DSC 152:
Applied Statistical Data Analysis and Inference

Lecture #10
Regression Model Diagnostics

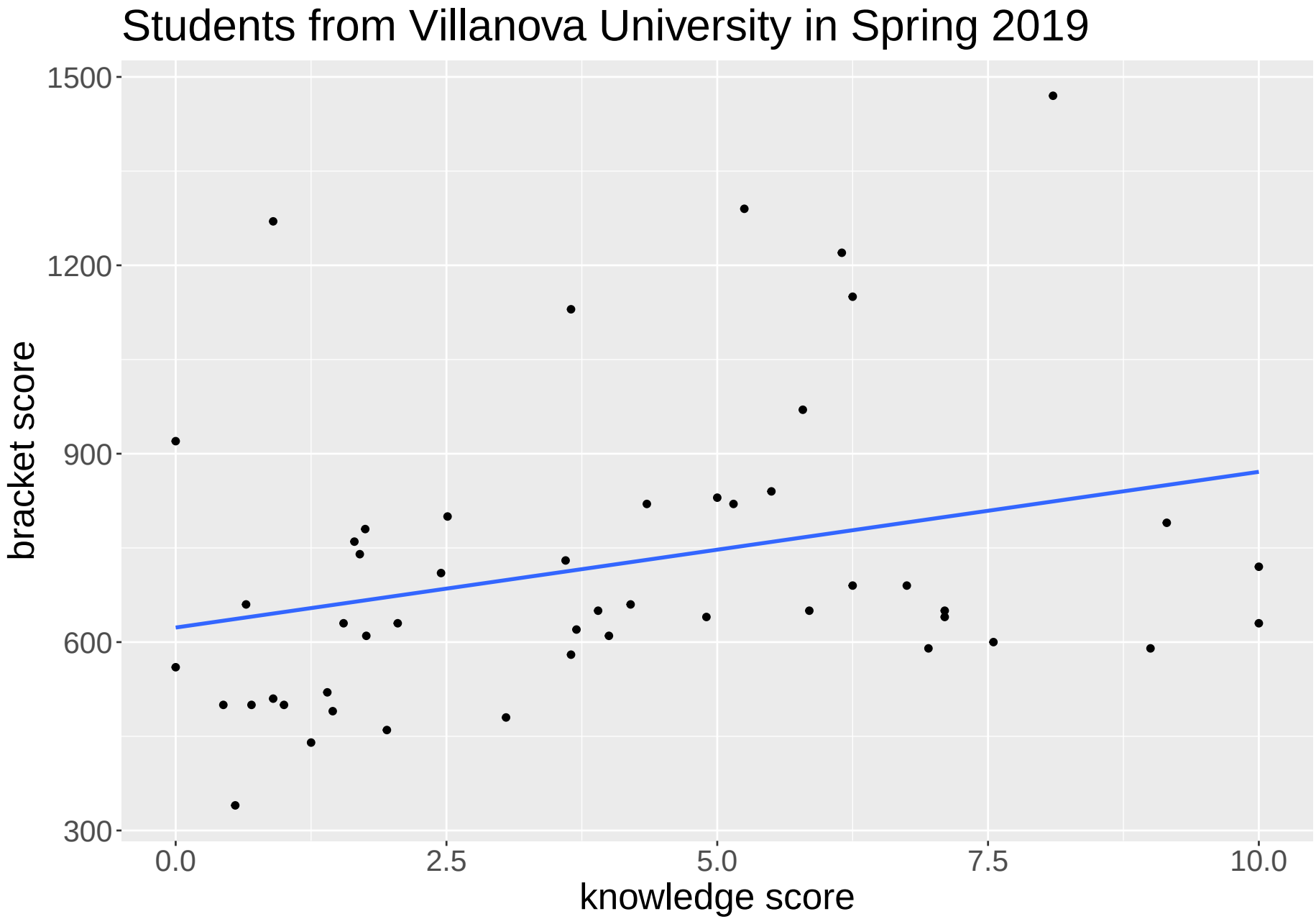
Thursday, April 30
Spring Quarter 2026
Peter Chi

Conditions for Validity in Linear Regression

Recall: these are the required conditions for statistical inference in the context of a linear model to be valid:

- The relationship between X and Y, if there is one, is actually **L**inear
  - e.g. not quadratic, exponential, etc.
- **I**ndependence of observations
- **N**ormality of ϵ_i
  - Note that this can also be achieved, due to the central limit theorem, with a large sample size even if ϵ_i does not follow a normal distribution
- **E**qual variance across all values of X
  - Also known as homoskedasticity

Example: March Madness knowledge vs. bracket scores



Example: March Madness knowledge vs. bracket scores

The data

1. Students from a class I was teaching in Spring 2019 at Villanova University were asked to fill out March Madness brackets for the NCAA Men's College Basketball Championship (n=50).

2. They also filled out a questionnaire to assess their "basketball knowledge." The questions were things like:
   - How many men's college basketball games did you watch this season?
   - On how many teams in the tournament can you name at least one player? Two players? The head coach?
   - When watching a basketball game, how often are you able to identify any particular strategies that are being used?

3. Their responses to the questions were summarized in a "knowledge score" from 0 to 10.

4. Brackets were scored according to default ESPN settings (10 pts for Round 1 picks, 20 pts for Round 2 picks, etc).

Example: March Madness knowledge vs. bracket scores

In Lab 4, you investigated the consequences of violations to the required conditions for validity on Type I Error rates in the context of simple linear regression, and also did a small bit of model diagnostics (specifically, the Q-Q plot).

So what are we doing today?

Today we will run through a complete set of model diagnostics for each of the LINE conditions

Note: we can never actually know for sure if the conditions are met or not (just like we can never know if we had made a Type I Error). But, we CAN (and should) look at what the data suggest.

Quote of the Day

Machine learning is statistics minus any checking of models and ~~assumptions~~ conditions.

Dr. Brian Ripley
Member of R Core Development Team
Professor of Applied Statistics (retired)
University of Oxford

Residual Analyses

Why do we do model diagnostics when doing statistical inference?

Statistical tests have conditions that need to be satisfied in order for the test to be valid.

Why don't we need model diagnostics when doing machine learning?

The goal in that setting is model building and prediction! These things do not require any conditions to be satisfied.

On the other hand, note that violation of conditions can lead to greater prediction error. However, as your machine learning workflow/algorithms will by definition basically seek to minimize prediction error, you don't usually care too much about how you got there.

Residual Analyses

Typically, regression model diagnostics proceed by utilizing the residuals of the model. What's a residual again??

Recall from Lecture #7:

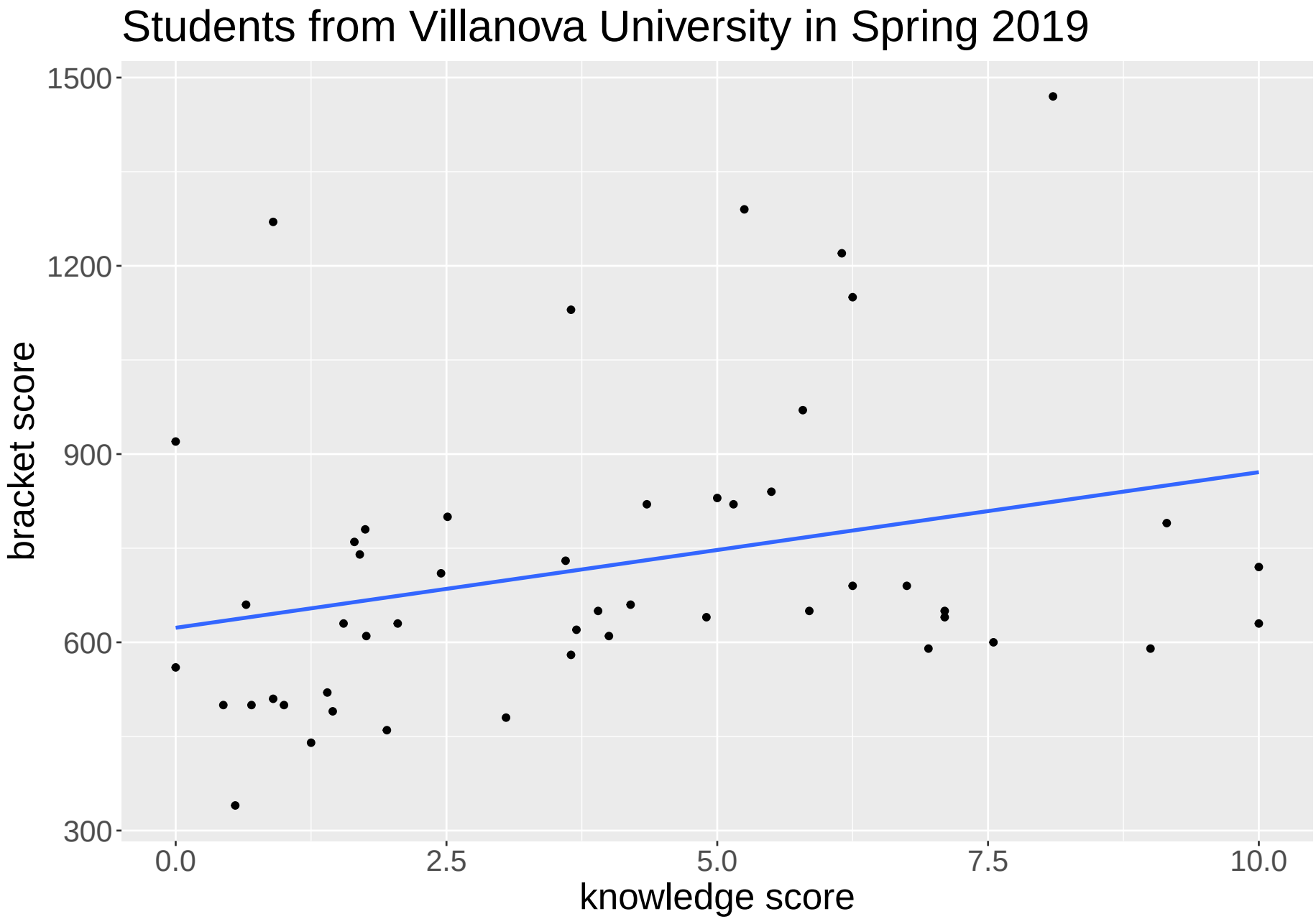
The Ordinary Least Squares Criterion is:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Each $\epsilon_i = y_i - \hat{y}_i$ is a residual.
- $\hat{y}_i$ is the fitted value of y for the i^{th} point, according to the ordinary least squares regression line.

What are the residuals for the March Madness data?

Residual Analysis



Residual Analysis

Recall that the equation for the blue line came from:

```
model1 <- lm(bracket_score ~ knowledge, data=march_madness)
summary(model1)
```

```
##
## Call:
## lm(formula = bracket_score ~ knowledge, data = march_madness)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -296.80 -146.10  -68.58   82.26  645.99
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   623.16      56.76  10.978 1.09e-14 ***
## knowledge      24.80      11.72   2.116   0.0395 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 230.1 on 48 degrees of freedom
## Multiple R-squared:  0.08534,    Adjusted R-squared:  0.06629
## F-statistic: 4.479 on 1 and 48 DF,  p-value: 0.03954
```

Residual Analysis

$$\hat{y}_i = 623.16 + 24.8 \cdot x_i$$

where x_i is the knowledge score of the i^{th} participant. Here's a snapshot of the first few rows of the `march_madness` dataframe:

bracket_score	knowledge
830	5.00
610	1.76
690	6.75
730	3.60
820	4.35
510	0.90

What are the residuals? https://pollev.com/chi

Residual Analysis

Thankfully, we don't actually have to do that every time (I just wanted us to do it to make sure we know how to get them in principle). The `model1` object has them for us:

```
names(model1)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

Let's grab the first 6 just to compare to the table:

```
model1$residuals[1:6]
```

```
##         1         2         3         4         5         6
##  82.86029 -56.80121 -100.53242  17.57446  88.97758 -135.47679
```

Residual Analysis

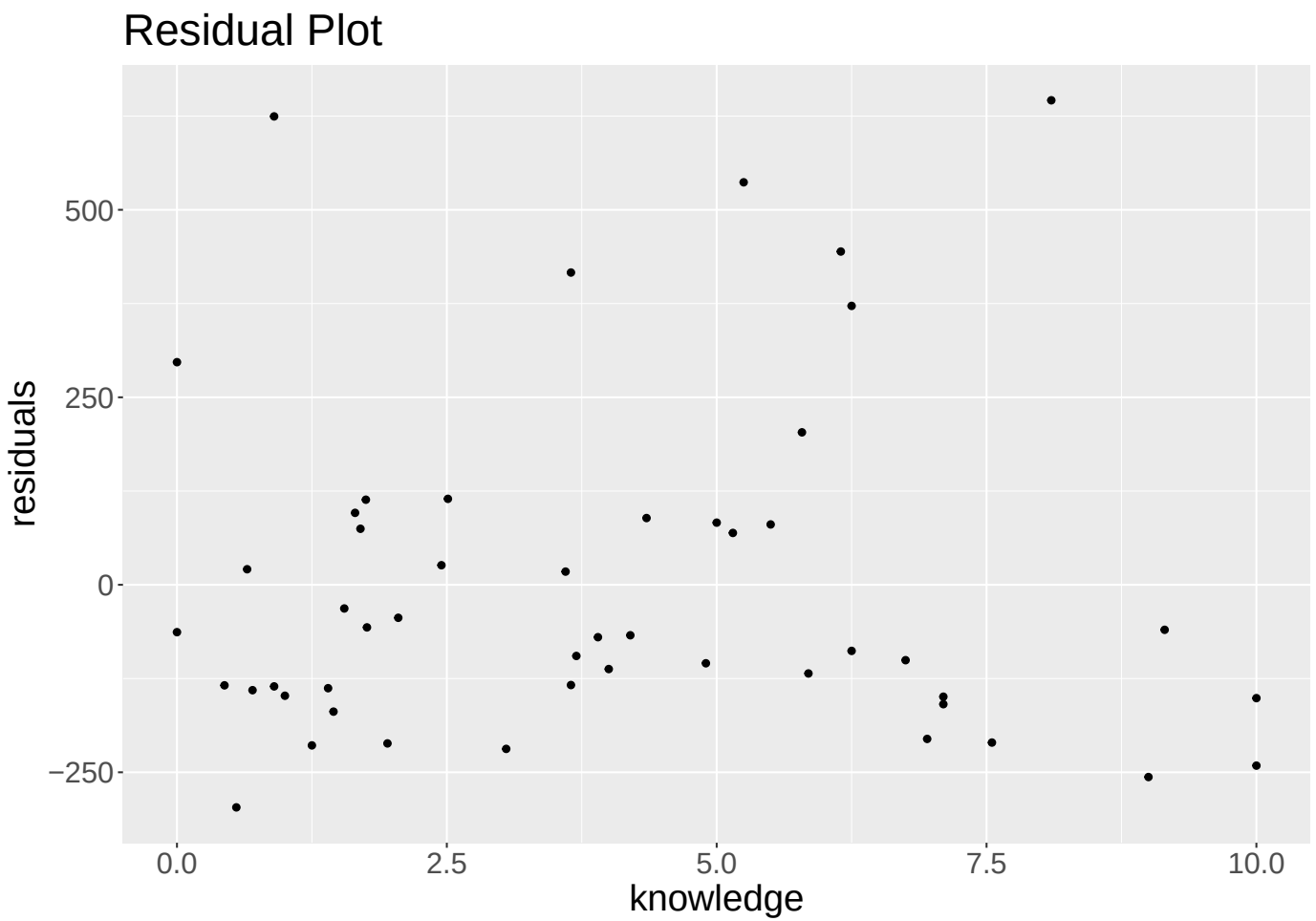
Now, how do we use them?

- The relationship between X and Y, if there is one, is actually **L**inear
  - e.g. not quadratic, exponential, etc.
- **I**ndependence of observations
- **N**ormality of ϵ_i
  - Note that this can also be achieved, due to the central limit theorem, with a large sample size even if ϵ_i does not follow a normal distribution
- **E**qual variance across all values of X
  - Also known as homoskedasticity

Residual Analysis

Equal variance across all values of X

```
march_madness$residuals <- model1$residuals
ggplot(data=march_madness, mapping=aes(x=knowledge, y=residuals)) +
  geom_point() + labs(title="Residual Plot")
```



Residual Analysis

Some Questions

- What does this tell us that just looking at the original scatterplot doesn't?
  - Answer: Often, not that much! But, it CAN make patterns more apparent.

Residual Analysis

Some Questions

- What does this tell us that just looking at the original scatterplot doesn't?
  - Answer: Often, not that much! But, it CAN make patterns more apparent.

- But wait so I'm supposed to just look at it and make a judgment call? What about a definitive answer??
  - Yeah... while you could conceivably cook up a statistical test for this question, it is not typically done.

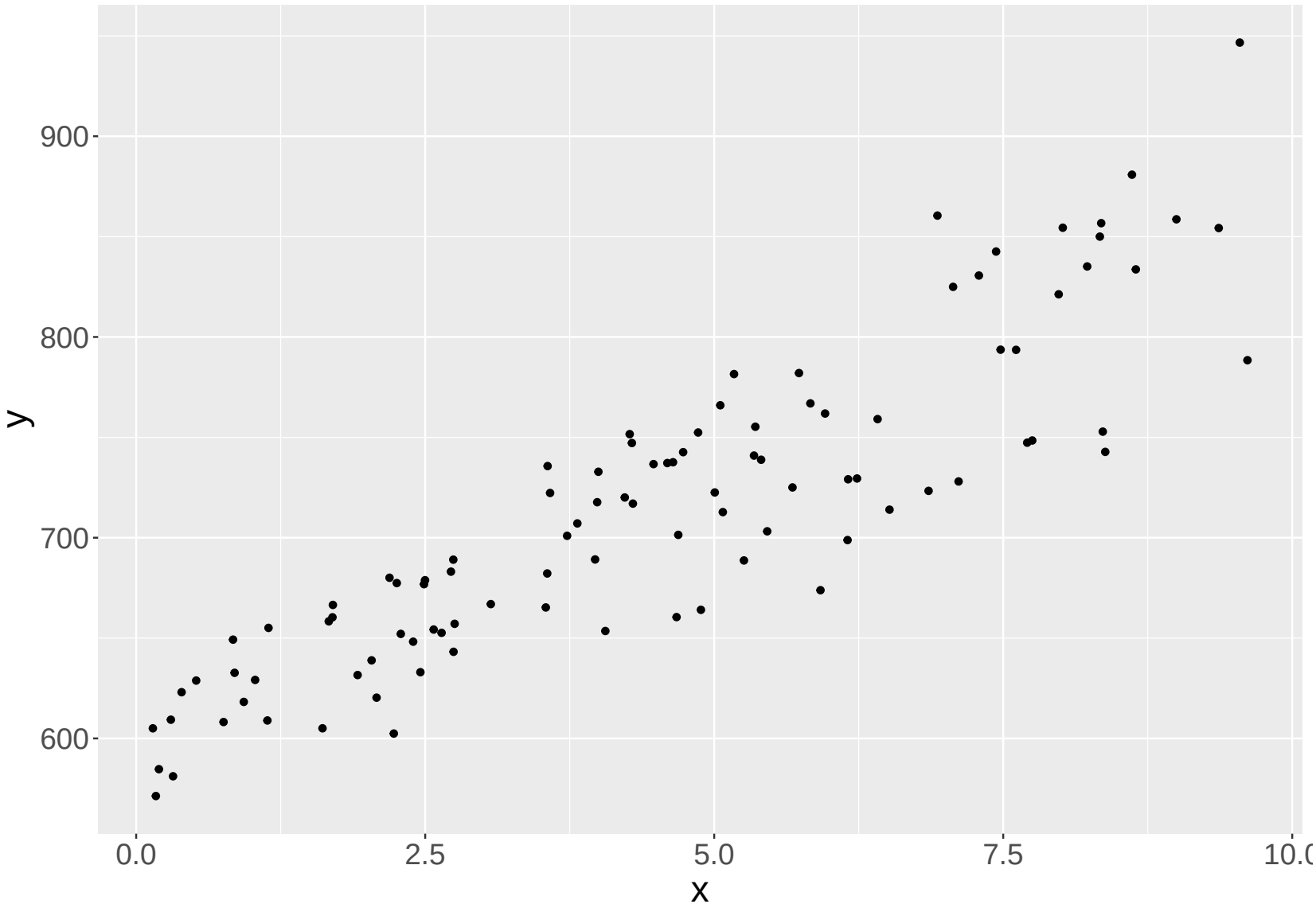
Residual Analysis

Some Questions

- What does this tell us that just looking at the original scatterplot doesn't?
  - Answer: Often, not that much! But, it CAN make patterns more apparent.
- But wait so I'm supposed to just look at it and make a judgment call? What about a definitive answer??
  - Yeah... while you could conceivably cook up a statistical test for this question, it is not typically done.
- Ok, so what judgement call would you make in this case?
  - Here, it looks like there is not a whole lot of difference in the vertical spread of the points across all values of x, so it looks like the condition of homoskedasticity is probably satisfied.

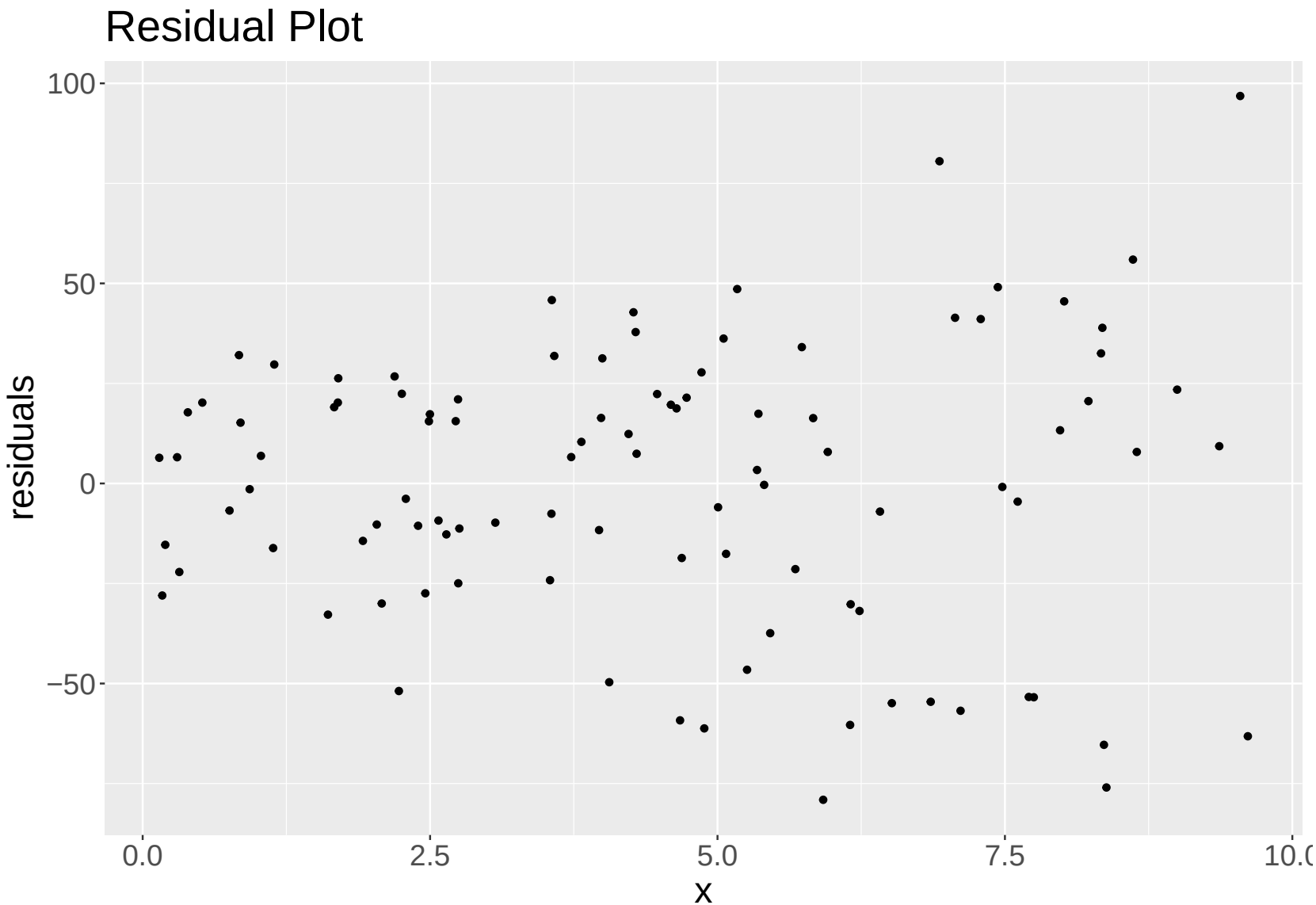
Residual Analysis

Here is another example:



Residual Analysis

And here the residual plot makes the problem more obvious:



Residual Analysis

Equal variance across all values of X

What do you think are the consequences of this condition being violated?

https://pollev.com/chi

How do we check if the parameter estimates are biased?

And what do we even mean by "biased?"
It is the difference in a statistic from the true parameter value it is trying to estimate. That is, e.g. for β_1:

$$bias = \hat{\beta}_1 - \beta_1$$

Wait but we can't ever know what this is right?

True. But, we can get simulated estimates under any given scenario.

Your Turn #1

Suppose that we have a true relationship of:

$$y_i = 600 + 25x_i + \epsilon_i$$

where $\epsilon_i \sim N(0, 5x_i)$. Then,

- We can simulate data according to this model
- Get $\hat{\beta}_1$ from the linear regression output
- Calculate $\hat{\beta}_1 - \beta_1$
- Do this lots of times to get an estimate of the bias

Residual Analysis

Equal variance across all values of X

Your Turn #1

```
reps <- 10000
beta_1 <- NULL

for(i in 1:reps){
  y <- ...
  dat <- data.frame(x=x, y=y)
  beta_1[i] <- ...
}

mean(...)
```

Residual Analysis

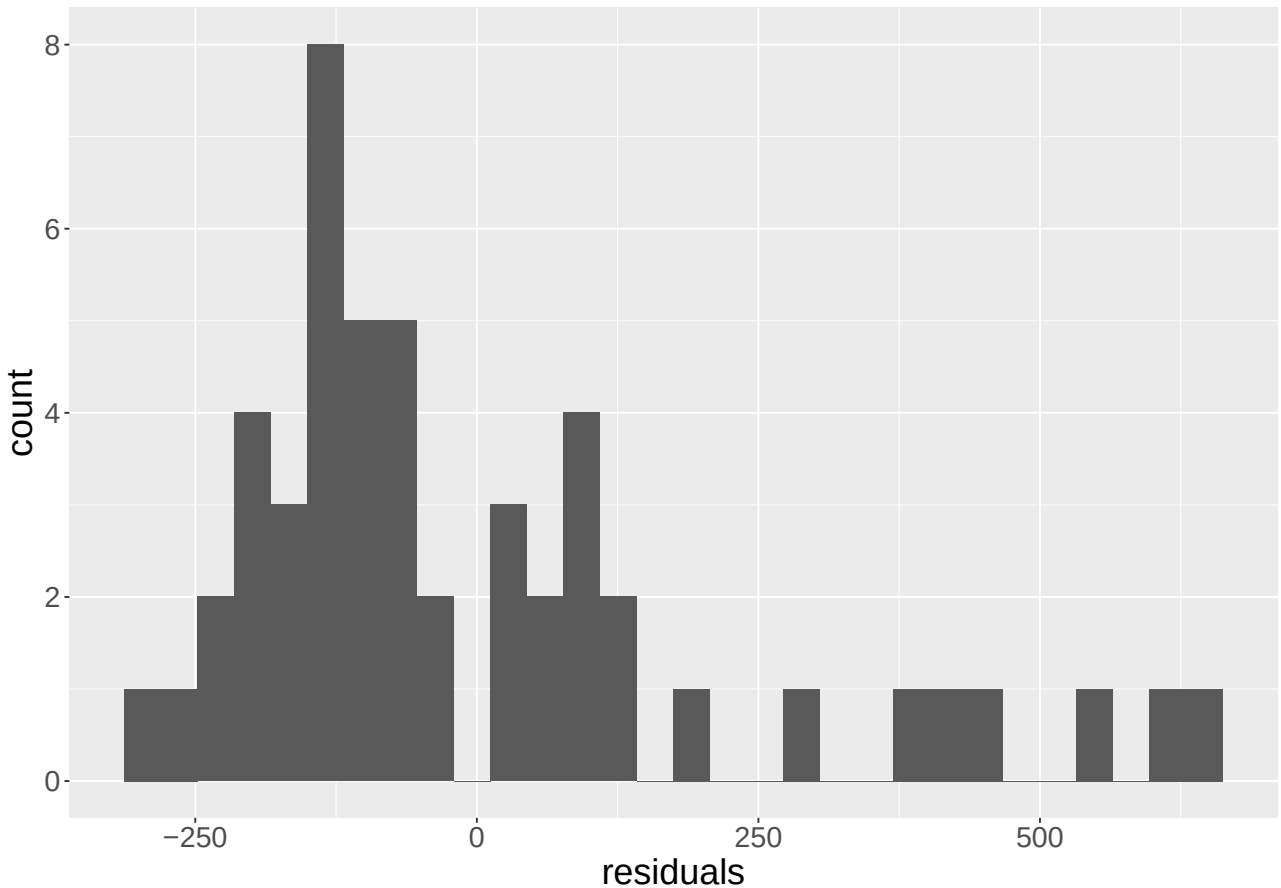
Now, how do we use them?

- The relationship between X and Y, if there is one, is actually **L**inear
  - e.g. not quadratic, exponential, etc.
- **I**ndependence of observations
- **N**ormality of ϵ_i
  - Note that this can also be achieved, due to the central limit theorem, with a large sample size even if ϵ_i does not follow a normal distribution
- **E**qual variance across all values of X
  - Also known as homoskedasticity

Residual Analysis

Normality of ϵ_i

```
ggplot(data=march_madness, aes(x=residuals)) +
  geom_histogram()
```



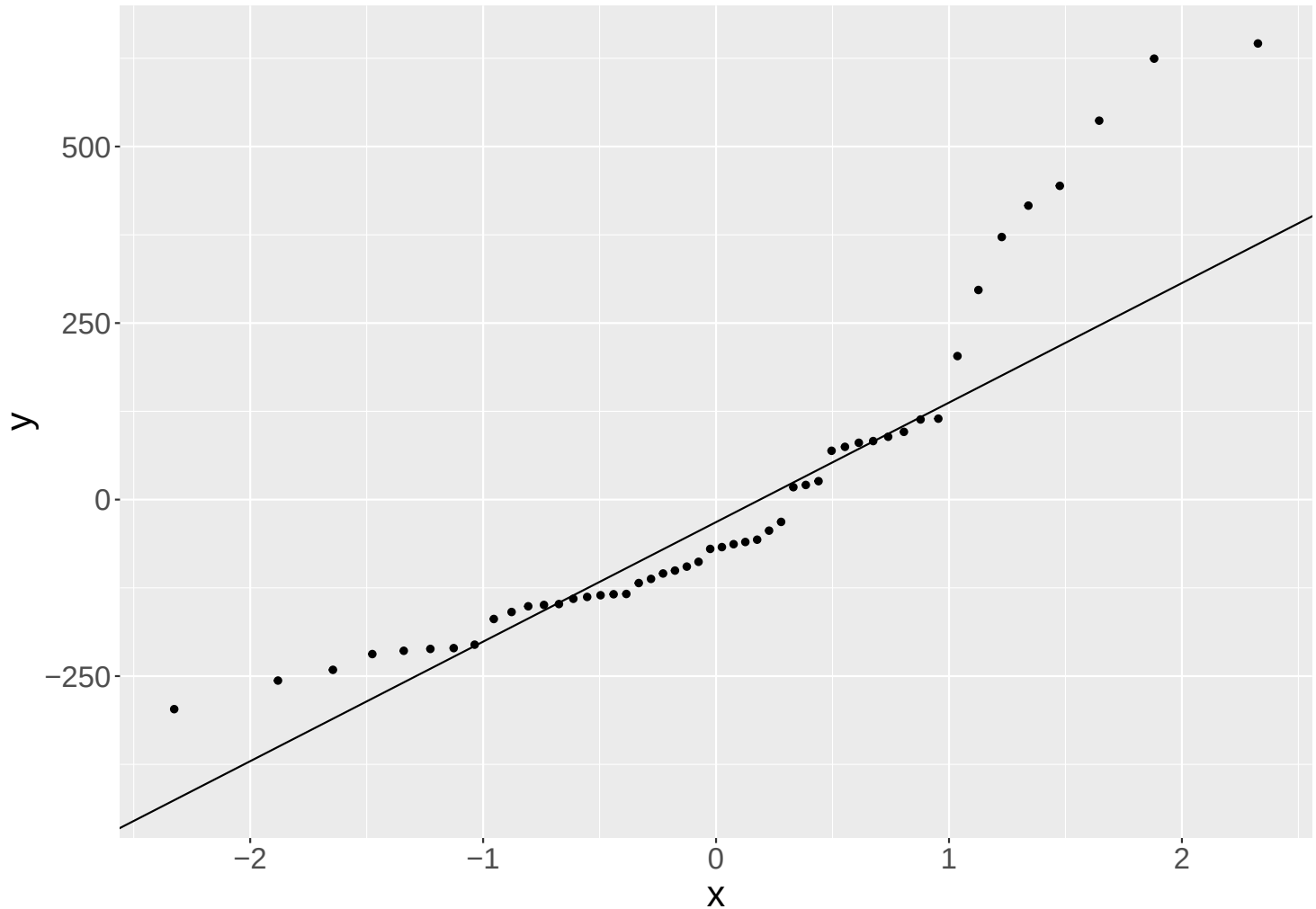
Some Questions

- Again, I'm supposed to just look at it and make a judgment call? What about a definitive answer??
  - For assessing normality, there actually is a statistical test that is often used!
  - Also, there's the QQ-plot

Residual Analysis

Normality of ϵ_i

```
ggplot(march_madness, aes(sample=residuals)) +
  stat_qq() + stat_qq_line()
```



Residual Analysis

Normality of ϵ_i

```
shapiro.test(march_madness$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  march_madness$residuals
## W = 0.85226, p-value = 1.782e-05
```

H_0 for this test is that the residuals do follow a normal distribution.

Residual Analysis

Now, how do we use them?

- The relationship between X and Y, if there is one, is actually **L**inear
  - e.g. not quadratic, exponential, etc.
- **I**ndependence of observations
- **N**ormality of ϵ_i
  - Note that this can also be achieved, due to the central limit theorem, with a large sample size even if ϵ_i does not follow a normal distribution
- **E**qual variance across all values of X
  - Also known as homoskedasticity

Residual Analysis

Independence of observations

- One way to check this is to see if there is any temporal trend in the data. That is, is there any relationship between datapoints over time?

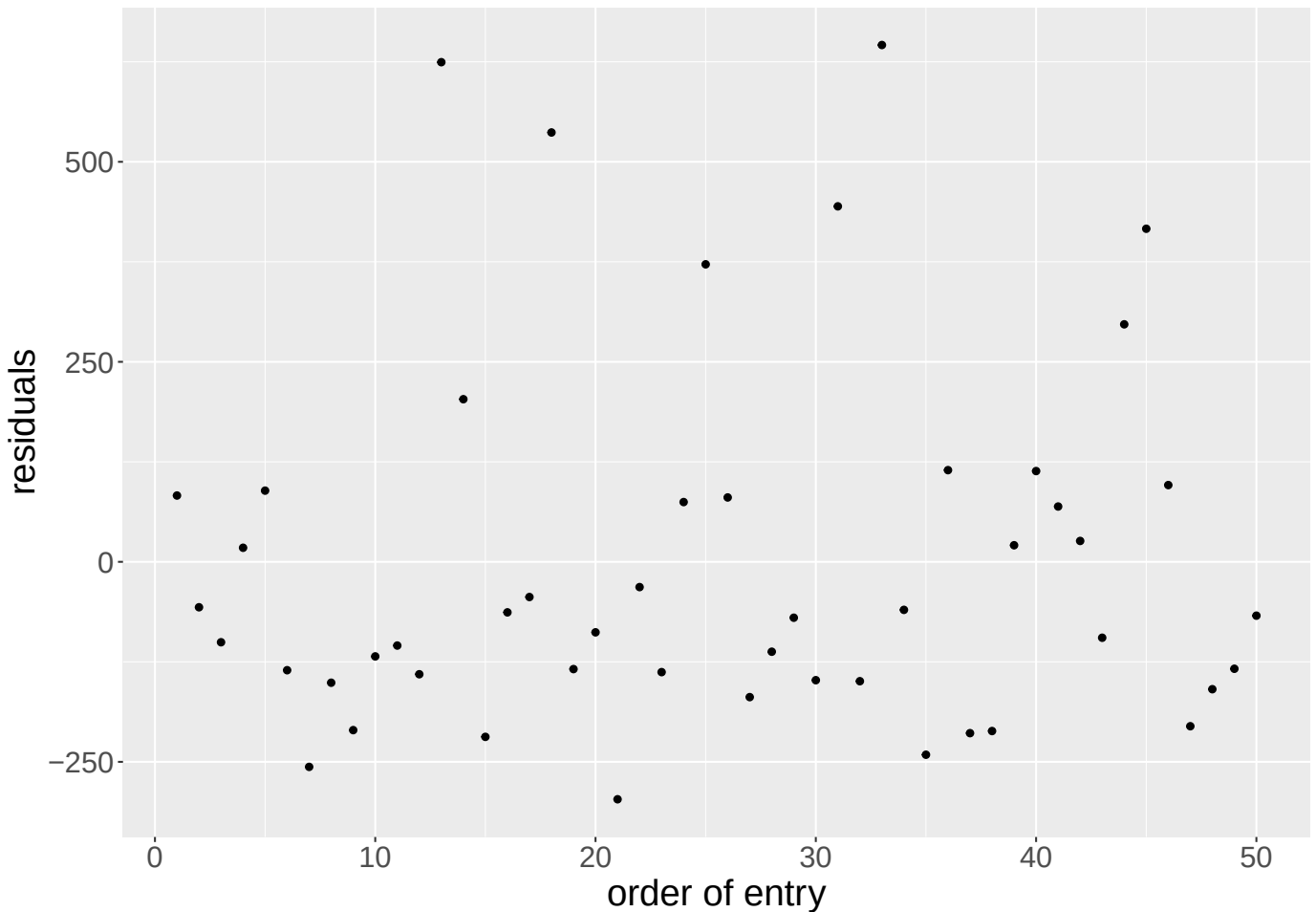
- Note that this is not the ONLY way in which the independence observation can be violated. But it is often the only one that is even possible to check.

- The March Madness dataset is in the order in which students signed up to be in the pool...

Residual Analysis

Independence of observations

```
march_madness$order <- 1:dim(march_madness)[1]
ggplot(march_madness, aes(x=order, y=residuals)) +
  geom_point() + labs(x="order of entry")
```



Residual Analysis

Independence of observations

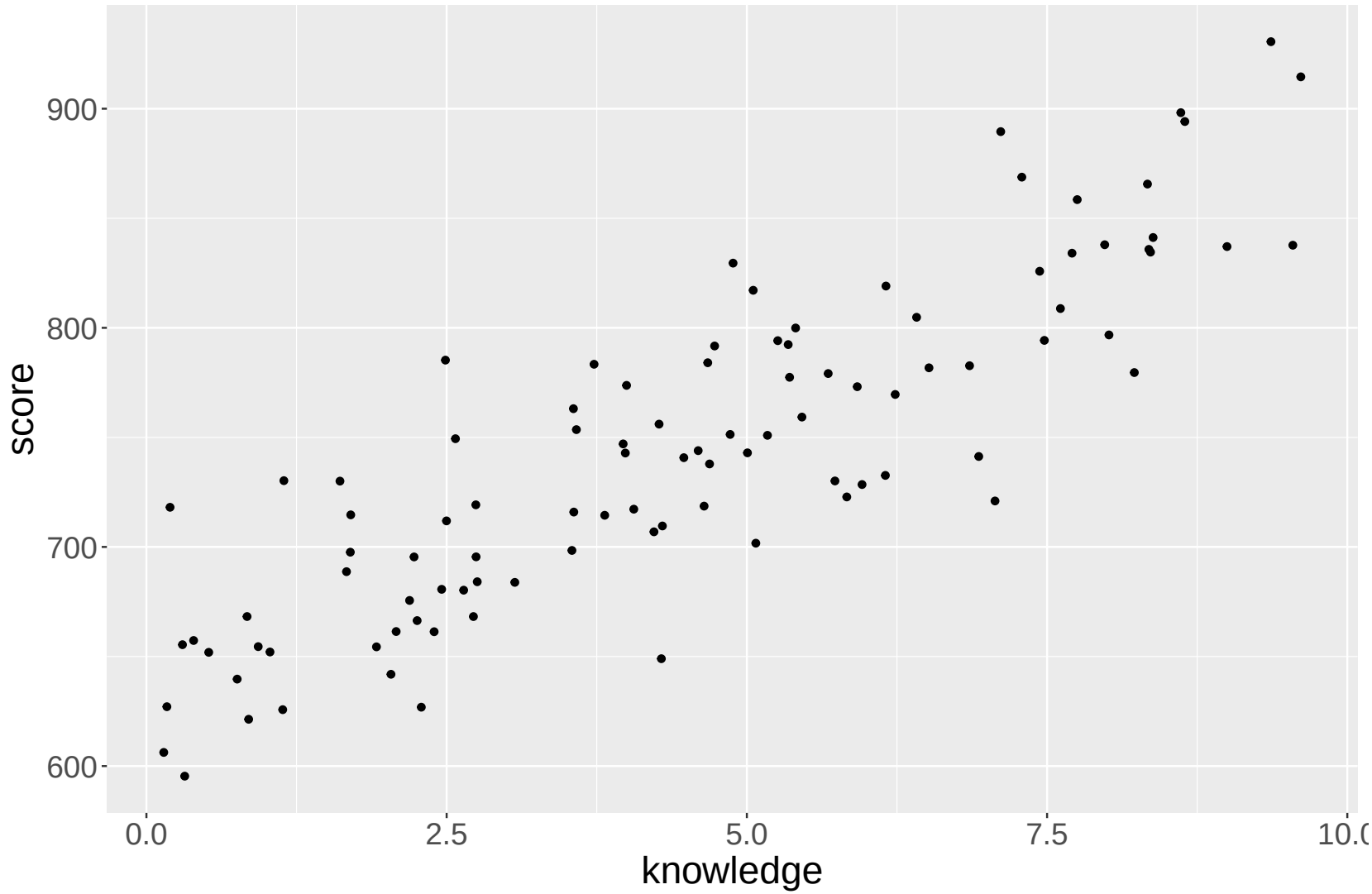
- There does not appear to be any discernible pattern in the residuals over time.

- What might it look like if there was a temporal trend?
  - That is, for example, what if those who signed up later tended to have higher scores?

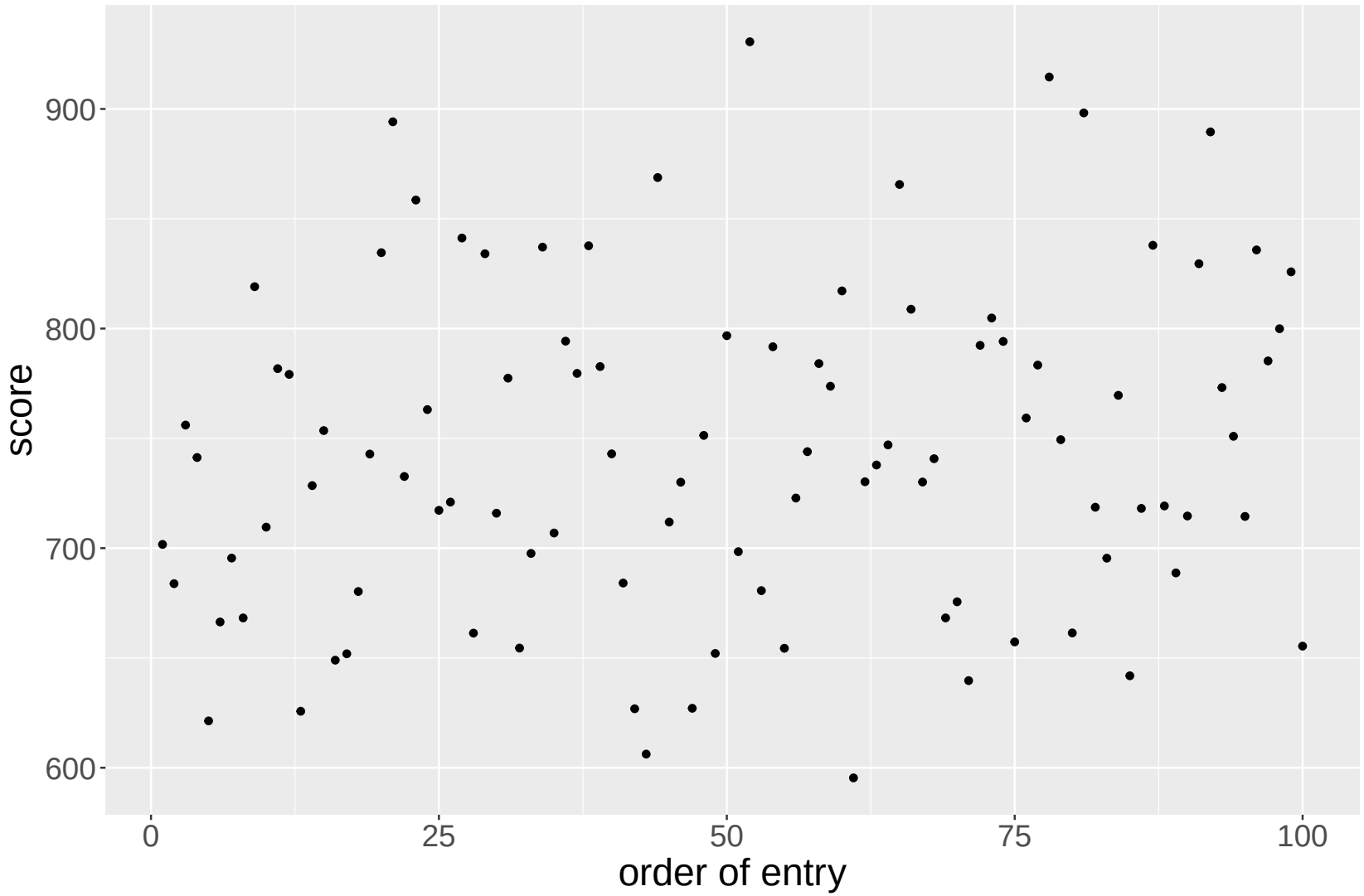
Residual Analysis

Independence of observations

From this plot, you can't tell that anything is wrong...



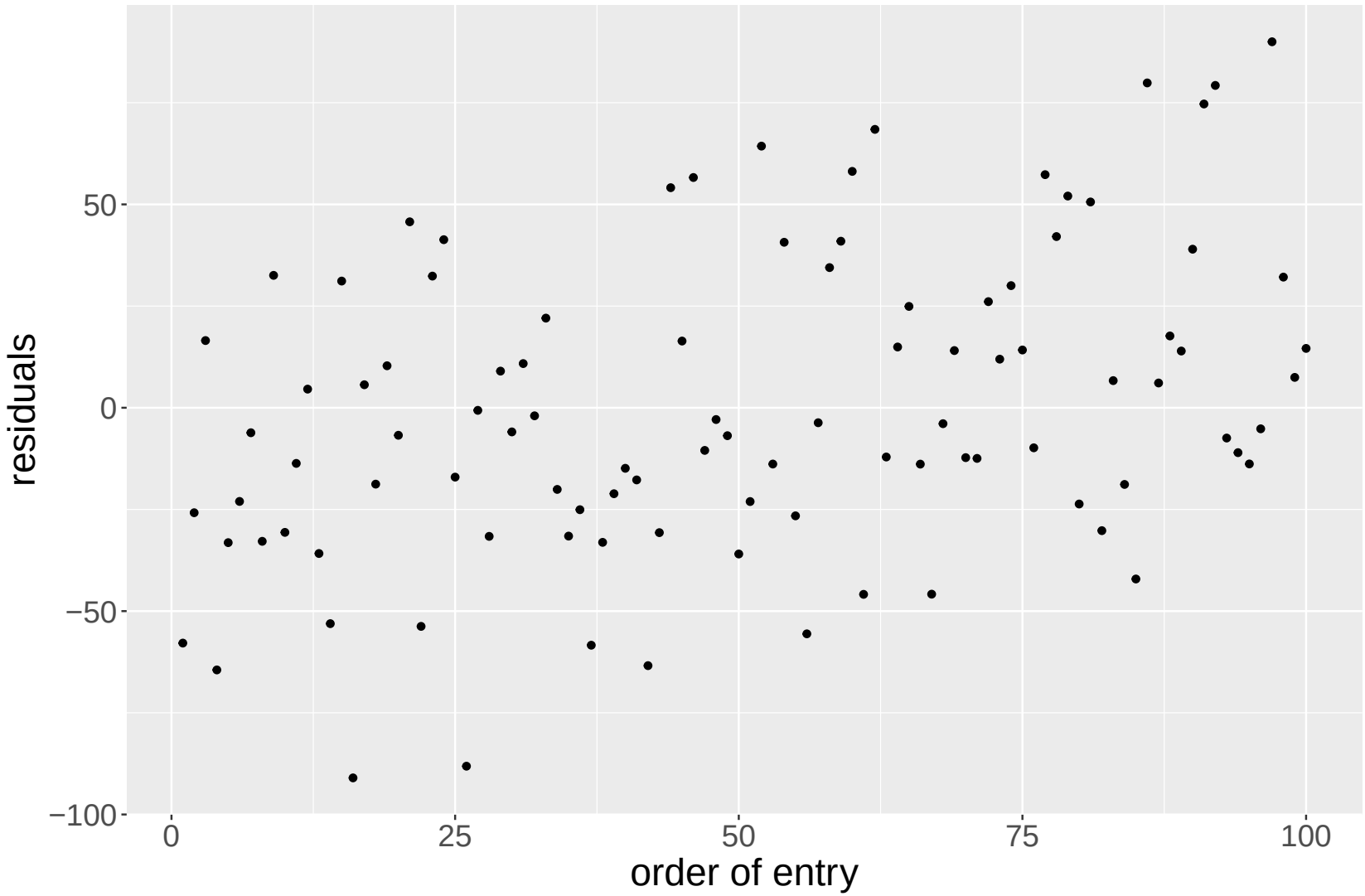
Nor can you with this one either



Residual Analysis

Independence of observations

But with the residual plot you can see the temporal trend



Residual Analysis

Your Turn #2

In Lab 4, you investigated the Type I Error rate in the presence of non-independence. Let's now investigate what happens to statistical power.

Simulation study:

- Simulate 100 samples of $x \sim Unif(0, 10)$
- Let t represent the order of data collection, and just go from 1 to 100
- Generate $y = x + 0.5t + \epsilon$ where $\epsilon_i \sim N(\mu = 0, \sigma = 10)$
- Run the following linear models:
  - $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  - $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 t$
- Check whether $H_0 : \beta_1 = 0$ is rejected in each case
- Do this repeatedly (1,000 times) to get an estimate of power in each scenario

Comment briefly on what you observe.

Residual Analysis

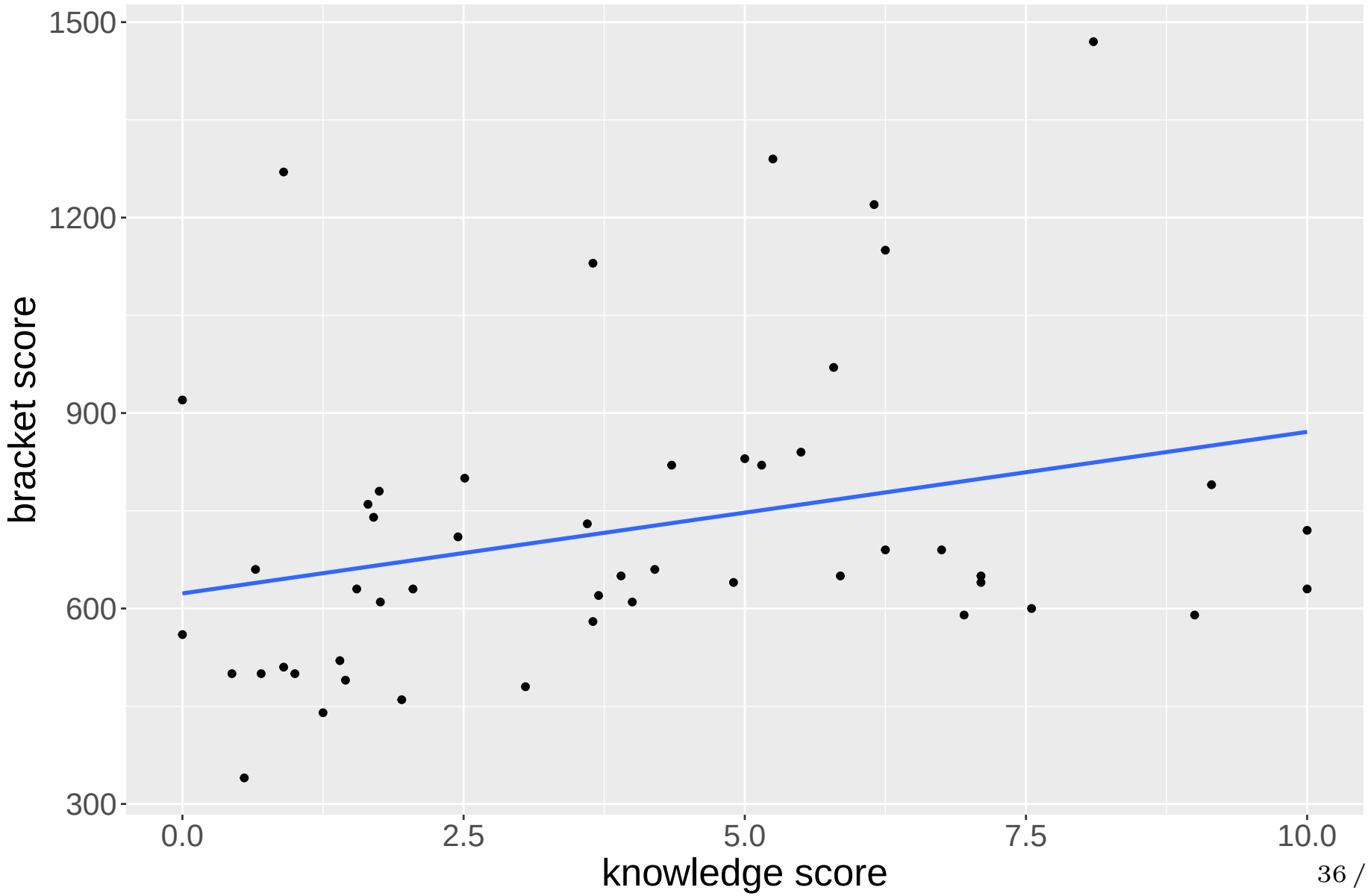
Now, how do we use them?

- The relationship between X and Y, if there is one, is actually **L**inear
  - e.g. not quadratic, exponential, etc.
- **I**ndependence of observations
- **N**ormality of ϵ_i
  - Note that this can also be achieved, due to the central limit theorem, with a large sample size even if ϵ_i does not follow a normal distribution
- **E**qual variance across all values of X
  - Also known as homoskedasticity

Residual Analysis

The relationship between X and Y, if there is one, is actually **Linear**

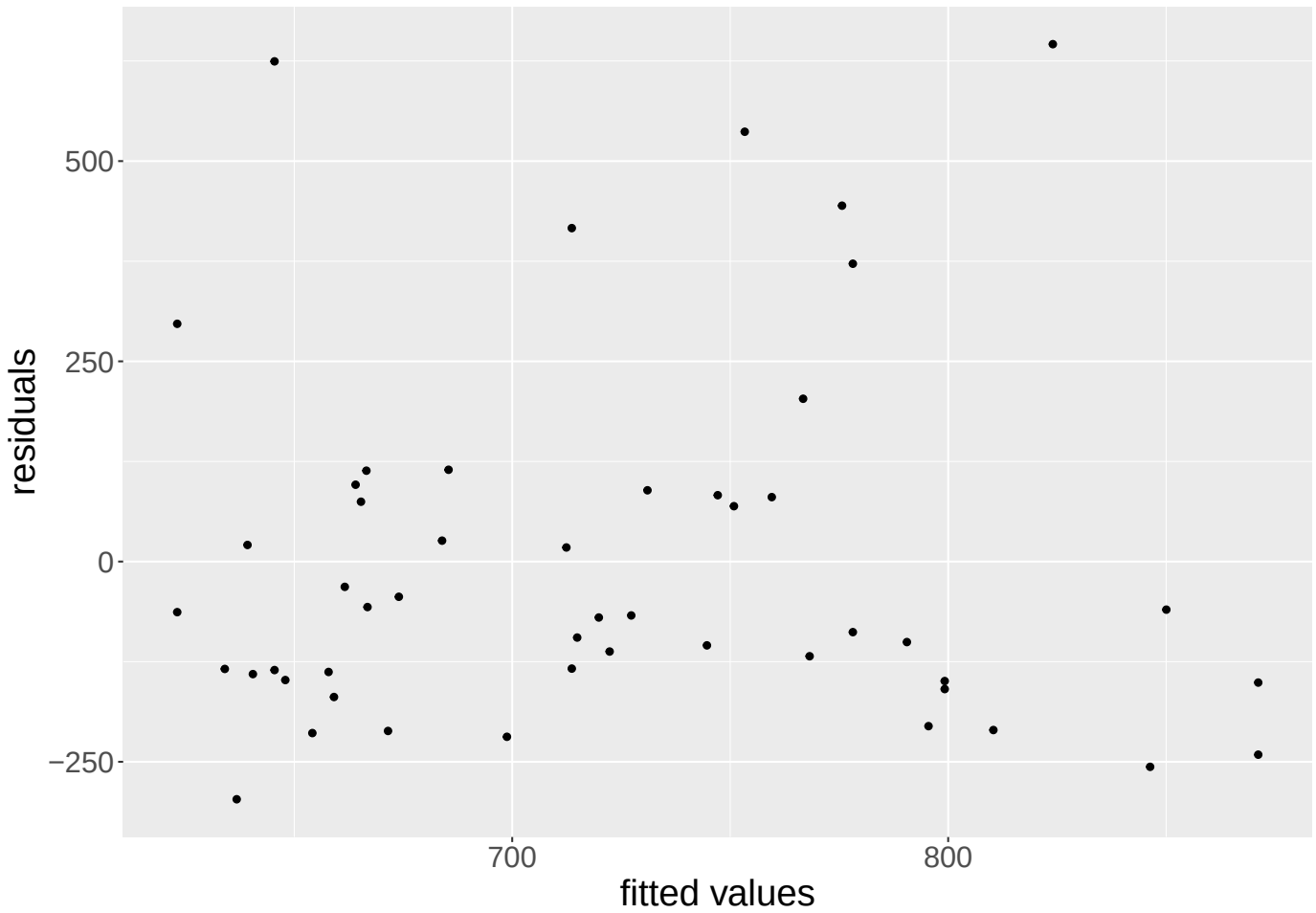
It looks fairly reasonable...



Residual Analysis

The relationship between X and Y, if there is one, is actually **Linear**

```
march_madness$fitted <- model1$fitted.values
ggplot(data=march_madness, aes(x=fitted, y=residuals)) +
  geom_point() + labs(x="fitted values")
```



Residual Analysis

The relationship between X and Y, if there is one, is actually **Linear**

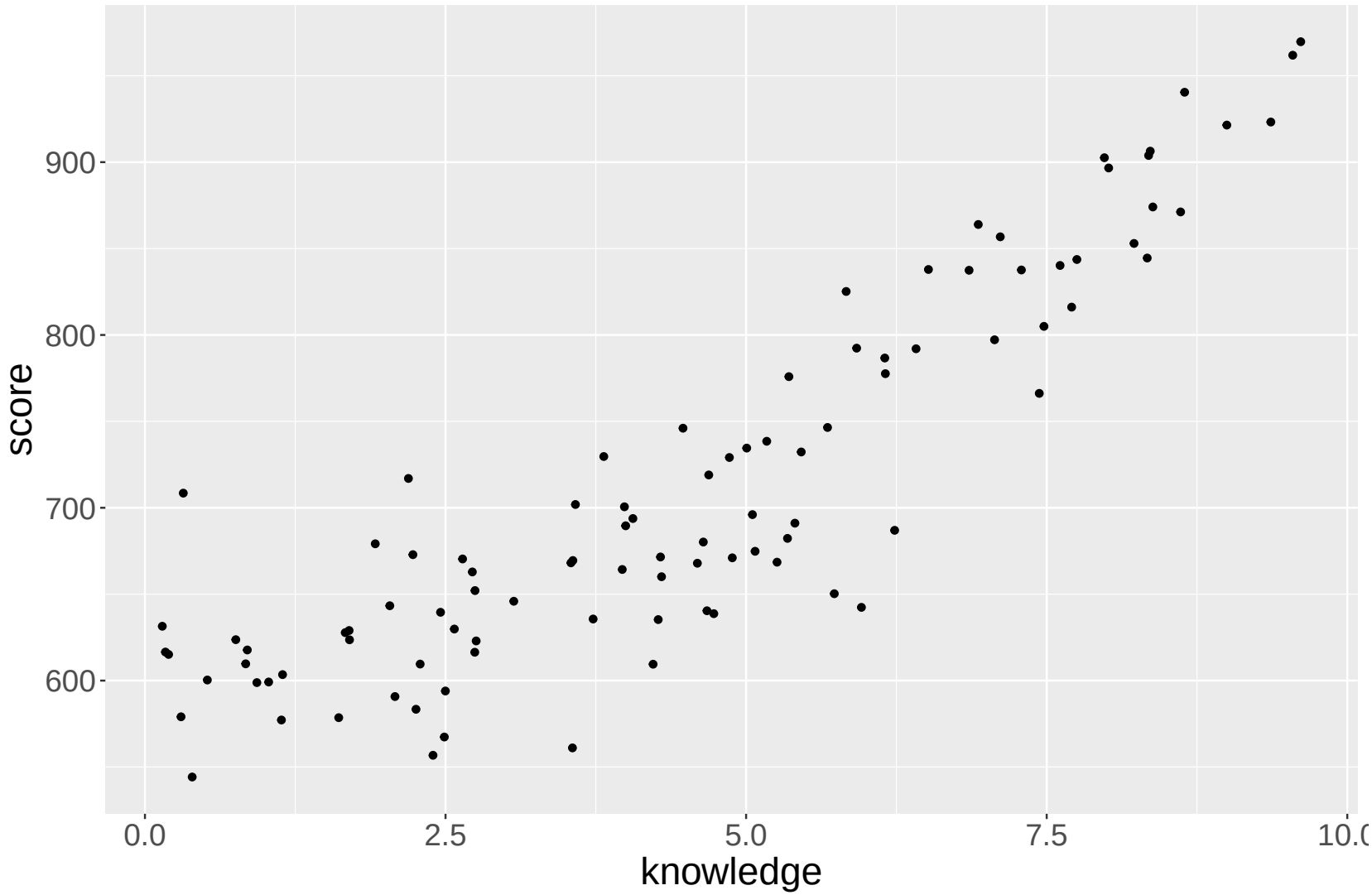
- There does not appear to be much of a pattern here

- What might it look like if the relationship was NOT linear?
  - For example, a quadratic relationship

Residual Analysis

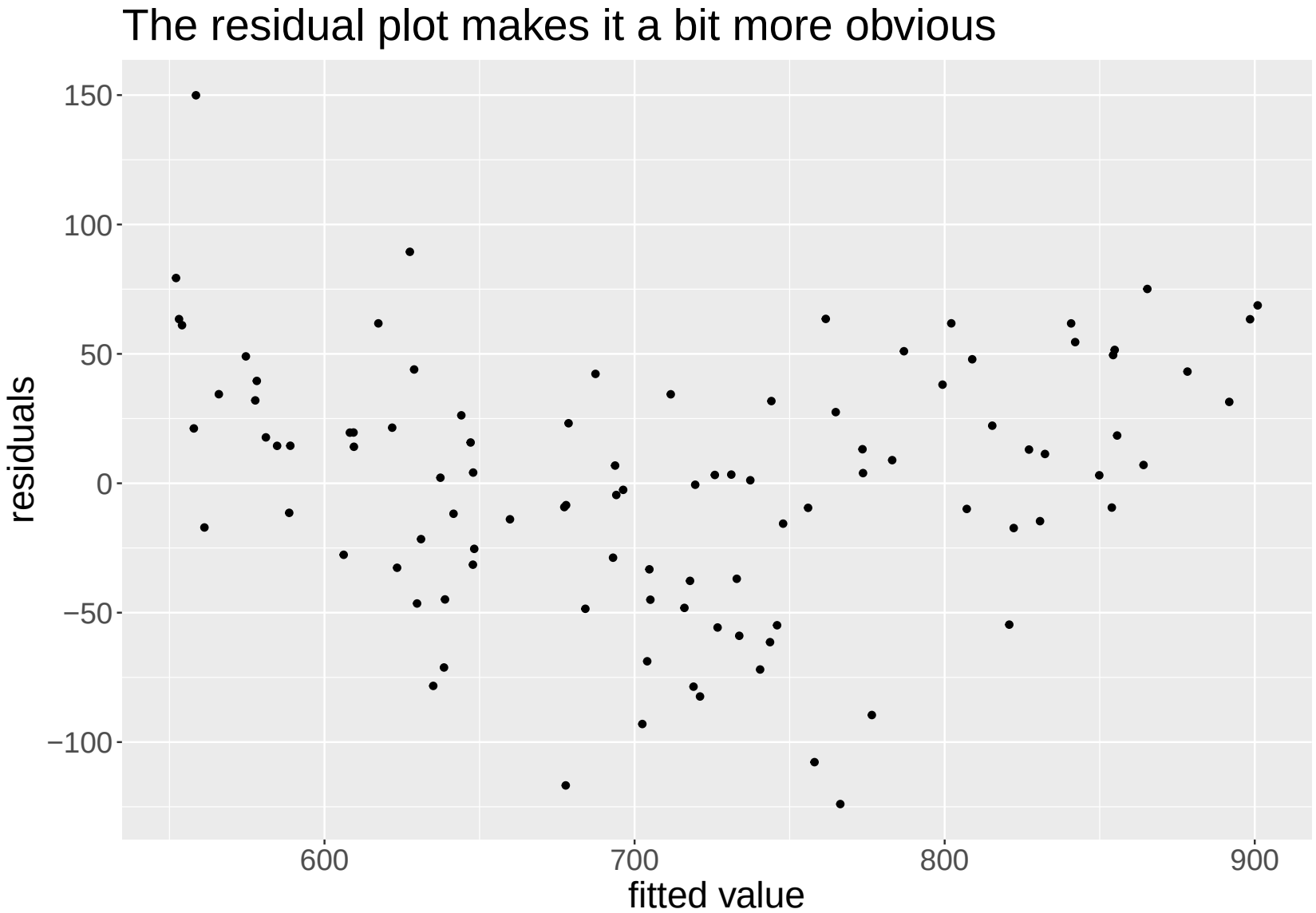
The relationship between X and Y, if there is one, is actually **Linear**

From this plot, it is not that obvious that anything is wrong...



Residual Analysis

The relationship between X and Y, if there is one, is actually **Linear**



Recap

A complete residual analysis would consist of all of the following:

- The relationship between X and Y, if there is one, is actually Linear
  - **Diagnostic: residuals vs. fitted values**
- Independence of observations
  - **Diagnostic: residuals vs. order of entry into study**
  - Reminder: this is only ONE possible dependence structure!!
- Normality of ϵ_i
  - **Diagnostics:**
    - **Histogram of residuals**
    - **QQ-plot**
    - **Shapiro-Wilk test**
  - (do all three of these!)
- Equal variance across all values of X
  - **Diagnostic: residuals vs. x values**

Recap

Today's Daily Check

Both of the Your Turns.