Regression Testing

Grinnell College

April 2026

Quick Review

If a sample is collected randomly....

then the calculated statistics are random...

and will have a sampling distribution which we can use for testing

Quick Review

In a t-test for one mean, we generally check to see if some mean μ is is some number. Eg...

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

And we use this to say something about where we think the mean of the population is(n't).

Quick Review

In a t-test for two means, we generally check to see if some mean μ is is some number. Eg...

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

And we use this to say something about where we think the mean of the population is(n't).

Expanding Our Horizons

A sample mean/proportion isn't the only statistic we can test.....

Thoughts on other things worth testing?

Expanding Our Horizons

A sample mean/proportion isn't the only statistic we can test.....
Thoughts on other things worth testing?

Many choices.....

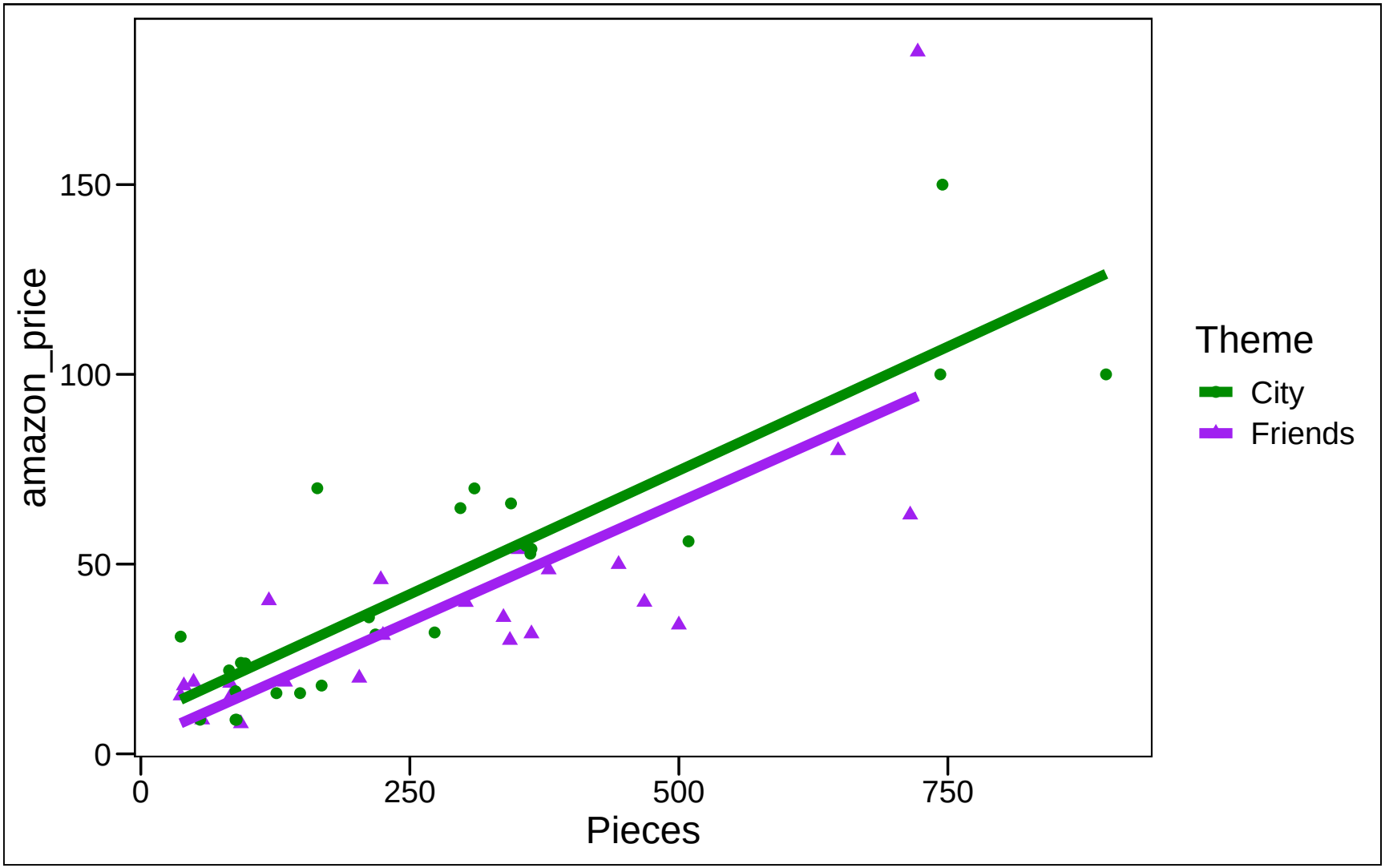
1. Sample variance (follows a χ^2 distribution)
2. Order statistics (eg estimating the largest value in the population)
3. Distances between two distributions

Most importantly for us, and more common, is testing regression coefficients (the β's in linear regression)

When would we ever use this?

Back to the lego data!

Are these lines different from each other?



Multiple Regression Framework

Skipping ahead to next week's notes multiple regression equations are written in the form...

$$\hat{y} = \beta_0 + \beta_1(\text{Variable 1}) + \beta_2(\text{Variable 2}) + \beta_3(\text{Variable 3}) + \ldots$$

- Subscript represents where it is in the equation
- β_i is used to denote a generic β not tied to a specific example
- The left hand side is our predicted mean price given variable 1, 2, 3, ...

  2 variables with the same scale and spread, the β that is closer to 0 indicates that variable is less useful in explaining the relationship

Back to the lego data!

Linear regression equation:

$$\widehat{\text{Amazon Price}} = \beta_0 + \beta_1 \text{Pieces} + \beta_2 \mathbb{I}_{\text{Friends}}$$

Estimated linear regression equation:

$$\widehat{\text{Amazon Price}} = \hat{\beta}_0 + \hat{\beta}_1 \text{Pieces} + \hat{\beta}_2 \mathbb{I}_{\text{Friends}}$$

Estimated linear regression equation:

$$\widehat{\text{Amazon Price}} = 10.02 + 0.13 \text{Pieces} - 7.32 \mathbb{I}_{\text{Friends}}$$

where $\mathbb{I}$ is 1 if the set if "Friends" themed and 0 otherwise.

Sampling Distributions for Coefficents

To keep it very brief the same concept applies as we have been talking about.

1. We take a sample
2. We fit a linear regression model
3. We record and plot on a histogram the value of a coefficient we are interested in
4. Do this again an infinite number of times
5. This is our sampline distribution for the coefficient

Test for Coefficients

We have seen this before!

1. State your hypothesis
2. Visualize your data
3. Check your assumptions
4. Calculate your test statistic and p-value
5. Decision

Test for Coefficients

We have seen this before!

1. State your hypothesis
   - Almost always it's H_0: $\beta_i = 0$
2. Visualize your data
   - Should have a scatterplot well before testing coefficients
3. Check your assumptions
   - Random, IID, Normal or large n
   - AND the assumptions from linear regression (revisisted)
4. Calculate your test statistic and p-value
   - Again, computer's will help
5. Decision

Test for Coefficients

We have seen this before!

1. State your hypothesis
   - Almost always it's H_0: $\beta_i = 0$
2. Visualize your data
   - Should have a scatterplot well before testing coefficients
3. Check your assumptions
   - Random, IID, Normal or large n
   - AND the assumptions from linear regression (revisited)
4. Calculate your test statistic and p-value
   - Again, computer's will help
5. Decision

Or we could do a confidence interval....

Hypothesis Statements

Generally, but not required, we want to see if a coefficient is different than 0.

Why?

Hypothesis Statements

Generally, but not required, we ant to see if a coefficient is different than 0.

Why?

Because if $\beta_i = 0$, then we don't think there is linear relationship between the response and the explanatory variable

Hypothesis Statements

Generally, but not required, we ant to see if a coefficient is different than 0.

Why?

Because if $\beta_i = 0$, then we don't think there is linear relationship between the response and the explanatory variable

---

$$\widehat{\text{Number of Birds}} = \beta_0 + \beta_1(\text{Number of times I tie my boots per day})$$

If we estimate β_1 to be close to 0 then it doesn't look like my boot tying affects avian populations averages in a linear way

Hypothesis Statements

Our starting assumption is that there is no linear relationship between the mean response and the explanatory variable being tested after accounting for the other variables...

$$\widehat{\text{weight}} = \beta_0 + \beta_1 \text{age}$$

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

$H_0 : \beta_1 = 0$ implies age has no linear effect on the average weight

WARNING: Hypothesis Statement Interpretations

If a quantitative explanatory variable is a monomial with order greater than 1 then we refer to the effect as being that order....

$$\widehat{\text{weight}} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2$$

Here H_0: $\beta_2 = 0$ implies there is no *quadratic* effect of age on average weight after accounting for the *linear* effect of age

---

If it's two indicator variables multiplied together, we refer to it as an *interaction* of explanatory variables A and B....

$$\widehat{\text{weight}} = \beta_0 + \beta_1 \mathbb{I}_{\text{Girl}} + \beta_2 \mathbb{I}_{\text{American}} + \beta_3 \mathbb{I}_{\text{American}} \mathbb{I}_{\text{Girl}}$$

Here H_0: $\beta_3 = 0$ implies there is no *interaction* effect of American and Girl on average weight after accounting for the *main effects* of Girl and American

Lego Example

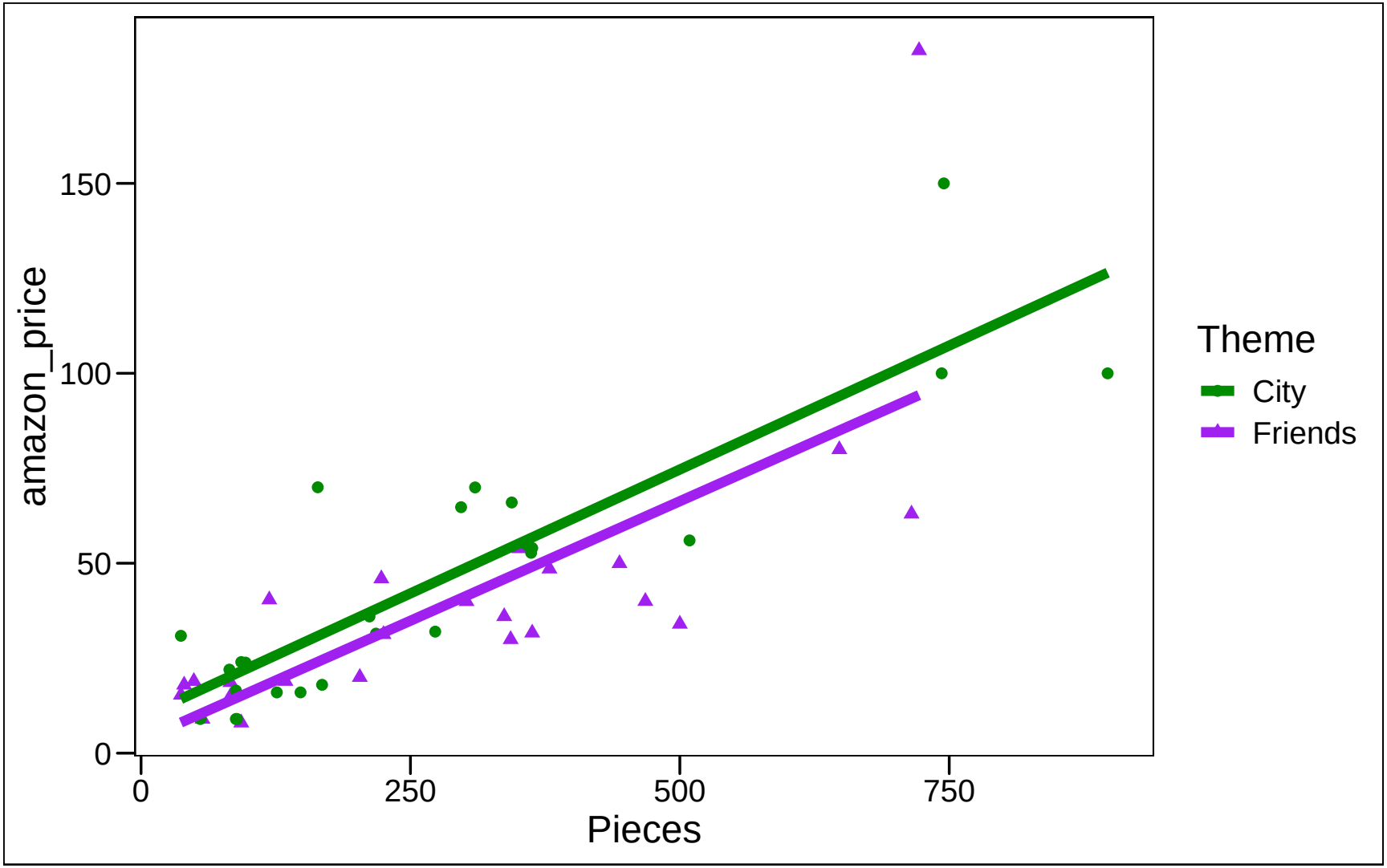
I'm interested to see if the effect of theme, after accounting for the number of pieces, is not 0 so....

$$\widehat{\text{Amazon Price}} = \hat{\beta}_0 + \hat{\beta}_1 \text{Pieces} + \hat{\beta}_2 \mathbb{I}_{\text{Friends}}$$

$$H_0 : \beta_2 = 0$$
$$H_A : \beta_2 \neq 0$$

NOTE: We don't use ^s (which are our estimates that we know)

Visualize


amazon_price
150
100
50
0
0
250
500
750
Pieces
Theme
City
Friends

Assumptions: The Big 3

1. Random: The data was randomly collected
   - Randomly collected surveys for samples
   - Randomly assigned treatments for experiments

2. Independent and Identically Distributed: referencing the residuals
   - Eg homoskedasticity (equal variance in your residual plot) is a requirement for identically distributed
   - "Linear" assumption is rolled into this one...if your line doesn't fit your residuals won't be IID

3. Population (of residuals) is Normal or n is large
   - Want a bell curve of our residuals around 0
     - ★ Seen and done this in unit 1
   - If that fails Central Limit Theorem still applies

WARNING

I love the CLT for testing means via t-tests.....

BUT

....often times I feel it's invoked in linear regression when the model is wrong and is used to mask modeling inadequacies.

WARNING

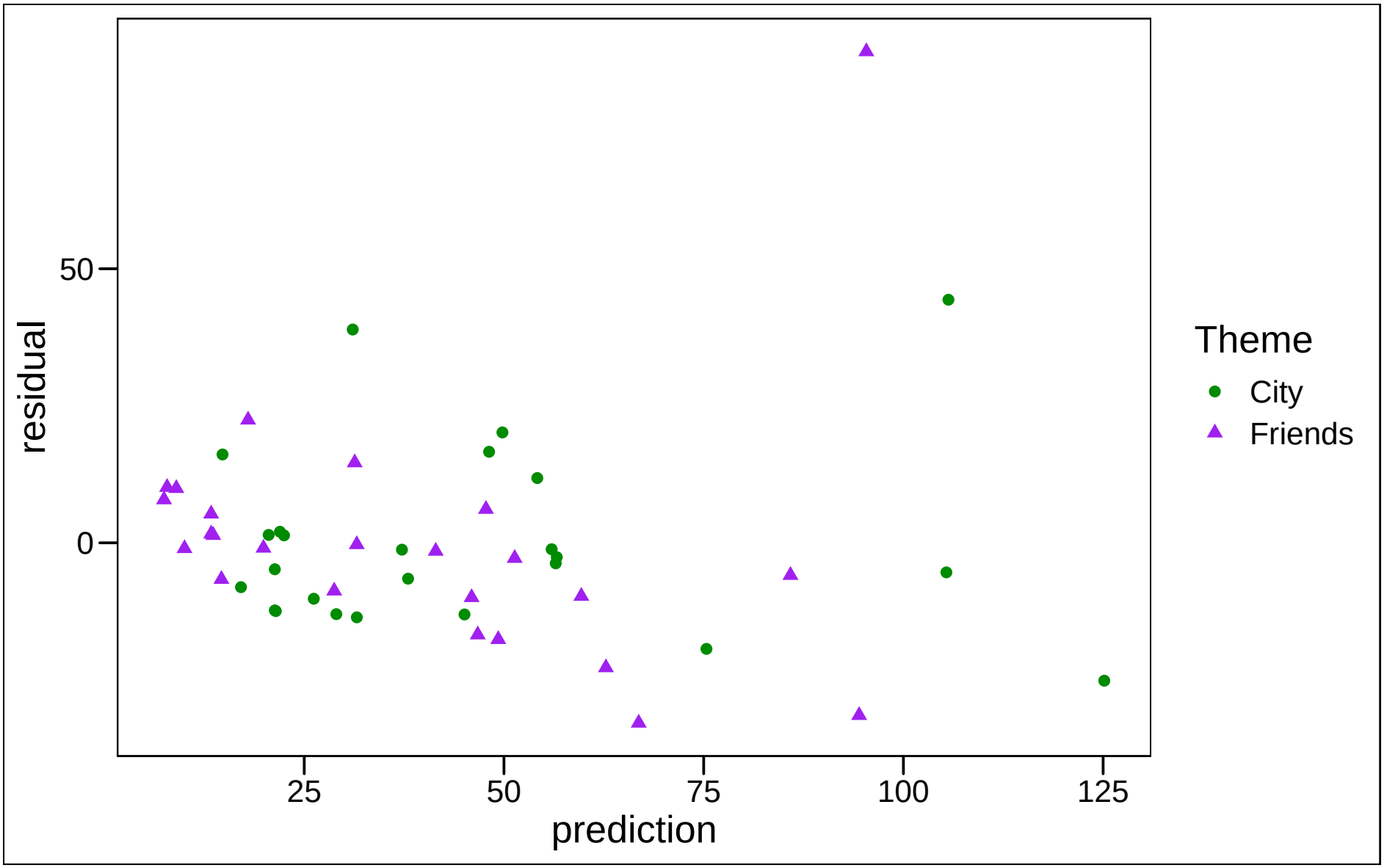
I love the CLT for testing means via t-tests.....

BUT

....often times I feel it's invoked in linear regression when the model is wrong and is used to mask modeling inadequacies.

Are you missing a variable you should have? A higher order term like x^2? Should you use a more advanced technique like a glm (generalized linear models)? Transformations?

Legos: Assumptions



Legos: Assumptions

1. Random: Yes, the legos were randomly selected

2. Independent and identically distributed
   - Independent is probably fine
   - identically distributed still fails...why?

3. Normal or large n

Legos: Assumptions

1. Random: Yes, the legos were randomly selected

2. Independent and identically distributed
   - Independent is probably fine
   - identically distributed still fails...why?
   - The data fans out going to the right
   - The difference between purple triangles and green circles is less pronounced than in your lab

3. Normal or large n
   - Normal but the spread is changing
   - n is 50 which is a healthy size but....

Legos: Assumptions

1. Random: Yes, the legos were randomly selected

2. Independent and identically distributed
   - Independent is probably fine
   - identically distributed still fails...why?
   - The data fans out going to the right
   - The difference between purple triangles and green circles is less pronounced than in your lab

3. Normal or large n
   - Normal but the spread is changing
   - n is 50 which is a healthy size but....
   - We have already explored that both amazon price and the number of pieces should be on the log scale (more later)

Test Statistic and P-value

It's still

$$\frac{(Observed) \; - \; (Hypothesis)}{Standard \;\; Error}$$

but the standard error becomes so annoying to calculate even the online textbook taps out:

> We will rely on statistical software to compute the standard error and leave the explanation of how this standard error is determined to a second or third statistics course. Table 24.3 shows software

Test Statistic and P-value

For R it's in the summary() function again when you input a model

```
> mod <- lm(amazon_price ~ Pieces + Theme, data = small_bricks)
> summary(mod)

Call:
lm(formula = amazon_price ~ Pieces + Theme, data = small_bricks)

Residuals:
    Min      1Q  Median      3Q     Max
-32.869 -10.142  -2.064   5.923  89.628

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.02122    5.42889   1.846   0.0712 .
Pieces        0.12834    0.01314   9.767 6.83e-13 ***
ThemeFriends -7.32372    5.73584  -1.277   0.2079
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.28 on 47 degrees of freedom
Multiple R-squared:  0.673,     Adjusted R-squared:  0.6591
F-statistic: 48.36 on 2 and 47 DF,  p-value: 3.915e-12
```

Test Statistic and P-value

Focusing on the "Coefficients" table...

- Estimate: is the estimated value of your coefficient
- Std. Error: estimated standard error for the coefficient
- t value: the test statistic
  - Yes, all of these are t-tests
- Pr(> | t |): P-value
  - ONLY for two sided tests (ie "not equal to")
  - Most common?
  - If you want one-sided you'll need to play around with the shapes or use pt()

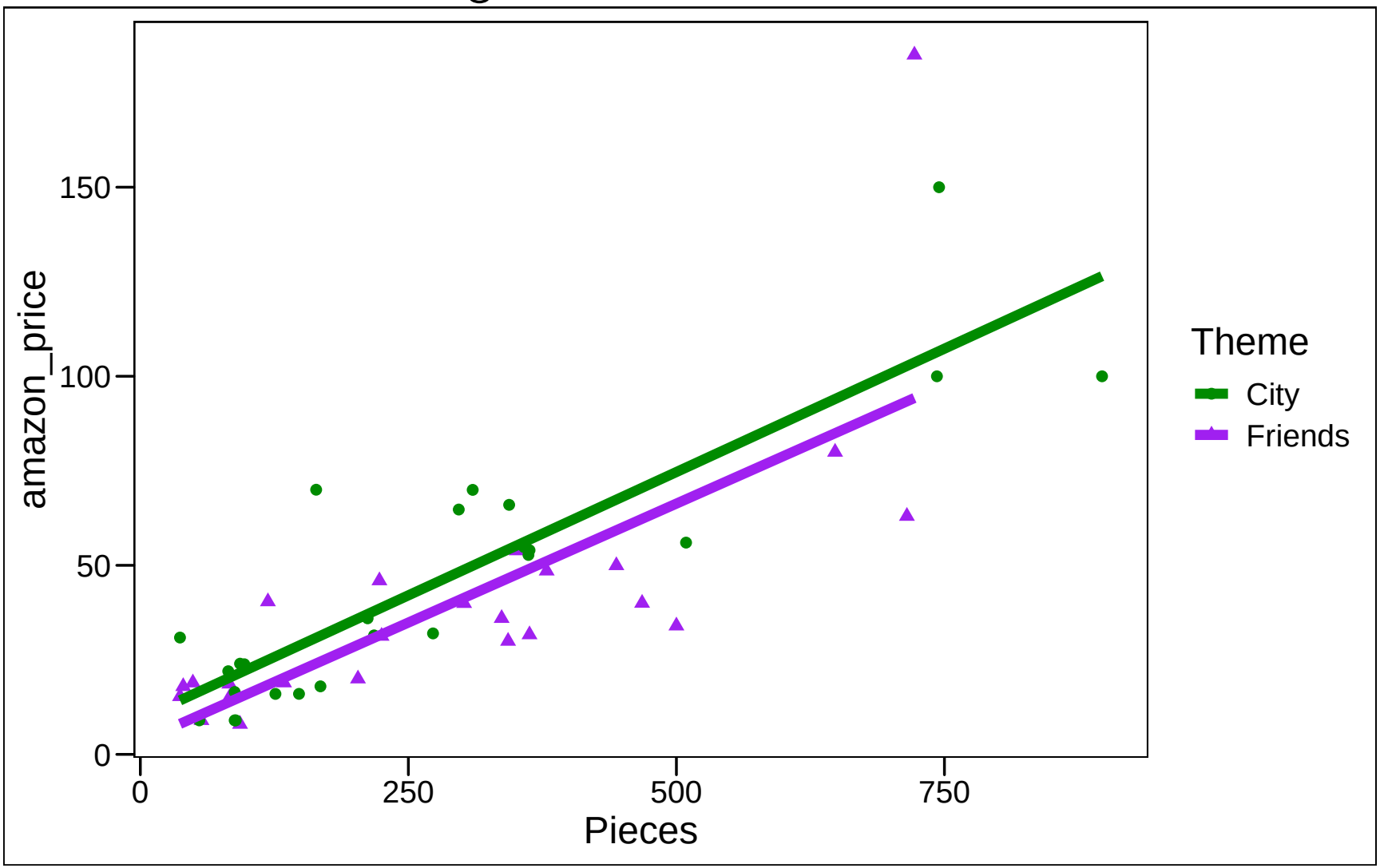
Legos: Test Statistic, p-value, and decision

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.02122    5.42889   1.846   0.0712 .
Pieces        0.12834    0.01314   9.767 6.83e-13 ***
ThemeFriends -7.32372    5.73584  -1.277   0.2079
```

- Test Statistic: -1.277
- p-value: .2079
- Decision 1: We have little to no evidence to suggest that there is a linear relationship between Theme and amazon price after accounting for the number of pieces OR
- Decision 2: We have little to no evidence to suggest that the main effect of Theme on amazon price is not 0 after accounting for the number of pieces
  - *main effect* is the term used when talking about the effect of an indicator variable (tied in with experiment jargon) going from 0 to 1

Legos

So we don't have strong beleif those two lines are different



But WHY do all this?

"...after accounting for the number of pieces."

In an earlier lab we failed the independent and identically distributed assumption because pieces affected price

- We have now accounted for the effect of pieces
- Our test for the difference in the 2 themes is *unbiased* by that (no longer lurking) variable
- THIS is how you deal with lurking variables, broadly
  - Just include them in your model somehow

Brief Comparison

Our estimated average difference between lego sets with City and Friends themes is -$7.32

Simply averaging the two groups (ignoring the number of pieces) gives a difference of -$6.62

More so, the p-value, when using a t-test, more than doubles....

0.2079 vs 0.5059

Useful Guideline

What should you do if a parameter is close to 0 with a large p-value? Can I remove it?

Useful Guideline

What should you do if a parameter is close to 0 with a large p-value?

I will suggest not to remove it because...

Useful Guideline

What should you do if a parameter is close to 0 with a large p-value? I will suggest not to remove it because...

1. You believed it was important a priori
2. A large p-value does NOT mean the coefficient is 0
3. **Kempthorne Princple**: A statistical analysis needs to reflect the physical realities of the scientific mechanism
   - These lego sets are sold under different collections
   - Different target audiences
   - Even if the difference is tiny we should still make our estimates unbiased

(NOTE: That is what this was called at ISU but he was a prof there for like 30 years so idk if this is a thing or an in-house name for it....)

Confidence Intervals

Again, use R and the interpretation carries the same generic format (be sure to keep the "after accounting for...")

We are (BLANK) % confident the true (STATISTIC) is between (LOWER) and (UPPER).

We are 95% confident the true difference in mean cost going from City to Friends legos after accounting for the number of pieces is between -$18.87 and $4.22

```
> confint(mod)
                   2.5 %     97.5 %
(Intercept)   -0.9002932 20.9427279
Pieces         0.1019074  0.1547805
ThemeFriends -18.8627565  4.2153068
```

Confidence Interval Comparison

Linear Regression Interval: -$18.87 and $4.22

t-test interval: -$26.48 to $13.24

Why the large difference in widths?

Confidence Interval Comparison

Linear Regression Interval: -$18.87 and $4.22

t-test interval: -$26.48 to $13.24

Why the large difference in widths?

The number of pieces is explaining changes in amazon price we would assume was random noise otherwise.

- That is our residuals are getting smaller
- The sum of the squares of the errors (adding up the residuals^2) is also getting smaller....(foreshadowing)

IMPORTANT: Lurking variables don't just bias your estimated means but also bias your estimate of variance