MONASH University

ETC5521: Exploratory Data Analysis

Using computational tools to determine whether what is seen in the data can be assumed to apply more broadly

Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 4 - Session 1

Revisiting

hypothesis testing

(Frequentist) hypothesis testing framework

- Suppose X is the number of heads out of n independent tosses.
- Let p be the probability of getting a  for this coin.

Hypotheses	$H_0 : p = 0.5$ vs. $H_a : p > 0.5$. Note $p_0 = 0.5$. Alternative H_a is saying we believe that the coin is biased to heads. Alternative needs to be decided before seeing data.
Assumptions	Each toss is independent with equal chance of getting a head.
Test statistic	$X \sim B(n, p)$. Recall $E(X \mid H_0) = np_0$. We observe $n, x, \hat{p}$. Test statistic is $\hat{p} - p_0$.
P-value (or critical value or confidence interval)	$P(X \geq x \mid H_0)$
Conclusion	Reject null hypothesis when the p-value is less than some significance level α. Usually $\alpha = 0.05$.

Testing coin bias Part 1/2

- Suppose I have a coin that I'm going to flip 
- If the coin is unbiased, what is the probability it will show heads?
- Yup, the probability should be 0.5.
- So how would I test if a coin is biased or unbiased?
- We'll collect some data.
- **Experiment 1**: I flipped the coin 10 times and this is the result:

- The result is 7 head and 3 tails. So 70% are heads.
- Do you believe the coin is biased based on this data?

Testing coin bias Part 2/2

- **Experiment 2**: Suppose now I flip the coin 100 times and this is the outcome:

- We observe 70 heads and 30 tails. So again 70% are heads.
- Based on this data, do you think the coin is biased?

Calculate it

Experiment 1 (n=10)

- We observed $x = 7$, or $\hat{p} = 0.7$.
- Assuming H_0 is true, we expect $np = 10 \times 0.5 = 5$.
- Calculate the $P(X \geq 7)$

```
sum(dbinom(7:10, 10, 0.5))

## [1] 0.171875
```

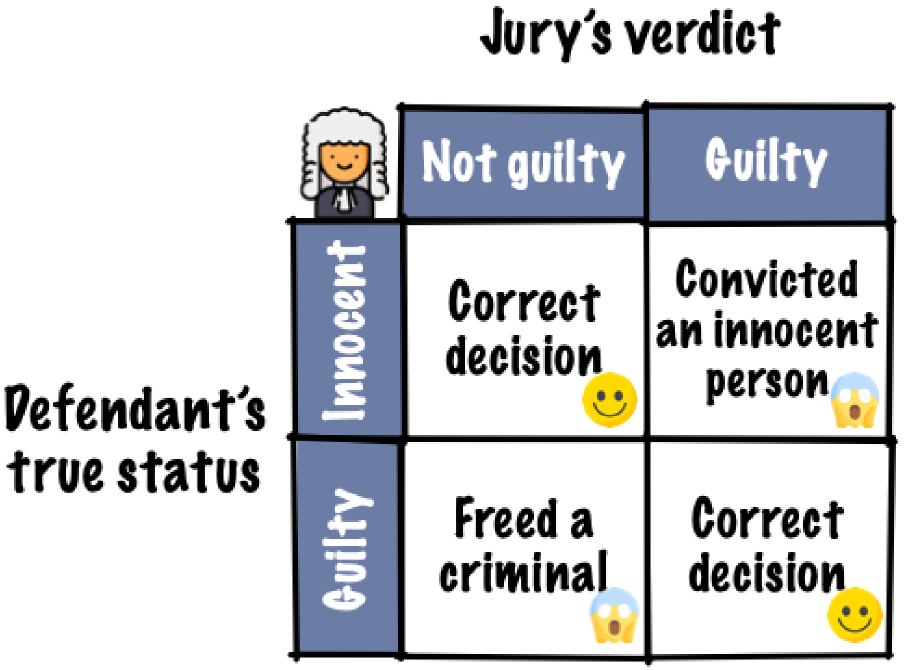
Experiment 1 (n=100)

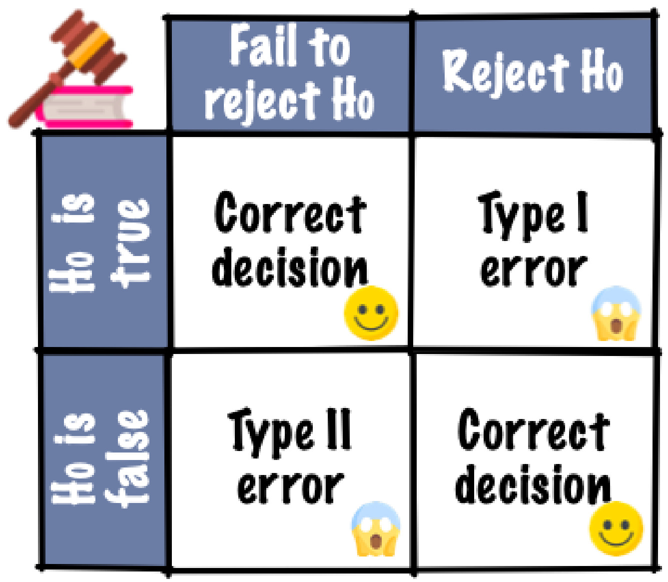
- We observed $x = 70$, or $\hat{p} = 0.7$.
- Assuming H_0 is true, we expect $np = 100 \times 0.5 = 50$.
- Calculate the $P(X \geq 70)$

```
sum(dbinom(70:100, 100, 0.5))

## [1] 3.92507e-05
```

Judicial system





- Evidence by test statistic
- Judgement by p-value, critical value or confidence interval

Does the test statistic have to be a *numerical summary statistic*?

Visual inference

Visual inference

- Hypothesis testing in visual inference framework is where:
  - the *test statistic is a plot* and
  - judgement is by human visual perception.
- You (and many other people) actually do visual inference many times but generally in an informal fashion.
- Here, we are making an inference on whether the residual plot has any patterns based on a single data plot.

- Data plots tend to be over-interpreted

- Reading data plots require calibration

Visual inference more formally

1. State your null and alternate hypotheses.
2. Define a **visual test statistic**, $V(.)$, i.e. a function of a sample to a plot.
3. Define a method to generate **null data**, $\boldsymbol{y}_0$.
4. $V(\boldsymbol{y})$ maps the actual data, $\boldsymbol{y}$, to the plot. We call this the **data plot**.
5. $V(\boldsymbol{y}_0)$ maps a null data to a plot of the same form. We call this the **null plot**. We repeat this $m-1$ times to generate $m-1$ null plots.
6. A **lineup** displays these m plots in a random order.
7. Ask n human viewers to select a plot in the lineup that looks different to others without any context given.

Suppose x out of n people detected the data plot from a lineup, then

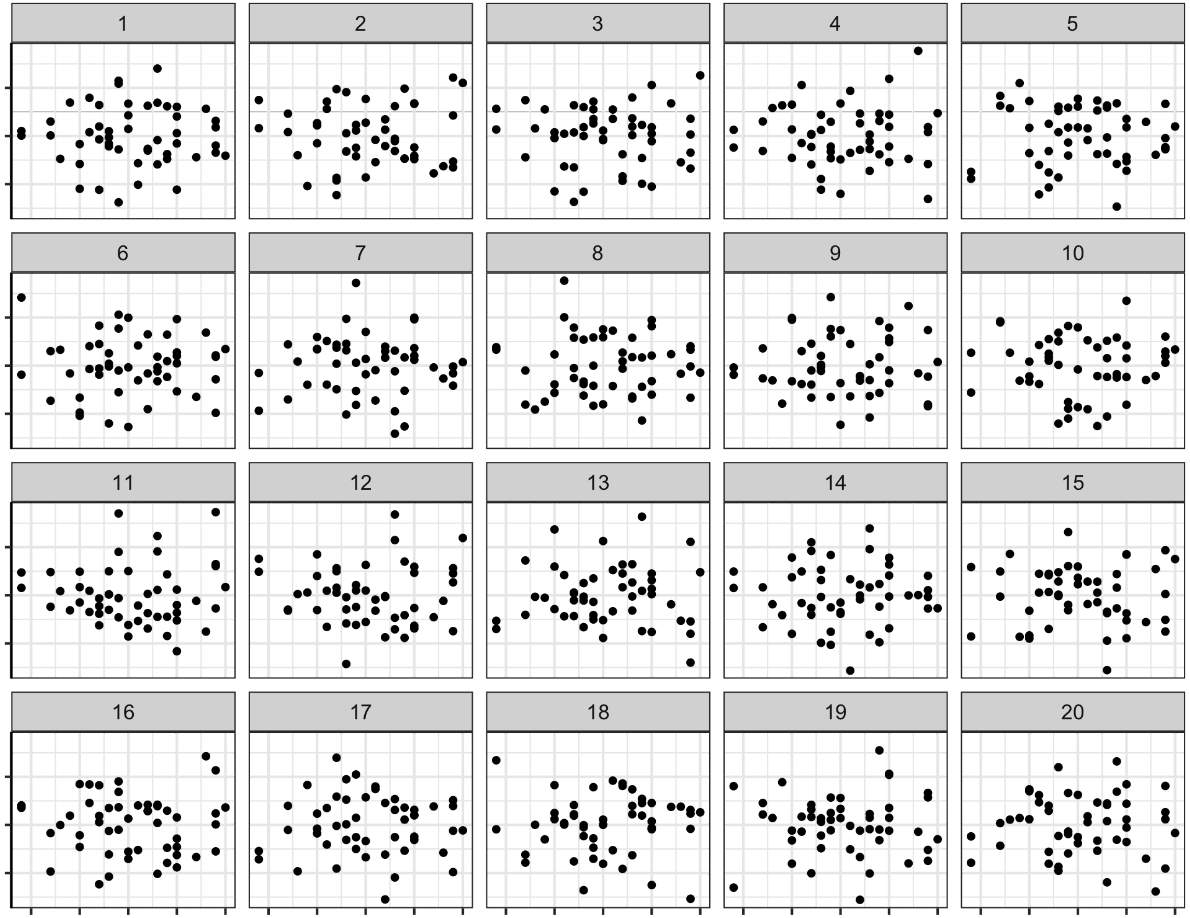
- the **visual inference p-value** is given as

$$P(X \geq x)$$

where $X \sim B(n, 1/m)$, and

- the **power of a lineup** is estimated as x/n.

Lineup 1 In which plot has a pattern that is different from other plots?



Recall the linear model for cars shown in week 3.

```
lm(dist ~ speed, data = cars)
```

- This is a lineup of the residual plot
- Which plot (if any) looks different from the others?
- Why do you think it looks different?
- Note: there is no correct answer here.

```
> decrypt("clZx bKhK oL 3OHohoOL 0B'
[1] "True data in position  11"
```

How do we calculate statistical significance from this?

Visual inference p-value (or "see"-value)

- So x out of n people chose the data plot.
- So the visual inference p-value is $P(X \geq x)$ where $X \sim B(n, 1/10)$.
- In R, this is

```
1 - pbinom(x - 1, n, 1/20)
# OR
nullabor::pvisual(x, n, 20)
```

- The calculation is made with the assumption that the chance of a single observer randomly chooses the true plot is 1/20.

Suppose $x = 2$ out of $n = 16$ people chose plot 11 (previous slide).

The probability that this happens by random guessing (p-value) is

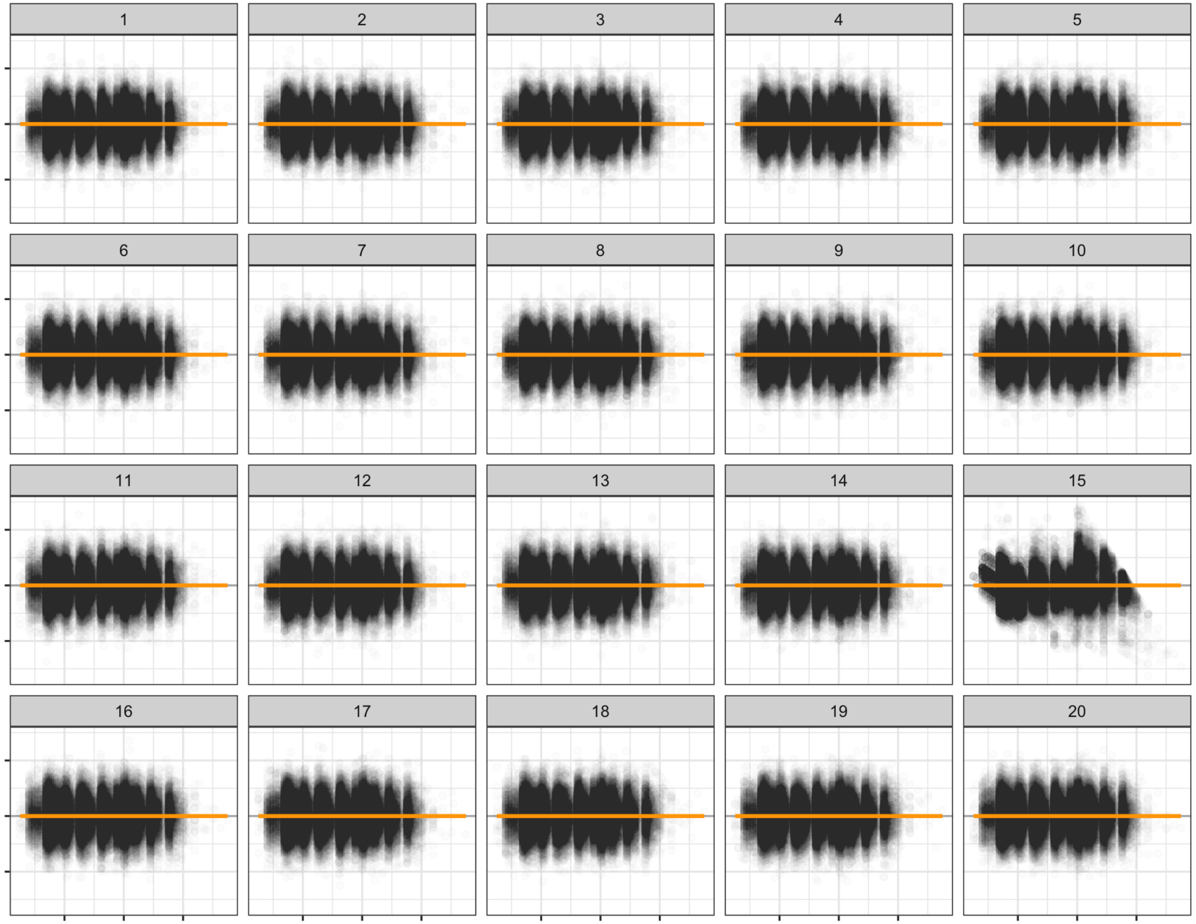
```
1 - pbinom(2 - 1, 16, 1/20)

## [1] 0.1892403

nullabor::pvisual(2, 16, 20)

##      x simulated     binom
## [1,] 2     0.204 0.1892403
```

Lineup 2 In which plot has a pattern that is different from other plots?



Recall the linear model for diamonds shown in week 3.

```
d_fit <- lm(lprice ~ lcarat, data=d:
```

- This is a lineup of the residual plot for the model where both carat and price are log-transformed
- Which plot (if any) looks different from the others?
- Why do you think it looks different?
- Note: there is no correct answer here.

```
> decrypt("clZx bKhK oL 3OHohoOL 0Q'
[1] "True data in position  15"
```

Visual inference p-value (or "see"-value)

Suppose $x = 8$ out of $n = 12$ people chose plot 15 (previous slide).

The probability that this happens by random guessing (p-value) is

```
1 - pbinom(8 - 1, 12, 1/20)

## [1] 1.612352e-08

nullabor::pvisual(8, 12, 20)

##      x simulated        binom
## [1,] 8         0 1.612352e-08
```

This is basically impossible to happen by chance.

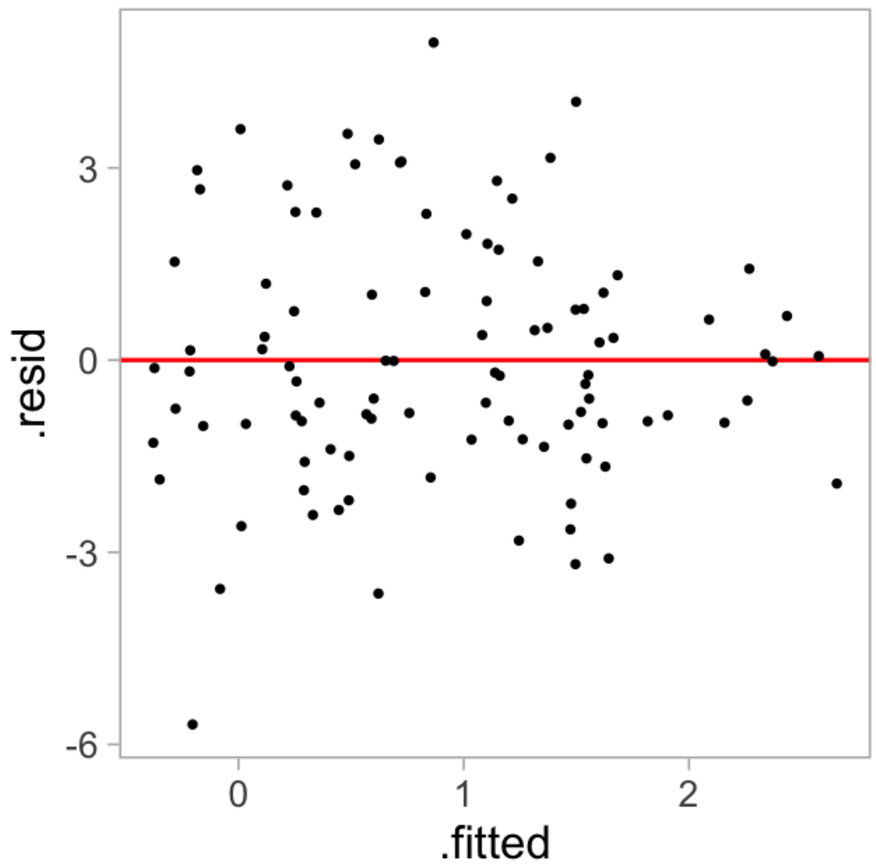
Next, how the residuals are different from "good" residuals has to be determined by the follow-up question: how did you decide your chosen plot was different?

Plot 15 has a different variance pattern, it's not the regular up-down pattern seen in the other plots. This suggests that there is some heteroskedasticity in the data that is not captured by the error distribution in the model.

Why?

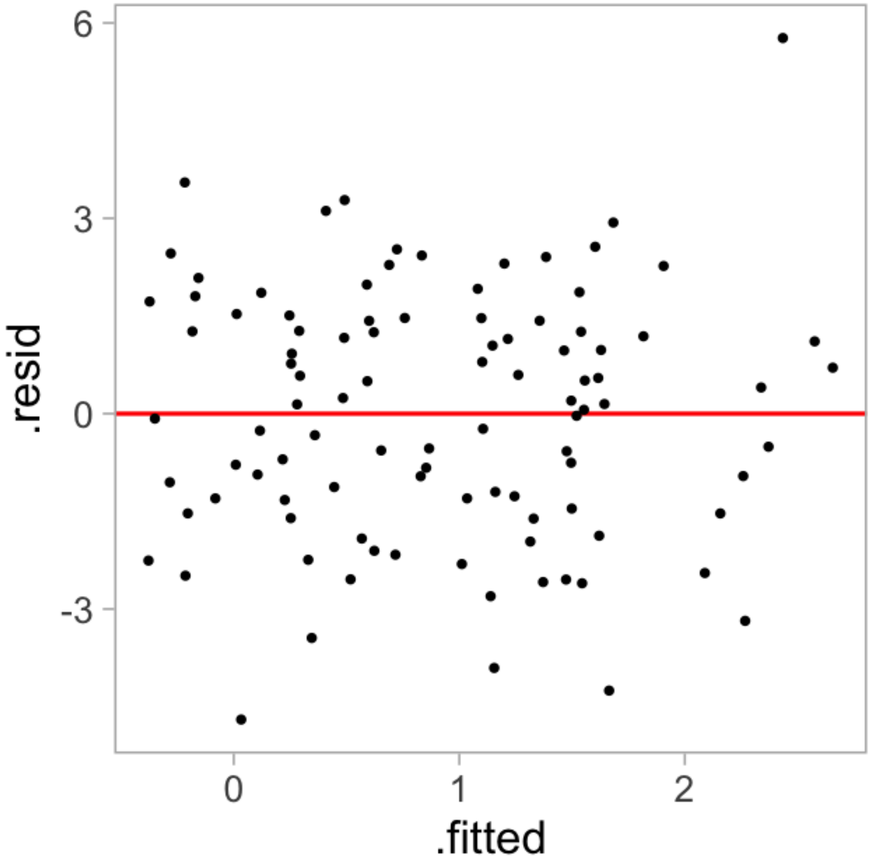
Residual plot (1/3)

Is there a problem with the model?



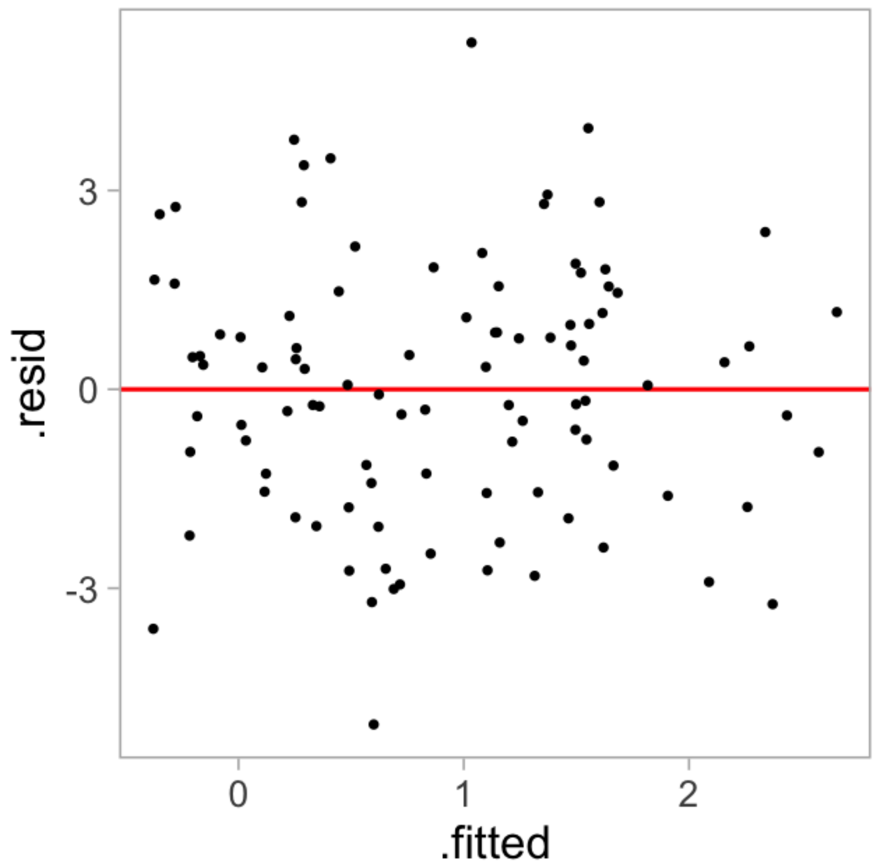
Residual plot (2/3)

Is there a problem with the model?



Residual plot (3/3)

Is there a problem with the model?

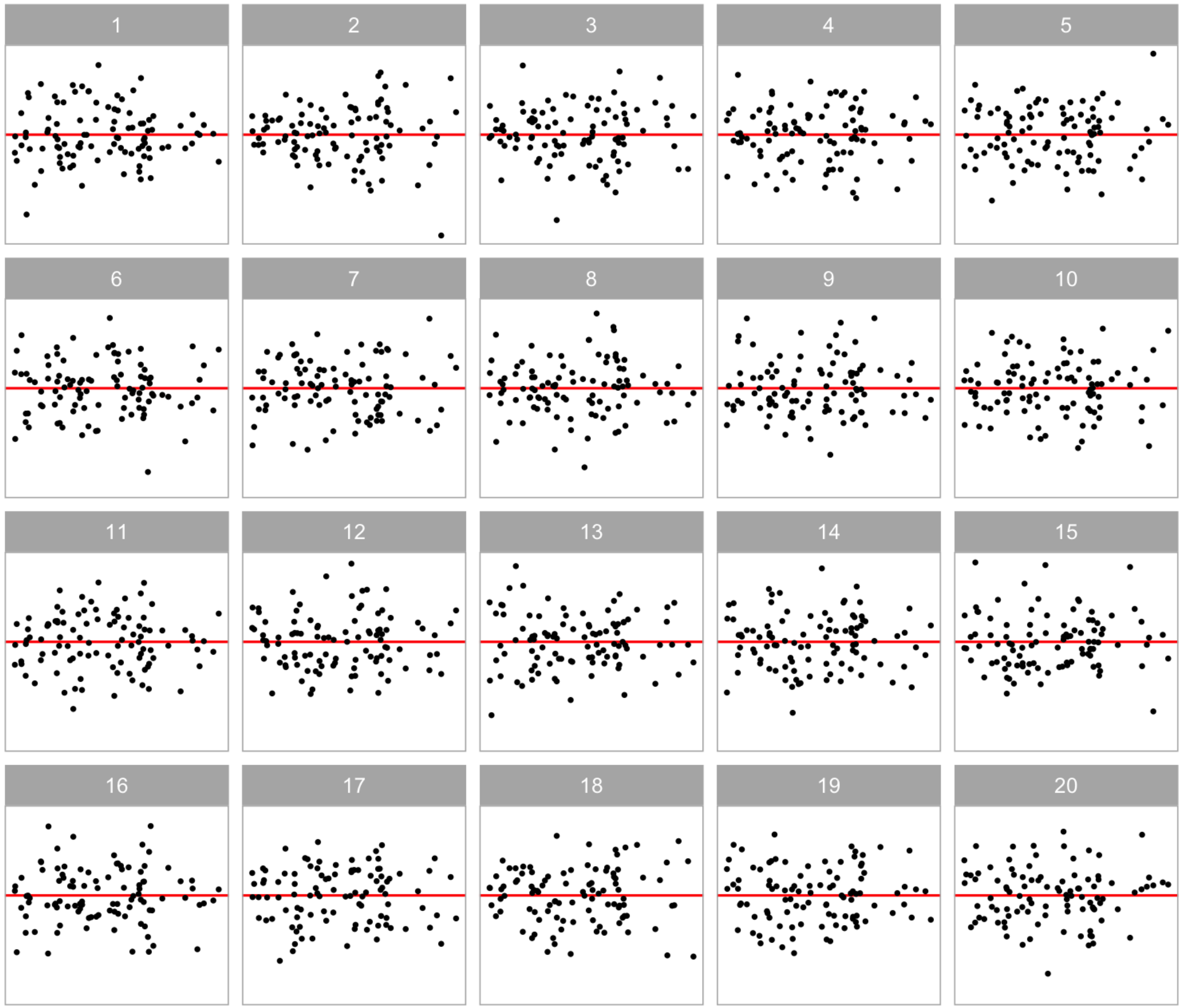


Residual plots need context

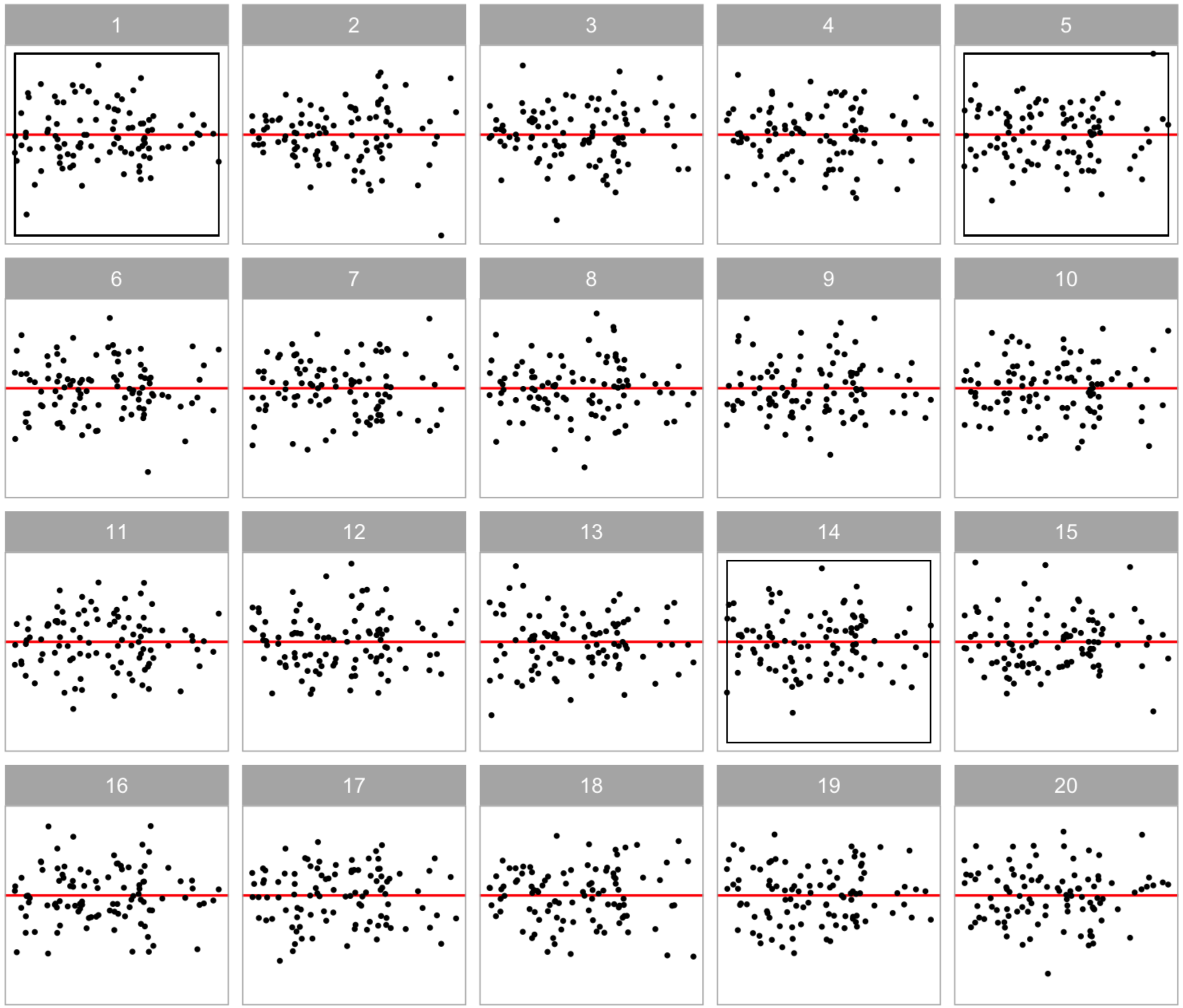
It's really hard to decide that there is NO PATTERN!

Residual plots are better when viewed in the context of good residual plots, where we know the assumptions of the model are satisfied.

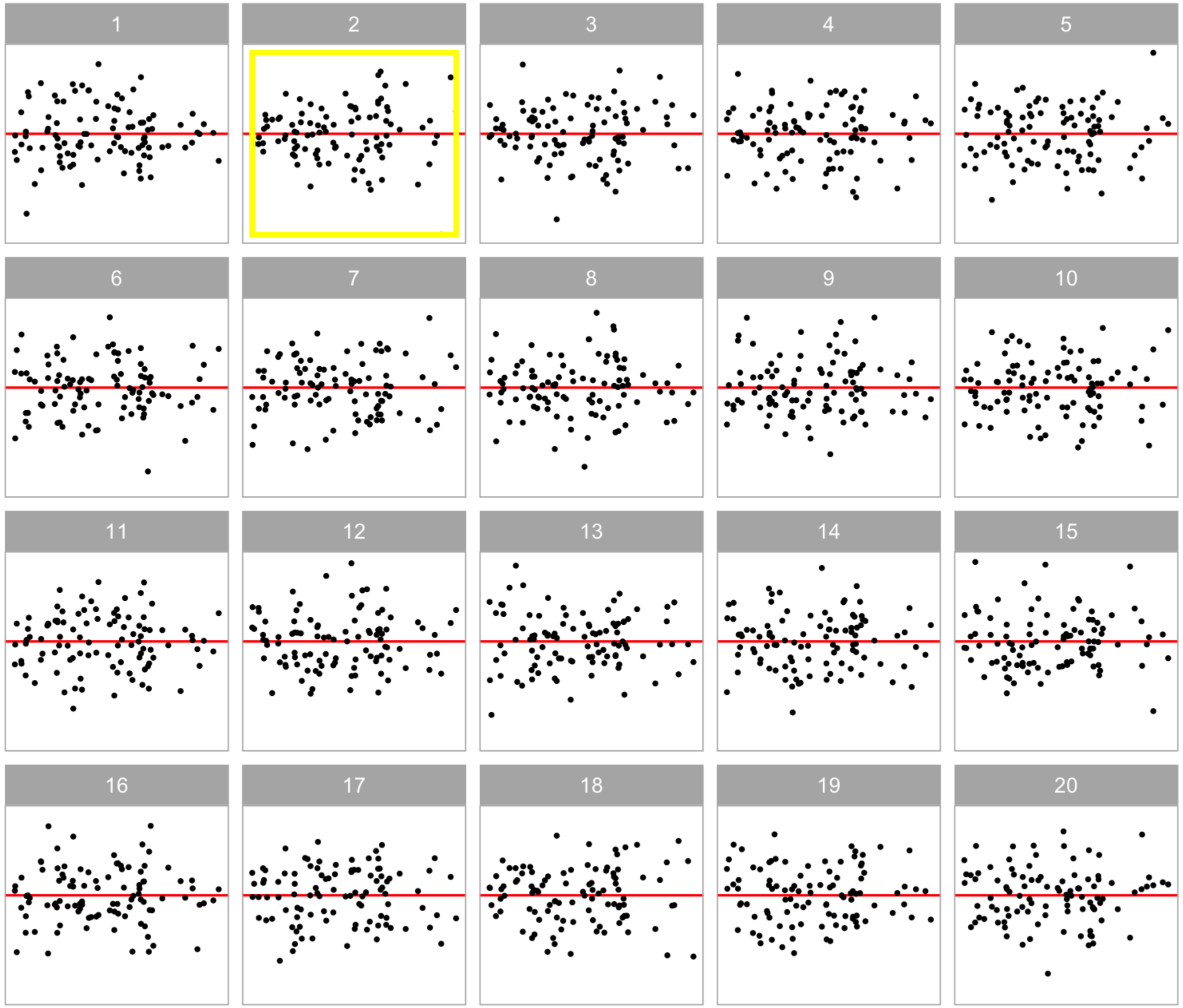
Which is the worst residual plot?



All of the previous residual plots shown were NULL plots

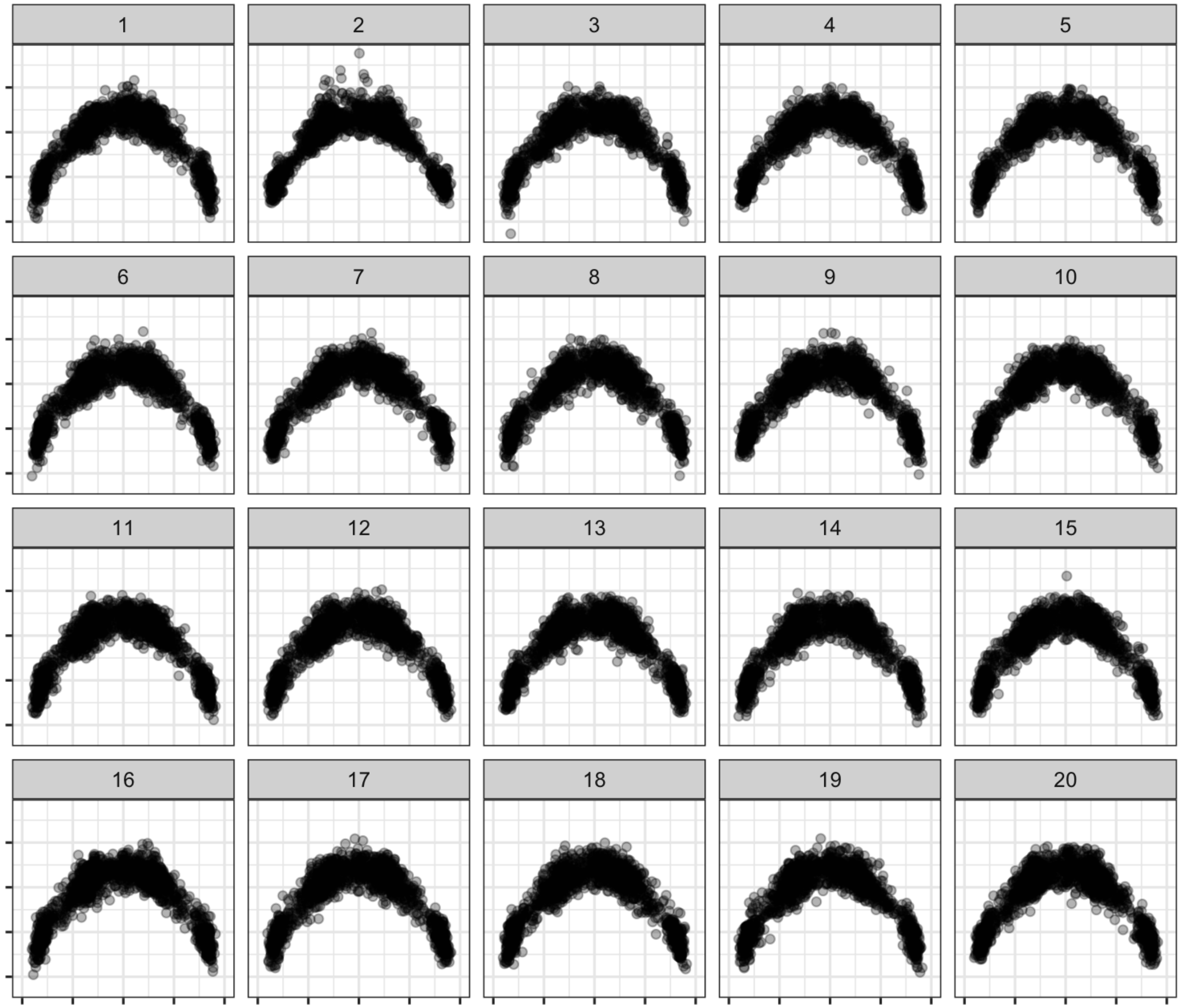


The actual residual plot is

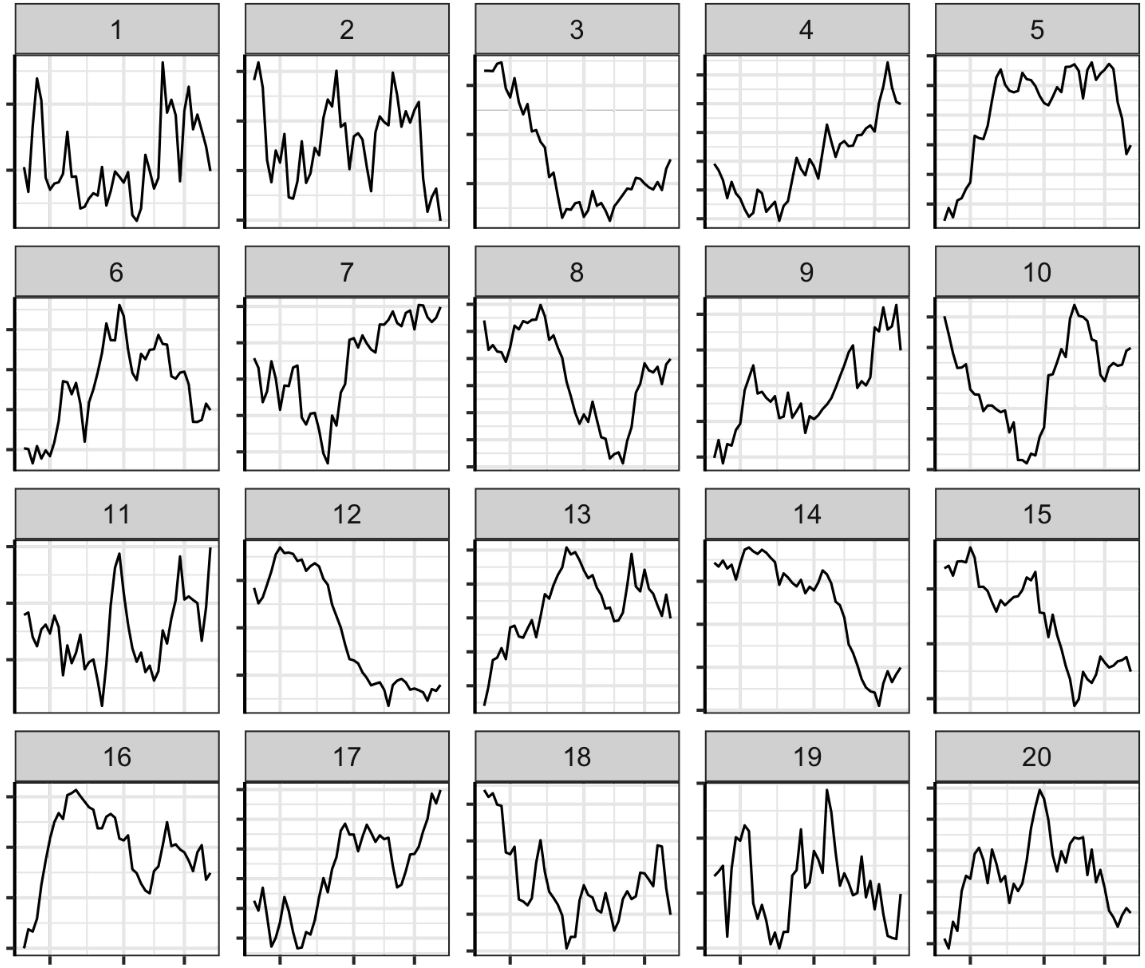


It's not only for residual plots

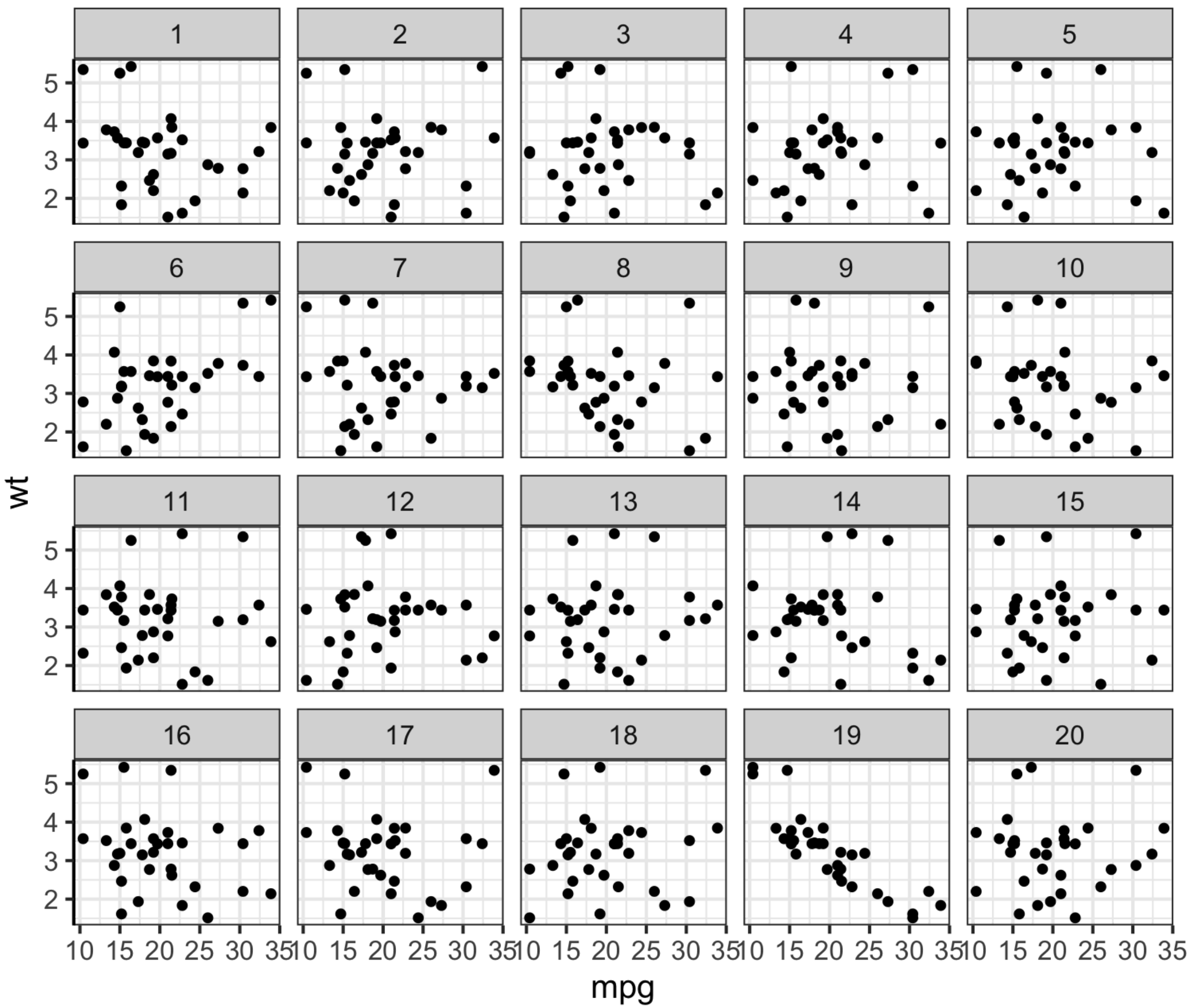
Which plot is most different?



Which plot is most different?



Which plot is most different?



Reading any plot can benefit from the context of null plots

Resources and Acknowledgement

- Buja, Andreas, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F. Swayne, and Hadley Wickham. 2009. “Statistical Inference for Exploratory Data Analysis and Model Diagnostics.” Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 367 (1906): 4361–83.
- Wickham, Hadley, Dianne Cook, Heike Hofmann, and Andreas Buja. 2010. “Graphical Inference for Infovis.” IEEE Transactions on Visualization and Computer Graphics 16 (6): 973–79.
- Hofmann, H., L. Follett, M. Majumder, and D. Cook. 2012. “Graphical Tests for Power Comparison of Competing Designs.” IEEE Transactions on Visualization and Computer Graphics 18 (12): 2441–48.
- Majumder, M., Heiki Hofmann, and Dianne Cook. 2013. “Validation of Visual Statistical Inference, Applied to Linear Models.” Journal of the American Statistical Association 108 (503): 942–56.

- Data coding using `tidyverse` suite of R packages
- Slides originally written by Emi Tanaka and constructed with `xaringan`, remark.js, `knitr`, and R Markdown.

Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 4 - Session 1