MEDB 5501, Module 06

2024-09-24

Topics to be covered

- What you will learn
  - Diagnostic plots
  - R code for diagnostic plots
  - Influence measures
  - R code for influence measures
  - Log transformations
  - R code for log transformations
  - Your homework

The population model

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ i = 1, \ldots, N$
  - ϵ_i is an unknown random variable
    - Mean 0, standard deviation, σ
    - Often assumed to be normal
  - β_0 and β_1 are unknown parameters
  - b_0 and b_1 are estimates from the sample

Speaker notes

You only have access to data from a sample, but it is important to recognize that the linear regression estimates from the sample are actually estimates of parameters from the population. In the population model, the subscript i goes from 1 to N. We capitalize N to emphasize that it is much larger than lower case n, the size of your sample.

The term beta0 represents the intercept, the population average value of Y when X=0. It is not an estimate because the population is fixed (no sampling error). The term beta1 represents the slope, the change in the population average of Y when X increases by one unit. Again, this is a fixed value because the population has no sampling error. The term epsilon represents how much an individual Y value in the population deviates from beta0+beta1 times the corresponding X value in the population.

If you want to compute confidence intervals and test hypotheses involving beta0 and beta1, you need to make some assumptions about the epsilons. You have to assume that the epsilons are normally distributed, with a mean of zero and a common standard deviation, sigma. You also have to assume that the epsilons are independent of one another.

Violations of this model

- Nonlinearity
- Heterogeneity
- Non-normality
- Lack of independence

Speaker notes

There are four things that you need to check. The first is non-linearity. Maybe the relationship between X and Y is more complex than a straight line.

The second is heterogeneity. Each epsilon has to have the same standard deviation.

The third is non-normality. The distribution of the epsilons might not follow a bell shaped curve.

The fourth is lack of independence. The epsilons might be related to one another.

Using residuals for diagnostic plots

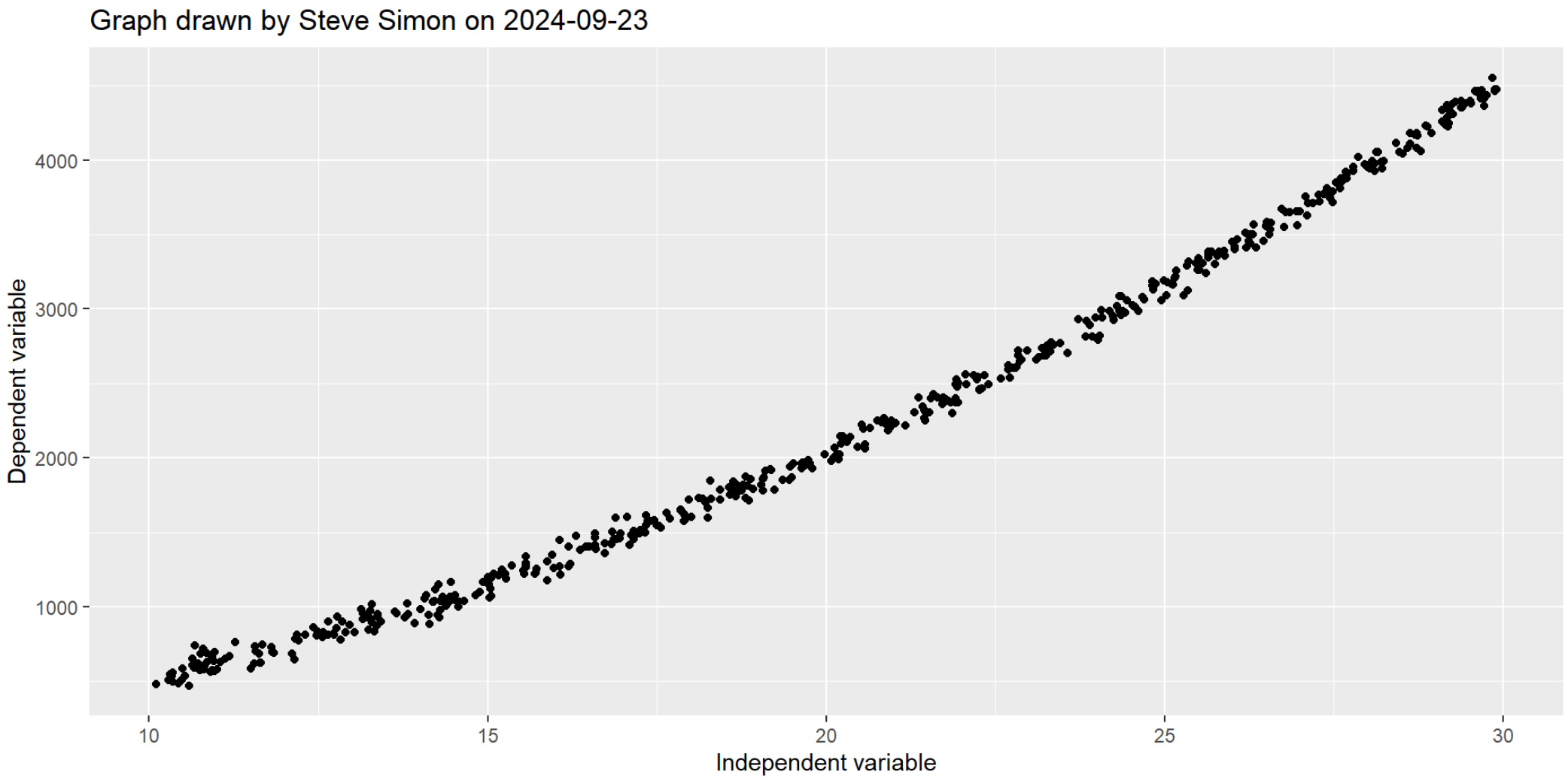
- ϵ_i is unknown, but e_i is known
  - $\epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$
  - $e_i = Y_i - (b_0 + b_1 X_i)$
  - Are there problems with the e_i
    - Indirect evidence of problems with the ϵ_i
- Residuals show patterns more clearly than the original data

Speaker notes

You can not get a direct assessment of whether the epsilons have problems with non-linearity, heterogeneity, non-normality, or lack of independence. The epsilons are unknown, unless you have access to the entire population. You can compute the residuals from the sample using the estimated regression coefficients. It’s not perfect, but if you notice problems in how the residuals behave, that is indirect evidence of problems with the epsilons.

A key point that you will see if that patterns in the dependent variable are often subtle and easy to overlook. This can often occur when there is a strong linear trend. The patterns that indicate problems with the assumptions are often far more apparent when you examine the residuals.

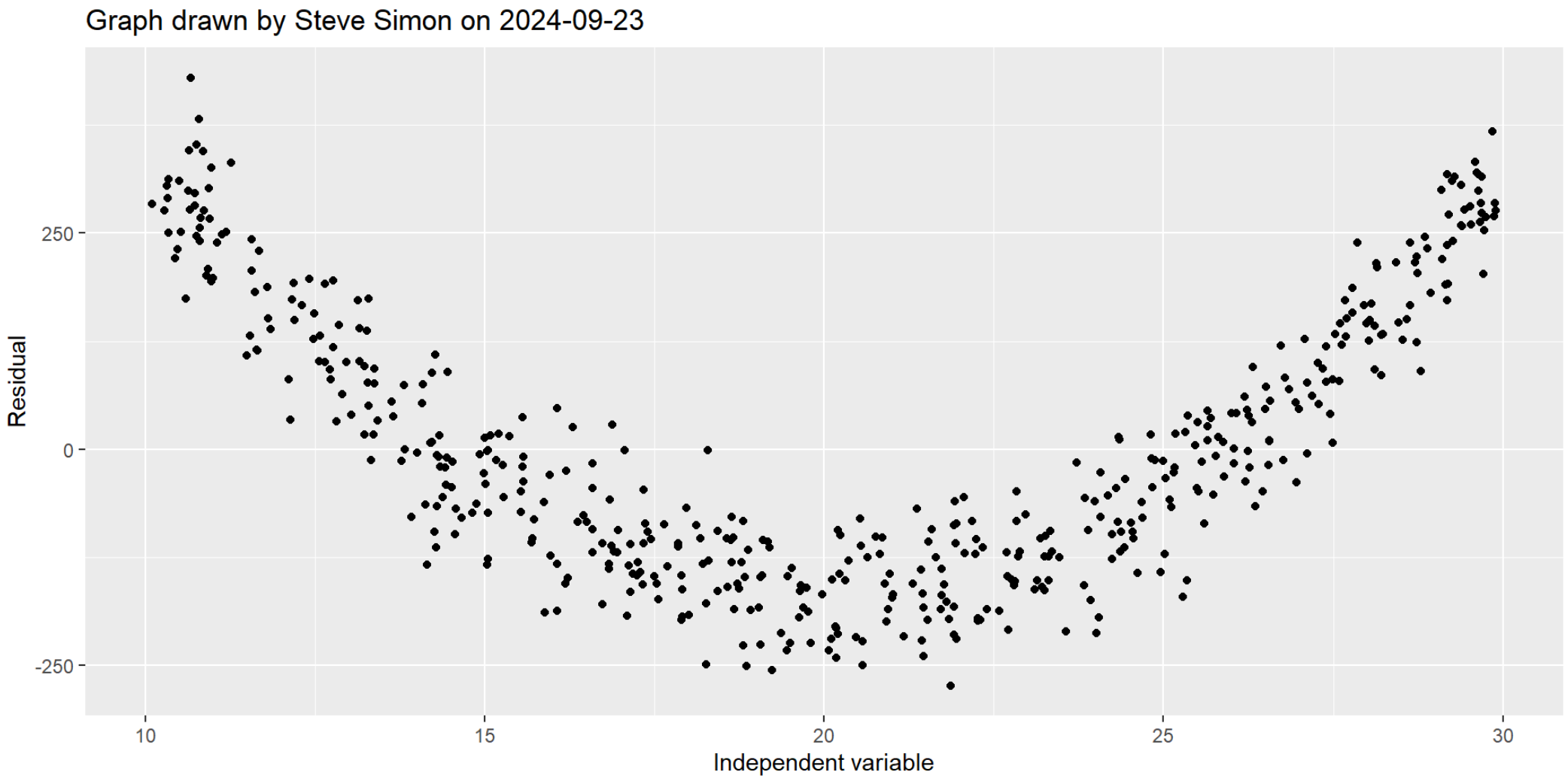
Diagnosing non-linearity, 1



Speaker notes

Here is some artificial data. It shows a strong relationship, and if you look closely, you might detect a slight bend in the data.

Diagnosing non-linearity, 2



Speaker notes

Here is the same data with the residuals on the Y-axis. A slight and very subtle bend becomes hard-to-miss evidence of non-linearity.

Remedies for non-linearity

- Largely beyond the scope of this class
  - Ignore the non-linearity
  - Add a quadratic term
  - Consider a non-linear model
  - Use a spline model

Speaker notes

If you see problems with non-linearity, you have several options. Some of these are beyond the scope of this class.

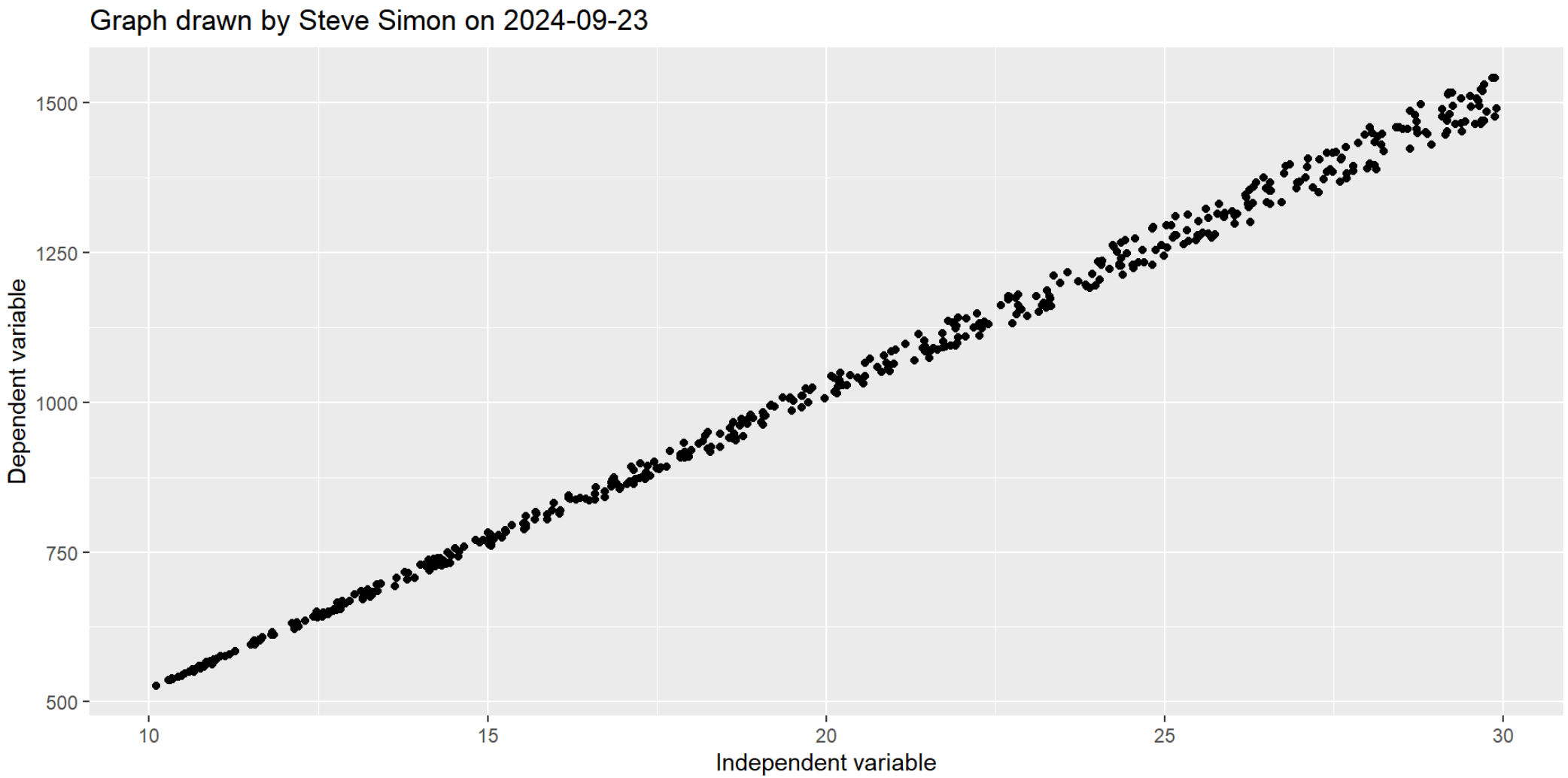
First, it your linear trend is very strong, maybe you can ignore the non-linearity. It means a less than perfect fit, but maybe it would be close enough in a real-world setting.

Second, you might find that a linear term plus a quadratic term does a good job. I will show how to add a quadratic term in the next module.

Third, you might have some information based on your knowledge of research context that would allow you to specify a non-linear function. This is a topic that I might cover in MEDB 5502, Applied Biostatistics II.

Fourth, you might consider a spline model. This is also a topic that I might cover in MEDB 5502.

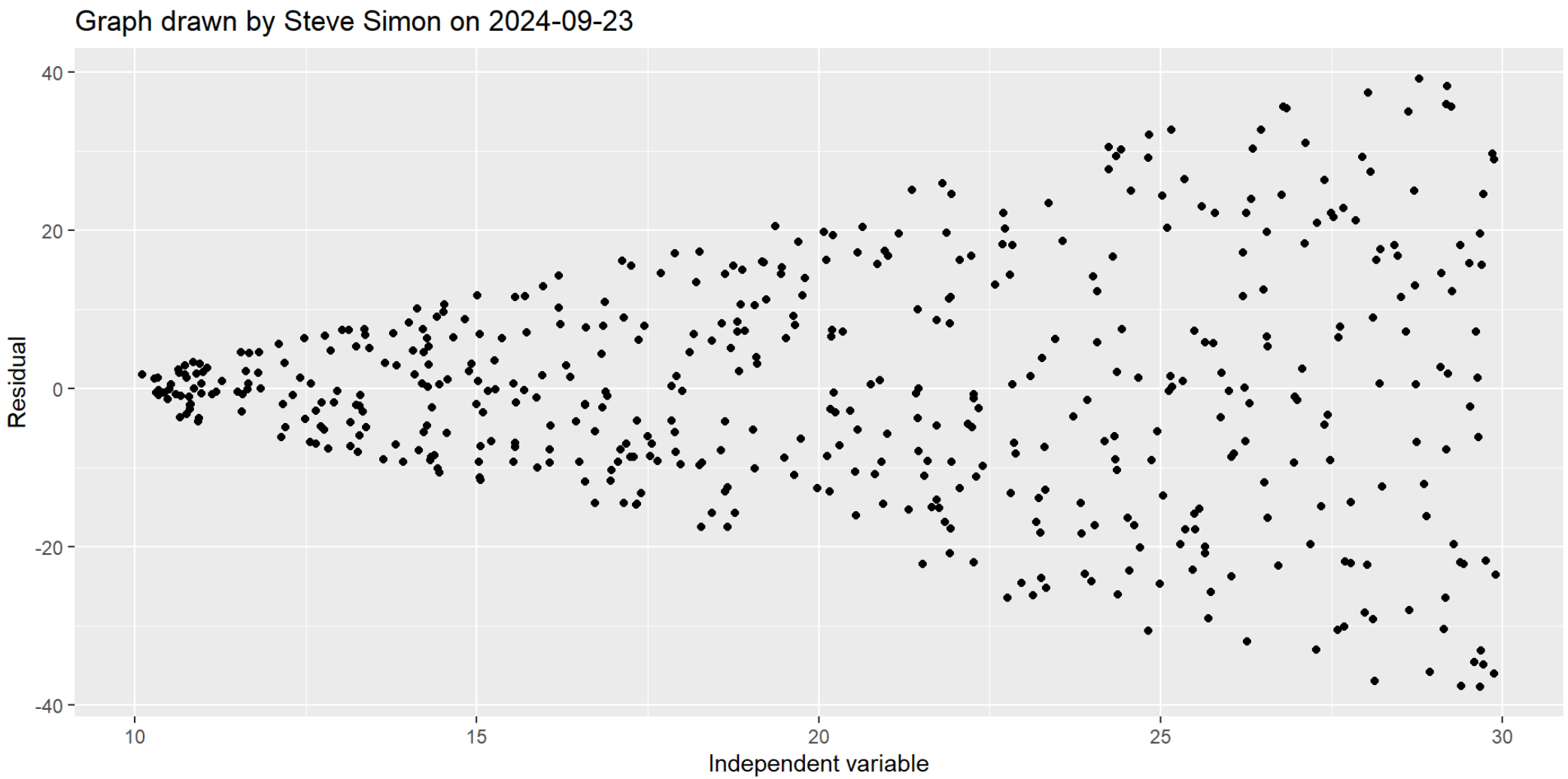
Diagnosing heterogeneity, 1



Speaker notes

Here’s another example of some artificial data. You can see a bit more variation on the high end of the X axis.

Diagnosing heterogeneity, 2



Speaker notes

The change in variation is much more apparent when you look at the residuals. This fanning out pattern is quite common in residual diagnostics.

Remedies for heterogeneity

- Ignore the heterogeneity
- Log transformation
  - Especially useful if larger values have more variation
- Weighted regression

Speaker notes

There are several choices facing you if you observe heterogeneity in your data.

First, a small amount of heterogeneity can be safely ignored. If the variation differs by a factor of two or less, then it may not be worth addressing.

Second, the log transformation can sometimes help. It stretches the small values and squeezes the large values. This can help quite a bit when larger values in your data exhibit larger amounts of variation. I'll show how the log transformation works later in this lecture.

Third, you can assign weights giving greater emphasis to data points with more precision (smaller variation) and lesser emphasis to data points with less precision (larger variation). This is not done too often in practice, but it can sometimes greatly improve the overall stability of your estimates.

Diagnosing non-normality, 1

- Independent variable (X), non-normality is okay
- Dependent variable (Y), non-normality is okay
- Population residual (ϵ), non-normality is not okay
  - Confidence intervals, hypothesis tests not valid
  - Less of a concern with large sample sizes

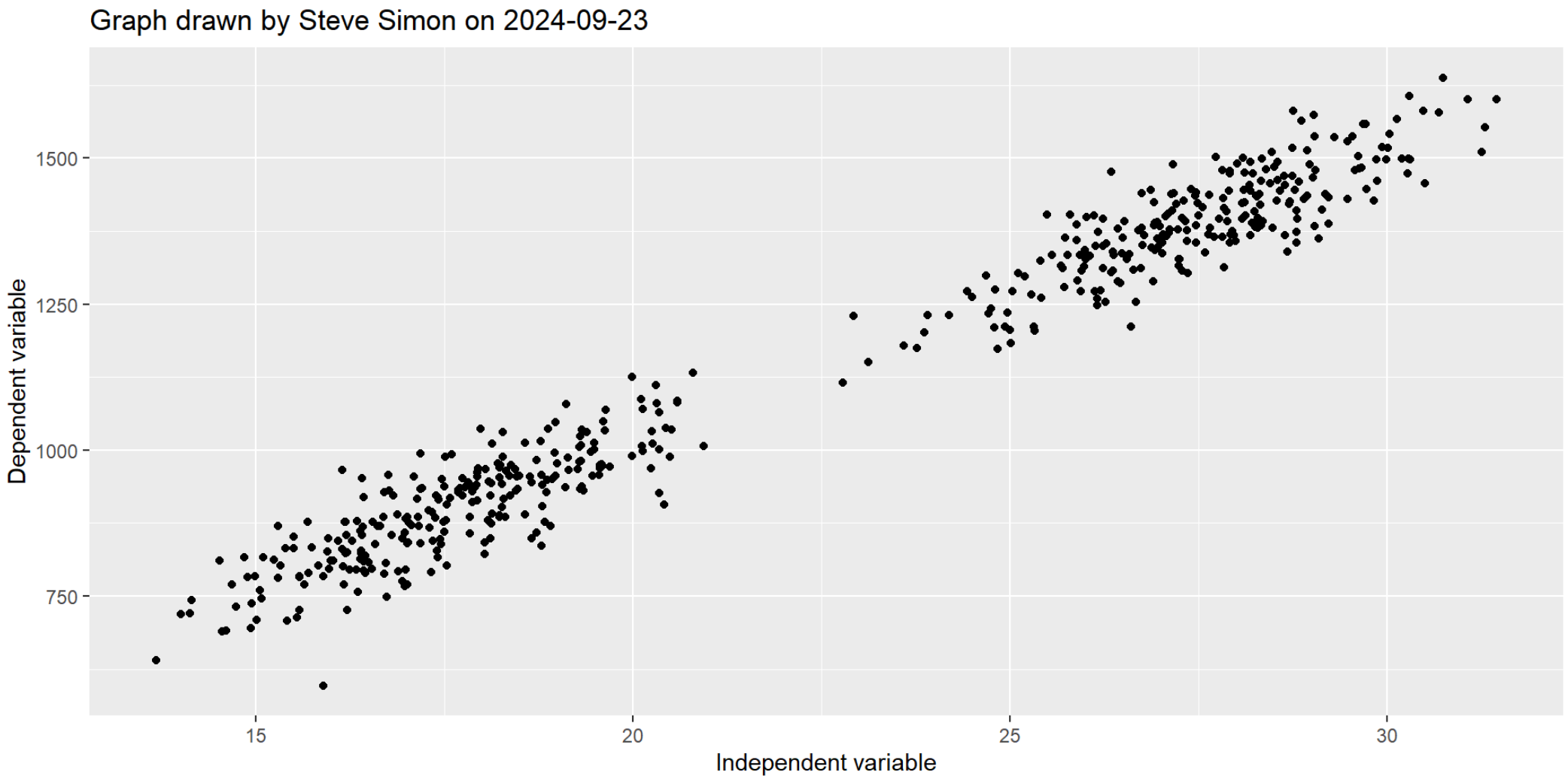
Speaker notes

When you think about non-normality, you must be careful. A non-normal independent variable is okay. It could be bimodal, skewed, or have outliers. None of this matters.

A non-normal dependent variable is also okay. This occurs often when the independent variable has some strange non-normal pattern.

The only thing that needs to be normal is your epsilons. And you diagnose that assumption of normality with the residuals.

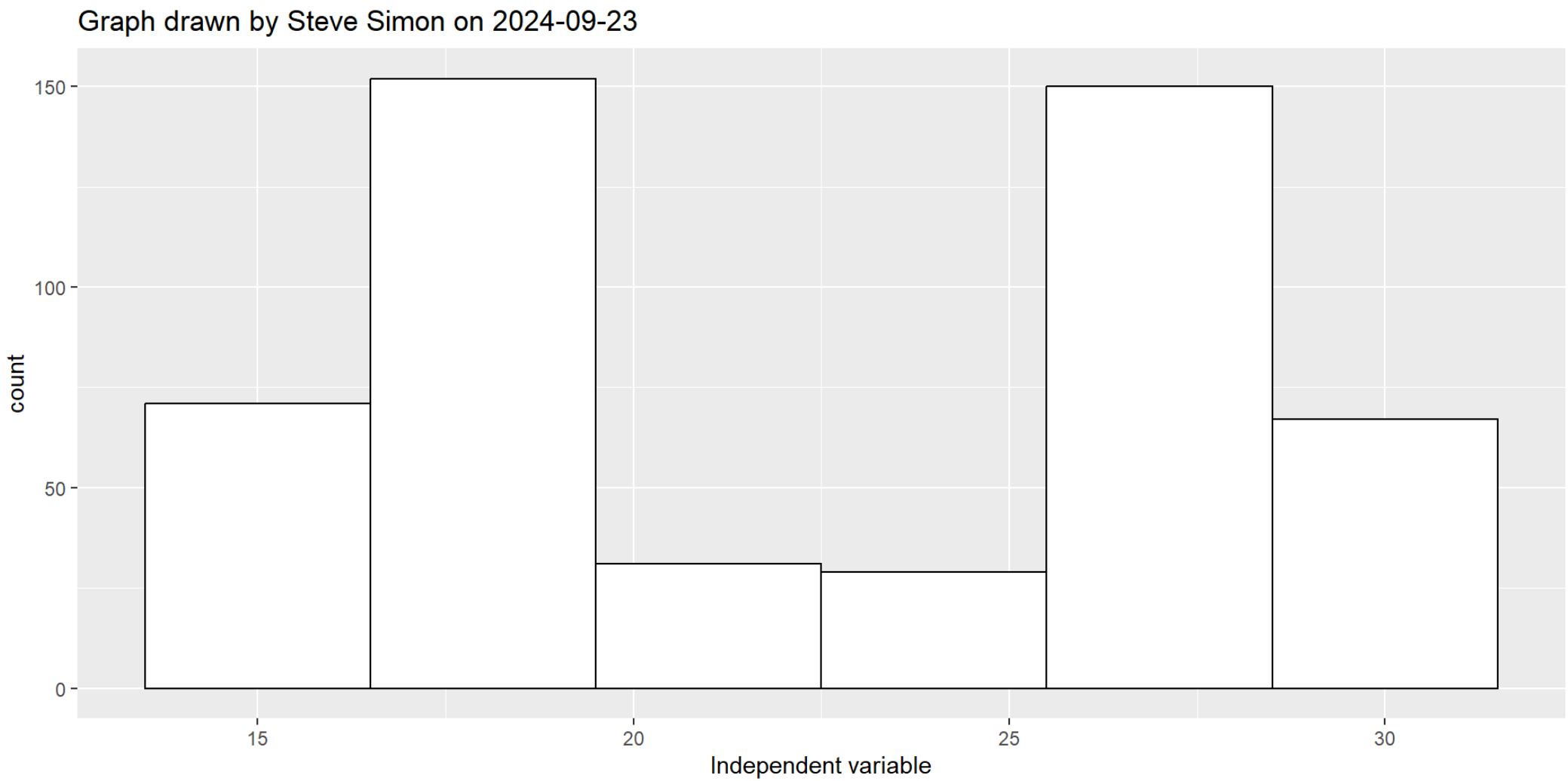
Diagnosing non-normality, 2



Speaker notes

Here is some more artificial data to illustrate where normality is an issue. Notice a gap near the middle. This can happen for a variety of reasons.

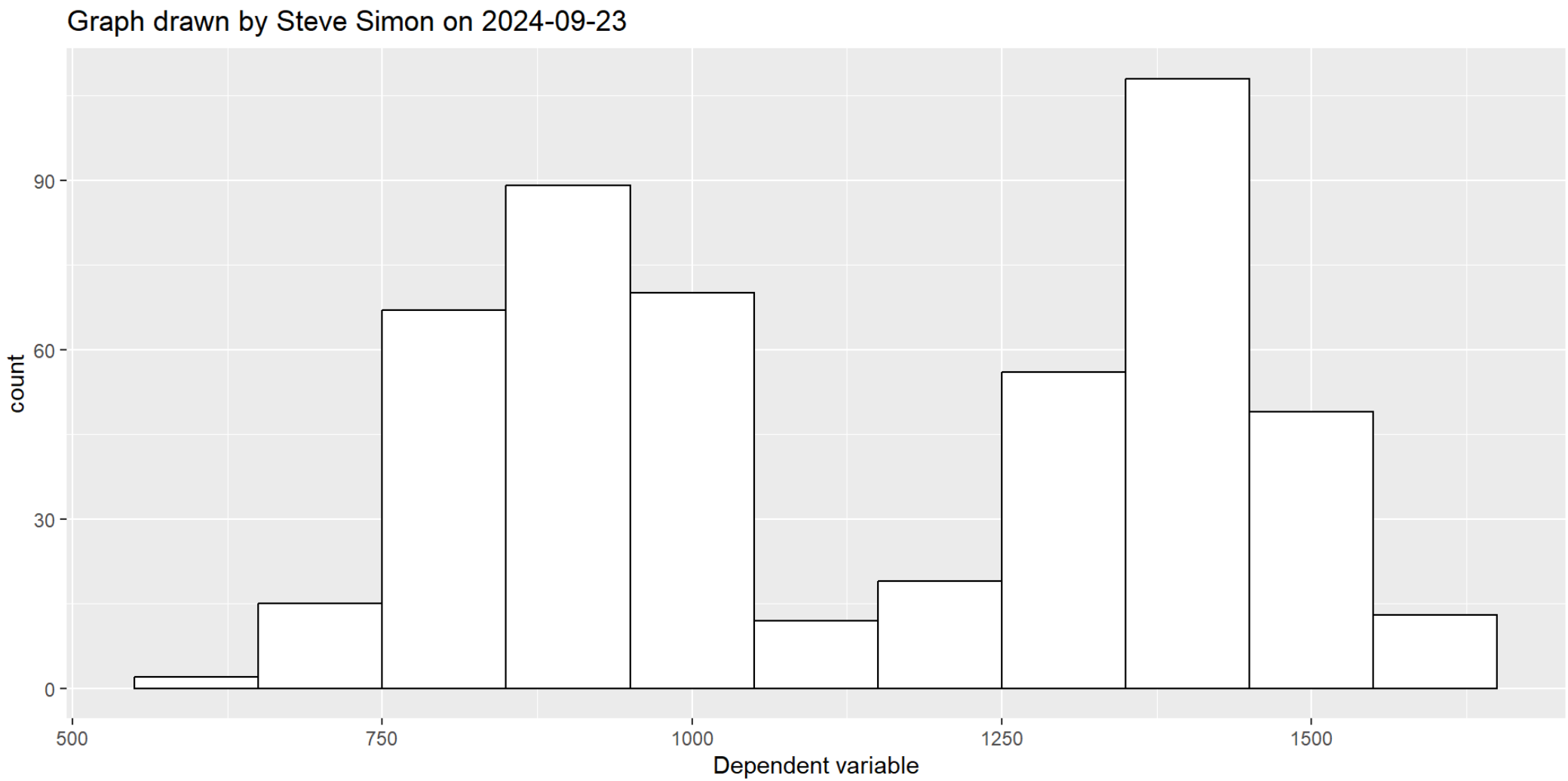
Diagnosing non-normality, 2



Speaker notes

The histogram of the independent variable shows a strong bimodal pattern. This is not a violation of assumptions.

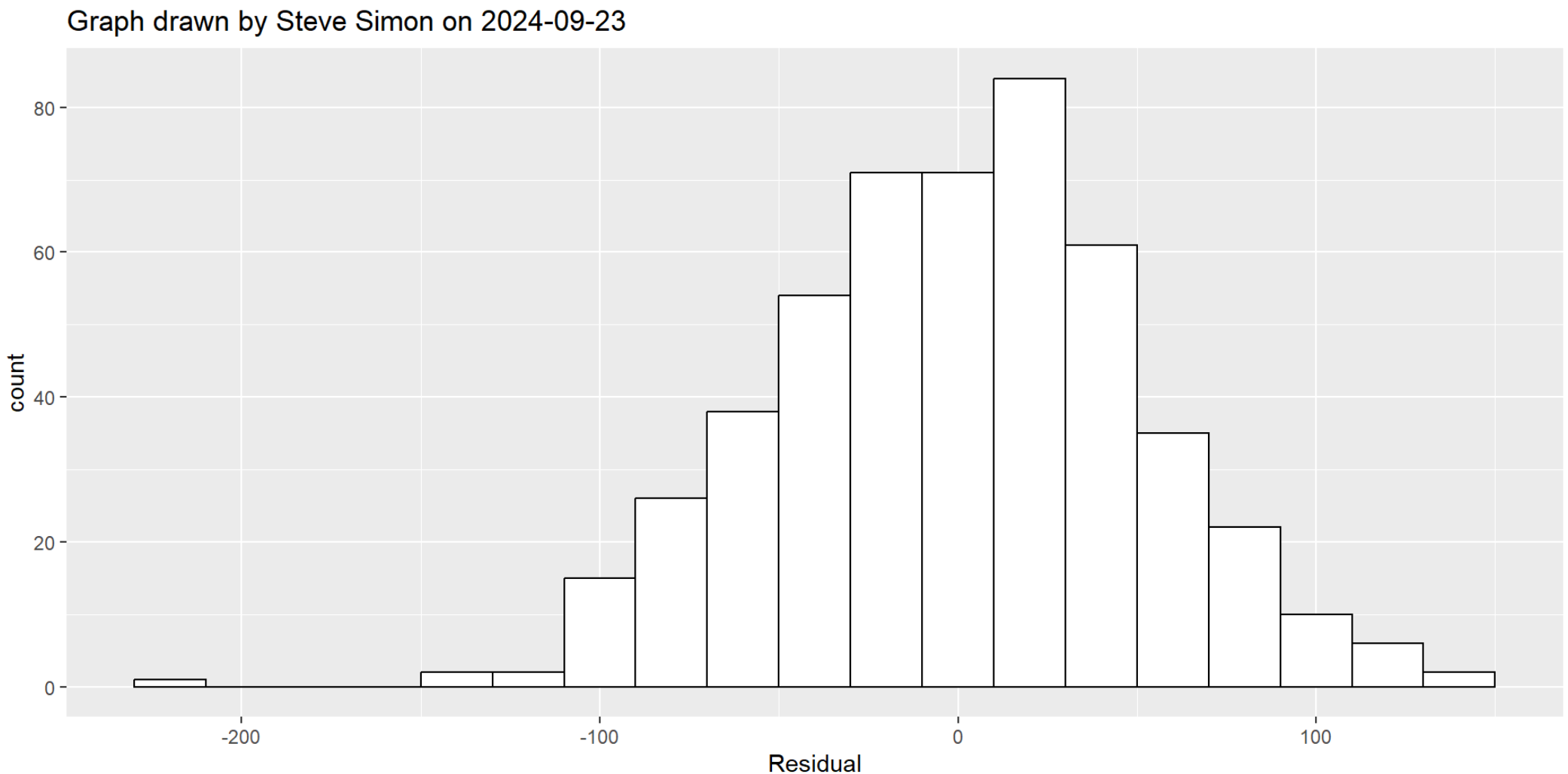
Diagnosing non-normality, 3



Speaker notes

The histogram of the dependent variable also shows a strong bimodal pattern. This is not a violation of assumptions.

Diagnosing non-normality, 4



Speaker notes

It is only a plot of the residuals that is important. In this example, the residual histogram is fairly close to a bell shaped curve, indicating that non-normality is not a concern.

Remedies for non-normality

- Ignore the non-normality
  - Large sample sizes
  - Certain types of non-normality
- Log transformation
  - Especially useful for right-skewed data
- Generalized linear models

Speaker notes

If you see non-normality in the residuals, you have several options.

First, you can sometimes ignore the non-normality. Non-normality is less of a concern for large sample sizes because of the Central Limit Theorem. How large is large? it depends on the type of non-normality. With certain types of non-normality, especially light tailed distributions, you're fine with a few dozen data points. For highly skewed data, or data with many outliers, you may need hundreds of data points before you can safely ignore non-normality.

Second, the log transformation may help here. It works best with fixing problems associated with right-skewed data. Recall that right skewed data has a greater tendency to produce outliers on the high end. This can be fixed by the tendency of the log function to square large values and stretch small values.

General thoughts on assumption tests

- Don’t over-interpret small departures from assumptions
- Normality is not an important assumption with large sample sizes
- A violation of assumptions is actually an opportunity
  - A more complex model could greatly improve things
- Linear regression is always okay as a descriptive tool

Speaker notes

In general, you should take remedial action only for large departures from the underlying assumptions. The main reason for this is that sampling error might produce a false warning.

In particular, the normality assumption is not really as important as the others. Do worry about highly skewed data and extreme outliers. Other types of non-normality are less concerning. Even large deviations from normality are not a serious concern for large sample sizes (several hundred might be sufficient).

Don’t think of a violation of assumptions as a death warrant or an insurmountable barrier. Think of it as an opportunity. The relationship between the independent variable and the dependent variable needs a more complex model. This is more work but it is also an opportunity to produce interesting and valid results.

Break #1

- What you have learned
  - Diagnostic plots
- What’s coming next
  - R code for diagnostic plots

albuquerque-housing, 1

```
data_dictionary: albuquerque-housing
format:
  txt: tab-delimited
  csv: comma-delimited
  sas7bdat: proprietary SAS
  sav: proprietary SPSS
varnames: first row of data
missing_value_code: '.'
```

albuquerque-housing, 2

```
description: |
  From the original source (no longer available)  A random sample of records
of resales of homes from Feb 15 to Apr 30, 1993 from the files  maintained by
the Albuquerque Board of Realtors. This type of data is collected by multiple
listing agencies in many cities and is used by  realtors as an information
base.

download_url:
https://raw.githubusercontent.com/pmean/datasets/master/albuquerque-
housing.csv
```

albuquerque-housing, 3

```
source: |
  DASL (Data and Story Library), a repository for various data sets useful for
teaching. This file was lost in the transition of DASL from statlib to
datadescription.

copyright: |
  Unknown. You should be able to use this data for individual educational
purposes under the Fair Use guidelines of U.S. copyright law.

size:
  rows: 117
  columns: 7
```

albuquerque-housing, 4

```
price:
  label: Sales price of house
  scale: ratio
  unit: dollars
sqft:
  label: Square footage of house
  scale: ratio
  unit: square feet
age:
  label: Age of house
  scale: ratio
  unit: years
```

albuquerque-housing, 5

```
features:
  label: Number of features of house
  scale: ratio
  range: 0 to 13

northeast:
  label: Is house located in Northeast Albuquerque?
  scale: nominal
  value: yes/no
```

albuquerque-housing, 6

```
custom_build:
  label: Is the house custom built?
  scale: nominal
  value: yes/no
corner_lot:
  label: Is the house on a corner lot?
  scale: nominal
  value: yes/no
```

simon-5501-06-albuquerque.qmd, 1

```
---
title: "Regression analysis and diagnostics for Albuquerque housing prices"
author: "Steve Simon"
format:
  html:
    embed-resources: true
date: 2024-08-18
---

This program reads data on housing prices in Albuquerque, New Mexico in 1993.
Find more information in the [data dictionary][dd].

[dd]: https://github.com/pmean/datasets/blob/master/albuquerque-housing.yaml

This code is placed in the public domain.
```

Speaker notes

The first few lines are the documentation header

simon-5501-06-albuquerque.qmd, 2

````
## Load the tidyverse library

For most of your programs, you should load the tidyverse library. The broom
package provides a nice way to compute residuals and predicted values. The
messages and warnings are suppressed.

```{r setup}
#| message: false
#| warning: false
library(broom)
library(tidyverse)
```
````

simon-5501-06-albuquerque.qmd, 3

````
## Read the data and view a brief summary

Use the read_csv function to read the data. The glimpse function will produce
a brief summary.

```{r read}
alb <- read_csv(
  file="../data/albuquerque-housing.csv",
  col_names=TRUE,
  col_types="nnnnccc",
  na=".")
glimpse(alb)
```
````

simon-5501-06-albuquerque.qmd, 4

```
## m1: regression analysis using features to predict price

You might expect that a house with more features would have a higher sales
price. Your first steps are to compute simple descriptive statistics for both
the independent variable (features) and the dependent variable (price). Then
you should plot the data.
```

simon-5501-06-albuquerque.qmd, 5

````
## m1: Calculate descriptive statistics for number of features

```{r features-means}
alb |>
  summarise(
    features_mn=mean(features, na.rm=TRUE),
    features_sd=sd(features, na.rm=TRUE),
    features_min=min(features, na.rm=TRUE),
    features_max=max(features, na.rm=TRUE),
    n_missing=sum(is.na(features)))
```
````

The average number of features is small (3.5) and the standard deviation (1.4) indicates very little variation. At least one house has zero features and no house has all 13 features.

simon-5501-06-albuquerque.qmd, 6

````
## m1: Calculate descriptive statistics for price

```{r price-means}
alb |>
  summarize(
    price_mn=mean(price, na.rm=TRUE),
    price_sd=sd(price, na.rm=TRUE),
    price_min=min(price, na.rm=TRUE),
    price_max=max(price, na.rm=TRUE),
    n_missing=sum(is.na(price)))
```

The average price is low (\$106,000), but the standard deviation (\$38,000)
shows that there is a fair amount of variation. The cheapest house (\$54,000)
is a bargain by today's standards.
````

simon-5501-06-albuquerque.qmd, 7

```
## m1: Plot features versus price

```{r scatterplot-1}
alb |>
  ggplot(aes(features, price)) +
    geom_point() +
    geom_smooth(method="lm", se=FALSE) +
      ggtitle("Plot drawn by Steve Simon on 2023-09-24") +
      xlab("Number of features") +
      ylab("Price in dollars")
```

There is a weak positive relationship between the number of features and the
price of a house.
```

simon-5501-06-albuquerque.qmd, 8

```
## m1: Use features to predict price

```{r regression-1}
m1 <- lm(price~features, data=alb)
m1
```

The estimated average sales price for a house with no features is \$66,000. This not an extrapolation beyond the range of the data. The estimated average sales price increases by \$11,000 for each additional feature. This is surprisingly large when you look at what the features are. Perhaps houses with more features are bigger and newer.
```

simon-5501-06-albuquerque.qmd, 9

```
## Skip some of the functions for hypothesis tests and p-values

Normally, you would follow this up with various functions like anova(),
confint(), or tidy(). This program skips those steps to focus on the
diagnostic plots of the residuals.
```

simon-5501-06-albuquerque.qmd, 10

```
## m1: Calculate residuals and predicted values

```{r residuals-1}
r1 <- augment(m1)
glimpse(r1)
```

You could have also used the resid() and predict() functions. No interpretation is needed here, as these numbers are better reviewed using various graphical displays.
```

simon-5501-06-albuquerque.qmd, 11

```
## m1: Normal probability plot for residuals

```{r qqplot-1}
qqnorm(r1$.resid)
```

The normal probability plot deviates markedly from a straight line, indicating some possible issues with the normality assumption.

Note that you cannnot use ggtitle, xlab, or ylab with the qqnorm function.
```

simon-5501-06-albuquerque.qmd, 12

````
## m1: Histogram for residuals

```{r histogram-1}
r1 |>
  ggplot(aes(.resid)) +
    geom_histogram(
      binwidth=10000,
      color="black",
      fill="white") +
      ggtitle("Plot drawn by Steve Simon on 2023-09-24") +
      xlab("Residuals from m1")
```

The histogram reinforces these concerns. It looks like the data is skewed to
the right.
````

simon-5501-06-albuquerque.qmd, 13

````
## m1: Plot residuals versus features

```{r residual-scatterplot-1}
r1 |>
  ggplot(aes(features, .resid)) +
    geom_point() +
      ggtitle("Plot drawn by Steve Simon on 2023-09-24") +
      xlab("Number of features") +
      ylab("Residuals from m1")
```

This plot is difficult to interpret. There is some evidence of heterogeneity.
It looks, perhaps, like houses with more features also tend to exhibit more
variation. There is no evidence of non-linearity.
````

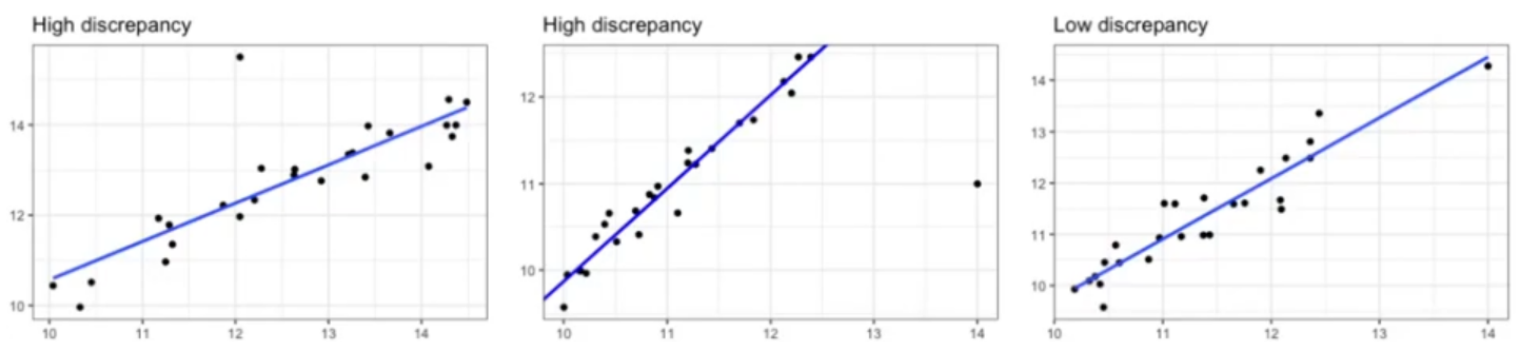
Break #2

- What you have learned
  - R code for diagnostic plots
- What’s coming next
  - Influence measures

Deleted Predicted value

- $\hat{Y}_{j(i)}$ is the predicted value of Y at $X = X_j$ using all the data except (X_i, Y_i)

In a two-variable data set, individual points may be unusual in their *x*-values, their *y*-values, or both. Technically, an **outlier** in such a set is a point with high **discrepancy**, that is, one whose *y*-value is far from the general trend of the data. In practice, the word *outlier* is often used more loosely.

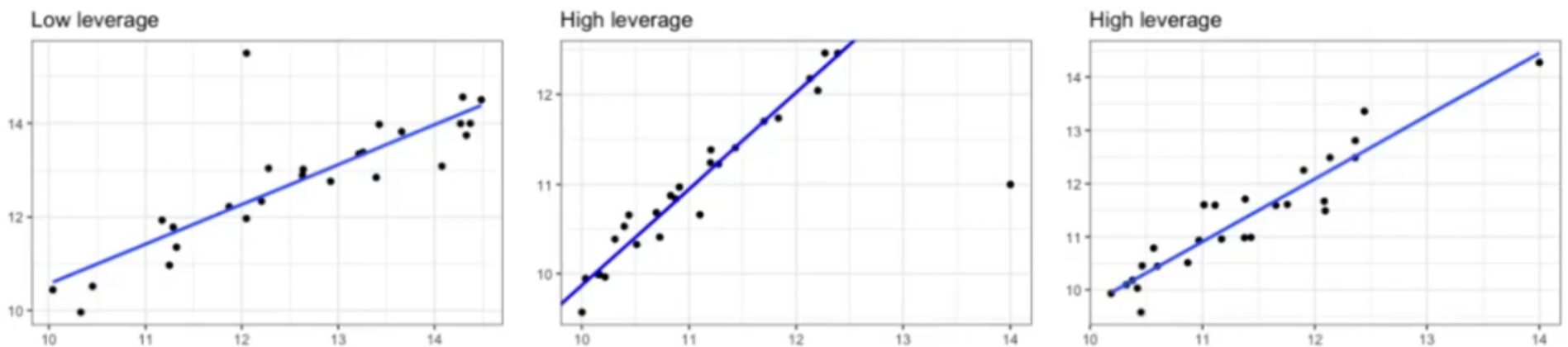


Each of these plots includes an unusual observation. Technically only the first two show outliers.

Leverage

- $h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma(X_j - \bar{X})^2}$

Observations with unusual x-values have greater potential to affect the fit of the model. Such observations are said to have high **leverage**. The leverage of a point only depends on its x-value.

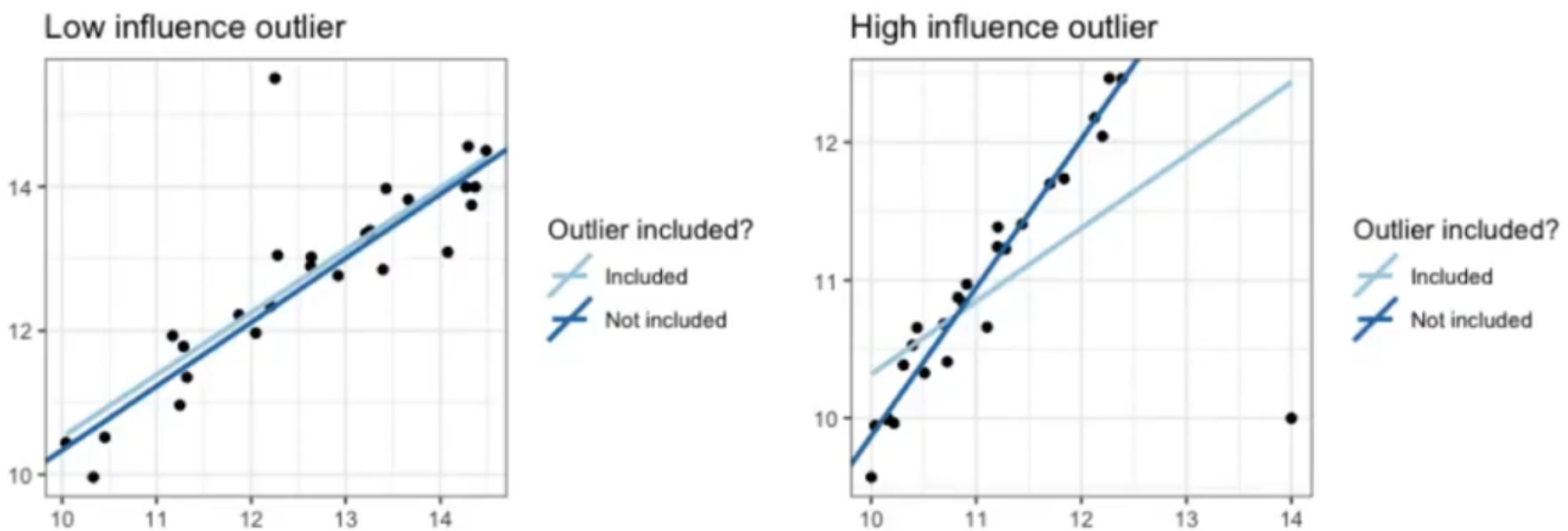


Only the second two plots show observations with high leverage.

Speaker notes

(Add notes)

An observation that substantially changes the fit of a model is said to have high **influence**.



Typically, an observation must have high discrepancy *and* high leverage to be influential.

Studentized residual

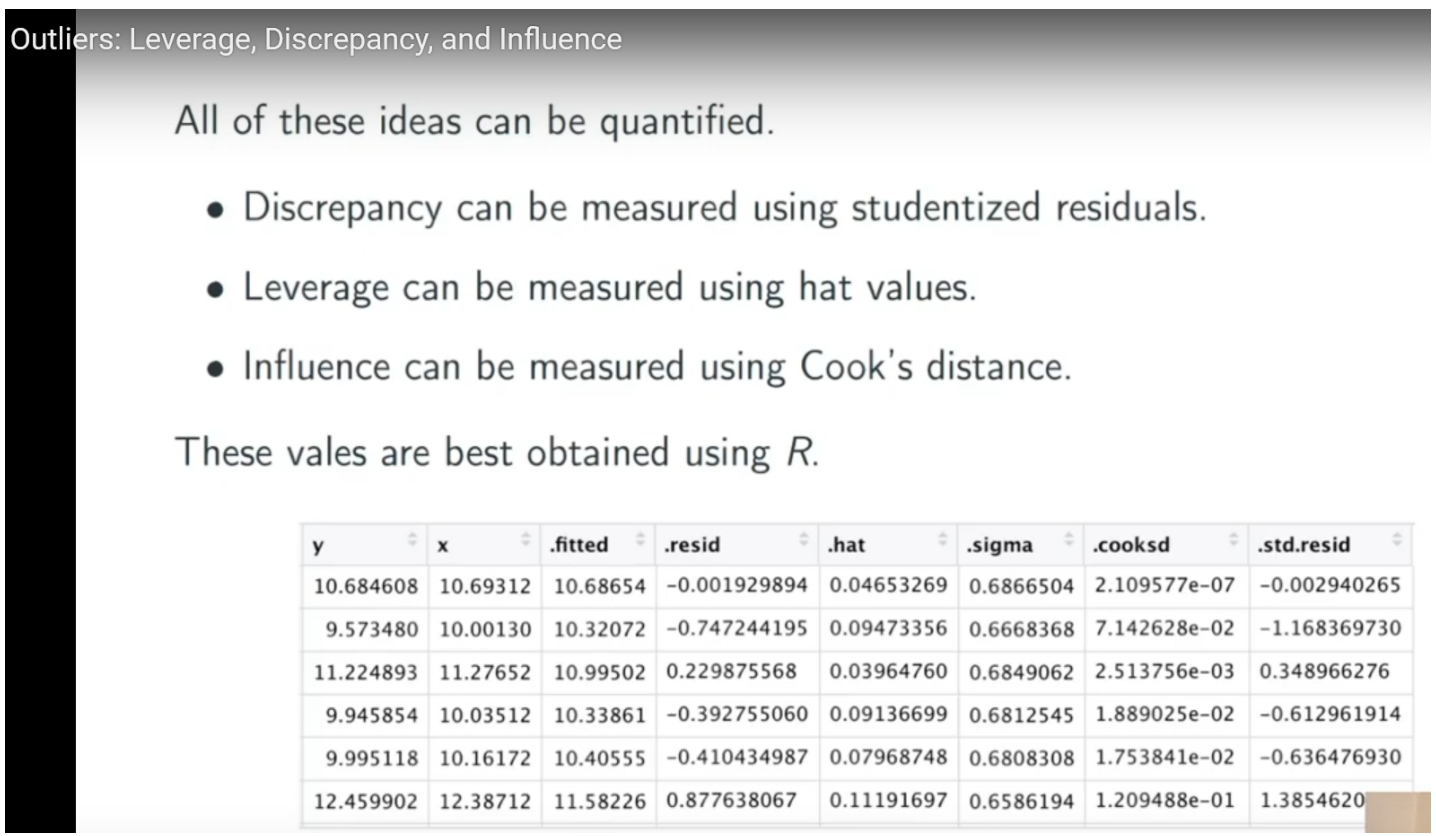
- $Z_i = \frac{e_i}{\sqrt{(1-h_{ii})MSE}}$

Speaker notes

(Add notes)

Cook's distance

- $D_i = \frac{\Sigma(\hat{Y}_j - \hat{Y}_{j(i)})^2}{(k+1)\sqrt{MSE}}$
  - where k is the number of independent variables



y	x	.fitted	.resid	.hat	.sigma	.cooksd	.std.resid
10.684608	10.69312	10.68654	-0.001929894	0.04653269	0.6866504	2.109577e-07	-0.002940265
9.573480	10.00130	10.32072	-0.747244195	0.09473356	0.6668368	7.142628e-02	-1.168369730
11.224893	11.27652	10.99502	0.229875568	0.03964760	0.6849062	2.513756e-03	0.348966276
9.945854	10.03512	10.33861	-0.392755060	0.09136699	0.6812545	1.889025e-02	-0.612961914
9.995118	10.16172	10.40555	-0.410434987	0.07968748	0.6808308	1.753841e-02	-0.636476930
12.459902	12.38712	11.58226	0.877638067	0.11191697	0.6586194	1.209488e-01	1.3854620

Break #3

- What you have learned
  - Influence measures
- What’s coming next
  - R code for influence measures

simon-5501-06-albuquerque.qmd, 14

````
## m1: Leverage values

```{r leverage-1}
n <- nrow(r1)
r1 |> filter(.hat > 3*2/n)
```

There are four data points with high leverage. These correspond to the houses
with the most and the fewest features.
````

simon-5501-06-albuquerque.qmd, 15

````
## m1: Studentized deleted residual

```{r studentized-1}
r1 |>
  filter(abs(.std.resid) > 3)
```

Only one house, with only an average number of features (3) but with the highest sales price (\$215,000), might be considered an outlier.
````

simon-5501-06-albuquerque.qmd, 16

````
## m1: Cook's distance

```{r cook-1}
r1 |>
  filter(.cooksd > 1)
```
````

No houses had a large value for Cook's distance. Even though there are a few high leverage points and one outlier, no single data point has unusually high influence on the predicted values.

Break #4

- What you have learned
  - R code for influence measures
- What's coming next
  - Log transformations

The log transformation

- Useful for correcting non-normality
  - Only if skewed right
- Useful for correcting heterogeneity
  - Only when larger values have larger variation

Which logarithm?

- Natural logarithm, ln, or log_e
  - R function is just log()
  - Difficult to interptet
  - Simple from a mathematics perspective
- Base 10 logarithm, log_{10}
  - R function is log10()
  - Easiest to interpret, powers of 10
- Base 2 logarithm, log_2 - R function is log2() - Powers of 2, common in genomics

Back transformation, 1

- Puts regression coefficients back on original scale
- Antilog inverts log
  - exp() inverts log()
  - 10^ inverts log10()
  - 2^ inverts log2()

Back transformation, 2

- $Y_i^* = log_{10}(Y_i)$ or or
-
- - Estimated average value of Y when X=0
  - Actually a geometric mean

Back transformation, 3

- - Estimated average relative change in Y when X increases by one unit
  - 1.1 represents a 10% increase
  - 0.9 represents a 10% decrease

Break #5

- What you have learned
  - Log transformations
- What's coming next
  - R code for log transformations

simon-5501-06-albuquerque.qmd, 17

m2: Using features to predict log(price)

Because there are some concerns about non-normality and heterogeneity, you might consider using a log transformation for price. In this example, a base 10 logarithm is a reasonable choice.

simon-5501-06-albuquerque.qmd, 18

````
## m2: scatterplot

```{r scatterplot-2}
alb$log_price <- log10(alb$price)
alb |>
  ggplot(aes(features, log_price)) +
    geom_point() +
    geom_smooth(method="lm", se=FALSE) +
      ggtitle("Plot drawn by Steve Simon on 2023-09-24") +
      xlab("Number of features") +
      ylab("Log base 10 of price in dollars")
```

There is a weak positive linear relationship between log price and features.
````

simon-5501-06-albuquerque.qmd, 19

````
## m2: linear regression on log transformed price

```{r regression-2}
m2 <- lm(log_price~features, data=alb)
m2
```

The estimated average log price is 4.8 for a house with no features. The estimated average log price increases by 0.043 for each additional feature. These numbers are easier to interpret when transformed back to the original scale.
````

simon-5501-06-albuquerque.qmd, 20

````
## m2: Coefficients back transformed to original scale

```{r back-transform-2}
10^(coef(m2))
```

The estimated average price is \$71,000 for a house with no features. The
estimated average price increases by 1.10 (10%) for each additional feature.
````

simon-5501-06-albuquerque.qmd, 21

```
## m2: Normal probability plot

```{r qqplot-2}
r2 <- augment(m2)
qqnorm(r2$.resid)
```

The normal probability plot is close to a straight line, indicating a
reasonably close fit to a normal distribution.
```

simon-5501-06-albuquerque.qmd, 22

````
## m2: Histogram of residuals

```{r histogram-2}
r2 |>
  ggplot(aes(.resid)) +
    geom_histogram(
      binwidth=0.05,
      color="black",
      fill="white") +
      ggtitle("Plot drawn by Steve Simon on 2023-09-24") +
      xlab("Residuals from m2")
```

The histogram of residuals also indicates a close fit to a normal
distribution. The regression model using log price does a better job meeting
the normality assumption.
````

simon-5501-06-albuquerque.qmd, 23

````
## m2: Plot residuals versus features

```{r residual-scatterplot-2}
r2 |>
  ggplot(aes(features, .resid)) +
    geom_point() +
      ggtitle("Plot drawn by Steve Simon on 2023-09-24") +
      xlab("Number of features") +
      ylab("Residuals from m2")
```

This plot is difficult to interpret. There is certainly no evidence of non-linearity, but perhaps the problems with heterogenity persist even after the log transformation. Houses with zero or one features seem to have less variation than the rest of the data.
````

Break #6

- What you have learned
  - R code for log transformations
- What’s coming next
  - Your homework

simon-5501-06-directions.qmd, 1

```
---
title: "Directions for 5501-01 programming assignment"
author: "Steve Simon"
format:
  html:
    embed-resources: true
date: 2024-08-18
---

This code is placed in the public domain.
```

simon-5501-06-directions.qmd, 2

```
## Program

-   Download [simon-5501-06-albuquerque.qmd][tem]
    -   Store it in your src folder
-   Modify the file name
    -   Use your last name instead of "simon"
-   Modify the documentation header
    -   Add your name to the author field
    -   Optional: change the copyright statement

[tem]: https://github.com/pmean/classes/blob/master/biostats-1/01/src/simon-
5501-01-template.qmd
```

simon-5501-06-directions.qmd, 3

```
## Data

-   Download [albuquerque-housing-prices.csv][dat]
    -   Store it in your data folder

[dat]: https://github.com/pmean/datasets/blob/master/albuquerque-housing.csv
```

simon-5501-06-directions.qmd, 4

Question 1

Calculate descriptive statistics (mean, standard deviation, minimum, and maximum for sqft. Interpret these numbers

Question 2

Draw a plot with price on the y-axis and sqft on the x-axis. Include a linear regression line, but do not extend it beyond the range of the data. Interpret this plot.

Question 3

Calculate a linear regression model using sqft to predict price. Interpret the slope and intercept.

Question 4

Draw a normal probability plot and a histogram for the residuals (.resid). Interpret these plots.

simon-5501-06-directions.qmd, 5

Question 5

Draw a scatterplot of sqft on the x-axis and the residuals on the y-axis. Is there evidence of non-linearity or heterogeneity?

Question 6

Display the data (if any) for leverage values greater than 3*2/n. Describe where these leverage values are found relative to the independent and/or dependent variables.

Question 7

Display the data (if any) for studentized deleted residuals (.std.resid) values greater than 3. Describe where these leverage values are found relative to the independent and/or dependent variables.

Question 8

Display the data (if any) for Cook's distance (.cooksd) values greater than 1. Describe where these leverage values are found relative to the independent and/or dependent variables.

simon-5501-06-directions.qmd, 6

Question 9

Calculate the regression equation predicting log10 of price using sqft. Transform the coefficients back to the original scale of measurement and interpret these values.

Question 10

Calculate diagnostic plots (normal probability plot, histogram, and sqft versus residuals). Do these plots show that a model using log10 price better meets the assumptions for linear regression?

Your submission

- Save the output in html format
- Convert it to pdf format.
- Make sure that the pdf file includes
  - Your last name
  - The number of this course
  - The number of this module
- Upload the file

If it doesn't work

If your program has any errors or fails
to produce the output that you desire

and you can't resolve the problem,
upload the program file along with the
pdf file to help us figure out what
went wrong. You will get a chance to
resubmit the assignment if needed.

Summary

- What you have learned
  - Diagnostic plots
  - R code for diagnostic plots
  - Influence measures
  - R code for influence measures
  - Log transformations
  - R code for log transformations
  - Your homework