

Learning Objectives

- **Contrast** Pearson r vs. Spearman r_s
- **Review** Models (central tendency, variability, correlation) and describe them with reference to psychological processes
- **Describe** goal of statistical inference

This week in *Statistics Drama*!



The Average and the Outliers



380



What happens if the explanatory and response variables are sorted independently before regression?

regression

correlation

Suppose we have data set (X_i, Y_i) with n points. We want to perform a linear regression, but first we sort the X_i values and the Y_i values independently of each other, forming data set (X_i, Y_j) . Is there any meaningful interpretation of the regression on the new data set? Does this have a name?

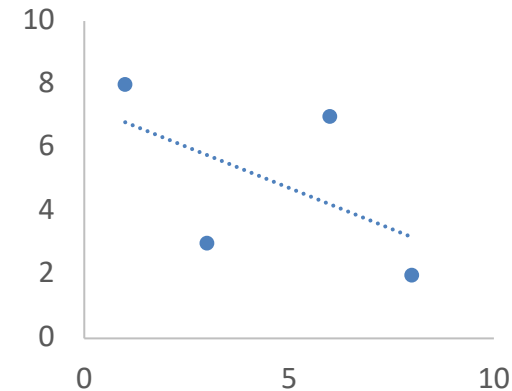
I imagine this is a silly question so I apologize, I'm not formally trained in statistics. In my mind this completely destroys our data and the regression is meaningless. But my manager says he gets "better regressions most of the time" when he does this (here "better" means more predictive). I have a feeling he is deceiving himself.

X_i	Y_i
1	8
3	3
7	7
8	2

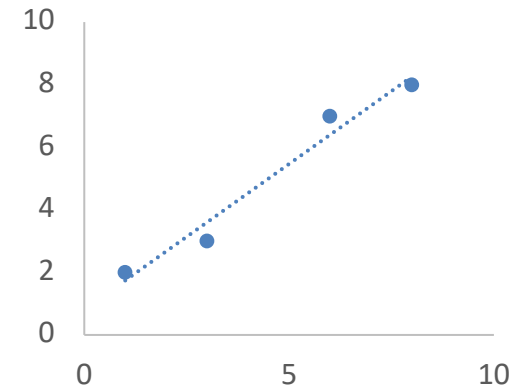


X_i	Y_j
1	2
3	3
6	7
8	8

Scatterplot $X_i \sim Y_i$



Scatterplot $X_i \sim Y_j$



The Average and the Outliers



380



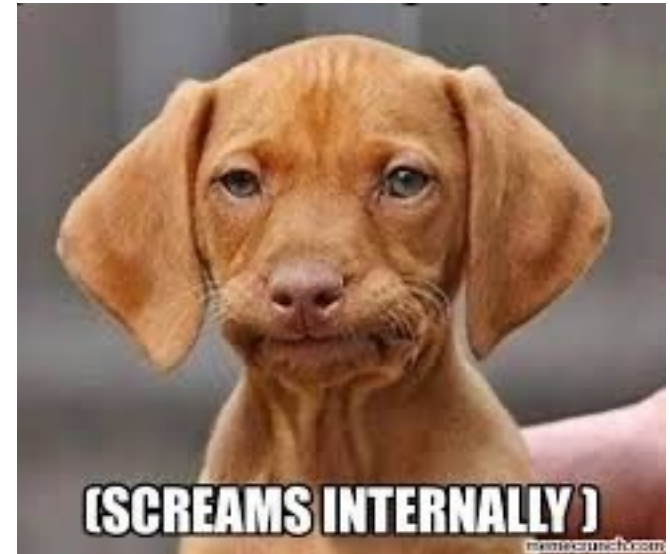
What happens if the explanatory and response variables are sorted independently before regression?

regression

correlation

Suppose we have data set (X_i, Y_i) with n points. We want to perform a linear regression, but first we sort the X_i values and the Y_i values independently of each other, forming data set (X_i, Y_j) . Is there any meaningful interpretation of the regression on the new data set? Does this have a name?

I imagine this is a silly question so I apologize, I'm not formally trained in statistics. In my mind this completely destroys our data and the regression is meaningless. But my manager says he gets "better regressions most of the time" when he does this (here "better" means more predictive). I have a feeling he is deceiving himself.



EDIT: Thank you for all of your nice and patient examples. I showed him the examples by @RUser4512 and @gung and he remains staunch. He's becoming irritated and I'm becoming exhausted. I feel crestfallen. I will probably begin looking for other jobs soon.

Pearson r vs. Spearman r_s (pp 141-144)

- Pearson r assumes:
 - Equal intervals (interval/ratio)
 - Normal distribution of X & Y
 - Absence of multivariate outliers
 - Linear relationship $X \sim Y$
 - If assumptions violated, model may be biased
- Spearman r_s (aka, ρ) assumes:
 - Ranked data (ordinal/interval/ratio)
 - Normal distribution unnecessary
 - Multivariate outliers are OK(ish)
 - Monotonic relationship $X \sim Y$ (*need not be linear!*)

Pearson r vs. Spearman r_s (pp 141-144)

- Spearman ρ advantages:
 - Robust to non-normal data (e.g., skewed distributions) & outliers
 - Can model not-quite-linear relationships
- Spearman ρ disadvantages:
 - If X & Y have equal intervals, ranking loses information
 - If X & Y are normally distributed, Spearman r_s is less powerful (may fail to find true $X \sim Y$ relationship)

Jamovi Demonstration

(spearman_demo.omv)

“What is the relationship between self-reported hours of studying (X) & self-reported grade (Y)?”

First

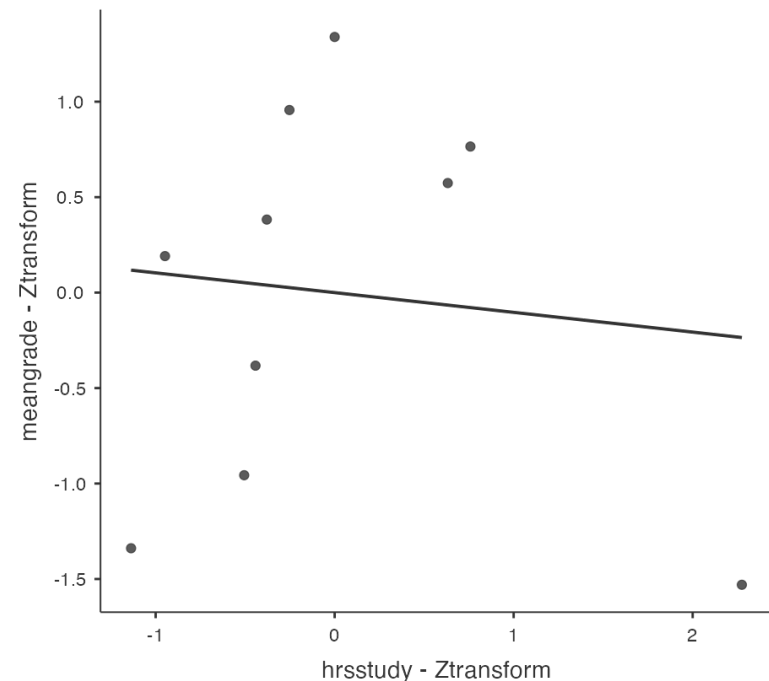
- Visualize distribution of X & Y
- Visualize scatterplot of $X \sim Y$

Pearson r

1. Standardize (z-score) X & Y
2. Plot $Z_X \sim Z_Y$
3. Calculate r

Spearman r_s

1. Rank X & Y
2. Plot $X_{\text{rank}} \sim Y_{\text{rank}}$
3. Calculate r_s



How do they 'see' data?

Unstandardized regression uses raw data
(w/original units)

Pearson r uses standardized data
(z-scores)

Spearman r_s uses ranked data

 hrsstudy ...	 hrsstudy ...	 hrsstudy ...
2	-1.137	1
5	-0.948	2
12	-0.506	3
13	-0.442	4
14	-0.379	5
16	-0.253	6
20	0.000	7
30	0.632	8
32	0.758	9
56	2.275	10



Prediction using Models

(harry_learns.omv)

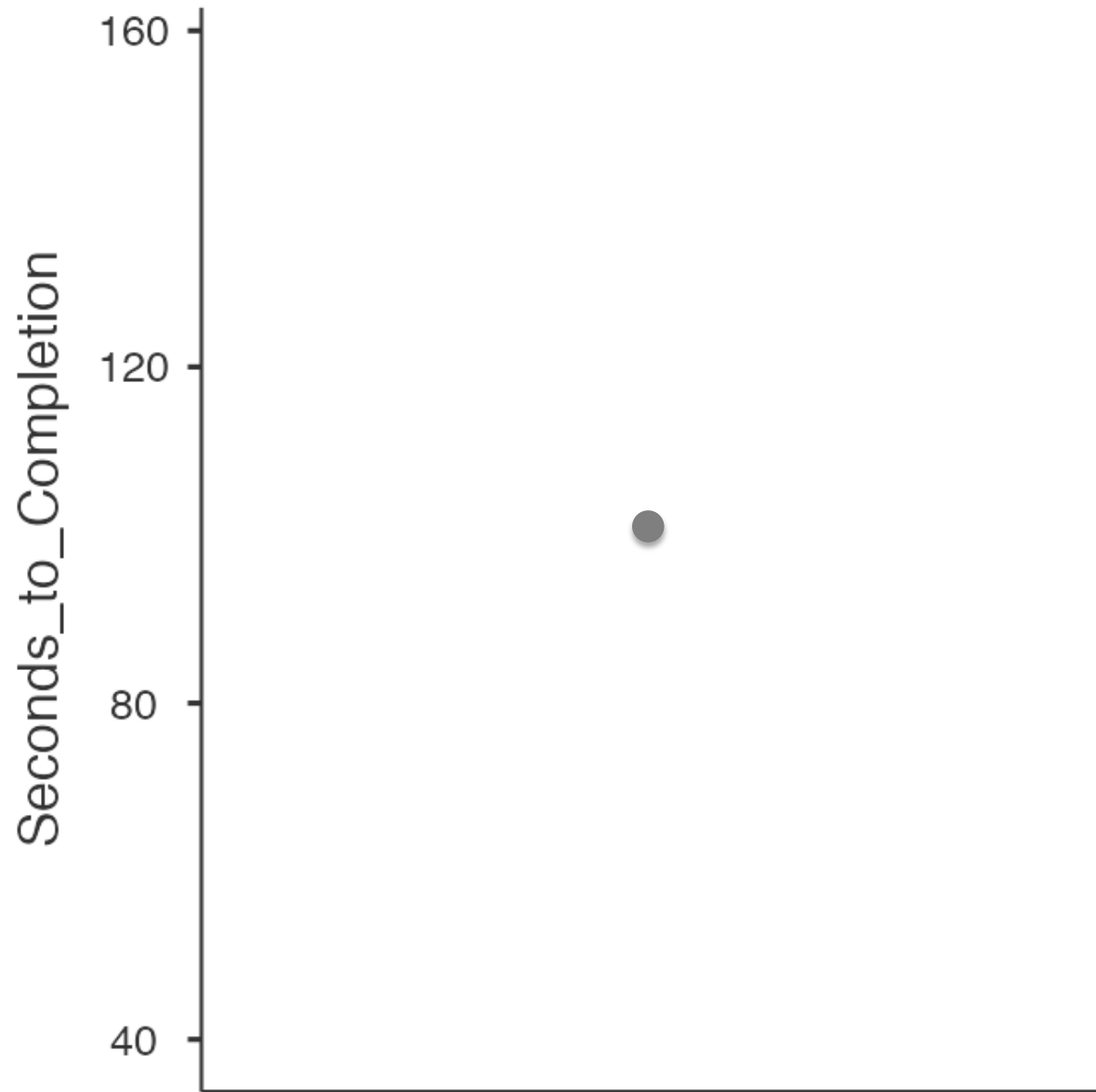
- We use *models* to simplify our data, and to make predictions
 - We've already covered many prediction models
- Let's predict using statistical models!!



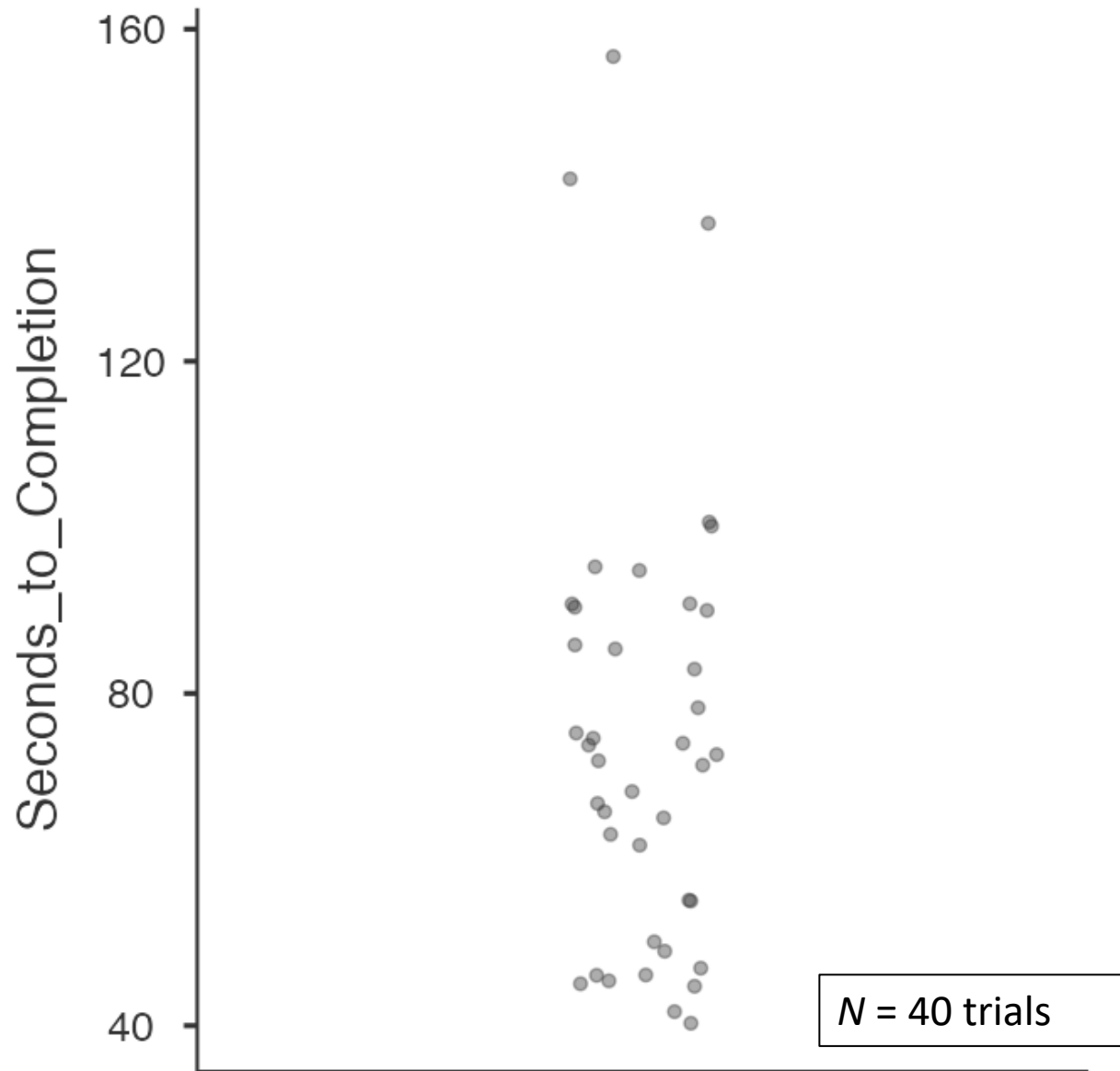
Data Characteristics

	 Encounter...	 Seconds_to_Completion
1	1	157
2	2	86
3	3	95
4	4	100
5	5	142
6	6	95
7	7	137
8	8	101
9	9	74
10	10	72

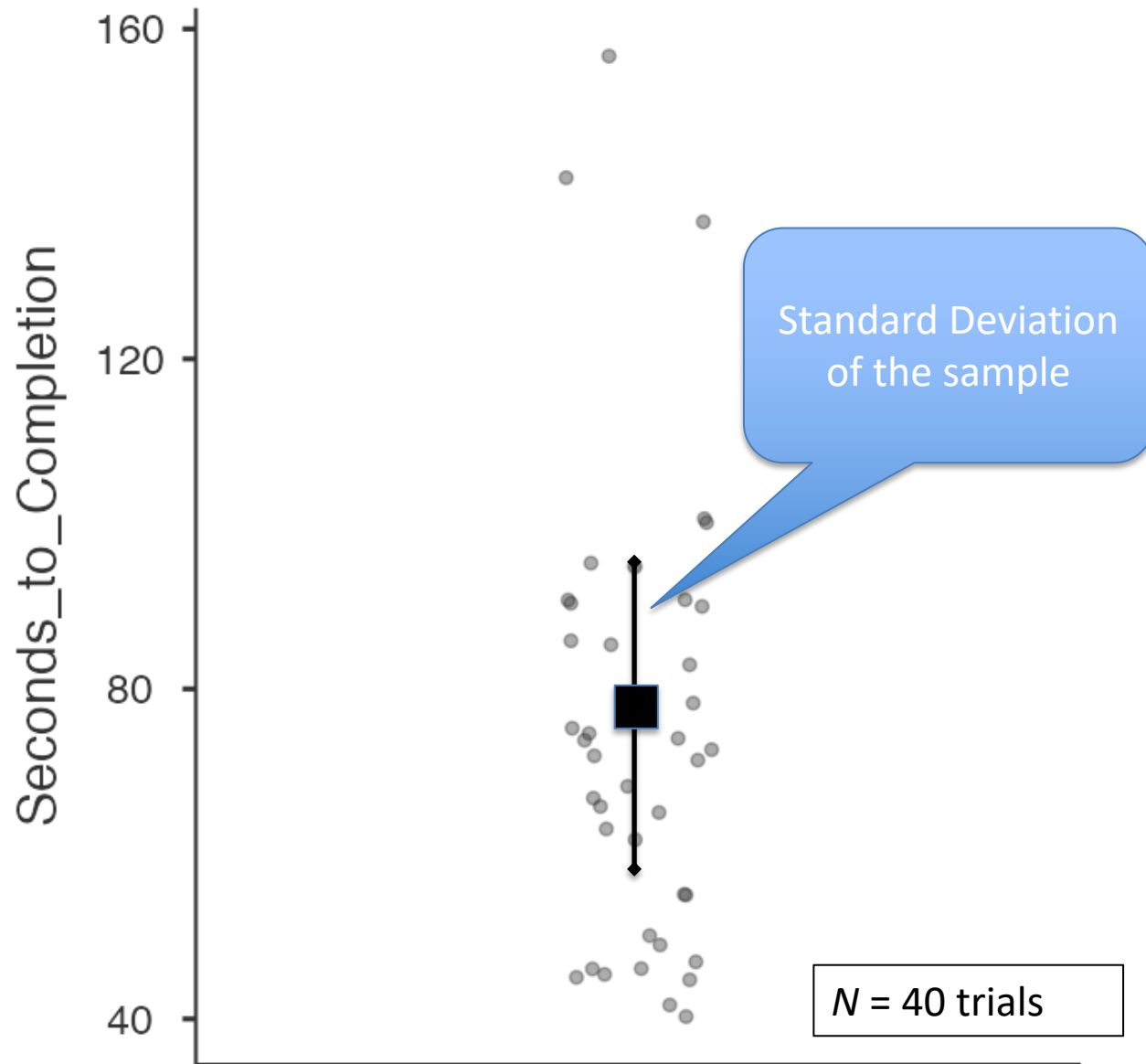
Single variable plot: Seconds to complete puzzle



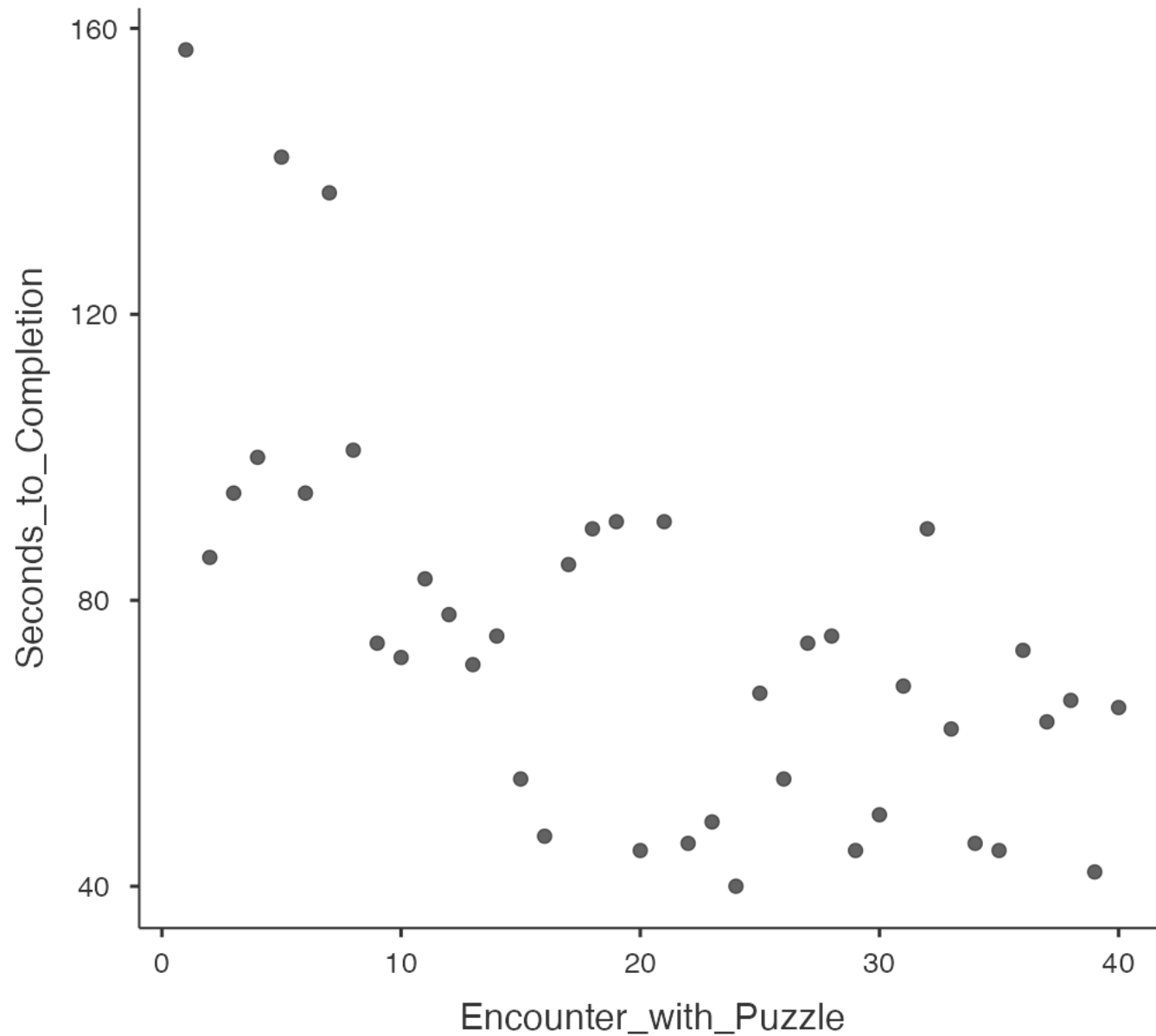
Single variable plot: Seconds to complete puzzle



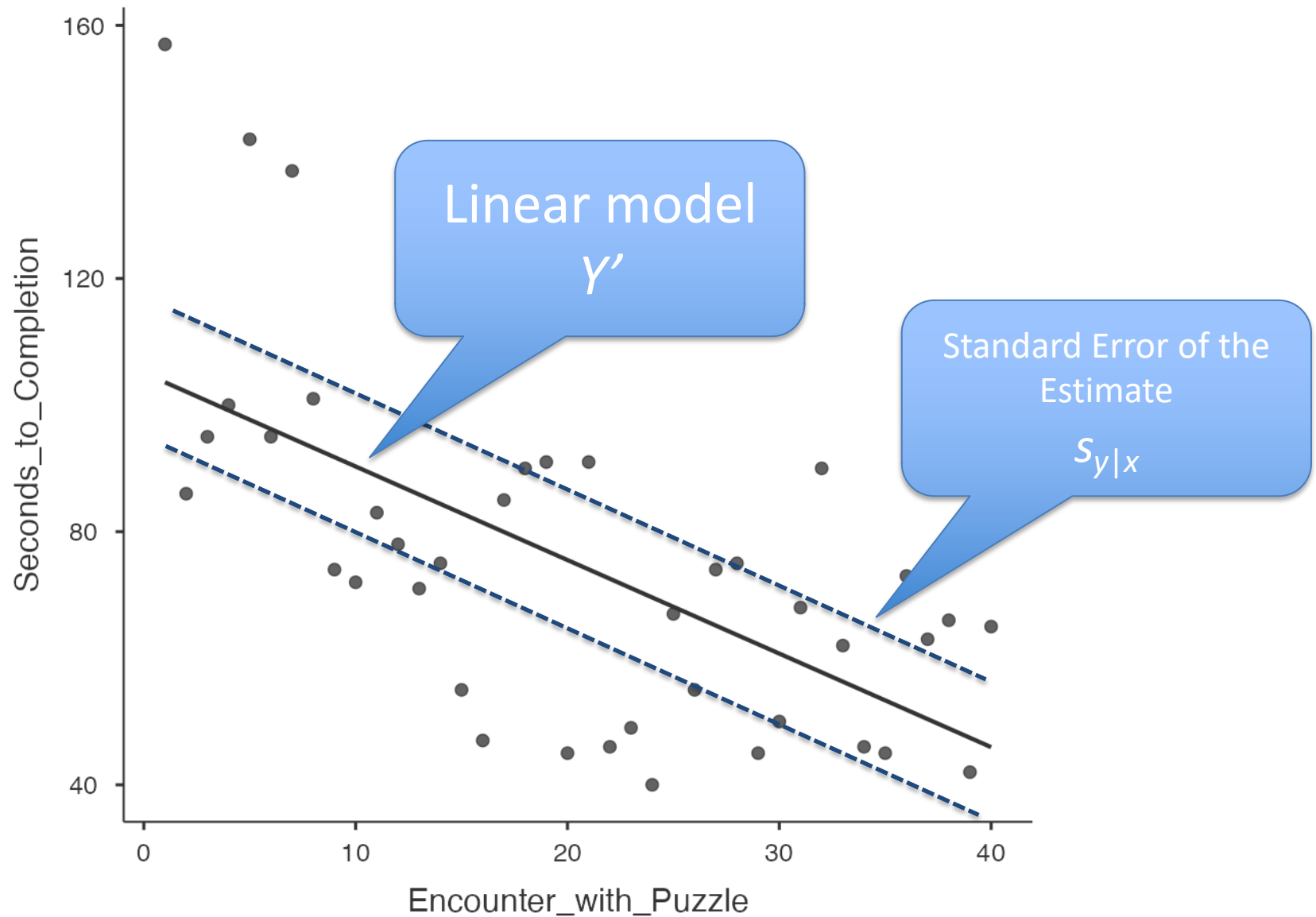
Single variable plot: Seconds to complete puzzle



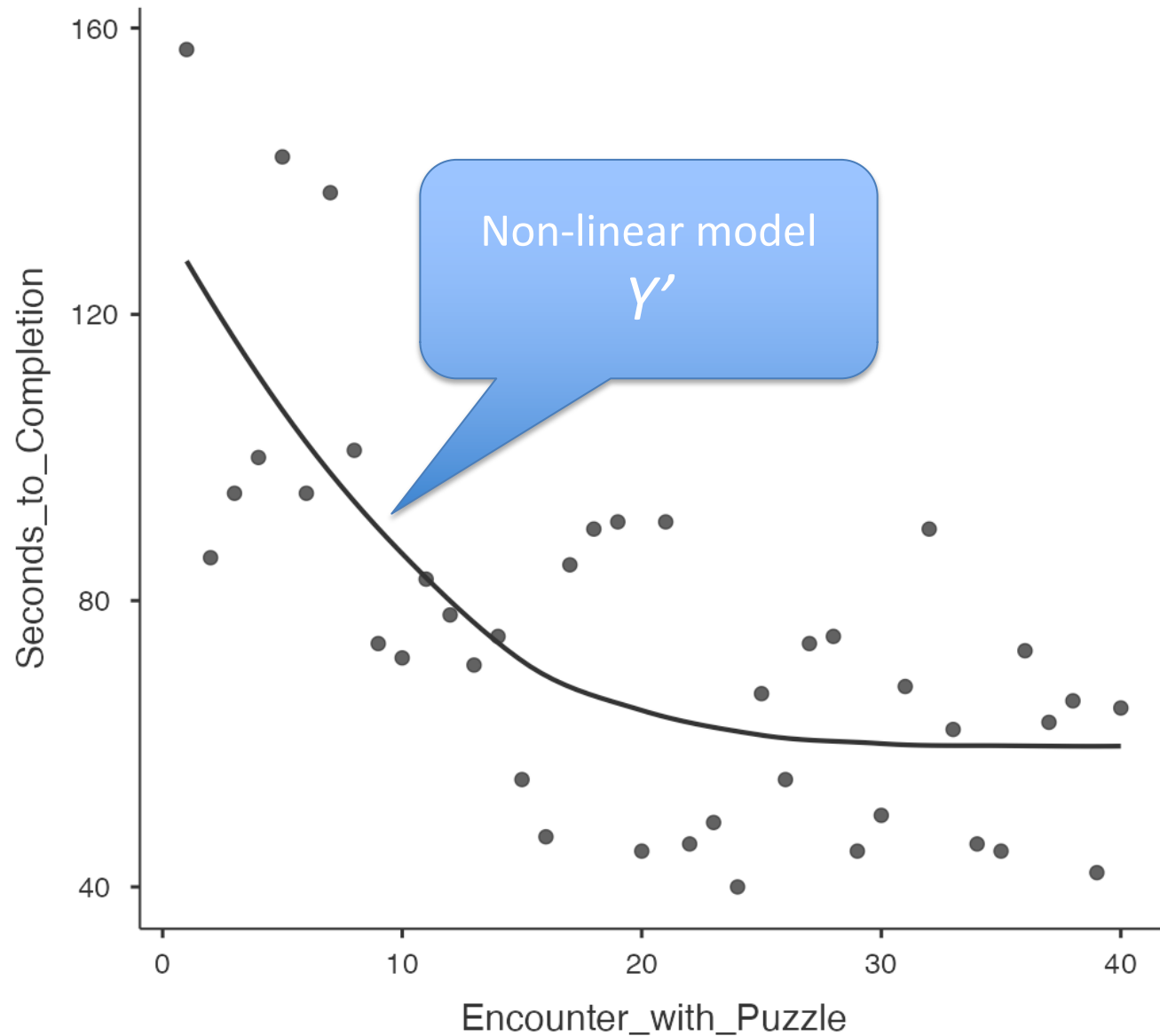
Scatterplot: Seconds to complete puzzle by number of puzzle encounters



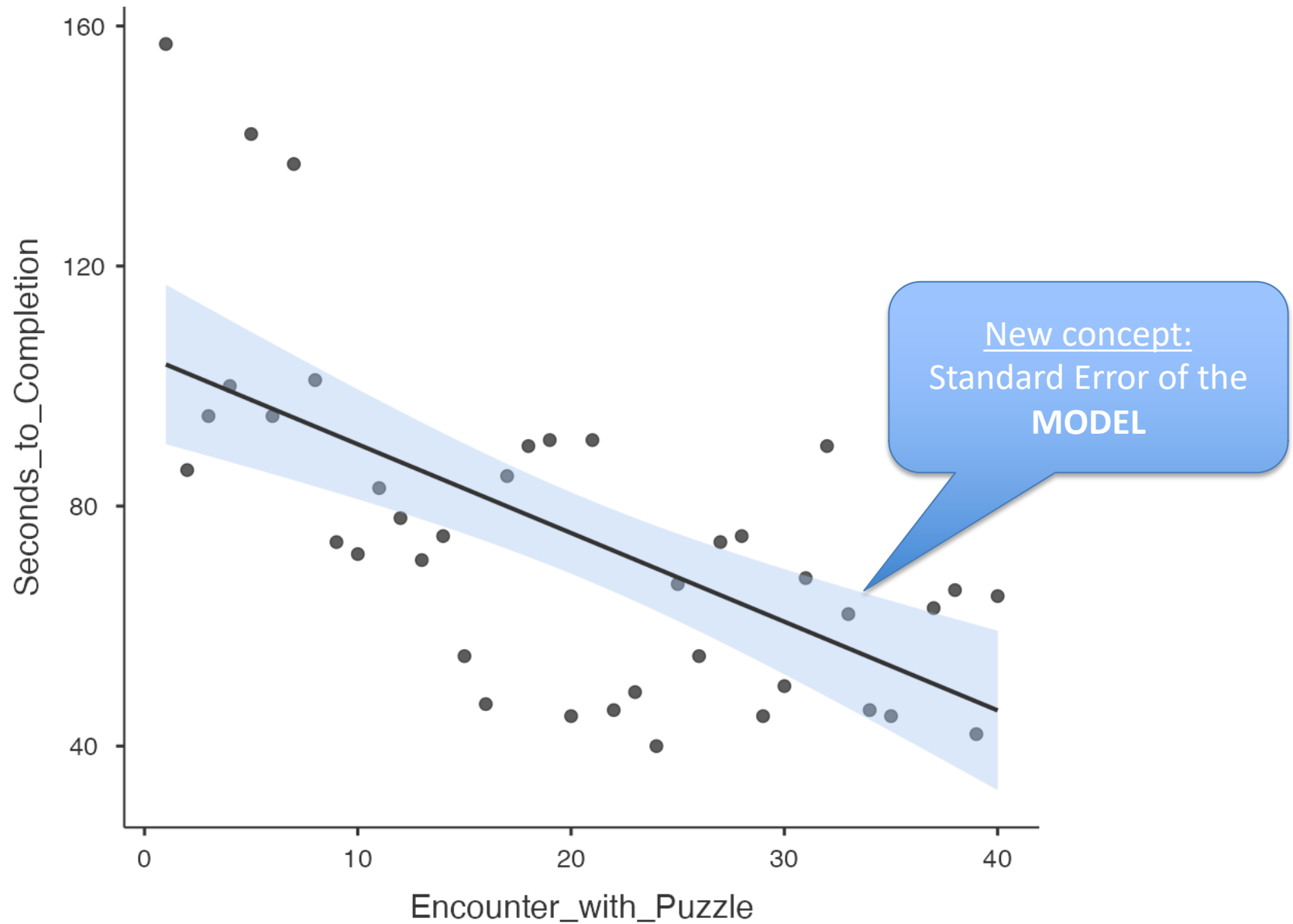
Scatterplot: Seconds to complete puzzle by number of puzzle encounters



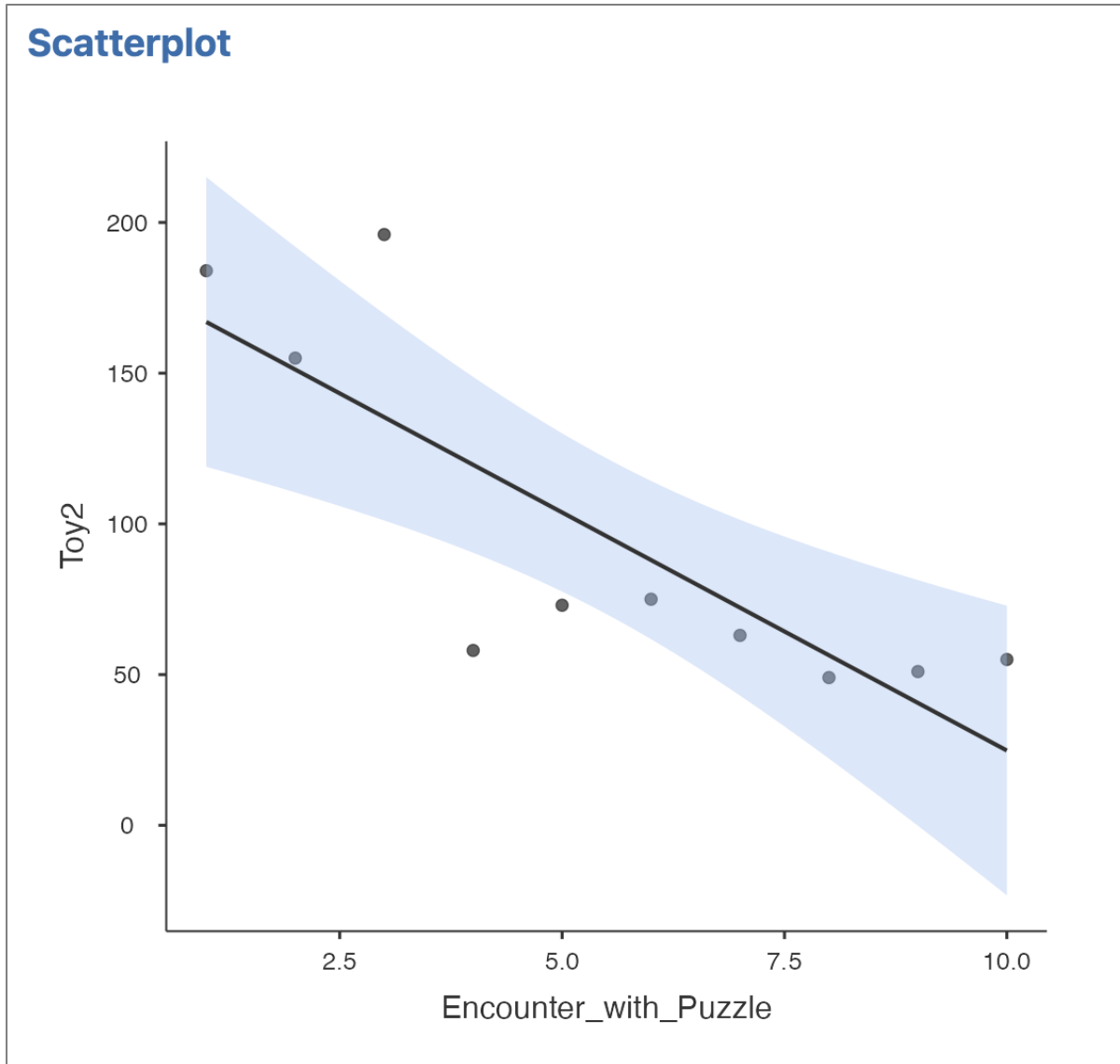
Scatterplot: Seconds to complete puzzle by number of puzzle encounters

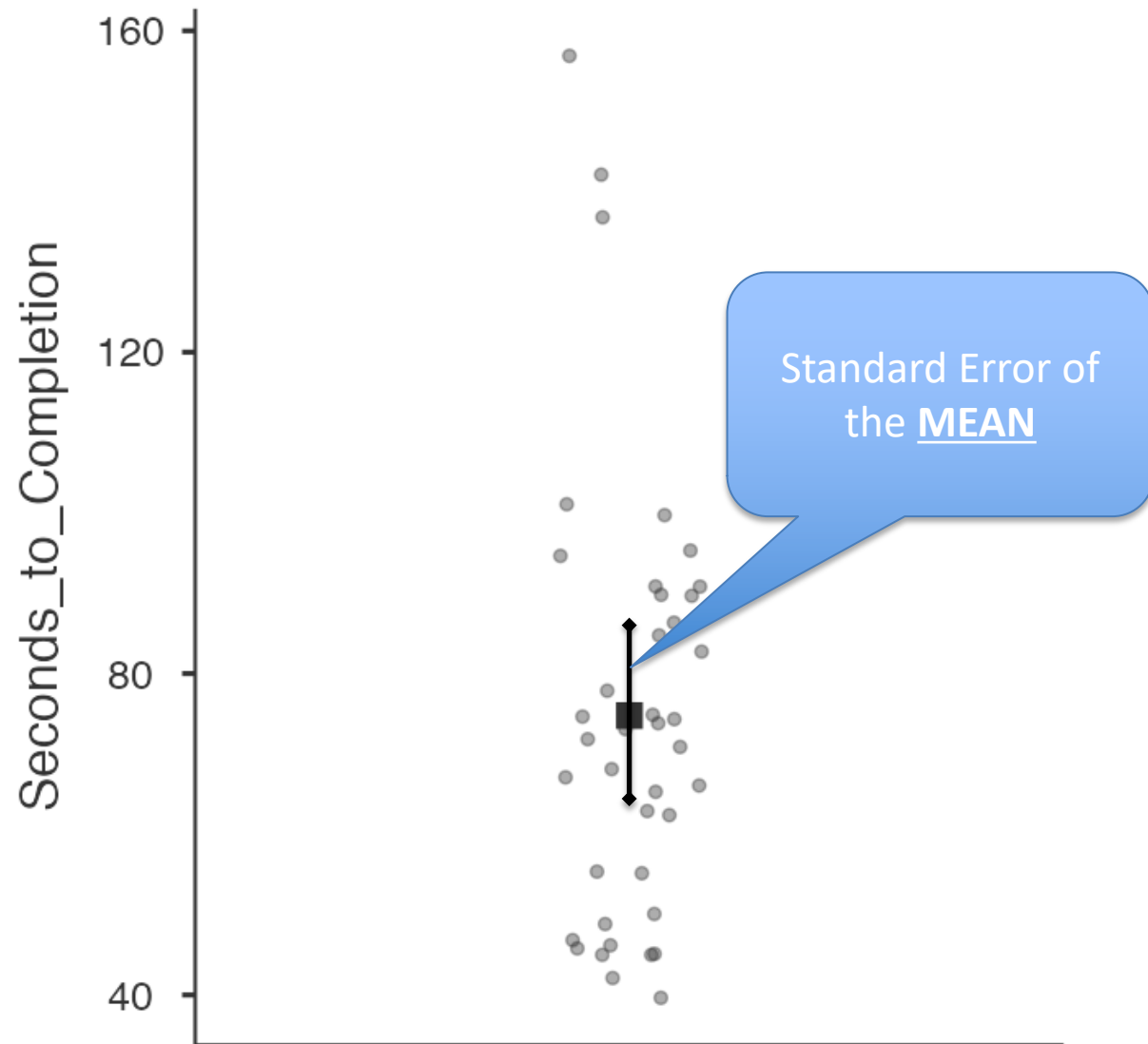


Scatterplot: Seconds to complete puzzle by number of puzzle encounters



Inference





What do you think?



Inferential Statistics Starts This Week!