Simple Linear Regression, Assumptions

CSU Chico, Math 314

2018-12-02

outline

outline

recap, linear regression assumptions

There are four assumptions about linear regression, three of which are easy to check

- Linearity – the data should show a linear trend.
- Independent observations – no two points are dependent on each other.
- Constant Variability – variation of points around least squares line remains roughly constant.
- Normality – the residuals should be nearly normal.

outline

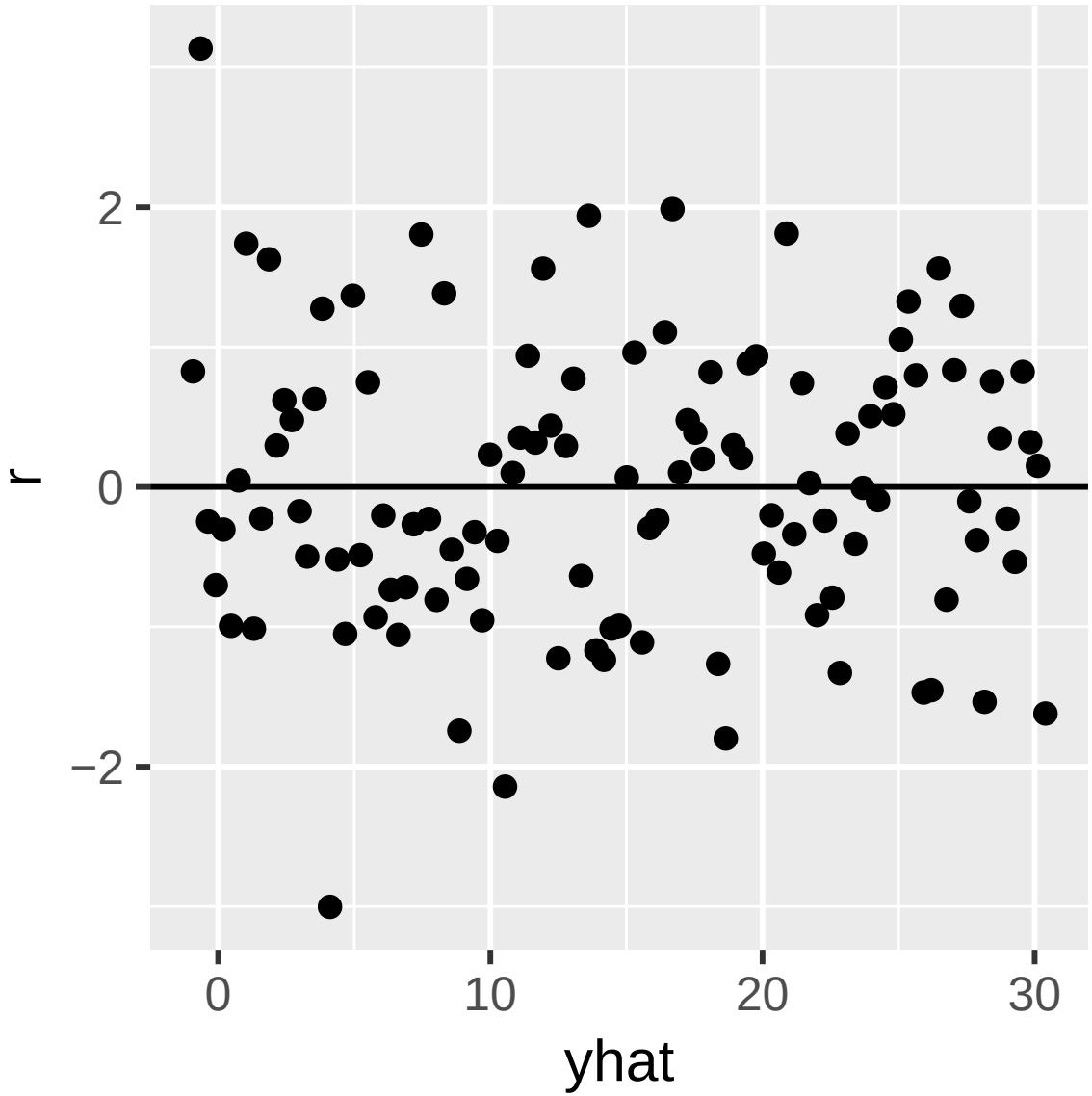
Residuals on Fitted Values

Scatter plots of (standardized) residuals on fitted values help you check the assumptions

1. linearity
2. constant variation

Linearity, residuals on fitted values

Good.



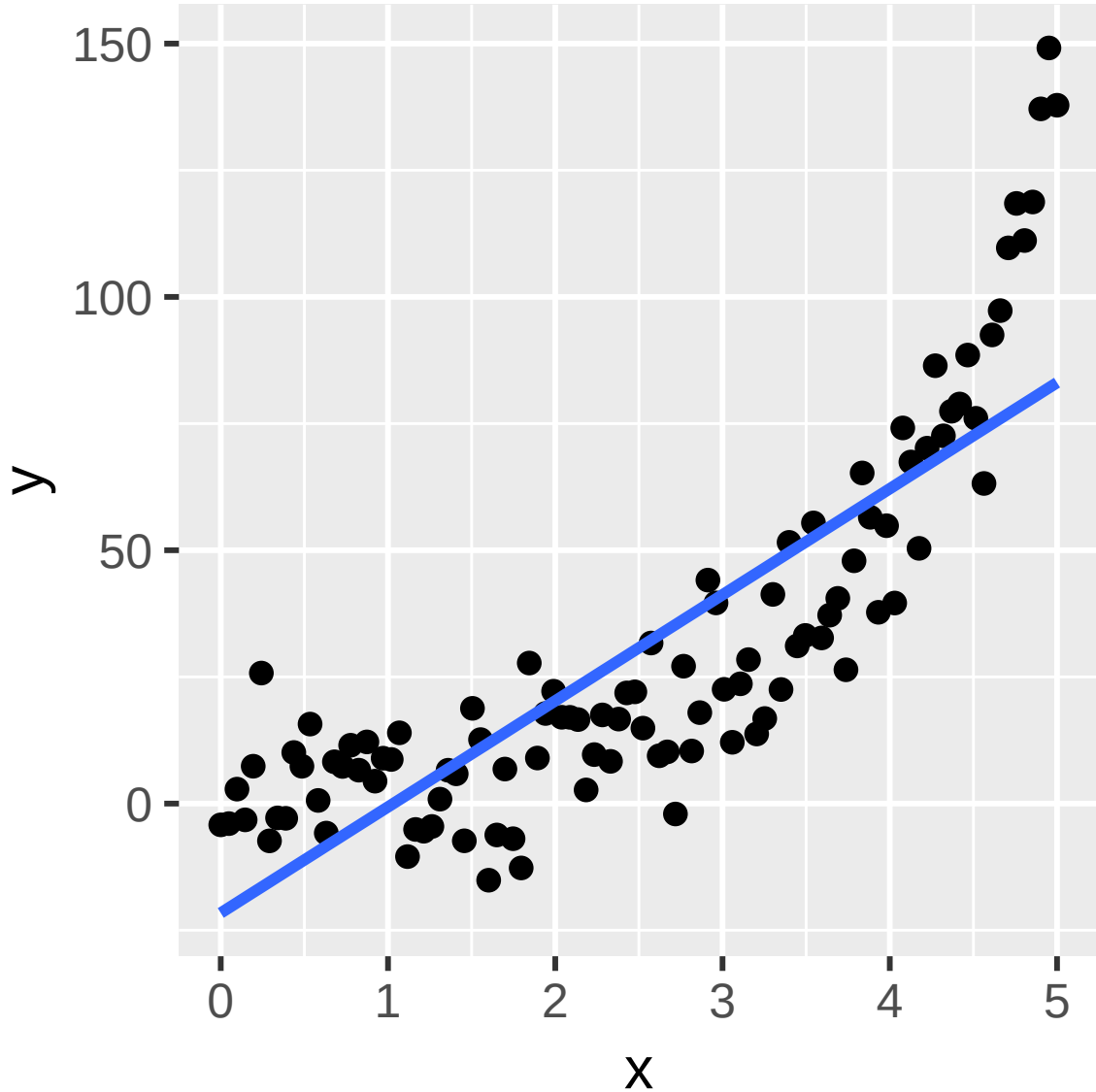
Linearity, residuals on fitted values

What do we like about the above plot?

- Linearity? Yes, because no consistent pattern to *these* data.

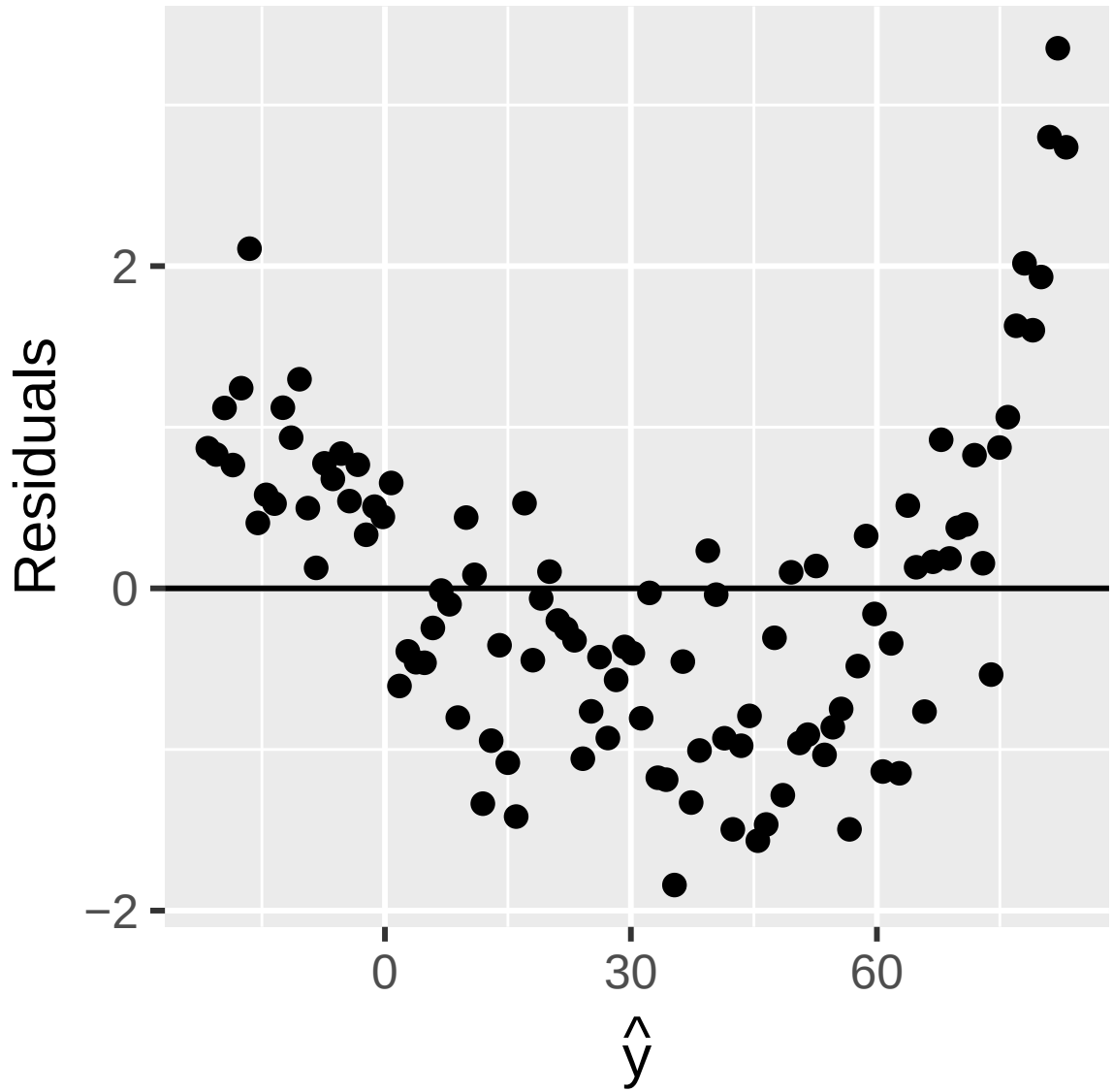
Linearity, residuals on fitted values

Imagine we fit linear regression to non-linear data.



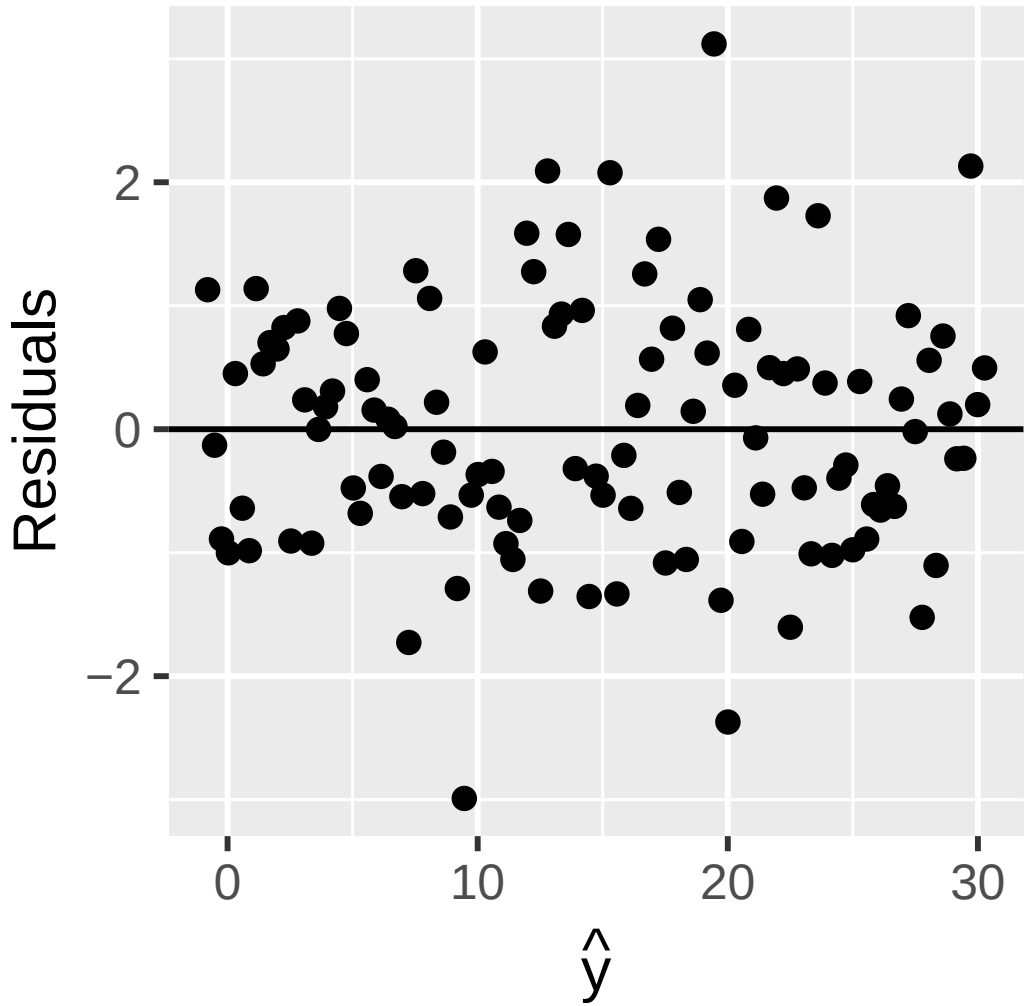
Linearity, residuals on fitted values

Bad. The residuals are not randomly scattered about, but instead have a clear pattern.



Linearity, residuals on fitted values

Good.



- Residuals show no clear pattern as a function of $\hat{y}$.

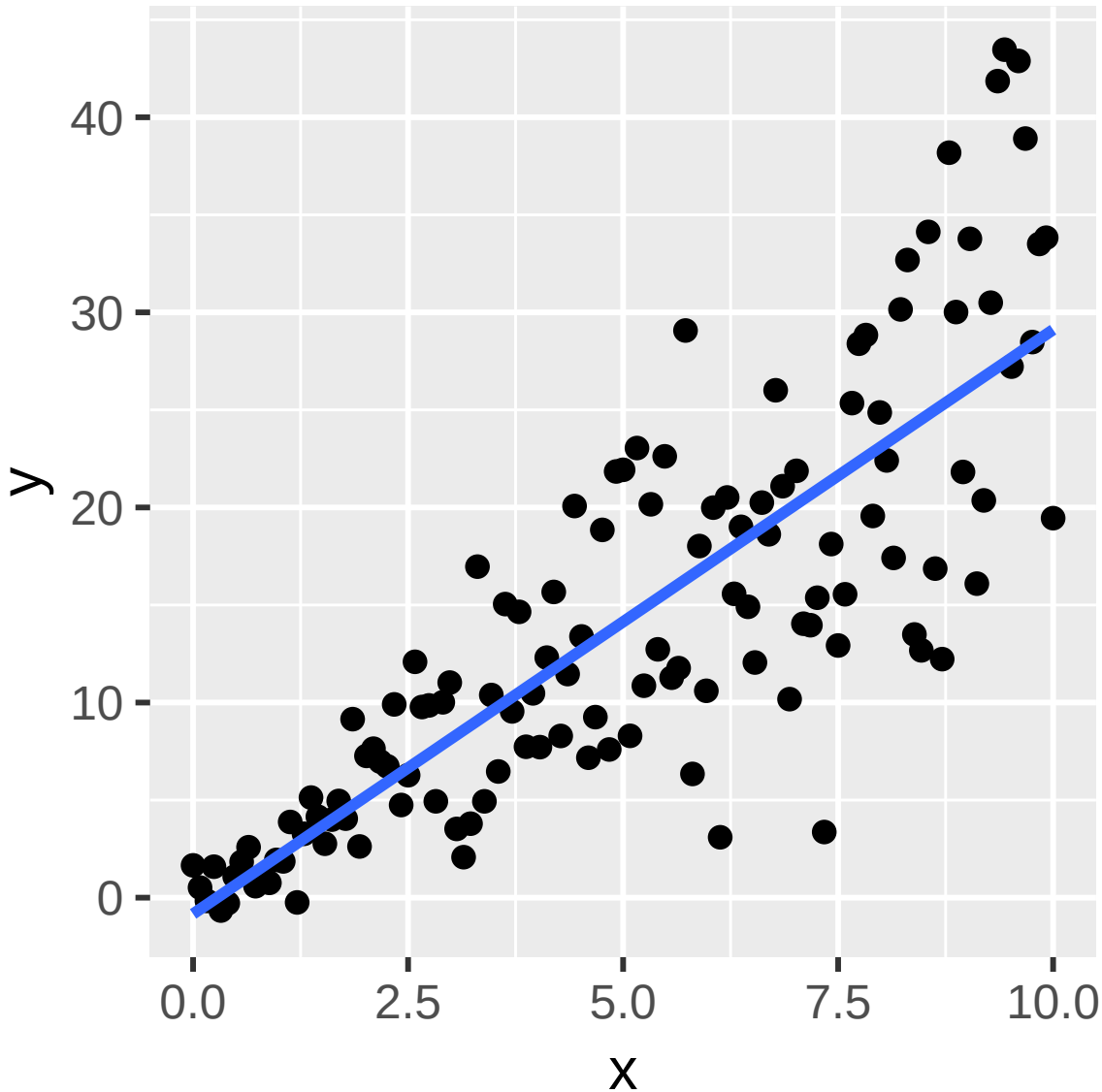
Linearity and Constant Variation, residuals on fitted values

What do we like about the above plot?

- Linearity? Yes, because no consistent trend to the data.
- Constant Variability? Yes, because no (horizontal) megaphone/cone (nor alligator/pacman mouth).

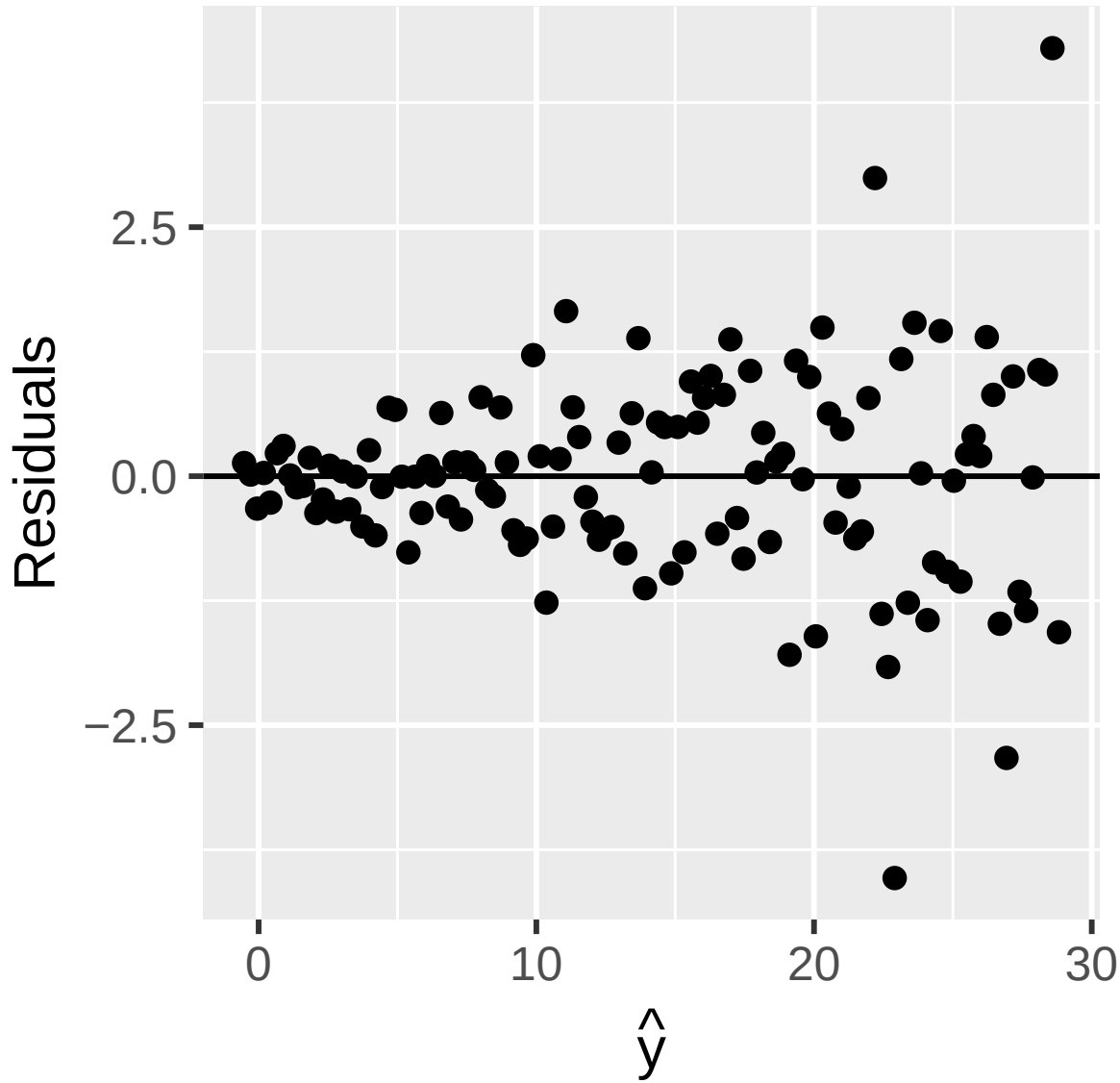
Constant Variation, residuals on fitted values

Imagine we fit linear regression to data with non-constant variability.



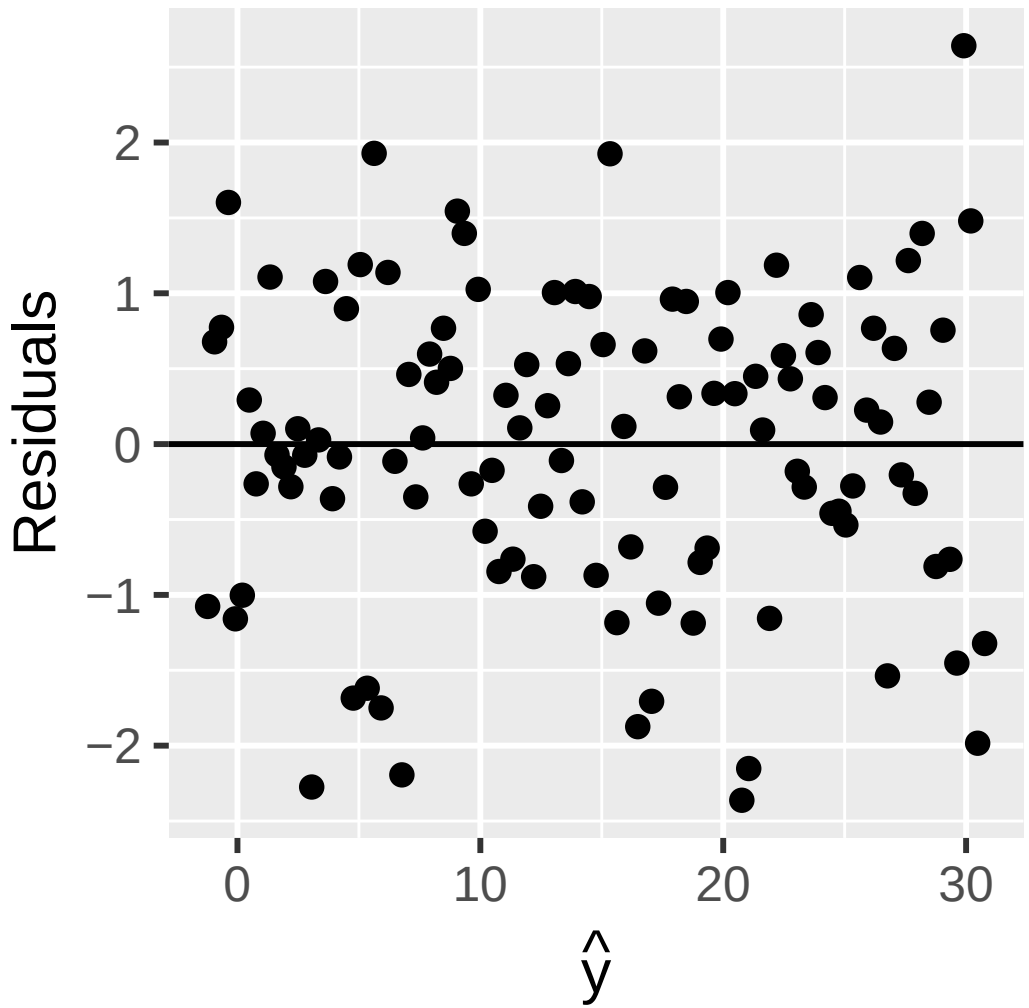
Constant Variation, residuals on fitted values

Bad. The residuals have non-constant variation along the fitted values.



Linearity and Constant Variation, residuals on fitted values

Good.



- residuals show no clear pattern as a function of $\hat{y}$, and
- all residuals have roughly equal (vertical) width.

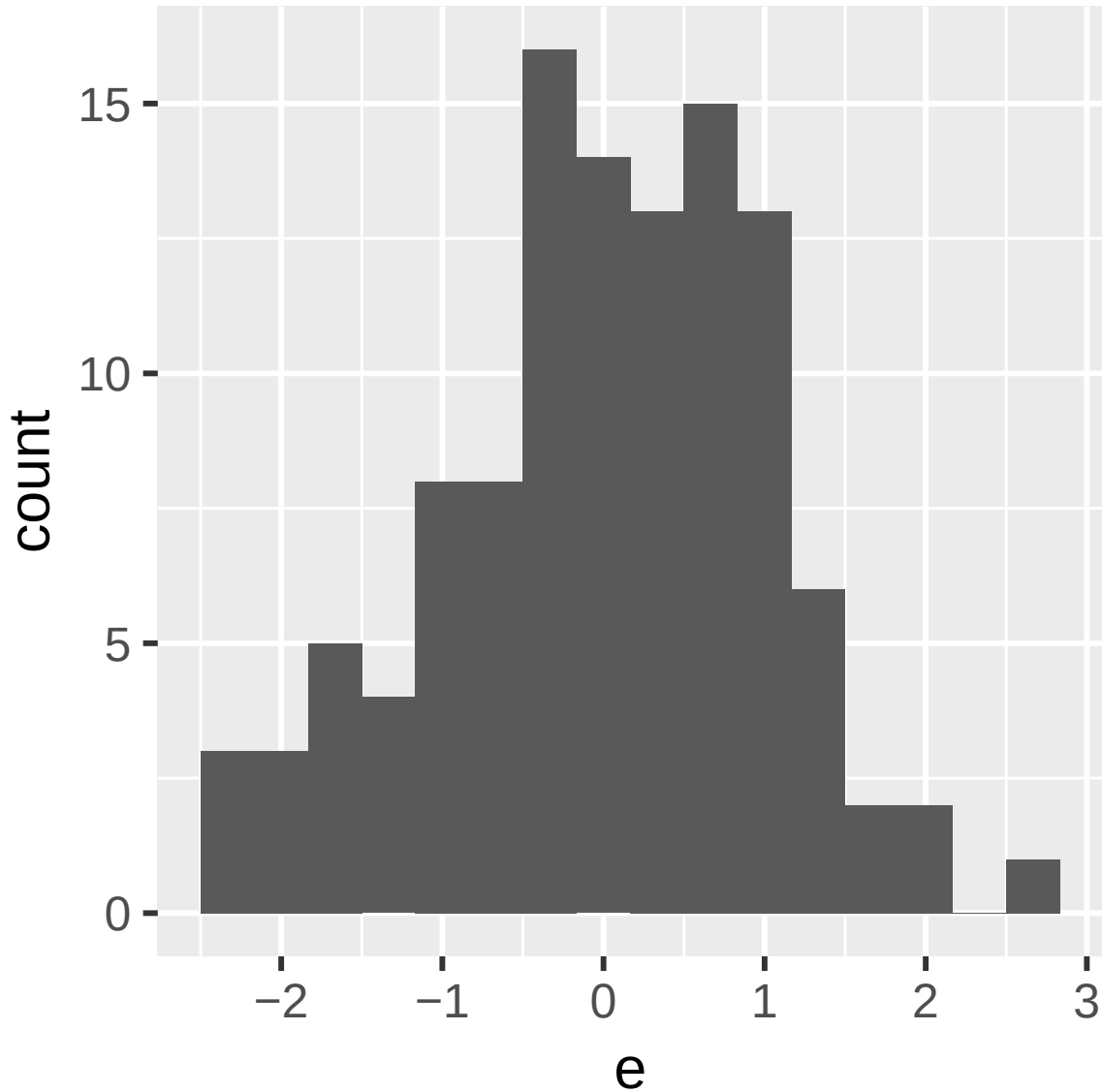
Residuals on Fitted Values

To make plots of the residuals on the fitted values,

```
## fake code
model <- lm(y~x, data=data) # first fit a model
df <- data.frame(e=rstandard(model), # vector of standardi
                 yhat=fitted(model)) # vector of fitted va
ggplot(df, aes(yhat, e)) +
    geom_hline(aes(yintercept=0))
```

Histogram of Residuals

To check the normality of the residuals, make a histogram of the residuals. Ask, do they seem normal-ish?



Histogram of Residuals

To make histograms of standardized residuals,

```
## fake code
model <- lm(y~x, data=data) # first fit a model
df <- data.frame(e=rstandard(model)) # vector of standardi
ggplot(df, aes(e)) +
    geom_histogram(binwidth=1/3)
```

Independence, residuals by time

If the information is available, you could plot residuals in the order that they were recorded, though this information is not always available.

Independence, up to you

Often times you simply need think carefully about how the data were collected.

Potential Outliers

We could define outliers as observations that have large residuals. Naturally then, the next question is, "How large is large?" We use the normality assumption to help answer this question.

Potential Outliers

Let's standardize the residuals to the standard normal distribution, $N(0, 1)$. Since the mean of the residuals will always[1] be equal to zero, we simply divide by the appropriate standard deviation

$$r_i = \frac{e_i}{\sigma_{e_i}}.$$

The `R` function `rstandard` will do this for you.

[1] This is a mathematical fact.

Potential Outliers

Any observation more than three standard deviations away from the mean could be considered an outlier. It isn't difficult to find such standardized residuals, but it is difficult to find the observations these large residuals correspond to.

```
## fake code
e <- rstandard(model) # vector of standardized residuals
## named vector of indices where expression is true
which(abs(e) > 3)
```

Potential Outliers, what to do

Outliers in linear regression are tough. Sometimes they heavily influence your least squares line. General recommendations:

- fit linear regression with and without outliers
- report qualitative and quantitative differences in the models
- if you are convinced that the outlier is in error
  - you better have good reason to justify its exclusion, state the reason
  - not liking the model with the point(s) included is not good reason

outline

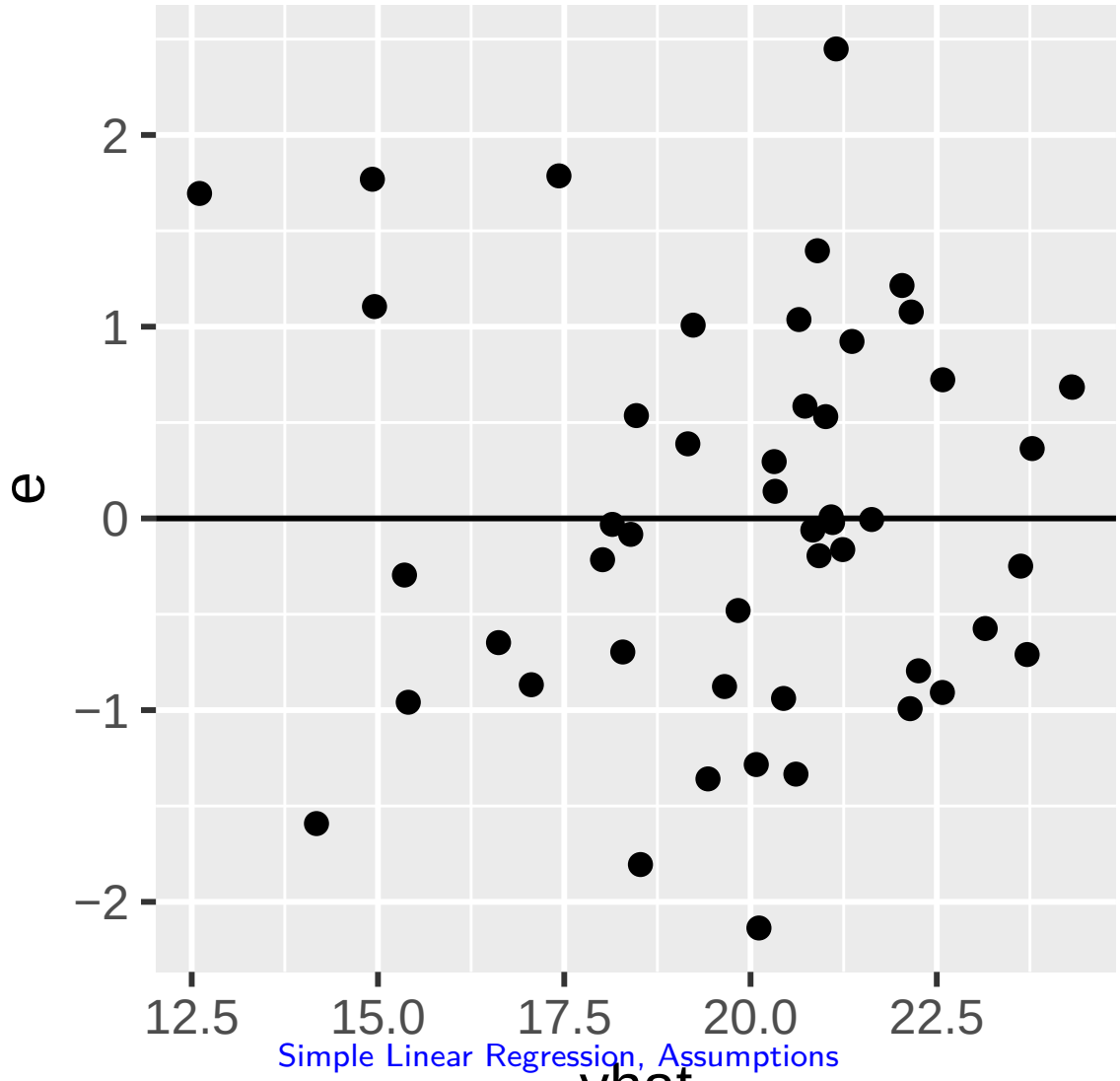
Elmhurst College Data

Let's return to the data frame `openintro::elmhurst`. First use linear regression to predict `gift_aid` with `family_income`, then calculate the data we need.

```
suppressMessages(library(ggplot2))
url <- "https://roualdes.us/data/elmhurst.csv"
elmhurst <- read.csv(url)
fit <- lm(gift_aid ~ family_income, data=elmhurst)
e <- rstandard(fit)
yhat <- fitted(fit)
df <- data.frame(e=e, yhat=yhat)
```

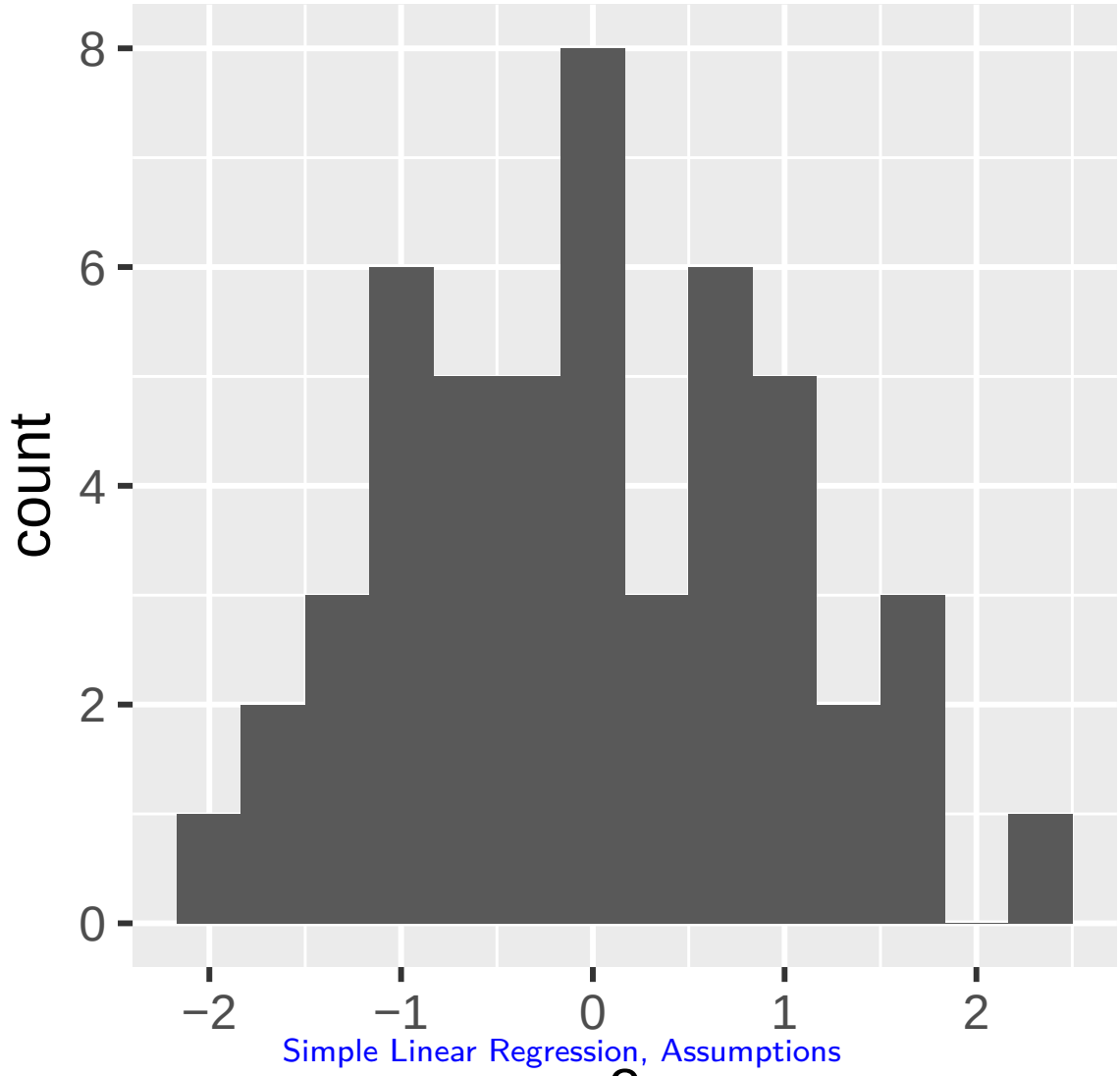
Residuals on Fitted Values

```
ggplot(df, aes(yhat, e)) +
    geom_point() + geom_hline(aes(yintercept=0))
```



Histogram of Residuals

```
ggplot(df, aes(e)) +
    geom_histogram(binwidth=1/3)
```

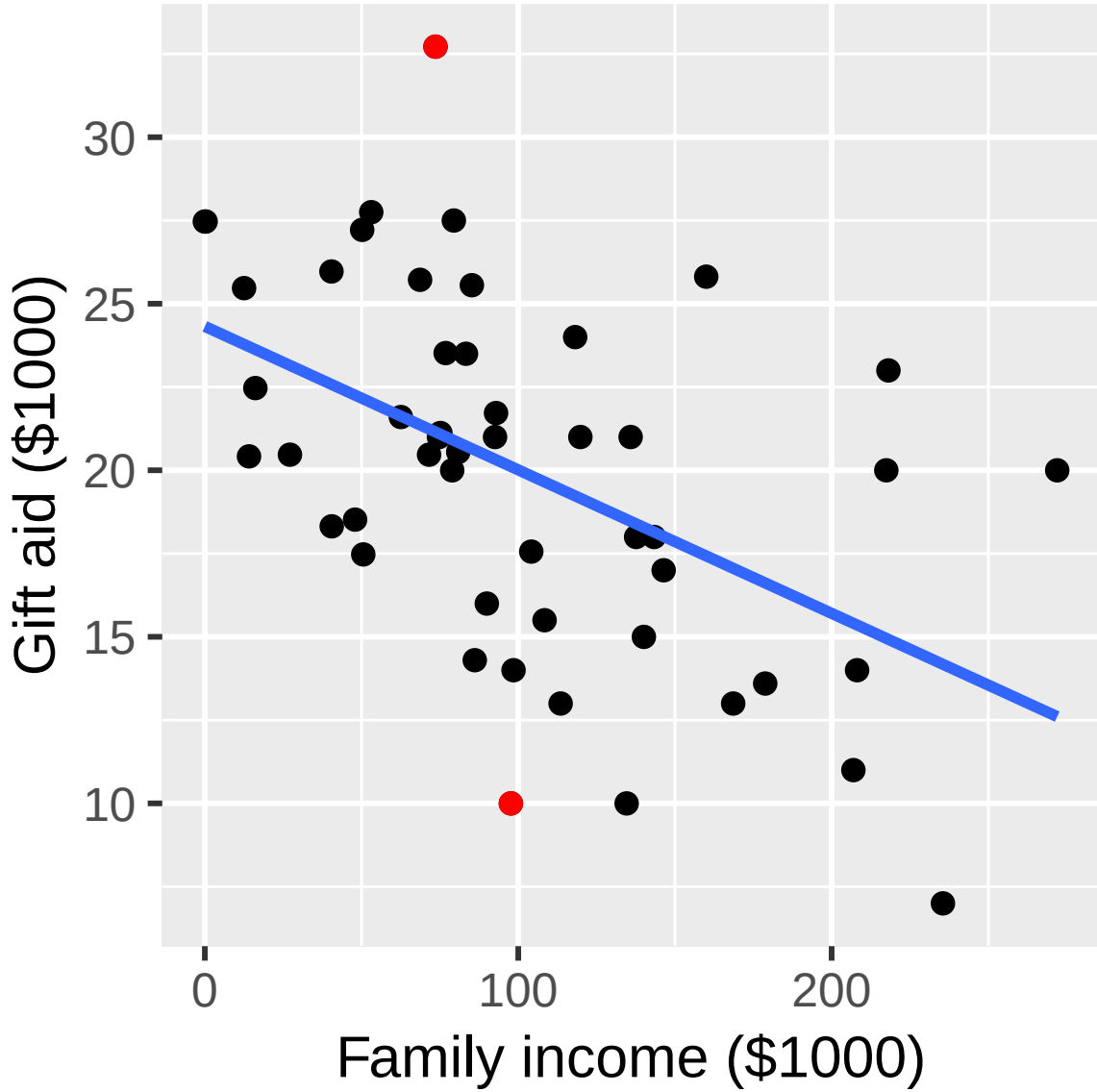


Identifying Outliers

Any outliers?

```
(idx <- which(abs(e) > 3))

## named integer(0)

(jdx <- which(abs(e) > 2))

## 16 34
## 16 34

(xout <- elmhurst[jdx, "family_income"])

## [1] 73.598 97.664

(yout <- elmhurst[jdx, "gift_aid"])

## [1] 32.72 10.00
```

Identifying Outliers



outline

Take Away

Checking model assumptions is not easy. Plots, yet again, help.

- Use standardized residuals.
- Common plots to help check assumptions
  - Residuals on fitted – scatter plot
  - histogram of residuals
- Outliers are tough.
  - You better have good, explicitly stated reason to report only the data set and subsequent model with them removed.

outline

references I

David M Diez, Christopher D Barr, and Mine Cetinkaya-Rundel. *OpenIntro Statistics*. CreateSpace independent publishing platform, third edition, 2015.

Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer Science and Business Media, 2009.