

(Linear) Modeling

Winter Institute in Data Science

Ryan T. Moore

2025-12-16

Table of contents I

Modeling

The Linear Model in R

Modeling

The Two¹ Steps

- ▶ Choose a generic functional form

¹For a large value of two.

The Two¹ Steps

- ▶ Choose a generic functional form
- ▶ Estimate parameters of that form with data

¹For a large value of two.

The Two¹ Steps

- ▶ Choose a generic functional form
- ▶ Estimate parameters of that form with data
- ▶ (Evaluate and improve)

¹For a large value of two.

The Canonical Linear Form

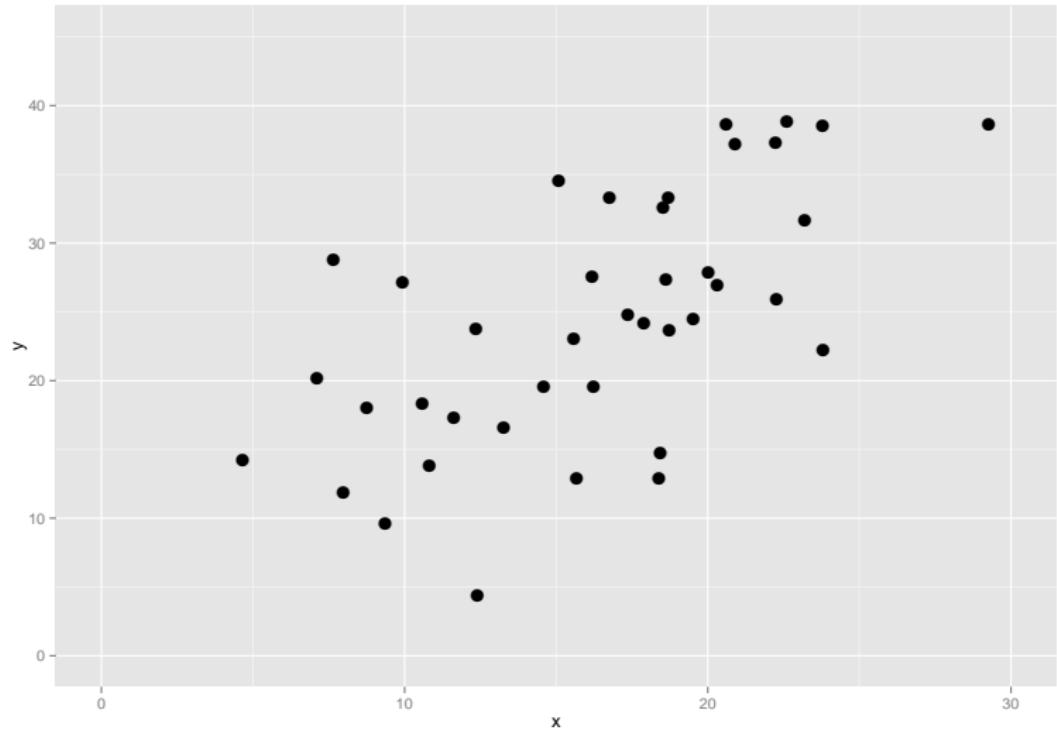
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

The Model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

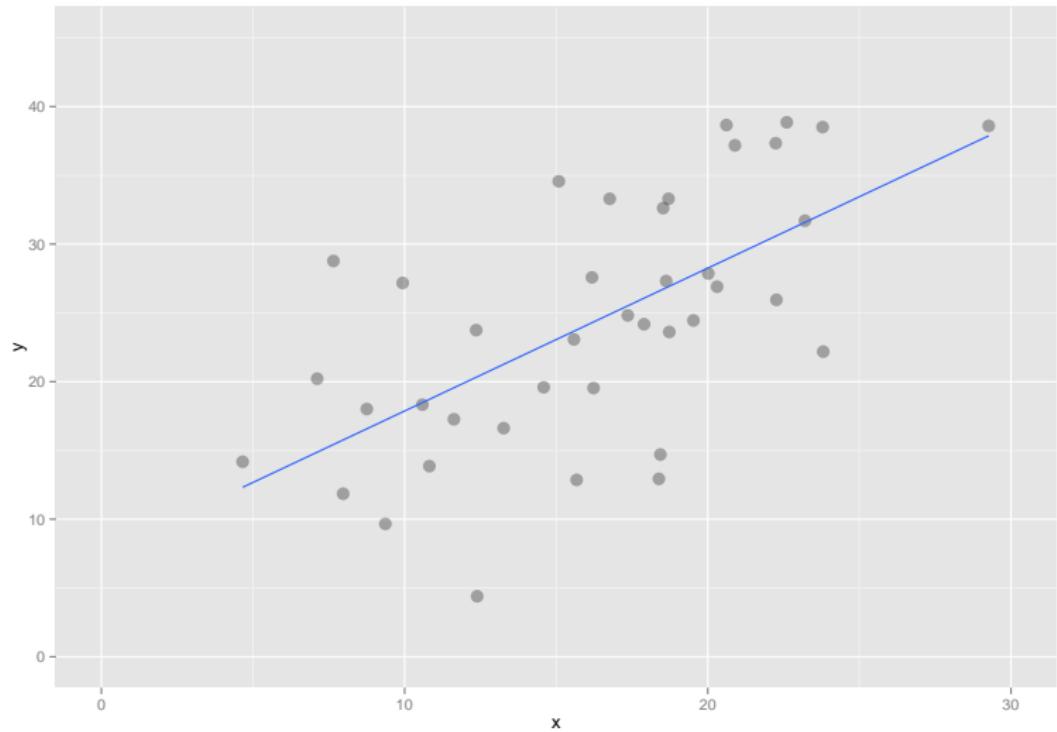
The Model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



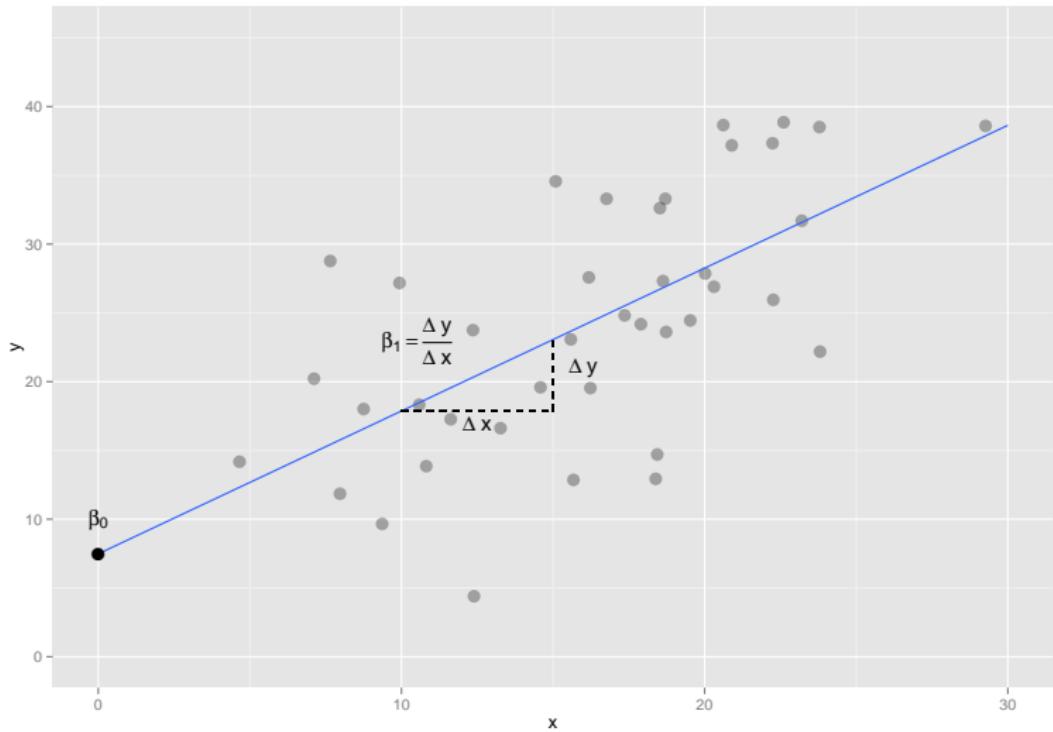
The Model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



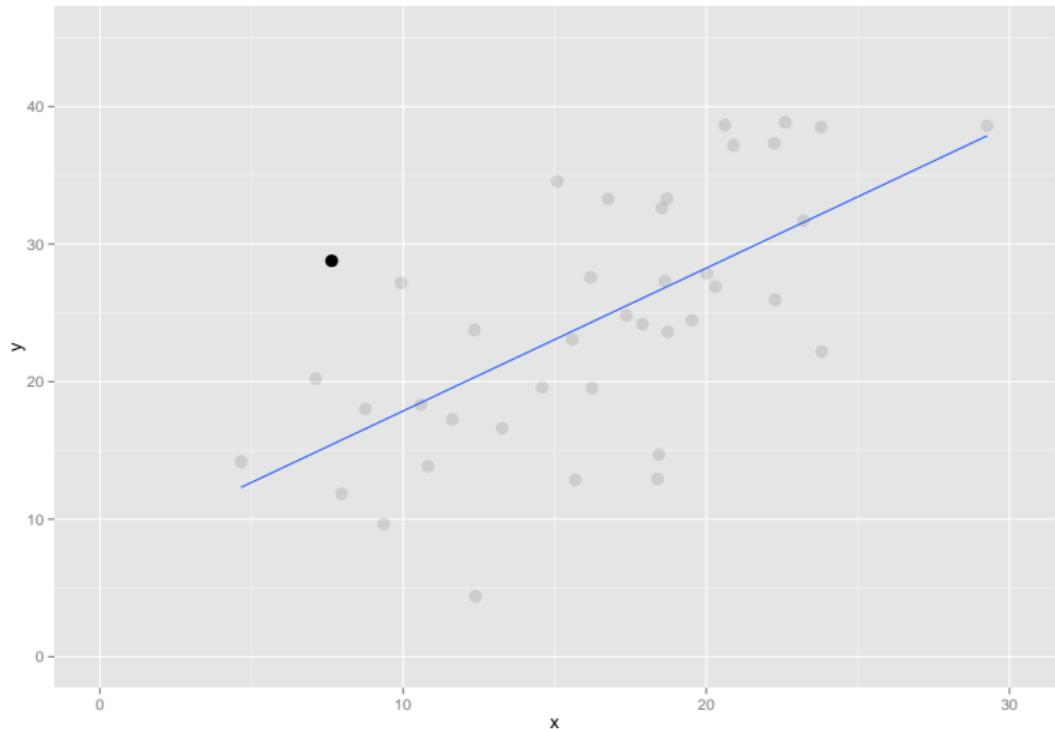
The Model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



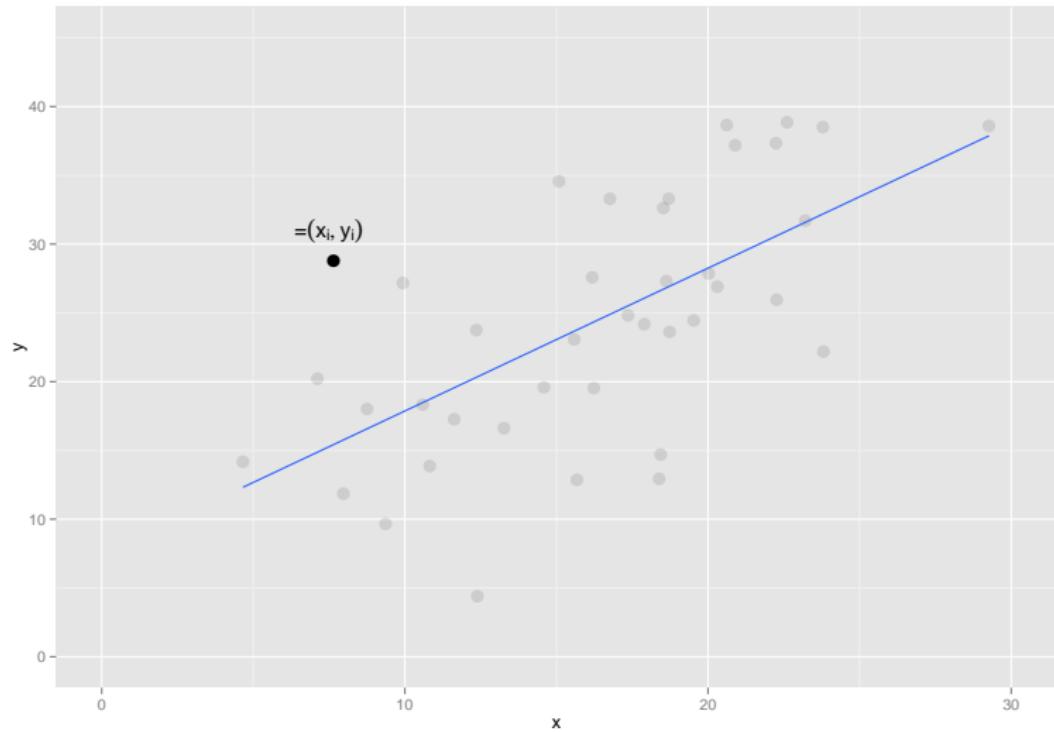
The Model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



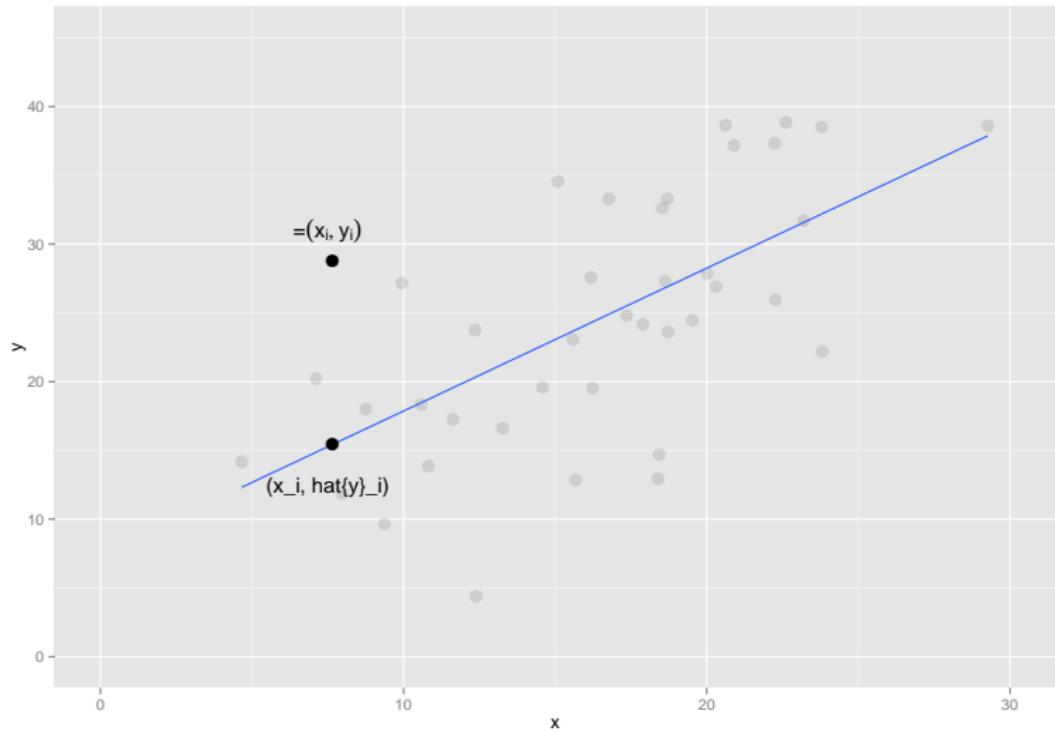
The Model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



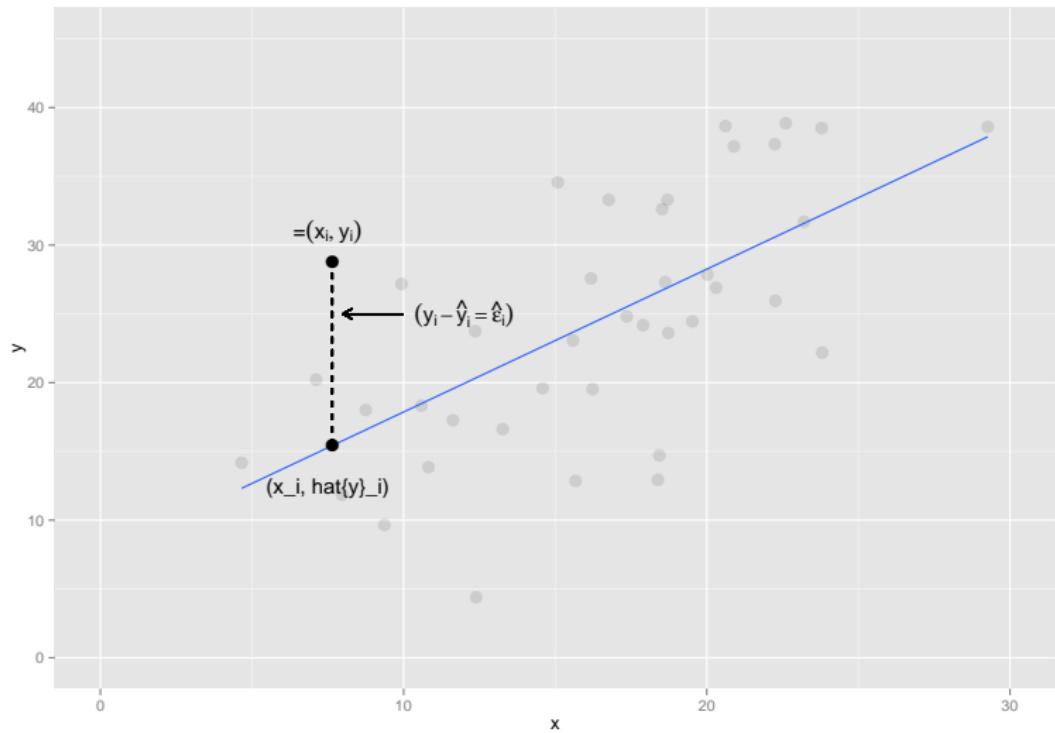
The Model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



The Model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



The Model

$$\underbrace{y_i}_{\text{outcome}} = \beta_0 + \beta_1 x_i + \epsilon_i$$

The Model

$$\underbrace{y_i}_{\text{outcome}} = \beta_0 + \beta_1 \underbrace{x_i}_{\text{predictor}} + \epsilon_i$$

The Model

$$\underbrace{y_i}_{response} = \beta_0 + \beta_1 \underbrace{x_i}_{stimulus} + \epsilon_i$$

The Model

$$\underbrace{y_i}_{\text{dependent var}} = \beta_0 + \beta_1 \underbrace{x_i}_{\text{independent var}} + \epsilon_i$$

The Model

$$\underbrace{y_i}_{\text{outcome}} = \underbrace{\beta_0}_{y\text{-intercept}} + \underbrace{\beta_1}_{\text{slope}} \underbrace{x_i}_{\text{predictor}} + \epsilon_i$$

The Model

$$\underbrace{y_i}_{\text{outcome}} = \underbrace{\beta_0}_{y\text{-intercept}} + \underbrace{\beta_1}_{\text{slope}} \underbrace{x_i}_{\text{predictor}} + \underbrace{\epsilon_i}_{\text{error}}$$

The Model

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{systematic}} + \epsilon_i$$

The Model

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{systematic}} + \underbrace{\epsilon_i}_{\text{stochastic}}$$

The Model

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{signal}} + \epsilon_i$$

The Model

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{signal}} + \underbrace{\epsilon_i}_{\text{noise}}$$

The Model

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{pattern}} + \underbrace{\epsilon_i}_{\text{randomness}}$$

The Model

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{what we model}} + \underbrace{\epsilon_i}_{\text{what we ignore}}$$

The Model

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{what we model}} + \underbrace{\epsilon_i}_{\text{what we ignore}}$$

Have we ignored important things?

The Model

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{what we model}} + \underbrace{\epsilon_i}_{\text{what we ignore}}$$

Have we ignored important things?

Is there a pattern in the ϵ_i ?

The Posited Model

- Model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

The Posited Model

- ▶ Model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- ▶ Errors (amt “true” model off by):
 $\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$

The Estimated Model

- $\hat{\beta}_0$: the estimated intercept

The Estimated Model

- ▶ $\hat{\beta}_0$: the estimated intercept
- ▶ $\hat{\beta}_1$: the estimated slope

The Estimated Model

- ▶ $\hat{\beta}_0$: the estimated intercept
- ▶ $\hat{\beta}_1$: the estimated slope
- ▶ $\hat{\epsilon}_i$: amt *estimated* model off by,
“residual”

The Estimated Model

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i$$

The Estimated Model

$$\underbrace{y_i}_{\text{observed}} = \hat{\beta}_0 + \hat{\beta}_1 \underbrace{x_i}_{\text{observed}} + \hat{\epsilon}_i$$

The Estimated Model

$$y_i = \underbrace{\hat{\beta}_0}_{\text{estimated}} + \underbrace{\hat{\beta}_1}_{\text{estimated}} x_i + \underbrace{\hat{\epsilon}_i}_{\text{residual}}$$

The Estimated Model

$$y_i = \underbrace{\hat{\beta}_0}_{\text{estimated}} + \underbrace{\hat{\beta}_1}_{\text{estimated}} x_i + \underbrace{\hat{\epsilon}_i}_{\text{residual}}$$

$$y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}_{\text{prediction}} + \underbrace{\hat{\epsilon}_i}_{\text{residual}}$$

The Estimated Model

$$\underbrace{\hat{\epsilon}_i}_{\text{residual}} = y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{\text{prediction}}$$

The Estimated Model

$$\underbrace{\hat{\epsilon}_i}_{\text{residual}} = y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{\text{fitted value}}$$

The Estimated Model

$$\underbrace{\hat{\epsilon}_i}_{\text{residual}} = y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{\hat{y}_i}$$

The Estimated Model

$$\underbrace{\hat{\epsilon}_i}_{\text{residual}} = y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{\hat{y}_i}$$

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

Interpretation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- ▶ Intercept: when $x_i = 0$, predicted \hat{y}_i is $\hat{\beta}_0$

Interpretation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- ▶ Intercept: when $x_i = 0$, predicted \hat{y}_i is $\hat{\beta}_0$
- ▶ Slope:

Interpretation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- ▶ Intercept: when $x_i = 0$, predicted \hat{y}_i is $\hat{\beta}_0$
- ▶ Slope:
 - ▶ Continuous x : if x is 1 unit greater, expect \hat{y}_i greater by $\hat{\beta}_1$

Interpretation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- ▶ Intercept: when $x_i = 0$, predicted \hat{y}_i is $\hat{\beta}_0$
- ▶ Slope:
 - ▶ Continous x : if x is 1 unit greater, expect \hat{y}_i greater by $\hat{\beta}_1$
 - ▶ Binary $x \in \{0, 1\}$: if $x_i = 1$, expect \hat{y}_i gr than $x_i = 0$ by $\hat{\beta}_1$

Interpretation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- ▶ Intercept: when $x_i = 0$, predicted \hat{y}_i is $\hat{\beta}_0$
- ▶ Slope:
 - ▶ Continous x : if x is 1 unit greater, expect \hat{y}_i greater by $\hat{\beta}_1$
 - ▶ Binary $x \in \{0, 1\}$: if $x_i = 1$, expect \hat{y}_i gr than $x_i = 0$ by $\hat{\beta}_1$
 - ▶ Binary treatment indicator: average treatment effect is $\hat{\beta}_1$

The Linear Model in R

Estimating a Linear Model in R

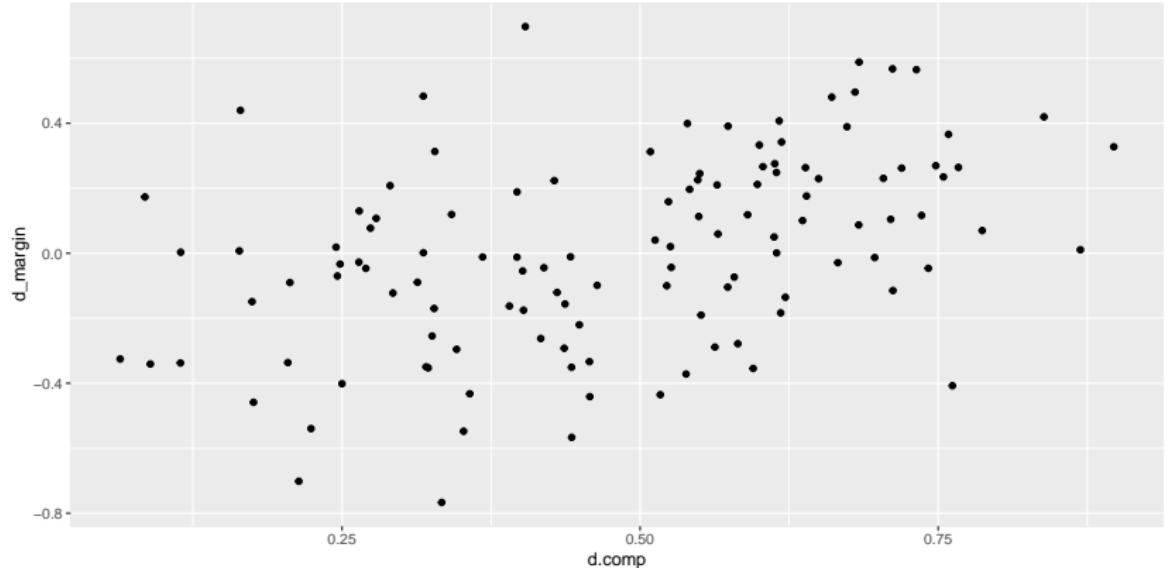
```
library(qss)
data(face)
head(face)
```

	year	state	winner	loser	w.party	l.party	d.comp	r.comp	d.
1	2000	CA	Feinstein	Campbell	D	R	0.5645676	0.4354324	57
2	2000	DE	Carper	Roth	D	R	0.3419122	0.6580878	1
3	2000	FL	Nelson	McCollum	D	R	0.6123680	0.3876320	29
4	2000	GA	Miller	Mattingly	D	R	0.5415328	0.4584672	13
5	2000	HI	Akaka	Carroll	D	R	0.6802323	0.3197677	2
6	2000	IN	Lugar	Johnson	R	D	0.3205024	0.6794976	6

	r.votes
1	3779325
2	142683
3	2703608
4	933698
5	84657
6	1419629

Estimating a Linear Model in R

```
face <- face |>  
  mutate(d_margin = (d.votes - r.votes) / (d.votes + r.votes))  
ggplot(face, aes(d.comp, d_margin)) + geom_point()
```



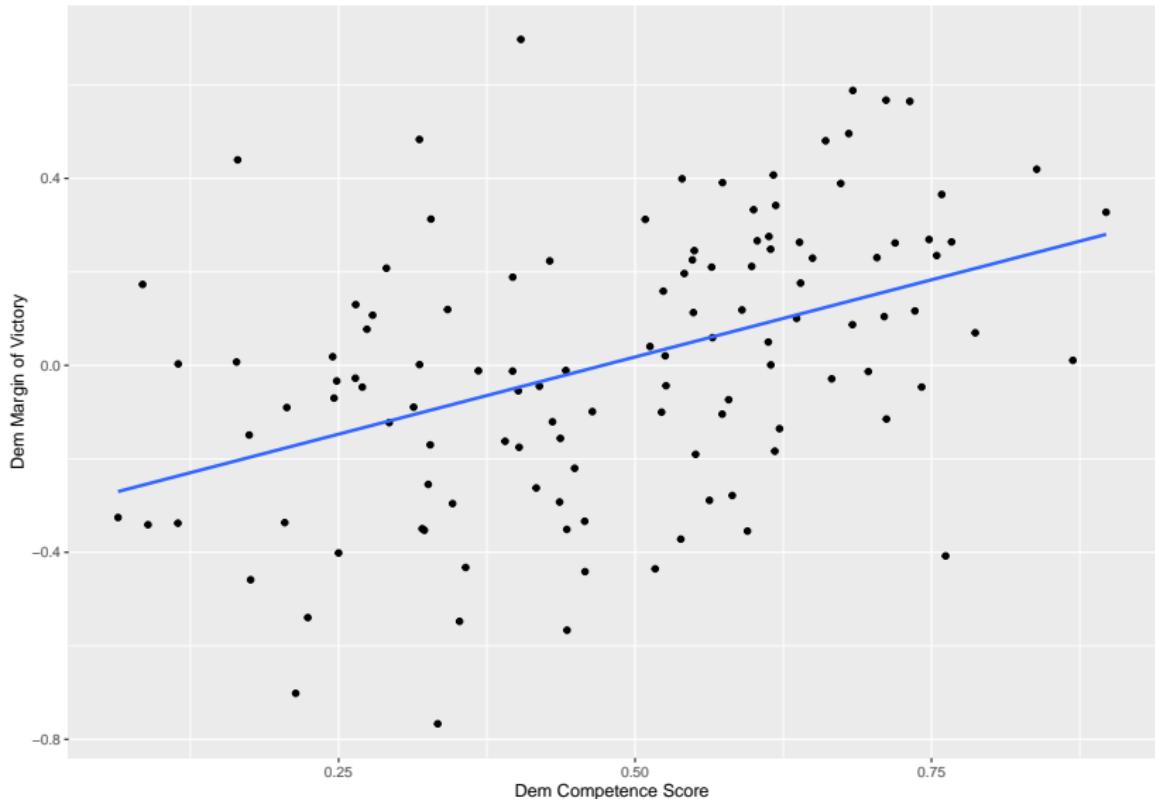
```
lm_out <- lm(d_margin ~ d.comp, data = face)  
lm_out
```

Call:

```
lm(formula = d_margin ~ d.comp, data = face)
```

Coefficients:

(Intercept)	d.comp
-0.3122	0.6604



```
summary(lm_out)
```

Call:

```
lm(formula = d_margin ~ d.comp, data = face)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.67487	-0.16600	0.01399	0.17741	0.74297

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	-0.31223	0.06596	-4.733	6.24e-06 ***							
d.comp	0.66038	0.12718	5.193	8.85e-07 ***							

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'. '	0.1	' '	1

Residual standard error: 0.2665 on 117 degrees of freedom

Multiple R-squared: 0.1873, Adjusted R-squared: 0.1803

F-statistic: 26.96 on 1 and 117 DF, p-value: 8.854e-07