Regression

Recap

- Correlations are their own effect size
- On a scale of -1 to 1
- Useful for depicting relationships

Today

Regression

- What is it? Why is it useful
- Nuts and bolts
  - Equation
  - Ordinary least squares
  - Interpretation

Regression

- Regression is an umbrella term -- lots of things fall under "regression"
- This system can handle a variety of forms of relations, although all forms have to be specified in a *linear* way.

The output of regression includes both effect sizes and statistical significance. We can also incorporate multiple influences (IVs) and account for their intercorrelations.

Regression

- **Scientific** use: explaining the influence of one or more variables on some outcome.
    - Does this intervention affect reaction time?
    - Does self-esteem predict relationship quality?
- **Prediction** use: We can develop models based on what's happened in the past to predict what will happen in the figure.
    - Insurance premiums
    - Graduate school... success?
- **Adjustment**: Statistically control for known effects
    - If everyone had the same level of SES, would abuse still be associated with criminal behavior?

How does Y vary with X?

- The regression of Y (DV) on X (IV) corresponds to the line that gives the mean value of Y corresponding to each possible value of X
- "Our best guess" regardless of whether our model includes categories or continuous predictor variables

Regression Equation

$$Y = b_0 + b_1 X + e$$

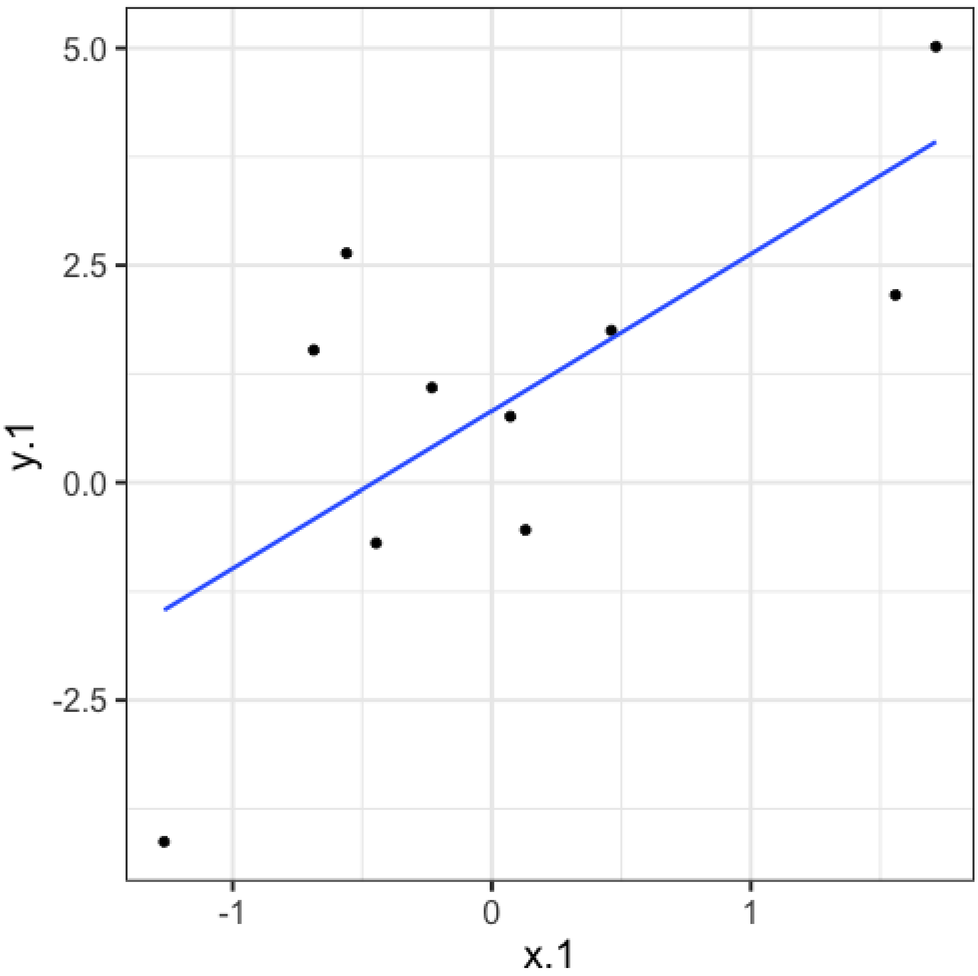
$$\hat{Y} = b_0 + b_1 X$$

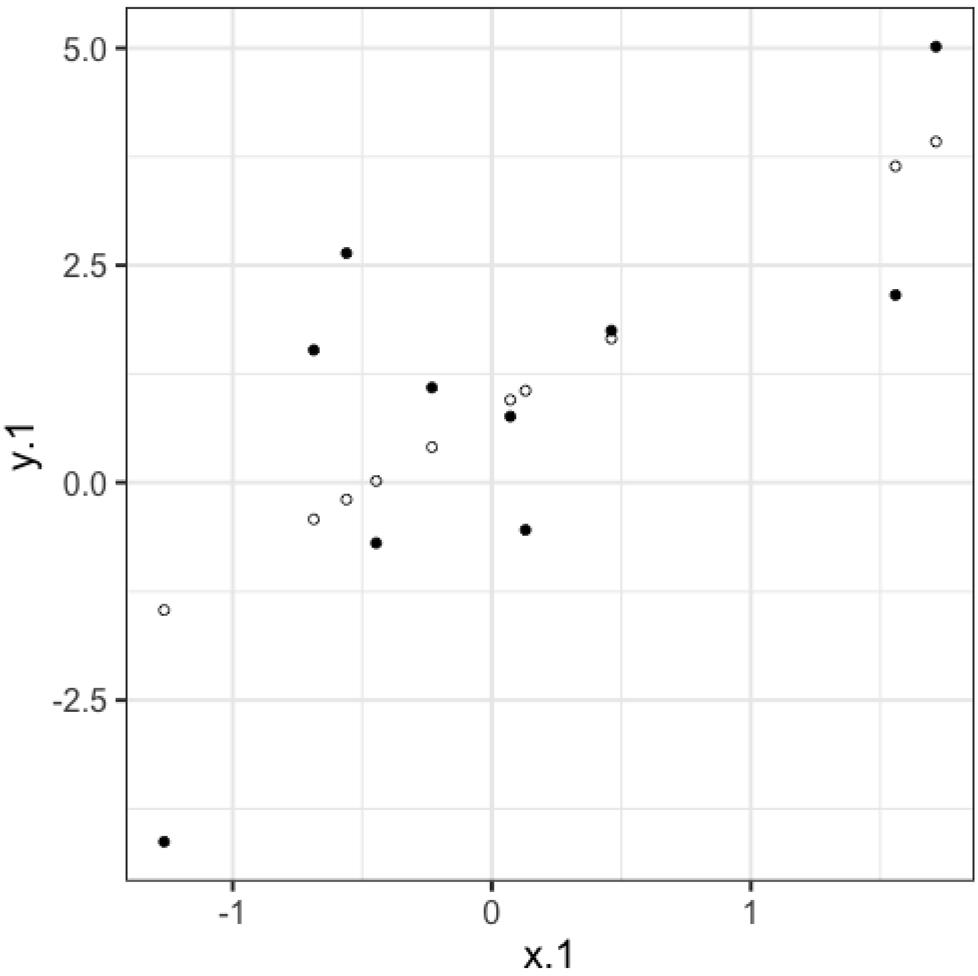
OLS

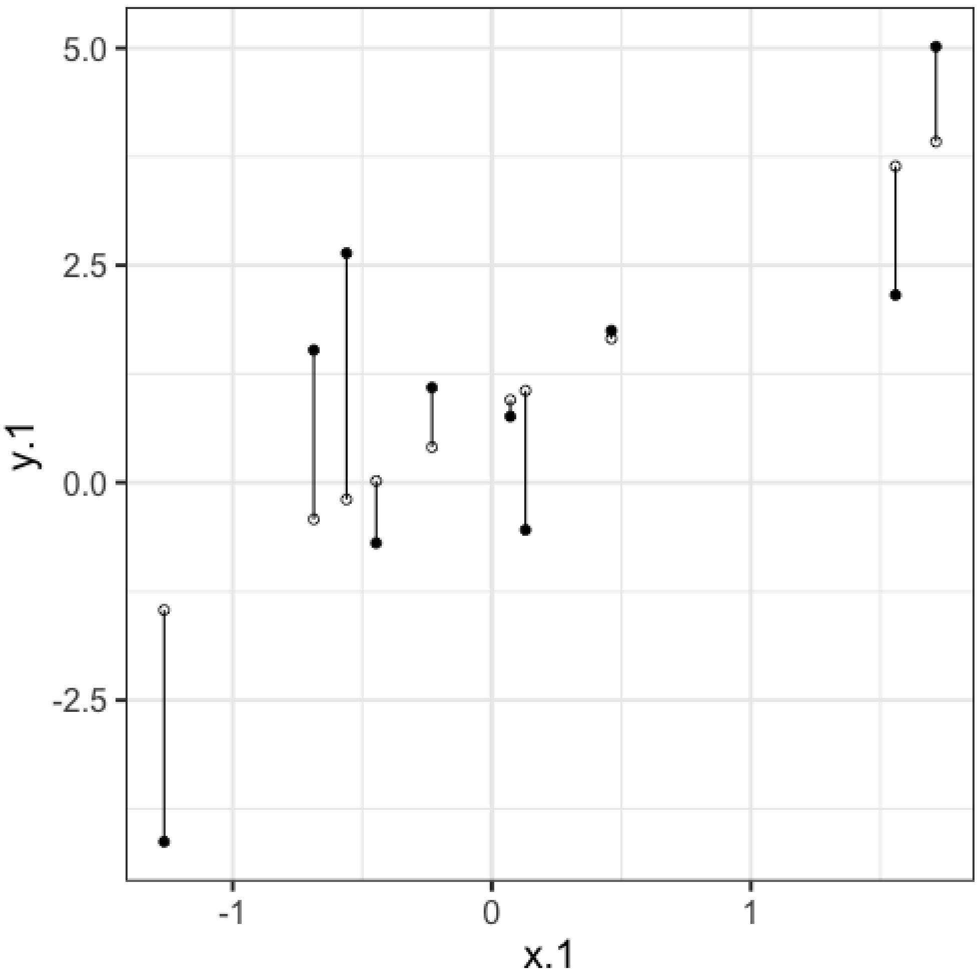
- How do we find the regression estimates?
- Ordinary Least Squares (OLS) estimation
- Minimizes deviations

$$min \sum (Y_i - \hat{Y})^2$$

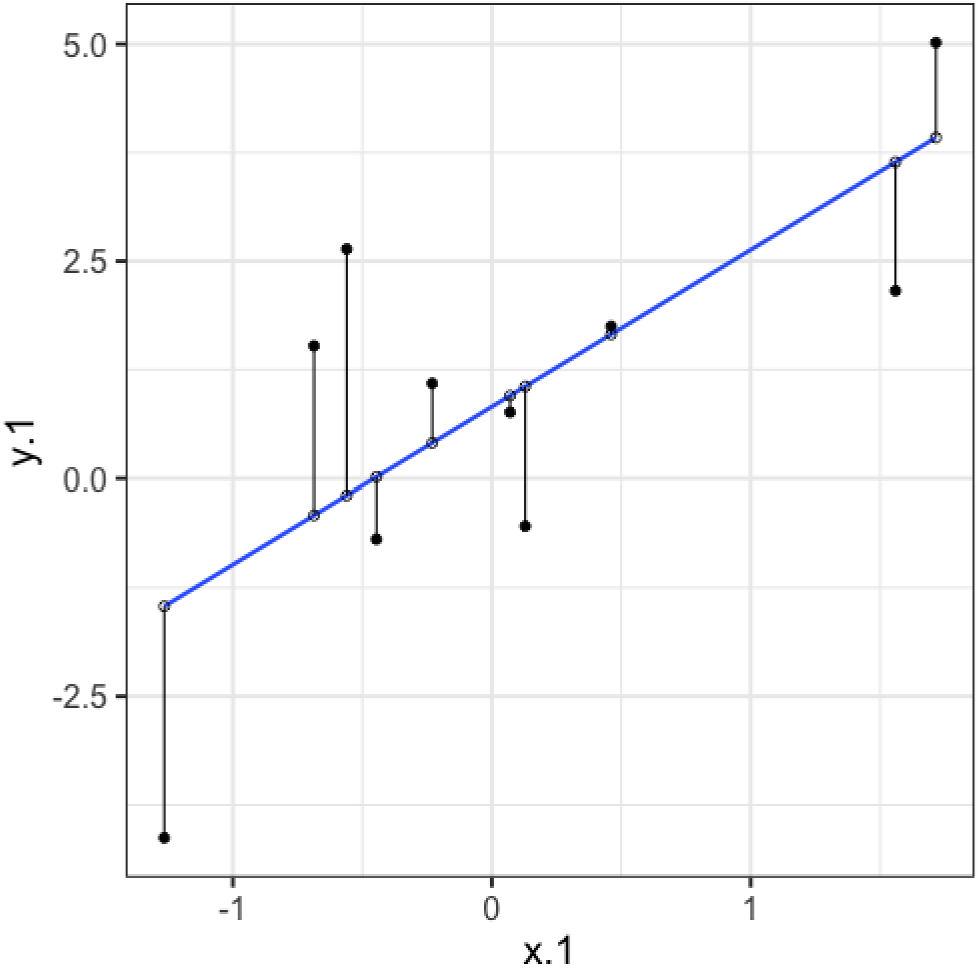
- Other estimation procedures possible (and necessary in some cases)

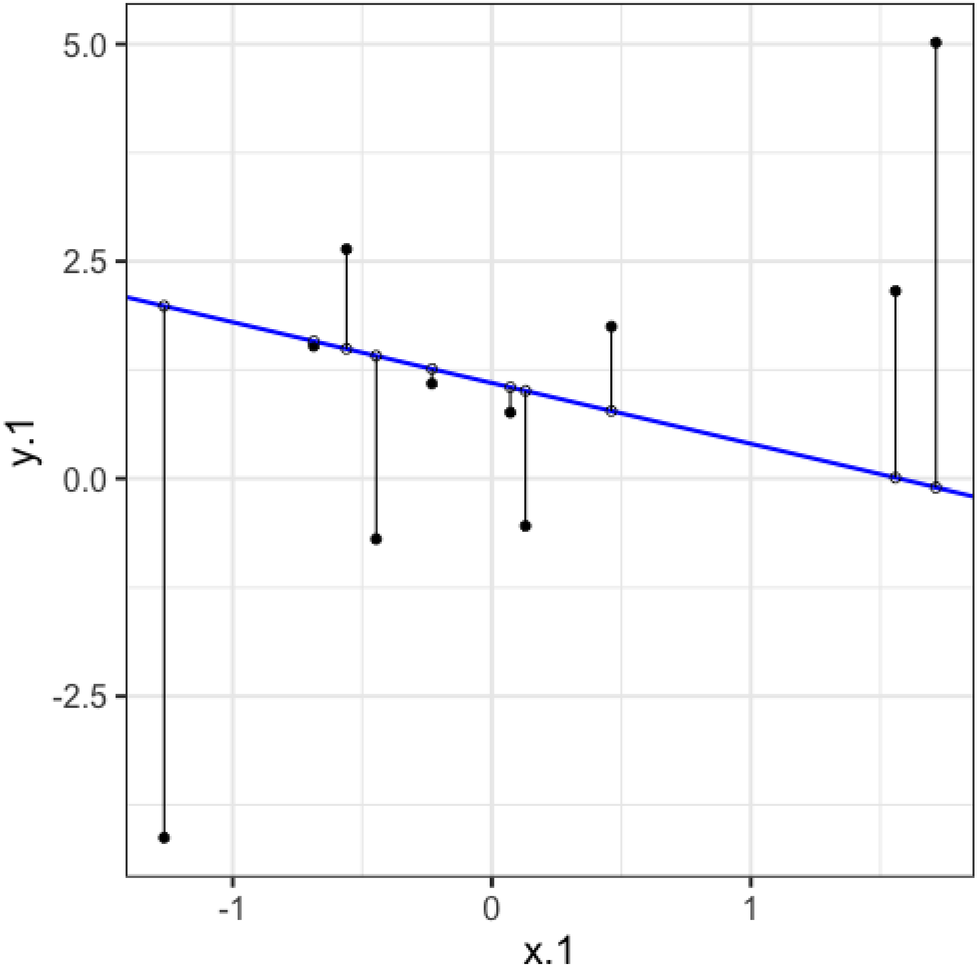

5.0
2.5
0.0
-2.5
y.1
-1
0
1
x.1


5.0
2.5
0.0
-2.5
y.1
-1
0
1
x.1



5.0
2.5
0.0
-2.5
y.1
-1
0
1
x.1


5.0
2.5
0.0
-2.5
y.1
-1
0
1
x.1



compare to bad fit

$$Y = b_0 + b_1 X + e$$

$$\hat{Y} = b_0 + b_1 X$$

$$Y_i = \hat{Y}_i + e_i$$

$$e_i = Y_i - \hat{Y}_i$$

OLS

The line that yields the smallest sum of squared deviations

$$\Sigma(Y_i - \hat{Y}_i)^2$$
$$= \Sigma(Y_i - (b_0 + b_1X_i))^2$$
$$= \Sigma(e_i)^2$$

In order to find the OLS solution, you could try many different coefficients $(b_0 \text{ and } b_1)$ until you find the one with the smallest sum squared deviation. Luckily, there are simple calculations that will yield the OLS solution every time.

Example

```
galton.data <- psychTools::galton
head(galton.data)
```

```
##   parent child
## 1   70.5  61.7
## 2   68.5  61.7
## 3   65.5  61.7
## 4   64.5  61.7
## 5   64.0  61.7
## 6   67.5  62.2
```

```
describe(galton.data, fast = T)
```

```
##        vars   n  mean   sd  min  max range   se
## parent    1 928 68.31 1.79 64.0 73.0     9 0.06
## child     2 928 68.09 2.52 61.7 73.7    12 0.08
```

```
cor(galton.data)
```

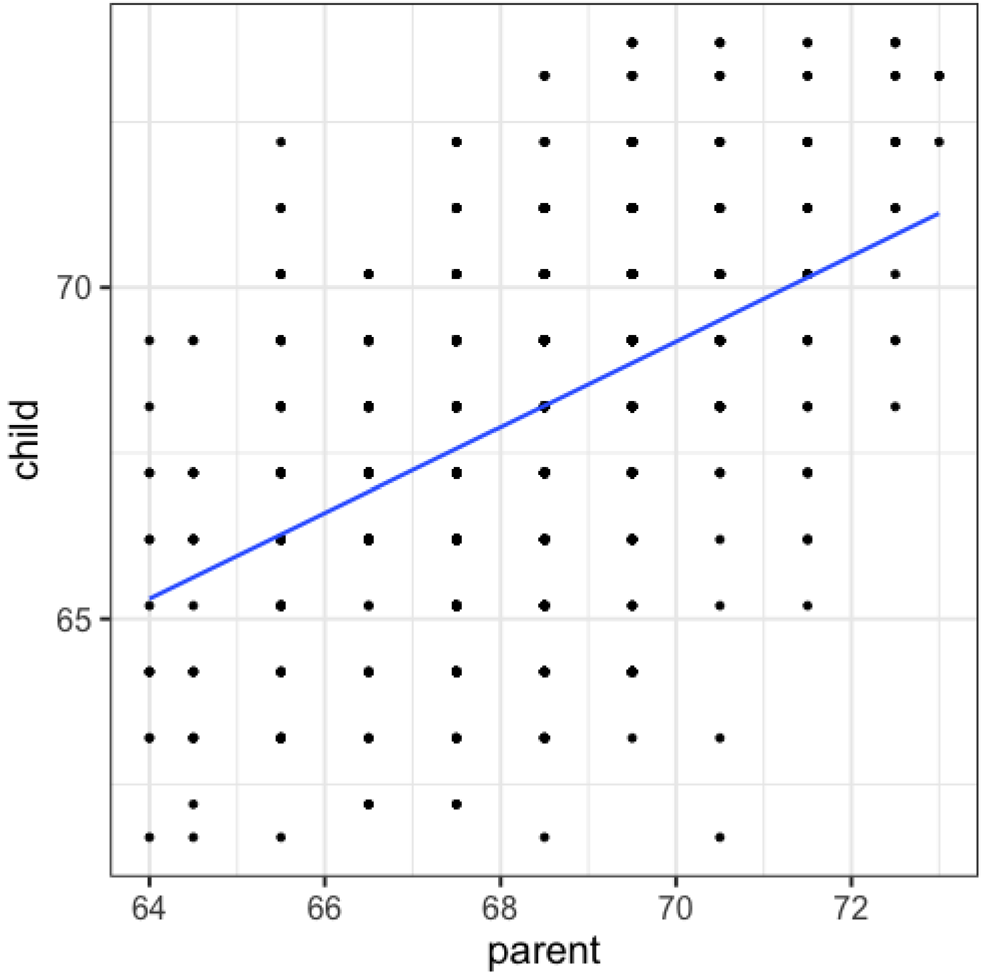
```
##           parent     child
## parent 1.0000000 0.4587624
```

In R

What if we regress parent height onto child height?

```
fit.1 <- lm(child ~ parent, data = galton.data)
summary(fit.1)
```

```
##
## Call:
## lm(formula = child ~ parent, data = galton.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8050 -1.3661  0.0487  1.6339  5.9264
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.94153    2.81088   8.517   <2e-16 ***
## parent       0.64629    0.04114  15.711   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.239 on 926 degrees of freedom
## Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
## F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```



child
70
65
64
66
68
70
72
parent

Data, predicted, and residuals

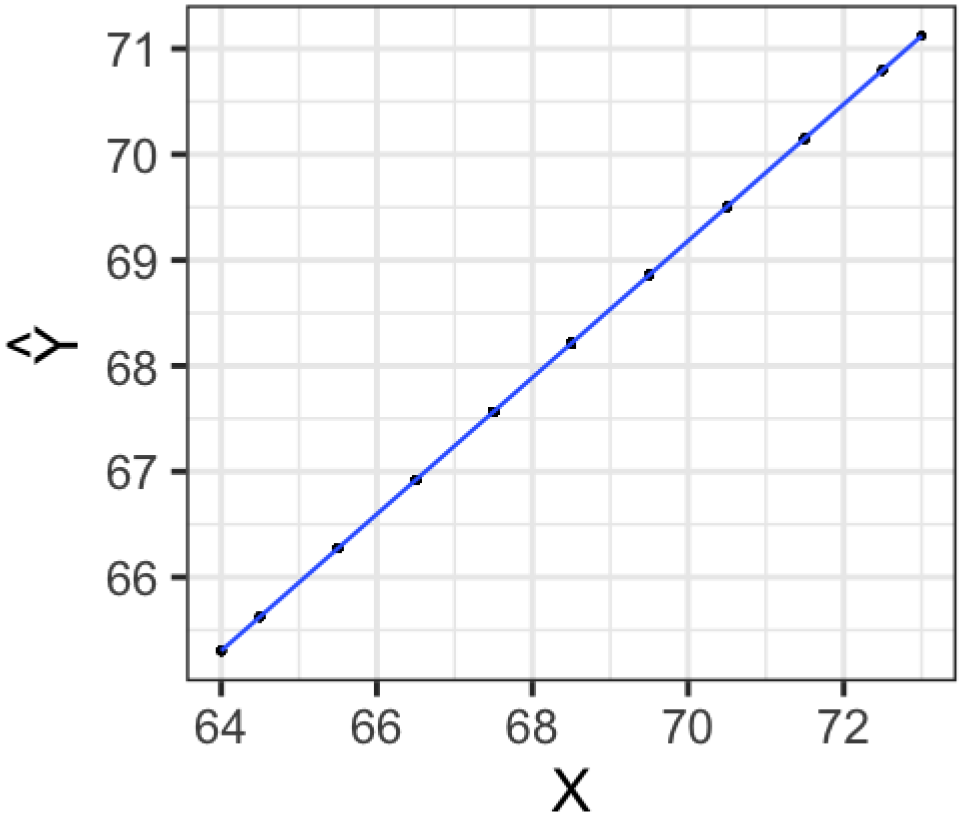
```r
library(broom)
model_info = augment(fit.1)
head(model_info)
```

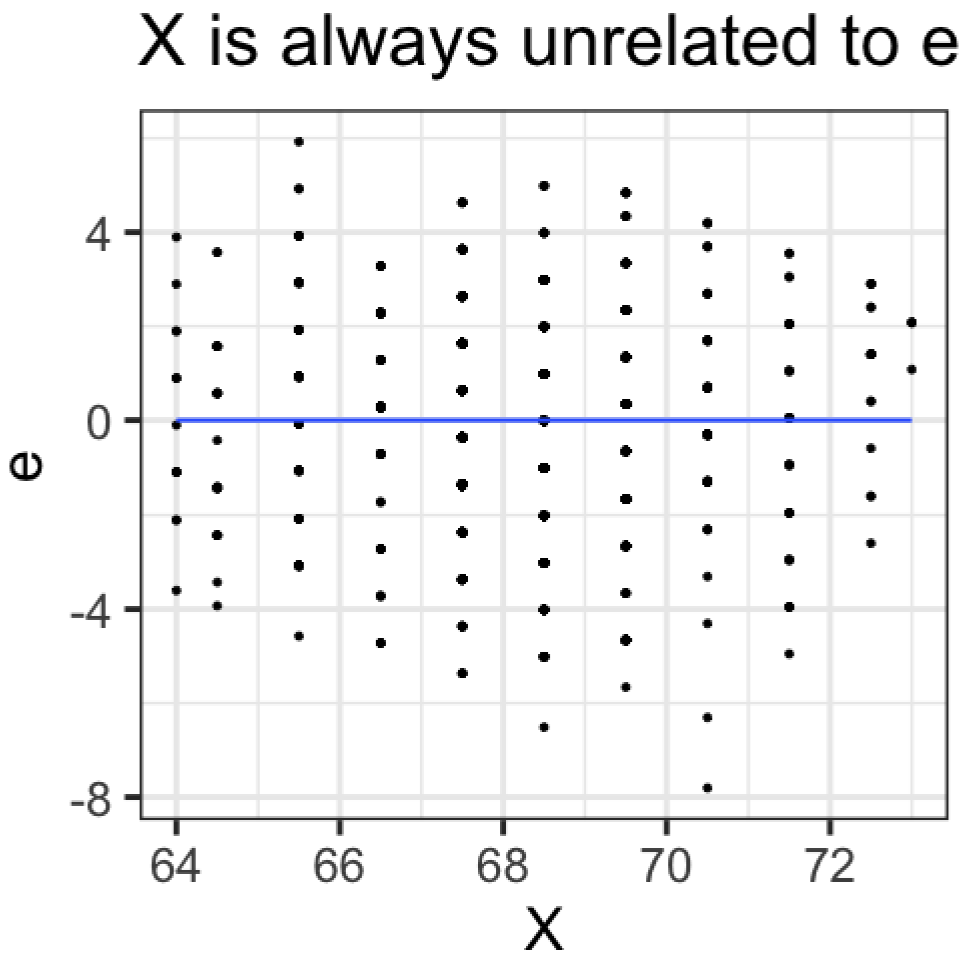
```
## # A tibble: 6 x 9
##   child parent .fitted .se.fit .resid     .hat .sigma .cooksd .std.resid
##   <dbl>  <dbl>   <dbl>   <dbl>  <dbl>    <dbl>  <dbl>   <dbl>      <dbl>
## 1  61.7   70.5    69.5  0.116   -7.81 0.00270    2.22 0.0165       -3.49
## 2  61.7   68.5    68.2  0.0739  -6.51 0.00109    2.23 0.00462      -2.91
## 3  61.7   65.5    66.3  0.137   -4.57 0.00374    2.23 0.00787      -2.05
## 4  61.7   64.5    65.6  0.173   -3.93 0.00597    2.24 0.00931      -1.76
## 5  61.7   64      65.3  0.192   -3.60 0.00735    2.24 0.00966      -1.62
## 6  62.2   67.5    67.6  0.0807  -5.37 0.00130    2.23 0.00374      -2.40
```

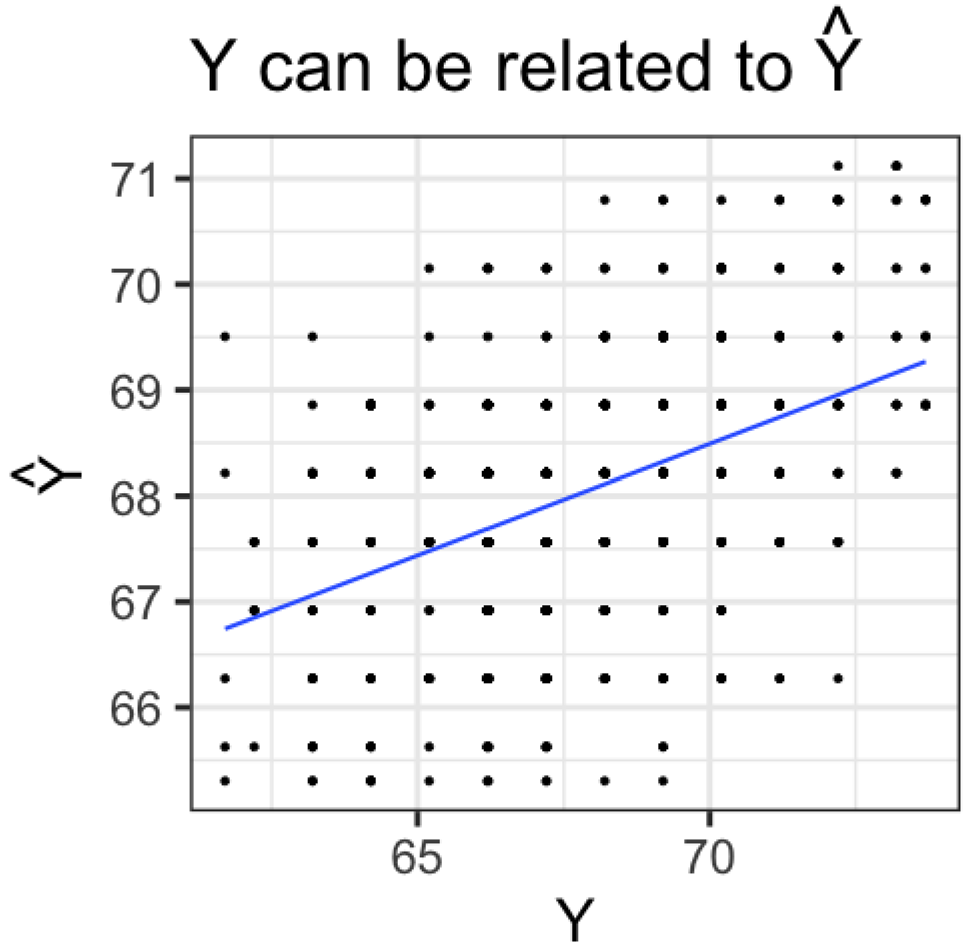
```r
describe(model_info)
```

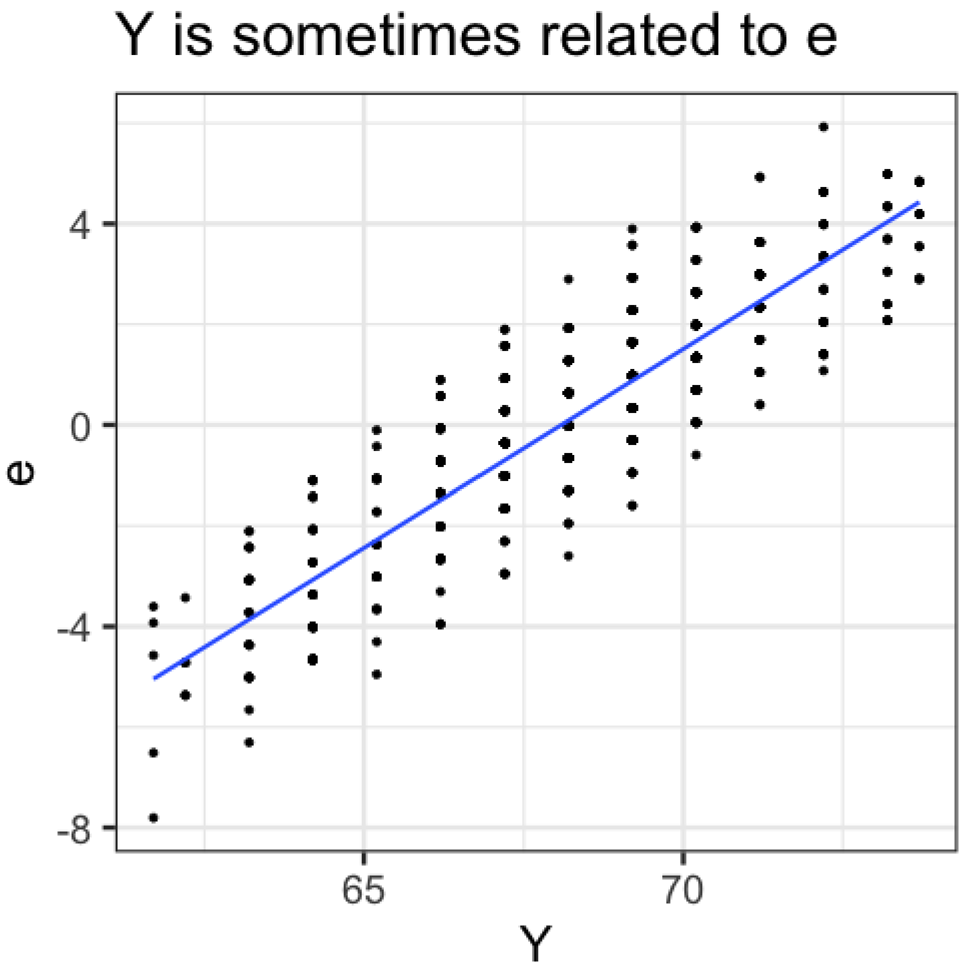
```
##            vars   n  mean   sd median trimmed  mad   min   max range  skew
## child         1 928 68.09 2.52  68.20   68.12 2.97 61.70 73.70 12.00 -0.09
## parent        2 928 68.31 1.79  68.50   68.32 1.48 64.00 73.00  9.00 -0.04
## .fitted       3 928 68.09 1.16  68.21   68.10 0.96 65.30 71.12  5.82 -0.04
## .se.fit       4 928  0.10 0.03   0.09    0.09 0.02  0.07  0.21  0.13  1.53
## .resid        5 928  0.00 2.24   0.05    0.06 2.26 -7.81  5.93 13.73 -0.24
```

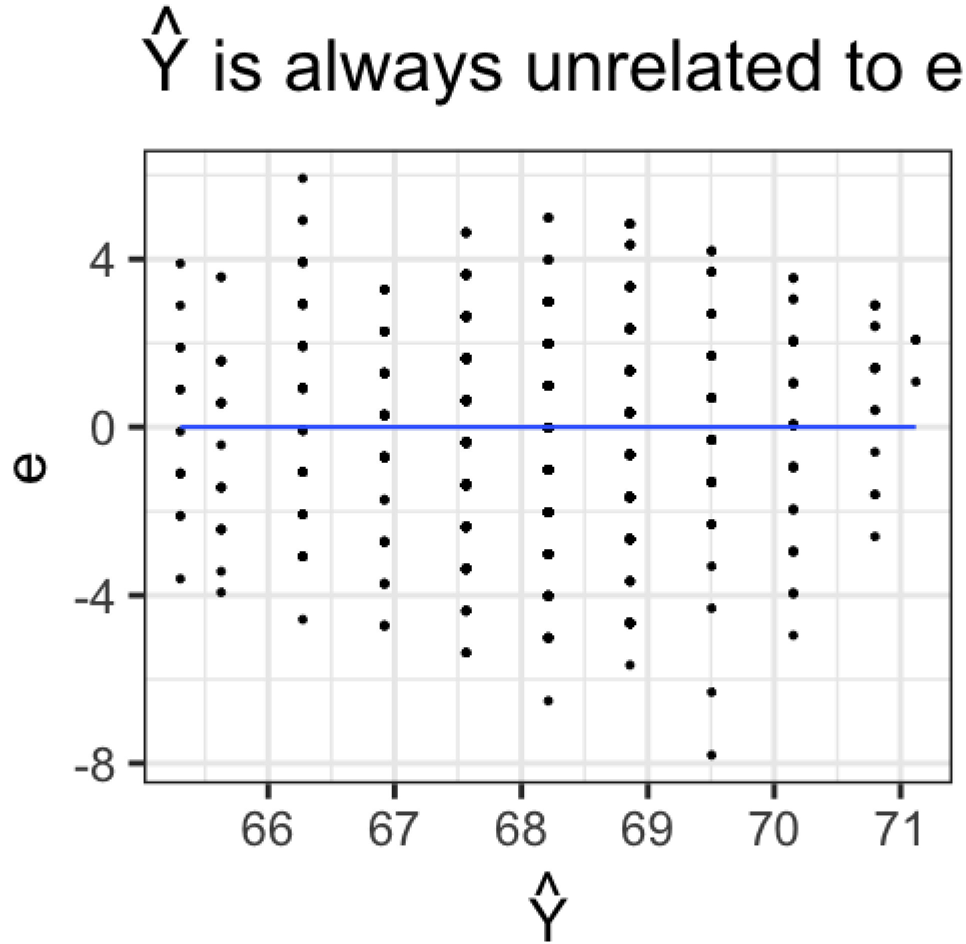
X is related to $\hat{Y}$

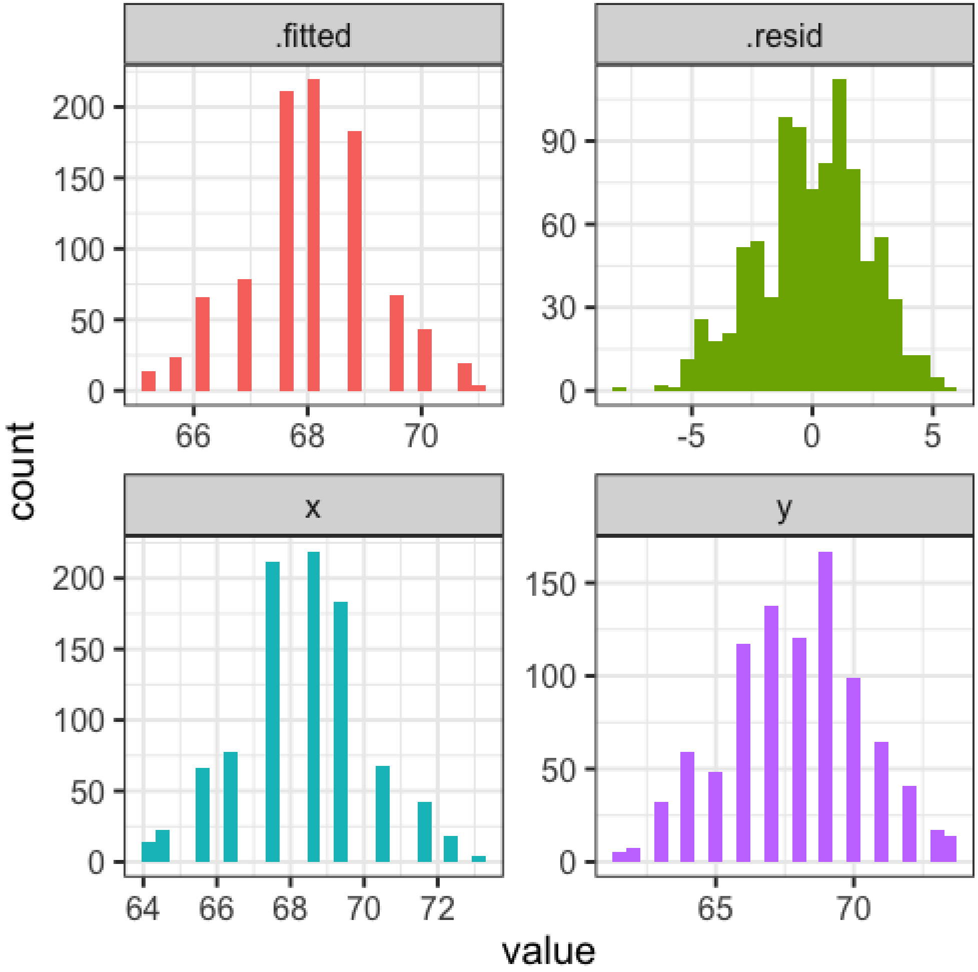



X is always unrelated to e
4
0
-4
-8
e
64
66
68
70
72
X


Y can be related to Ŷ
71
70
69
68
67
66
Ŷ
65
70
Y


Y is sometimes related to e
4
0
-4
-8
e
65
70
Y


Ŷ is always unrelated to e
4
0
-4
-8
e
66
67
68
69
70
71
Ŷ



.fitted
.resid
x
y
count
value
200
150
100
50
0
66
68
70
90
60
30
-5
5
64
72
65