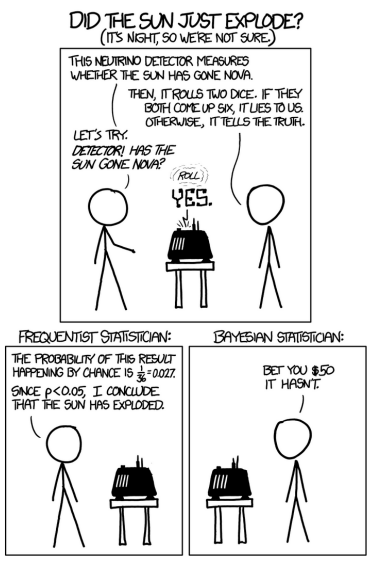
Econometrics I

Lecture 4: Inference and Standard Errors

Chris Conlon

Fall 2025



Source: xkcd.com

Hypothesis testing

- We are often interested in testing theories, or testing hypotheses about the values of certain parameters
- Simplest example: testing whether mean of a variable $\mu_x \equiv E[X]$ is different from a particular value:

$$\begin{aligned} H_0 : &\quad \mu_x = a \\ H_1 : &\quad \mu_x \neq a \end{aligned}$$

- A hypothesis test typically involves a **null hypothesis** and **alternative hypothesis**. The alternative hypothesis could also be about a particular value ($H_1 : \mu_x = b$, or a one-sided rejection of the null ($H_1 : \mu_x > a$).

Review: z test

- If X_i is i.i.d. normal with *known* variance σ^2, then

$$\overline{X} \sim \mathcal{N}\left(\mu_x, \sigma^2/n\right)$$

- In this case, we know the distribution of our estimate $\overline{X}$. We can test

$$H_0 : \mu_x = a \qquad H_1 : \mu_x \neq a$$

  using a **z test**.
- We construct the test statistic

$$z = \frac{\overline{X} - a}{n^{-1/2}\sigma}$$

  which under the null hypothesis has the standard normal distribution:

$$z \sim \mathcal{N}(0, 1)$$

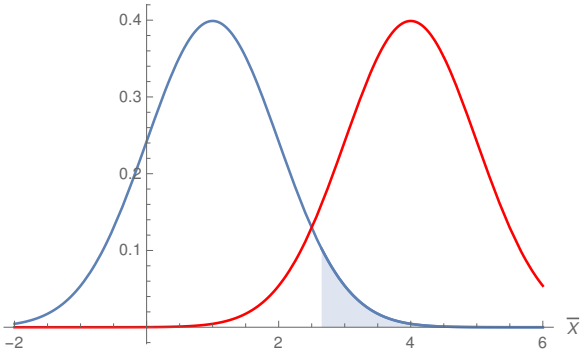
Level and Size of Test

- The **size** (or **level**) of a test is the probability of rejection if the null hypothesis is true. *The size is the rate of false positives or type I errors.*
- When hypothesis testing, we make it hard to reject the null hypothesis. We typically choose the size of the test to be small (most commonly, .01 or .05).

Power of Test

- We typically want to reject only for the outcomes that are least likely under the null hypothesis (or relatively more likely under the alternative hypothesis than the null). For the z test above, we reject only in the tails of the normal distribution. See: Neyman-Pearson Lemma.
- Choosing the rejection region appropriately maximizes the test's **power**, the probability of rejecting the null hypothesis when it is indeed false. Power is often harder to quantify and not something we typically choose. Power is one minus the rate of type II errors (**false negatives**), or failures to reject the null hypothesis when it is false.

Rejection Region and Power

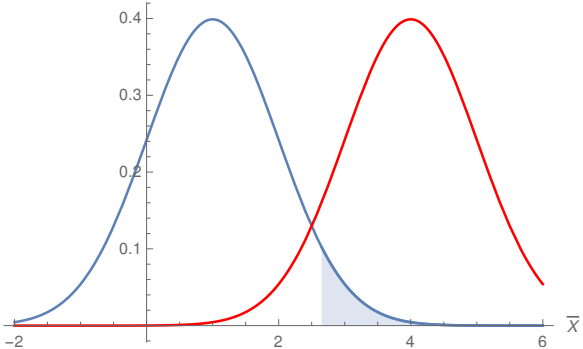


- Suppose $\bar{X}$ is normally distributed with $Var\left(\bar{X}\right) = 1$
- We want to test

$$H_0: \quad \mu_X = 1$$
$$H_1: \quad \mu_X = 4$$

- The blue and red lines are the PDF of $\bar{x}$ under the null and alternative hypotheses, respectively
- The shaded region is the rejection region with level $\alpha = .05$ that maximizes power. Note that this is for $\bar{X} \geq 2.65$.

Rejection Region and Power



- Note that the rejection region is the region where the PDF of the alternative hypothesis is high relative to the null hypothesis.
- The maximum power test with level .05 is the test that rejects for the 5% of the null-hypothesis PDF in which H_1's likelihood (probability density) is highest relative to H_0's.
- We often take for granted that rejection regions are in the tails of the null-hypothesis PDF; this is why.

Review: t Statistics

- Let's return to testing the value of a normally distributed random variable's mean, but now let's suppose that σ^2 is not known (which is typically the case).
- Our test statistic instead is

$$t = \frac{\overline{X} - a}{n^{-1/2} s}$$

where

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}$$

- Here, t has a t-distribution with $n - 1$ degrees of freedom.

Testing Paradigm

- We focus on different versions of **Wald tests**, which are based on test statistics that are (approximately) normally distributed.
- Other paradigms:
  - Likelihood Ratio tests and goodness-of-fit-based tests. The idea here is to compare how well different models fit the data.
  - Lagrange multiplier test: for example, testing whether residuals from a restricted model are correlated with excluded variables.

Motivating Small Sample t-Tests

- Last week we learned that if N is large then,

$$\mathbf{b}_{OLS} \overset{a}{\sim} \mathcal{N}\left(\beta, Var(\mathbf{b}_{OLS})\right)$$

  - This hinges on knowing $Var(\mathbf{b}_{OLS})$
  - We rarely know this in practice — we estimate it instead
  - Like testing the mean of a normal random variable, estimating the variance of the test statistic puts us in a t-test situation.

t-Statistics for OLS Parameters (homoscedasticity)

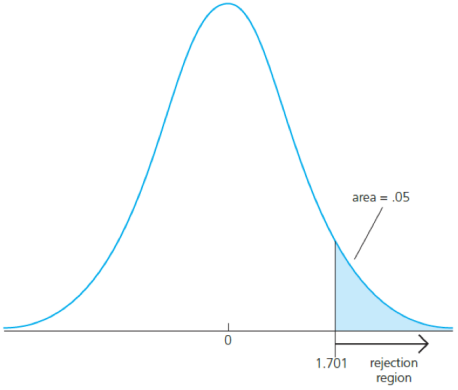
$$\frac{\mathbf{b}_{OLS,k} - \beta_k}{\sqrt{s^2 \, (\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim t_{n-K}$$

- where K is the number of parameters, s^2 is the estimator of the variance of ε, and

$$(\mathbf{X}'\mathbf{X})^{-1}_{kk}$$

  refers to the kth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.
- Note that the denominator of the above formula is the standard error for the kth estimated parameter $\mathbf{b}_{OLS,k}$.

The t-Distribution



- Similar to the $\mathcal{N}(0,1)$ but parametrized by degrees of freedom
- The tails are fatter but become $\mathcal{N}(0,1)$ as df go to ∞

Example of Reading a t-Table

Example of a table of critical values for t distribution from a textbook:

Degrees of Freedom	.10	.05
1	6.31	12.71
2	2.92	4.30
⋮		
28	1.70	2.05
⋮		
∞	1.65	1.96

- If N were very large we would use the $\mathcal{N}(0,1)$ approximation which is exactly the case that $df = \infty$
- If $N < \infty$ we can use a table like this, or a computer does it for us
- *Example:* If $N = 30$, $K = 2$, then $df = N - K = 28$ the 5% cutoff value is 2.05

An Example (Bivariate Regression)

Suppose I have the following estimated parameters on 30 observations

$$b_1 = 1.00$$

$$\sum_{i=1}^{N}(X_i - \bar{X})^2 = 14$$

$$\sum_{i=1}^{N} e_i^2 = 100$$

1. First, state the hypothesis:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Joint hypotheses

- Sometimes we want to test multiple parameters:

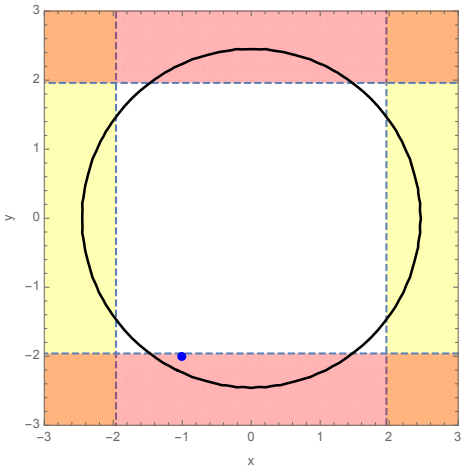
$$H_0: \quad \beta_{exp} = 0 \quad \text{AND} \quad \beta_{exp2} = 0$$

$$H_1: \quad \beta_{exp} \neq 0 \quad \text{OR} \quad \beta_{exp2} \neq 0$$

- Note that we do not want to do two separate t-tests for this hypothesis.

Illustration of Two t-Tests Failing

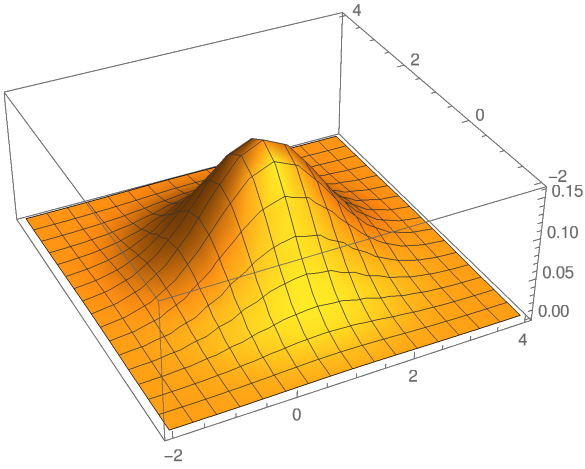
Suppose we have a large sample and t-statistics are -1 and -2. Do we reject null?

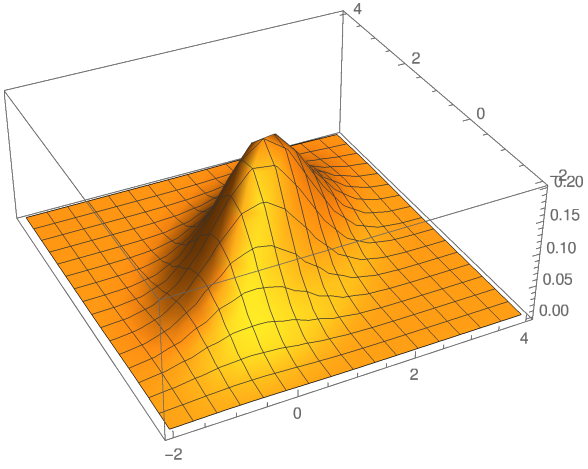


If the t's are independent:

- The circle contains 95% of the probability for two independent t-statistics; the area outside it is the rejection rejoin for the joint t-test.
- The dashed lines are the rejection regions for each of the individual t-tests. (5% level)
- Even though naively would reject, in actuality *not significant*
- What happens with correlated normal RVs?

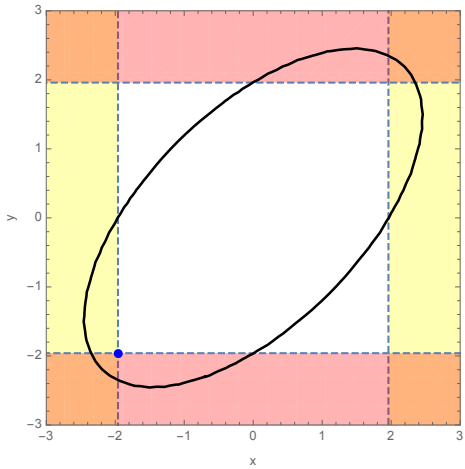
Bivariate normal: independent and correlated





T-Tests with correlation

If the t's are correlated this is the picture:



- Now, the area outside the ellipse is the rejection region for the joint t-test (5% level)
- The dashed lines are the rejection regions for each of the individual t-tests. (5% level)
- Now, notice that even with $t_1 = -2$, $t_2 = -2$, which would be a rejection according to each of the individual tests, is not a rejection of the joint test.

Correcting for Correlation: The F-Test

The issues we have are:

1. Testing a joint hypothesis with independent tests will not give the correct type 1 error
2. Correlated $\hat{\beta}$'s make things very messy

How can we solve this?

- First get a statistic that combines both hypotheses
  - Should be "big" when either t_1 or t_2 or both are big
  - Should include both t's
- Natural candidate:

$$F = t_1^2 + t_2^2$$

  - Always positive and only big when t's are big
  - If t_1 and t_2 are independent normals, then $F \sim \chi^2_2$
  - If we divide by 2 we have F_2 distribution

Correcting for Correlation: The F-Test, Cont'd

Our candidate test:

$$\frac{1}{2} \times (t_1^2 + t_2^2)$$

- Has a well understood distribution *when* t's are independent
- If not, we can *rotate* the t's so they are
  - Non-matrix formula (for 2 parameters):

$$F = \frac{1}{2} \times \frac{t_1^2 + t_2^2 - 2\rho_{t_1,t_2} t_1 t_2}{1 - \rho_{t_1,t_2}}$$

  - Matrix version (for k parameters):

$$\hat{\beta} - \beta \sim \mathcal{N}\left(0, \Sigma_{\hat{\beta}}\right) \Rightarrow \Sigma_{\hat{\beta}}^{-1/2} \times \left(\hat{\beta} - \beta\right) \sim \mathcal{N}(0, I)$$

This implies,

$$(\hat{\beta} - \beta)' \Sigma^{-1} (\hat{\beta} - \beta)/k \sim \chi_k^2/k = F_k$$

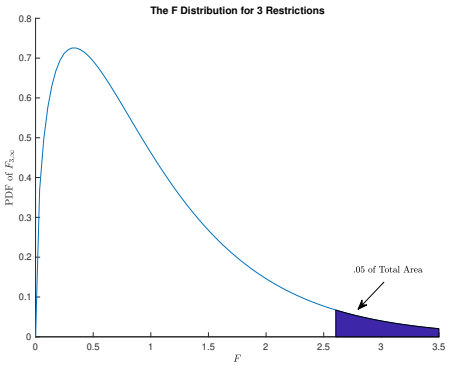
What is the F Distribution?

New test statistic:

$$F = \frac{1}{2} \times \frac{t_1^2 + t_2^2 - 2\rho_{t_1,t_2} t_1 t_2}{1 - \rho_{t_1,t_2}}$$

- Almost always requires a computer
- Ugly formula that follows a simple distribution
- In general, for q restrictions, we will calculate the F statistic and it will be distributed F_q ($F_{q,\infty}$ sometimes)
- Related to take the sum of squared normal random variables
- Critical values will depend on the number of restrictions
- Fun fact: for 1 restriction $F = t^2$

Critical Values of the F



- The distribution looks different than the *t*
- But the testing procedure is the same!
  - Find a critical value so that $P(F > cv) = .05$
  - If *F* is large *given the null* then null is unlikely to be true
  - Critical value depends on number of restrictions, *q*

F-tests: General Definition

- We are interested in testing the following linear restrictions on the parameters:

$$\mathrm{R}\boldsymbol{\beta} = \mathrm{q},$$

  where usually $\mathbf{q} = \mathbf{0}$, but not always.
- What would $\mathbf{R}$ and $\mathbf{q}$ be if we were testing whether two slopes were equal?
- The F statistic (or feasible Wald statistic):

$$F = \frac{(Rb_{OLS} - q)' \left\{ R \left[s^2 (X'X)^{-1} \right] R' \right\}^{-1} (Rb_{OLS} - q)}{J},$$

  which has a $F[J, n-K]$ distribution, where J is the number of rows of $\mathbf{R}$ (the number of restrictions).

F-tests: equivalent definitions

$$F = \frac{(Rb_{OLS} - q)' \left\{ R \left[s^2 (X'X)^{-1} \right] R' \right\}^{-1} (Rb_{OLS} - q)}{J},$$

- We could also write:

$$F = \frac{SSE_{CLS} - SSE_{OLS}}{Js^2},$$

  where SSE_{OLS} is the sum of squared residuals for the (unrestricted) OLS estimator, SSE_{CLS} is the sum of squared residuals under the **constrained least squares** estimator with constraints $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$.
- We are not going to delve into constrained least squares, but the point here is that there are two equivalent ways to think about the F-statistic: (1) as a Wald statistic, which is based on comparing an asymptotically normal parameter estimate to a hypothesized value, and (2) as an assessment of how much better the unconstrained model fits than the constrained model.

Non-Nested Models

- We have considered only nested models thus far. When testing

$$R\beta = q,$$

  we are testing a restricted linear model against alternative hypothesis of an unrestricted linear model, *which includes the restricted model as a special case*.

- Sometimes we want to compare non-nested models, which brings us to model selection. The main idea is to balance the model's goodness of fit and number of parameters: see adjusted R^2, **Akaike Information Criterion**, **Bayesian Information Criterion**. Machine learning approaches typically try to compare models by directly assessing out-of-sample performance. More on this stuff later and/or next semester.

OLS standard errors

- A useful identity for linear algebra:

$$\text{Var}(a\mathbf{Z}) = a^2 \text{Var}(\mathbf{Z})$$
$$\text{Var}(A\mathbf{Z}) = A \text{Var}(\mathbf{Z})A'$$

- Since $\mathbf{b}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$,

$$\text{Var}(\mathbf{b}_{OLS}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \text{Var}(\mathbf{y}|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

- Recalling that $\text{Var}(\mathbf{y}|\mathbf{X}) = \text{Var}(\varepsilon|\mathbf{X})$,

$$\text{Var}(\mathbf{b}_{OLS}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \text{Var}(\varepsilon|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

OLS standard errors

$$\text{Var}(\mathbf{b}_{OLS}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\text{Var}(\varepsilon|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

- With homoscedasticity, $\text{Var}(\varepsilon|\mathbf{X}) = \sigma^2\mathbf{I}$, and this simplifies to

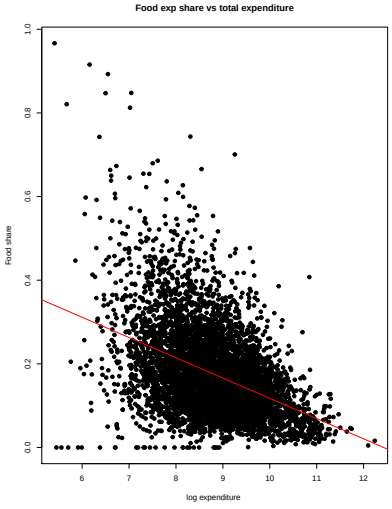
$$\text{Var}(\mathbf{b}_{OLS}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- Without homoscedasticity, we can still use the “sandwich” formula above.

Example: Engel Curves

- Engel curves refer to the relationship between a household's expenditure share on a good and income (or total expenditure).
- Engel curves for food are typically downward sloping – as total expenditure of a household increases, the proportion of its expenditure dedicated to food falls.
  - Expenditure on food still rises as total expenditure rises, but less than proportionally, so that food's expenditure *share* falls.

Food Engel Curves



- If we plotted total food expenditure (rather than the expenditure share), the heteroscedasticity would go in the other direction.

Data source: BLS Consumer Expenditure Survey data

Heteroscedasticity Robust Standard Errors I

- It is common to compute Eicker-Huber-White standard errors, which is a different estimator of $\mathbf{\Sigma}$ that is consistent even if each observation has a different variance σ_i^2:

$$(X'X)^{-1}\left(X'\text{diag}\left(e_1^2, e_2^2, \ldots, e_n^2\right)X\right)(X'X)^{-1}$$

- Statistical software typically makes it easy to use this estimator for $\mathbf{\Sigma}$ instead of the standard homoscedastic estimator.
- Heteroscedasticity-robust standard errors still perform well even if homoscedasticity holds, so there's little reason to even assume homoscedasticity when computing standard errors.

Heteroscedasticity Robust Standard Errors II

- We can rewrite the **heteroscedasticity-consistent** (or **heteroscedasticity-robust**) standard error estimator as:

$$n^{-1}\left(n^{-1}\mathbf{X}'\mathbf{X}\right)^{-1}\left(n^{-1}\mathbf{X}'\operatorname{diag}\left(e_1^2, e_2^2, \ldots, e_n^2\right)\mathbf{X}\right)\left(n^{-1}\mathbf{X}'\mathbf{X}\right)^{-1},$$

where the middle piece of this "sandwich" estimator can be written as

$$n^{-1}\sum_i \mathbf{x}_i\mathbf{x}_i' e_i^2$$

- Notice that this is the sample analog of $V\left[\mathbf{x}_i\varepsilon_i\right]$. What's going on with the robust standard error formula is we're constructing an estimate of

$$n^{-1}E\left[\mathbf{x}_i\mathbf{x}_i'\right]^{-1}V\left[\mathbf{x}_i\varepsilon_i\right]E\left[\mathbf{x}_i\mathbf{x}_i'\right]^{-1}.$$

Doing a Heteroscedastic F-Test in R

Let us revisit the wage equation:

$$\log(Wage_i) = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Educ_i + \varepsilon_i$$

- New question: does experience/age matter at all?
- New hypothesis:

$$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0$$
$$H_a : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0$$

- How do we test in R using heteroscedastic robust standard errors?

We need a new command:

- First, we need the `car` package

Doing a Heteroscedastic F-Test in R, Cont'd

Let us revisit the wage equation:

$$\log(Wage_i) = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Educ_i + \varepsilon_i$$

```
> m1 = lm(formula = lwage~age+age2+education, data=wageData)
> linearHypothesis(m1, c("age =0", "age2=0"), vcov = vcovHC(m1, type = "HC1"))
Linear hypothesis test

Hypothesis:
age = 0
age2 = 0

Model 1: restricted model
Model 2: lwage ~ age + age2 + education

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1   5357
2   5355  2 335.03 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Red box is the name of the model
- Green box is the list of hypotheses:
  - List enclosed by the `c()` command
  - Each restriction is enclosed in quotes, one equal sign and uses the names of the variables from the model
  - Don't forget to separate commands with commas
- Purple box is the variance-covariance argument

Command without boxes:

Outliers

- **Outliers** refer to observations that are “far away” from the rest of the data. They can be due to errors in the data. There is no standard formal definition.
- What to do? Greene:“*It is difficult to draw firm general conclusions... It remains likely that in very small samples, some caution and close scrutiny of the data are called for.*” I’d say that’s true even in large samples, but there isn’t a generally accepted way of quantifying what counts as appropriate “caution and close scrutiny.”

Removing Outliers?

- Removing extreme outliers (in x) from datasets is often considered good practice. But we should be mindful about why as dropping observations creates the potential for manipulation.
- Sometimes extreme outliers are just errors, in which case they should almost certainly be dropped.
- Even if they aren't errors, they may reflect a different mode in the data generating process. They may require a different or more general model to account for them properly. Consider the justificaiton of a linear model based on Taylor's theorem (local linear approximation). With such a justification for your modeling strategy, it would not make sense to include an outlier in x.
- It's important to be transparent about how dropped outliers affect results.

Outliers and Leverage

- One way to find influential observations is to calculate the **leverage** of each observation *i*. We begin with the hat matrix:

$$P = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

and consider the diagonal elements, which are labeled h_{ii}

$$h_{ii} = \mathbf{x_i}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}$$

- This tells us how influential an observation is in our estimate of $\mathbf{b}_{OLS}$. Particularly important for $\{0, 1\}$ dummy variables with uneven groups.

Leave One Out Regression

- This is sometimes called the **Jackknife**
- Sometimes it is helpful to know what would happen if we omitted a single observation i
- Turns out we don't need to run N regressions

$$\begin{aligned}\mathbf{b}_{-i} &= (\mathbf{X}'_{-i}\mathbf{X}_{-i})^{-1}\mathbf{X}'_{-i}\mathbf{y}_{-i} \\ &= \mathbf{b}_{OLS} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\tilde{e}_i \quad \text{where } \tilde{e}_i = (1-h_{ii})^{-1}e_i\end{aligned}$$

- $\tilde{e}_i$ has the interpretation of the LOO prediction error.
- high leverage observations move $\mathbf{b}_{OLS}$ a lot.

Heteroskedasticity Consistent (HC) Variance Estimates

What we need is a consistent estimator for e_i^2.

$$\mathbf{V}_{OLS}^{HC0} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'e_i^2\right)(\mathbf{X}'\mathbf{X})^{-1}$$

$$\mathbf{V}_{OLS}^{HC1} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'e_i^2\right)(\mathbf{X}'\mathbf{X})^{-1}\cdot\left(\frac{n}{n-k}\right)$$

Could use $\tilde{e}_i$ instead of e_i for a better estimate

$$\mathbf{V}_{OLS}^{HC2} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{N}(1-h_{ii})^{-1}\mathbf{x}_i\mathbf{x}_i'e_i^2\right)(\mathbf{X}'\mathbf{X})^{-1}$$

$$\mathbf{V}_{OLS}^{HC3} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{N}(1-h_{ii})^{-2}\mathbf{x}_i\mathbf{x}_i'e_i^2\right)(\mathbf{X}'\mathbf{X})^{-1}$$

Heteroskedasticity Consistent (HC) Variance Estimates

- We know that $\mathbf{V}_{OLS}^{HC3} > \mathbf{V}_{OLS}^{HC2} > \mathbf{V}_{OLS}^{HC0}$ because $(1 - h_{ii}) < 1$.
- `Stata` uses *HC*1 as the default and it is what most people refer to when they say **robust standard errors**. These are the Eicher-Huber-White SE's.
- Note: failure to correct for heteroscedasticity can lead to size distortions (a rate of type I errors that differs from what you intend).
- *HC*3 are the most conservative and also place the most weight on potential outliers.

Heteroskedasticity Consistent (HC) Variance Estimates

To read about SE's in `estimatr`:

https://declaredesign.org/r/estimatr/articles/mathematical-notes.html

```
dat <- data.frame(X = matrix(rnorm(2000*5), 2000), y = rnorm(2000))
hc0<-lm_robust(y ~ ., data = dat, se_type="HC0")$std.error
hc1<-lm_robust(y ~ ., data = dat, se_type="HC1")$std.error
hc2<-lm_robust(y ~ ., data = dat, se_type="HC2")$std.error
hc3<-lm_robust(y ~ ., data = dat, se_type="HC3")$std.error
all(hc2 > hc0 )
[1] TRUE
all(hc3> hc2 )
[1] TRUE
```

Heteroscedasticity vs. Correlation

- Recall that we defined the homoscedasticity assumption as:

$$Var(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$$

  this assumption has two aspects:

  1. The disturbance for each observation has the same variance
  2. Imposing zero correlation between disturbances for different observations

- The terminology can be misleading here, because what people refer to as "heteroscedasticity-robust" standard errors (the variance estimators on the previous slide) are robust to violations of 1 but not 2.
- We need to do a bit more to estimate standard errors in a way that is robust to correlated data.

Correlation I

- The baseline assumptions of the linear regression framework imply that the disturbances are uncorrelated across observations. There are many ways for this to be violated.
  - Example 1: we might have county-level data for a regression and be concerned that different counties within a given state have correlated disturbances because all counties are subject to the same (unobserved) state-level policies.
  - Example 2: time series data (asset prices), and we are worried that some unobserved factors within the disturbances are serially correlated
  - Example 3: county level data again, and we are worried about geographically correlated factors such as weather.

Correlation II

Different correlation patterns call for different estimators of $\mathbf{\Sigma}$, the variance of $\mathbf{b}_{OLS}$ Some common alternatives to the no-correlation baseline:

1. Clustered standard errors, when there is correlation between observations within well-defined groups, but no correlation between observations in different groups.
2. Newey-West standard errors (and extensions) to deal with serial correlation in time series data.
3. Conley-Newey-West standard errors that allow for correlation in multiple dimensions (especially popular in the context of spatially explicit models).

Clustering I

- Suppose data are organized into distinct groups $g = 1, 2, \ldots, G$. Let $g(i)$ be the group identity of observation i.
  - e.g., with county-level data, we have $g(\textit{Manhattan}) = \textit{NY}$.
- We assume $E\left[\varepsilon_i \varepsilon_j\right] = 0$ as long as $g(i) \neq g(j)$, and we do not restrict the correlation $E\left[\varepsilon_i \varepsilon_j\right]$ for observations within the same group.
- Intuition: the linear regression framework with no correlation in observations will overstate the precision of our estimates. If we add another observation within a cluster, and that observation is highly correlated with the other observations, it's not actually as good as adding another independent observation.

Clustering II

- Recall the sandwich formula for standard errors:

$$n^{-1}E\left[\mathbf{x}_i\mathbf{x}_i'\right]^{-1}V\left[\mathbf{x}_i\varepsilon_i\right]E\left[\mathbf{x}_i\mathbf{x}_i'\right]^{-1}.$$

- The estimator for the middle part without clustering was

$$\mathbf{V}\left[\mathbf{x}_i\varepsilon_i\right] = n^{-1}\sum_i \mathbf{x}_i\mathbf{x}_i'e_i^2$$

- With clustering, it will be

$$\mathbf{V}_{clu} = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{x}_i\mathbf{x}_j'e_ie_j\mathbf{I}\left[g\left(i\right) == g\left(j\right)\right]$$

where the $\mathbf{I}$ function is 1 when i, j come from the same group and zero otherwise.

Clustering III

- The cluster-robust estimate of standard errors will be consistent as the number of groups gets large.
- Note that this estimator adds extra terms (covariance terms) to the estimate of variance, so this is going to make standard errors larger as long as covariances $E\left[\varepsilon_i \varepsilon_j\right]$ are positive.
- Thus, if standard formulas are used in the presence of cluster-correlated disturbances, standard errors will be too small.
- Statistical software packages typically make it easy to compute cluster-robust errors.
- Clustering often makes a **huge** difference in standard errors.

Correlation III

Consider the cluster-robust estimator of the "meat" part of the sandwich estimator:

$$\mathbf{V}_{clu} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{x}_i \mathbf{x}_j' e_i e_j \mathbf{I}\left[g\left(i\right) == g\left(j\right)\right]$$

For Conley-Newey-West standard errors (where there is correlation between "nearby" observations), procedure is similar.

The difference is that instead of the 1/0 indicator function for $\mathbf{I}$, we will have a weighting (or kernel) function which takes on values ≈ 1 for "nearby" observations and goes to zero for observations that are far apart.

Clustered SE's

$$\widehat{V}_{OLS}^{\text{CR1}} = (X'X)^{-1}\left(\sum_{g=1}^{G} X_g' e_g e_g' X_g\right)(X'X)^{-1}$$

$$\widehat{V}_{OLS}^{\text{CR3}} = (X'X)^{-1}\left(\sum_{g=1}^{G} X_g' \tilde{e}_g \tilde{e}_g' X_g\right)(X'X)^{-1}$$

- Can replace $\mathbf{e}_g \longrightarrow \tilde{\mathbf{e}}_g$ for leave-one out like *HC*3 (these are called *CR*3).

Clustering in R

```
lm_robust(y~ x1 + x2, data=df, se_type="CR0", cluster=group_id )
lm_robust(y~ x1 + x2, data=df, se_type="CR2", cluster=group_id )
lm_robust(y~ x1 + x2, data=df, se_type="CR1", cluster=group_id )
```

Bootstrap I

- Another approach to estimating the standard errors of $\mathbf{b}_{OLS}$ is the **bootstrap**
- The basic idea:
  1. Simulate a new data set (same number of observations) by sampling (with replacement) from the original data set
  2. Estimate $\mathbf{b}_{OLS,s}$ for the new data set.
  3. Repeat lots of times, resulting in a bunch of different estimates of $\mathbf{b}_{OLS,s}$, say $s = 1, \ldots, 10000$
  4. Look at the variance of the $\mathbf{b}_{OLS,s}$ estimates across the various simulated data sets. This is your estimate of $\mathbf{\Sigma}$, or $Var(\mathbf{b}_{OLS})$

Bootstrap II

- The bootstrap's main appeal is that it can provide a better finite-sample approximation of the distribution of the parameter estimates.
  - Note that the Eicker-Huber-White standard errors estimates are *consistent*, but not generally *unbiased* in finite samples
  - The bootstrap is probably worth trying if you're ever working with non-linear estimators (which can be consistent but are typically not unbiased in finite samples).
- Also, it can potentially deliver good estimates of standard errors even with correlated errors, but this depends on the version of the bootstrap (see **block bootstrap**). Exploring formally the conditions under which the bootstrap works well is beyond our scope.

Confidence Intervals I

- Note that if

$$\mathsf{b} \sim \mathcal{N}\left(\boldsymbol{\beta}, \Sigma\right)),$$

then

$$b_k \sim \mathcal{N}\left(\beta_k, \Sigma_{kk}\right),$$

and

$$Pr\left[b_k - z_{(1-\alpha/2)}\sqrt{\Sigma_{kk}} \leq \beta_k \leq b_k + z_{(1-\alpha/2)}\sqrt{\Sigma_{kk}}\right] = \alpha$$

where $z_{(1-\alpha/2)}$ is the value such that the CDF of the standard normal distribution is $1 - \alpha/2$.

Confidence Intervals II

- Similarly, when

$$\frac{b_k - \beta_k}{\sqrt{\hat{\Sigma}_{kk}}} \sim t_{n-K}$$

because the variance Σ_{kk} has to be estimated, then

$$Pr\left[b_k - t_{(1-\alpha/2),n-K}\sqrt{\hat{\Sigma}_{kk}} \leq \beta_k \leq b_k + t_{(1-\alpha/2),n-K}\sqrt{\hat{\Sigma}_{kk}}\right] = \alpha$$

where $t_{(1-\alpha/2),n-K}$ is the value such that the CDF of the t-distribution with $n - K$ degrees of freedom is $1 - \alpha/2$.

Confidence Intervals III

- We define the $1-\alpha$ **confidence interval** for b_k as

$$\left(b_k - t_{(1-\alpha/2),n-K}\sqrt{\hat{\Sigma}_{kk}}, b_k + t_{(1-\alpha/2),n-K}\sqrt{\hat{\Sigma}_{kk}}\right)$$

- Note that this confidence interval is a function of the data – the end points of the confidence interval are statistics and therefore random variables in their own right.
- Defining the confidence interval in this way, the probability that the confidence interval contains the true parameter is $1-\alpha$ (if the asymptotic distribution of the estimator is taken as the true distribution). That is, if $\alpha = .05$, this is called a 95% confidence interval, and there is a 95% chance it will contain the true parameter (assuming that the asymptotic approximation holds).

Summary

- Linear regression theory gives us formulas for estimating $Var(\mathbf{b}_{OLS})$
- We can use that variance estimator to test hypotheses about parameters (using t-Tests and f-Tests) as well as construct confidence intervals.
- When the baseline assumptions of the linear regression model are violated (due to correlation or heteroscedasticity), we need to use somewhat more complex formulas to estimate $Var(\mathbf{b}_{OLS})$.