Univariate regression

Last time

- Correlation as inferential test
- Fisher's r to z transformation
- Correlation matrices
- Interpreting effect size

Today

Regression

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

GLM/Regression

GLM is a general data analytic system, meaning lots of things fall under the umbrella of regression.

- *General* - broad set of similar models; can be applied to almost any context
  - IVs can be continuous, categorical, nominal, ordinal....
- *Linear* - We try to understand our dependent variable (DV) via a linear combination predictor variables (add & multiply)

GLM/Regression

What do we get?

- effect sizes
- statistical significance
- incorporate multiple IVs
- account for intercorrelations

Regression

- **Scientific** use: explaining the influence of one or more variables on some outcome.
- **Prediction** use: We can develop models based on what's happened in the past to predict what will happen in the figure.
- **Adjustment**: Statistically control for known effects
  - If everyone had the same level of SES, would abuse still be associated with criminal behavior?

Regression equation

What is a regression equation?

- Functional relationship
    - Ideally like a physical law $(E = MC^2)$
    - In practice, it's never as robust as that

How do we uncover the relationship?

How does Y vary with X?

- $E(Y|X)$
- "Our best guess" regardless of whether our model includes categories or continuous predictor variables
- We will evaluate our guesses based on how far away we are from the mean. But how do we come up with those guesses in the first place?

Regression Equation

$$Y_i = b_0 + b_1 X_i + e_i$$

$$\hat{Y}_i = b_0 + b_1 X_i$$

$\hat{Y}$ signifies the predicted score -- no error

The difference between the predicted and observed score is the residual (e_i)

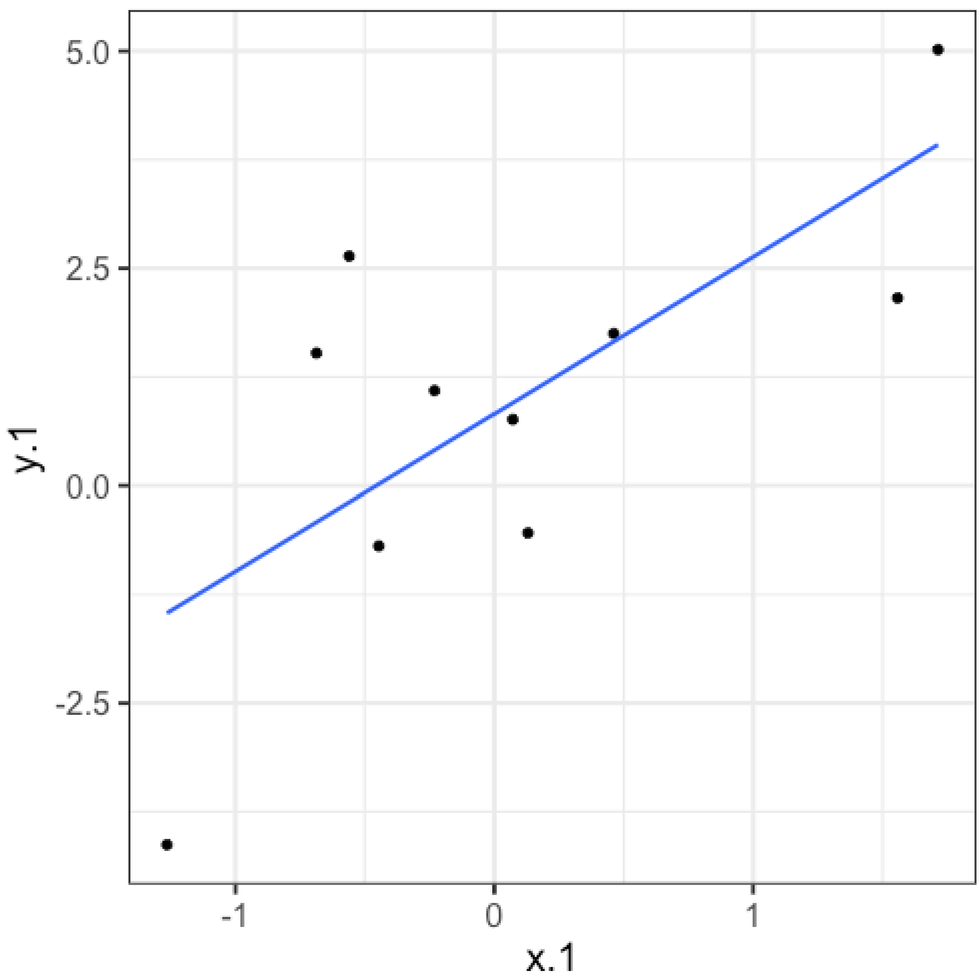
There is a different e value for each observation in the dataset

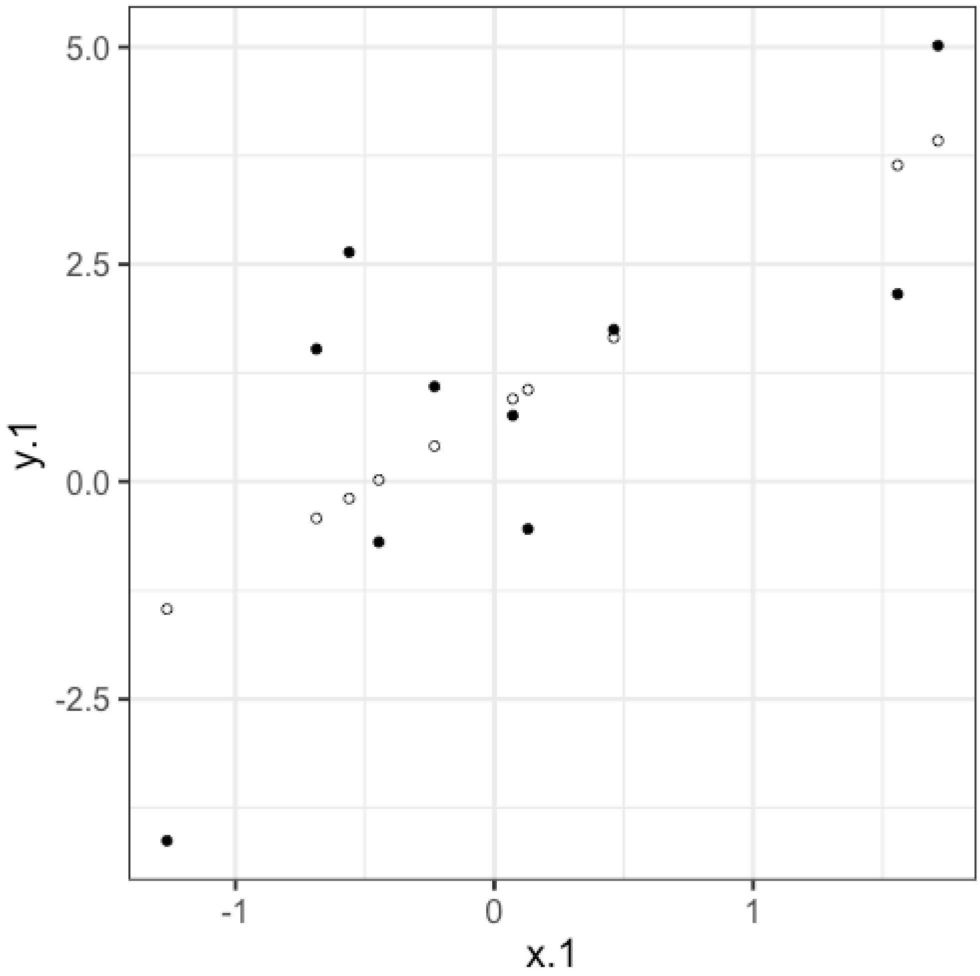
OLS

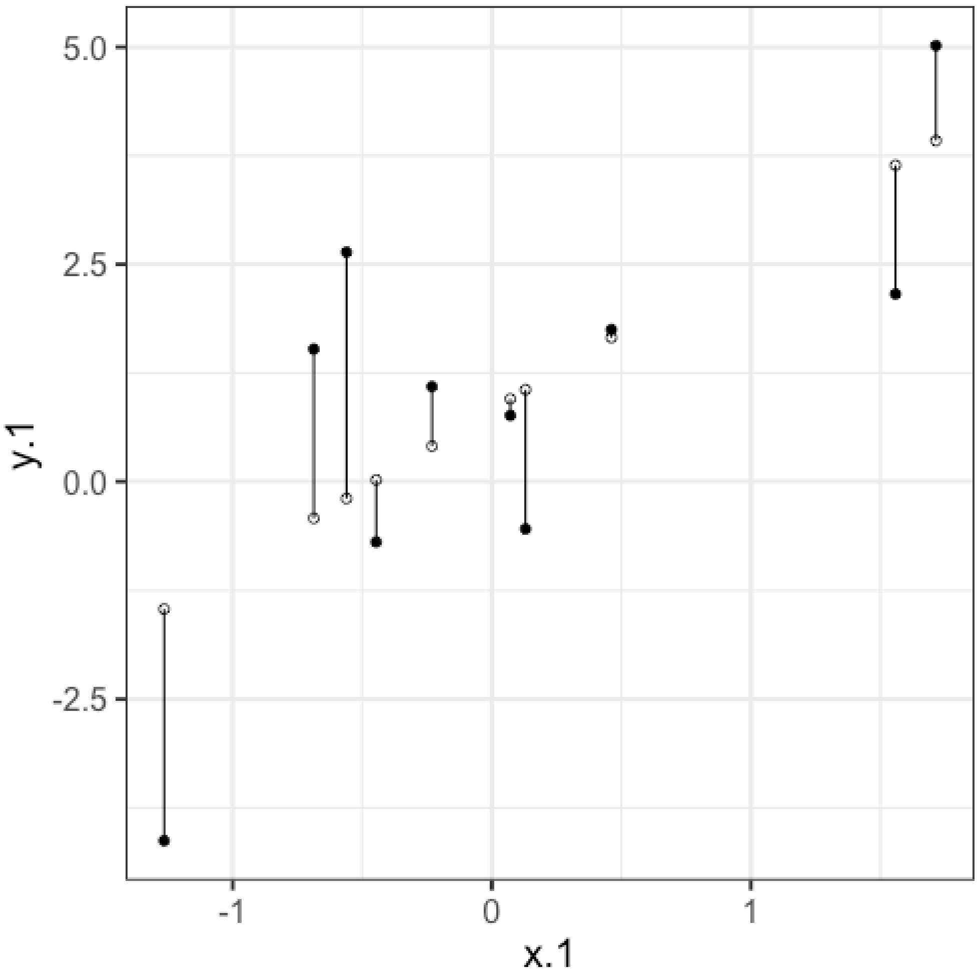
- How do we find the regression estimates?
- Ordinary Least Squares (OLS) estimation
- Minimizes deviations

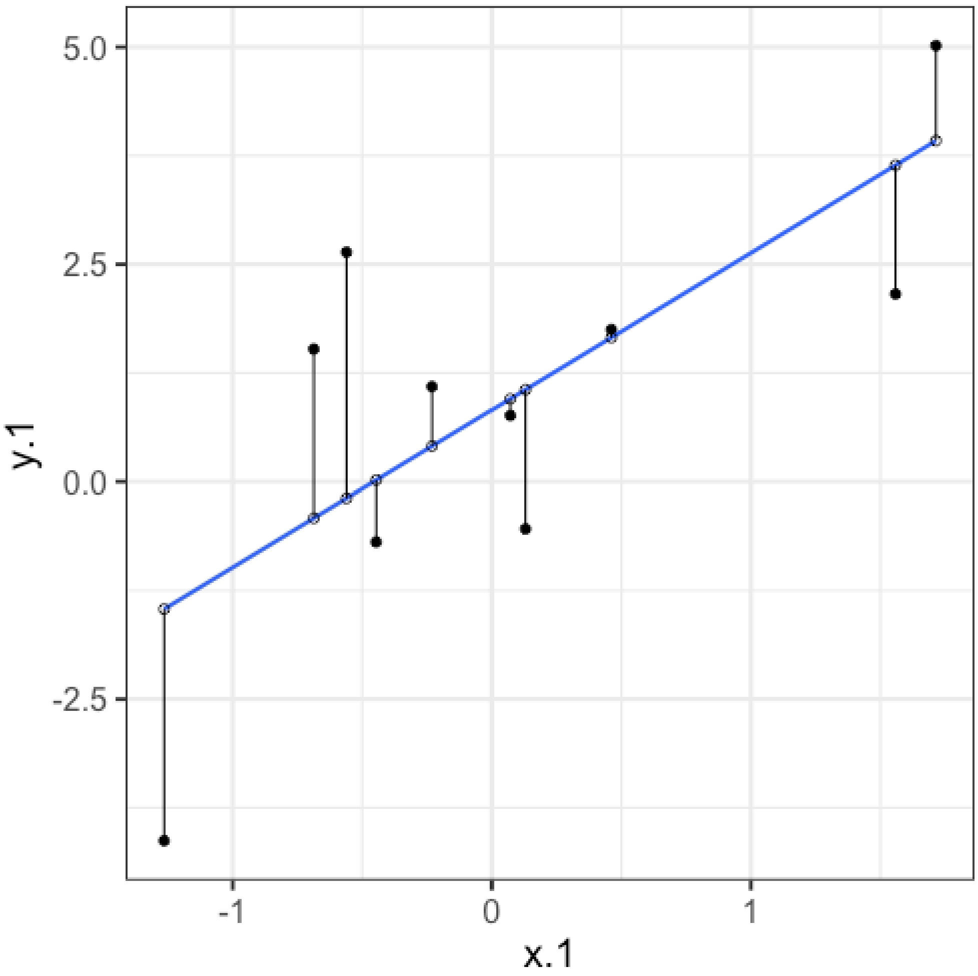
$$min \sum (Y_i - \hat{Y})^2$$

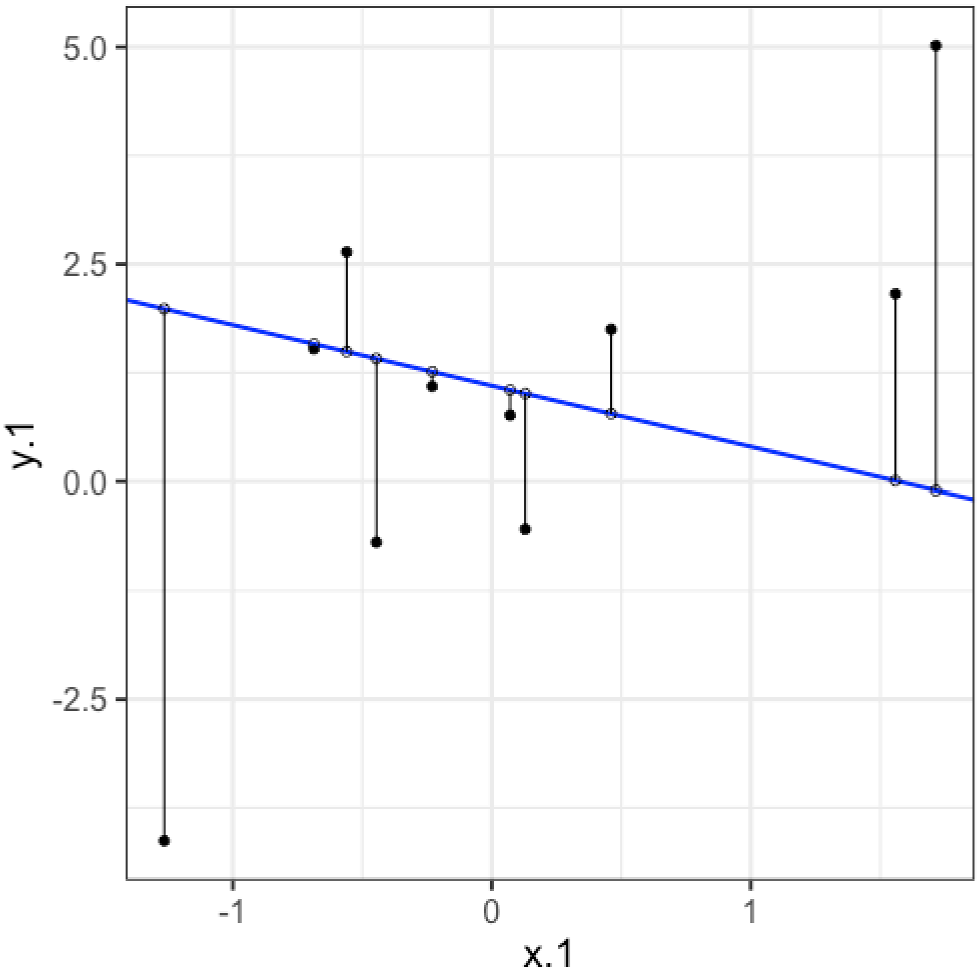
- Other estimation procedures possible (and necessary in some cases)


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

What is error?

$$Y_i = b_0 + b_1 X_i + e_i$$

$$\hat{Y}_i = b_0 + b_1 X_i$$

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

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

OLS

The line that yields the smallest sum of squared deviations

$$\Sigma(Y_i - \hat{Y}_i)^2$$

$$= \Sigma(Y_i - (b_0 + b_1 X_i))^2$$

$$= \Sigma(e_i)^2$$

In order to find the OLS solution, you could try many different coefficients $(b_0 \text{ and } b_1)$... or not

Regression coefficient, b_1

$$b_1 = \frac{cov_{XY}}{s_x^2} = r_{xy}\frac{s_y}{s_x}$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Suggested Practice: Go in R and see if you can prove to yourself that these equations are the same!

Standardized regression

- Regression using z-scores for Y and X
- Correlation equals standardized regression coefficient

$$b_1 = r_{xy} \frac{s_y}{s_x}$$

$$r_{xy} = b_1 \frac{s_x}{s_y}$$

Standardized regression

If the variance of both X and Y is equal to 1 (as in z-scores):

$$\beta_1 = b_1^* = r_{xy}$$

Standardized regression equation

$$Y = b_1^* X + e$$

$$b_1^* = b_1 \frac{s_x}{s_y}$$

According to this regression equation, when $X = 0, Y = 0$. Our interpretation of the coefficient is that a one-standard deviation increase in X is associated with a b_1^* standard deviation increase in Y. Our regression coefficient is equivalent to the correlation coefficient *when we have only one predictor in our model.*

Estimating the intercept raw b_0

- Re-write equation to include $\bar{X}$ & $\bar{Y}$
- Intercept serves to adjust for differences in means between X and Y

$$\hat{Y} = \bar{Y} + r_{xy}\frac{s_y}{s_x}(X - \bar{X})$$

- If standardized, intercept drops out
- Otherwise, intercept is where regression line crosses the y-axis at X = 0
- When $X = \bar{X}$ the regression line goes through $\bar{Y}$. This is true for all regressions -- line must pass through $\bar{X}$ and $\bar{Y}$

```r
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
```

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

```
##        vars   n  mean   sd median  min  max range  skew kurtosis   se
## parent    1 928 68.31 1.79   68.5 64.0 73.0     9 -0.04     0.05 0.06
## child     2 928 68.09 2.52   68.2 61.7 73.7    12 -0.09    -0.35 0.08
```

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

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

If we regress child height (Y) onto parents' (X):

```
r = cor(galton.data)[2,1]
m_parent = mean(galton.data$parent)
m_child = mean(galton.data$child)
s_parent = sd(galton.data$parent)
s_child = sd(galton.data$child)

(b1 = r*(s_child/s_parent))
```

```
## [1] 0.6462906
```

```
(b0 = m_child - b1*m_parent)
```

```
## [1] 23.94153
```

How will this change if we regress parent height onto child height?

```r
(b1 = r*(s_child/s_parent))
```

```
## [1] 0.6462906
```

```r
(b0 = m_child - b1*m_parent)
```

```
## [1] 23.94153
```

```r
(b1 = r*(s_parent/s_child))
```

```
## [1] 0.3256475
```

```r
(b0 = m_parent - b1*m_child)
```

```
## [1] 46.13535
```

In R

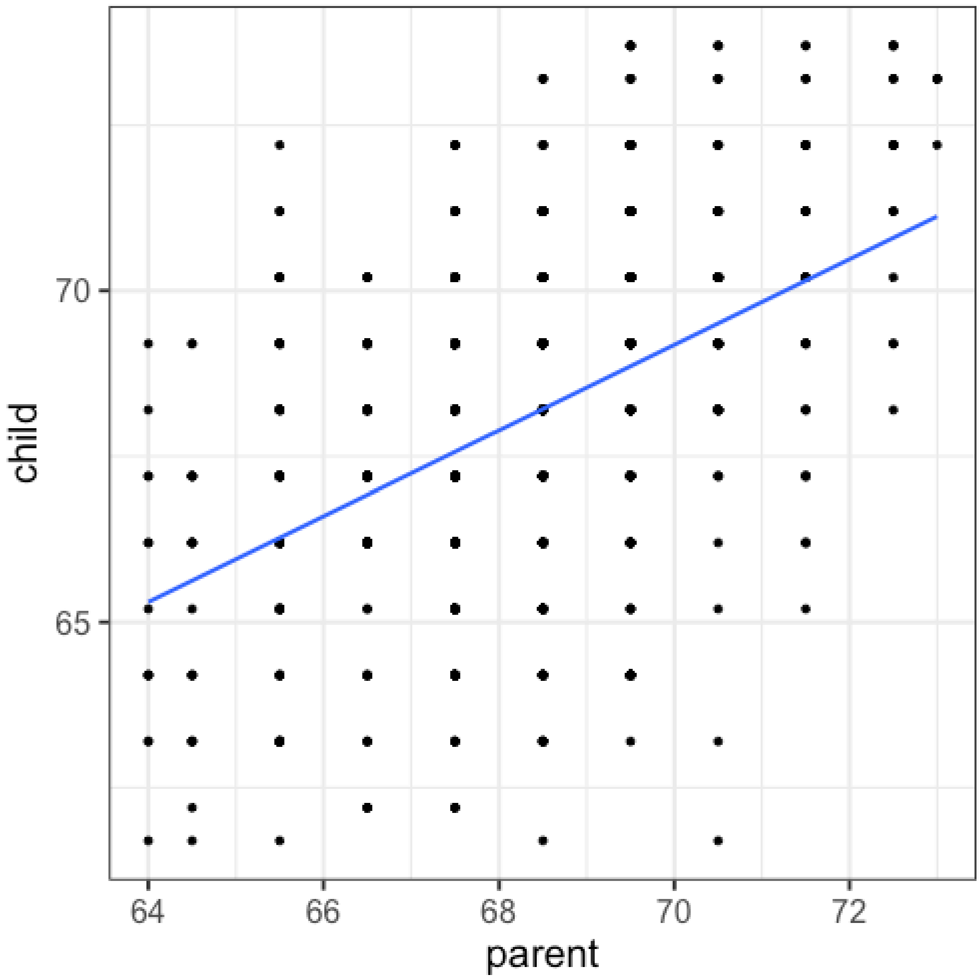
```
fit.1 <- lm(child ~ parent, data = galton.data)
summary(fit.1)
```

```
##
## 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
```

Reversed

```
summary(lm(parent ~ child, data = galton.data))
```

```
##
## Call:
## lm(formula = parent ~ child, data = galton.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6702 -1.1702 -0.1471  1.1324  4.2722
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.13535    1.41225   32.67   <2e-16 ***
## child        0.32565    0.02073   15.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.589 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
```



70
65
child
64
66
68
70
72
parent

Data, predicted, and residuals

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

```
## # A tibble: 6 × 8
##   child parent .fitted .resid    .hat .sigma .cooksd .std.resid
##   <dbl>  <dbl>   <dbl>  <dbl>   <dbl>  <dbl>   <dbl>      <dbl>
## 1  61.7   70.5    69.5  -7.81 0.00270   2.22 0.0165       -3.49
## 2  61.7   68.5    68.2  -6.51 0.00109   2.23 0.00462      -2.91
## 3  61.7   65.5    66.3  -4.57 0.00374   2.23 0.00787      -2.05
## 4  61.7   64.5    65.6  -3.93 0.00597   2.24 0.00931      -1.76
## 5  61.7   64      65.3  -3.60 0.00735   2.24 0.00966      -1.62
## 6  62.2   67.5    67.6  -5.37 0.00130   2.23 0.00374      -2.40
```

```
describe(model_info)
```

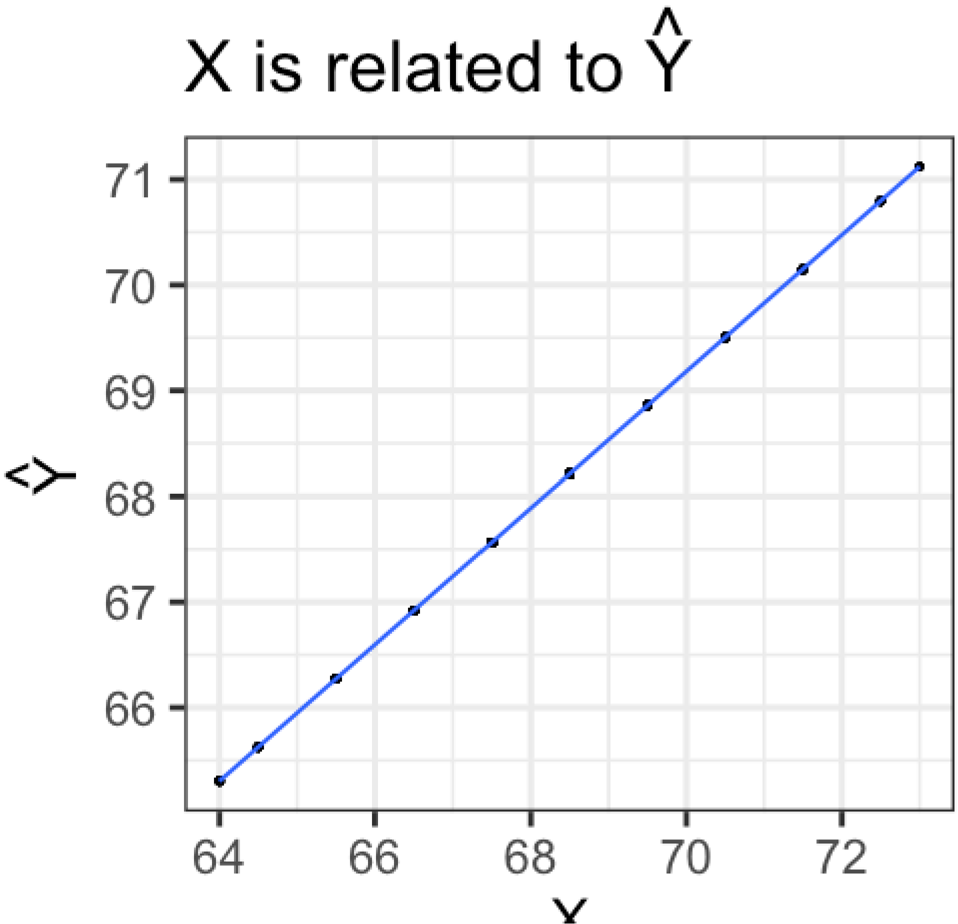
```
##            vars   n  mean   sd median trimmed  mad   min   max range -
## child         1 928 68.09 2.52  68.20   68.12 2.97 61.70 73.70 12.00 -
## parent        2 928 68.31 1.79  68.50   68.32 1.48 64.00 73.00  9.00 -
## .fitted       3 928 68.09 1.16  68.21   68.10 0.96 65.30 71.12  5.82 -
## .resid        4 928  0.00 2.24   0.05    0.06 2.26 -7.81  5.93 13.73 -
```

Residuals

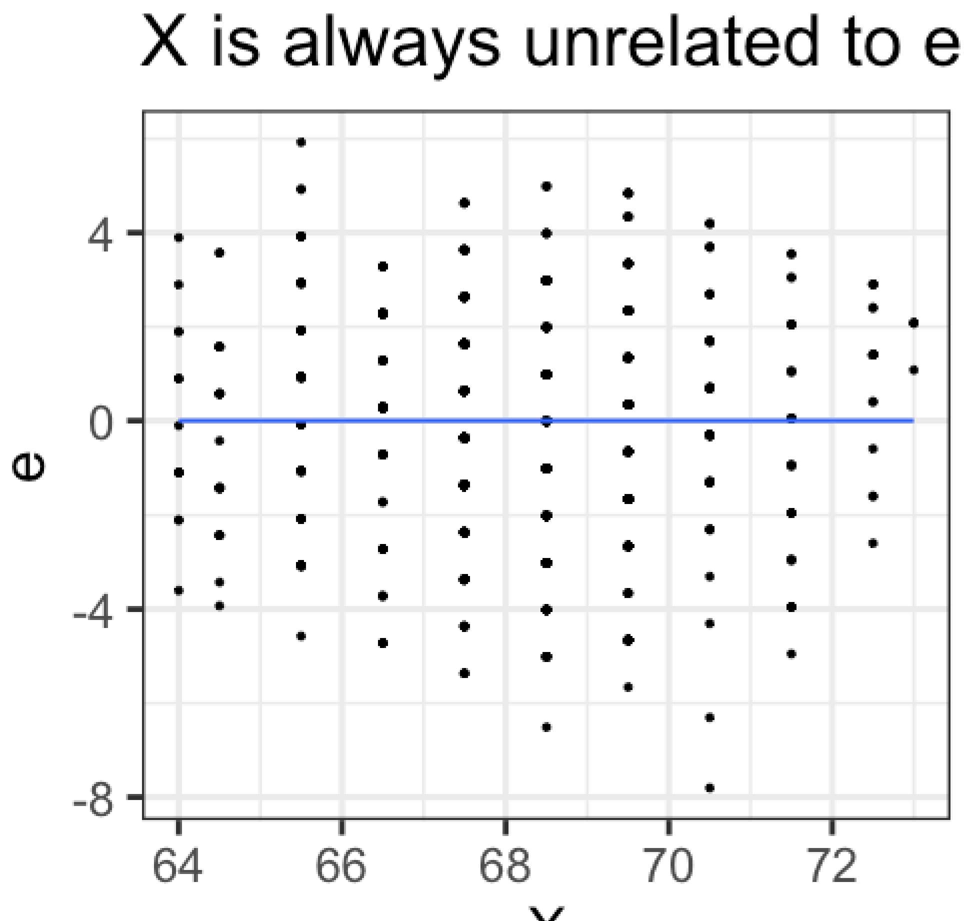
- Dispersion of residuals can be thought of as what is left over in Y that is *not* explained by our model. As residuals get smaller on average, so will the SD of the residuals.

- Sigma (σ) is the SD of residuals. It can be thought of as how much left over in Y that we cannot explain by our model.

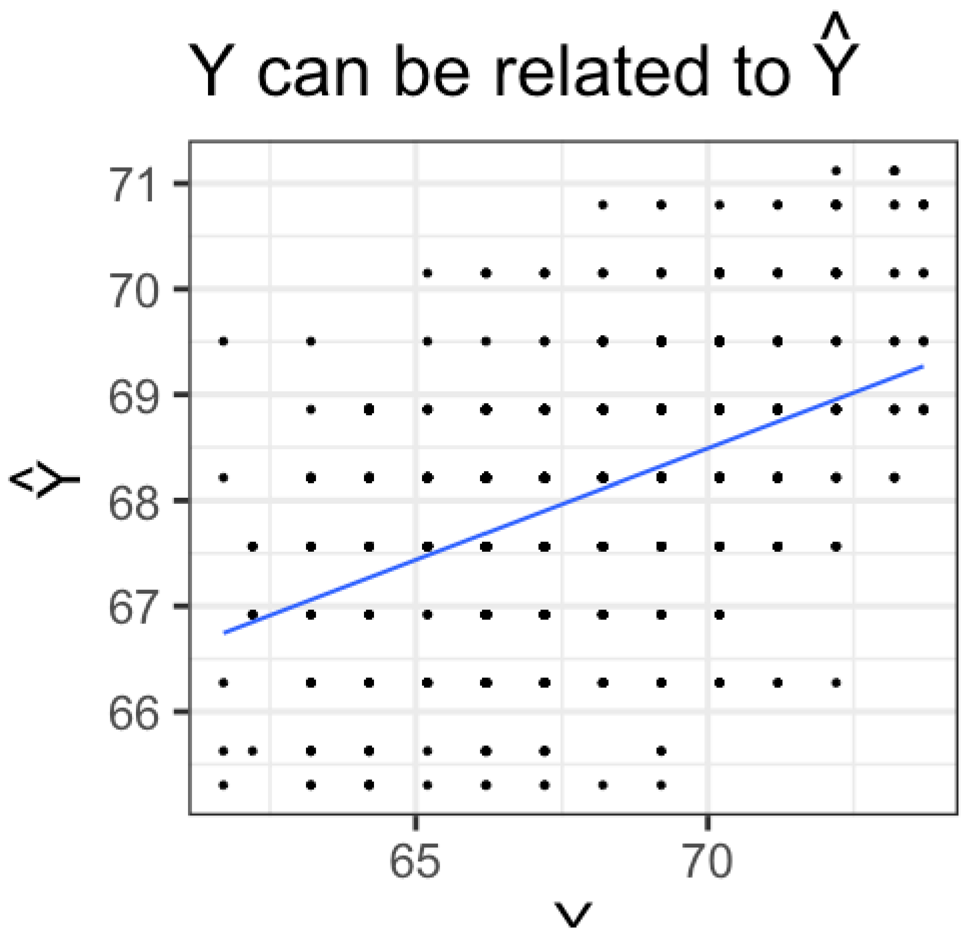
```
model_info %>% ggplot(aes(x = parent, y = .fitted)) +
  geom_point() + geom_smooth(se = F, method = "lm") + ggtitle
  scale_x_continuous("X") + scale_y_continuous(expression(hat
```



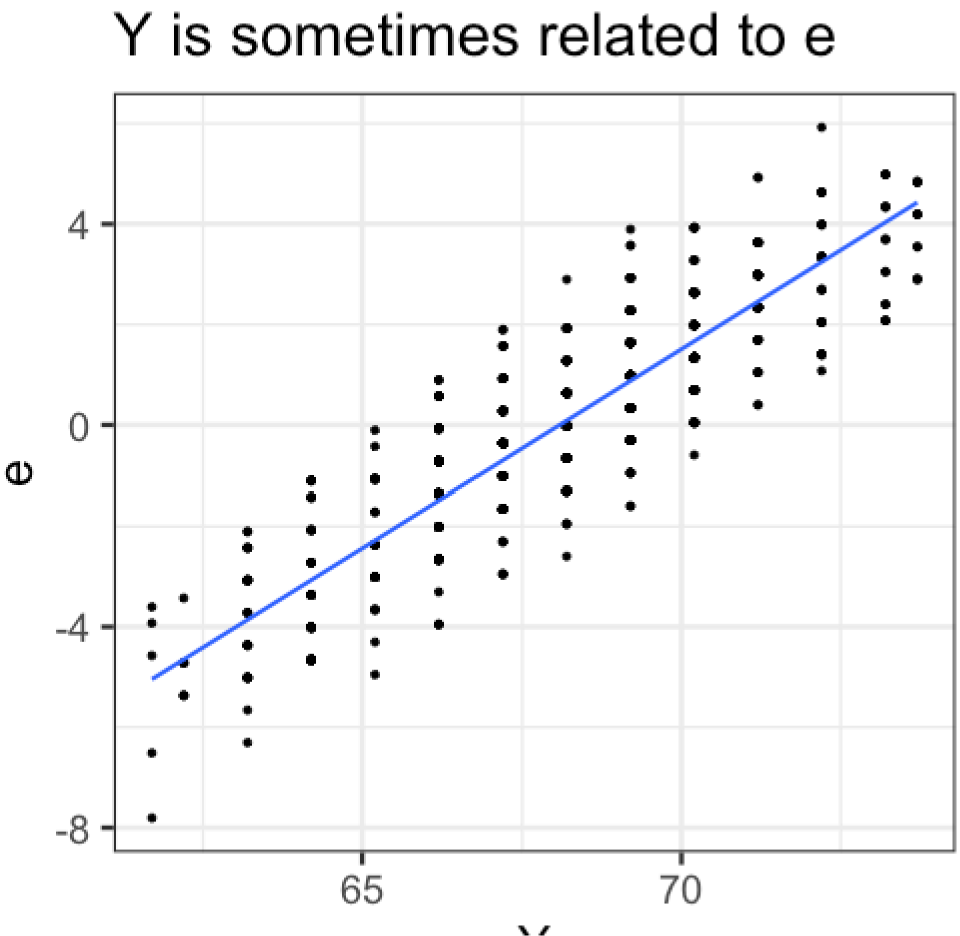
```
model_info %>% ggplot(aes(x = parent, y = .resid)) +
  geom_point() + geom_smooth(se = F, method = "lm") + ggtitle
  scale_x_continuous("X") + scale_y_continuous("e") + theme_b
```



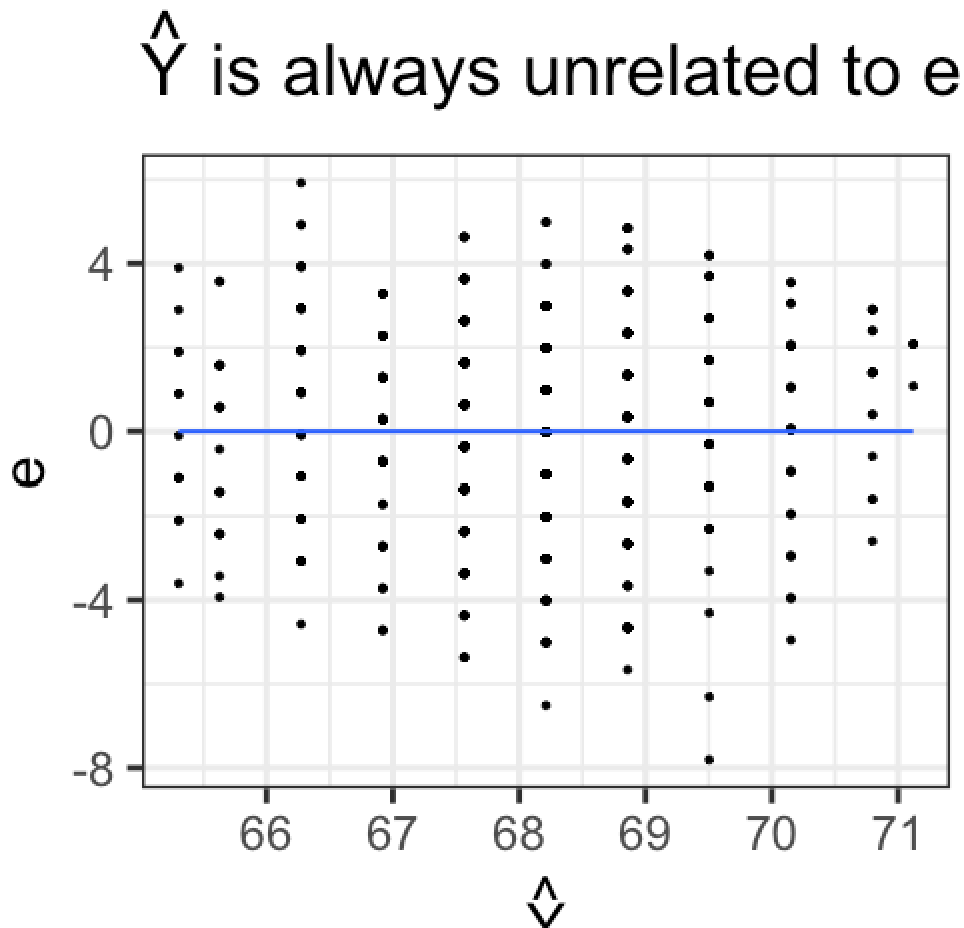
```
model_info %>% ggplot(aes(x = child, y = .fitted)) +
  geom_point() + geom_smooth(se = F, method = "lm") + ggtitle
  scale_x_continuous("Y") + scale_y_continuous(expression(hat
```



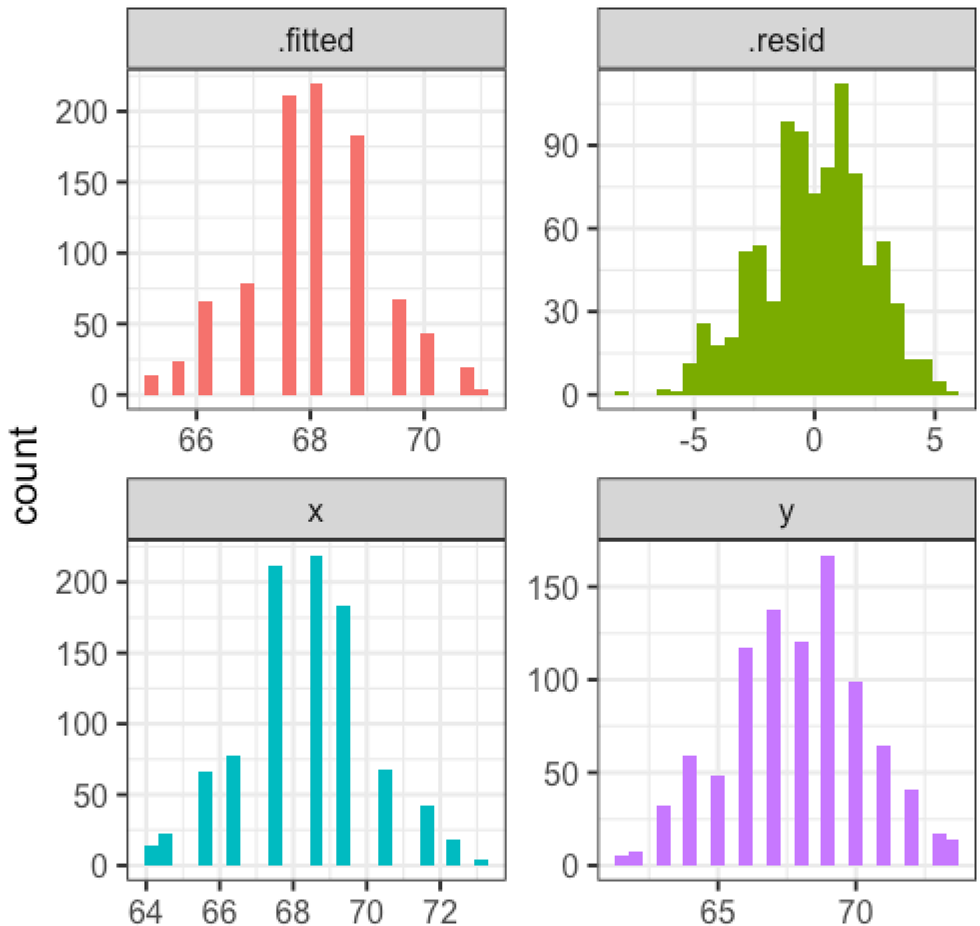
```
model_info %>% ggplot(aes(x = child, y = .resid)) +
  geom_point() + geom_smooth(se = F, method = "lm") + ggtitle
  scale_x_continuous("Y") + scale_y_continuous("e") + theme_b
```



```
model_info %>% ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() + geom_smooth(se = F, method = "lm") + ggtitle
  scale_y_continuous("e") + scale_x_continuous(expression(hat
```



```r
model_info %>% rename(y = child, x = parent) %>% select(x,y,.
  ggplot(aes(value, fill = key)) + geom_histogram(bins = 25)
  facet_wrap(~key, scales = "free") + theme_bw(base_size = 20
```



Residuals Summary

- Residuals are not correlated with X and $\hat{Y}$ because those two are perfectly correlated with one another (that is, $r_{\text{fitted,x}} = 1$)
- X and $\hat{Y}$ represent the *same* information. We use our model (X) to make a prediction ($\hat{Y}$).
- No correlation between residuals with X and $\hat{Y}$ because they are created by subtracting them out of Y. ($\epsilon = Y - \hat{Y}$)
- σ (SD of residuals) can be thought of as how much left over in Y after we take out all of the information our model provides.

Next time...

Statistical inferences with regression