Today We Regress

Karl Pearson

not this guy

Sigmund Freud

this guy

not this guy
But First We Correlate

- Values are correlated when big values of X go with big values of Y (and small values with small values) - positive correlation
  -or-
- When big values of X go with small values of Y (and vice versa) - negative correlation
- Correlation is always a number between -1 and +1.

![Graph showing positive, negative, and no correlation with correlation coefficients r = 0.4, r = 0, and r = -0.4 respectively.](image-url)
Why Are Variables Correlated?

Remember this! It’s important!

- **Spurious relationship**
- **Direct causal relationship**
- **Mediated causal relationship**
- **Common cause**

---

**Why Correlation Does Not Imply Causation**

**Someone sent me another anonymous email with a link to an article about the world’s worst bosses.**

**I get one of those emails every time I leave your cubicle. Did you think I wouldn’t notice the correlation?**

**Correlation does not imply causation.**

**I used to think correlation implied causation.**

**Then I took a statistics class. Now I don’t.**

**Sounds like the class helped. Well, maybe.**
Correlation does not necessarily imply causation.
Correlation Example

These tests can be directional.

95% CI for population correlation

scatterplot
Problems With Correlation

- Correlation (Pearson correlation) works well as a measure of relationship between two numeric variables only when the relationship between them is linear.

- A scatterplot will often reveal problems.
Regression Analysis

done with the “linear model” function in R

response (or criterion) variable - DV
explanatory (or predictor) variable - IV

> lm(son ~ father, data=KP)

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

Coefficients:
  (Intercept)    father
     33.893       0.514

intercept          slope
____________________
regression coefficients

regression equation

\[ \hat{Y} = 33.893 + 0.514 \times X \]

predicted height for son

height of father
Using the Regression Equation

- for description
- for prediction
- a father is 6 ft. tall - what is the predicted height of his son?

\[
\hat{Y} = 33.893 + 0.514 \times X
\]

\[
= 33.893 + 0.514 \times 72
\]

\[
= 70.901
\]

Notice the son is predicted to be not quite as tall as the father. “regression towards the mean”
Another way to think of the regression line is that it gives the (estimated) mean height of sons for fathers of a certain height (the descriptive value of the regression equation).

Question: Is it possible to use the regression equation or regression line to predict $X$ from $Y$?
Testing Regression Coefficients

What are residuals?

$t$-tests on coefficients
null: true coef = 0

Standard error of estimate

$R^2$

F-test on complete model

(Also for the time being, the same as the test on the correlation coefficient.)
Confidence Intervals

for regression line

(hey! everybody makes mistakes!)

Thanks to [https://rpubs.com/Bio-Geek/71339](https://rpubs.com/Bio-Geek/71339) for this method.

(from a smaller dataset)

for coefficients
Regression Diagnostics

- testing the assumptions
- linear relationship (residuals plot)
- normal distribution of residuals (normal probability plot)
- homogeneity of variance or homoscedasticity
- no influential cases (Cook’s distance)
Another Example: Scottish Hill Racing

Scottish hill races are long-distance foot races that have to be run up hill. The variables: dist = distance of race in miles, climb = height of hill that must be ascended in feet, time = record time for the race in minutes (old data as of about 1975; there is a website if you want newer values).

(https://scottishhillracing.co.uk/)
Relationship Between Time and Distance

By the way, we say “time is being regressed on distance” or “time by distance”.

\[ \text{time-hat} = -4.8407 + 8.3305 \times \text{dist} \]

\[ R^2 = 0.85 \]

(proportion of explained variability)

---

n = 35 races

"All models are wrong, but some are useful."
-George Box
Regression Diagnostics

```r
> par(mfrow=c(2,2))
> plot(lm.out, 1:4)
> ## three flagged cases ##
> hills["Bens of Jura",]
> Bens of Jura
> dist climb time
> Lairig Ghru 16 7500 204.617
> hills["Lairig Ghru",]
> Lairig Ghru 28 2100 192.667
> hills["Knock Hill",]
> Knock Hill 3 350 78.65
```
Multiple Regression

> lm.out = lm(time ~ dist + climb, data=hills)  # additive model; order doesn't matter
> summary(lm.out)

Call:
  lm(formula = time ~ dist + climb, data = hills)

Residuals:
     Min      1Q  Median      3Q     Max
-16.215  -7.129  -1.186   2.371  65.121

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.992039   4.302734  -2.090  0.0447 *
dist        6.217956   0.601148  10.343  9.86e-12 ***
climb       0.011048   0.002051   5.387  6.45e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.68 on 32 degrees of freedom
Multiple R-squared:  0.9191,  Adjusted R-squared:  0.914
F-statistic: 181.7 on 2 and 32 DF,  p-value: < 2.2e-16

An ominous portent of the future: this is called standard multiple regression or “simultaneous regression”. Every term is treated as if entered last. (i.e., everything is controlled for everything else, like Type III sums of squares!)

Interpret!
Diagnostics For This Model

- Residuals vs Fitted
- Normal Q-Q
- Scale-Location
- Cook's distance
Multiple Regression with Interaction

```r
> lm.out = lm(time ~ dist * climb, data=hills)
> summary(lm.out)
```

```
Call:
  lm(formula = time ~ dist * climb, data = hills)

Residuals:
     Min      1Q  Median      3Q     Max
-25.994  -4.968  -2.220   2.381  56.115

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              9.3954374  6.8790233  1.366  0.18183
dist                      4.1489201  0.8352489  4.967 2.36e-05 ***
climb                    -0.0009710  0.0041648 -0.233  0.81718
dist:climb                0.0009831  0.0003070  3.203  0.00314 **

---
Signif. codes:  
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.92 on 31 degrees of freedom
Multiple R-squared:  0.9392,   Adjusted R-squared:  0.9333
F-statistic: 159.6 on 3 and 31 DF,  p-value: < 2.2e-16
```
sometimes called a modifier effect moderator
Interaction means as the value of one variable changes, the slope ("effect") of the other variable changes (is modified or moderated).

\[
time-hat = 9.395 + 4.149*\text{dist} - 0.001*\text{climb} + 0.001*\text{dist*climb}
\]

- if climb = 0: \( \text{time-hat} = 9.395 + 4.149*\text{dist} \)
- if climb = 500 ft: \( \text{time-hat} = 8.895 + 4.649*\text{dist} \)
- if climb = 1000 ft: \( \text{time-hat} = 8.395 + 5.149*\text{dist} \)
- if climb = 2000 ft: \( \text{time-hat} = 7.395 + 6.149*\text{dist} \)
Third Variable Effects

- mediator - carries or transmits effect of IV
- moderator - changes effect of IV
- confounder - obscures effect of IV
- covariate - something else that affects the DV
- suppressor - to be defined later!