Classes 6 & 7: Linear regression

Alison Presmanes Hill

What is linear regression?

- A method to model the relationship between:
  - An outcome/dependent/predicted variable y…
  - As a function of a covariate/independent/predictor variable x

- The variability in y has 2 components:
  - Systematic variation (what we can model as a function of predictors)
  - Random variation (what is left over from the model)

Data are paired observations of X and Y

$$(x_1, y_1), \ldots, (x_n, y_n)$$

The formula

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Deconstructing the formula

$$Y = \overbrace{\beta_0}^{intercept} + \overbrace{\beta_1}^{slope} x + \overbrace{\varepsilon}^{residuals}$$

observed y = predicted + residuals

Actual observed Y_i's

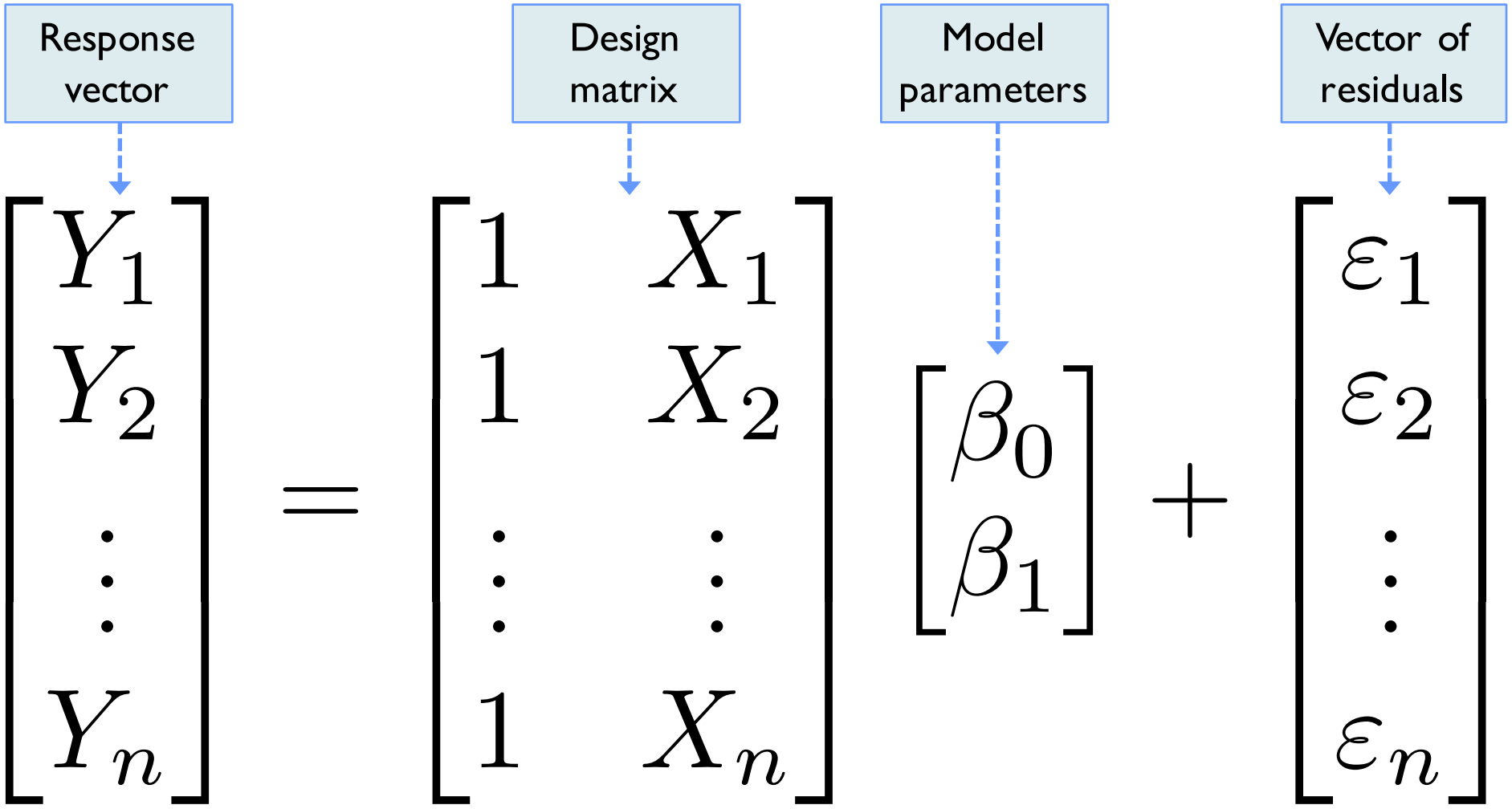
$$Y_i = \underbrace{\beta_0 + \beta_1 x_i} + \underbrace{\varepsilon_i}$$

$$= \underbrace{\hat{Y}_i} + \underbrace{\varepsilon_i}$$

Predicted or fitted Y_i's

residuals

The formula in matrix notation



For > 1 predictors, note that number of columns in design matrix must equal number of rows of model parameters.

Least squares criterion

$$g(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum (Y_i - \beta_0 - \beta_1 x_i)^2$$

Betas must minimize this sum

Linear models in R

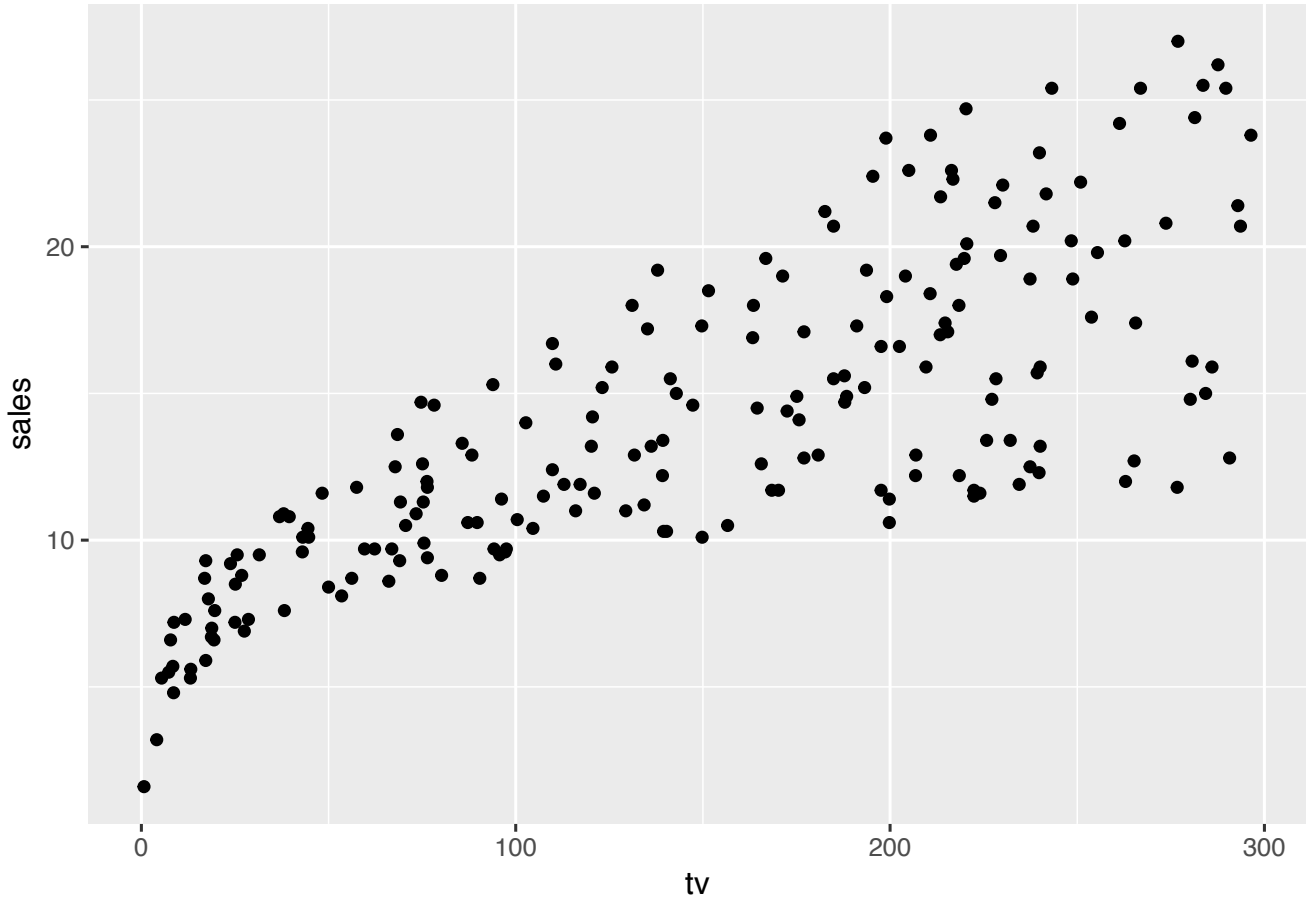
$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

```
lm(y ~ x, data = dataframe)
```

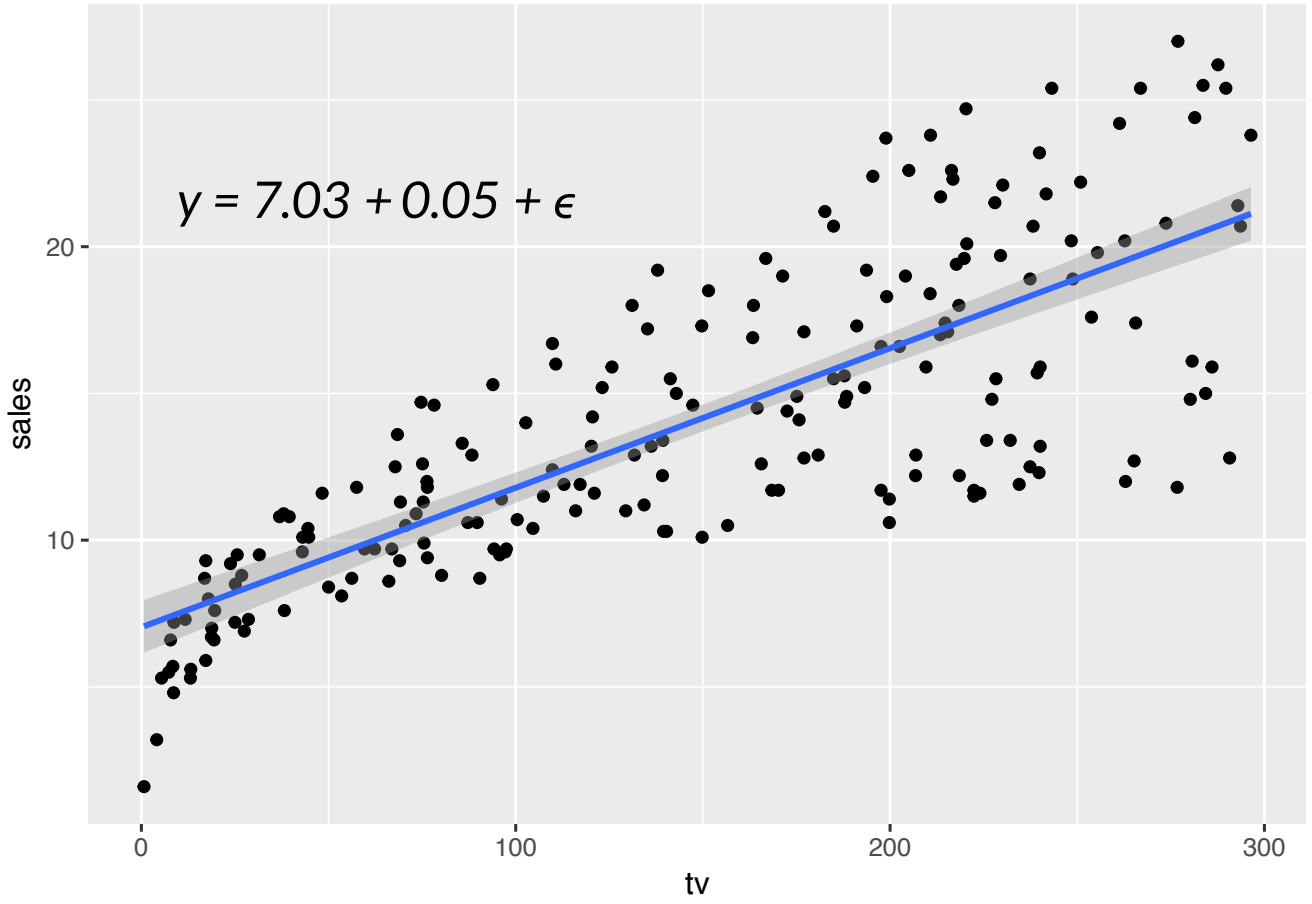
(R formulas are expressed in 'Wilkinson-Rogers' notation)

Predicting sales from TV ad budgets



Predicting sales from TV ad budgets

$$Y = \beta_0 + \beta_1 x + \varepsilon$$



Assumptions

- Importantly, linear regression does not assume anything about the distributions of *X* and *Y*
- All assumptions center around the behavior of the residuals:

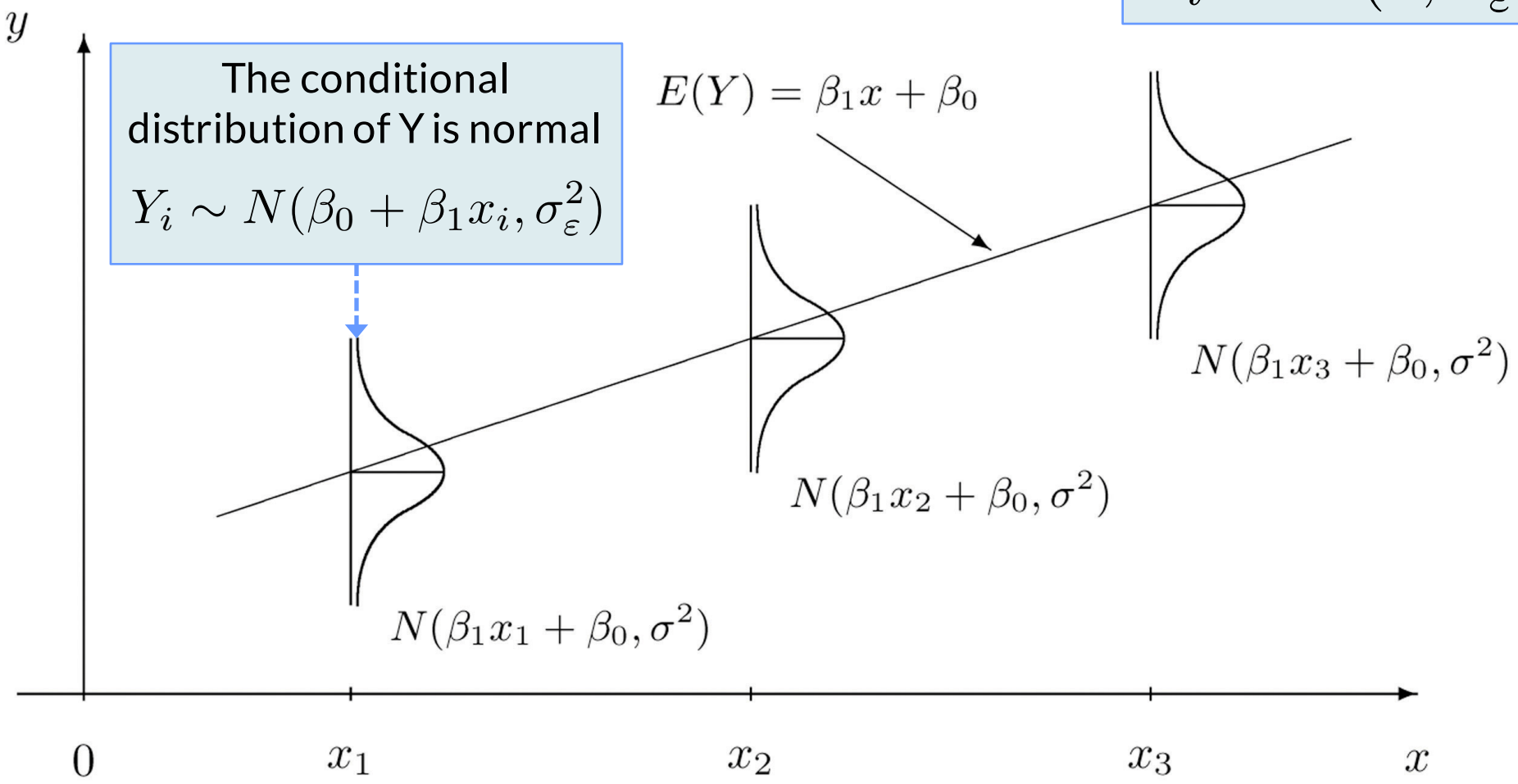
$$E(\varepsilon_i) = E(\varepsilon|x_i) = 0$$

$$Var(\varepsilon_i) = Var(\varepsilon|x_i) = Var(Y|x_i) = \sigma_\varepsilon^2$$

Assumptions: Normality

Equivalent statement: Residuals are normally distributed

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$



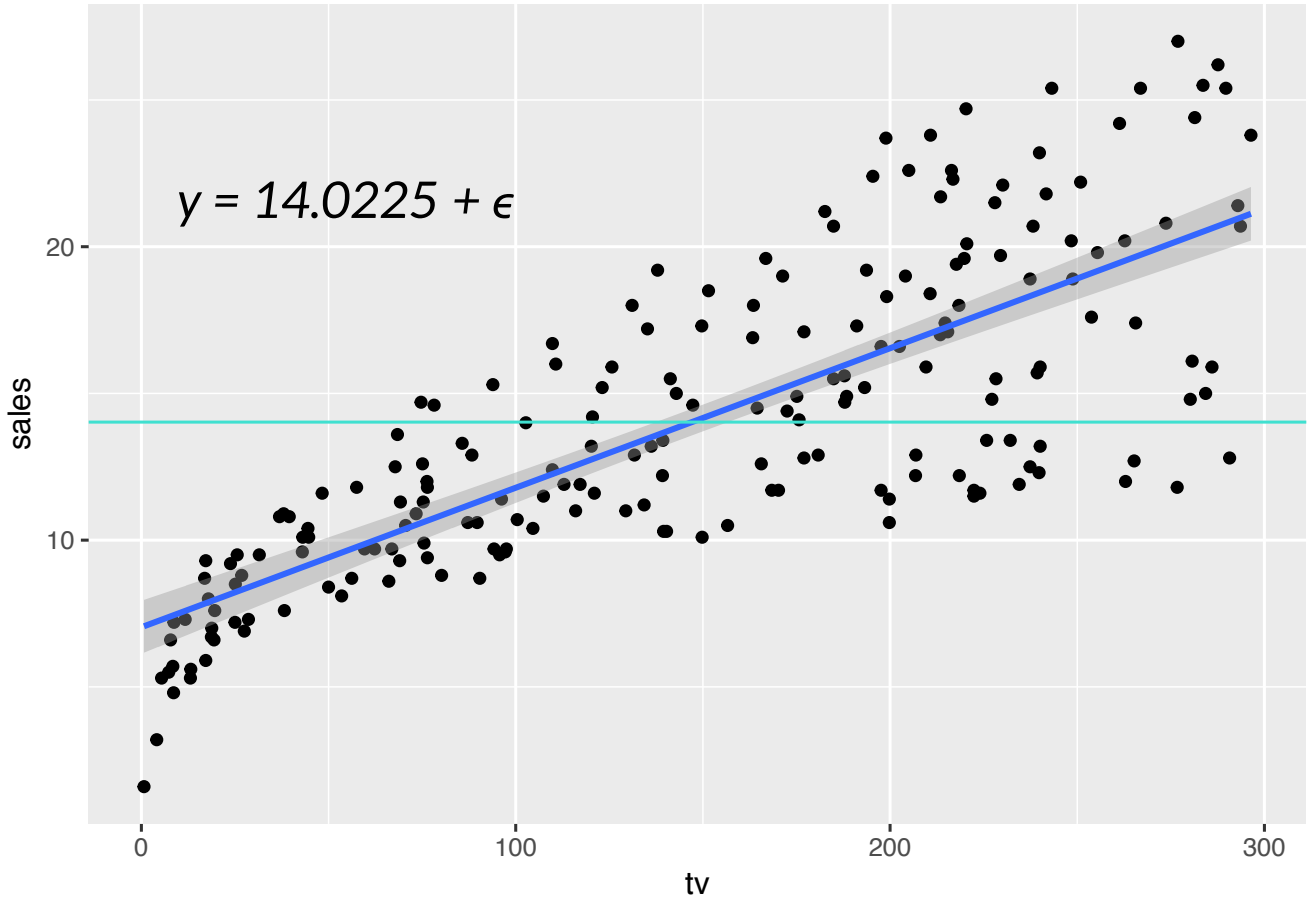
http://statwiki.ucdavis.edu/Textbook_Maps/General_Statistics/Shafer_and_Zhang's_Introductory_Statistics/10%3A_Correlation_and_Regression/10.3_Modelling_Linear_Relationships_with_Randomness_Present

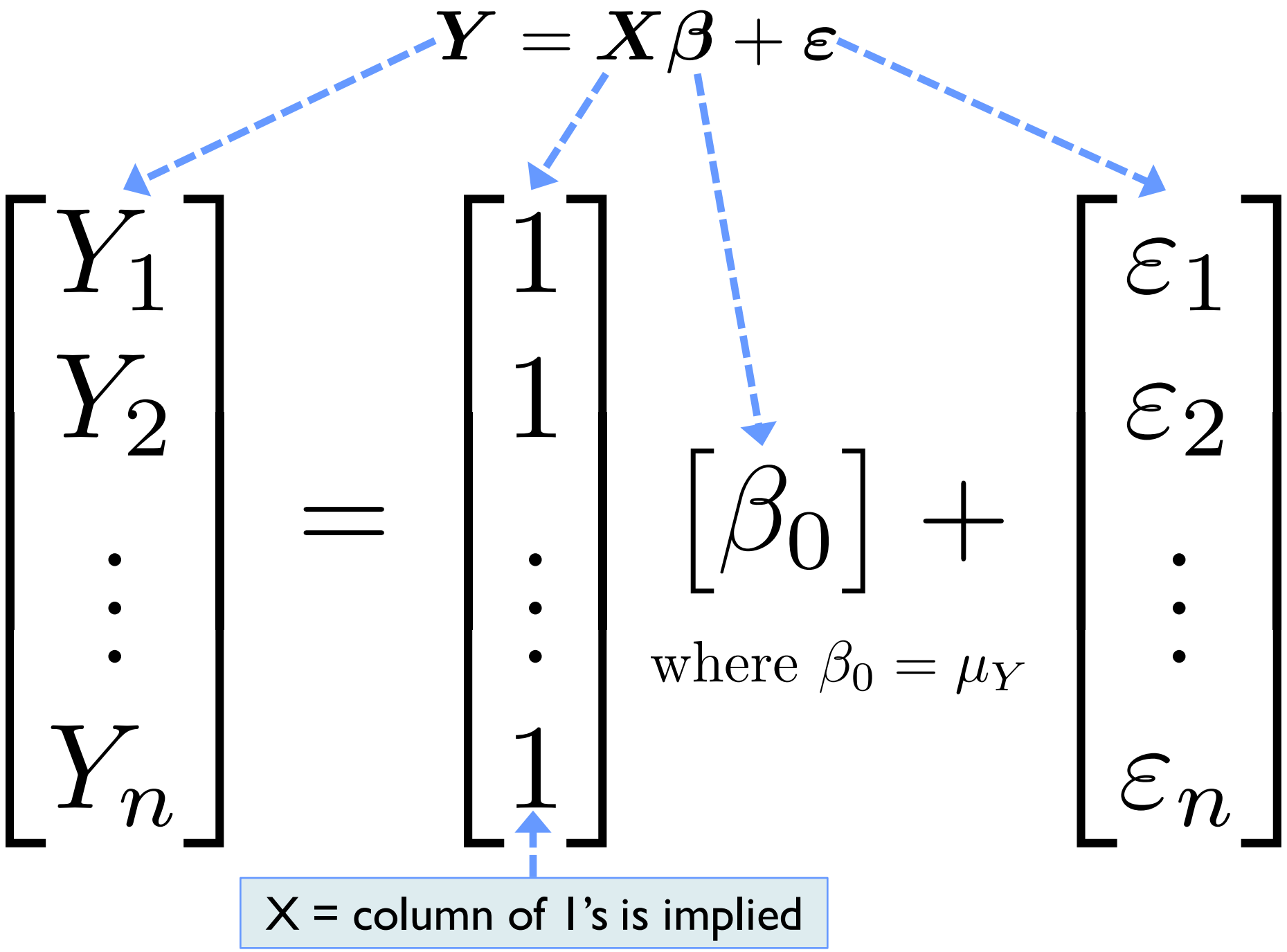
Assumptions continued

- x_i's are non-identical. That is, X has variance > 0.
- x_i's are measured without error (!!!!)
- Observations are sampled independently, so $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\varepsilon}_j$ are independent for i ≠ j

What if we left x out of the equation?

$$Y = \beta_0 + \beta_1 x + \varepsilon$$



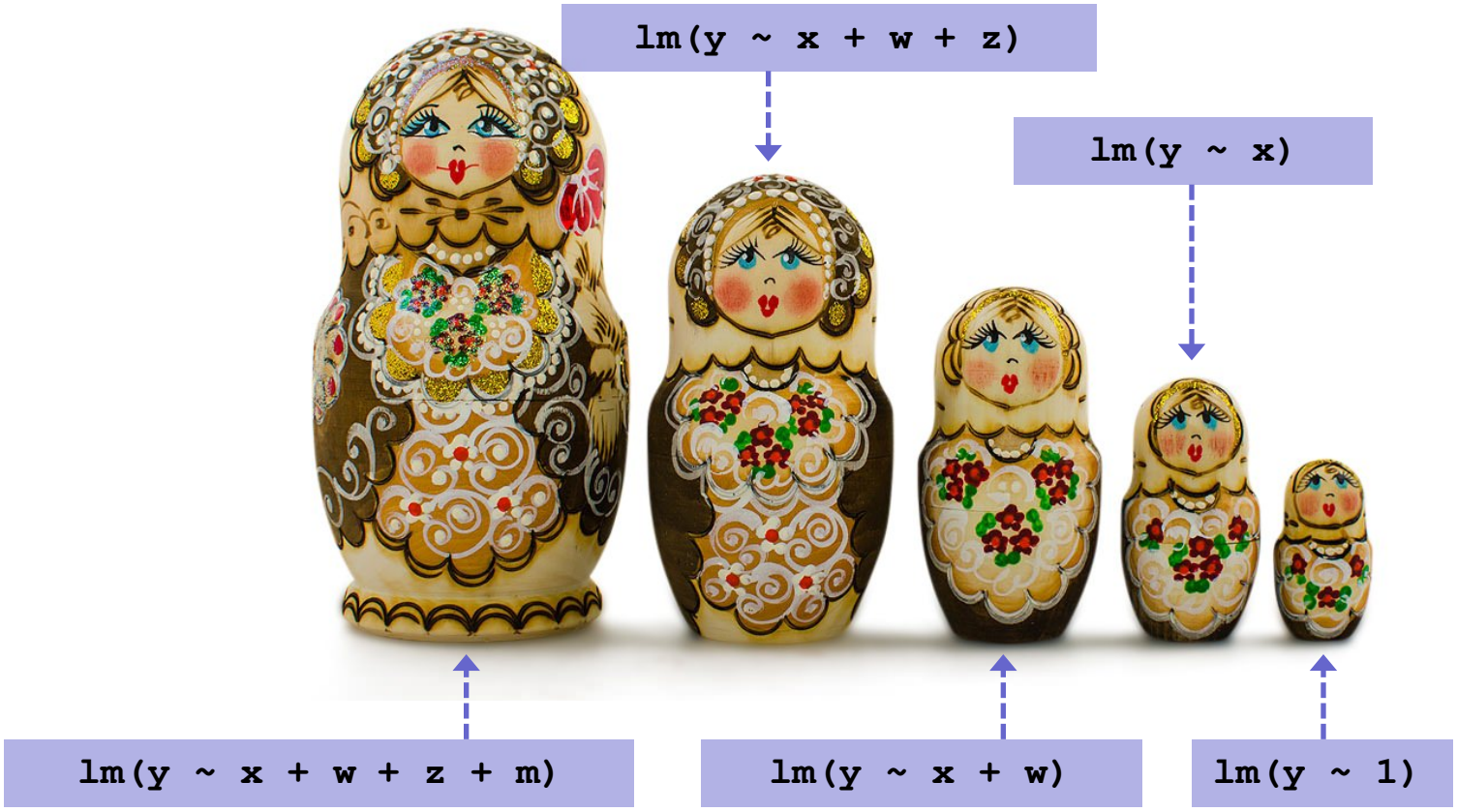

$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} [\beta_0] + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$
where $\beta_0 = \mu_Y$
X = column of 1's is implied

Sequences of nested models

- The more complex of two nested models will always fit at least as well as the less complex model.

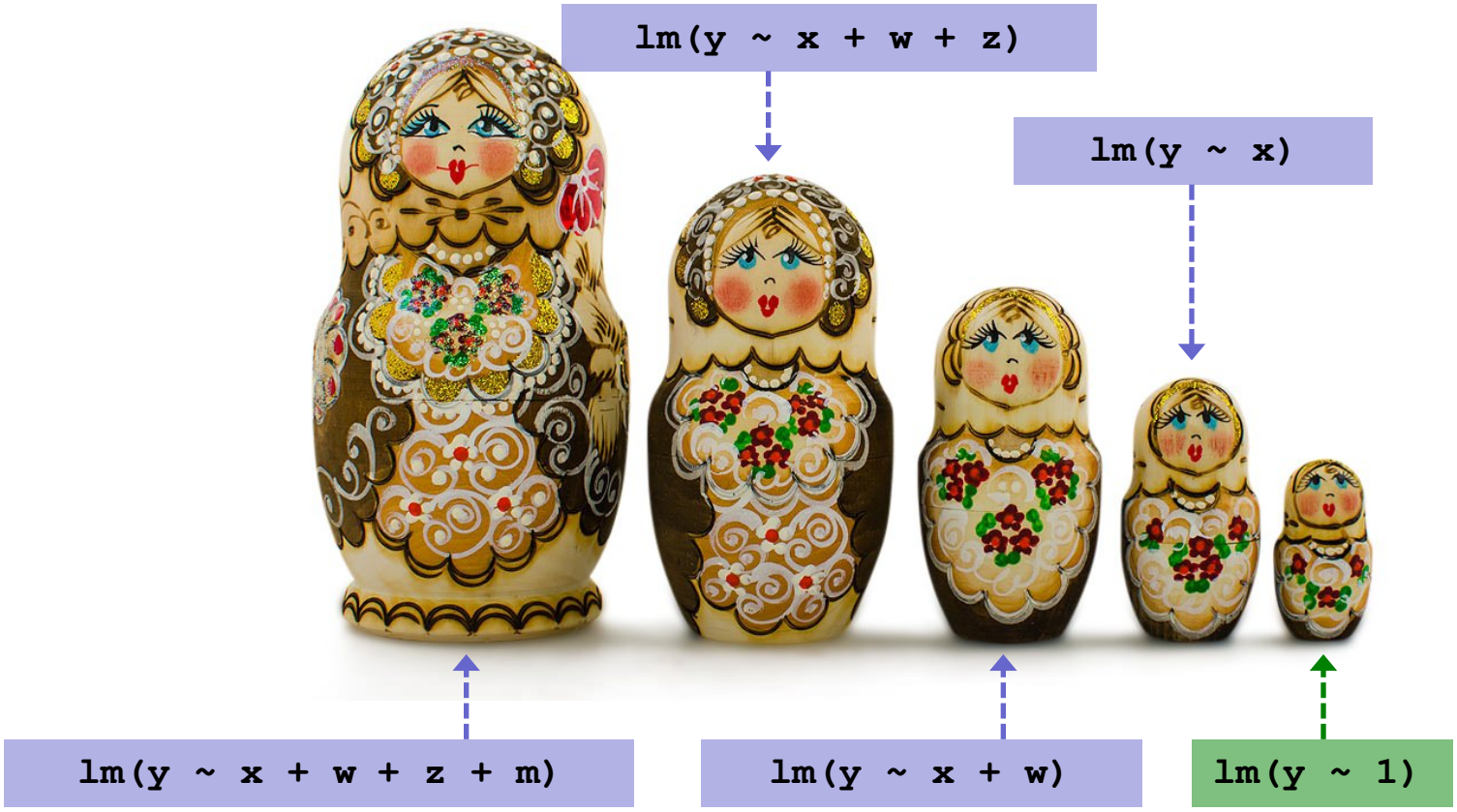
Sequences of nested models

- The "1" is implicitly included in your **`lm()`** formula



Sequences of nested models

- **`lm()`** takes ANY model you specify and compares it to the intercept-only linear model.

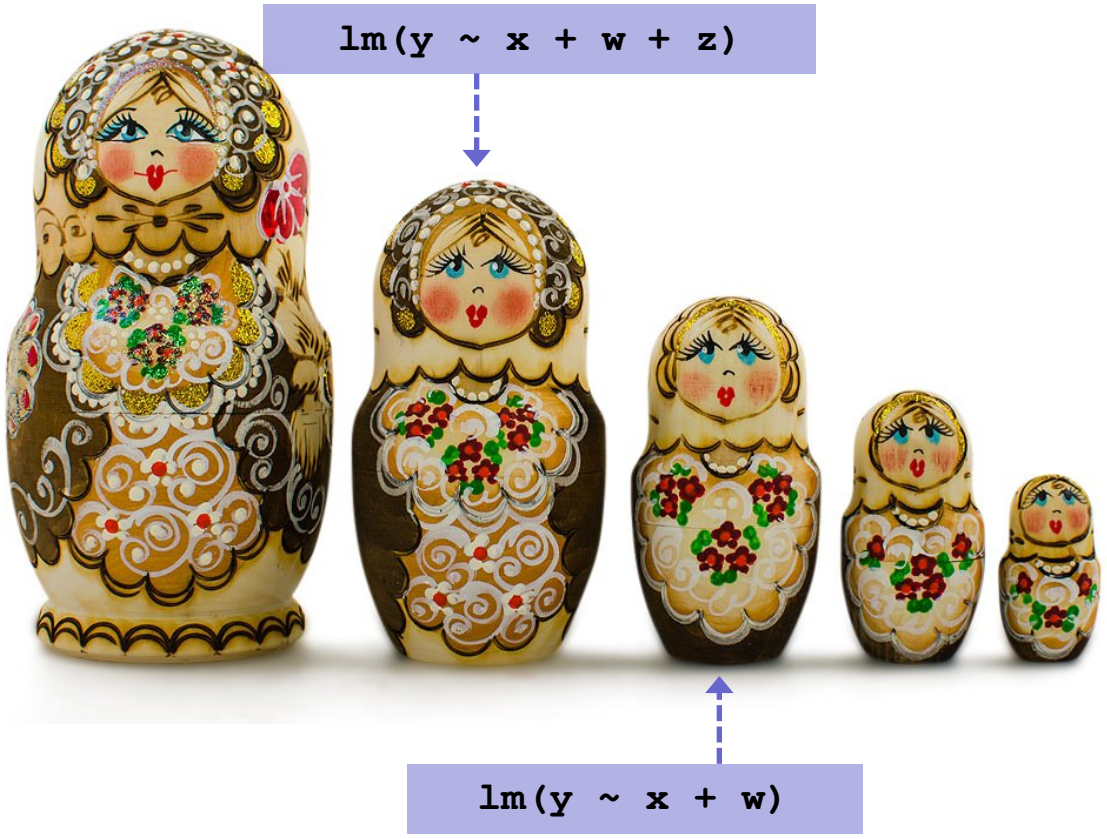


Sequences of nested models

- But you can use **`anova()`** to compare any two nested models!

Sequences of nested models

- But you can use **`anova()`** to compare any two nested models!

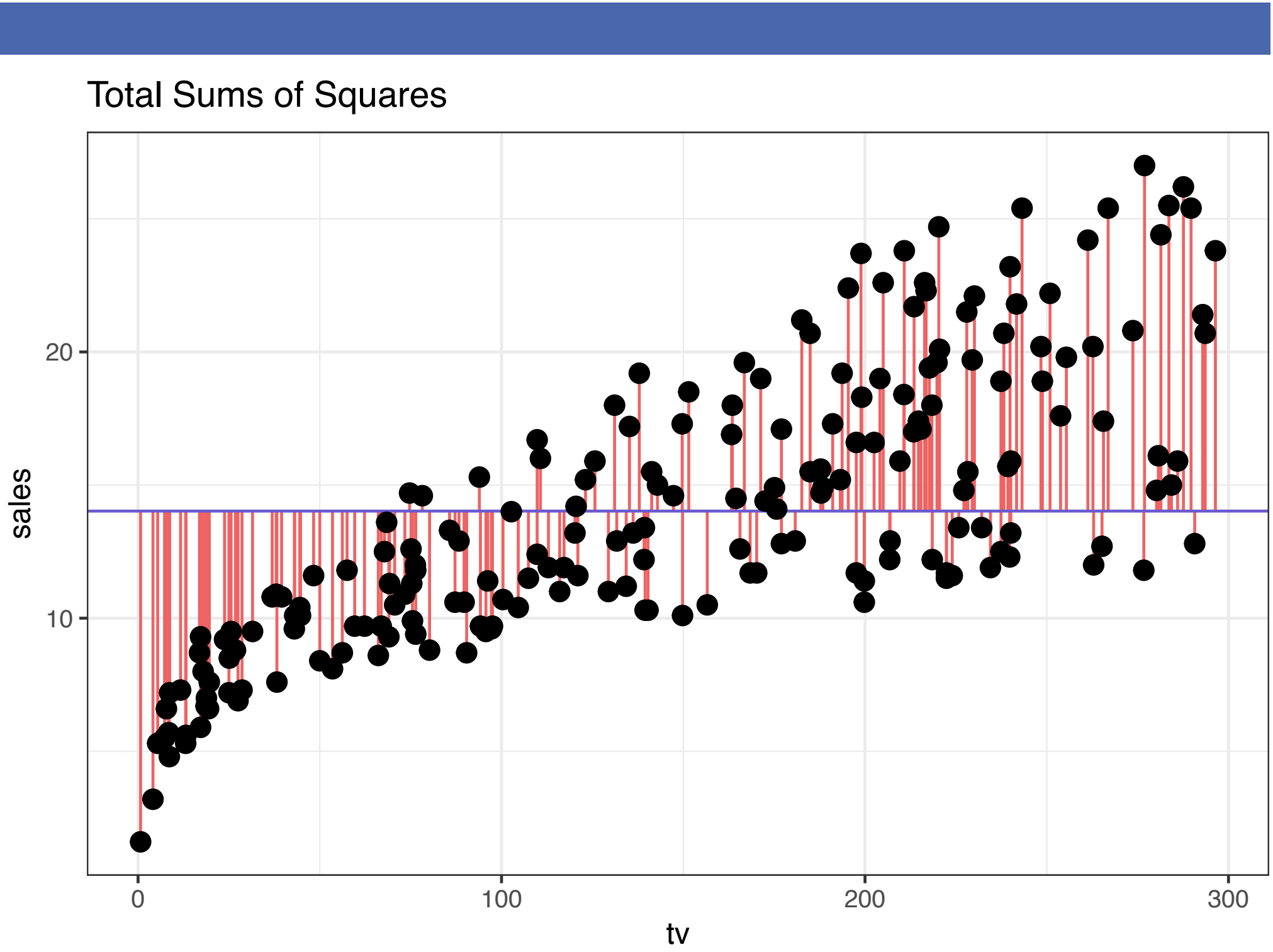


Summing up nested models

- **`lm()`** in R compares your stated model (with however many predictors you choose, can be > 1!) against the intercept-only model.

- You can directly compare any two nested linear models using the **`anova()`** command

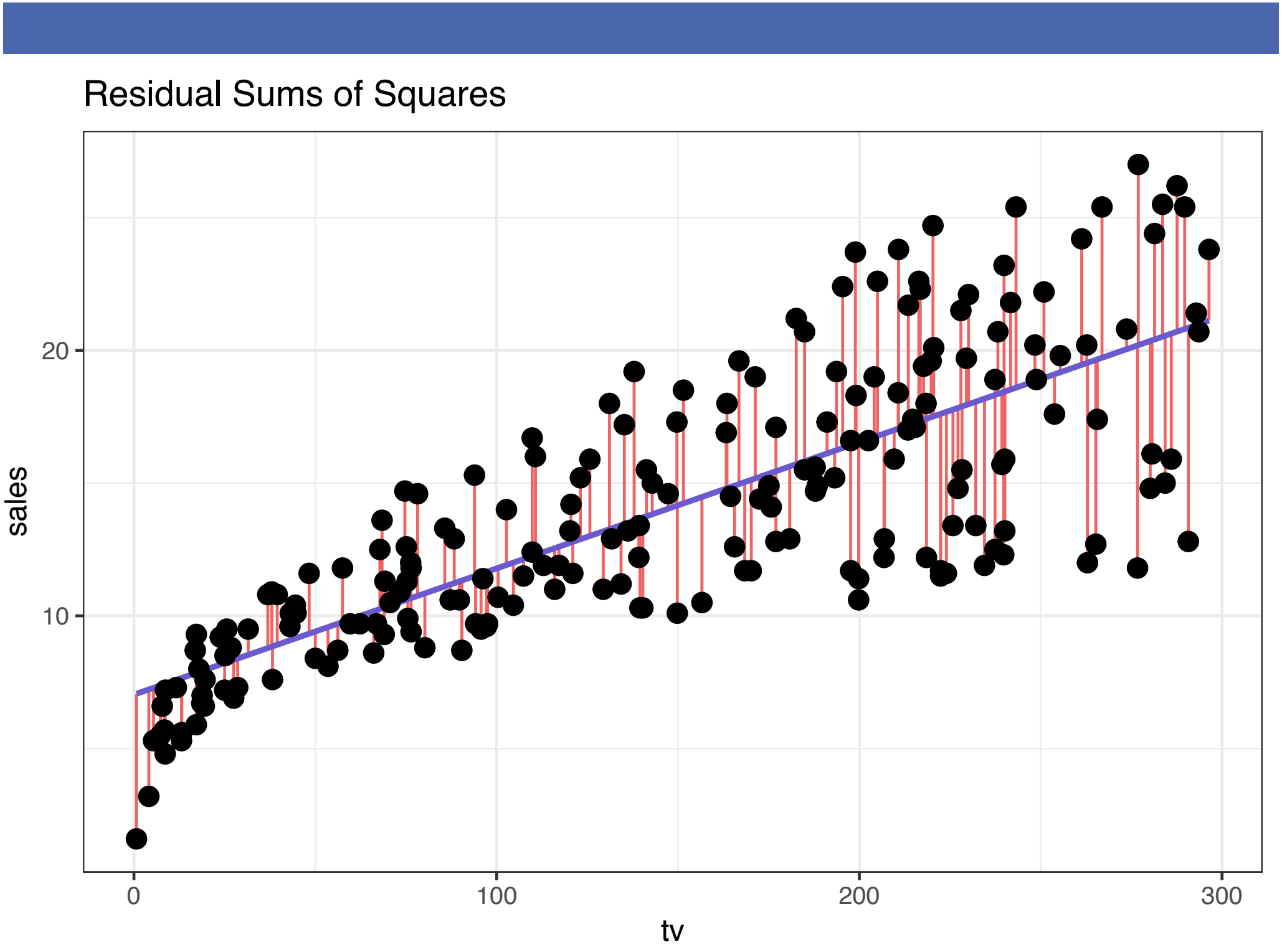
Understanding anova output


Total Sums of Squares
sales
20
10
0
100
200
300
tv

Total sums of squares (TSS) for *Y*

observed – model

$$\sum \varepsilon_i'^2 = \sum (y_i - \overline{y})^2$$


Residual Sums of Squares
sales
20
10
0
100
200
300
tv

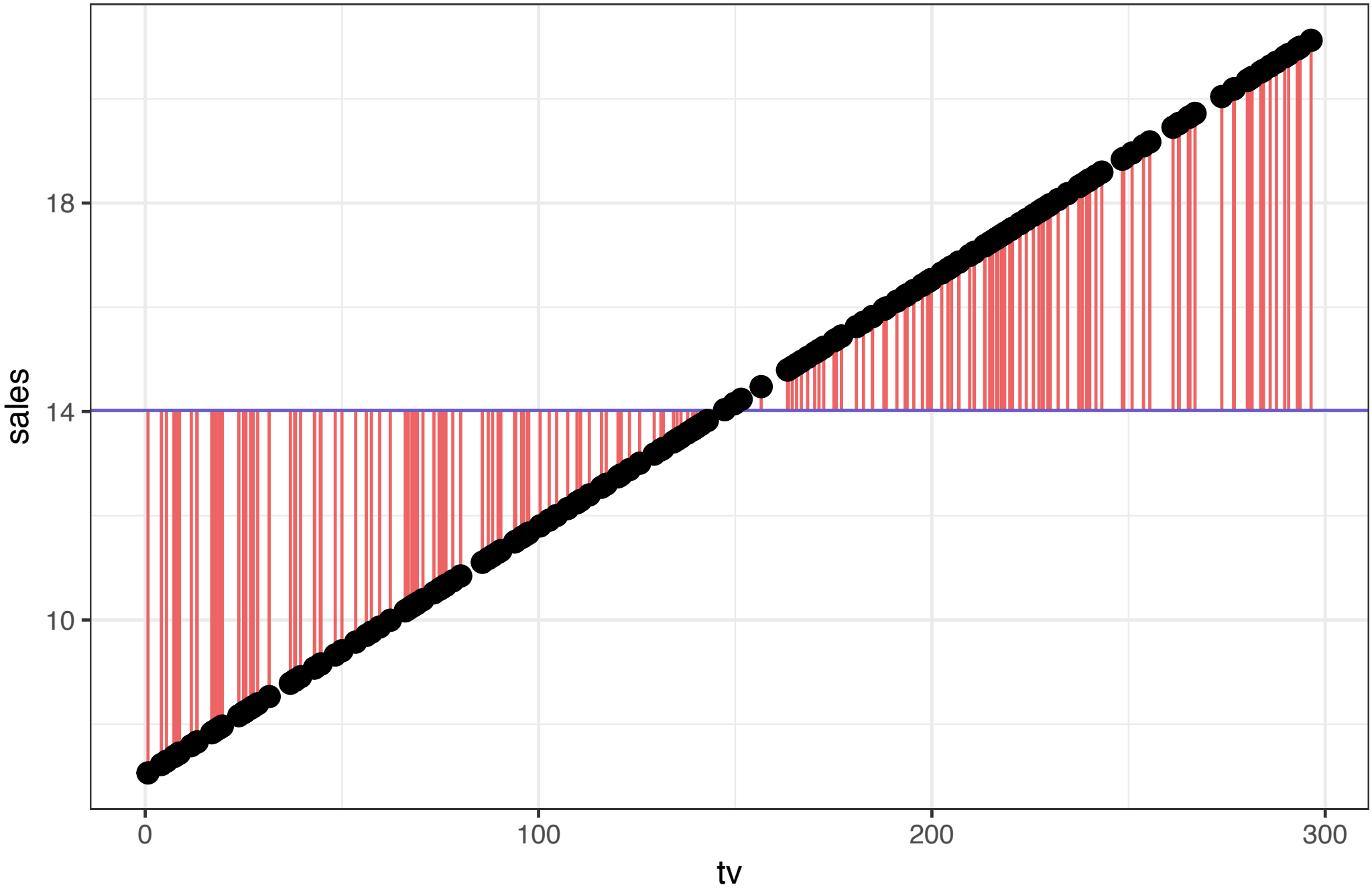
Residual sums of squares* (RSS) for Y

observed – model

$$\sum \varepsilon_i^2 = \sum (y_i - \hat{y}_i)^2$$

* This is sometimes called the sums of squares for errors (SSE)- try to mentally switch to the residual sums of squares terms, as that notation will confuse you eventually

Model Sums of Squares



Model sums of squares

observed – model

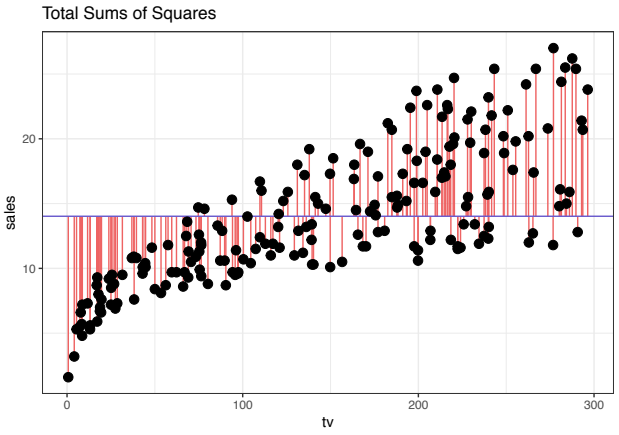
$$ModelSS = \sum (\hat{y}_i - \overline{y})^2$$

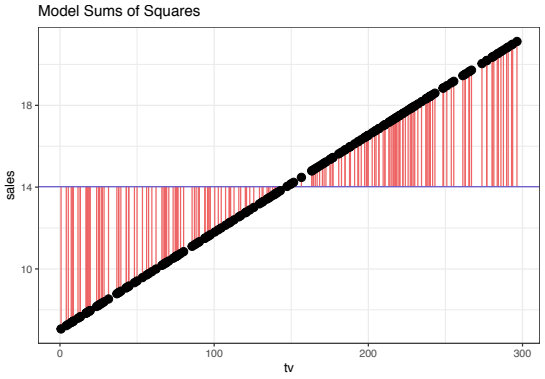
More like:
Linear regression model – mean model

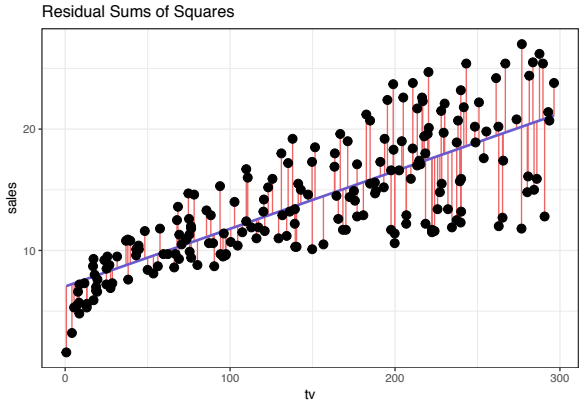
Sums of squares

Total sums of squares	Model sums of squares	Residual sums of squares
$\sum(y_i - \overline{y})^2$	$\sum(\hat{y}_i - \overline{y})^2$	$\sum(y_i - \hat{y}_i)^2$

total variation = "explained" variation + residual variation







Sums of squares

Total sums of squares	Model sums of squares	Residual sums of squares
$\sum(y_i - \overline{y})^2$	$\sum(\hat{y}_i - \overline{y})^2$	$\sum(y_i - \hat{y}_i)^2$

total variation	=	"explained" variation	+	residual variation
41.2	=	27.5	+	13.7

	set	tot_ss	res_ss	mod_ss
1	I	41.27269	13.76269	27.51000
2	II	41.27629	13.77629	27.50000
3	III	41.22620	13.75619	27.47001
4	IV	41.23249	13.74249	27.49000

Squared multiple correlation, R^2

Total sums of squares	Model sums of squares	Residual sums of squares
$\sum(y_i - \overline{y})^2$	$\sum(\hat{y}_i - \overline{y})^2$	$\sum(y_i - \hat{y}_i)^2$

total variation = "explained" variation + residual variation

$$R^2 = \frac{\text{explained variation}}{\text{total variation}}$$

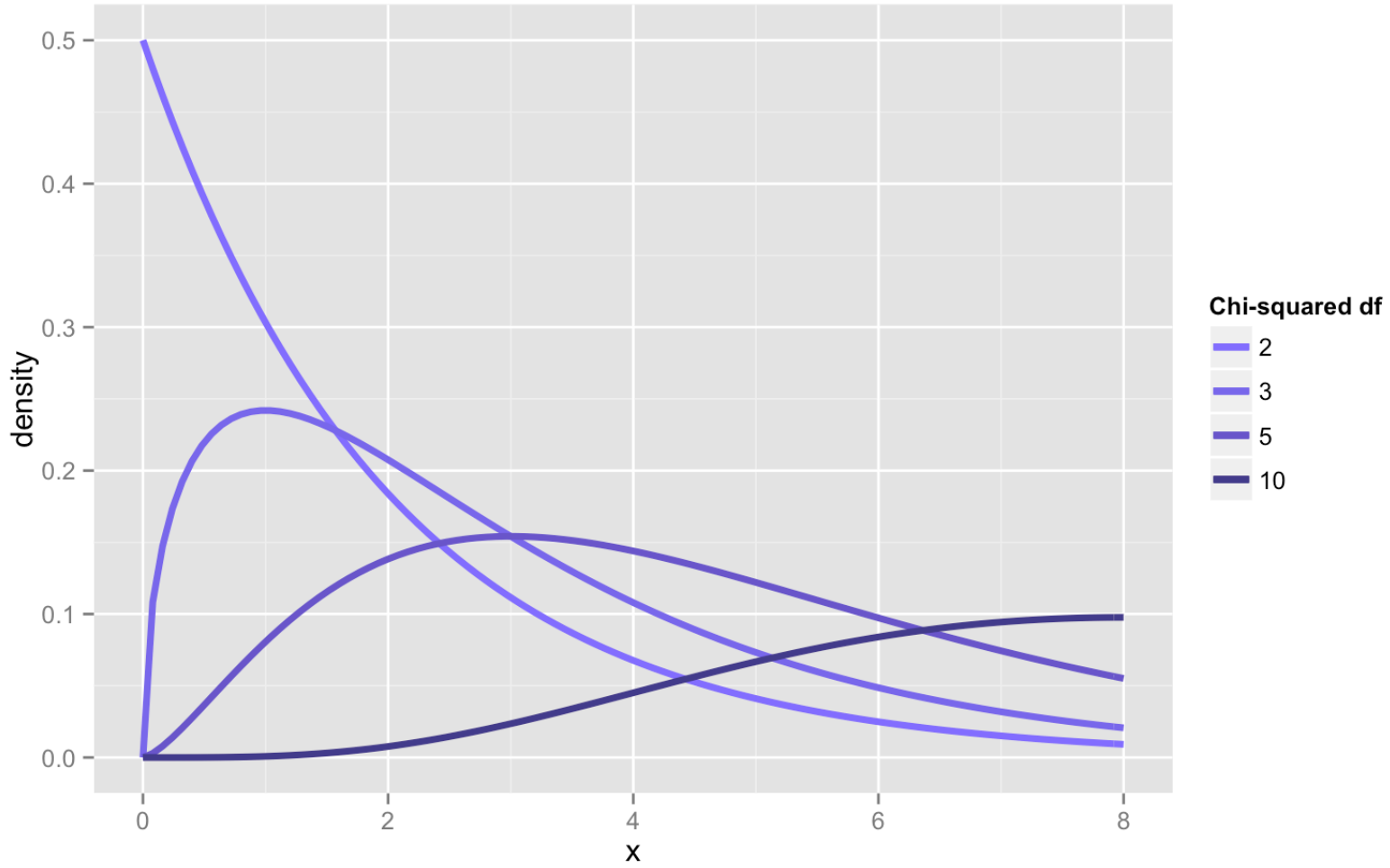
$$= \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2}$$

$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y})^2}$$

$$\frac{27.5}{41.2} = .667$$

New distribution: χ^2

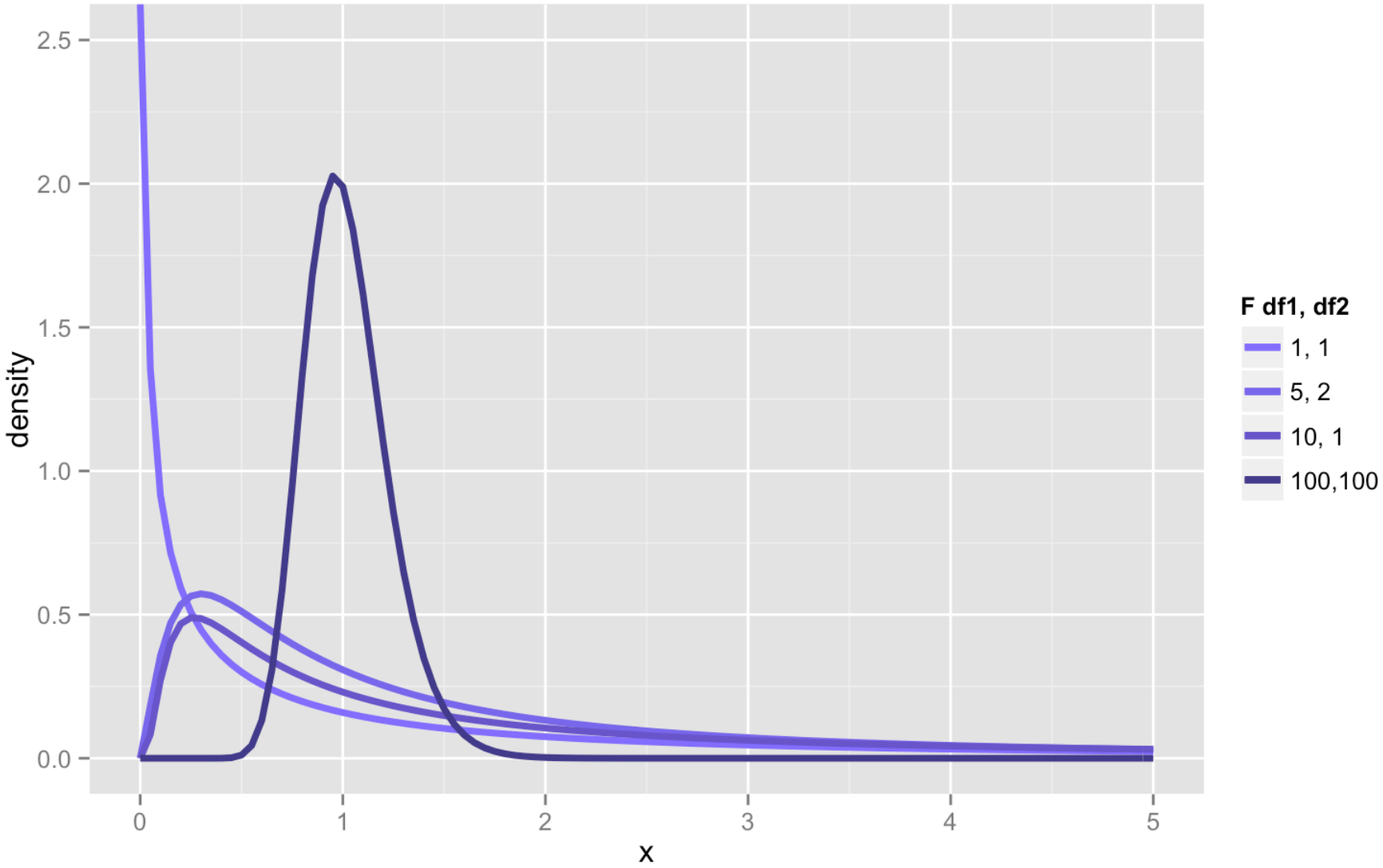
- Distribution of a **sum of the squares** of k independent standard normal random variables (k degrees of freedom)



New distribution: F

- Distribution of a **ratio of two χ^2 distributed variables** (divided by their degrees of freedom)
- Asymmetric, minimum = 0, no maximum
- Two degrees of freedom:
  - one for numerator
  - one for denominator

New distribution: F



Nested models

```
> intmod <- lm(sales ~ 1, data = ad) # model 1
> admod <- lm(sales ~ tv, data = ad) # model 2
> anova(intmod, admod)
Analysis of Variance Table
Model 1: sales ~ 1
Model 2: sales ~ tv
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    199 5417.1
2    198 2102.5  1    3314.6 312.14 < 2.2e-16
***---Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F_{df_{m1}-df_{m2},df_{m2}} = \frac{\frac{RSS_{m1}-RSS_{m2}}{p_{m2}-p_{m1}}}{\frac{RSS_{m2}}{df_{m2}}}$$

Let's return to our linear model

Because of this formula:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Running a regression model produces the following for every (x_i, y_i):

- Predicted or "fitted" values of y_i: $\hat{Y}_i$
- The residuals of y_i: ε_i

Since both of these are **statistics**, they each have:

- Expected values
- Standard errors (recall this is the standard deviation of the sampling distribution of a statistic)

Predicted values of Y

$$\hat{Y}_i = \beta_0 + \beta_1 x_i$$

Predicted or fitted Y_i's

Expectation and variance of predicted values

Expectation:

The conditional mean of *Y* is linear in *X*, with an intercept of β_0 and a slope of β_1

$$E(Y_i|X = x_i) = \beta_0 + \beta_1 x_i$$

Variance

The conditional variance of *Y* is constant with respect to *X*

$$Var(Y|X = x) = \sigma^2$$

Residual values of *Y* (reversing previous formula)

Actual observed Y_i's

$$\varepsilon_i = Y_i - \underbrace{\hat{Y}_i}$$

$$= Y_i - \underbrace{(\beta_0 + \beta_1 x_i)}$$

Predicted or fitted Y_i's

Expectation and variance of residual values

Expectation:

The values of the residuals are unrelated to X, such that if we plot the residuals vs. the x's, we see a null scatterplot with no patterns

$$E(\varepsilon_i) = E(\varepsilon|x_i) = 0$$

Variance

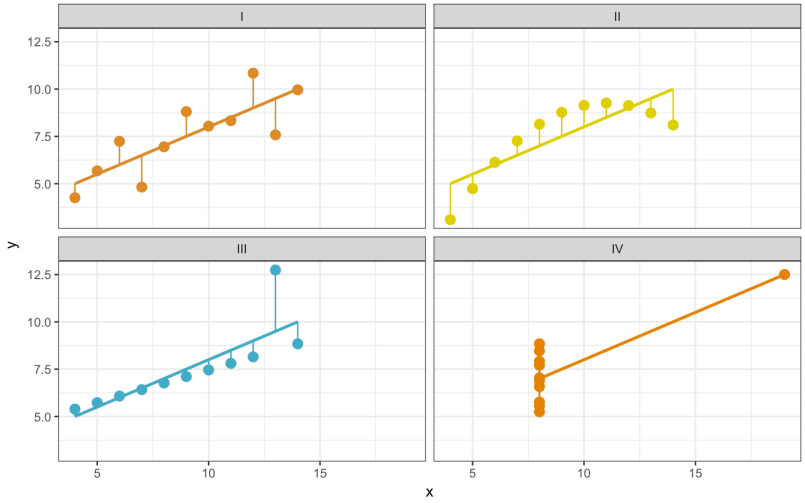
The conditional variance of the residuals is constant with respect to *X*

$$Var(\varepsilon_i) = Var(\varepsilon|x_i) = Var(Y|x_i) = \sigma_\varepsilon^2$$

Residual standard error

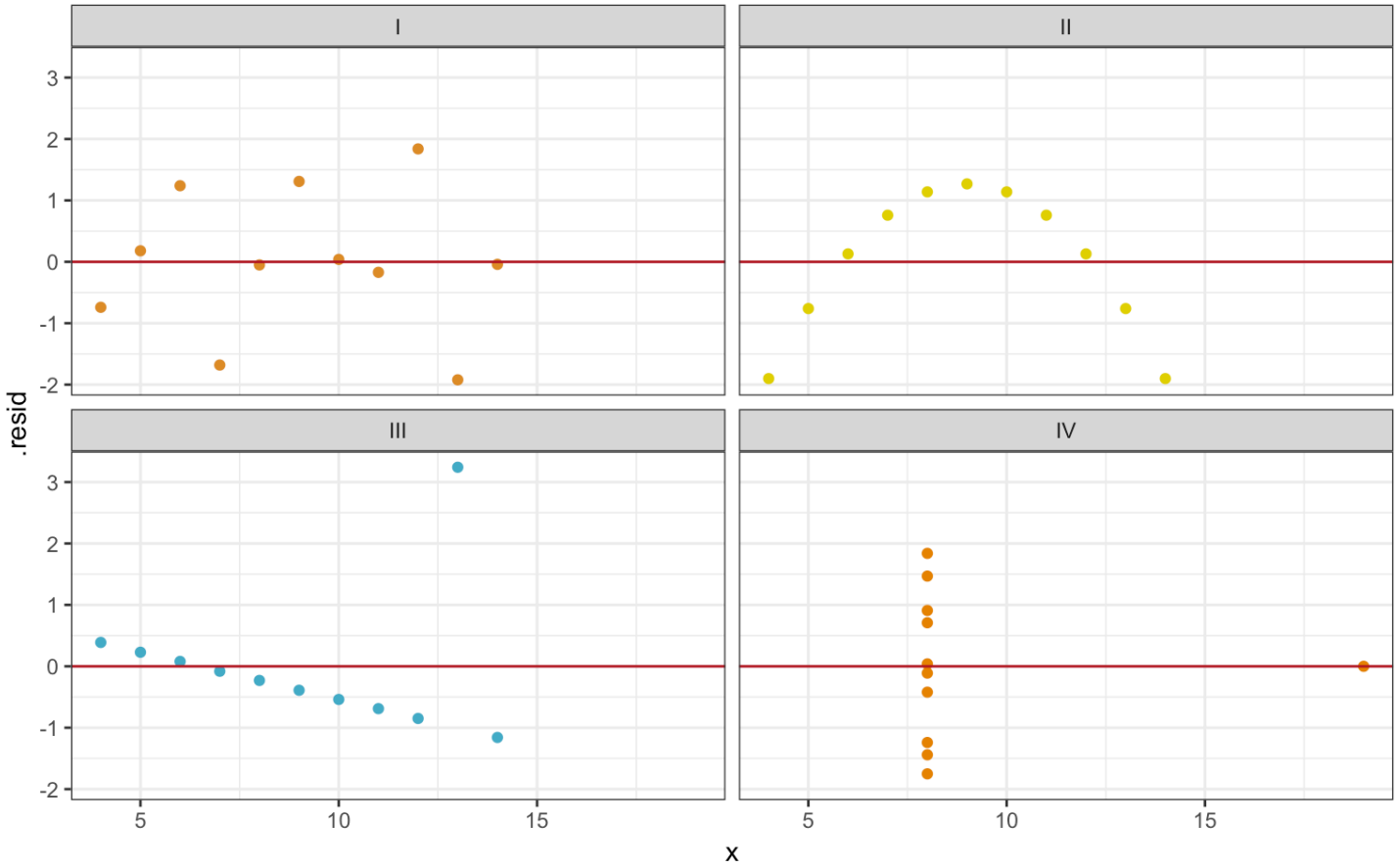
$$SE_{resid} = \sqrt{\frac{\sum E_i^2}{(n-2)}}$$

$$= \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)}}$$



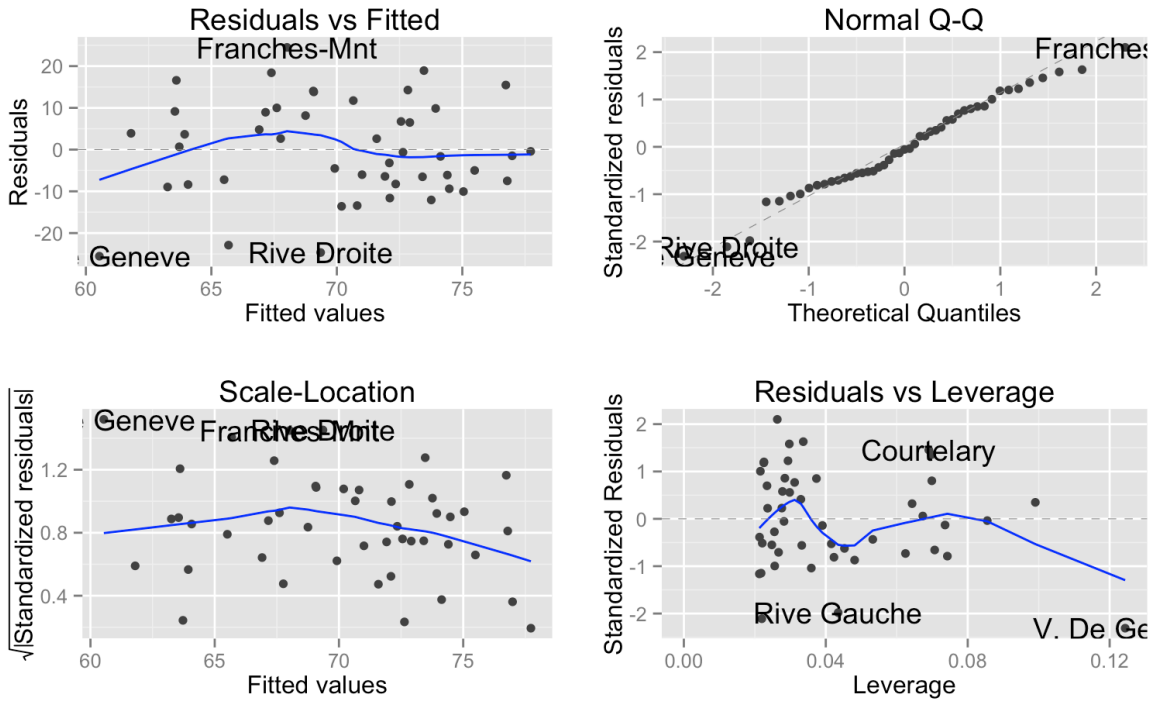
Always plot residuals against *x*!

What you want to see: no pattern at all



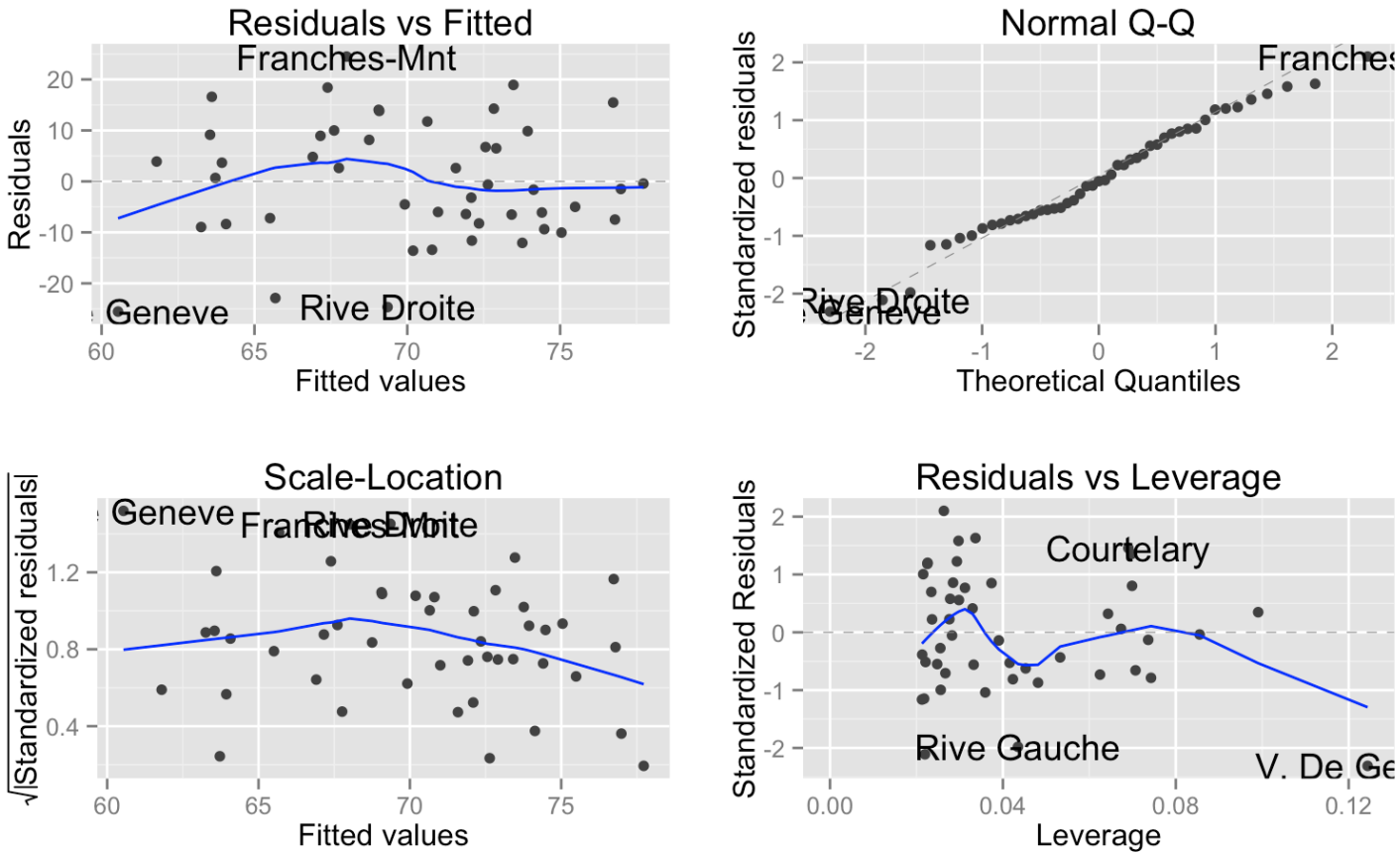
How to read these plots (clockwise)

- The plot in the upper left shows the residual errors plotted versus their fitted values. The residuals should be randomly distributed around the horizontal line representing a residual error of zero; that is, there should not be a distinct trend in the distribution of points.



How to read these plots

- The plot in the upper right is a standard Q-Q plot, which should suggest that the residual errors are normally distributed.



Brief aside: QQ plots

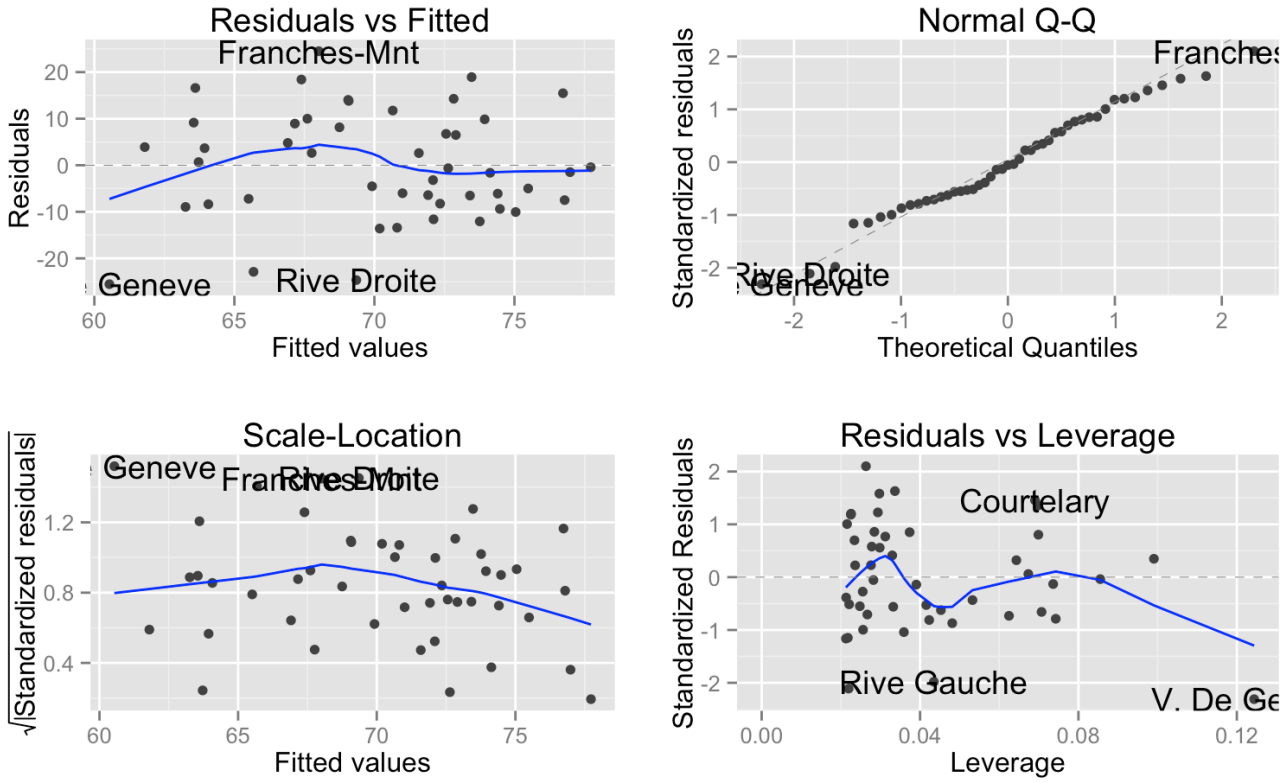
- A Q-Q plot displays quantiles of one distribution against quantiles of another. What this means is that the data are ranked and sorted.
- A normal Q-Q plot displays quantiles of the normal distribution on the *x*-axis against quantiles of the empirical (i.e., the observed) distribution on the *y*-axis.
- A straight line is typically plotted through the points corresponding to the 1st and 3rd quantiles of each variable. If the empirical data is normally distributed, all the points on the normal Q-Q plot will form a perfectly straight line.

Brief aside: QQ plots

Description of Point Pattern	Possible Interpretation
all but a few points fall on line	outliers in the data
left end sags below line; right end lifts above line	long tails at both ends of the data distribution
left end lifts above line; right end sags below line	short tails at both ends of the data distribution
curved pattern with increasing slope (L to R)	data distribution is skewed to the right
curved pattern with decreasing slope (L to R)	data distribution is skewed to the left
staircase pattern (plateaus and gaps)	data have been rounded or are discrete

How to read these plots

- The scale-location plot in the lower left shows the square root of the standardized residuals (sort of a square root of relative error) as a function of the fitted values. Again, there should be no obvious trend in this plot.



How to read these plots

- Finally, the plot in the lower right shows each points' leverage.