R Stats Bootcamp

2.10 - Regression

Megan Lewis

2025-03-06

R stats bootcamp - Module 2

Schedule:

- ~~Session 7: Explore data~~
- ~~Session 8: Distributions~~
- ~~Session 9: Correlation~~
- **Session 10: Regression**
- Session 11: T-test
- Session 12: ANOVA

R Stats Bootcamp

We should be suspicious if the data points all fall exactly on the straight line of prediction

Session 9 objectives:

- The question of simple regression
- Data and assumptions
- Graphing
- Tests and alternatives
- Practice exercises

Regression to the mean

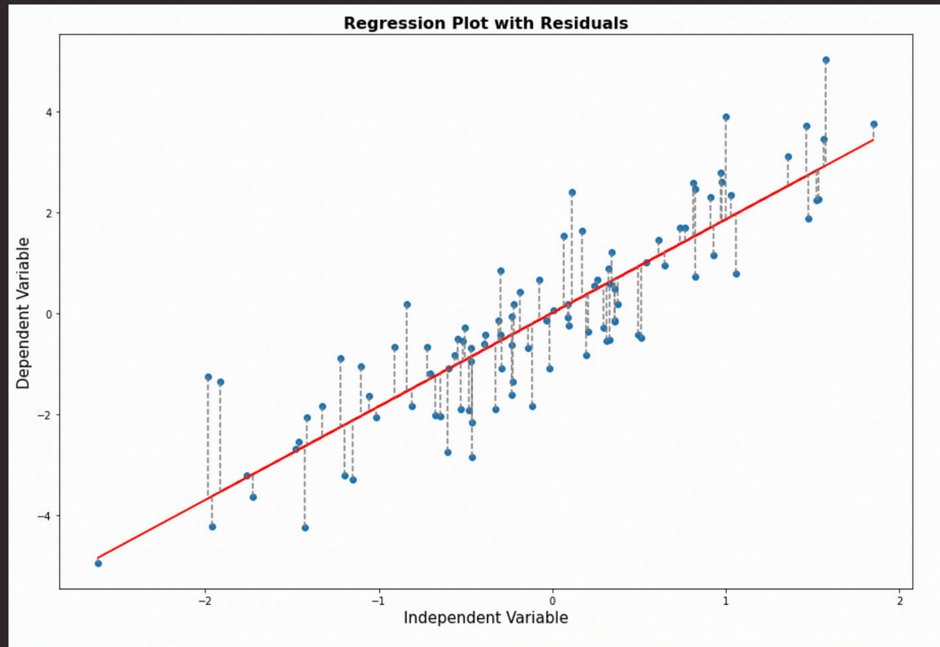
“The general rule is straightforward but has surprising consequences: whenever the correlation between two scores is imperfect, there will be regression to the mean” — Francis Galton

Regression to the mean

- Wrote a book
- Correlated human height based on average height of parents
- People tend to be shorter than average of human height
- Regression to mean refers to this phenomenon

Regression to the mean

- How we measure error in regressions
- The way that data points are scattered around the line



The question of simple regression

- Motivation:
  - Related the value of a numeric variable to that of another variable
  - May be several objectives to the analysis:

The question of simple regression

- Motivation:
  - Related the value of a numeric variable to that of another variable
  - May be several objectives to the analysis:
    - **Predict the value of the variable** based on the value of another
    - **Quantify variation** observed in one variable attributable to another
    - **Quantify the degree of change** in one variable attributible to another
    - **Null Hypothesis Significance Testing** for aspects of these relationships

A few definitions

$$(1)\ y_i = \alpha + \beta x_i + \epsilon_i$$

- Classic linear regression model
  - α (alpha, intercept) and β (beta, slope) = regression parameters
  - y and x = dependent and predictor variables
  - ϵ (epsilon) = residual error
    - error not accounted for by model

A few definitions

- Different equations in different fields…

$$y_i = \alpha + \beta x_i + \epsilon_i$$

$$y = m + \alpha X$$

$$y_1 = \beta_0 + \beta_1 x_1 + \epsilon_i$$

A few definitions

$$(2)\ \epsilon_i\ Gaussian(0, \sigma^2)$$

- Assumption for the residual error
- Gaussian with a mean of 0 and a variance we estimate with our model

A few definitions

- Sum of squares (SS) error for the residual
  - variance of residuals is the SSres/(n-2)
  - where n is ouor sample size

$$(3)\ SS_{res} = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

A few definitions

- $\hat{\beta}$ is our estimate of the slope

$$(4)\ \hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

- $\hat{\alpha}$ is our estimate of the intercept

$$(5)\ \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Data and assumptions

- Linear relationship
- Numeric continuous data for dependent y variable
- numeric continuous (or numeric ordinal) for predictor x variable
- Independence of observations
- Gaussian distribution of residuals
  - not the same as assuming raw data is Gaussian!
- Homoscedasticity
  - Residual variance is approximately the same along the x variable axis

Data and assumptions

- Explore in R with kaggle fish market dataset

Graphing

- Scatterplot
- dependent variable on y axis
- predictor variable on x axis
- Regression equation can be used to estimate **line of best fit**

Regression in R

- `lm()` simple regression function in R

Testing the assumptions

- Validating statistical model
- Part of exploratory data analysis
- Subjective and subtle
- Gaussian residual distribution
- Homoscedasticity

Testing the assumptions - R demo

Closer look at the residual distribution

- Histogram
- QQ plots
- Formal test for normality

Diagnosis - take 1

- The histogram is “shaped a little funny” for Gaussian
- Slightly too many points in the middle, slightly too few between the mean and the extremes in the histogram
- Very slight right skew in the histogram
- Most points are very close to the line on the q-q plot, but there are a few at the extremes that veer off
- Two points are tagged as outliers a little outside the error boundaries on the q-q plot (rows 118 and 124, larger than expected observations)

Diagnosis - take 2

- Near the mean, our residual density is slightly higher than expected under theoretical Gaussian
- Between -0.5 and -1 and also between 0.5 and +1 our residual density is lower than expected under theoretical Gaussian
- Overall the differences are not very extreme
- The distribution is mostly symmetrical around the mean

Formal test of assumption

- Shapiro Wilk test
- Do our residuals deviate from Gaussian?
- Tests like this are a bit atypical
  - Here we test against the null of NO DIFFERENCE

Tyranny of the p-value

- Traditionally when $P < 0.05$ we reject null hypothesis
- But when testing assumptions of no difference, we still use 0.05
  - Here when $p>0.05$, we interpret this as a lack of evidence tha there is a difference
- P value can often be misinterprested or relied on too heaily (see boot camp page for further reading)

Reporting the test of assumptions

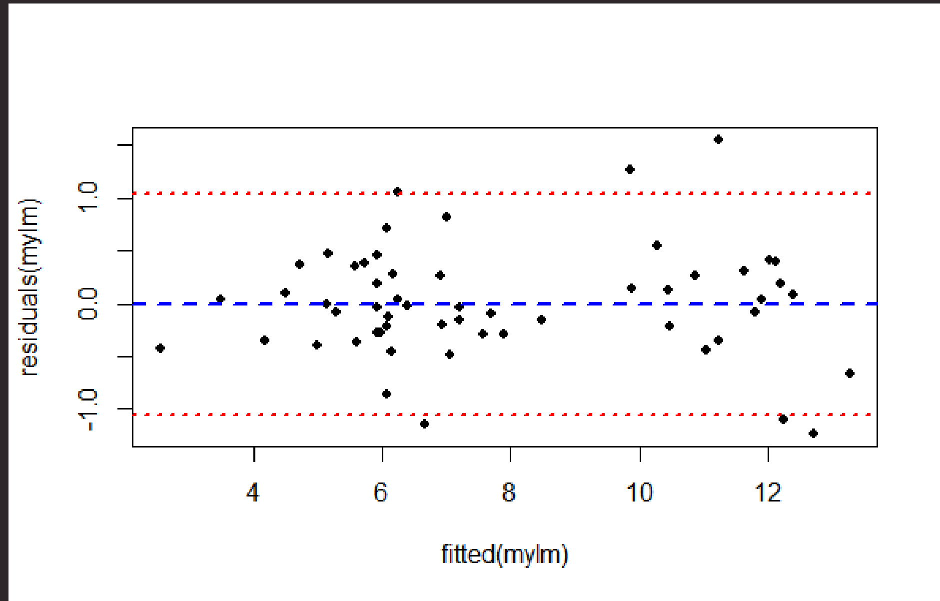
- Reporting of evidence supporting claims that assumptions underlying stats tests have been tested is ok
- Often understated despite an important part of the process
- Based on results of a Shapiro-Wilk test
  - We found no evidence our assumption of Gaussian residual distribution was violated (Shapiro-Wilk: $W = 0.97$, $n = 56$, $p = 0.14$)

Diagnostic plots and heteroscedasticity

- Heteroscedasticity: Variance of residuals are not constant across predicted values
- Homoscedasticity: Variance of residuals are constant across predicted values
- Looking for:
  - Even spread of residuals across x axis
  - Absence of systematic pattern in the data that might indicate lack of independence

Diagnostic plots and heteroscedasticity

- We see:
  - Not a perfect spread across whole x axis
  - Appears to be two groupings
    - For each group, residual spread appears similar
  - Low residual variance on left hand side - but only a few data points
  - Might be inclined to proceed, concluding no



Tests and output

- The `summary()` function provides different output depending on the `class()` of object passed to it

Tests and output

- Output:
  - **Call**: The formula representing the model
  - **Residuals**: Summary stats of residuals
  - **Coefficients**: includes estimate and std. err of estimates for regression coefficients
    - For `intercept` and slope for `Width`, the y coeff is 0.30 and slope 1.59.
  - **P-values**: Associated with parametric estimates
    - Intercept is 0.16 - thereore no evidence that the intercept is different to 0
    - Slope value (`width`) is <0.0001 - Width is significant

Reporting results

- We found a significant linear relationship for Weight predicting Height in perch (regression: R-squared = 0.97, df = 1,54, P < 0.0001).
- Acommpanied by appropriate graph
- Don’t just copy & paste summarized results

Alternatives to regression

- Many alternative options
- Some quite advanced - beyond scope of boot camp
- Data transformation
- Spearman Rank Correlation (if ok to just demonstrate a relationship)
- Intermediate difficulty - Kendal-Theil-Siegel nonparametric regression

Practice Exercises

Practice exercise 1

- Test whether the assumption of Gaussian residuals holds for the R formula `Weight ~ Length1` for perch in the fish dataset.
- Describe the evidence for why or why not; show your code.

Practice exercise 2

- Perform the regression for `Weight ~ Height` for the species `Bream.`
- Assess whether the residuals fit the Gaussian assumption.
- Present any graphical tests or other results and your conclusion in the scientific style.

Practice exercise 3

- For the analysis in #2 above present the results of your linear regression (if the residuals fit the Gaussian assumption) or a Spearman rank correlation (if they did not).

Practice exercise 4

- Plot `perch$Weight ~ perch$Length2`.
- The relationship is obviously not linear but curved. Devise and execute a solution to enable the use of linear regression, possibly by transforming the data.
- Show any relevant code and briefly explain your results and conclusions.

Practice exercise 5

- Explore the data for `perch` and describe the covariance of all of the morphological, numeric variables using all relevant means, while being as concise as possible. Show your code.

Practice exercise 6

- Write a plausible practice question involving the the exploration or analysis of regression.
- Make use of the fish data from any species except for Perch.