432 Class 13

Thomas E. Love, Ph.D.

2024-02-27

## Today’s Agenda

1. Evaluation of a Study through Retrospective Design
   * Gelman and Carlin: Type S and Type M errors
   * The retrodesign() function
2. Robust Linear Regression
   * The **crimestat** data
   * Using Huber weights, then using bisquare weights
   * Quantile Regression on the Median

## Today’s R Setup

knitr::opts\_chunk$set(comment = NA)  
  
library(janitor)  
library(broom)  
library(gt)  
library(MASS) ## fitting robust linear models  
library(quantreg) ## fitting quantile regressions  
library(tidyverse)  
  
theme\_set(theme\_bw())

# Evaluation through Retrospective Design

## On “post hoc” power analysis

Suppose you read a study, and the result appears to not meet the standard of statistical significance that you wish it did. The idea is to show that a “non-significant” hypothesis test failed to achieve significance because it wasn’t powerful enough.

* Knowing what we have learned from this new study, what sort of power did the study have to detect the effect that we saw? (a *post hoc* power calculation)
* What sort of sample size might we use in a new study? (maybe useful?)
* Was this study dead in the water before we did it?

## To be clear, Post hoc power calculations are not useful

[Post Hoc Power Calculations Are Not Useful](https://library.virginia.edu/data/articles/post-hoc-power-calculations-are-not-useful) from University of Virginia’s Research Data Services cites these papers.

* [Althouse A (2021)](https://pubmed.ncbi.nlm.nih.gov/32814615/)
* [Goodman SN Berlin JA (1994)](https://pubmed.ncbi.nlm.nih.gov/8017747/)
* [Hoenig JM Heisey DM (2001)](https://www.tandfonline.com/doi/abs/10.1198/000313001300339897)

This last piece is from *The American Statistician* and is entitled “The abuse of power: The pervasive fallacy of power calculations for data analysis.” Much wisdom there.

## Andrew Gelman’s blog

Blog is [Statistical Modeling, Causal Inference, and Social Science](https://statmodeling.stat.columbia.edu/)

* [Statistics is like basketball, or knitting](https://statmodeling.stat.columbia.edu/2016/03/11/statistics-is-like-basketball-or-knitting/) from 2016-03-11
* The post discusses [this article](https://pubmed.ncbi.nlm.nih.gov/26510427/), shown on the next slide.

Headline Finding: A sample of ~500 men from America and India shows a significant relationship between sexist views and the presence of facial hair.



## Facial Hair and Sexist Attitudes

Excerpt 1:

Since a linear relationship has been found between facial hair thickness and perceived masculinity . . . we explored the relationship between facial hair thickness and sexism. . . . Pearson’s correlation found no significant relationships between facial hair thickness and hostile or benevolent sexism, education, age, sexual orientation, or relationship status.

## Facial Hair and Sexist Attitudes

Excerpt 2:

We conducted pairwise comparisons between clean-shaven men and each facial hair style on hostile and benevolent sexism scores. . . . For the purpose of further analyses, participants were classified as either clean-shaven or having facial hair based on their self- reported facial hair style . . . There was a significant Facial Hair Status by Sexism Type interaction . . .

## Gelman, 2016-03-11 Blog

So their headline finding appeared only because, after their first analysis failed, they shook and shook the data until they found something statistically significant.

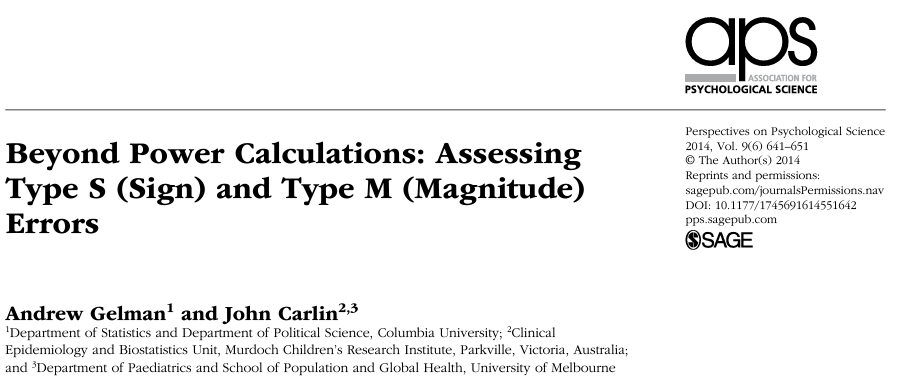
* All credit to the researchers for admitting that they did this, but poor practice of them to present their result in the abstract to their paper without making this clear, and too bad that the journal got suckered into publishing this.

## How should we react to this?

Gelman:

* Statisticians such as myself should recognize that the point of criticizing a study is, in general, to shed light on statistical errors, maybe with the hope of reforming future statistical education.
* Researchers and policymakers should not just trust what they read in published journals.

## Gelman and Carlin (2014) [PDF Link](http://www.stat.columbia.edu/~gelman/research/published/retropower_final.pdf)



## Specifying effect sizes - how? (1/2)

When doing a power calculation, how do people specify an effect size of interest? Two main approaches…

1. **Empirical**: assuming an effect size equal to the estimate from a previous study or from the data at hand (if performed retrospectively).
   * generally based on small samples
   * when preliminary results look interesting, they are more likely biased towards unrealistically large effects

## Specifying effect sizes - how? (2/2)

When doing a power calculation, how do people specify an effect size of interest? Two main approaches…

1. **On the basis of goals**: assuming an effect size deemed to be substantively important or more specifically the minimum effect that would be substantively important.
   * Can also lead to specifying effect sizes that are larger than what is likely to be the true effect.

Both approaches lead to performing studies that are too small or misinterpretation of findings after completion.

## What is a design analysis?

* The idea of a **design analysis** is to improve the design and evaluation of research, when you want to summarize your inference through concepts related to statistical significance.
* Type 1 and Type 2 errors are tricky concepts and aren’t easy to describe before data are collected, and are very difficult to use well after data are collected.

## Why a design analysis?

* The previous slide’s problems are made worse when you have:
  + Noisy studies, where the signal may be overwhelmed,
  + Small Sample Sizes
  + No pre-registered (prior to data gathering) specifications for analysis
* Top statisticians avoid “post hoc power analysis”…
  + Why? It’s usually crummy.

## Why not post hoc power analysis?

You collected data and analyzed the results. Now you want to do an after data gathering (post hoc) power analysis.

1. What will you use as your “true” effect size?
   * Often, point estimate from data - results very misleading - power is usually seriously overestimated when computed on the basis of “significant” results.
   * Much better (but rarer) to identify plausible effect sizes based on external information rather than on your sparkling new result.

## Why not post hoc power analysis?

1. What are you trying to do? (too often)
   * get researcher off the hook (I didn’t get p < 0.05 because I had low power - an alibi to explain away non-significant findings) or
   * encourage overconfidence in the finding.

A broader notion of design, though, can be useful before and after data are gathered.

## Broader Design Ideas

Gelman and Carlin recommend design calculations to estimate

1. Type S (sign) error - the probability of an estimate being in the wrong direction, and
2. Type M (magnitude) error, or exaggeration ratio - the factor by which the magnitude of an effect might be overestimated.

## The Value of Type S and Type M error

These ideas can (and should) have value **both** before data collection/analysis and afterwards (especially when an apparently strong and significant effect is found.)

* The big challenge remains identifying plausible effect sizes based on external information. Crucial to base our design analysis on an external estimate.

## Building Blocks (1/2)

You perform a study that yields estimate *d* with standard error *s*. Think of *d* as an estimated mean difference, for example.

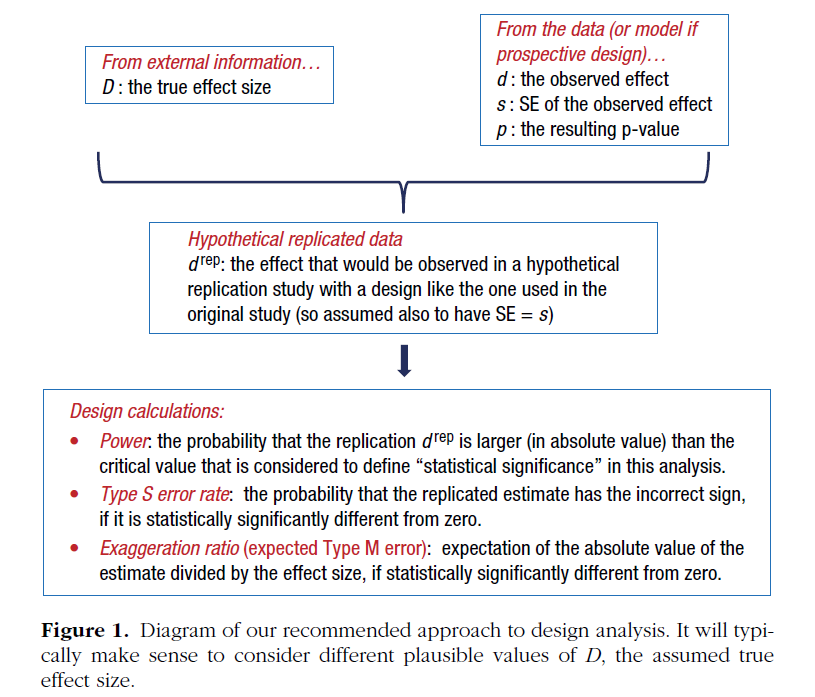
* Looks significant if , which roughly corresponds to *p* < 0.05. Inconclusive otherwise.
* Now, consider a true effect size *D* (the value that *d* would take if you had an enormous sample)
* *D* is hypothesized based on *external* information (Other available data, Literature review, sometimes Modeling, etc.)

## Building Blocks (2/2)

You perform a study that yields estimate *d* with standard error *s*. Think of *d* as an estimated mean difference, for example.

* Define as the estimate that would be observed in a hypothetical replication study with a design identical to our original study.

## Design Analysis (Gelman and Carlin)



## Retrodesign function (R code coming)

Inputs to the function:

* D, the hypothesized true effect size
* s, the standard error of the estimate
* alpha, the statistical significance threshold (default 0.05)
* df, the degrees of freedom (default assumption: infinite)

## Retrodesign function (R code coming)

Output:

* the power
* the Type S error rate
* the exaggeration ratio

## Retrodesign function

* built by Gelman and Carlin

retrodesign <- function(D, s, alpha=.05, df=Inf, n.sims=10000)  
 {  
 z <- qt(1-alpha/2, df)  
 p.hi <- 1 - pt(z-D/s, df)  
 p.lo <- pt(-z-D/s, df)  
 power <- p.hi + p.lo  
 typeS <- p.lo/power  
 estimate <- D + s\*rt(n.sims,df)  
 significant <- abs(estimate) > s\*z  
 exaggeration <- mean(abs(estimate)[significant])/D  
 return(list(power=power, typeS=typeS,   
 exaggeration=exaggeration))  
}

## What if we have a beautiful, unbiased study?

Suppose the true effect is 2.8 standard errors away from zero, in a study built to have 80% power for that effect with 95% confidence.

set.seed(20240227)  
retrodesign(D = 28, s = 10, alpha = 0.05)

$power  
[1] 0.7995569  
  
$typeS  
[1] 1.210843e-06  
  
$exaggeration  
[1] 1.128872

## A beautiful, unbiased study?

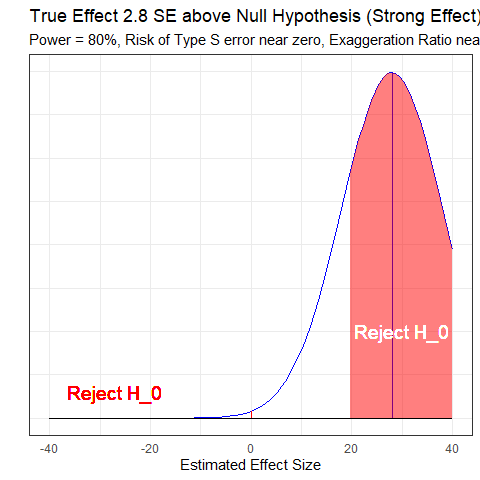
| power | typeS | exaggeration |
| --- | --- | --- |
| 0.79956 |  | 1.13 |

* With the power this high (80%), we have a type S error rate of and an expected exaggeration factor of 1.13.
* Nothing to worry about with either direction of a statistically significant estimate and the overestimation of the magnitude of the effect will be small.
* What does this look like?

## 80% power; large effect (2.8 SE above )

x <- seq(-40, 40, length = 100)  
hx0 <- dnorm(x, mean = 0, sd = 10)  
hx3 <- dnorm(x, mean = 3, sd = 10)  
hx12 <- dnorm(x, mean = 12, sd = 10)  
hx28 <- dnorm(x, mean = 28, sd = 10)  
hx2215 <- dnorm(x, mean = 22.15, sd = 10)  
dat <- data.frame(x, hx0, hx3, hx12, hx28, hx2215)

ggplot(dat, aes(x, hx28)) +  
 geom\_line(col = "blue") +  
 geom\_segment(aes(x = 0, xend = 0, y = 0, yend = dnorm(0, mean = 28, sd = 10)), col = "red") +  
 geom\_segment(aes(x = 28, xend = 28, y = 0, yend = dnorm(28, mean = 28, sd = 10)), col = "blue") +  
 geom\_segment(aes(x = -40, xend = 40, y = 0, yend = 0)) +  
 geom\_ribbon(data = subset(dat, x > 19.6), aes(ymax = hx28), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_ribbon(data = subset(dat, x < -19.6), aes(ymax = hx28), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_text(x = 30, y = 0.01, label = "Reject H\_0", col = "white", size = 5) +  
 geom\_text(x = -27, y = 0.003, label = "Reject H\_0", col = "red", size = 5) +  
 theme(axis.ticks = element\_blank(), axis.text.y = element\_blank()) +  
 labs(x = "Estimated Effect Size", y = "", title = "True Effect 2.8 SE above Null Hypothesis (Strong Effect)", subtitle = "Power = 80%, Risk of Type S error near zero, Exaggeration Ratio near 1")



## retrodesign for Zero Effect

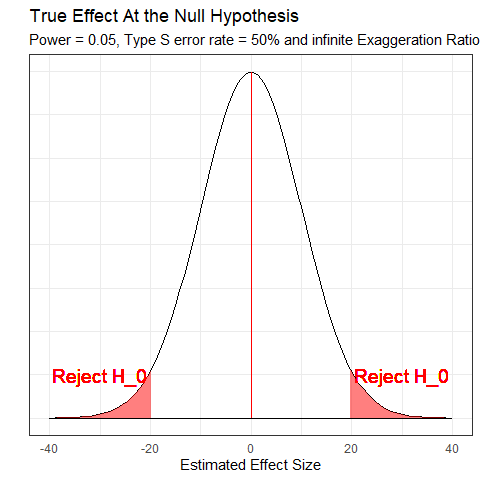
set.seed(202402272)  
retrodesign(D = 0, s = 10)

$power  
[1] 0.05  
  
$typeS  
[1] 0.5  
  
$exaggeration  
[1] Inf

* Power = 0.05, Pr(Type S error) = 0.5, Exaggeration Ratio is infinite.

## Power, Type S and Type M Errors: Zero Effect

ggplot(dat, aes(x, hx0)) +  
 geom\_line() +  
 geom\_segment(aes(x = 0, xend = 0, y = 0, yend = dnorm(0, mean = 0, sd = 10)), col = "red") +  
 geom\_segment(aes(x = -40, xend = 40, y = 0, yend = 0)) +  
 geom\_ribbon(data = subset(dat, x > 19.6), aes(ymax = hx0), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_ribbon(data = subset(dat, x < -19.6), aes(ymax = hx0), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_text(x = 30, y = 0.005, label = "Reject H\_0", col = "red", size = 5) +  
 geom\_text(x = -30, y = 0.005, label = "Reject H\_0", col = "red", size = 5) +  
 theme(axis.ticks = element\_blank(), axis.text.y = element\_blank()) +  
 labs(x = "Estimated Effect Size", y = "", title = "True Effect At the Null Hypothesis", subtitle = "Power = 0.05, Type S error rate = 50% and infinite Exaggeration Ratio")



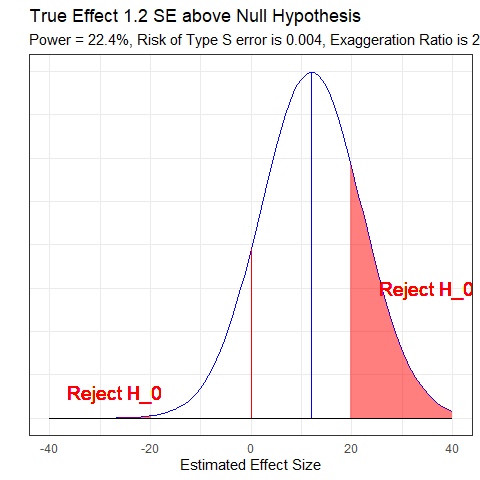
## Retrodesign for a true effect 1.2 SE above

set.seed(202402273)  
retrodesign(D = 12, s = 10)

$power  
[1] 0.224427  
  
$typeS  
[1] 0.003515367  
  
$exaggeration  
[1] 2.105828

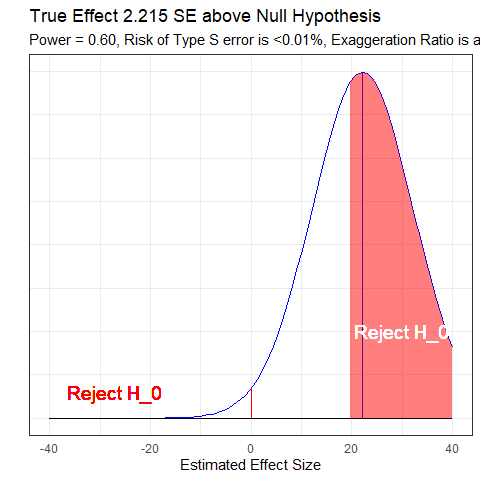
## What 22.4% power looks like…

ggplot(dat, aes(x, hx12)) +  
 geom\_line(col = "blue") +  
 geom\_segment(aes(x = 0, xend = 0, y = 0, yend = dnorm(0, mean = 12, sd = 10)), col = "red") +  
 geom\_segment(aes(x = 12, xend = 12, y = 0, yend = dnorm(12, mean = 12, sd = 10)), col = "blue") +  
 geom\_segment(aes(x = -40, xend = 40, y = 0, yend = 0)) +  
 geom\_ribbon(data = subset(dat, x > 19.6), aes(ymax = hx12), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_ribbon(data = subset(dat, x < -19.6), aes(ymax = hx12), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_text(x = 35, y = 0.015, label = "Reject H\_0", col = "red", size = 5) +  
 geom\_text(x = -27, y = 0.003, label = "Reject H\_0", col = "red", size = 5) +  
 theme(axis.ticks = element\_blank(), axis.text.y = element\_blank()) +  
 labs(x = "Estimated Effect Size", y = "", title = "True Effect 1.2 SE above Null Hypothesis", subtitle = "Power = 22.4%, Risk of Type S error is 0.004, Exaggeration Ratio is 2.12")

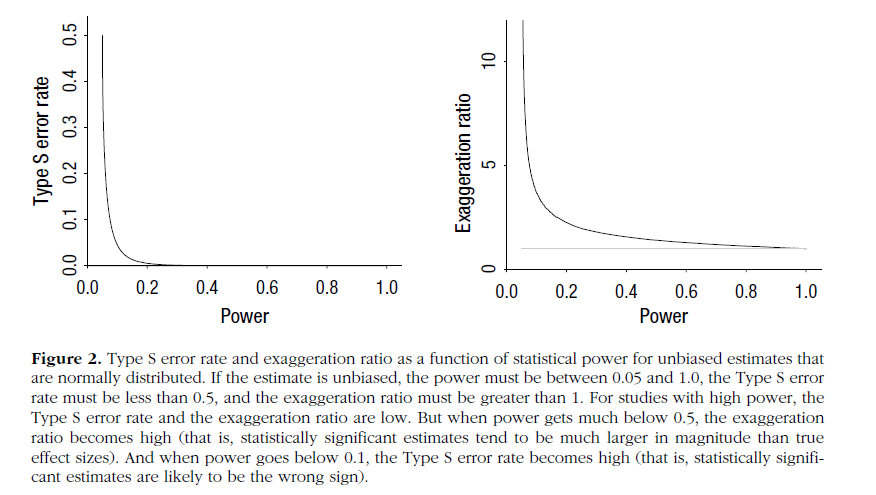


## What 60% Power Looks Like

ggplot(dat, aes(x, hx2215)) +  
 geom\_line(col = "blue") +  
 geom\_segment(aes(x = 0, xend = 0, y = 0, yend = dnorm(0, mean = 22.15, sd = 10)), col = "red") +  
 geom\_segment(aes(x = 22.15, xend = 22.15, y = 0, yend = dnorm(22.15, mean = 22.15, sd = 10)), col = "blue") +  
 geom\_segment(aes(x = -40, xend = 40, y = 0, yend = 0)) +  
 geom\_ribbon(data = subset(dat, x > 19.6), aes(ymax = hx2215), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_ribbon(data = subset(dat, x < -19.6), aes(ymax = hx2215), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_text(x = 30, y = 0.01, label = "Reject H\_0", col = "white", size = 5) +  
 geom\_text(x = -27, y = 0.003, label = "Reject H\_0", col = "red", size = 5) +  
 theme(axis.ticks = element\_blank(), axis.text.y = element\_blank()) +  
 labs(x = "Estimated Effect Size", y = "", title = "True Effect 2.215 SE above Null Hypothesis", subtitle = "Power = 0.60, Risk of Type S error is <0.01%, Exaggeration Ratio is about 1.3")



## Gelman & Carlin, Figure 2



## Example: Beauty and Sex Ratios

Kanazawa study of 2972 respondents from the National Longitudinal Study of Adolescent Health

* Each subject was assigned an attractiveness rating on a 1-5 scale and then, years later, had at least one child.
* Of the first-born children with parents in the most attractive category, 56% were girls, compared with 48% girls in the other groups.
* So the estimated difference was 8 percentage points with a reported *p* = 0.015
* Kanazawa stopped there, but Gelman and Carlin don’t.

## Beauty and Sex Ratios

We need to postulate an effect size, which will not be 8 percentage points. Instead, Gelman and colleagues hypothesized a range of true effect sizes using the scientific literature.

There is a large literature on variation in the sex ratio of human births, and the effects that have been found have been on the order of 1 percentage point (for example, the probability of a girl birth shifting from 48.5 percent to 49.5 percent).

## More from Gelman et al.

Variation attributable to factors such as race, parental age, birth order, maternal weight, partnership status and season of birth is estimated at from less than 0.3 percentage points to about 2 percentage points, with larger changes (as high as 3 percentage points) arising under economic conditions of poverty and famine.

(There are) reliable findings that male fetuses (and also male babies and adults) are more likely than females to die under adverse conditions.

## So, what is a reasonable effect size?

* Small observed differences in sex ratios in a multitude of studies of other issues (much more like 1 percentage point, tops)
* Noisiness of the subjective attractiveness rating (1-5) used in this particular study

So, Gelman and colleagues hypothesized three potential effect sizes (0.1, 0.3 and 1.0 percentage points) and under each effect size, considered what might happen in a study with sample size equal to Kanazawa’s study.

## How big is the standard error?

* From the reported estimate of 8 percentage points and p value of 0.015, the standard error of the difference is 3.29 percentage points.
  + If *p* value = 0.015 (two-sided), then Z score = qnorm(p = 0.015/2, lower.tail=FALSE) = 2.432
  + Z = estimate/SE, and if estimate = 8 and Z = 2.432, then SE = 8/2.432 = 3.29

## Retrodesign Results: Option 1

* Assume true difference D = 0.1 percentage point (probability of girl births differing by 0.1 percentage points, comparing attractive with unattractive parents).
* Standard error assumed to be 3.29, and = 0.05

set.seed(201803164)  
retrodesign(D = 0.1, s = 3.29, alpha = 0.05)

$power  
[1] 0.05010584  
  
$typeS  
[1] 0.4645306  
  
$exaggeration  
[1] 76.93614

## Option 1 Conclusions

Assuming the true difference is 0.1 means that probability of girl births differs by 0.1 percentage points, comparing attractive with unattractive parents.

If the estimate is statistically significant, then:

1. There is a 46% chance it will have the wrong sign (from the Type S error rate).

## Option 1 Conclusions

Assuming the true difference is 0.1 means that probability of girl births differs by 0.1 percentage points, comparing attractive with unattractive parents.

If the estimate is statistically significant, then:

1. The power is 5% and the Type S error rate of 46%. Multiplying those gives a 2.3% probability that we will find a statistically significant result in the wrong direction.

## Option 1 Conclusions

Assuming the true difference is 0.1 means that probability of girl births differs by 0.1 percentage points, comparing attractive with unattractive parents.

If the estimate is statistically significant, then:

1. We thus have a power - 2.3% = 2.7% probability of showing statistical significance in the correct direction.

## Option 1 Conclusions

Assuming the true difference is 0.1 means that probability of girl births differs by 0.1 percentage points, comparing attractive with unattractive parents.

If the estimate is statistically significant, then:

1. In expectation, a statistically significant result will be 78 times too high (the exaggeration ratio).

## Retrodesign Results: Options 2 and 3

| Assumption | Power | Type S | Exaggeration Ratio |
| --- | --- | --- | --- |
| D = 0.1 | 0.05 | 0.46 | 78 |
| D = 0.3 | 0.05 | 0.39 | 25 |
| D = 1.0 | 0.06 | 0.19 | 7.8 |

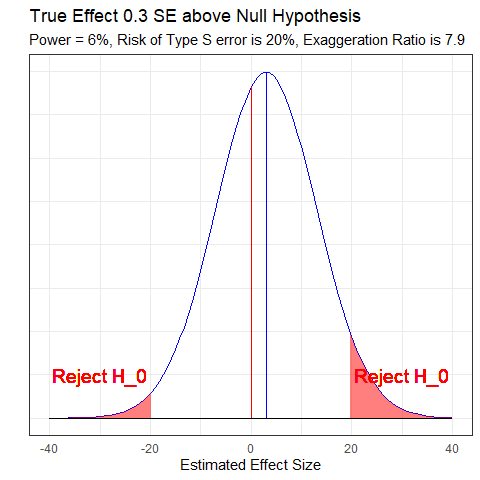
## What if true D = 1.0 percentage points?

Under a true difference of 1.0 percentage point, there would be

* a 4.9% chance of the result being statistically significantly positive and a 1.1% chance of a statistically significantly negative result.
* A statistically significant finding in this case has a 19% chance of appearing with the wrong sign, and
* the magnitude of the true effect would be overestimated by an expected factor of 8.

## What 6% power looks like…

ggplot(dat, aes(x, hx3)) +  
 geom\_line(col = "blue") +  
 geom\_segment(aes(x = 0, xend = 0, y = 0, yend = dnorm(0, mean = 3, sd = 10)), col = "red") +  
 geom\_segment(aes(x = 3, xend = 3, y = 0, yend = dnorm(3, mean = 3, sd = 10)), col = "blue") +  
 geom\_segment(aes(x = -40, xend = 40, y = 0, yend = 0)) +  
 geom\_ribbon(data = subset(dat, x > 19.6), aes(ymax = hx3), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_ribbon(data = subset(dat, x < -19.6), aes(ymax = hx3), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_text(x = 30, y = 0.005, label = "Reject H\_0", col = "red", size = 5) +  
 geom\_text(x = -30, y = 0.005, label = "Reject H\_0", col = "red", size = 5) +  
 theme(axis.ticks = element\_blank(), axis.text.y = element\_blank()) +  
 labs(x = "Estimated Effect Size", y = "", title = "True Effect 0.3 SE above Null Hypothesis", subtitle = "Power = 6%, Risk of Type S error is 20%, Exaggeration Ratio is 7.9")



## Gelman’s Chief Criticism: 6% Power = D.O.A.

Their effect size is tiny and their measurement error is huge. My best analogy is that they are trying to use a bathroom scale to weigh a feather … and the feather is resting loosely in the pouch of a kangaroo that is vigorously jumping up and down.



## What to do?

In advance, **and** after the fact, think hard about what a plausible effect size might be. Then…

* Analyze *all* your data.
* Present *all* your comparisons, not just a select few.
  + A big table, or even a graph, is what you want.
* Make your data public.
  + If the topic is worth studying, you should want others to be able to make rapid progress.

## But I do studies with 80% power? (1/2)

Based on some reasonable assumptions regarding main effects and interactions (specifically that the interactions are half the size of the main effects), you need **16 times** the sample size to estimate an interaction that you need to estimate a main effect.

And this implies a major, major problem with the usual plan of designing a study with a focus on the main effect, maybe even preregistering, and then looking to see what shows up in the interactions.

## But I do studies with 80% power? (2/2)

Or, even worse, designing a study, not finding the anticipated main effect, and then using the interactions to bail you out. The problem is not just that this sort of analysis is “exploratory”; it’s that these data are a lot noisier than you realize, so what you think of as interesting exploratory findings could be just a bunch of noise.

* Gelman [2018-03-15](http://andrewgelman.com/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/)

# A New Topic: Introducing Robust Linear Regression Methods

## The crimestat data

For each of 51 states (including the District of Columbia), we have the state’s ID number, postal abbreviation and full name, as well as:

* **crime** - the violent crime rate per 100,000 people
* **poverty** - the official poverty rate (% of people living in poverty in the state/district) in 2014
* **single** - the percentage of households in the state/district led by a female householder with no spouse present and with her own children under 18 years living in the household in 2016

## The crimestat data set

crimestat <- read\_csv("c13/data/crimestat.csv",   
 show\_col\_types = FALSE)  
crimestat

# A tibble: 51 × 6  
 sid state crime poverty single state.full   
 <dbl> <chr> <dbl> <dbl> <dbl> <chr>   
 1 1 AL 427. 19.2 9.02 Alabama   
 2 2 AK 636. 11.4 7.63 Alaska   
 3 3 AZ 400. 18.2 8.31 Arizona   
 4 4 AR 480. 18.7 9.41 Arkansas   
 5 5 CA 396. 16.4 7.25 California   
 6 6 CO 309. 12.1 6.75 Colorado   
 7 7 CT 237. 10.8 8.04 Connecticut   
 8 8 DE 489. 13 6.52 Delaware   
 9 9 DC 1244. 18.4 8.41 District of Columbia  
10 10 FL 540. 16.6 8.29 Florida   
# ℹ 41 more rows

## Modeling crime with poverty and single

Our main goal will be to build a linear regression model to predict **crime** using centered versions of both **poverty** and **single**.

crimestat <- crimestat |>  
 mutate(pov\_c = poverty - mean(poverty),  
 single\_c = single - mean(single))

## Our original (OLS) model

Note the sneaky trick with the outside parentheses…

(mod1 <- lm(crime ~ pov\_c + single\_c, data = crimestat))

Call:  
lm(formula = crime ~ pov\_c + single\_c, data = crimestat)  
  
Coefficients:  
(Intercept) pov\_c single\_c   
 364.41 16.11 23.84

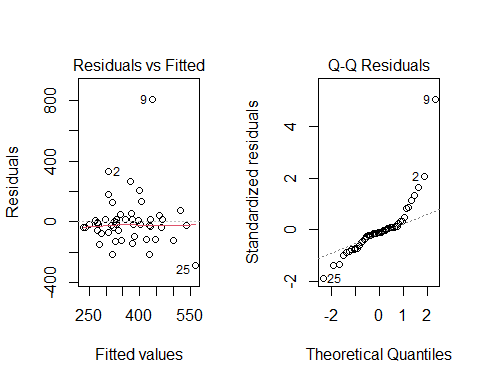
## Coefficients?

tidy(mod1, conf.int = TRUE) |>  
 select(term, estimate, std.error,   
 p.value, conf.low, conf.high) |>  
 gt() |> fmt\_number(decimals = 3) |> tab\_options(table.font.size = 24)

| term | estimate | std.error | p.value | conf.low | conf.high |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | 364.406 | 22.933 | 0.000 | 318.297 | 410.515 |
| pov\_c | 16.115 | 9.616 | 0.100 | -3.219 | 35.448 |
| single\_c | 23.843 | 18.384 | 0.201 | -13.121 | 60.807 |

## OLS Residuals

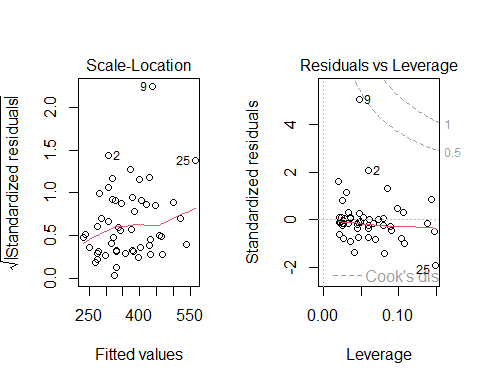
par(mfrow=c(1,2))  
plot(mod1, which = c(1:2))



par(mfrow = c(1,1))

## Remaining Residual Plots from OLS

par(mfrow=c(1,2))  
plot(mod1, which = c(3, 5))



par(mfrow = c(1,1))

## Which points are of special interest?

Several points show up in the residual plots.

crimestat |>  
 slice(c(2, 9, 25))

# A tibble: 3 × 8  
 sid state crime poverty single state.full pov\_c single\_c  
 <dbl> <chr> <dbl> <dbl> <dbl> <chr> <dbl> <dbl>  
1 2 AK 636. 11.4 7.63 Alaska -3.47 -0.0588  
2 9 DC 1244. 18.4 8.41 District of Columbia 3.53 0.721   
3 25 MS 278. 21.9 11.4 Mississippi 7.03 3.67

## Robust Linear Models (Huber weights)

There are several ways to do robust linear regression using M-estimation, including weighting using Huber and bisquare strategies.

* Robust linear regression here will make use of a method called iteratively re-weighted least squares (IRLS) to estimate models.
* M-estimation defines a weight function which is applied during estimation.

## Robust Linear Models (Huber weights)

* The weights depend on the residuals and the residuals depend on the weights, so an iterative process is required.

We’ll fit the model, using the default weighting choice: what are called Huber weights, where observations with small residuals get a weight of 1, and the larger the residual, the smaller the weight.

## Our robust model (using MASS::rlm)

rob.huber <-   
 rlm(crime ~ pov\_c + single\_c, data = crimestat)

## Summary of the robust (Huber weights) model

tidy(rob.huber) |>  
 gt() |> fmt\_number(decimals = 3) |> tab\_options(table.font.size = 24)

| term | estimate | std.error | statistic |
| --- | --- | --- | --- |
| (Intercept) | 343.798 | 13.131 | 26.182 |
| pov\_c | 11.910 | 5.506 | 2.163 |
| single\_c | 30.987 | 10.527 | 2.944 |

Now, *both* predictors appear to have estimates that exceed twice their standard error. So this is a very different result than ordinary least squares gave us.

## Glance at the robust model (vs. OLS)

glance(mod1) |> gt() |> fmt\_number(decimals = 3) |>  
 tab\_options(table.font.size = 24)

| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residual | nobs |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.197 | 0.163 | 163.771 | 5.883 | 0.005 | 2.000 | -330.842 | 669.684 | 677.411 | 1,287,404.929 | 48.000 | 51.000 |

glance(rob.huber) |> gt() |> fmt\_number(decimals = 3) |>  
 tab\_options(table.font.size = 24)

| sigma | converged | logLik | AIC | BIC | deviance | nobs |
| --- | --- | --- | --- | --- | --- | --- |
| 59.145 | TRUE | -331.378 | 670.757 | 678.484 | 1,314,783.620 | 51.000 |

## Understanding the Huber weights

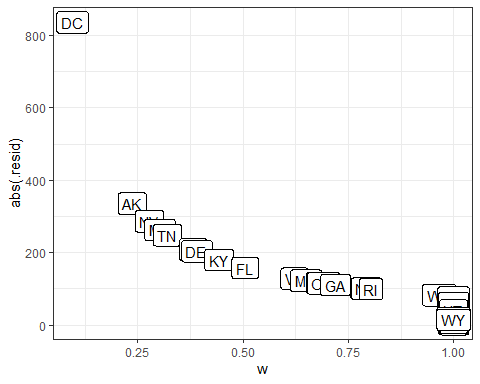
Let’s augment the data with results from this model, including the weights.

crime\_with\_huber <- augment(rob.huber, crimestat) |>  
 mutate(w = rob.huber$w) |> arrange(w)   
  
crime\_with\_huber |>   
 select(sid, state, w, crime, pov\_c, single\_c, everything()) |>  
 head()

# A tibble: 6 × 15  
 sid state w crime pov\_c single\_c poverty single state.full .fitted  
 <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <dbl>  
1 9 DC 0.0951 1244. 3.53 0.721 18.4 8.41 District of … 408.  
2 2 AK 0.237 636. -3.47 -0.0588 11.4 7.63 Alaska 301.  
3 29 NV 0.278 636. 0.527 -0.0288 15.4 7.66 Nevada 349.  
4 25 MS 0.303 278. 7.03 3.67 21.9 11.4 Mississippi 541.  
5 43 TN 0.321 608. 3.33 -0.749 18.2 6.94 Tennessee 360.  
6 46 VT 0.382 99.3 -2.67 -0.139 12.2 7.55 Vermont 308.  
# ℹ 5 more variables: .resid <dbl>, .hat <dbl>, .sigma <dbl>, .cooksd <dbl>,  
# .std.resid <dbl>

## Are cases with large residuals down-weighted?

ggplot(crime\_with\_huber, aes(x = w, y = abs(.resid))) +  
 geom\_label(aes(label = state))



## Conclusions from the Plot of Weights

* District of Columbia will be down-weighted the most, followed by Alaska and then Nevada and Mississippi.
* But many of the observations will have a weight of 1.
* In ordinary least squares, all observations would have weight 1.
* So the more cases in the robust regression that have a weight close to one, the closer the results of the OLS and robust procedures will be.

## summary(rob.huber)

summary(rob.huber)

Call: rlm(formula = crime ~ pov\_c + single\_c, data = crimestat)  
Residuals:  
 Min 1Q Median 3Q Max   
-262.751 -45.641 1.762 36.732 836.244   
  
Coefficients:  
 Value Std. Error t value   
(Intercept) 343.7982 13.1309 26.1823  
pov\_c 11.9098 5.5058 2.1631  
single\_c 30.9868 10.5266 2.9437  
  
Residual standard error: 59.14 on 48 degrees of freedom

## Robust Linear Regression with the biweight

As mentioned there are several possible weighting functions - we’ll next try the **biweight**, also called the bisquare or Tukey’s bisquare, in which all cases with a non-zero residual get down-weighted at least a little.

## Robust Linear Model with biweight

Here is the resulting fit…

(rob.biweight <- rlm(crime ~ pov\_c + single\_c,  
 data = crimestat, psi = psi.bisquare))

Call:  
rlm(formula = crime ~ pov\_c + single\_c, data = crimestat, psi = psi.bisquare)  
Converged in 13 iterations  
  
Coefficients:  
(Intercept) pov\_c single\_c   
 336.17015 10.31578 34.70765   
  
Degrees of freedom: 51 total; 48 residual  
Scale estimate: 67.3

## Coefficients and Standard Errors

tidy(rob.biweight) |>   
 gt() |> fmt\_number(decimals = 3) |> tab\_options(table.font.size = 24)

| term | estimate | std.error | statistic |
| --- | --- | --- | --- |
| (Intercept) | 336.170 | 12.673 | 26.526 |
| pov\_c | 10.316 | 5.314 | 1.941 |
| single\_c | 34.708 | 10.160 | 3.416 |

## Understanding the biweights weights a bit

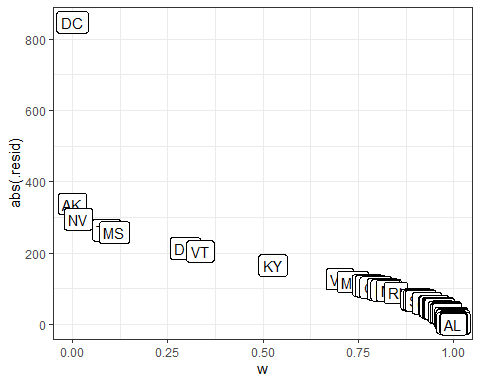
Let’s augment the data, as above

crime\_with\_biweights <-   
 augment(rob.biweight, newdata = crimestat) |>  
 mutate(w = rob.biweight$w) |>   
 arrange(w)  
  
head(crime\_with\_biweights, 3)

# A tibble: 3 × 11  
 sid state crime poverty single state.full pov\_c single\_c .fitted .resid  
 <dbl> <chr> <dbl> <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
1 2 AK 636. 11.4 7.63 Alaska -3.47 -0.0588 298. 337.  
2 9 DC 1244. 18.4 8.41 District of C… 3.53 0.721 398. 847.  
3 29 NV 636. 15.4 7.66 Nevada 0.527 -0.0288 341. 295.  
# ℹ 1 more variable: w <dbl>

## Relationship of Weights and Residuals

ggplot(crime\_with\_biweights, aes(x = w, y = abs(.resid))) +  
 geom\_label(aes(label = state))



## Conclusions from the biweights plot

Again, cases with large residuals (in absolute value) are down-weighted generally, but here, Alaska and Washington DC receive no weight at all in fitting the final model.

* We can see that the weight given to DC and Alaska is dramatically lower (in fact it is zero) using the bisquare weighting function than the Huber weighting function and the parameter estimates from these two different weighting methods differ.
* The maximum weight (here, for Alabama) for any state using the biweight is still slightly smaller than 1.

## summary(rob.biweight)

summary(rob.biweight)

Call: rlm(formula = crime ~ pov\_c + single\_c, data = crimestat, psi = psi.bisquare)  
Residuals:  
 Min 1Q Median 3Q Max   
-257.58 -40.53 8.01 45.30 846.81   
  
Coefficients:  
 Value Std. Error t value   
(Intercept) 336.1702 12.6733 26.5259  
pov\_c 10.3158 5.3139 1.9413  
single\_c 34.7077 10.1598 3.4162  
  
Residual standard error: 67.27 on 48 degrees of freedom

## Comparing OLS and the two weighting schemes

glance(mod1) |> select(1:6)

# A tibble: 1 × 6  
 r.squared adj.r.squared sigma statistic p.value df  
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 0.197 0.163 164. 5.88 0.00518 2

glance(mod1) |> select(7:12)

# A tibble: 1 × 6  
 logLik AIC BIC deviance df.residual nobs  
 <dbl> <dbl> <dbl> <dbl> <int> <int>  
1 -331. 670. 677. 1287405. 48 51

## Comparing OLS and the two weighting schemes

glance(rob.biweight) # biweights

# A tibble: 1 × 7  
 sigma converged logLik AIC BIC deviance nobs  
 <dbl> <lgl> <logLik> <dbl> <dbl> <dbl> <int>  
1 67.3 TRUE -331.8601 672. 679. 1339850. 51

glance(rob.huber) # Huber weights

# A tibble: 1 × 7  
 sigma converged logLik AIC BIC deviance nobs  
 <dbl> <lgl> <logLik> <dbl> <dbl> <dbl> <int>  
1 59.1 TRUE -331.3785 671. 678. 1314784. 51

## Quantile Regression on the Median

We can use the rq function in the quantreg package to model the **median** of our outcome (violent crime rate) on the basis of our predictors, rather than the mean, as is the case in ordinary least squares.

rob.quan <- rq(crime ~ pov\_c + single\_c, data = crimestat)  
  
glance(rob.quan)

# A tibble: 1 × 5  
 tau logLik AIC BIC df.residual  
 <dbl> <logLik> <dbl> <dbl> <int>  
1 0.5 -315.7569 638. 643. 48

## summary(rob.quan)

summary(rob.quan <- rq(crime ~ pov\_c + single\_c,   
 data = crimestat))

Call: rq(formula = crime ~ pov\_c + single\_c, data = crimestat)  
  
tau: [1] 0.5  
  
Coefficients:  
 coefficients lower bd upper bd   
(Intercept) 344.75658 336.94534 366.23603  
pov\_c 10.54757 3.06714 28.95962  
single\_c 32.27249 4.45889 48.18925

## Estimating a different quantile (tau = 0.70)

In fact, if we like, we can estimate any quantile by specifying the tau parameter (here tau = 0.5, by default, so we estimate the median.)

(rob.quan70 <- rq(crime ~ pov\_c + single\_c, tau = 0.70,  
 data = crimestat))

Call:  
rq(formula = crime ~ pov\_c + single\_c, tau = 0.7, data = crimestat)  
  
Coefficients:  
(Intercept) pov\_c single\_c   
 379.72818 19.30376 32.15827   
  
Degrees of freedom: 51 total; 48 residual

## Comparing our Four Models

**Estimating the Mean**

| Fit | Intercept CI | pov\_c CI | single\_c CI |
| --- | --- | --- | --- |
| OLS | (318.6, 410.2) | (-3.13, 35.35) | (-12.92, 60.6) |
| Robust (Huber) | (320, 367.6) | (0.89, 22.93) | (9.93, 52.05) |
| Robust (biweight) | (310.7, 361.5) | (-0.3, 20.94) | (14.39, 55.03) |

**Note**: CIs estimated for OLS and Robust methods as point estimate 2 standard errors

**Estimating the Median**

| Fit | Intercept CI | pov\_c CI | single\_c CI |
| --- | --- | --- | --- |
| Quantile (Median) Reg | (336.9, 366.2) | (3.07, 28.96) | (4.46, 48,19) |

## Comparing AIC and BIC

| Fit | AIC | BIC |
| --- | --- | --- |
| OLS | 669.7 | 677.4 |
| Robust (Huber) | 670.8 | 678.5 |
| Robust (biweight) | 671.7 | 679.4 |
| Quantile (median) | 637.5 | 643.3 |

## Some Closing Thoughts (1/2)

1. When comparing the results of a regular OLS regression and a robust regression for a data set which displays outliers, if the results are very different, you will most likely want to use the results from the robust regression.
   * Large differences suggest that the model parameters are being highly influenced by outliers.

## Some Closing Thoughts (2/2)

1. Different weighting functions have advantages and drawbacks.
   * Huber weights can have difficulties with really severe outliers.
   * Bisquare weights can have difficulties converging or may yield multiple solutions.
   * Quantile regression approaches have some nice properties, but describe medians (or other quantiles) rather than means.