MODELS AND STATISTICAL INFERENCE

A statistical model is a *simplification* and *idealization* of reality that captures our key assumptions about some process underlying our data (the **data generating process** or DGP).

WHY DO WE USE MODELS?

- Making predictions (forecasting)
- Exploration and discovery
- Hypothesis testing

STEPS IN STATISTICAL ANALYSIS

1. Import
2. Transform
3. Visualize
4. Specify
5. Estimate
6. Validate
7. Interpret
8. Report
9. Archive

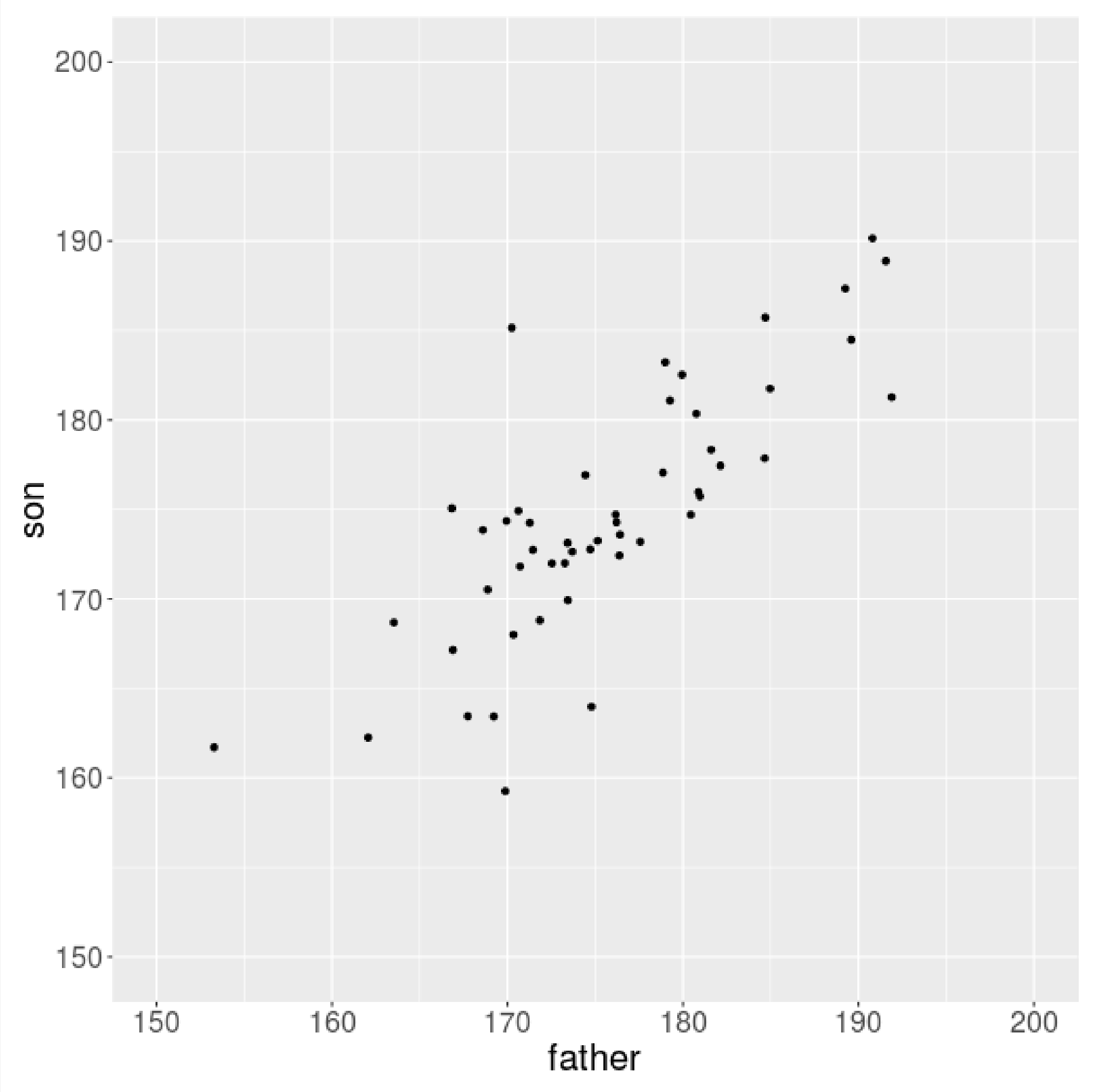
STATISTICAL RECIPES

- t-test
- correlation & regression
- multiple regression
- Analysis of Variance
- mixed-effects modeling

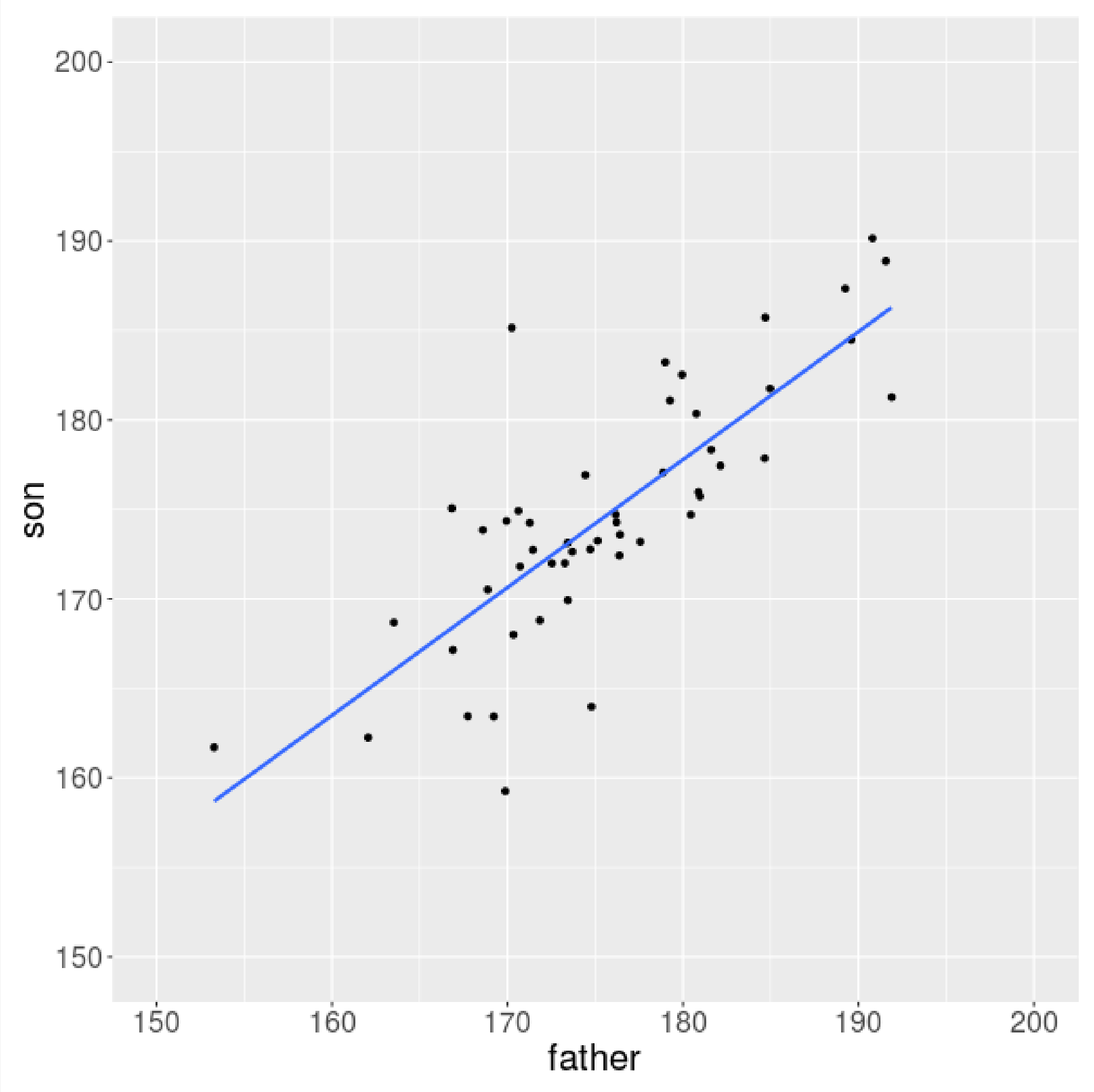
- *All of these are special cases of the General Linear Model (GLM).*

SOME GLM EXAMPLES

regression	$Y_i = \beta_0 + \beta_1 X_i + e_i$
t-test	$Y_i = \mu + A_i + e_i$
one-way ANOVA	$Y_i = \mu + A_i + e_i$
factorial ANOVA	$Y_{ij} = \mu + A_i + B_j + AB_{ij} + e_{ij}$
multiple regression	$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$



father	son
169	174
153	162
179	177
180	175
182	178
173	172
167	167
176	174



father	son
169	174
153	162
179	177
180	175
182	178
173	172
167	167
176	174

SPECIFYING THE MODEL

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

Y_i	response	son's height (observed)
X_i	predictor	father's height (observed)
β_0	intercept	prediction when $X_i = 0$
β_1	slope	increase in Y for each increase in X
e_i	residual	observed minus predicted

$$e_i \sim N(0, \sigma^2)$$

ESTIMATING MODEL PARAMETERS IN R

```r
mod <- lm(son ~ father, hgt)
```

```
> ?lm

lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)

‘lm’ is used to fit linear models.  It can be used to carry out
regression, single stratum analysis of variance and analysis of
covariance (although ‘aov’ may provide a more convenient interface
for these).
```

INTERPRETING R OUTPUT

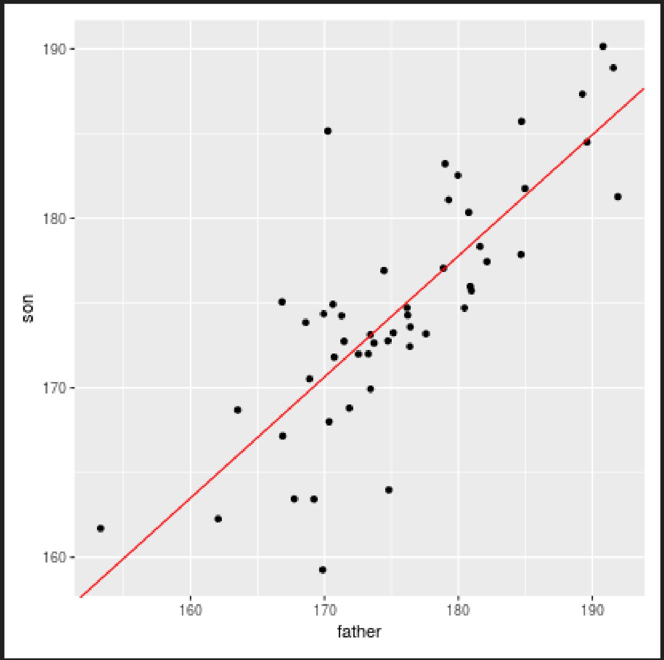
```
summary(mod)
```

```

Call:
lm(formula = son ~ father, data = hgt)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.287  -2.740  -0.395   2.918  14.332 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 49.18917   13.77029   3.572 0.000818 ***
father       0.71441    0.07831   9.122 4.69e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.331 on 48 degrees of freedom
Multiple R-squared:  0.6342,	Adjusted R-squared:  0.6266 
F-statistic: 83.22 on 1 and 48 DF,  p-value: 4.688e-12
```

MODEL VALIDATION

Are the predictions of your model sensible?

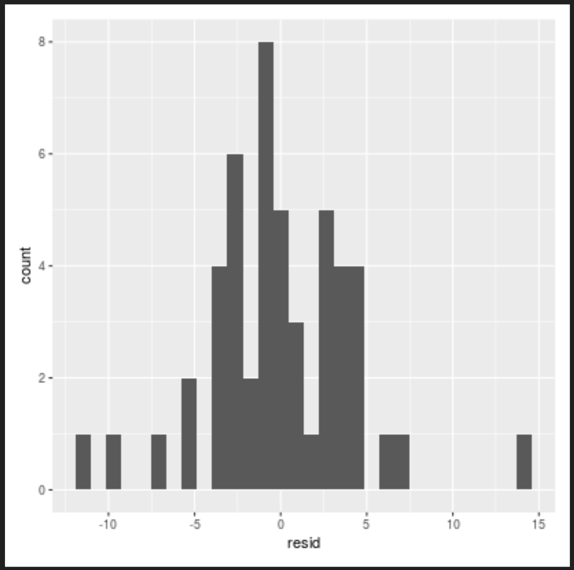
```
ggplot(hgt, aes(father, son)) +
  geom_point() +
  geom_abline(slope = coef(mod)[2],
              intercept = coef(mod)[1],
              color = "red")
```



MODEL VALIDATION

Are the residuals normally distributed?

```r
hgt2 <- mutate(hgt, resid = residuals(mod))

ggplot(hgt2, aes(resid)) +
  geom_histogram()
```



MAKING PREDICTIONS

```
coef(mod)
```

```
(Intercept)      father 
 49.1891719   0.7144094 
```

$$\hat{\beta}_0 = 49.19; \hat{\beta}_1 = 0.71$$
$$\hat{Y}_i = 49.19 + 0.71X_i$$

We want to use the model to predict the heights of boys when they grow up. We have measured their fathers’ heights.

$$\hat{Y}_i = 49.19 + 0.71X_i$$

by "hand"

```
int <- coef(mod)[1]
slp <- coef(mod)[2]

dads <- c(176, 198, 160)

preds <- int + slp * dads
preds
```

```
[1] 174.9252 190.6422 163.4947
```

using `predict()`

```
dad2 <- tibble(father = c(176,
                          198,
                          160))

predict(mod, newdata = dad2)
```

```
       1        2        3 
174.9252 190.6422 163.4947 
```

INTERPRETING AND REPORTING

> *Adult sons' heights were related to their fathers' heights by the formula SON = 49.19 + 0.71 × FATHER. The model explained about 63% of the variance, and the association was significant,* $F(1, 48) = 83.22$, $p < .001$.

RELATIONSHIP BETWEEN CORRELATION AND REGRESSION

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

$$e_i \sim N\left(0, \sigma^2\right)$$

$$\beta_1 = \rho_{xy} \frac{\sigma_y}{\sigma_x}$$

$$\beta_0 = \mu_y - \beta_1 \mu_x$$

IMPLICATIONS

$$Y_i = \beta_0 + \beta_1 X_i + e_i \qquad \beta_1 = \rho_{xy} \frac{\sigma_y}{\sigma_x}$$

$$e_i \sim N\left(0, \sigma^2\right) \qquad \beta_0 = \mu_y - \beta_1 \mu_x$$

- $\beta_1 > 0$ implies $\rho > 0$, since standard deviations can't be negative.
- $\beta_1 < 0$ implies $\rho < 0$, for the same reason.
- Rejecting $H_0 : \beta_1 = 0$ is the same as rejecting $H_0 : \rho = 0$.
    - also, same p-values for β_1 in `lm()` as for r in `cor.test()`.

REGRESSION FROM CORRELATION

A study of student performance obtains a correlation of .16 between final exam score and number of lectures attended. The mean score on the final exam was 70 ($SD = 10$), and the mean number of courses attended was 6 ($SD = 2$).

Write the regression equation predicting exam score from attendance.

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

$$e_i \sim N\left(0, \sigma^2\right)$$

$$\beta_1 = \rho_{xy} \frac{\sigma_y}{\sigma_x}$$

$$\beta_0 = \mu_y - \beta_1 \mu_x$$

CORRELATION FROM REGRESSION

A study on the relationship between wellbeing and hours spent on social media (per week) yields the following regression:

wellbeing = 62 - .3 hours

with 5 for the standard deviation of wellbeing and .1 for the standard deviation of number of hours.

What is the correlation?

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

$$e_i \sim N\left(0, \sigma^2\right)$$

$$\beta_1 = \rho_{xy} \frac{\sigma_y}{\sigma_x}$$

$$\beta_0 = \mu_y - \beta_1 \mu_x$$