Week 5: Simple linear regression as an intro to general linear models

ANTH 674: Research Design & Analysis in Anthropology
Professor Andrew Du
Andrew.Du2@colostate.edu

Office Hours: Thursdays, 9:00am–12:00pm
In person: GSB 312
Virtual: https://tinyurl.com/F22ANTH674

1

Statistical vignette

What do these two have in common?

The Madden Curse

Curse "record": 24–0
https://www.wyexpect.com/stories/is-the-madden-curse-real

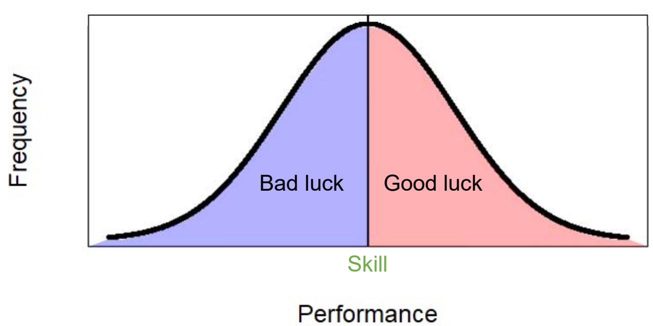
Training Israeli Air Force (1960s)

Praise → worse performance
Scold → better performance

2

Regression to the mean

Performance = skill + luck



3

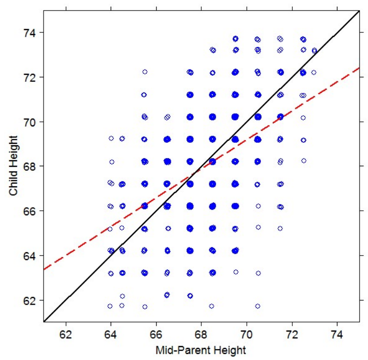
Cf. central limit theorem

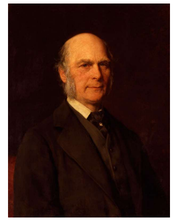
https://www.youtube.com/watch?v=6YDHBFVIvIs

4

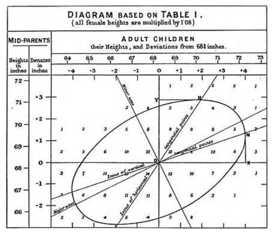
Galton & regression

Regression to the mean





Sir Francis Galton

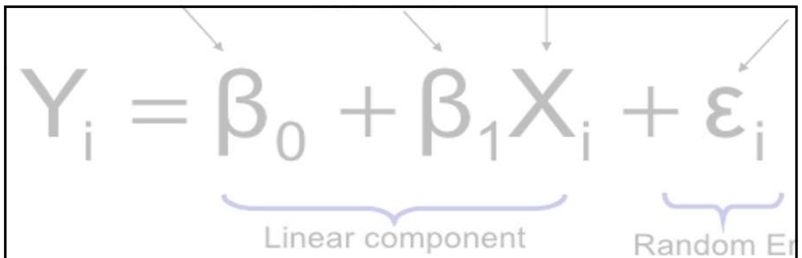


Original 1886 figure

https://www.youtube.com/watch?v=lOWjRaCKbQg

Lecture outline

1. Quick intro to general linear models
2. Simple linear regression
   1. What is it? What does it do?
   2. Using transformed variables
   3. Goals of regression
   4. Assumptions
   5. Diagnostics to assess validity of model
3. Correlation coefficients



Intro to general linear models

What are they? What are they used for?

What is a model?

- What do you think of when someone says “model” in data analysis?

$$\frac{dLs}{dt} = N * K_1 - (d_1 + r) * Ls + \delta_1$$

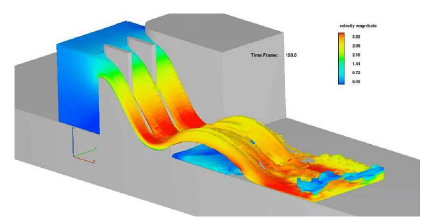
$$\frac{dLux}{dt} = \left(\frac{K_2}{1 + A * Ls1^2}\right) * N - (d_2 + r) * Lux + \delta_2$$

$$Z = Z_1 + Z_2$$

$$Z_1 = K_3 * Lux$$

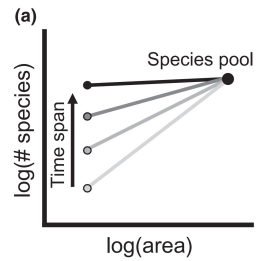
$$Z_2 = \iint_{0,0}^{2\pi,D} e^{-K_4 s} * Z_{i,j} ds$$

$$\frac{dLs1}{dt} = Ls * \left(1 - \frac{Ls}{K5}\right) * Z * \left(1 - \frac{Z}{K6}\right) \qquad (1)$$



What is a model?

- A model is any description of how the natural world might work
- Can be verbal description, graphs, equations, computer simulations, and many more!



9

What is a model?

- A model is any description of how the natural world might work
- Can be verbal description, graphical, equations, computer simulations
- In statistics, we model one variable (*dependent/response* variable) as a function of another (*independent/predictor* variable)
- IV gets input into model and get an output (DV)

IV (input) → Model → DV (output)

10

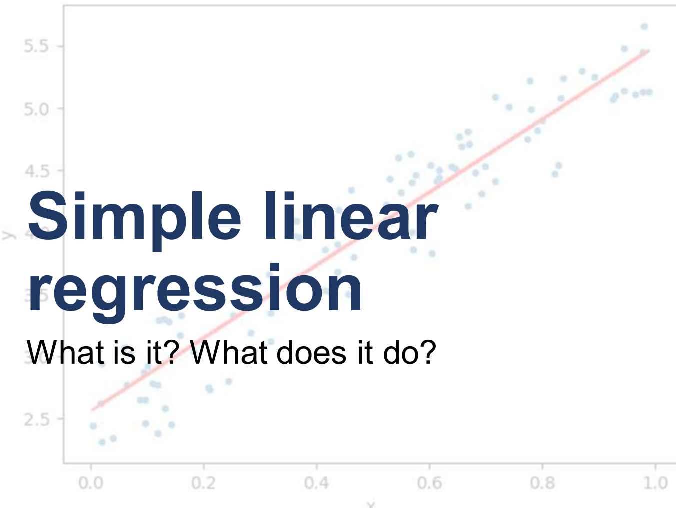
What are general linear models?

- Models DV as a *linear/additive* function of one or more IV
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots$
- Dependent & independent variables can be continuous/discrete/ordinal/categorical
- t-tests, ANOVAs, linear regression, logistic regression, and others are all GLMs
- Will introduce GLMs with simple linear regression

11

Questions?

12



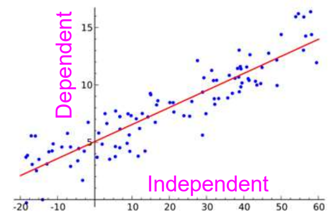
Simple linear regression

What is it? What does it do?

13

What is simple linear regression?

- Models one continuous DV as a linear function of one continuous IV
- E.g., how does femur length increase as a function of body size?
- Also known as a “linear model”



14

The linear regression equation

Dependent variable

Independent variable

$$Y = \beta_0 + \beta_1 X_1$$

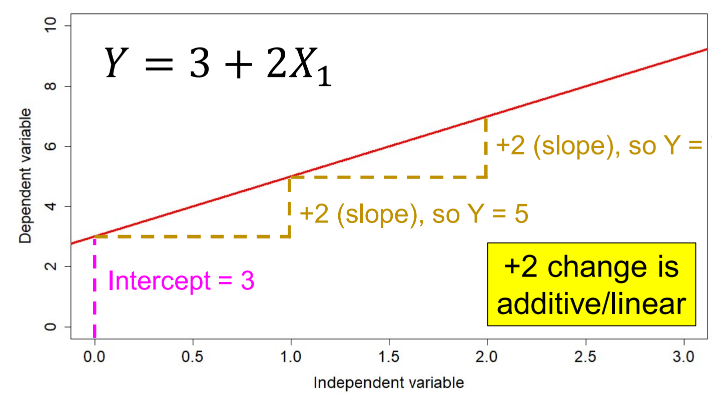
Intercept

Slope

- Intercept: Value of DV when IV = 0 (in units of DV)
  - $Y = \beta_0 + \beta_1 \times 0 \rightarrow Y = \beta_0$
- Slope: Change in DV when IV increases by 1
  - $Y = \beta_0 + \beta_1 X_1$
  - $Y = \beta_0 + \beta_1 (X_1 + 1) \rightarrow Y = \beta_0 + \beta_1 X_1 + \beta_1$

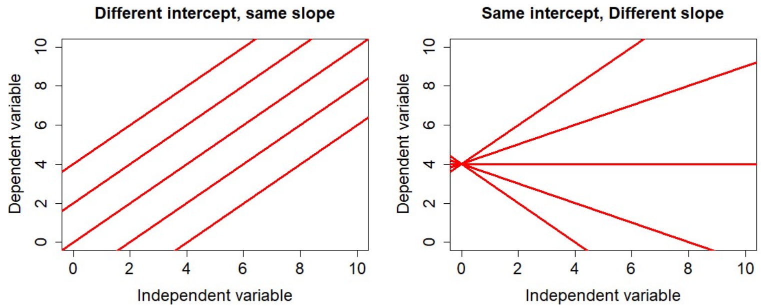
15

The linear regression equation



16

Intercept and slope



17

Estimating parameters

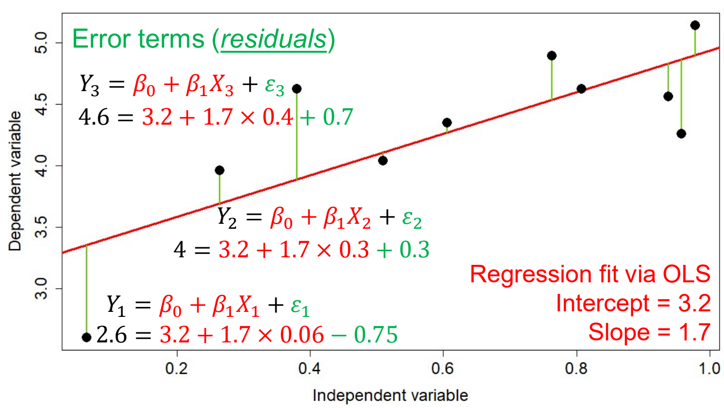
- The intercept and slope are *parameters,* population unknowns estimated from the data
- Estimated parameters in regression are also known as *coefficients*
- Parameters are estimated using the *ordinary least squares* method
- But first, let's slightly modify our regression equation, so it applies to data:

The error term

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

18

Fitting regression to data



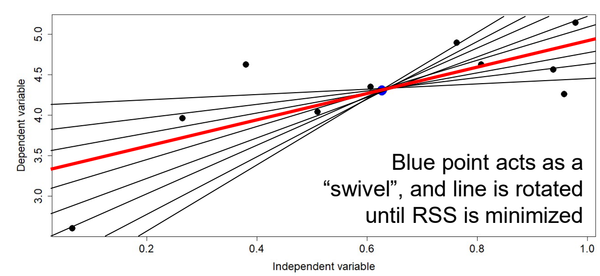
https://shiney.zoology.ubc.ca/whitlock/Residuals/

19

https://seeing-theory.brown.edu/regression-analysis/index.html#section1

Ordinary least squares

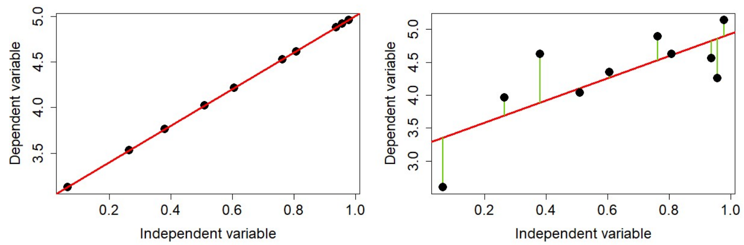
- Fit by minimizing residuals, specifically the *residual sum of squares* ($\sum \varepsilon_i^2$)
- OLS line *must* go through mean of DV and mean of IV (blue point)



20

How to interpret residuals?

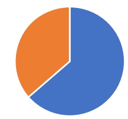
- Signal in DV not accounted for by IV
- Extra noise due to unmeasured factors
- E.g., if DV = femur length, IV = body size, perhaps points below line are a different species with shorter legs

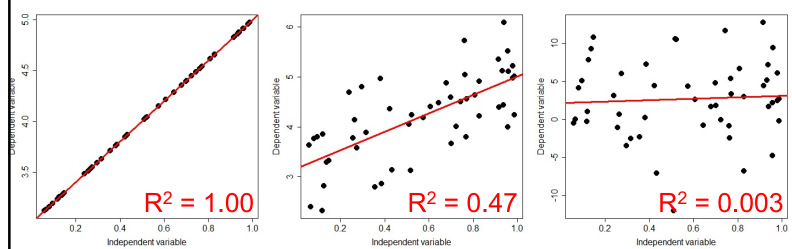


Coefficient of determination (R^2)

- Proportion of variation in DV attributed to IV

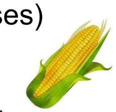
$$\frac{SS_{reg}}{SS_{reg} + RSS}$$





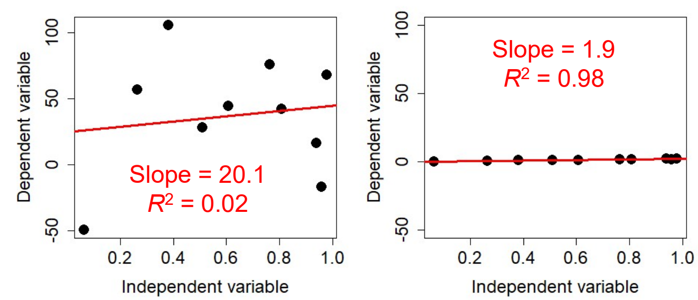
Effect size & goodness of fit

- *Effect size* (measure of magnitude of a pattern)
- Slope (how quickly DV changes as IV increases)
  - E.g., how much your crop yield increases as a function of fertilizer amount
- *Goodness of fit* (how well model fits the data)
- R^2 (how much variation in DV attributed to IV)
  - E.g., how much variation in crop yield is attributed to fertilizer amount → how predictable is crop yield as a function of fertilizer amount)

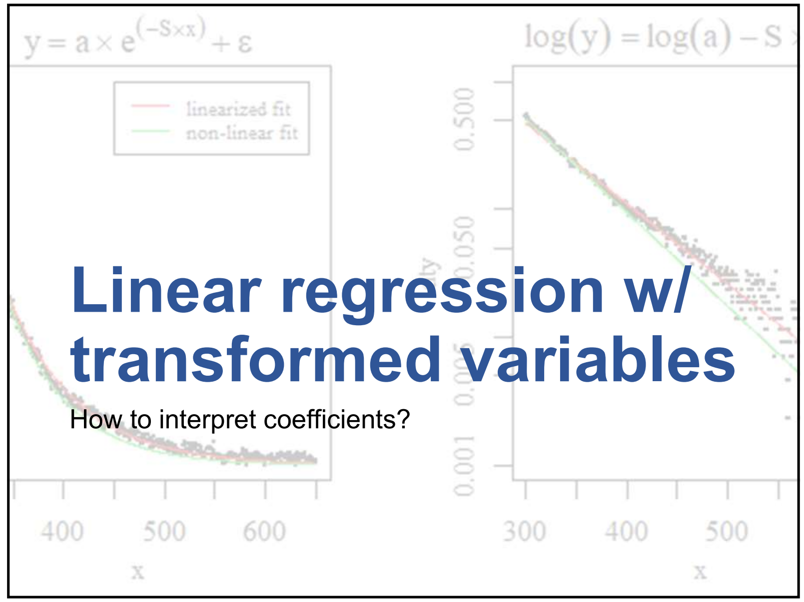


Effect size & goodness of fit

- Theoretically independent: can have large slopes and small R^2, and vice versa



Questions?



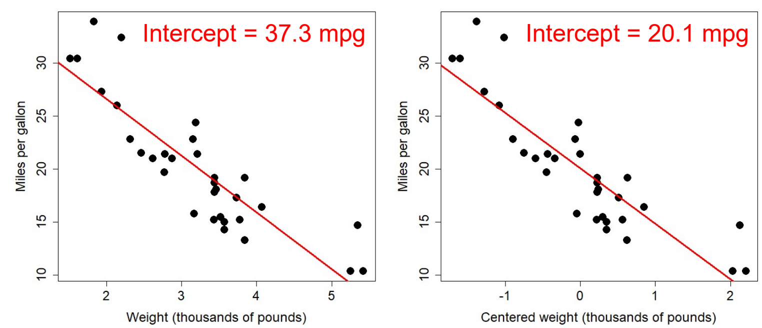
Linear regression w/ transformed variables

How to interpret coefficients?

Centering (mean → 0)

- Many times, the interpretation of an intercept is meaningless
- E.g., if IV is `mtcars$wt` and DV is `mtcars$mpg`, what does it mean to have a certain `mpg` when `wt` is zero?
- Can center IV to mean = 0, so now intercept is interpreted as expected DV for mean IV

Centering (mean → 0)

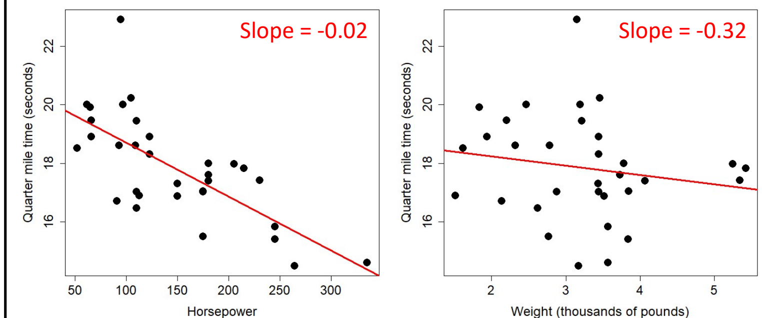


Scaling (SD → 1)

- Scaling transforms variables to have SD = 1
- Useful for comparing variables measured in different units or if they differ by orders of magnitude
- Usually used when comparing slopes from different regressions
- E.g., if DV is `mtcars$qsec` (speed), I want to know if `mtcars$hp` or `mtcars$wt` (IVs) has a bigger effect

29

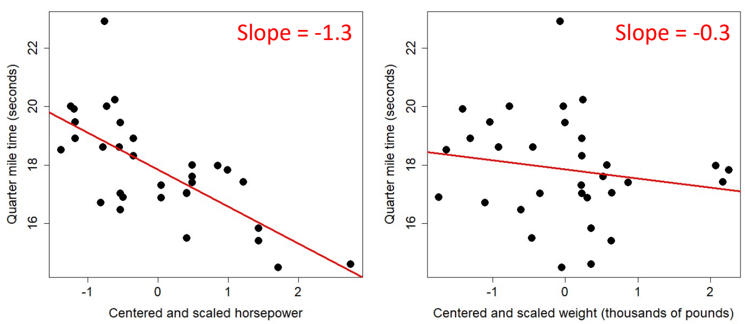
Scaling (SD → 1)



30

Scaling (SD → 1)

- Have to center also in R!



31

Log-transformations

- log2-transformations → one unit increase = one doubling
  - E.g., 1 = log2(100) – log2(50)
- log10-transformations → one unit increase = one order of magnitude increase
  - E.g., 1 = log10(1000) – log10(100)
- In general, how one interprets change in DV and/or IV for slope (more difficult for natural log)

32

Log-transformed DV
$\log(Y) = \beta_0 + \beta_1 X$

- Intercept is $\log(Y)$ when $X = 0$
- If antilog of slope is taken, it is interpreted as the proportional change in unlogged Y as X increases by 1
- E.g., if estimated slope is 0.69 (natural log), then antilog is 2, which means unlogged Y doubles every time X increases by 1
- **Works for all log-transformations!**

33

Log-transformed IV
$Y = \beta_0 + \beta_1 \log(X)$

- Intercept is Y when $\log(X) = 0$
- 1% increase in unlogged X → approximate $\beta_1/100$ change in Y
- Z% increase in unlogged X → exact $\beta_1 \times \log(1.Z)$ change in Y
  - E.g., 10% increase in unlogged $X \rightarrow Y$ changes by $\beta_1 \times \log(1.1)$ exactly
- **Slope interpretations work for natural log only!**

34

Log-transformed IV & DV
$\log(Y) = \beta_0 + \beta_1 \log(X)$

- Intercept is $\log(Y)$ when $\log(X) = 0$
- β_1 is approx. % change in unlogged Y for every 1% increase in unlogged X
- For a $Z\%$ increase in unlogged X, unlogged Y changes approx. by a percentage equal to $(1.Z^{\beta_1} - 1) \times 100$
- E.g., a 50% increase in X results in an approx. $(1.5^{\beta_1} - 1) \times 100$ % increase in Y
- **Slope interpretations work with natural log only!**

35

Questions?

36

Goals of regression

What is regression used for?

Three different goals of regression

1. Exploration
2. Testing null hypotheses
3. Prediction

1. Exploration

- Just want to know what the intercept and slope is
- E.g., at what rate does femur length increase with body size (slope)?
- Use OLS to estimate parameters

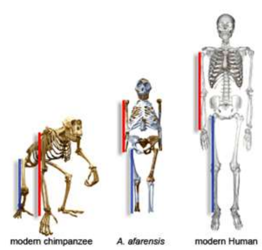


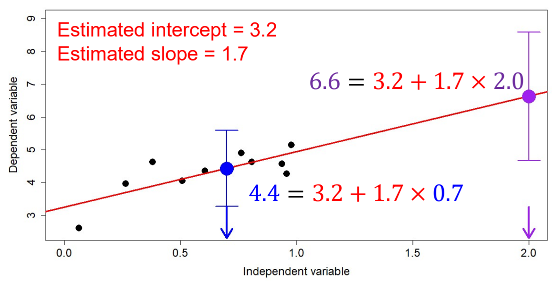
Illustration of comparative limb proportions. Red lines represent the forelimbs (arms), while blue lines represent the lower limbs (legs).

2. Testing null hypotheses

- Do samples of intercept and slope come from populations where these parameters equal zero?
- What are the 95% CI and P-values for these two estimated parameters?
- Easily done in R with `lm()` and `confint()` functions

3. Prediction

- Want to know predicted DV value, corresponding to IV value not in your data
- E.g., predict femur length using body mass for an individual that has no femur preserved



Interpolation
Extrapolation

41

Questions?

42

Assumptions of linear regression

What are they? How do they affect results?

43

Linear regression assumptions

- Like all models, linear regression has assumptions
- Assumptions allow us to simplify reality and bring data into the realm of logic and math
- Violations of assumptions affect results in different ways
- So a violation(s) does not mean your results are automatically meaningless!

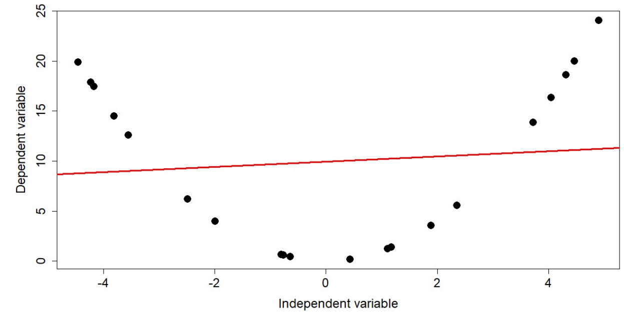
44

Linear regression assumptions

1. Relationship between DV and IV is linear
2. IV measured without error
3. Error terms have mean = 0 and are normally distributed
4. Error terms drawn from population with the same variance
5. Error terms are independent

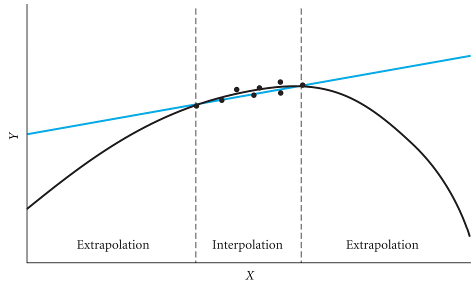
1. Relationship is linear

- Otherwise, intercept and slope are meaningless
- And predicted DVs are meaningless



1. Relationship is linear

- Can transform variables to linearize relationship or fit a different model
- Or focus on linear part of relationship only

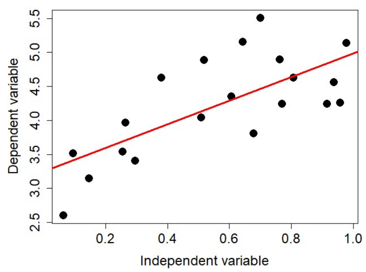


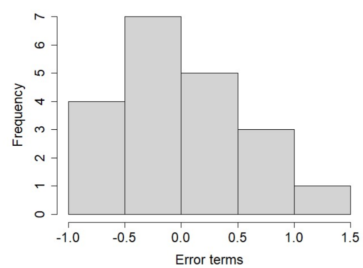
2. IV is measured w/o error

- Error is assumed to be wholly due to DV, so error in IV is not good
- This assumption is rarely not violated and is usually ignored (e.g., I often ignore it)

3. Errors mean = 0 & normal

- This assumption is necessary for robust CI and P-values
- Transforming the DV can normalize errors or need to fit another model (e.g., Monte Carlo)

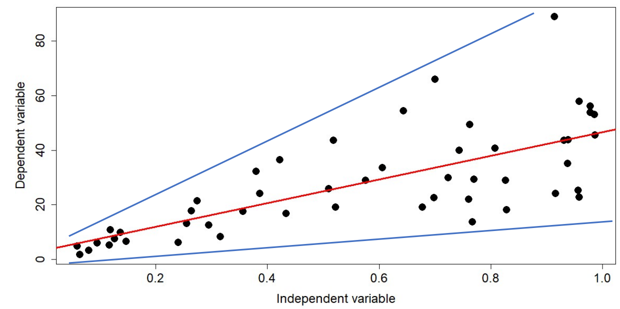




49

4. Error terms have constant variance

- Violation of assumption is known as *heteroscedasticity*



50

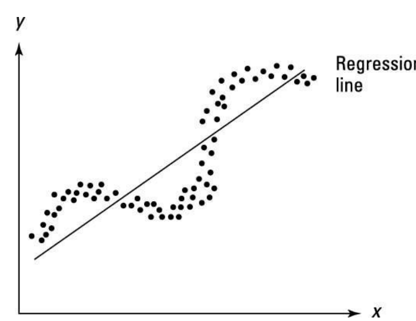
4. Error terms have constant variance

- Violation of assumption is known as *heteroscedasticity*
- Affects P-values & CI, but coefficient estimates are *unbiased* (hits the true value on average)
- Can transform DV, include missing IV, calculate robust standard errors, or need a different model (e.g., weighted least squares)

51

5. Error terms are independent

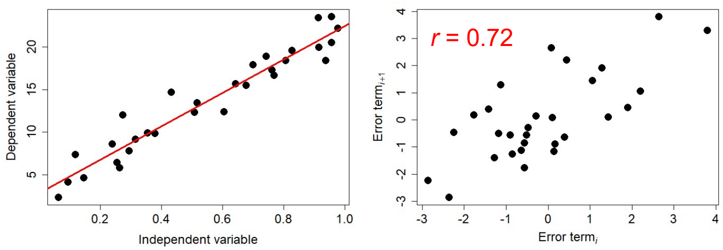
- Value of one error term is not a function of another (i.e., error terms are uncorrelated)



52

5. Error terms are independent

- Value of one error term is not a function of another (i.e., error terms are uncorrelated)



53

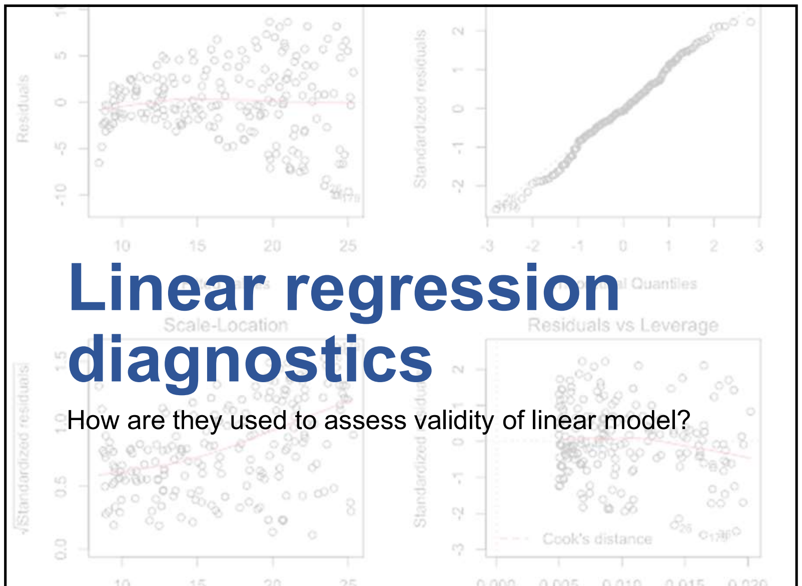
5. Error terms are independent

- Value of one error term is not a function of another (i.e., error terms are uncorrelated)
- Violated w/ spatial autocorrelation, temporal autocorrelation, phylogenetic autocorrelation
- P-values and CI are too small, but coefficients are *unbiased*
- Need to add IV to account for autocorrelation or use another model (e.g., generalized least squares)

54

Questions?

55



Linear regression diagnostics

How are they used to assess validity of linear model?

56

https://shiney.zoology.ubc.ca/whitlock/Residuals/ (2nd tab)

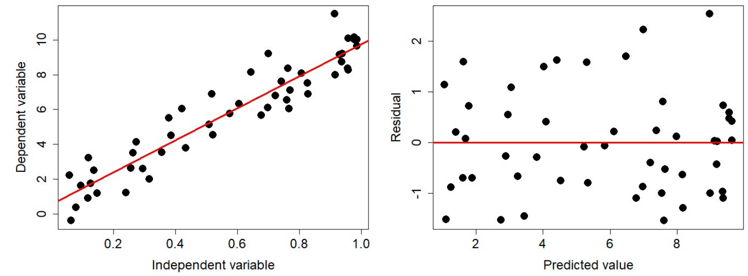
Regression diagnostics

- Assesses whether assumptions are grossly violated
- Most commonly done visually with *residual plots* (plots of residuals as a function of predicted values from the linear regression)
- Easily done in R w/ `plot(lm(y ~ x))`

57

Residual plots

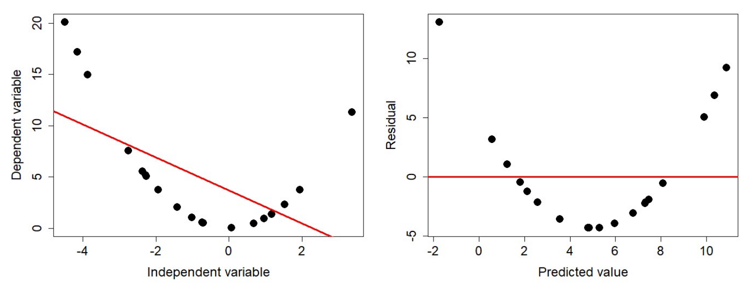
- A good model fit has residuals showing a horizontal band of randomly distributed points surrounding zero on the Y-axis



58

Residual plots

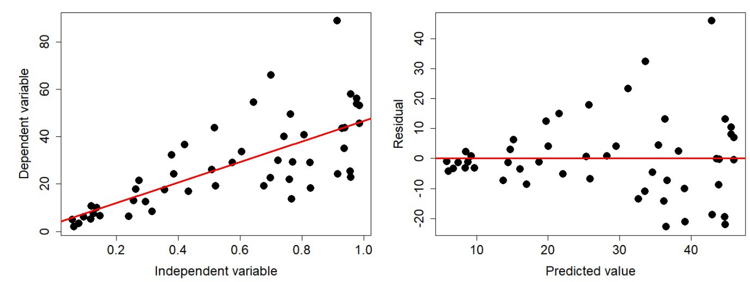
- Non-linear relationship between DV and IV



59

Residual plots

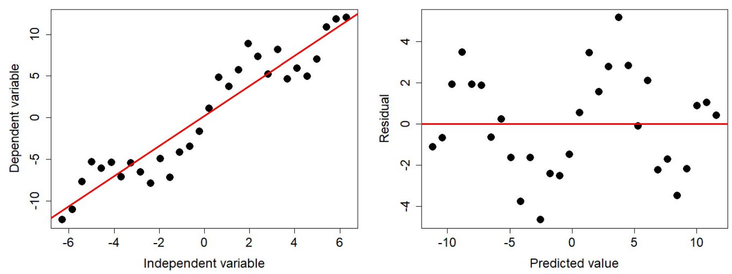
- Heteroscedasticity



60

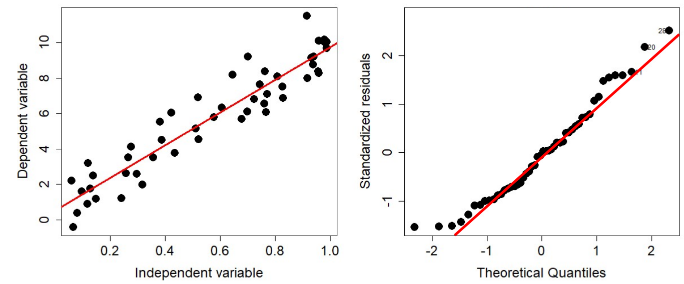
Residual plots

- Non-independent errors



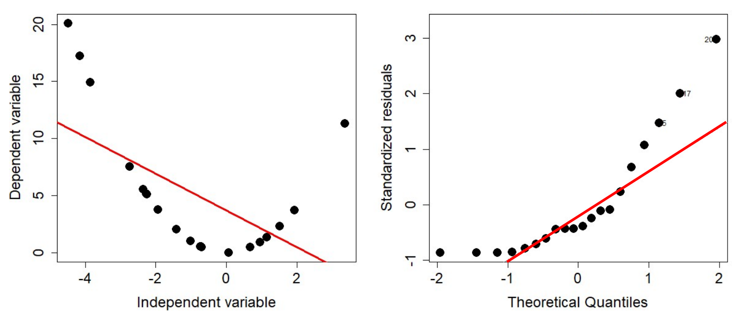
Normal Q-Q plot

- Used to assess normality of errors
- Can also plot a histogram

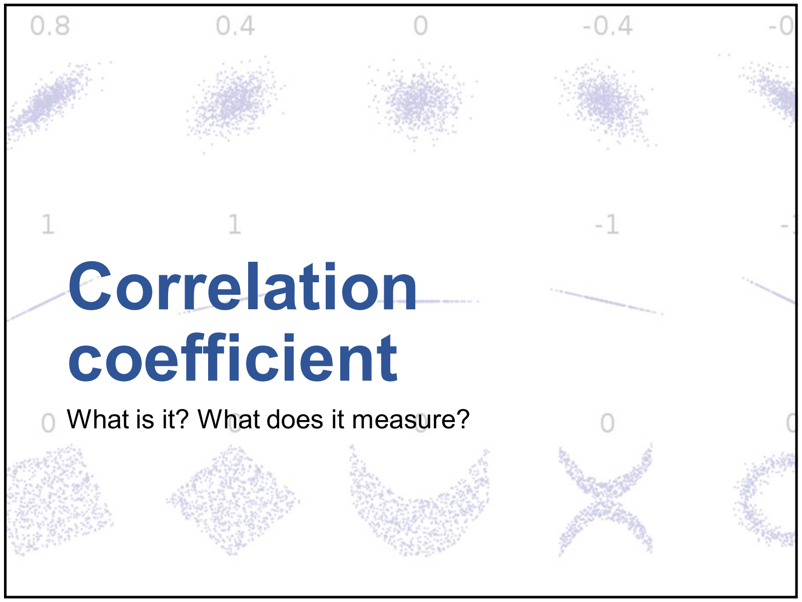


Normal Q-Q plot

- What it looks like w/ a non-linear relationship



Questions?



Correlation coefficient

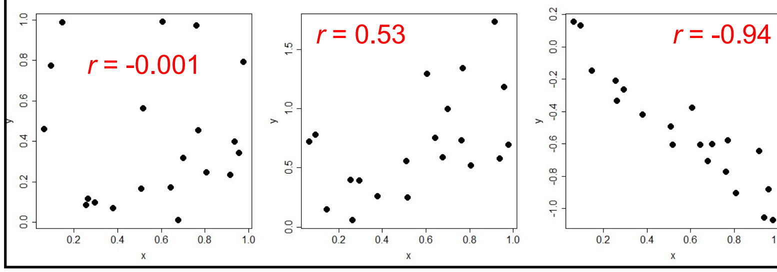
What is it? What does it measure?

65

https://shiney.zoology.ubc.ca/whitlock/Guessing_correlation/

Correlation coefficient

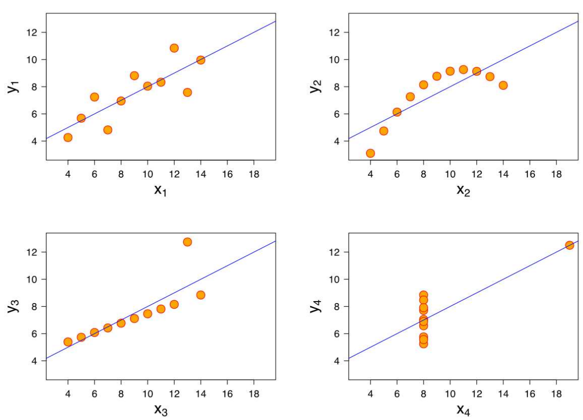
- Measures how tightly two variables covary & the direction (ranges from -1 to 1)
- Most common measure is Pearson's correlation coefficient (*r*) → linear correlation



66

Anscombe's quartet

- r = 0.816 for <u>ALL</u> plots
- <u>**ALWAYS**</u> plot your data!



67

Relationship w/ other measures

- Is also the square-root of coefficient of determination (R^2)!
- Is also the standardized slope of a linear regression (DV and IV centered and scaled)!

68

Null hypothesis test

- Does sample's *r* come from population where *r* equals zero?
- What are the 95% CI and P-value of estimated *r* ?
- Easily done in R with `cor.test()`

69

Correlation vs. linear regression

- Often said that linear regression assumes a causal, directional relationship: IV → DV
- And that correlation doesn't care about such directions
- My view: linear regression doesn't necessarily imply causation; just describes rate of DV change w/ increase in IV (slope)
- If interested in slope (or predicting DV), use linear regression; if interested in how tightly two variables covary, use correlation

70

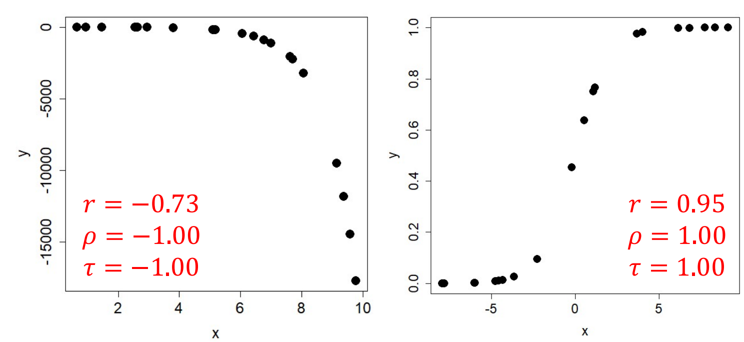
Non-parametric alternatives

- What if interested in "tightness" of *non-linear*, monotonic relationship?
1. Spearman's rho (ρ)
   - Transforms variables into ranks and calculates *r*
   - {4.4, 9.0, 3.2}, {0.8, 8.2, 9.0} → {2, 3, 1}, {1, 2, 3}
2. Kendall's tau (τ)
   - Interpreted roughly as probability that ranks of variables correspond
- These measures are less sensitive to outliers compared to Pearson's *r*

71

Non-parametric alternatives

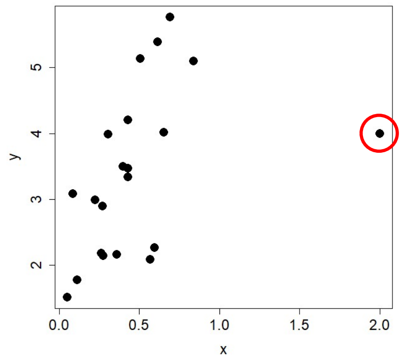
- Non-linear, monotonic relationships



72

Non-parametric alternatives

- Robust to outliers



All data

$r = 0.44$
$\rho = 0.68$
$\tau = 0.49$

W/o outlier

$r = 0.67$
$\rho = 0.67$
$\tau = 0.49$

73

Spearman's or Kendall's?

- Kendall's has agreed upon formula for standard error → more robust CI and P-values, especially with smaller sample sizes
- Spearman's is more appropriate when there is less certainty about the reliability of close ranks
- Spearman's is more popular
- Both usually lead to the same inference & conclusions

74

Questions?

75

Summary

- Linear regression is a general linear model where IV and DV are continuous
- Regression gives us an estimated intercept, slope (effect size), R^2 (goodness of fit), & P-value (null hypothesis test)
- Can transform variables to make coefficients more interpretable and/or to satisfy model assumptions
- Three different goals: exploration, hypothesis test, prediction
- Violations of model assumptions affect results in different ways

76