratio** is the ratio of the variances between the groups with the largest and smallest variances
- reporting Levene's test is written as $F(\text{df1}, \text{df2}) = \text{F value}$, p *value*
  - $F(1, 124) = 3.17$, $p = .03$

Reducing Bias

- trim the data
  - delete cases
- winsorizing
  - bring those same cases in a little with the most extreme **okay** value
- apply robust estimation method
  - bootstrapping
- transform data
  - apply mathematical function to scores to correct problem (we'll cover one)