DSC 152:
Applied Statistical Data Analysis and Inference

Lecture #7
Simple Linear Regression

Tuesday, April 21
Spring Quarter 2026
Peter Chi

What is Simple Linear Regression?

In DSC 40A...

You likely saw a simple linear regression model written as

$$H^*(x) = w_0^* + w_1^* x$$

Statisticians tend to write it as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

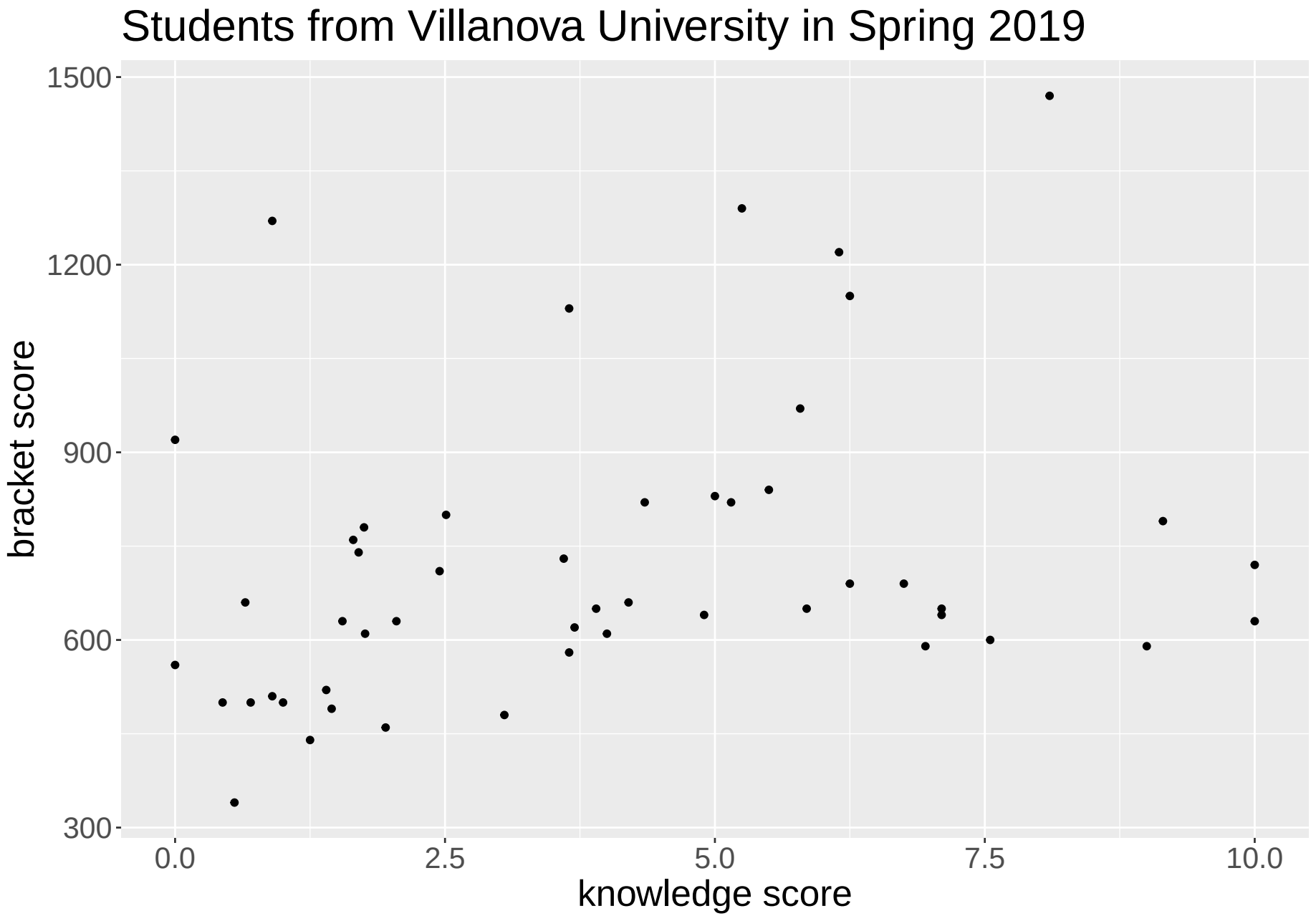
Typically, when we think of simple linear regression, we think of:

- A quantitative outcome variable
- A quantitative predictor variable (or primary covariate)

The "simple" refers to the fact that there is only ONE predictor variable (if there are more, then we are doing multiple linear regression, which we will get to later).

Then, with two quantitative variables we can plot a scatterplot...

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).

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?

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.

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?

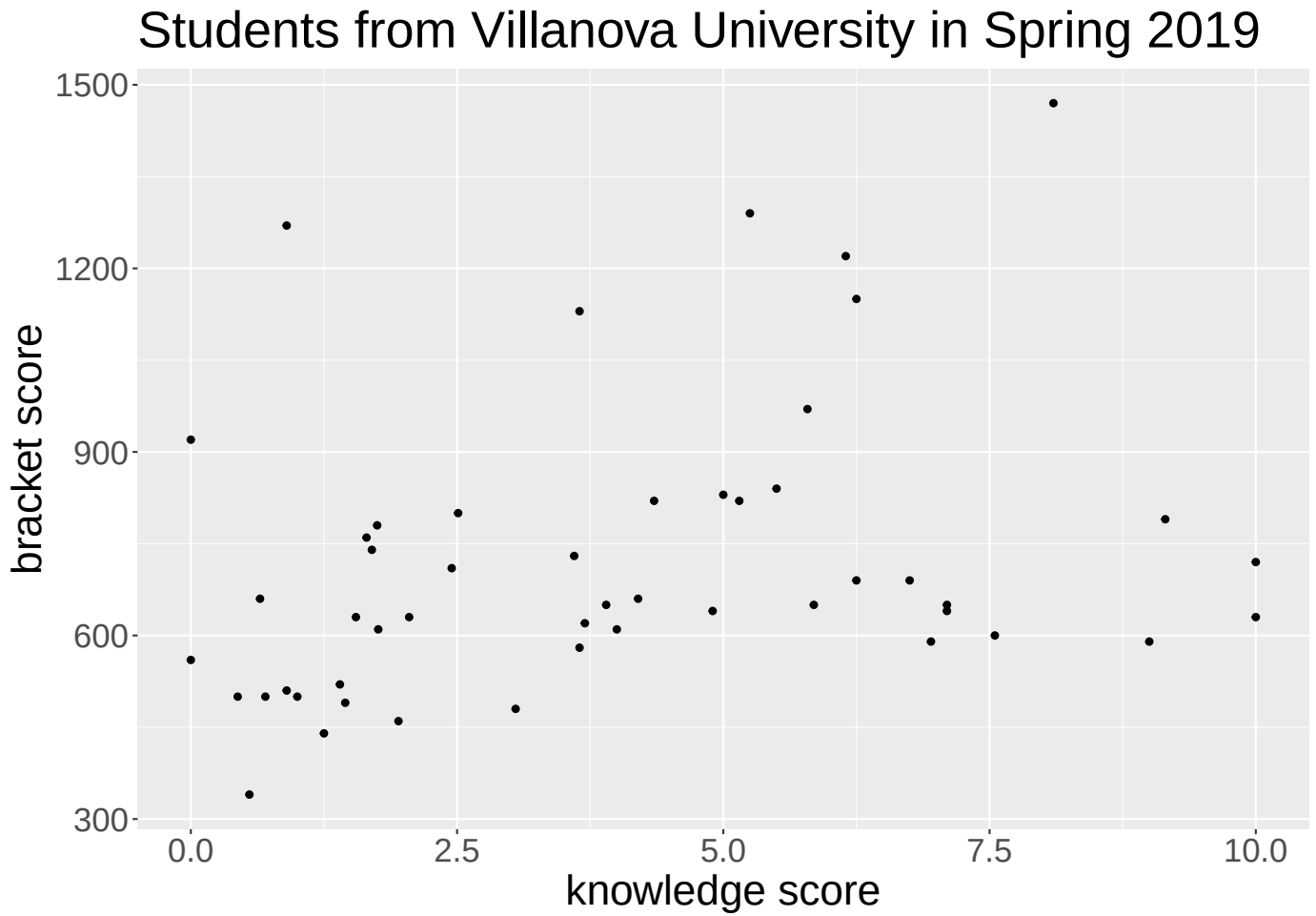
3. Their responses to the questions were summarized in a "knowledge score" from 0 to 10.

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

So,

- x_i is the knowledge score of the i^{th} participant
- y_i is the bracket score of the i^{th} participant



Least Squares Solutions

Then, if we want to model their relationship...

In DSC 40A you derived the least squares solutions:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

(how the heck did we get these again?)

Least Squares Solutions

Then, if we want to model their relationship...

In DSC 40A you derived the least squares solutions:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

(how the heck did we get these again?)

Quote of the Day:

Least Squares Solutions

Then, if we want to model their relationship...

In DSC 40A you derived the least squares solutions:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

(how the heck did we get these again?)

In this particular case these would give:

```
beta1 <- sum((knowledge-mean(knowledge))*(bracket_score-mean(bracket_score))) /
  sum((knowledge-mean(knowledge))^2)
beta1
```

```
## [1] 24.79583
```

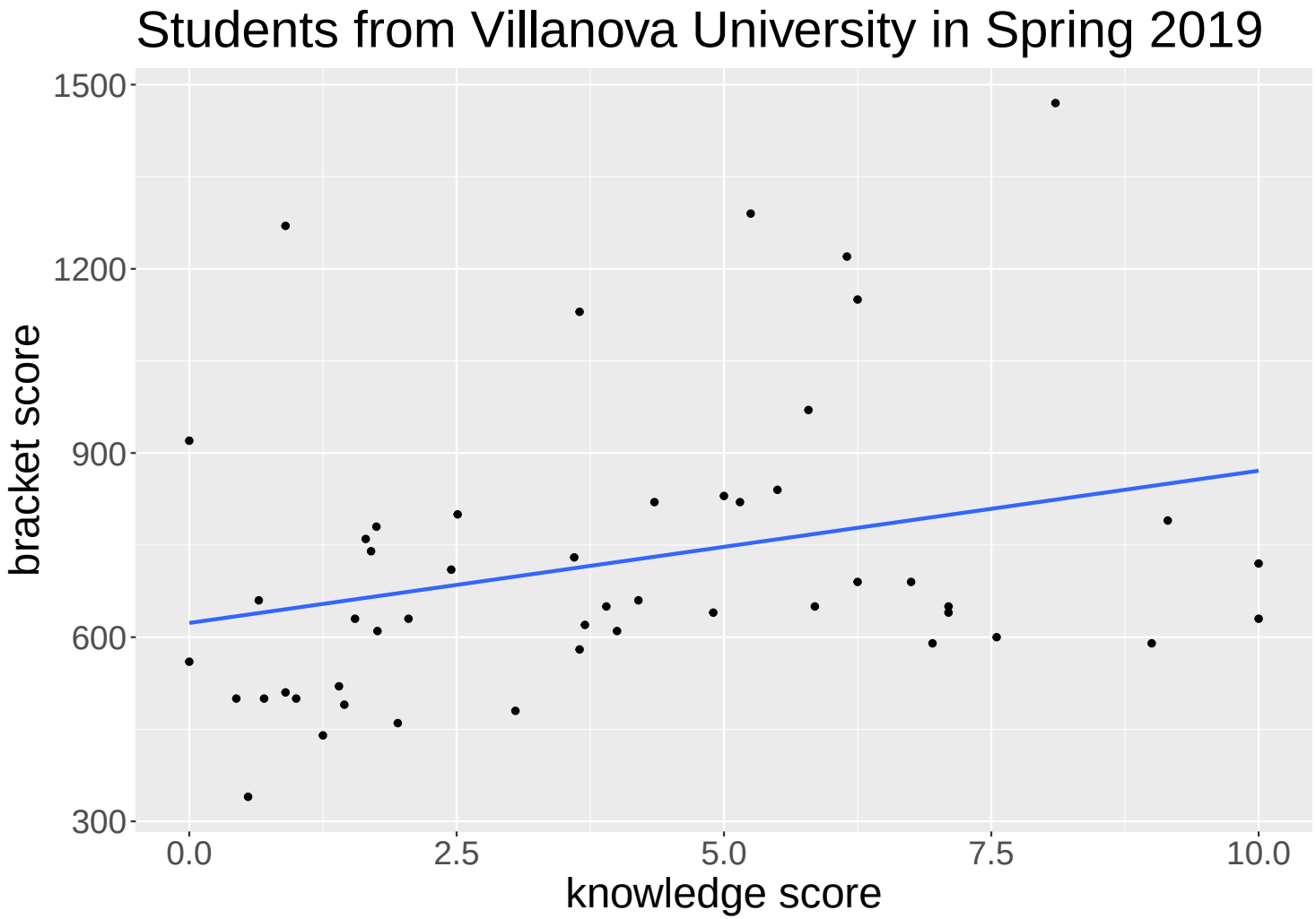
and

```
beta0 <- mean(bracket_score) - beta1*mean(knowledge)
beta0
```

```
## [1] 623.1605
```

Least Squares Regression Line

We can of course then draw the line $\hat{y} = 623.16 + 24.8x$ on the scatterplot:



Least Squares Regression Line

The blue line drawn on the scatterplot is the best fitting line to the data, in the sense that the quantity

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

is the smallest it could possibly be, with this line as the model.

Some notes:

- RSS = residual sum of squares
- In DSC 40A, you may have seen this as:

$$R_{sq}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$$

But it's ultimately the same thing, and either way the point is that the line that we get by minimizing this quantity is the "best" line that could be fit to these data (according to squared error loss).

Sidenote: notation

- y vs. $\hat{y}$
  - and vs. $H(x)$ from 40A
- β_i vs. $\hat{\beta}_i$
  - and vs. w_i and w_i^* from 40A
- What is a "parameter"?
  - ML definition vs. statistical definition
    - In ML, the word "parameter" is often used to refer to any input value for a model.
    - This is NOT how we use the word parameter in statistics!

Sidenote: notation

From DSC 40A, Fall 2024 (and probably most other quarters)

Minimizing multivariate functions

- Our goal is to find the parameters w_0^* and w_1^* that minimize mean squared error:

$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$$

- R_{sq} is a function of two variables: w_0 and w_1.
- To minimize a function of multiple variables:
  - Take partial derivatives with respect to each variable. $\frac{\partial R_{sq}}{\partial w_0}$, $\frac{\partial R_{sq}}{\partial w_1}$
  - Set all partial derivatives to 0. $\frac{\partial R_{sq}}{\partial w_0}(\quad) = 0$, $\frac{\partial R_{sq}}{\partial w_1}(\quad) = 0$
  - Solve the resulting system of equations.
  - Ensure that you've found a minimum, rather than a maximum or saddle point (using the second derivative test for multivariate functions).

R_{sq} is parabolic, convex → single minimum

8

This usage of the word "parameter" is ok for ML, but is *incorrect* in a statistical context! Why?

Sidenote: notation

Statistical definition of parameter:

A population quantity that is typically unknown, but is what we are trying to estimate from any given model.

This is in contrast to a statistic, which is an estimate of a parameter that we can calculate from any given dataset.

Examples:

- $\overline{x}$ is a statistic that estimates the parameter μ
- s^2 is a statistic that estimates the parameter σ^2
- $\hat{\beta}_i$ is a statistic that estimates the parameter β_i

Incidentally, this is how the word "parameter" is also used in DSC 10, and in Justin's notes for DSC 140A.

Sidenote: notation

Why does this distinction matter??

One reason: Hypothesis testing

In DSC 40A and probably several other classes, you learned a lot about how to do modeling with regression, but likely little to no statistical inference for regression. That is, something like:

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

Note carefully that β_1 does not have a hat on it. Again, why does this matter?

And what are we even trying to do here and how is it different from modeling?

Statistical Inference for Regression

Model Building

When doing modeling (such as in a machine learning context), the goal is to find the "best" model according to some criteria (R^2, RMSE, etc). Then we might use that model to do prediction.

Example: if I know a March Madness participant's knowledge score, what is their predicted bracket score?

Statistical Inference

In contrast, when we are doing statistical inference, we have a specific question about the population, and we want to answer it with our sample.

Example: is there an association between a March Madness participant's basketball knowledge and their bracket score?

These are two different questions!

Statistical Inference for Regression

If there is an association between a participant's basketball knowledge and their bracket score, then from our simple linear model, this would mean that the slope of the regression line is not equal to 0.

(Note that this specifically investigates a linear association).

The slope of the regression line is β_1. Since the question is about the population, there is no hat:

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

The one with the hat, $\hat{\beta}_1$, refers to the actual value that is calculated from the data. We use that to then get our p-value! But how?

Statistical Inference for Regression

```r
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
```

Sidenote: notice that we get the same values from this output as we did from the previous manual calculation:

$\hat{\beta}_1$:

```
beta1 <- sum((knowledge-mean(knowledge))*(bracket_score-mean(bracket_score))) /
  sum((knowledge-mean(knowledge))^2)
beta1
```

```
## [1] 24.79583
```

```
summary(model1)$coefficients[2,1]
```

```
## [1] 24.79583
```

Sidenote: notice that we get the same values from this output as we did from the previous manual calculation:

$\hat{\beta}_0$:

```
beta0 <- mean(bracket_score) - beta1*mean(knowledge)
beta0
```

```
## [1] 623.1605
```

```
summary(model1)$coefficients[1,1]
```

```
## [1] 623.1605
```

Statistical Inference for Regression

From the R output, the p-value for $H_0 : \beta_1 = 0$ is:

```
summary(model1)$coefficients[2,4]
```

```
## [1] 0.03953505
```

and the heading of `Pr(>|t|)` in the output table suggests that this p-value comes from a t-Test. But why/how?

Statistical Inference for Regression

Recall: the one-sample t-Test statstic

$$t_s = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

Now, here is the test statistic for $H_0 : \beta_1 = 0$:

$$\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

- And recall that for t_s to have a t-Distribution, it relied on $\overline{x}$ having a normal distribution.
- So, it follows that here we need $\hat{\beta}_1$ to have a normal distribution!

Does it??

Yes it does, if we assume normally distributed residuals in our model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$. What exactly does this mean?

Normally distributed residuals

The assumption of normally distributed residuals means that each y_i value deviates from the model by an amount that follows a normal distribution, centered at 0.

It follows somewhat intuitively that if ϵ_i follows a normal distribution then $\hat{\beta}_1$ also will, but a rigorous proof of that is beyond the scope of this course.

Conditions for Validity

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 linear
  - e.g. not quadratic, exponential, etc.
- Independence of observations
- Normality of $\hat{\beta}_1$
  - 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
- Equal variance across all values of X
  - Also known as homoskedasticity

Conditions for Validity

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 $\hat{\beta}_1$
  - 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

Conditions for Validity

As we have said, "validity" of a statistical test means that its Type I Error rate will be equal to its nominal significance level of α (typically 0.05).

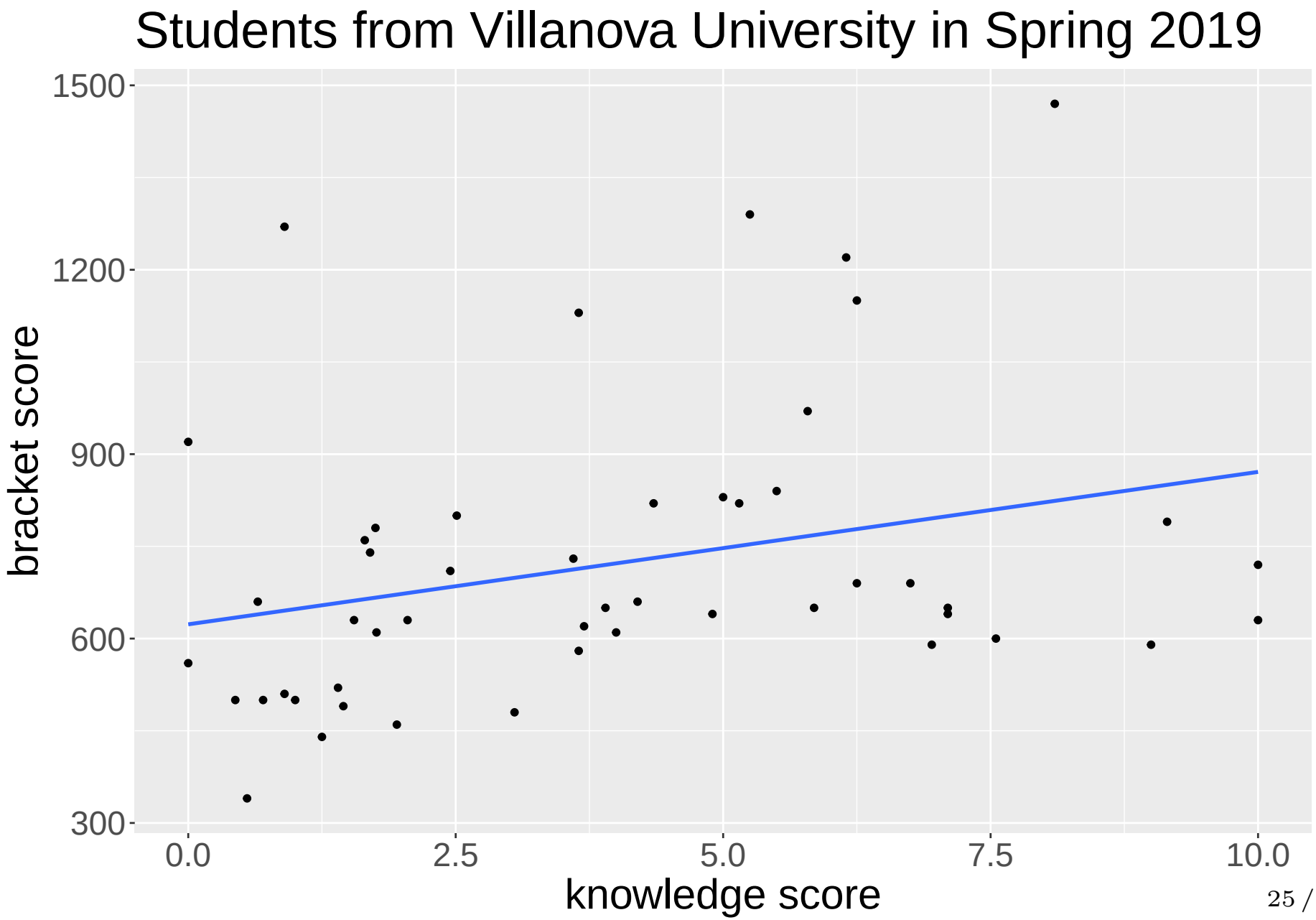
You will investigate Type I Error rates with in the presence of violations to these conditions on Lab 4

In class today, we will instead investigate statistical power

Recall:

- Effect Size
- Variance
- Sample Size
- Significance Level

First: what is an effect size in the context of a linear model?


Students from Villanova University in Spring 2019
bracket score
1500
1200
900
600
300
0.0
2.5
5.0
7.5
10.0
knowledge score

- $\hat{\beta}_0 \approx 623.1605$; what is its interpretation?

- $\hat{\beta}_1 \approx 24.7958$; what is its interpretation?

Statistical Power

We will take a simulation approach to power estimation in the linear model setting, as we are getting into a realm where ready-made routines are either:

- difficult to find
- difficult to use/understand what they do

So, what do we simulate?

Effect size and variance

As a starting point, consider:

- a linear increase by 20 points,
- per increase in knowledge score by 1 point

What about the variance?

Statistical Power

The X values (knowledge score) can be randomly or deterministically generated; it doesn't make much of a difference as long as there is decent coverage along the range of interest:

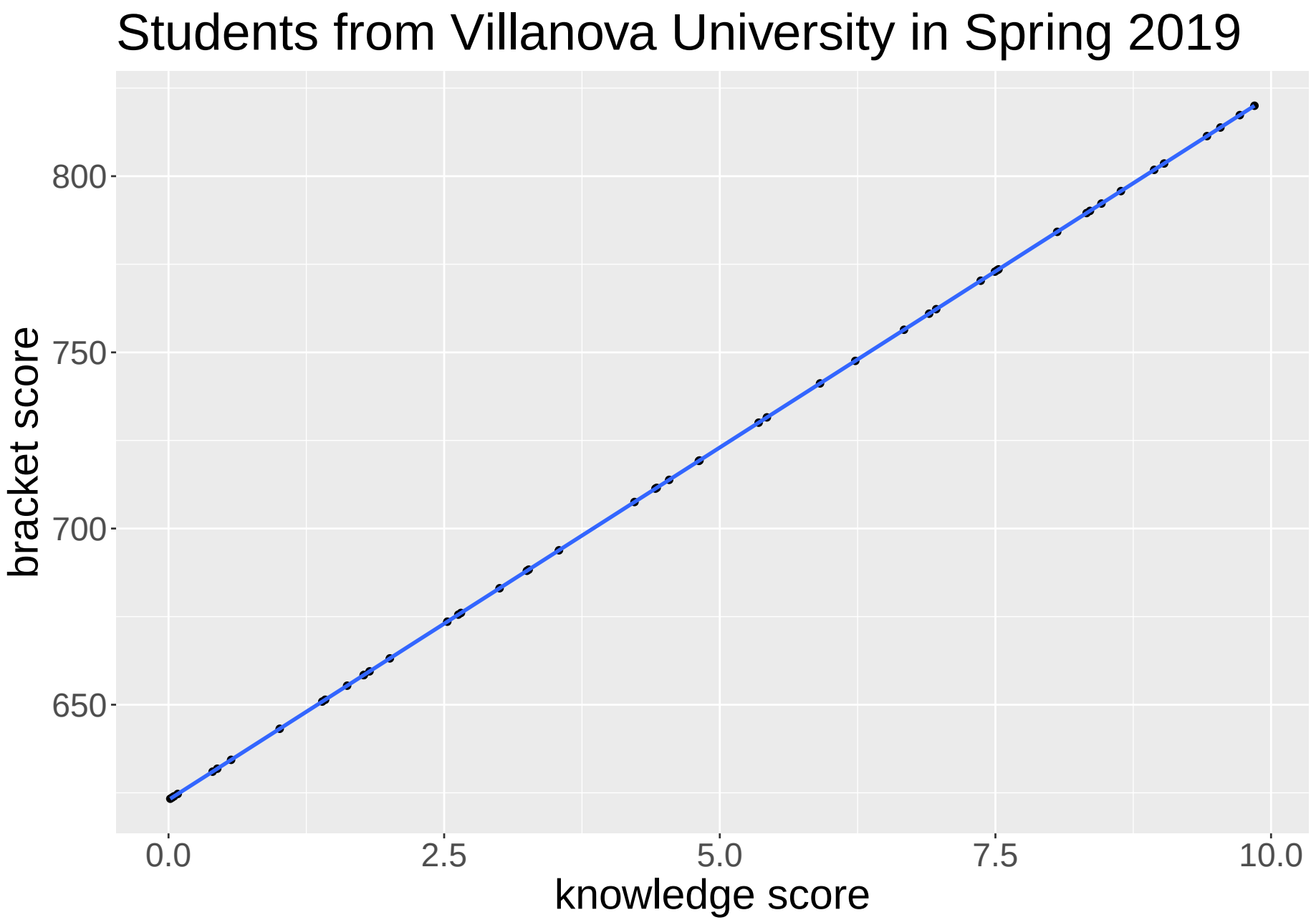
```
march_madness$knowledge <- runif(50, min=0, max=10)
```

The, the Y values are generated as a function of the X values:

```
march_madness$bracket_score <- 623 + march_madness$knowledge
```

- The intercept of 623 does not matter at all; we could completely omit it if we wanted to
- We're not done yet; we need to add some noise. But first, let's look at the simulated data we just generated:

Statistical Power



Statistical Power

Now, how much noise should we add?

We can get an idea of how much variability is present in the system by looking at the graph of our actual data (note that this is all getting very dangerously into the territory of "post-hoc" power calculations but alas...)

Statistical Power

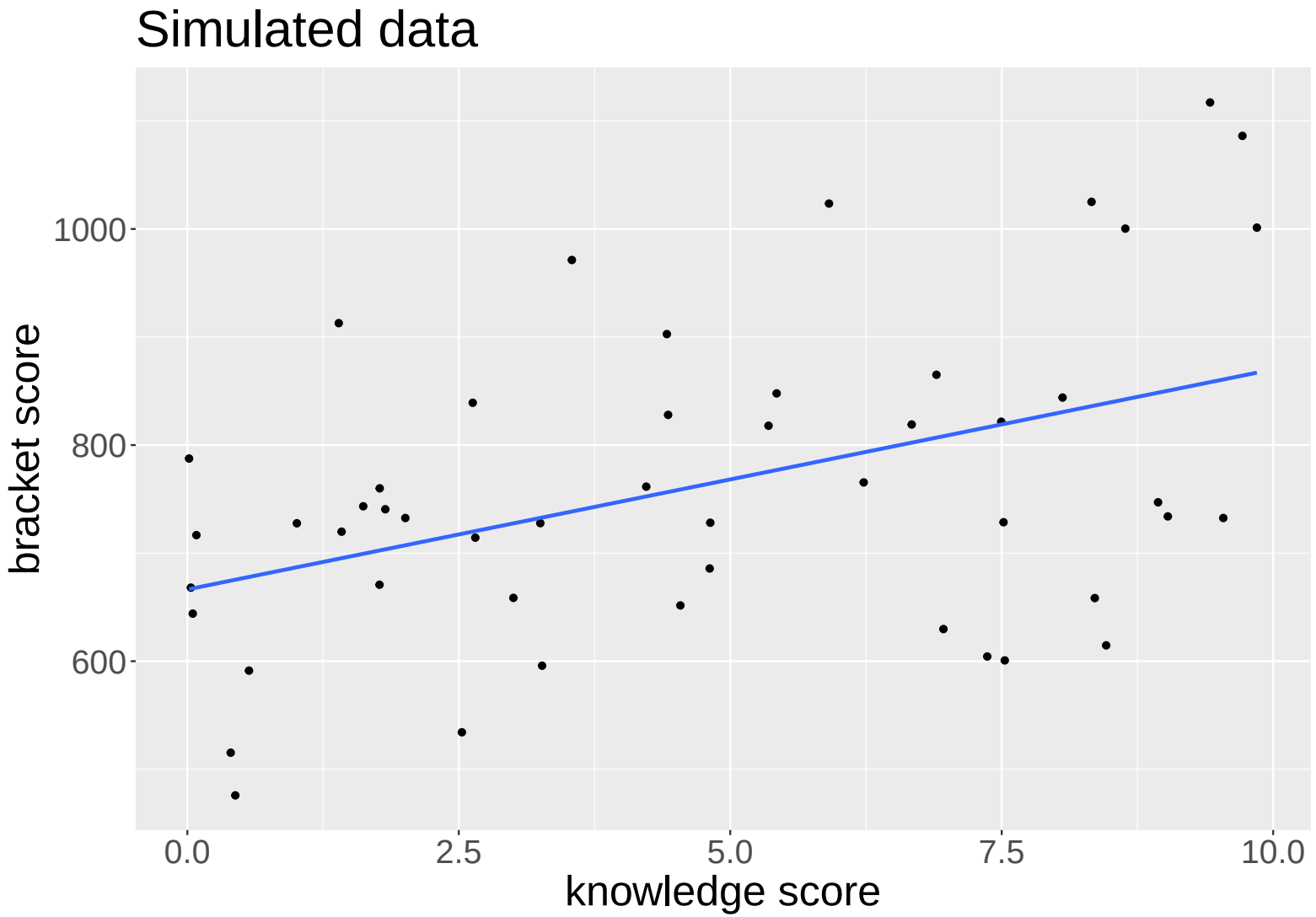
Now, how much noise should we add?

We can get an idea of how much variability is present in the system by looking at the graph of our actual data (note that this is all getting very dangerously into the territory of "post-hoc" power calculations but alas...)

From inspection of the graph, it looks like $\sigma = 150$ might be reasonable (based on 68% of observations falling within 1 standard deviation)

Statistical Power

```
march_madness$bracket_score <- 623 + march_madness$knowledge * 20 +
  rnorm(50, 0, 150)
```



Statistical Power

Now, how do we estimate power?

Reminder: what is statistical power?

Statistical power is the probability of correctly rejecting H_0.

So, here if $\beta_1 = 20$, then we should reject H_0.

Statistical Power

Your Turn

Now, how do we estimate power?

- Simulate samples of size $n = 50$ under $\beta_1 = 20, \epsilon \sim N(0, \sigma = 150)$.
- Run the statistical test for $H_0 : \beta_1 = 0$ on these data
- Determine whether $p < 0.05$
- Repeat many times, count the proportion of the time that H_0 is rejected

Recap and Looking Ahead

Today's Daily Check

Just the Your Turn on the previous slide

Recap

- Simple Linear Regression refers to the scenario in which we have one quantitative outcome variable and one quantitative predictor variable
- Statistical inference in this setting asks questions about β_1, the slope
- Conditions for validity are summarized by LINE

Looking Ahead

Multiple Linear Regression!

Quiz Review