# Social Science Inquiry II

## Week 8: Inference for multivariate regression, part II

Molly Offer-Westort

Department of Political Science,
University of Chicago

Winter 2022

# Loading packages for this class

```
> set.seed(60637)
> # For plotting:
> library(ggplot2)
> # library(devtools)
> # devtools::install_github("wilkelab/ungeviz")
> library(ungeviz)
> library(ggridges)
```

- Housekeeping

# P-hacking

# P-values

Suppose $\hat{\theta}$ is the general form for an estimate produced by our estimator, and $\hat{\theta}^*$ is the value we have actually observed.

# P-values

- A two-tailed p-value under the null hypothesis is

$$p = \mathrm{P}_0[|\hat{\theta}| \geq |\hat{\theta}^*|]$$

i.e., the probability *under the null distribution* that we would see an estimate of $\hat{\theta}$ as or more extreme as what we saw from the data.

- Suppose we have some data, $(Y, X_1, X_2, \ldots X_K)$.

- Suppose we have some data, $(Y, X_1, X_2, \dots X_K)$.
- Suppose the null distribution represents the truth.

- Suppose we have some data, $(Y, X_1, X_2, \ldots X_K)$.
- Suppose the null distribution represents the truth.
- If we test one hypothesis, what is the probability that we will find something that is statistically significant at $p \leq 0.05$?

- Suppose we have some data, $(Y, X_1, X_2, \ldots X_K)$.
- Suppose the null distribution represents the truth.
- If we test one hypothesis, what is the probability that we will find something that is statistically significant at $p \leq 0.05$?
- If we test two unrelated hypotheses, what is the probability that we will find something that is statistically significant at $p \leq 0.05$?

- A: event we reject hypothesis 1 at $p \leq 0.05$

- A: event we reject hypothesis 1 at $p \leq 0.05$
- B: event we reject hypothesis 2 at $p \leq 0.05$

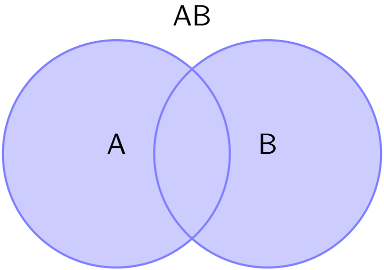
- A: event we reject hypothesis 1 at $p \leq 0.05$
- B: event we reject hypothesis 2 at $p \leq 0.05$
- $A \perp\!\!\!\perp B$: the two events are independent

- A: event we reject hypothesis 1 at $p \leq 0.05$
- B: event we reject hypothesis 2 at $p \leq 0.05$
- $A \perp\!\!\!\perp B$: the two events are independent
- $\mathrm{P}[A] = 0.05$

- A: event we reject hypothesis 1 at $p \leq 0.05$
- B: event we reject hypothesis 2 at $p \leq 0.05$
- $A \perp\!\!\!\perp B$: the two events are independent
- $\mathrm{P}[A] = 0.05$
- $\mathrm{P}[B] = 0.05$

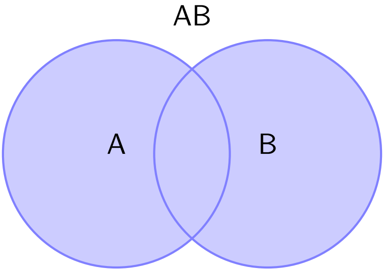
- A: event we reject hypothesis 1 at $p \leq 0.05$
- B: event we reject hypothesis 2 at $p \leq 0.05$
- $A \perp\!\!\!\perp B$: the two events are independent
- $\mathrm{P}[A] = 0.05$
- $\mathrm{P}[B] = 0.05$
- $\mathrm{P}[AB]$? The probability we see event A OR B?

$$\mathrm{P}[\mathsf{AB}] = \mathrm{P}[\mathsf{A}] + \mathrm{P}[\mathsf{A}^{\mathsf{C}}] \times \mathrm{P}[\mathsf{B}|\mathsf{A}^{\mathsf{C}}]$$



$$\mathrm{P}[\mathsf{AB}] = \mathrm{P}[\mathsf{A}] + \mathrm{P}[\mathsf{A}^{\mathsf{C}}] \times \mathrm{P}[\mathsf{B}|\mathsf{A}^{\mathsf{C}}]$$

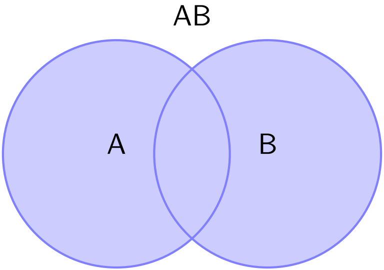
A and B are independent, so $\mathrm{P}[\mathsf{B}|\mathsf{A}^{\mathsf{C}}] = \mathrm{P}[\mathsf{B}]$



$$\mathrm{P}[\mathsf{AB}] = \mathrm{P}[\mathsf{A}] + \mathrm{P}[\mathsf{A}^\mathsf{C}] \times \mathrm{P}[\mathsf{B}|\mathsf{A}^\mathsf{C}]$$

A and B are independent, so $\mathrm{P}[\mathsf{B}|\mathsf{A}^\mathsf{C}] = \mathrm{P}[\mathsf{B}]$

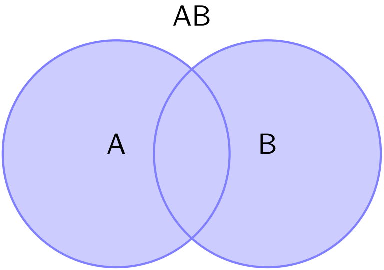
$$\mathrm{P}[\mathsf{AB}] = \mathrm{P}[\mathsf{A}] + \mathrm{P}[\mathsf{A}^\mathsf{C}] \times \mathrm{P}[\mathsf{B}]$$



$$P[AB] = P[A] + P[A^C] \times P[B|A^C]$$

A and B are independent, so $P[B|A^C] = P[B]$

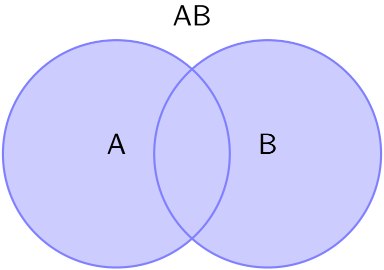
$$P[AB] = P[A] + P[A^C] \times P[B]$$
$$P[AB] = 0.05 + 0.95 \times 0.05$$



$$P[AB] = P[A] + P[A^C] \times P[B|A^C]$$

A and B are independent, so $P[B|A^C] = P[B]$

$$P[AB] = P[A] + P[A^C] \times P[B]$$
$$P[AB] = 0.05 + 0.95 \times 0.05$$
$$P[AB] = 0.0975$$



If the null is true . . .

- and we conduct three independent tests, the probability that *at least one of them* will be statistically significant at $p \leq 0.05$ is $1 - 0.95^3 = 0.1426$

If the null is true . . .

- and we conduct three independent tests, the probability that *at least one of them* will be statistically significant at $p \leq 0.05$ is $1 - 0.95^3 = 0.1426$
- and we conduct four independent tests, the probability that *at least one of them* will be statistically significant at $p \leq 0.05$ is $1 - 0.95^4 = 0.1855$

If the null is true . . .

- and we conduct three independent tests, the probability that *at least one of them* will be statistically significant at $p \leq 0.05$ is $1 - 0.95^3 = 0.1426$
- and we conduct four independent tests, the probability that *at least one of them* will be statistically significant at $p \leq 0.05$ is $1 - 0.95^4 = 0.1855$
- and we conduct ten independent tests, the probability that *at least one of them* will be statistically significant at $p \leq 0.05$ is $1 - 0.95^{10} = 0.4013$

If the null is true . . .

- and we conduct three independent tests, the probability that *at least one of them* will be statistically significant at $p \leq 0.05$ is $1 - 0.95^3 = 0.1426$
- and we conduct four independent tests, the probability that *at least one of them* will be statistically significant at $p \leq 0.05$ is $1 - 0.95^4 = 0.1855$
- and we conduct ten independent tests, the probability that *at least one of them* will be statistically significant at $p \leq 0.05$ is $1 - 0.95^{10} = 0.4013$

This becomes a real problem when researchers run many tests in their papers!

| | **Fail to reject null hypothesis** ($p > 0.05$) | **Reject null hypothesis** ($p \leq 0.05$) |
|---|---|---|
| **Null hypothesis true** | True negative | Type I error, false positive |
| **Null hypothesis false** | Type II error, false negative | True positive |

- **Type I error:** (false positive) we see an effect, where one doesn't really exist

| | **Fail to reject null hypothesis** ($p > 0.05$) | **Reject null hypothesis** ($p \leq 0.05$) |
|---|---|---|
| **Null hypothesis true** | True negative | Type I error, false positive |
| **Null hypothesis false** | Type II error, false negative | True positive |

- **Type I error:** (false positive) we see an effect, where one doesn't really exist
- **Type II error:** (false negative) we didn't see an effect, but one really does exist

# Multiple testing scenarios

- Comparisons across multiple treatments; A to B, B to C, A to C...

# Multiple testing scenarios

- Comparisons across multiple treatments; A to B, B to C, A to C...
- Multiple outcomes

# Multiple testing scenarios

- Comparisons across multiple treatments; A to B, B to C, A to C...
- Multiple outcomes
- Heterogeneous treatment effects (where is the cut point)

# Multiple testing scenarios

- Comparisons across multiple treatments; A to B, B to C, A to C...
- Multiple outcomes
- Heterogeneous treatment effects (where is the cut point)
- Multiple regression specifications (specification search)

# Multiple testing scenarios

- Comparisons across multiple treatments; A to B, B to C, A to C...
- Multiple outcomes
- Heterogeneous treatment effects (where is the cut point)
- Multiple regression specifications (specification search)

These tests aren't all fully independent, but the more tests we do, the more likely we are to uncover a false positive.

# Ways to account for multiple testing

- Pre-specification of analyses

# Ways to account for multiple testing

- Pre-specification of analyses
- Separating data in training and testing sets (more on this with machine learning)

# Ways to account for multiple testing

- Pre-specification of analyses
- Separating data in training and testing sets (more on this with machine learning)
- *p-value adjustment*

# p-value adjustment

| | **Fail to reject null hypothesis** ($p > 0.05$) | **Reject null hypothesis** ($p \leq 0.05$) |
|---|---|---|
| **Null hypothesis true** | True negative | Type I error, false positive |
| **Null hypothesis false** | Type II error, false negative | True positive |

- Family-Wise Error Rate (FWER): the probability of falsely rejecting even one *true* null hypothesis;

# p-value adjustment

| | **Fail to reject null hypothesis** (p $>$0.05) | **Reject null hypothesis** (p $\leq$ 0.05) |
|---|---|---|
| **Null hypothesis true** | True negative | Type I error, false positive |
| **Null hypothesis false** | Type II error, false negative | True positive |

- Family-Wise Error Rate (FWER): the probability of falsely rejecting even one *true* null hypothesis; $\mathrm{P}[\text{Type I error} > 0]$

# p-value adjustment

| | **Fail to reject null hypothesis** (p $>$0.05) | **Reject null hypothesis** (p $\leq$ 0.05) |
|---|---|---|
| **Null hypothesis true** | True negative | Type I error, false positive |
| **Null hypothesis false** | Type II error, false negative | True positive |

- Family-Wise Error Rate (FWER): the probability of falsely rejecting even one *true* null hypothesis; $\mathrm{P}[\text{Type I error} > 0]$
- False Discovery Rate (FDR): expected proportion of false discoveries among all discoveries;

# p-value adjustment

| | **Fail to reject null hypothesis** (p $>$0.05) | **Reject null hypothesis** (p $\leq$ 0.05) |
|---|---|---|
| **Null hypothesis true** | True negative | Type I error, false positive |
| **Null hypothesis false** | Type II error, false negative | True positive |

- Family-Wise Error Rate (FWER): the probability of falsely rejecting even one *true* null hypothesis; $\mathrm{P}[\text{Type I error} > 0]$
- False Discovery Rate (FDR): expected proportion of false discoveries among all discoveries; $\mathrm{E}\left[\# \text{ False discoveries}/\# \text{ All discoveries}\right]$

# p-value adjustment

- Correcting FWER
  - Bonferroni correction: for $m$ hypotheses, for significance level $\alpha$, implement $\alpha/m$

# p-value adjustment

- Correcting FWER
  - Bonferroni correction: for *m* hypotheses, for significance level $\alpha$, implement $\alpha/m$
  - four independent tests, the probability that *at least one of them* will be statistically significant at $p \leq \alpha$ is $1 - (1 - \alpha)^4$
  - For $\alpha = 0.05$, $1 - 0.95^4 = 0.1855$

# p-value adjustment

- Correcting FWER
  - Bonferroni correction: for $m$ hypotheses, for significance level $\alpha$, implement $\alpha/m$
  - four independent tests, the probability that *at least one of them* will be statistically significant at $p \leq \alpha$ is $1-(1-\alpha)^4$
  - For $\alpha = 0.05$, $1 - 0.95^4 = 0.1855$
  - With Bonferroni correction: $1-(1-\alpha/4)^4 = 0.0491$

# p-value adjustment

- Correcting FWER
  - Bonferroni correction: for $m$ hypotheses, for significance level $\alpha$, implement $\alpha/m$
  - four independent tests, the probability that *at least one of them* will be statistically significant at $p \leq \alpha$ is $1-(1-\alpha)^4$
  - For $\alpha = 0.05$, $1 - 0.95^4 = 0.1855$
  - With Bonferroni correction: $1-(1-\alpha/4)^4 = 0.0491$
  - Ten independent tests: $1-(1-\alpha/10)^1 0 = 0.0489$

# p-value adjustment

What if tests are not independent?

# p-value adjustment

What if tests are not independent? Bonferroni is too aggressive.

$$\mathrm{P}[\mathsf{AB}] = \mathrm{P}[\mathsf{A}] + \mathrm{P}[\mathsf{A}^{\mathsf{C}}] \times \mathrm{P}[\mathsf{B}|\mathsf{A}^{\mathsf{C}}]$$

# p-value adjustment

What if tests are not independent? Bonferroni is too aggressive.

$$\mathrm{P}[\mathsf{AB}] = \mathrm{P}[\mathsf{A}] + \mathrm{P}[\mathsf{A}^\mathsf{C}] \times \mathrm{P}[\mathsf{B}|\mathsf{A}^\mathsf{C}]$$

If A and B are positively correlated $\mathrm{P}[\mathsf{B}|\mathsf{A}^\mathsf{C}] \leq \mathrm{P}[\mathsf{B}]$

# p-value adjustment

What if tests are not independent? Bonferroni is too aggressive.

$$\mathrm{P}[\mathsf{AB}] = \mathrm{P}[\mathsf{A}] + \mathrm{P}[\mathsf{A}^{\mathsf{C}}] \times \mathrm{P}[\mathsf{B}|\mathsf{A}^{\mathsf{C}}]$$

If A and B are positively correlated $\mathrm{P}[\mathsf{B}|\mathsf{A}^{\mathsf{C}}] \leq \mathrm{P}[\mathsf{B}]$

# p-value adjustment

What if tests are not independent? Bonferroni is too aggressive.

$$\mathrm{P}[\mathsf{AB}] = \mathrm{P}[\mathsf{A}] + \mathrm{P}[\mathsf{A}^\mathsf{C}] \times \mathrm{P}[\mathsf{B}|\mathsf{A}^\mathsf{C}]$$

If A and B are positively correlated $\mathrm{P}[\mathsf{B}|\mathsf{A}^\mathsf{C}] \leq \mathrm{P}[\mathsf{B}]$

- Correcting FWER
  - Holm correction: for *m* hypotheses, for significance level $\alpha$:
    - Order the *m* conventionally calculated p-values from smallest to largest

# p-value adjustment

What if tests are not independent? Bonferroni is too aggressive.

$$\mathrm{P}[\mathsf{AB}] = \mathrm{P}[\mathsf{A}] + \mathrm{P}[\mathsf{A}^{\mathsf{C}}] \times \mathrm{P}[\mathsf{B}|\mathsf{A}^{\mathsf{C}}]$$

If A and B are positively correlated $\mathrm{P}[\mathsf{B}|\mathsf{A}^{\mathsf{C}}] \leq \mathrm{P}[\mathsf{B}]$

- Correcting FWER
  - Holm correction: for *m* hypotheses, for significance level $\alpha$:
    - Order the *m* conventionally calculated p-values from smallest to largest
    - Find the *smallest* p-value indexed as *k* such that $p_k > \frac{\alpha}{m+1-k}$

# p-value adjustment

What if tests are not independent? Bonferroni is too aggressive.

$$\mathrm{P}[\mathsf{AB}] = \mathrm{P}[\mathsf{A}] + \mathrm{P}[\mathsf{A}^{\mathsf{C}}] \times \mathrm{P}[\mathsf{B}|\mathsf{A}^{\mathsf{C}}]$$

If A and B are positively correlated $\mathrm{P}[\mathsf{B}|\mathsf{A}^{\mathsf{C}}] \leq \mathrm{P}[\mathsf{B}]$

- Correcting FWER
  - Holm correction: for *m* hypotheses, for significance level $\alpha$:
    - Order the *m* conventionally calculated p-values from smallest to largest
    - Find the *smallest* p-value indexed as $k$ such that $p_k > \frac{\alpha}{m+1-k}$
    - Reject all p-values greater than or equal to $p_k$, accept all p-values less $p_k$

# p-value adjustment

- Correcting FDR
  - Benjamini-Hochberg correction: for *m* hypotheses, for significance level $\alpha$:
    - Order the *m* conventionally calculated p-values from smallest to largest

# p-value adjustment

- Correcting FDR
  - Benjamini-Hochberg correction: for $m$ hypotheses, for significance level $\alpha$:
    - Order the $m$ conventionally calculated p-values from smallest to largest
    - Find the *largest* p-value indexed as $k$ such that $p_k \leq \frac{k}{m}\alpha$

# p-value adjustment

- Correcting FDR
  - Benjamini-Hochberg correction: for $m$ hypotheses, for significance level $\alpha$:
    - Order the $m$ conventionally calculated p-values from smallest to largest
    - Find the *largest* p-value indexed as $k$ such that $p_k \leq \frac{k}{m}\alpha$
    - Reject all p-values greater than $p_k$, accept all p-values less than or equal to $p_k$

# p-value adjustment

- In either case, for more complex settings, try simulation.

# Multiple testing

- When can you consider tests as unrelated?

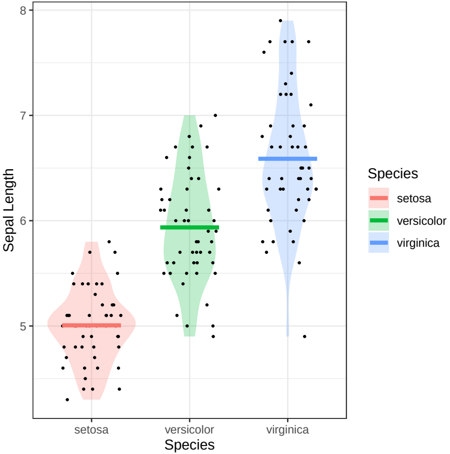
# Multiple testing

- When can you consider tests as unrelated?
- Exploratory vs. confirmatory hypotheses?

# Some alternatives to confidence intervals (via `ungeviz`)

Show the underlying data.

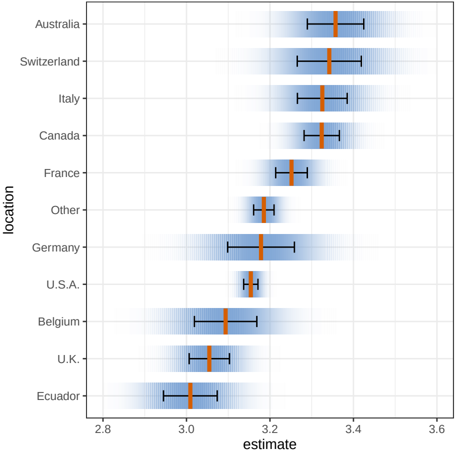
```
> ggplot(iris, aes(Species, Sepal.Length,fill = Species)) +
+     geom_violin(alpha = 0.25, color = NA) +
+     geom_point(position = position_jitter(width = 0.3, height = 0), size = 0.5) +
+     geom_hpline(aes(colour = Species), stat = "summary", width = 0.6, size = 1.5, fun = 'mean')
```



# Some alternatives to confidence intervals (via `ungeviz`)

## Shaded confidence strips.

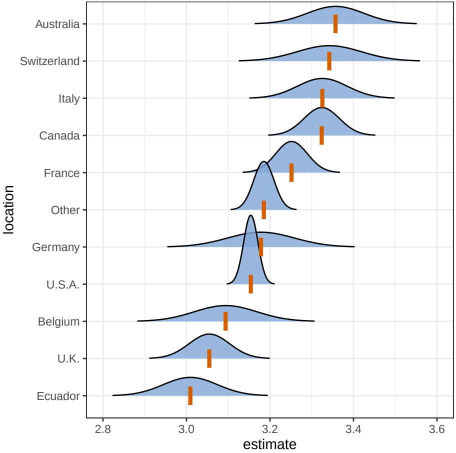
```
> ggplot(cacao_means, aes(x = estimate, y = location)) +
+     stat_confidence_density(aes(moe = std.error), confidence = 0.68, fill = "#81A7D6", height = 0.7) +
+     geom_errorbarh(aes(xmin = estimate - std.error, xmax = estimate + std.error), height = 0.3) +
+     geom_vpline(aes(x = estimate), size = 1.5, height = 0.7, color = "#D55E00")
```



# Some alternatives to confidence intervals (via `ungeviz`)

## Confidence densities.

```
> ggplot(cacao_means, aes(x = estimate, y = location)) +
+     stat_confidence_density(
+         aes(moe = std.error, height = stat(density)), geom = "ridgeline",
+         confidence = 0.68, fill = "#81A7D6", alpha = 0.8, scale = 0.08, min_height = 0.1) +
+     geom_vpline(aes(x = estimate), size = 1.5, height = 0.5, color = "#D55E00")
```



# References I

Clause Wilke: `https://wilkelab.org/ungeviz/index.html`