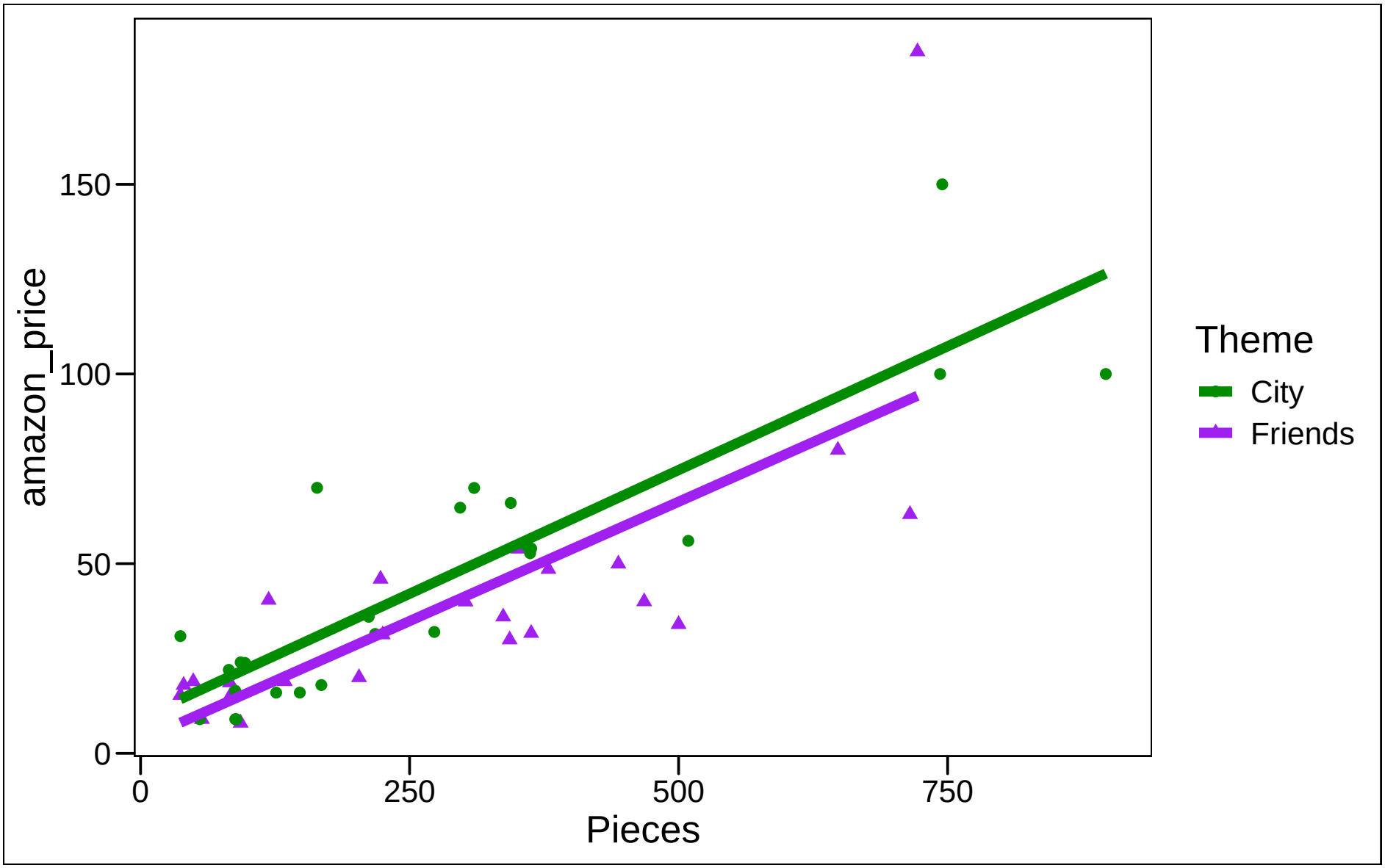
Mulitple Linear Regression

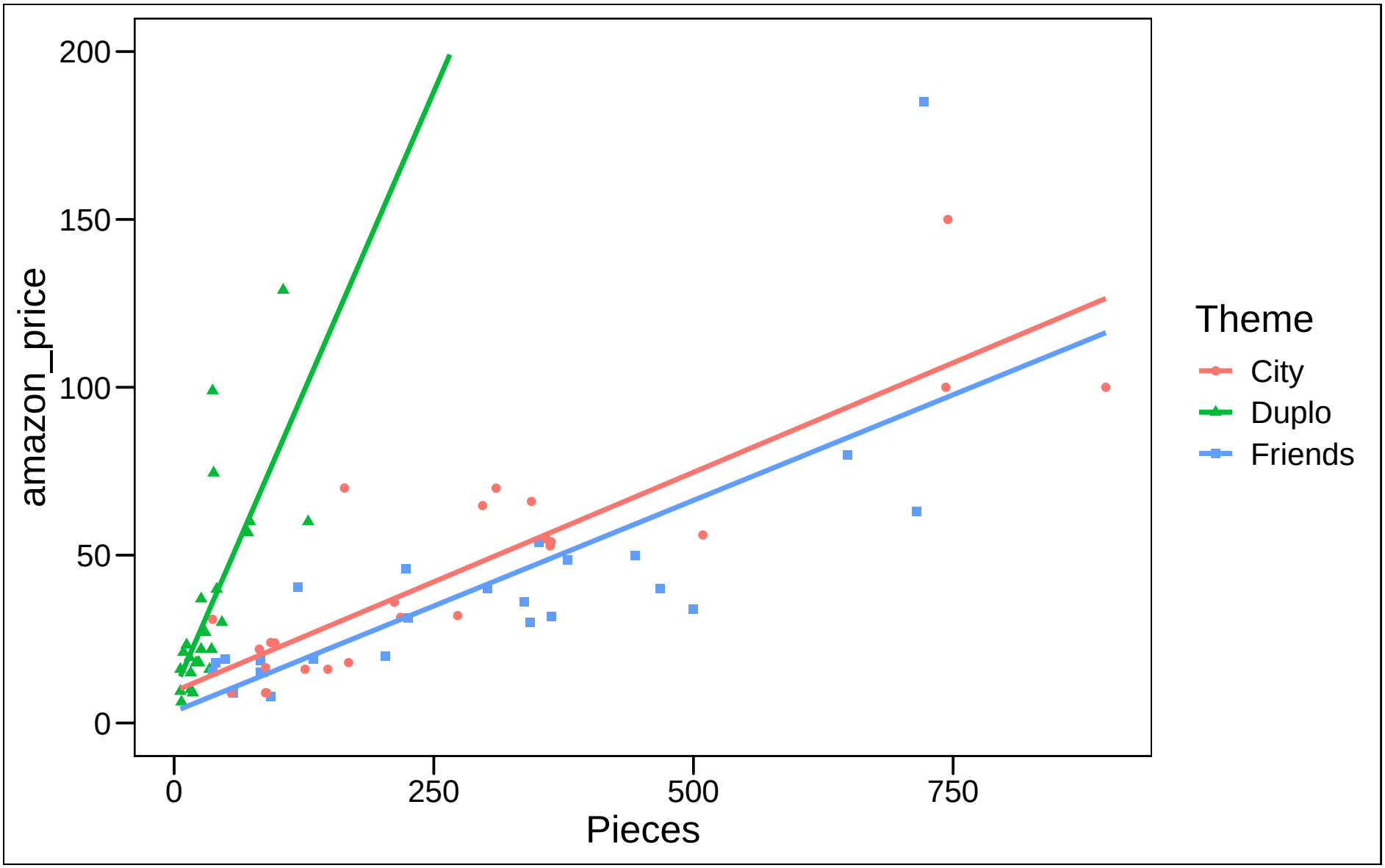
Vinny Paris

December 2025

Review


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

Review


200
150
100
50
0
amazon_price
0
250
500
750
Pieces
Theme
City
Duplo
Friends

Review

From your lego lab: Sometimes we want to fit best-fit-lines to multiple subpopulations

Review

From your lego lab: Sometimes we want to fit best-fit-lines to multiple subpopulations

This (mulitple regression) is how

Simple Linear Regression

Estimated Regression Equation:

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

where $\hat{\beta}$ is our estimated coefficients

- $\hat{\beta}_0$ is the estimated intercept
  - If x $= 0$, we expect the mean of y to be $\hat{\beta}_0$
- $\hat{\beta}_1$ is the estimated slope
  - If x increases by 1, we expect the mean of y to increase by $\hat{\beta}_1$

Simple Linear Regression

Simple linear regression is excellent in simpler cases where there is only one explanatory variable but...

- Doesn't work if there are multiple explanatory variables
  - No way to account for them
  - Introduces bias
  - Violates "identically distributed"
- Might not work if there is a curve
  - eg there needs to be a quadratic term
- Conclusions and generalizing become difficult if there are unaccounted for variables
  - If we know we have subpopulations (that are different) but we ignore that fact how strong are our results?

Mulitple Linear Regression

This is the generalization from SLR to allow for multiple explanatory variables with estimated regresson equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + ...$$

where $\hat{\beta}$ is our estimated coefficients

- Multiple explanatory variables taken in
- All used in one big equation to make our estimate
- The effect of each variable can be measured after accounting for the others
  - Our estimated coefficents are no longer bias
  - Nor is our variance estimate bias anymore

Lego Example

$$\widehat{\text{Amazon Price}} = 37.057 + 0.136\,(\text{Pieces}) - 5.68\,(\text{age})$$

```
> new_mod <- lm(amazon_price ~ Pieces + age, data = legos)
> summary(new_mod)

Call:
lm(formula = amazon_price ~ Pieces + age, data = legos)

Residuals:
    Min      1Q  Median      3Q     Max
-37.546 -13.348  -4.163   5.830  88.929

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.05710    6.81479   5.438 7.03e-07 ***
Pieces       0.13647    0.01639   8.325 3.79e-12 ***
age         -5.68257    1.79650  -3.163  0.00229 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.53 on 72 degrees of freedom
Multiple R-squared:  0.5141,    Adjusted R-squared:  0.5006
F-statistic: 38.09 on 2 and 72 DF,  p-value: 5.205e-12
```

Things to Discuss

1. Interpretation of parameters
2. What is linear model
3. Checking Residuals for MLR
4. Multicolinearity

Parameter Interpretation: Quantitative Variables

Working with the previous lego equation...

$$\widehat{\text{Amazon Price}} = 37.057 + 0.136\,(\text{Pieces}) - 5.68\,(\text{age})$$

We can now ask what does the slope in front of "Pieces" means/how to interpret it.

Guesses?

Parameter Interpretation: Quantitative Variables

Working with the previous lego equation...

$$\widehat{\text{Amazon Price}} = 37.057 + 0.136\,(\text{Pieces}) - 5.68\,(\text{age})$$

We can now ask what does the slope in front of "Pieces" means/how to interpret it.

For a one piece increase in the number of pieces, we think the mean lego set price will increase by \$.14 for a fixed age. OR

For a one piece increase in the number of pieces, we think the mean lego set price will, ceteris paribus, increase by \$.14.

(ceteris paribus: holding all other things equal/the same....so age is held constant)

Parameter Interpretation: Quantitative Variables

Working with the previous lego equation...

$$\widehat{\text{Amazon Price}} = 37.057 + 0.136\,(\text{Pieces}) - 5.68\,(\text{age})$$

So what is the interpretation of -5.68?

Parameter Interpretation: Quantitative Variables

Working with the previous lego equation...

$$\widehat{\text{Amazon Price}} = 37.057 + 0.136\,(\text{Pieces}) - 5.68\,(\text{age})$$

So what is the interpretation of -5.68?

For a one year increase in the recommended age we think the mean cost of a lego set will decrease by \$5.68, ceteris paribus.

Parameter Interpretation: Quantitative Variables

Working with the previous lego equation...

$$\widehat{\text{Amazon Price}} = 37.057 + 0.136\,(\text{Pieces}) - 5.68\,(\text{age})$$

What's the interpretation of the 37.057?

Parameter Interpretation: Quantitative Variables

Working with the previous lego equation...

$$\widehat{\text{Amazon Price}} = 37.057 + 0.136\ (\text{Pieces}) - 5.68\ (\text{age})$$

What's the interpretation of the 37.057?

If the explanatory variables Age and Number of Pieces are 0, we expect the mean amazon price to be $\hat{\beta}_0$

Parameter Interpretation: Qualitative Variables

$$\widehat{\text{Amazon Price}} = 10.02 + 0.128\,(\text{Pieces}) - 7.32\,\mathbb{I}_{Friends}$$

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

Parameter Interpretation: Qualitative Variables

$$\widehat{\text{Amazon Price}} = 10.02 + 0.128\ (\text{Pieces}) - 7.32\ \mathbb{I}_{\textit{Friends}}$$

Guesses on how to interpret the indicator's coefficient?

Parameter Interpretation: Qualitative Variables

$$\widehat{\text{Amazon Price}} = 10.02 + 0.128\ (\text{Pieces}) - 7.32\ \mathbb{I}_{Friends}$$

The estimated difference in the mean cost of a lego set going from a friend theme to a city theme is -$7.32, holding the number of pieces constant.
OR

The estimated difference in the mean cost of a lego set going from a friends theme to a city theme, ceteris paribus, is -$7.32.

Parameter Interpretations: Qualitative Variables

$$\widehat{\text{Amazon Price}} = 10.02 + 0.128\ (\text{Pieces}) - 7.32\ \mathbb{I}_{\textit{Friends}}$$

Interpretation of the 10.02?

Parameter Interpretations: Qualitative Variables

$$\widehat{\text{Amazon Price}} = 10.02 + 0.128\,(\text{Pieces}) - 7.32\,\mathbb{I}_{Friends}$$

Interpretation of the 10.02?

We predict the mean amazon price for a city lego set (baseline cateogry) with 0 pieces to be \$10.02.

Big thing is that you need both the numeric variables set to 0 AND identify the baseline category.

Parameter Interpretations: Numeric Deep Breath

Last few interpretations look very similar to our previous interpretations....

- **Simple Linear Regression**: If the (numeric) explanatory variable is 0, we believe the mean y will be $\hat{\beta}_0$
- **Mulitple Linear Regression**: If the (numeric) explanatory variableS are 0, we believe the mean y will be $\hat{\beta}_0$, holding other variables constant

---

- **Simple Linear Regression**: If the (numeric) variable increases by 1 unit, we expect the mean of y to increase by $\hat{\beta}_1$
- **Multiple Linear Regression**: If a (numeric) variable, call it A, increases by 1 unit, we expect the mean of y to increase by $\hat{\beta}_A$ holding other variables constant

Parameter Interpretations: Categorical Deep Breath

Last few interpretations look very similar to our previous interpretations....

- **Simple Linear Regression**: We expect the mean of y for our baseline category to be $\hat{\beta}_0$
- **Multiple Linear Regression**: We expect the mean of y for our baseline category to be $\hat{\beta}_0$ so long as the numeric variables are 0.

- **Simple Linear Regression**: The estimated difference between the mean of category A and our baseline is $\hat{\beta}_A$
- **Multiple Linear Regression**: The estimated difference between the mean of category A and our baseline is $\hat{\beta}_A$, holding other variables constant

Parameter Interpretation: Special Cases

Polynomial Effect: $\hat{y} = \beta_0 + \beta_1 X + \beta_2 X^2$

The interpretation of β_2 is difficult because it talks about the *quadratic* effect of X...It's coefficent sign (+/-) dictate if the parabola is a $\cup$ or $\cap$

Interaction Effect 1: $\hat{y} = \beta_0 + \beta_1 X_1 * X_2$

For a 1 unit increase in the product of X_1 and X_2 the mean response of y changes by β_1

Interaction Effect 2: $\hat{y} = \beta_0 + \beta_1 \mathbb{I}_{American} + \beta_2 \mathbb{I}_{Girl} + \beta_3 \mathbb{I}_{American} \mathbb{I}_{Girl}$

β_3 is the effect on the mean of the response when going from the sum of the average effects of being an American and being Girl to being (American & Girl)

But what counts as a linear model?

Two mathy ways:

1. Your estimated coefficient is a weighted average of your observed responses (better)

2. The first derivatives of all coefficients in your model are constants (a by-product of the above)

Honestly not super useful to us so instead....

But what counts as a linear model?

Two mathy ways:

1. Your estimated coefficient is a weighted average of your response variables

2. The first derivatives of all coefficients in your model are constants

Honestly not super useful so instead....

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + ...$$

But what counts as a linear model?

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \ldots$$

So long as it's always in the form of .. + (coefficient)(variable) + ... it's a linear model

Is this a linear model?

$$\hat{y} = \beta_0 + \frac{1}{(\beta_1 X_1 + 1)}$$

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_1^{\beta_3}$$

$$\hat{y} = \beta_0 + \beta_1 log(X_1)$$

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 \sin(X_1^2)$$

$$\hat{y} = \beta_0 + \beta_1 X_1 * X_2$$

Is this a linear model?

$$NO: \hat{y} = \beta_0 + \frac{1}{(\beta_1 X_1 + 1)}$$

$$YES: \hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$YES: \hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$

$$NO: \hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_1^{\beta_3}$$

$$YES: \hat{y} = \beta_0 + \beta_1 log(X_1)$$

$$YES: \hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 \sin(X_1^2)$$

$$YES: \hat{y} = \beta_0 + \beta_1 X_1 * X_2$$

Assumptions

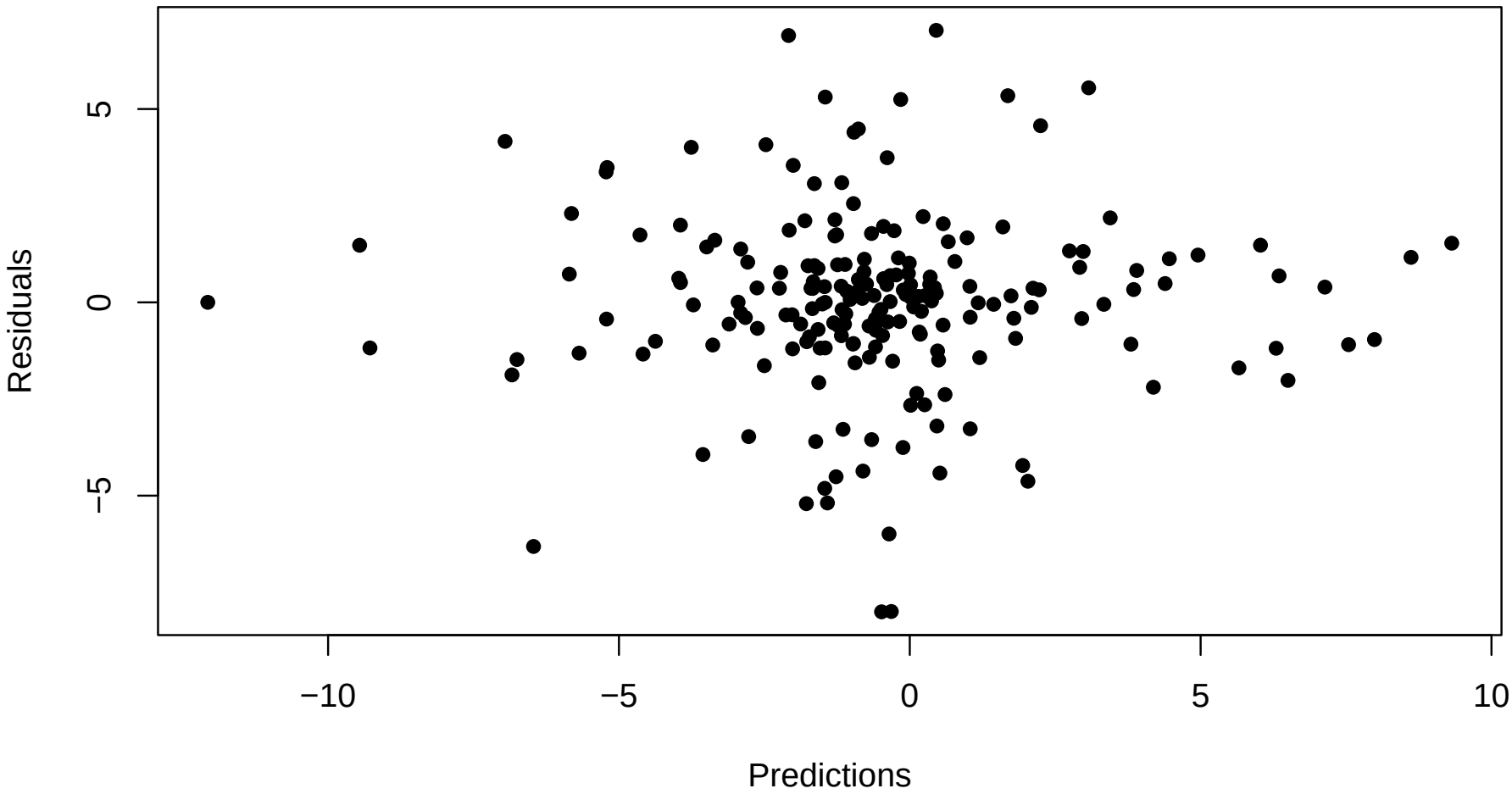
What we saw on Friday[1] with two notes....

[1] Random, IID, Large n or the pop is normal

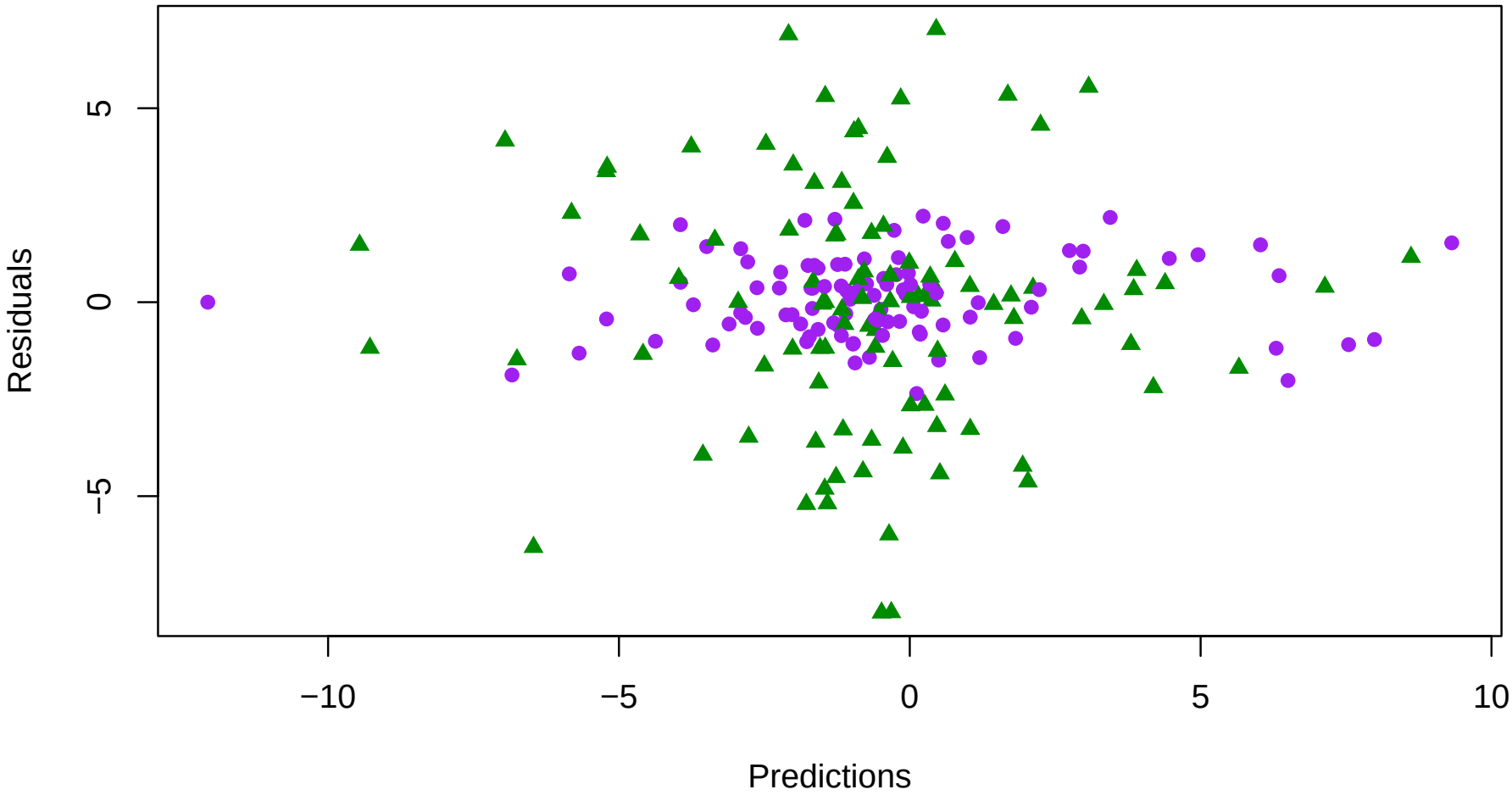
Assumptions: Note 1

ALWAYS COLOR/SHAPE YOUR RESIDUALS BY YOUR EXPLANATORY VARIABLES!!

Assumptions: Note 1



Assumptions: Note 1 with IID failed



Assumptions: Note 2

Recall that in SLR our coefficient was a function of the correlation between the response and the explanatory variable

In order to estimate the coefficient we needed a fair sense of the correlation between the explanatory variable (say x_1) and the response, y

Assumptions: Note 2

Recall that in SLR our coefficient was a function of the correlation between the response and the explanatory variable

In order to estimate the coefficient we needed a fair sense of the correlation between the explanatory variable (say x_1) and the response, y

This can break down though if x_1 and x_2 are strongly correlated

Assumptions: Note 2 Multicolinear

When two explanatory variables are closely correlated the linear model estimates will become unstable.

They are called **colinear**

Assumptions: Note 2 Multicolinear

When two explanatory variables are closely correlated the linear model estimates will become unstable.

They are called **colinear**

If there are multiple (3+) explanatory variables that are closely related they are called **multicolinear**

Assumptions: Note 2 Multicolinear

When two explanatory variables are closely correlated the linear model estimates will become unstable.

$$\widehat{\text{100m dash time}} = \beta_0 + \beta_1 \text{Weight}_{8am} + \beta_2 \text{Weight}_{8pm}$$

Given people's weights don't change that dramatically it'll be hard to...

1. Distinguish between weight at 8am and 8pm
2. Distinguish the different effects on running time

Assumptions: Note 2 Multicolinear

$$\widehat{\text{Boston Marathon Place}} = -566 + .2904\text{Finish (Gun)} - .2174\text{Finish (Net)}$$

- Finish (Gun) is time to finish since the starting gun
- Finish (Net) is the time to finish since crossing the start line

---

- Finish (Gun) has a small p-value (.03) while Finish (Net) is .10
- The 2 explanatory variables have a correlation of .99971
- The equation looks weird.....
  - To my eyes those coefficients feel strange....
  - Taking 3 seconds before you cross the start line bumps our estimate of your place up by almost 1 spot

Assumptions: Note 2 Multicolinear

Strategy to Check: Rerun the analysis with just one explanatory variable and compare the changes

---

$$\widehat{Place} = -578 + .0744 \text{ Net Finish Time}$$

$$\widehat{Place} = -576 + .0742 \text{ Gun Finish Time}$$

Both p-values have my computer bottoming out on how low it can estimate them (2.2×10^{-16}) and the coefficients have changed dramatically

Assumptions: Note 2 Multicolinear

What to do if you find it?

Assumptions: Note 2 Multicolinear

What to do if you find it?

I have no good answer because I have never found a good answer.

The general advice is to...

- Pick one of the two (or more) explanatory variables
- Tell people what you did and why
- Ideally include a small analysis on the side demonstrating the colinearity

Assumptions: Note 2 Multicolinear

What to do if you find it? Specifically...

- Pick only the one that has the best predicitons of the two explanatory variables
- Pick the one that drives the other explanatory variable if that situation exists
- Pick the one that is scientifically better supported
- Pick the one that better captures what you are trying to study

Ultimately we hope that whichever choice you go with that the results and end conclusions don't change or only change a very little.....

Assumptions: Note 2 Multicolinear

Most reasonable analysis should produce the same or similar results

-George Ostrouchov, paraphrased