

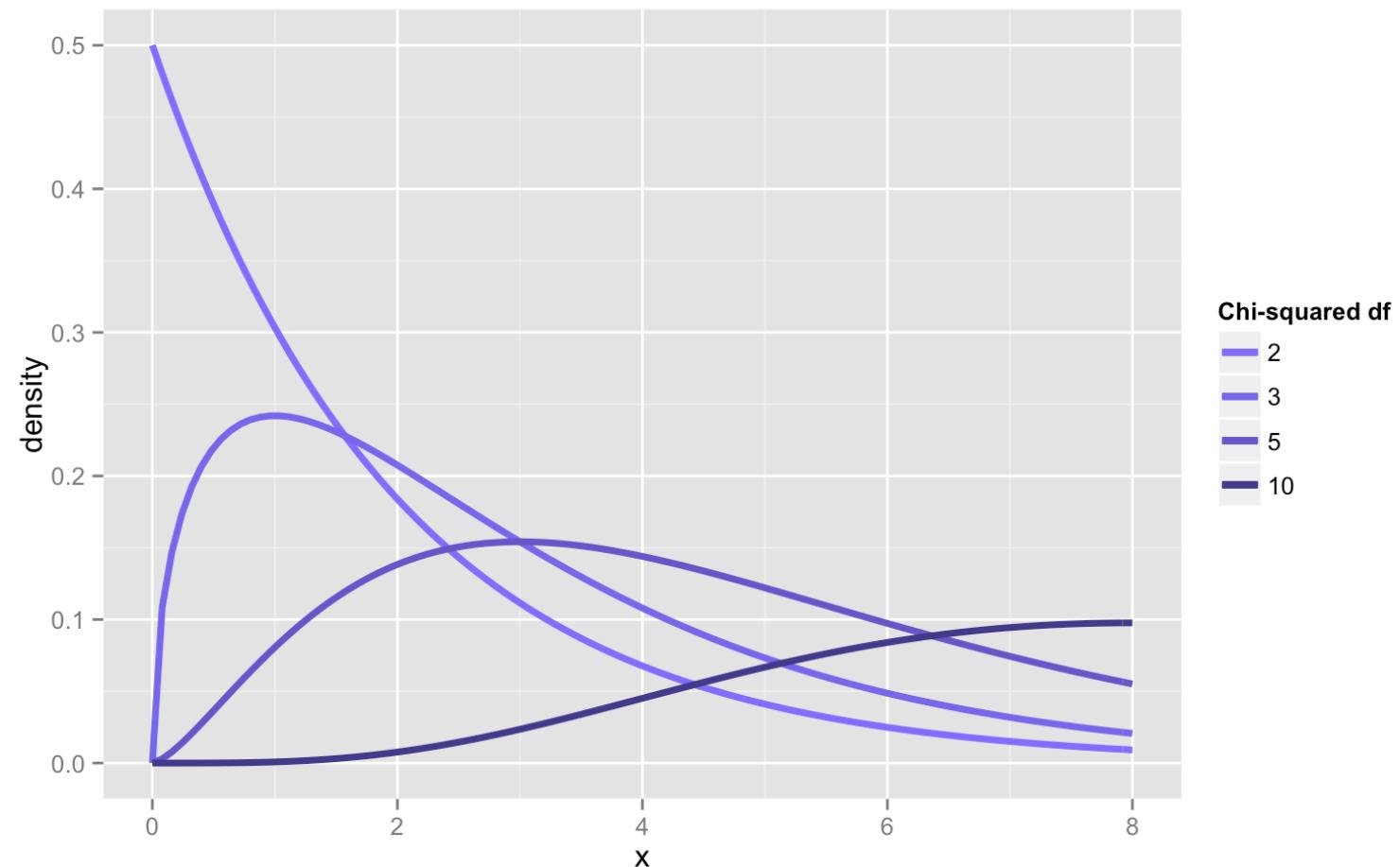
Class 9: Regression diagnostics & adding predictors

Alison Presmanes Hill

3 new distributions...

- If you see squares of normal random variables (or of their differences!), think **chi-squared**.
- If you see the ratio of an independent normal random variable (numerator) to the square-root of a chi-squared variable (denominator; any time you have an SE on the bottom!), think **Student's t**.
- If you see the ratio of two chi-squared random variables (i.e., sample variances from a normal distribution), think **F distribution**.

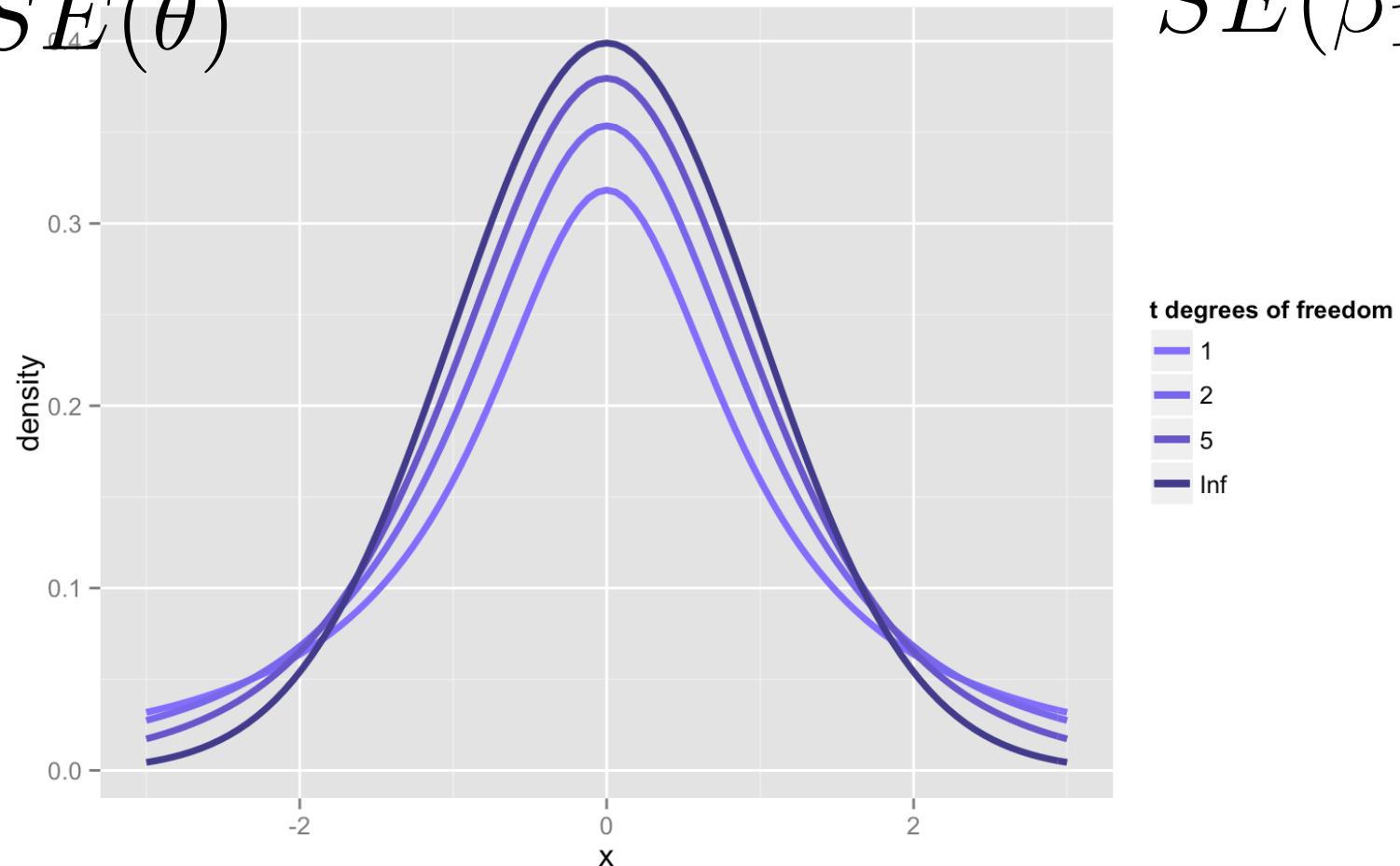
If you see squares of normal random variables, think chi-squared



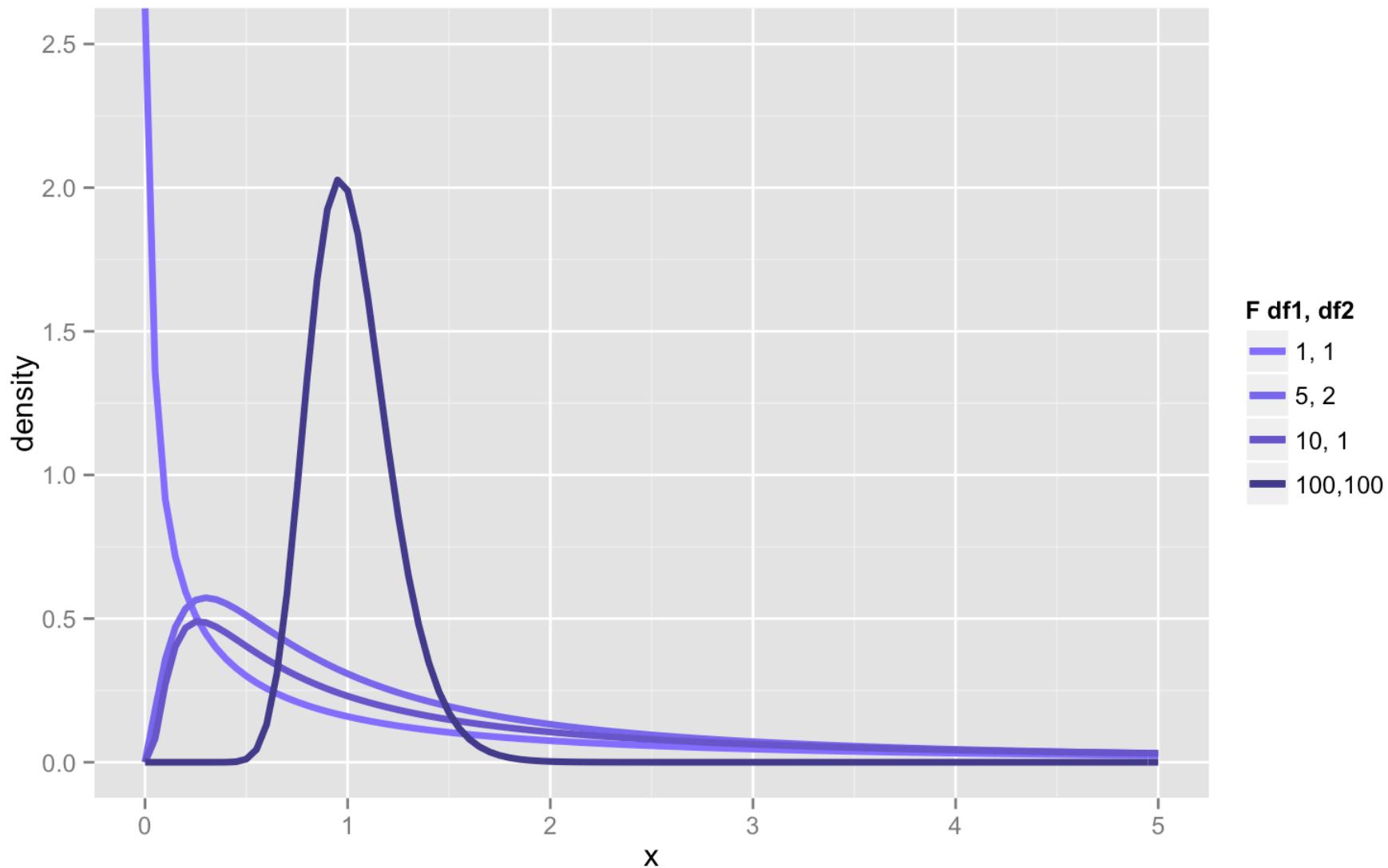
If you see the ratio of an independent normal random variable to the square-root of a chi-squared, think Student's t

$$T = \frac{\hat{\theta} - \theta}{SE(\hat{\theta})}$$

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$$



If you see the ratio of two chi-squared random variables (i.e., sample variances from a normal distribution), think F distribution



Swiss Fertility and Socioeconomic Indicators (1888)

Details (paraphrasing Mosteller and Tukey):

- Switzerland, in 1888, was entering a period known as the *demographic transition*; i.e., its fertility was beginning to fall from the high level typical of underdeveloped countries.
- The data collected are for 47 French-speaking “provinces” at about 1888.
- Here, all variables are scaled to [0, 100].



6 swiss variables

- Fertility: (common standardized measure)
- Agriculture: % of males involved in agriculture as an occupation
- Examination: % draftees received highest mark on army exam
- Education: % draftees educated beyond primary school
- Catholic: % Catholic (as opposed to Protestant)
- InfantMortality: Percent of live births who live less than one year



```
> summary(swiss)
```

fertility	agriculture	examination
Min. :35.00	Min. : 1.20	Min. : 3.00
1st Qu.:64.70	1st Qu.:35.90	1st Qu.:12.00
Median :70.40	Median :54.10	Median :16.00
Mean :70.14	Mean :50.66	Mean :16.49
3rd Qu.:78.45	3rd Qu.:67.65	3rd Qu.:22.00
Max. :92.50	Max. :89.70	Max. :37.00

education	catholic	infant.mortality
Min. : 1.00	Min. : 2.150	Min. :10.80
1st Qu.: 6.00	1st Qu.: 5.195	1st Qu.:18.15
Median : 8.00	Median : 15.140	Median :20.00
Mean :10.98	Mean : 41.144	Mean :19.94
3rd Qu.:12.00	3rd Qu.: 93.125	3rd Qu.:21.70
Max. :53.00	Max. :100.000	Max. :26.60


```
> with(swiss, cor.test(fertility, agriculture))
```

Pearson's product-moment correlation

```
data: fertility and agriculture
```

```
t = 2.5316, df = 45, p-value = 0.01492
```

```
alternative hypothesis: true correlation is not equal to 0
```

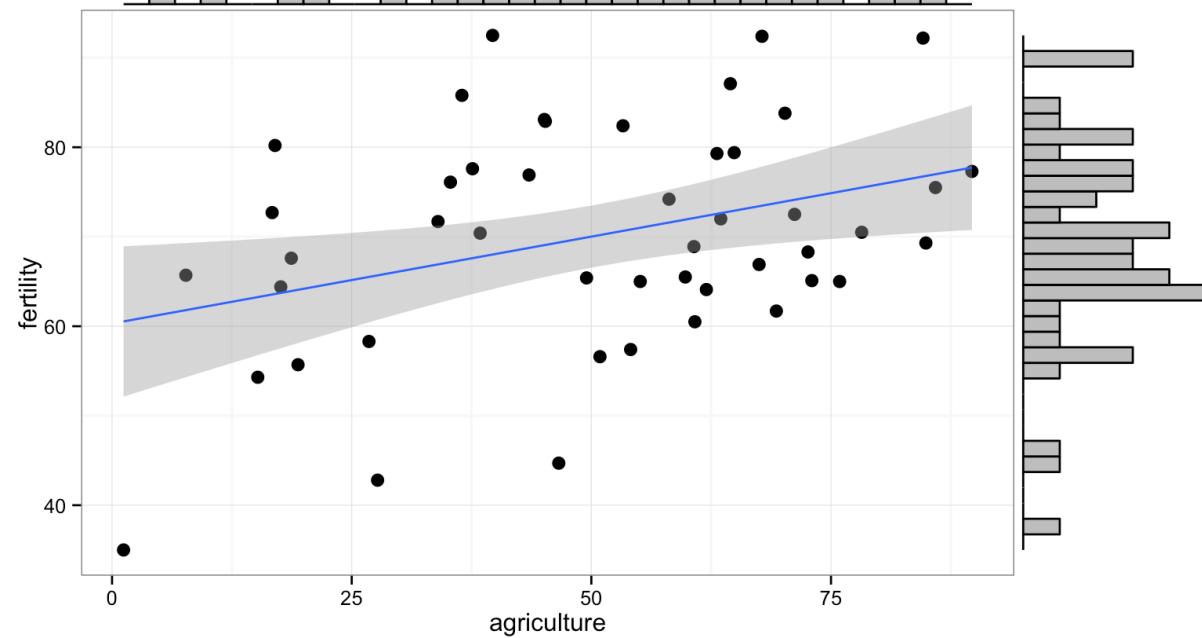
```
95 percent confidence interval:
```

```
0.07334947 0.58130587
```

```
sample estimates:
```

```
cor
```

```
0.3530792
```



```
> agr_mod <- lm(fertility ~ agriculture, data = swiss) # simple linear regression  
> summary(agr_mod)
```

Call:

```
lm(formula = fertility ~ agriculture, data = swiss)
```

Residuals:

Min	1Q	Median	3Q	Max
-25.5374	-7.8685	-0.6362	9.0464	24.4858

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	60.30438	4.25126	14.185	<2e-16 ***
agriculture	0.19420	0.07671	2.532	0.0149 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.82 on 45 degrees of freedom

Multiple R-squared: 0.1247, Adjusted R-squared: 0.1052

F-statistic: 6.409 on 1 and 45 DF, p-value: 0.01492



```
> agr_mod <- lm(fertility ~ agriculture, data = swiss)
> summary(agr_mod)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	60.30438	4.25126	14.185	<2e-16 ***
agriculture	0.19420	0.07671	2.532	0.0149 *

Signif. codes:	0	'****'	0.001	'***'
	0.01	'*' 0.05	'. '	0.1 ' '
	1			

Residual standard error: 11.82 on 45 degrees of freedom

Multiple R-squared: 0.1247, Adjusted R-squared: 0.1052

F-statistic: 6.409 on 1 and 45 DF, p-value: 0.01492

```
> int_mod <- lm(fertility ~ 1, data = swiss) # intercept ONLY
> anova(int_mod, agr_mod)
```

Analysis of Variance Table

Model 1: fertility ~ 1

Model 2: fertility ~ agriculture

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	46	7178.0				
2	45	6283.1	1	894.84	6.4089	0.01492 *

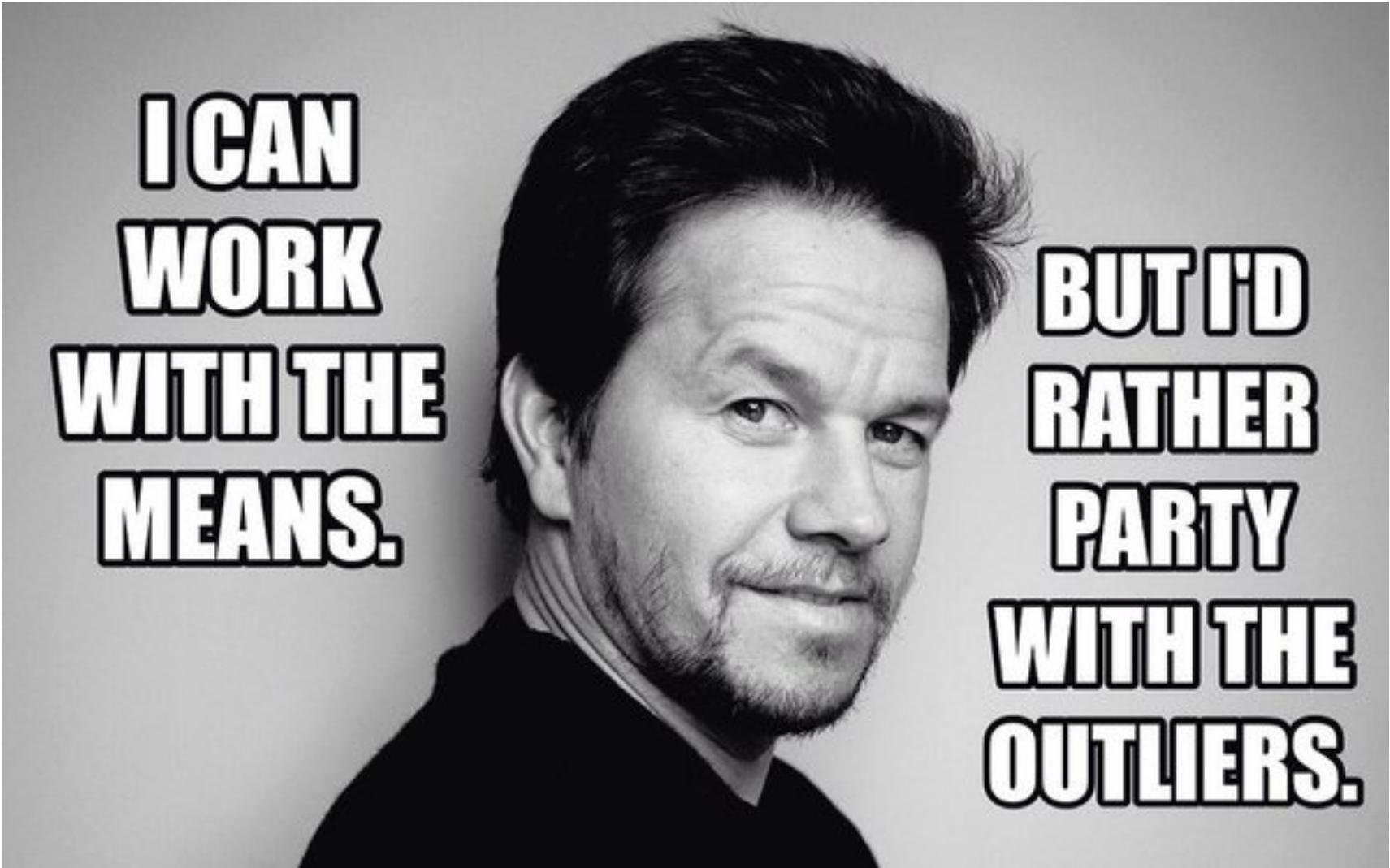
Signif. codes: 0 '****' 0.001 '***' 0.01 '*' 0.05 '. ' 0.1 ' ' 1



Sequences of nested models (revisited)

- Full(er) model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- (Default) reduced model: $Y_i = \beta_0 + \varepsilon_i$





**I CAN
WORK
WITH THE
MEANS.**

**BUT I'D
RATHER
PARTY
WITH THE
OUTLIERS.**

Types of influence

- Extremity on the x's (leverage)
- Extremity on y (discrepancy)
- Influence on the regression estimates
 - Global
 - Specific coefficients

```
agr_mod <- lm(fertility ~ agriculture, data = swiss) # original model  
slr_vars <- augment(agr_mod) # broom
```

Contaminated observation or special snowflake?

- A **contaminated observation** is one that has been damaged in some way. Some examples:
 - Error of execution of the research procedure.
 - Inaccurate measurement of the dependent measure.
 - Data entry error.
 - Error in calculating a measure.
 - Nonattentive or distracted participants.
- The outlier may simply be an **extremely rare case**. For example, a college freshman might be 12 years old and have an 800 SAT in math. Such an individual is extremely rare, but data is valid.



**SUP BRO! YEAH I SAW
YOU REMOVE MY
OUTLIERS WITHOUT
JUSTIFICATION.**

HILARIOUS.

Outliers



U.S. Department of Health & Human Services

www.hhs.gov



Home

Home

Contact Us

Google™ Custom Search



ORI
2.0

ORI found that Respondent knowingly falsified data by removing outlier values or replacing outliers with mean values to produce results that conform to predictions. Specifically, these falsifications appear in:

Case Summary: Anderson, David

DEPARTMENT OF HEALTH AND HUMAN SERVICES

Office of the Secretary

Findings of Research Misconduct

AGENCY: Office of the Secretary, HHS

ACTION: Notice.

SUMMARY: Notice is hereby given that the Office of Research Integrity (ORI) has taken final action in the following case:

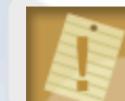
David Anderson, University of Oregon, Eugene: Based on an assessment conducted by the University of Oregon, Eugene (UOE), the Respondent's admission, and analysis conducted by ORI, ORI and UOE found that Mr. David Anderson, Graduate Student, UOE, engaged in



Newsletter



Follow Us on Twitter



PHS Administrative Action
Bulletin Board



Annual Report System



Extremity on the x's: leverage

- Standardized measure of how far the observed value for each observation is from the mean value on the set of x's
- Observations with high leverage have the potential to be influential, especially if also extreme on Y

Leverage

- Measure of how unusual the X value of a point is, relative to the X observations as a whole. Absolute minimum of 1/n. Leverage describes how unusual an observation is in predictor(s) data.
- If h_{ii} is large then the ith observation has considerable impact on the fitted value
- **broom:: .hat**

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum x^2}$$

Leverage

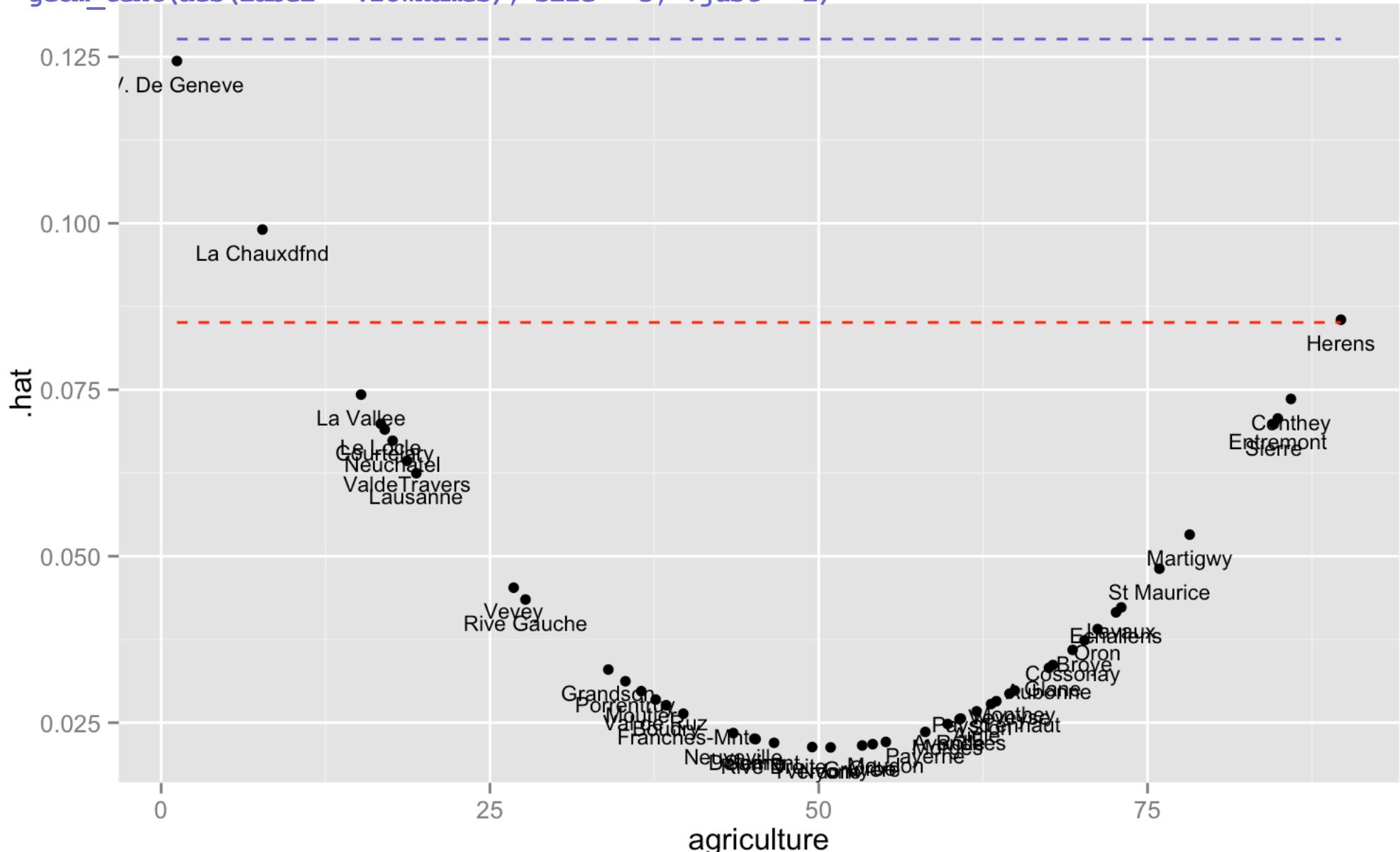
```
slr_vars %>%
  filter(.hat > (2*mean(.hat))) %>%
  select(.rownames, fertility, agriculture, .hat)

  .rownames fertility agriculture .hat
1      Herens      77.3        89.7 0.08551436
2 La Chauxdfnd      65.7        7.7 0.09905898
3 V. De Geneve      35.0        1.2 0.12437740

## small samples
slr_vars %>%
  filter(.hat > (3*mean(.hat))) %>%
  select(.rownames, fertility, agriculture, .hat)
[1] .rownames   fertility   agriculture .hat
<0 rows> (or 0-length row.names)
```

```
## plot leverage versus xs
```

```
ggplot(slr_vars, aes(x = agriculture, y = .hat)) +  
  geom_point() +  
  geom_line(aes(y = 2*mean(.hat)), colour = "red", lty = "dashed") +  
  geom_line(aes(y = 3*mean(.hat)), colour = "slateblue", lty = "dashed") +  
  geom_text(aes(label = .rownames), size = 3, vjust = 2)
```



Just the x's

```
slr_vars %>%
  select(.rownames, fertility, agriculture, .hat) %>%
  mutate(mean_dev = agriculture - mean(agriculture)) %>%
  arrange(mean_dev)

  .rownames  fertility  agriculture      .hat    mean_dev
1 V. De Geneve     35.0       1.2 0.12437740 -49.4595745
2 La Chauxdfnd    65.7       7.7 0.09905898 -42.9595745
3 La Vallee      54.3      15.2 0.07427080 -35.4595745
...
45 Entremont      69.3      84.9 0.07068941  34.2404255
46 Conthey        75.5      85.9 0.07361779  35.2404255
47 Herens         77.3      89.7 0.08551436  39.0404255
```

Extremity on the y's: discrepancy

- The discrepancy (or distance²) between each predicted and observed value of y_i
- A studentized residual is an observed residual divided by its standard error; two types:
 - Internally studentized (**rstandard**): re-normalize the residuals to have unit variance, using a measure of the error variance.
 - Also **broom::std.resid**
 - Externally studentized (**rstudent**): re-normalize the residuals to have unit variance, using a leave-one-out measure of the error variance. This is a measure of the size of the residual, standardized by the estimated standard deviation of residuals based on all the data but that observation. Sometimes called jackknifed residuals.

Internally studentized residuals

- **rstandard**
- Also **broom::augment(.std.resid)**

$$e'_i = \frac{e_i}{SE_e \sqrt{1 - h_i}}$$

Externally studentized residuals

- `rstudent`
- Sadly, not available in `broom::augment`

$$e_i^* = \frac{e_i}{SE_{e-i}\sqrt{1 - h_i}}$$

Save externally studentized residuals (ESR)

```
m_agr <- lm(fertility ~ agriculture, data = swiss) # original model
slr_vars <- augment(m1) # broom
slr_vars <- slr_vars %>%
  mutate(.extsr = rstudent(m1)) # add ESR

> slr_vars %>%
  select(.std.resid, .extsr) %>%
  head()
  .std.resid    .extsr
1  1.4554806 1.4743391
2  1.2015896 1.2076959
3  2.1000488 2.1864890
4  1.5814696 1.6091557
5  0.6977688 0.6937354
6  0.7687025 0.7651537
```

Expect 5% with $\text{abs}(\text{.extsr}) \geq 2$

```
> slr_vars %>%
  filter(abs(.extsr) >= 2) %>%
  select(.rownames, fertility, agriculture, .resid, .std.resid, .extsr)
```

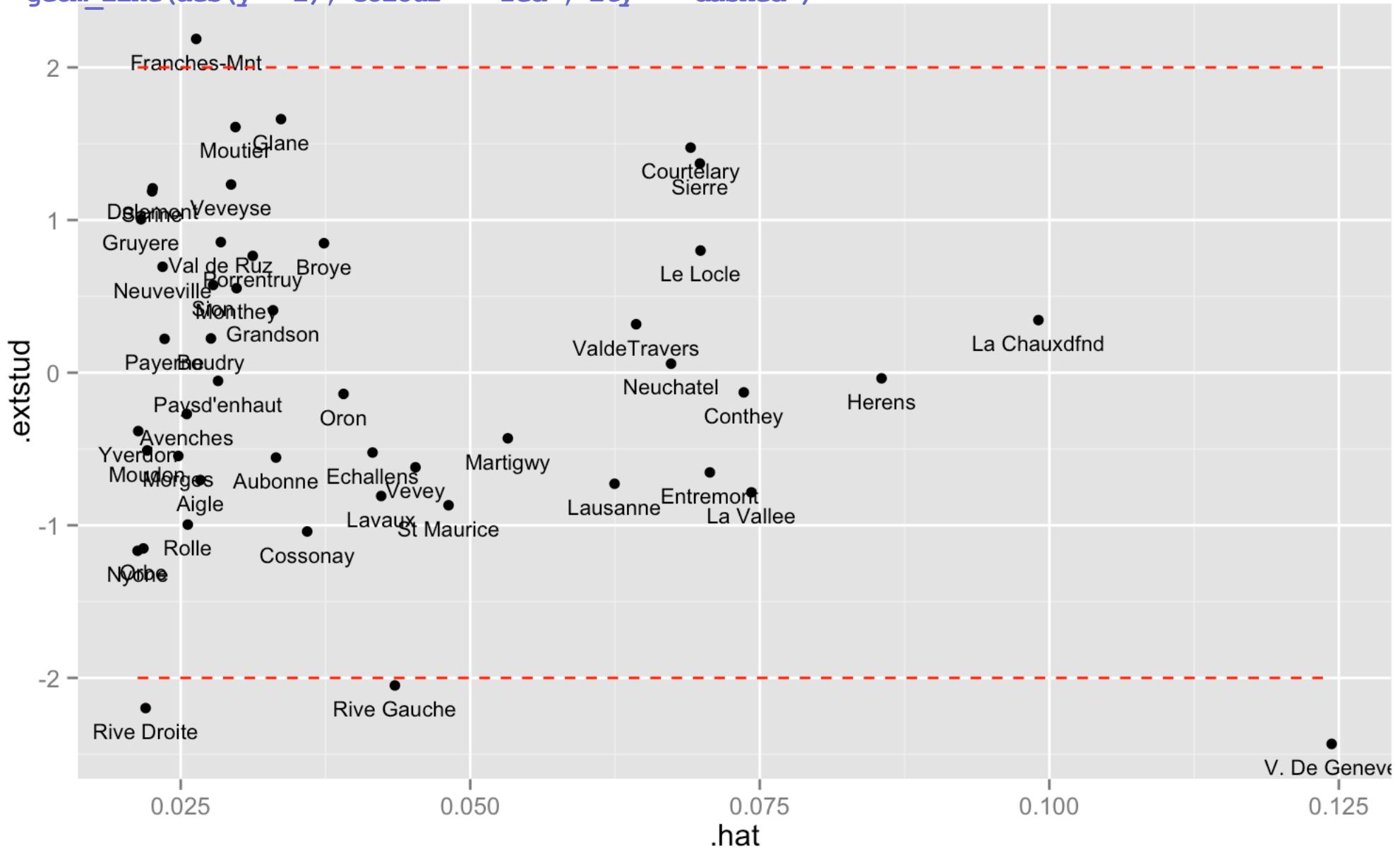
	.rownames	fertility	agriculture	.resid	.std.resid	.extsr
1	Franches-Mnt	92.5	39.7	24.48582	2.100049	2.186489
2	V. De Geneve	35.0	1.2	-25.53742	-2.309602	-2.432516
3	Rive Droite	44.7	46.6	-24.65418	-2.109762	-2.197709
4	Rive Gauche	42.8	27.7	-22.88376	-1.980169	-2.049364

4 out of 47:

8.5% of observations with ESR considered to be
relatively large
(expected ≈ 2 or 3 observations)

```
## plot leverage versus ESR
```

```
ggplot(slr_vars, aes(x = .hat, y = .extsr)) +  
  geom_point() +  
  geom_text(aes(label = .rownames), size = 3, vjust = 2) +  
  geom_line(aes(y = -2), colour = "red", lty = "dashed") +  
  geom_line(aes(y = 2), colour = "red", lty = "dashed")
```



car::outlierTest()

```
> outlierTest(m1) # library(car)
```

No Studentized residuals with Bonferroni p < 0.05

Largest |rstudent|:

	rstudent	unadjusted p-value	Bonferroni p
V. De Geneve	-2.432516	0.01913	0.89909

ESR follows t_{n-k-2} distribution (k = number of predictors, n = number of observations), t_{44}

Alpha = .025/n (two-tailed)

```
> qt(1-.025/47, 44) # 1 - (alpha/2)/n, df = n-k-2  
[1] 3.504708  
> pt(-2.432516, 44)*2  
[1] 0.01912966
```



Bonferroni adjustment

- Problem: if n is large, and we “threshold” at t_{n-k-2} we will get many outliers by chance alone, even if our model is correct
- Solution: adjust the “threshold” to reflect reality- we are screening all observed ESR values and looking for the max, so we should account for multiple testing somehow
- Recall: ESRs are t-distributed variables
- For nominal $\alpha = .05$, the nominal critical t-value value is t_{n-k-2} where k = number of regression coefficients (including intercept), n = number of observations
- Instead, we would set $\alpha = .05/n$ to find ESRs that are in fact outliers

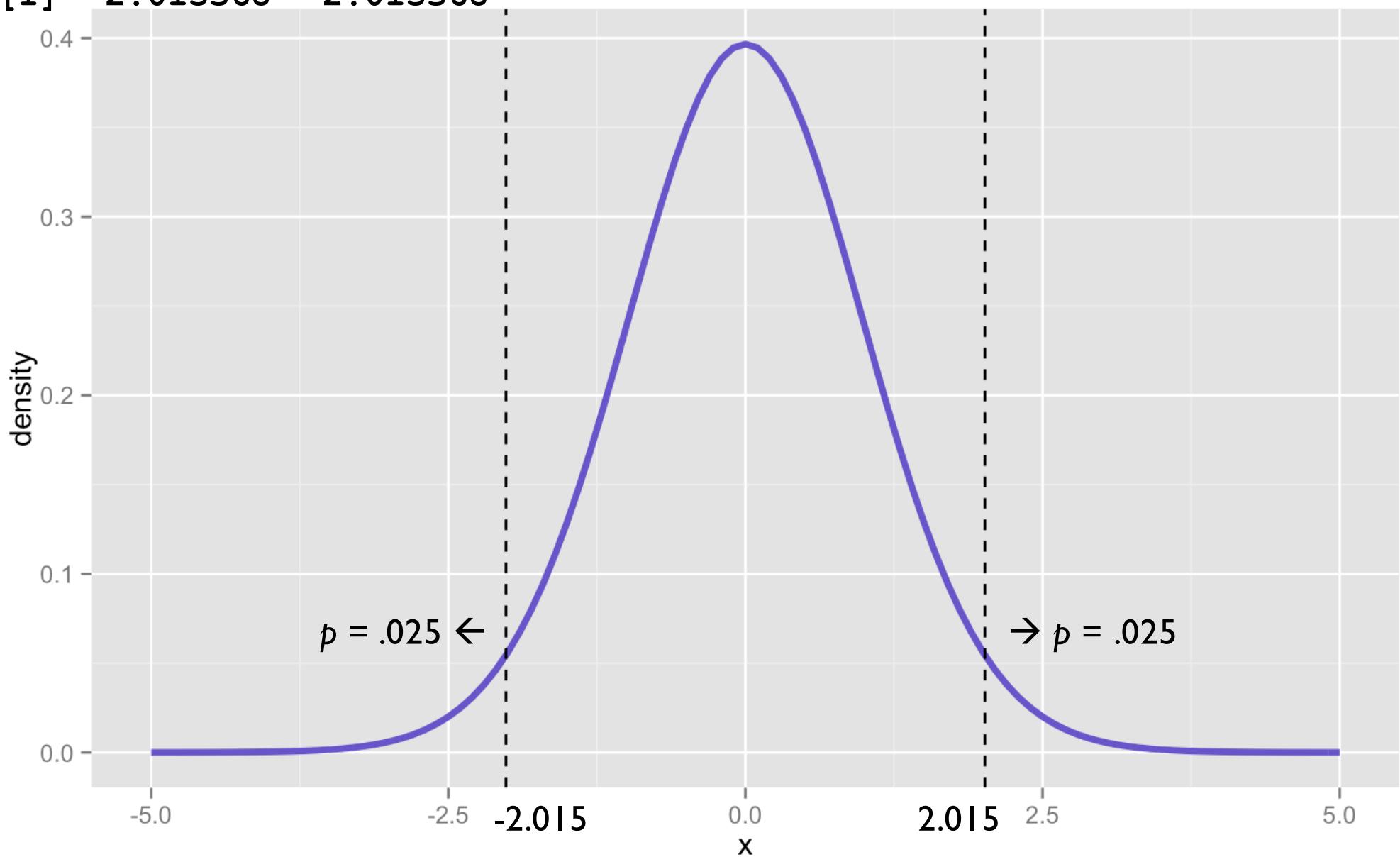
Bonferroni adjustment

- If I only wanted to know whether one datapoint was more extreme at $\alpha = .05$ (two-tailed), we compare to critical $t_{44} =$

```
> lowert <- qt(.025, 44)  
> uppert <- qt(1 - .025, 44)  
> c(lowert, uppert)  
[1] -2.015368 2.015368
```



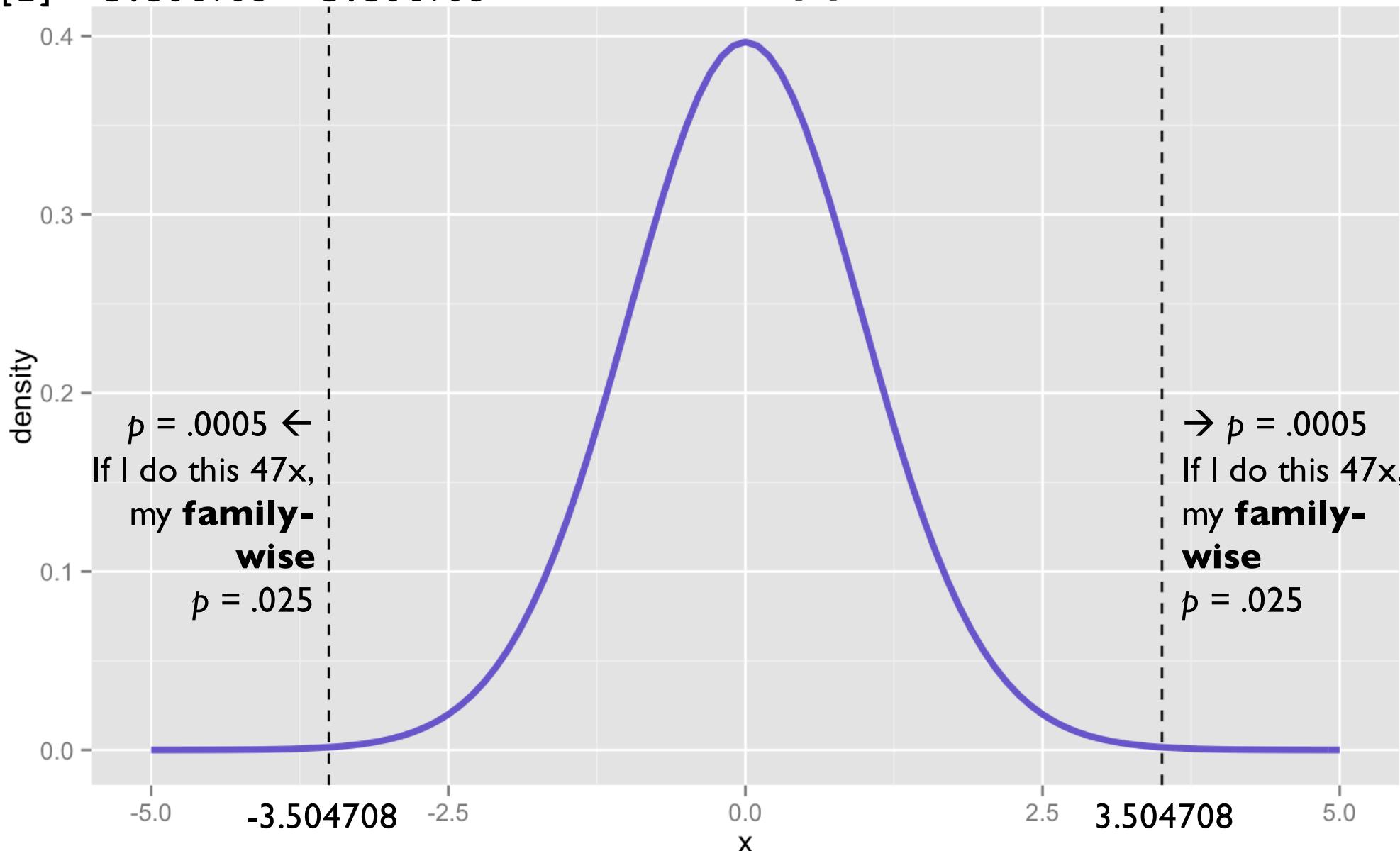
```
> lowert <- qt(.025, 44)
> uppert <- qt(1 - .025, 44)
> c(lowert, uppert)
[1] -2.015368  2.015368
```



```

> lowertstar <- qt(.025/47, 44)           > pstar <- 1 - pt(3.504708, 44)
> uppertstar <- qt(1-.025/47, 44)         > pstar_fam <- pstar*47*2
> c(lowertstar, uppertstar)                 > c(pstar, pstar_fam)
[1] -3.504708 3.504708                   [1] 0.0005319156 0.0500000646

```



A black and white photograph of actor Matt Damon. He is positioned in the center-right of the frame, looking directly at the camera with a slight smile. He has short brown hair and is wearing a dark, button-down shirt under a light-colored blazer. The background is a soft-focus outdoor scene with trees and foliage.

**YOU DON'T
NEED THAT
BONFERRONI,
GIRL.**

**THERE'S NO
COMPARISON.**

Influence on regression estimates

- Global: DFFITS and Cook's D provide information about how each case affects overall regression equation
 - Both are deletion statistics (like the externally studentized residuals)
 - Answer similar questions, but scale differently
- Specific: DFBETAS provide information about how each case affects each individual coefficient estimate (i.e., each beta)
 - Will return to this in multiple regression

DFFITS

- Compares predicted values of \hat{Y}_i when (1) observation is included and (2) when observation is excluded
- Number of standard deviations by which \hat{y}_{hat} would change for each case if it were deleted
- Thus: “difference in fit, standardized”
- Higher absolute values mean greater influence; 0 means no influence
- But: extreme outliers could also change the regression equation, and hence the predicted values, for **other** observations too...

Cook's distance (Cook's D)

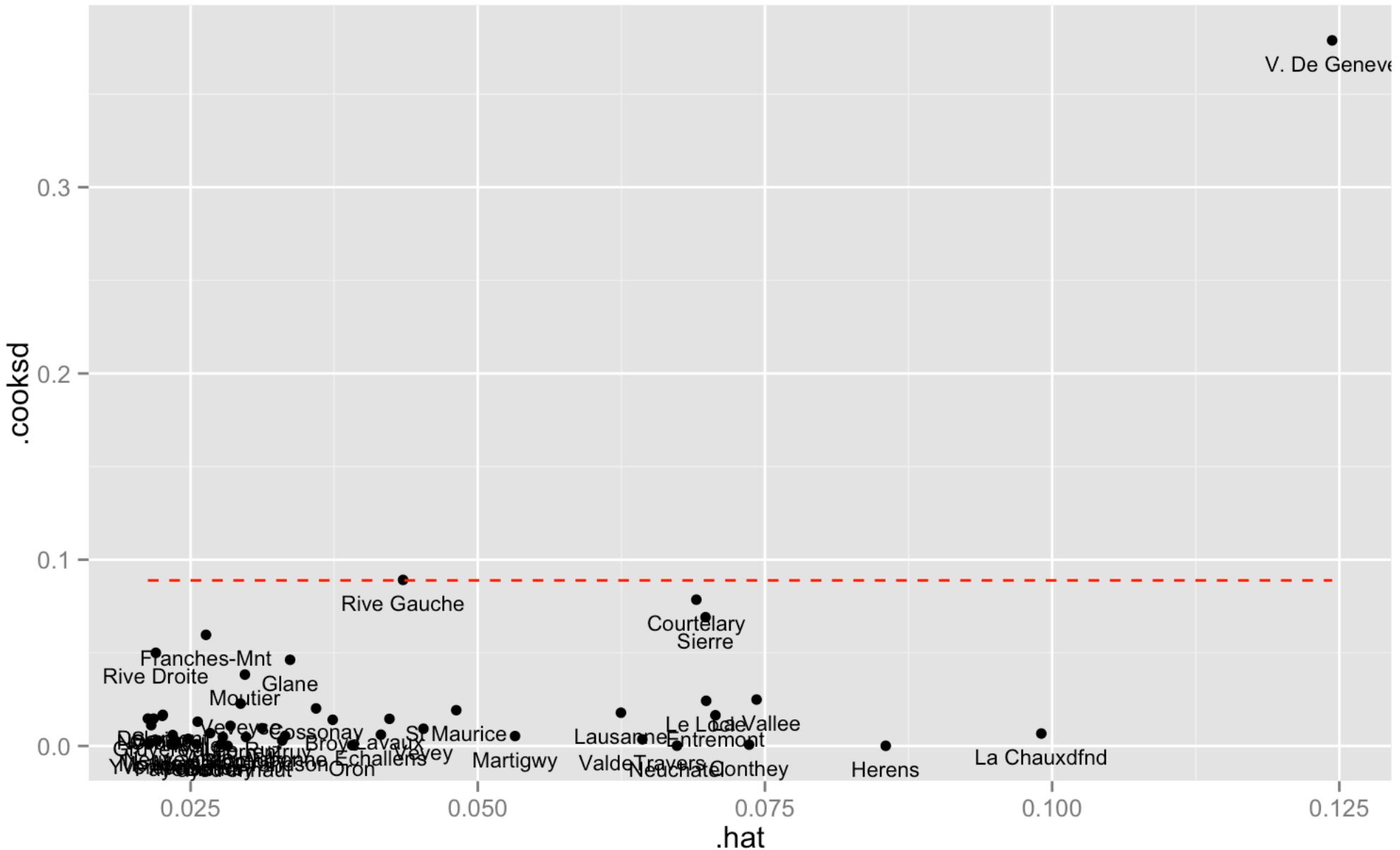
- Influence on regression line, measured by how much the regression line would change if the point were not included in the analysis.
- Cook's distance measures the influence of case i on *all n fitted values* y_i (not just the fitted value for case i as DFFITS)
- Cook's distance refers to how far, on average, *predicted y-values* will move if the observation in question is dropped from the data set
- With > 1 predictor, Cook's distance is presumably more important to you if you are doing predictive modeling, whereas dfbeta is more important in explanatory modeling.

Cook's distance

```
n <- nrow(slr_vars)
k <- 1 # predictors (not including intercept), number regressors
d <- 4/(n - k - 1)
d
[1] 0.08888889
slr_vars %>%
  filter(.cooksdi > d)
  .rownames fertility agriculture .fitted .se.fit   .resid      .hat   .sigma   .cooksdi .std.resid
1 V. De Geneve     35.0       1.2 60.53742 4.167274 -25.53742 0.12437740 11.21922 0.37885147 -2.309602
2 Rive Gauche     42.8       27.7 65.68376 2.464307 -22.88376 0.04349378 11.41733 0.08914847 -1.980169

## plot leverage versus cooks d
ggplot(slr_vars, aes(x = .hat, y = .cooksdi)) +
  geom_point() +
  geom_text(aes(label = .rownames), size = 3, vjust = 2) +
  geom_line(aes(y = d), colour = "red", lty = "dashed")
```

Cook's D



Specific measures of influence

- How each individual regression coefficient is changed by deleting each case from the dataset
- Number of DFBETAS is the number of predictors in your model
- Also for the intercept- generally not interesting one bit!
- Recommendation: I would look at DFBETAS after Cook's to isolate which x 's might be unduly influencing your overall regression equation
- How to: when we do multiple regression...

Summing up diagnostics

Influence	Recommended index	R	broom::augment ?
Leverage (extreme on x)	Hat values	<code>hatvalues(model)</code>	<code>.hat</code>
Discrepancy (extreme on y)	Externally studentized residual	<code>rstudent(model)</code>	Sadly no
Global influence	Cook's d	<code>cooks.distance(model)</code>	<code>.cooks.d</code>
Specific influence	DFBETAS	<code>dfbetas(model)</code>	Sadly no