Regression pt2

Jonathan A. Pedroza, PhD

Assessing Individual Predictors

- if there is a significant association between a predictor and your outcome
  - the b value should be different from zero
- the hypothesis is tested with a t-statistic

$$t = \frac{b_{observed} - b_{expected}}{SE_b} = \frac{b_{observed}}{SE_b}$$

- the $b_{expected}$ value is the value you would expect if the null hypothesis were true
- the t-statistic has a probability distribution that differs slightly for this test
  - $N - k - 1$ is now used for the degrees of freedom

Interpretation of Coefficients

- unstandardized regression coefficients (b)
  - "for a 1 unit/point increase in IV, there is a ________ unit/point increase/decrease in your DV"
  - scale dependent
- standardized regression coefficient (β)
  - standardized scale of standard deviations
  - "for a one standard deviation increase in IV, there is a ______ standard deviation increase/decrease in DV"

Bias

- the main points of bias are
  - whether there are outliers
  - if it generalizes to other samples

Outliers

- can look at outliers in scatterplots for each relationship
  - check every predictor and outcome
- JP: best to check scatterplots with and without outliers included
- to check the relationship between all the predictors and the outcome, you'll check the residuals
  - with multiple predictors, we can't look at scatterplots, so we look at the residuals

Outliers

- **unstandardized residuals** are raw differences between the predicted and observed values of the outcome
- **standardized residuals** are all coverted in to z-scores, so they are expressed in standard deviation units
- **studentized residuals** are unstandardized residuals divided by an estimate of its standard deviation that varies from point to point

Influential Cases

- if there are severe outliers think about either deleting them
- using Cook's distance is influence of cases on the model
  - some state over $|1|$ could be influential

Influential Cases

- Leverage is the influence of observed value on the outcome across the predicted values
  - influential is a value 2-3x grater than the average value
  - k = predictors, n = number of cases/observations

$$\frac{2(k+1)}{n}\ or\ \frac{3(k+1)}{n}$$

- Mahalanobis distance
  - distance from the mean (highest = bad)

Sample Size & Linear Model

- larger sample = better
- how many participants/cases/observations for each variable depends on the statistician
  - JP: ~20 participants per predictor
  - Book: 10-15 participants per predictor
  - Some state ~5 participants per predictor

Methods for Entering Predictors in Model

- hierarchical regression OR hierarchical linear regression
  - NOT hierarchical linear modeling
  - this is including predictors as steps
    - Block 1: control variables
    - Block 2: predictors of interest - main effects
    - Block 3: interactions
- simultaneous regression
  - all predictors included together

Methods for Entering Predictors in Model

- automated regression
  - lets the computers do everything for you in choosing predictors in a forward manner (searches for predictors that would be best) or backward (contains all predictors and removes useless predictors)
  - no theory
  - **do not use this method**

Model Comparisons

- we may be interested in comparing two multiple regression models
  - these models must be nested
- to put it simply **nested** models are when models contain all the same variables, with the second model containing additional variables
- good way to see if adding additional variables made your model better/account for more variation in your outcome
  - compares model by using ANOVA

Model Comparisons

- complicated fit criteria but to keep it simple, lower AIC = better fitting model
  - penalizes model for having more variables
- comparing these AIC values is interpretable
  - Recommendations by Burnham and Anderson (2002)

Multicollinearity

- **multicollinearity** is when one IV correlates strongly with another IV (r $> .7$)
- **variance inflation factor (VIF)** is when an IV has a strong linear relationship with one or more IV(s)
  - VIF $> 10 =$ concern in the model, diagnose it for multicollinearity
  - average VIF > 1 there may be bias in model
- **tolerance** is similar to VIF in that tolerance $= 1/$VIF
  - tolerance below .1 is a serious problem
  - tolerance below .2 may indicate bias in model

Reporting Linear Models

- The things that need to be reported
  - F test for the model (omnibus test)
  - if hierarchical regression
    - report changes in F and R^2
  - report b and β values and the p values
  - intercept is the average value of your outcome when all numeric predictors are zero for all the reference groups of your categorical variables

Reporting Linear Models

```
	Pearson's product-moment correlation

data:  penguins$bill_depth_mm and penguins$flipper_length_m
t = -13.261, df = 340, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6496752 -0.5093379
sample estimates:
       cor
-0.5838512
```

Reporting Linear Models

```
   value    numdf    dendf
539.8239   2.0000 339.0000
```

Reporting Linear Models

```
lm(formula = body_mass_g ~ bill_depth_mm + flipper_length_m
    data = penguins)
                  coef.est coef.se  t value  Pr(>|t|)
(Intercept)       -6541.91   540.75   -12.10     0.00
bill_depth_mm        22.63    13.28     1.70     0.09
flipper_length_mm    51.54     1.87    27.64     0.00
---
n = 342, k = 3
residual sd = 393.18, R-Squared = 0.76

Call:
lm(formula = body_mass_g ~ bill_depth_mm + flipper_length_m
    data = penguins)

Standardized Coefficients::
      (Intercept)     bill_depth_mm flipper_length_mm
        0.0000000         0.0557360         0.9037433
```

Reporting Linear Models

- The model including bill depth and flipper length was statistically significant; $F(2, 339) = 539.82$, $p < .001$.
- Bill depth and flipper length accounted for 76% of the variation in body mass.
- There was no evidence of a significant association between bill depth and body mass ($b = 22.63$, $\beta = .06$, $p = .09$)
  - For a one mm increase in bill depth, there is a 22.63 gram increase in body mass.
- There was a significant association between flipper length and body mass ($b = 51.54$, $\beta = .90$, $p < .001$).
  - For a one mm increase in flipper length, there is a 51.54 increase gram increase in body mass.