Session 5: Inference

MGT 581 | Introduction to econometrics

Michaël Aklin

PASU Lab | EPFL

Last time...

- Model specification

Today:

- Inference
- Hypothesis testing
- Confidence intervals

Readings:

- Stock and Watson (2011) ch5, 7
- Verbeek (2018) 2.4-2.6, 4

Inference

- So far: we focused on **point estimates**: scalars (or vectors) that captures our best guess of the quantity of interest
- But this doesn't tell us anything about **uncertainty** regarding this number
- At the very least: sampling variation alone means that point estimates will vary from one sample to another
- Also: point estimates don't tell us about the plausibility of a hypothesis
- Eg: does my new tax software reduce time to fill taxes? What if my $\widehat{\beta} = -10$minutes?
- We will always assume that sampling is **IID**: all observations are drawn **independently** from an **identical distribution**

Inference

- There exist several (sometimes related) approaches to inference
- Assessing uncertainty via **standard errors**
- Using standard errors to conduct **hypothesis tests**
- Computing **p values**
- Computing **confidence intervals**
- These are all connected to each other

How we will proceed...

1. We will figure out the standard errors of $\hat{\beta}$: the standard deviation of the sampling distribution of $\hat{\beta}$
2. We will then examine how to conduct **hypothesis testing** (incl. p-values)
3. We will discuss **confidence intervals**

Standard errors

- What we want: find the distribution of $\hat{\beta}_{ols}$.
- Two approaches: **finite sample** (fixed N) and **asymptotic** $(n \to \infty)$
- **Finite sample** requires additional assumptions...
- And in most cases, we work with large samples (say, $>$100)
- Thus: let's think about asymptotic distribution of $\hat{\beta}$

Asymptotic

- Asymptotics: we are interested in understanding the distribution of $\hat{\beta}$ as N tends to infinity
- As we saw earlier: $\hat{\beta}$ converges to a constant β if consistency is met
- Thus: not very helpful for infinite samples
- Instead: use a formulation of $\hat{\beta}_{ols}$ whose distribution we know
- One such case: $\sqrt{N}(\hat{\beta} - \beta) \sim N(0, \sigma^2)$ by the Central Limit Theorem (CLT)
- Let's derive this!

$$
\begin{aligned}
Var[\sqrt{N}(\hat{\beta}-\beta)] &= Var\left[E[X_i X_i']^{-1}\frac{1}{\sqrt{N}}\sum_i^n X_i\varepsilon_i\right] \\
&= E[X_i X_i']^{-1} Var\left[\frac{1}{\sqrt{N}}\sum_i^n X_i\varepsilon_i\right] E[X_i X_i']^{-1} \\
&= E[X_i X_i']^{-1} E[X_i X_i'\varepsilon_i^2] E[X_i X_i']^{-1} \\
&\equiv \Omega
\end{aligned}
$$

Then:

$$\sqrt{N}(\hat{\beta}-\beta) \sim N(0,\Omega)$$

Or:

$$\hat{\beta} \sim N\left(\beta, \frac{\Omega}{N}\right)$$

- Where do we go from here?
- By the **Central Limit Theorem**, we have:

$$\sqrt{N}(\hat{\beta} - \beta) \sim N(0, \Omega)$$

- Note that:

$$\begin{aligned} \sqrt{N}(\hat{\beta} - \beta) &\sim N(0, \Omega) \\ (\hat{\beta} - \beta) &\sim \frac{N(0, \Omega)}{\sqrt{N}} \\ \hat{\beta} &\sim N(\beta, \Omega/N) \end{aligned}$$

- Thus: we know the asymptotic distribution of $\hat{\beta}$!
- Only need to find out Ω (i.e. variance of $\sqrt{N}(\hat{\beta} - \beta)$)

- We had:

$$Var[\sqrt{N}(\hat{\beta} - \beta)] = E[X_i X_i']^{-1} E[X_i X_i' \varepsilon_i^2] E[X_i X_i']^{-1}$$
$$\equiv \Omega$$

- Known as **variance-covariance** matrix
- Variance of $\hat{\beta}_k$: kth diagonal value of Ω
- Standard error: square root of the variance
- Let's dig a bit deeper...

- Key: $E[X_i X_i' \varepsilon_i^2]$. Let's ignore $X_i X_i'$ for a sec:

$$E[\varepsilon_i^2] = \begin{bmatrix} E[\varepsilon_1^2] & E[\varepsilon_1 \varepsilon_2] & ... & E[\varepsilon_1 \varepsilon_n] \\ E[\varepsilon_2 \varepsilon_1] & E[\varepsilon_2^2] & ... & E[\varepsilon_2 \varepsilon_n] \\ ... & ... & ... & ... \\ E[\varepsilon_n \varepsilon_1] & ... & ... & E[\varepsilon_n^2] \end{bmatrix}$$

- This is the variance-covariance of residuals. Why "var-cov"?
- Recall that $E[\varepsilon_i] = 0$. Thus: $V(\varepsilon_i) \equiv E[(\varepsilon_i - \bar{u})^2] = E[\varepsilon^2]$.
- Likewise: $Cov(\varepsilon_i, \varepsilon_j) \equiv E[(\varepsilon_i - \bar{u})(\varepsilon_j - \bar{u})] = E[\varepsilon_i \varepsilon_j]$.
- But: we don't observe var-cov $E[\varepsilon_i^2]$. Need to estimate it.
- Two types of assumptions sometimes made

1. Independence (no cor across error terms)
2. Homoskedasticity

1. No spatial/temporal correlation:

$$E[\varepsilon_i^2] = \begin{bmatrix} E[\varepsilon_1^2] & 0 & ... & 0 \\ 0 & E[\varepsilon_2^2] & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & ... & E[\varepsilon_n^2] \end{bmatrix}$$

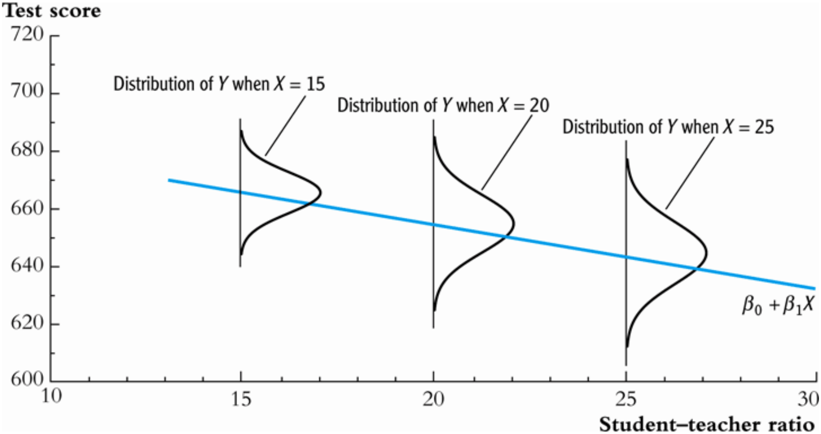
- This leads to Eicker-Huber-White (“robust”) standard errors (Huber 1967; White 1980)
- Concretely: variance of error terms can vary, but error terms aren’t correlated with each other
- Robust se: square each residual as estimate of $E[\varepsilon_i^2]$

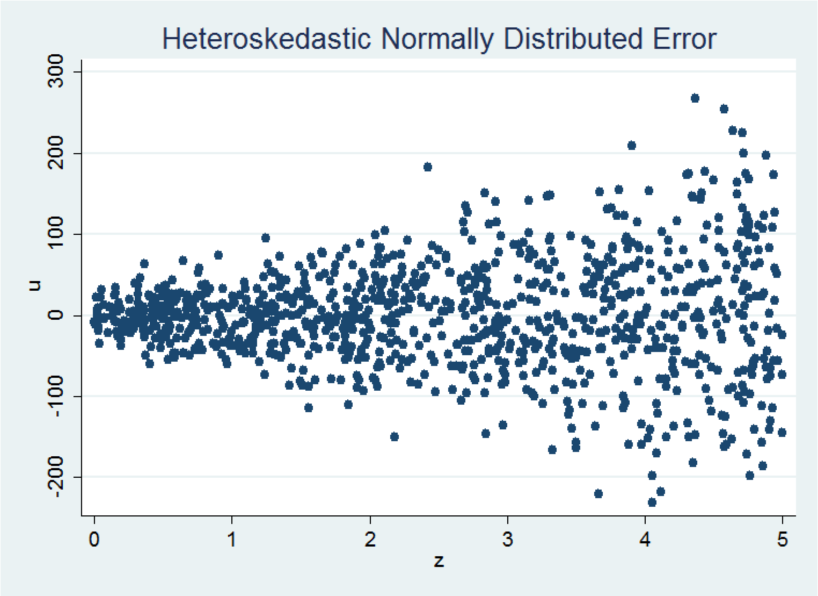
2. Homoskedasticity: variance of error terms is a constant (call it σ^2)

- Recall that $E[\varepsilon] = 0$... In that case: $E[\varepsilon^2] = Var(\varepsilon)$
- Then:

$$E[\varepsilon_i^2] = \begin{bmatrix} \sigma^2 & E[\varepsilon_1\varepsilon_2] & ... & E[\varepsilon_1\varepsilon_n] \\ E[\varepsilon_2\varepsilon_1] & \sigma^2 & ... & E[\varepsilon_2\varepsilon_n] \\ ... & ... & ... & ... \\ E[\varepsilon_n\varepsilon_1] & ... & ... & \sigma^2 \end{bmatrix}$$

- In practice: if you assume homoskedasticity, you generally assume no correlation
- $$E[\varepsilon_i^2] = \begin{bmatrix} \sigma^2 & 0 & ... & 0 \\ 0 & \sigma^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & ... & \sigma^2 \end{bmatrix}$$
- Nice if it happens (more in one sec)!
- But very unlikely


Test score
720
700
680
660
640
620
600
Distribution of Y when $X = 15$
Distribution of Y when $X = 20$
Distribution of Y when $X = 25$
$\beta_0 + \beta_1 X$
10
15
20
25
30
Student–teacher ratio


Heteroskedastic Normally Distributed Error
u
300
200
100
0
-100
-200
z
0
1
2
3
4
5

Nice! The variance becomes

$$
\begin{aligned}
Var[\sqrt{N}(\hat{\beta} - \beta)] &= E[X_i X_i']^{-1} E[X_i X_i' \varepsilon_i^2] E[X_i X_i']^{-1} \\
&= \sigma^2 E[X_i X_i']^{-1} E[X_i X_i'] E[X_i X_i']^{-1} \\
&= \sigma^2 E[X_i X_i']^{-1}
\end{aligned}
$$

Then we just have to estimate σ^2 with our observed residuals e_i:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k}$$

- *If* homoskedasticity and independence: opens the door of Gauss-Markov theorem
- Gauss-Markov: under assumptions of random sampling, correctly specified model, X is of full rank, exogeneity, homoskedasticity, independent error terms (no serial/spatial correlation), then OLS is the **b**est **l**inear **u**nbiased **e**stimator (BLUE)
- But: hard to make the assumption of homoskedasticity. Eg: non-linear relations create heteroskedasticity.
- That's why we often used "**robust**" standard errors (aka Eicker-Huber-White)

Model: $\text{Growth}_c = \alpha + \beta_1 \text{Schooling years}_c + \beta_2 \text{Trade (perGDP)}$

```
Call:
lm(formula = growth ~ yearsschool + tradeshare, data = growth)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4241 -0.9282 -0.2976  0.8468  5.5795 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.37015    0.56993  -0.649  0.51844   
yearsschool  0.25003    0.08286   3.018  0.00369 **
tradeshare   2.33129    0.72811   3.202  0.00216 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.685 on 62 degrees of freedom
Multiple R-squared:  0.2359,	Adjusted R-squared:  0.2113 
F-statistic: 9.571 on 2 and 62 DF,  p-value: 0.0002386
```

Figure 1: Homoskedastic standard errors

Model: $\text{Growth}_c = \alpha + \beta_1 \text{Schooling years}_c + \beta_2 \text{Trade (perGDP)}$

```
> model2 = lm_robust(data=growth, growth ~ yearsschool + tradeshare)
> summary(model2)

Call:
lm_robust(formula = growth ~ yearsschool + tradeshare, data = growth)

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
(Intercept)  -0.3702    0.59735 -0.6197 0.5377536 -1.56423   0.8239 62
yearsschool   0.2500    0.07621  3.2808 0.0017015  0.09769   0.4024 62
tradeshare    2.3313    0.63215  3.6879 0.0004781  1.06765   3.5949 62

Multiple R-squared:  0.2359 ,   Adjusted R-squared:  0.2113
F-statistic:  10.7 on 2 and 62 DF,  p-value: 0.000102
```

Figure 2: Heteroskedastic (Eicker-Huber-White) standard errors

Note that the standard error of the bivariate coefficient $\hat{\beta}$ with homoskedasticity can be written as:

$$se(\hat{\beta}) = \sqrt{\frac{\sigma_{\varepsilon}^2}{\sum(x - \bar{x})^2}}$$

Note: σ_{ε}^2 is the (unobservable) variance of the error term. We will see it again when we discuss model fit.

We can estimate it with:

$$s^2 = \frac{\sum(y - \hat{y})^2}{n - 2}$$

- Standard error decreases when e_i^2 decreases. Better model = smaller variance.
- Standard error decreases when sum of square X increases (denominator). More variance in treatments is good.
- Standard error decreases in n
- (Multiple regression) Standard error of $\widehat{\beta}_k$ increases when k is explained well by other variables (1st stage of Frisch-Waugh).

Hypothesis testing

- We now know how our estimate is distributed in large sample!
- Very useful for assessments regarding our causal effects
- We can engage in **hypothesis testing**
- Idea: can we say something about the plausibility of a hypothesis
- How: is our point estimate “far” from some hypothetical value?
- Careful: hypothesis testing is a very coarse instrument (and mildly controversial by now)

Step by step

1. Identify a null and a rival hypothesis

- Typically: my treatment has no effect: $\beta = 0$. But the benchmark could be something else (eg $\beta = 1$).
- This is the null hypothesis (H_0)
- Rival/alternative: my treatment has an effect $H_a : \beta \neq 0$ ("two-sided" test)
- Sometimes: my treatment has a positive (negative) effect ("one-sided" test). Eg: $H_a : \beta > 0$.
- This is the alternative (or rival) hypothesis (H_a or H_1)

2. Measure how 'far' from the null your point estimate is

- $t^* = \frac{\hat{\beta} - \beta_{H0}}{se(\hat{\beta})}$
- $se(\hat{\beta}) = \sqrt{\hat{\sigma}^2_{\beta}}$
- t has a *Student's t* distribution
- Note: since H_0 is generally $\beta = 0$, this simplifies to: $\frac{\hat{\beta}}{se(\beta)}$
- Tells you how many standard errors $\hat{\beta}$ is away from 0

FIGURE 5.1 Calculating the *p*-Value of a Two-Sided Test When $t^{act} = -4.38$

The *p*-value of a two-sided test is the probability that $|Z| > |t^{act}|$ where Z is a standard normal random variable and t^{act} is the value of the *t*-statistic calculated from the sample. When $t^{act} = -4.38$, the *p*-value is only 0.00001.

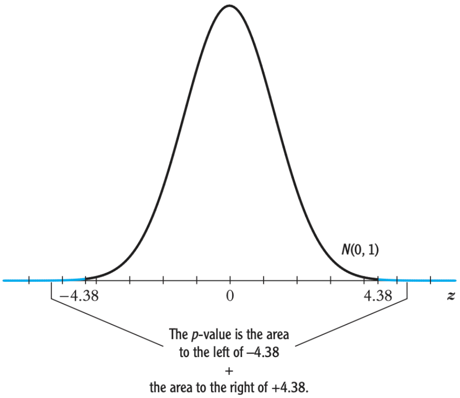


Figure 3: Example: $t^* = -4.38$

3. Select a critical value α at which you reject or not the null hypothesis.

- "How large does t^* need to be until I don't find the null hypothesis credible?"
- If it's very large: implausible.
- But: need a "critical" value. Typically: it must be so large that we only have an α probability of obtaining $\hat{\beta}$ by randomness.
- Compare t^* to critical t^c (value beyond which there is less than α of mass of probability distribution). Draw!
- Eg: if we set $\alpha = 0.05$: t^* is such that it would only be obtained 5% of the time, I deem my null hypothesis unlikely and **reject it**. $\hat{\beta}$ is "statistically significant".

Side note

- No rule as for ‘best’ value for α
- $\alpha \equiv Pr[\text{reject } H_0 | H_0 \text{ is true}]$
- Also called Type I error
- In contrast: Type II error: $Pr[\text{accept } H_0 | H_a \text{ is true}]$
- Useful concept: **power** of a text:
  $\pi = Pr(\text{reject } H_0 | H_a \text{ is true})$

4. **p value**: probability of getting a value of $|t^*|$ or greater.

- $p = Pr[|t| > |t^*|]$
- Small p: unlikely to get a such a t statistic by chance
- Thus: compare p to α. If smaller: **reject null hypothesis**. $\hat{\beta}$ is (statitically) significant.
- Note: comparing p to α is the same as comparing t to t^c

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Manually, in R

```
> qt(0.025, 1000, lower.tail=TRUE)
[1] -1.962339
```

```
> model = lm_robust(data=data_combined, pm25 ~ log(GDP_Per_Capita))
> summary(model)

Call:
lm_robust(formula = pm25 ~ log(GDP_Per_Capita), data = data_combined)

Standard error type:  HC2

Coefficients:
                    Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
(Intercept)           92.307     6.6385   13.90 1.063e-32   79.231  105.384 242
log(GDP_Per_Capita)   -7.309     0.7254  -10.08 3.631e-20   -8.738   -5.881 242

Multiple R-squared:  0.3113 ,   Adjusted R-squared:  0.3085
F-statistic: 101.5 on 1 and 242 DF,  p-value: < 2.2e-16
```

Note: $-7/0.72 = -10$

$p-$value attached to $|t| = 10$: 3.6e-20.

Thus: t^* is large, p is small: reject null hypothesis that $\beta = 0$.

- Null hypothesis testing responds to falsificationist approach: can you "reject" certain models (Popper)
- But: it is now highly criticized (Gelman 2013)
- Problems include...
- p-hacking: selecting models conditional on significance of parameter(s)
- Overly sharp statements: significant vs. non-significant (Gelman and Stern 2006)
- Economic vs. statistical significance

Confidence intervals

- CI: last way for now to assess uncertainty (Hoekstra et al. 2014)
- Different from hypothesis testing: we don't start from the idea that our model is wrong...
- Instead: we ask what is a plausible range around our point estimate $\hat{\beta}$?
- To do so: we can rely on t distribution centered around $\hat{\beta}$ (and not 0)

Formula

- (100-α)% CI $= \hat{\beta} \pm t^c_{\alpha/2} \times se(\hat{\beta})$
- Eg: 95% CI $= \hat{\beta} \pm 1.96 \times se(\hat{\beta})$
- Interpretation: in repeated sampling, this confidence interval will include the true value of β 95% of the time.
- NOT the same as: "β has a 95% chance of being in this interval." Either it is, or it isn't.

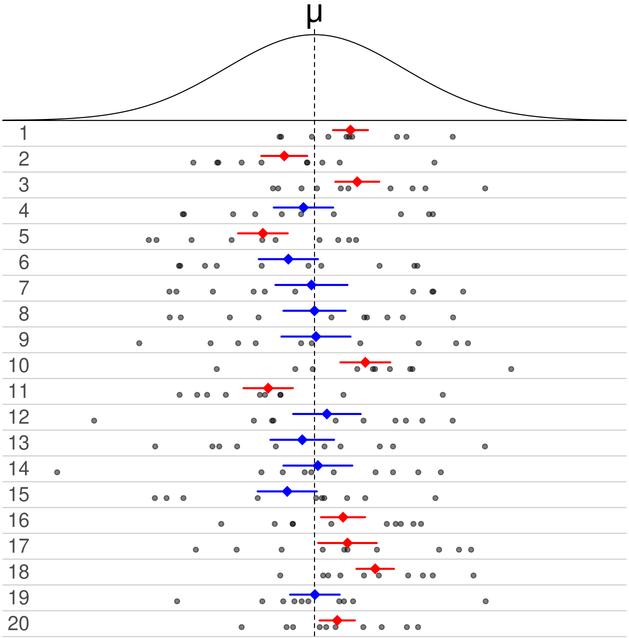


Figure 4: True value: $\beta = \mu$. 20 samples. Source: wikipedia

```
> model = lm_robust(data=data_combined, pm25 ~ log(GDP_Per_Capita))
> summary(model)

Call:
lm_robust(formula = pm25 ~ log(GDP_Per_Capita), data = data_combined)

Standard error type:  HC2

Coefficients:
                    Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
(Intercept)           92.307     6.6385   13.90 1.063e-32   79.231  105.384 242
log(GDP_Per_Capita)   -7.309     0.7254  -10.08 3.631e-20   -8.738   -5.881 242

Multiple R-squared:  0.3113 ,   Adjusted R-squared:  0.3085
F-statistic: 101.5 on 1 and 242 DF,  p-value: < 2.2e-16
```

Figure 5: 95% CI for effect of log(GDP): (-8.7, -5.8)

- CI are useful: they move away from overly unrealistic null hypothesis testing
- Also moves away from dichotomous answers ("significant" or not)
- But note: they share info. If CI includes 0, then $\widehat{\beta}$ will be "insignificant".
- Very good practice to report 95% CI.

Conclusion

- Estimation: obtain estimates of β. $\hat{\beta}$ will be normally distributed in large samples when standard assumptions are met.
- Hypothesis testing. Figure out if our $\hat{\beta}$ is "too large" (or too different) to be plausible under null hypothesis.
- Confidence intervals. What is a plausible range for β given my data?

Questions?

References

Gelman, Andrew. 2013. "Commentary: P: Values and Statistical Practice." *Epidemiology* 24 (1): 69–72.

Gelman, Andrew, and Hal Stern. 2006. "The Difference Between 'Significant' and 'Not Significant' Is Not Itself Statistically Significant." *The American Statistician* 60 (4): 328–31.

Hoekstra, Rink, Richard D. Morey, Jeffrey N. Rouder, and Eric-Jan Wagenmakers. 2014. "Robust Misinterpretation of Confidence Intervals." *Psychonomic Bulletin & Review* 21 (5): 1157–64.

Huber, Peter J. 1967. "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions." In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:221–33. 1. Berkeley, CA: University of California Press.

Stock, James H., and Mark W. Watson. 2011. *Introduction to Econometrics, 3rd Edition*. Pearson.

Verbeek, Marno. 2018. *A Guide to Modern Econometrics 5th Edition*. Wiley.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica: Journal of the Econometric Society*, 817–38.