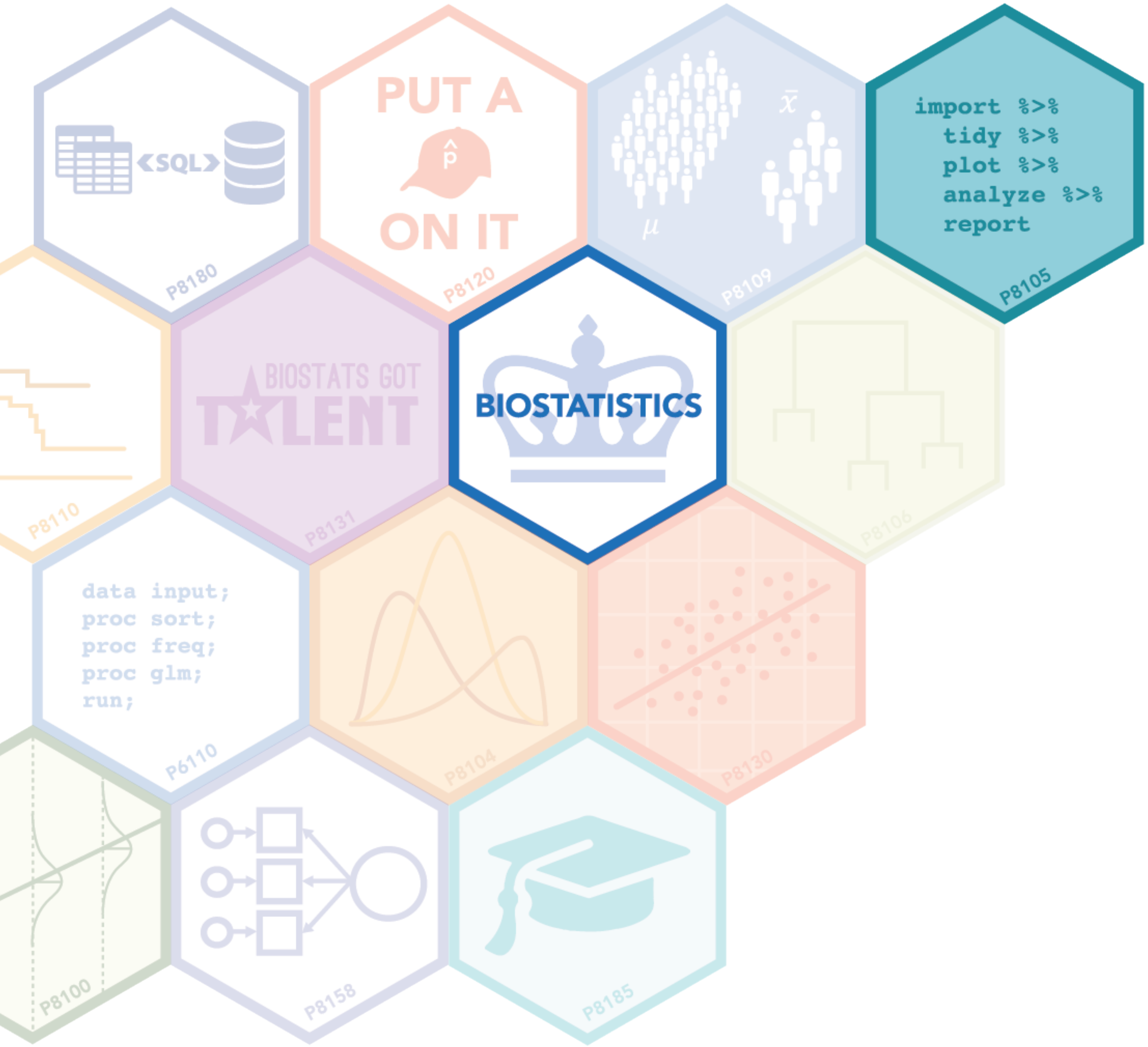
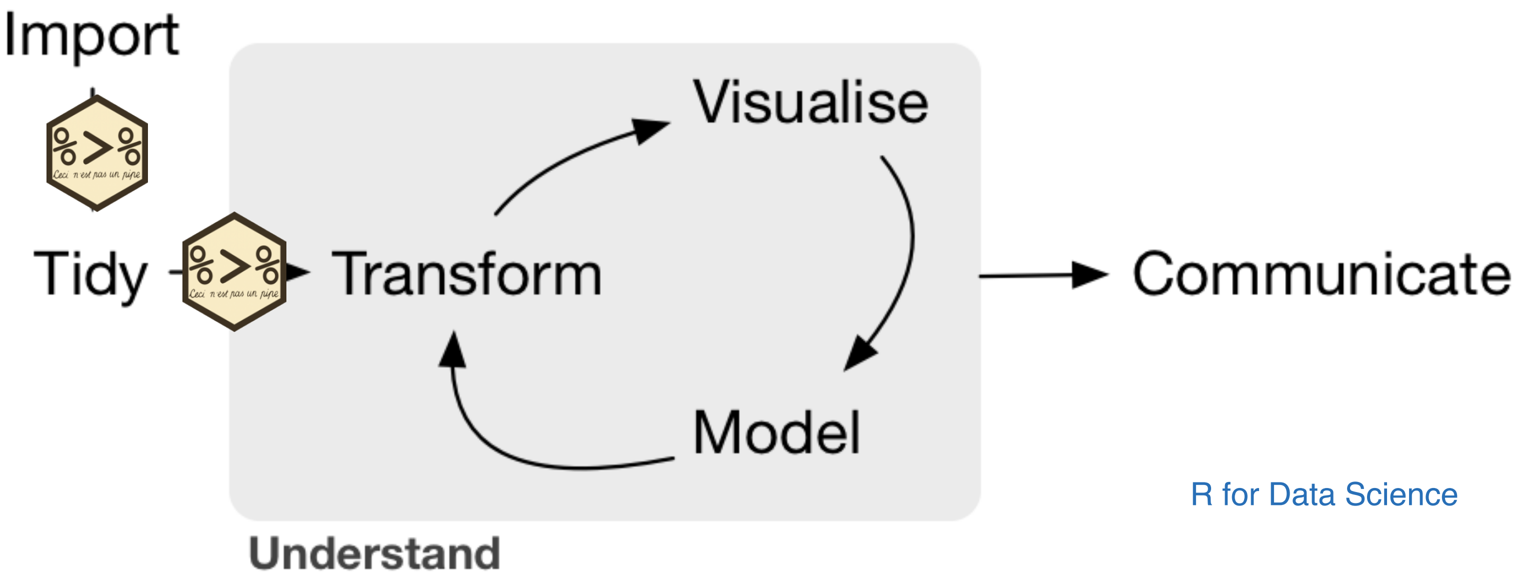


LINEAR REGRESSION

Jeff Goldsmith, PhD
Department of Biostatistics

Modeling

- Linear regression is one approach to modeling



R for Data Science

Regression is my favorite

- Like … seriously. I use regression for **everything**
- Regression covers simple stuff (t-tests) to complex stuff (automated variable selection via penalization)
  - Yes, I use regression for t-tests

Linear models

- Observe data $y_i, x_{i1}, \dots, x_{ip}$ for subjects 1 to n. Want to estimate $\beta_0, \beta_1, \dots, \beta_p$ in the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \epsilon_i \sim (0, \sigma^2)$$

- Assumptions: residuals have mean zero, constant variance, and are independent
- Estimate parameters using OLS

- This covers (well, really skips) a lot of ground -- general goodness of linear models, interpretation, inference, unbiasedness, …

Predictors

- Outcome is continuous; predictors can be anything

- Continuous predictors are added directly
- Categorical predictors require ~~dummy~~ indicator variables
  - For each non-reference group, a binary (0 / 1) variable indicating group membership for each subject is created and used in the model

Testing

- For a single regression coefficient, you can construct a test statistic using

$$t = \frac{\hat{\beta} - \beta}{\widehat{se}(\hat{\beta})}$$

- For large samples, this has a standard normal distribution

- To test multiple coefficients (i.e. those arising from the inclusion of a categorical variable with several predictors) you can use an F test / “ANOVA”

Diagnostics

- Many model assumptions (constant variance, model specification, etc) can be examined using residuals
  - Look at overall distribution (centered at 0? Skewed? Outliers?
  - Look at residuals vs predictors (any non-linearity? Trends? Non-constant residual variance?)

Generalized linear models

- Appropriate for non-continuous outcomes
- Common example is logistic regression:

$$logit\left(\frac{P(Y=1 \mid \boldsymbol{x})}{P(Y=0 \mid \boldsymbol{x})}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

- (GLMs count as part of my favorite)

Linear models in R

- `lm` for linear models
- `glm` for generalized linear models

- Arguments include
  - Formula: `y ~ x1 + x2`
  - Data

- Output is complex, and also kind of a mess
  - Use the `broom` package!