ANOVA

Team 14
Eileanor LaRocco, Erik Zwicklbauer, Junghee Mun, Sebastian Rincon

What is ANOVA?

- **ANOVA: AN**alysis **O**f **VA**riance
  - Analyzing variance to look for differences
- Testing how well nested models fit a dataset in comparing against models that add features
  - Less complex models (M_0) vs More complex models (M_1)
- R: anova() function performs a hypothesis test comparing two models
- Null hypothesis (H_0): $M_0 = M_1$
  - Model 0 and Model 1 explains the data equally well
    - p-value is the probability that the null is true given the data
    - If the p-value is very small (ie. 0.05), reject the H_A and accept H_0
- Alternative hypothesis (H_A): M_0 != M_1

ANOVA's Assumptions

- Models must be **nested**
  - The predictors in M_0 must be a subset of the predictors in M_1
- **Equal variances** between models (homoscedasticity)

Interpreting ANOVA

Hypothesis	Correct model?	R formula for correct model
Null	M0	Y ~ A + B
Alternative	M1	Y ~ A + B + C

- Null: $\hat{y}\_wage = \beta_1 age + \beta_2 age^2$
- Alternative: $\hat{y}\_wage = \beta_1 age + \beta_2 age^2 + \beta_3 age^3$

Contents of ANOVA (1)

- **Res. Df: Res**idual **D**egrees of **F**reedom
- **RSS: R**esidual **S**um of **S**quares

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

- **Df: D**egrees of **F**reedom
- **SS: S**um of **S**quares
  - SS model$_1$ = RSS model$_0$ - RSS model$_1$

```
Analysis of Variance Table

Model  1: wage ~ poly(age, 1)
Model  2: wage ~ poly(age, 2)
Model  3: wage ~ poly(age, 3)
Model  4: wage ~ poly(age, 4)
Model  5: wage ~ poly(age, 5)
Model  6: wage ~ poly(age, 6)
Model  7: wage ~ poly(age, 7)
Model  8: wage ~ poly(age, 8)
Model  9: wage ~ poly(age, 9)
Model 10: wage ~ poly(age, 10)
   Res.Df     RSS Df Sum of Sq        F    Pr(>F)
1    2998 5022216
2    2997 4793430  1    228786 143.7638 < 2.2e-16 ***
3    2996 4777674  1     15756   9.9005  0.001669 **
4    2995 4771604  1      6070   3.8143  0.050909 .
5    2994 4770322  1      1283   0.8059  0.369398
6    2993 4766389  1      3932   2.4709  0.116074
7    2992 4763834  1      2555   1.6057  0.205199
8    2991 4763707  1       127   0.0796  0.777865
9    2990 4756703  1      7004   4.4014  0.035994 *
10   2989 4756701  1         3   0.0017  0.967529
```

Contents of ANOVA (2)

- **F-statistic**:
  - SS.model1 <- RSS.model0 - RSS.model1
  - res <- RSS.model1/Res.Df.model1
  - diff <- SS.model1/(Res.Df.model0-Res.Df.model1)
  - F.stat <- diff/res

- **P-value:** Probability that the null is true given the data
  - P-value < 0.05: Model 1 and 2 are not statistically different
  - P-value > 0.05: Model 1 is an improvement from Model 0

```
Analysis of Variance Table

Model  1: wage ~ poly(age, 1)
Model  2: wage ~ poly(age, 2)
Model  3: wage ~ poly(age, 3)
Model  4: wage ~ poly(age, 4)
Model  5: wage ~ poly(age, 5)
Model  6: wage ~ poly(age, 6)
Model  7: wage ~ poly(age, 7)
Model  8: wage ~ poly(age, 8)
Model  9: wage ~ poly(age, 9)
Model 10: wage ~ poly(age, 10)
   Res.Df     RSS Df Sum of Sq        F    Pr(>F)
1    2998 5022216
2    2997 4793430  1    228786 143.7638 < 2.2e-16 ***
3    2996 4777674  1     15756   9.9005  0.001669 **
4    2995 4771604  1      6070   3.8143  0.050909 .
5    2994 4770322  1      1283   0.8059  0.369398
6    2993 4766389  1      3932   2.4709  0.116074
7    2992 4763834  1      2555   1.6057  0.205199
8    2991 4763707  1       127   0.0796  0.777865
9    2990 4756703  1      7004   4.4014  0.035994 *
10   2989 4756701  1         3   0.0017  0.967529
```

Check your ANOVA knowledge!