Week 5
Bivariate Regression Review

Applied Statistical Analysis/Quantitative Methods I

Jeffrey Ziegler, PhD

Assistant Professor in Political Science & Data Science
Trinity College Dublin

Fall 2024

Class Business

- Problem set #2 is out right now!
  - Due before class next class
  - Answer key for problem set #1 is up on GitHub
- Questions from last time?

Roadmap through Stats Land

Where we've been:

- We're learning how to make inferences about a population from a sample
- How to determine if two samples are different or independent (diff-in-means, contingency tables)
- Last 2 weeks: We learned about bivariate correlation and regression (correlation, parameters, prediction)

Outline for today:

- Partitioning our error
- Review for exam

Part of the story: Estimating σ^2

The linear model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

- Assumption: Variance for each of the conditional distributions of $Y|X$ is the same at all x values
- Best "guess"/estimate of variance?
  - We can pool all errors to common estimate for σ^2, which is the residual sums of squares

$$\hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

  - The degrees of freedom is $n-2$ because we've used 2 parameters for estimating the $\hat{\beta}$ and $\hat{\alpha}$

Now we have an association, how good is model?

- The strength of the fit of a linear model is most commonly evaluated using R^2
- This can be calculated two ways:
    1. $R^2 =$ square of correlation coefficient (r)
    2. $R^2 = \dfrac{\text{explained variability}}{\text{total variability}}$
- Interpreted as % of variability in y explained by x
- Bounded between [0, 1]

Intuition behind (un)explained variability

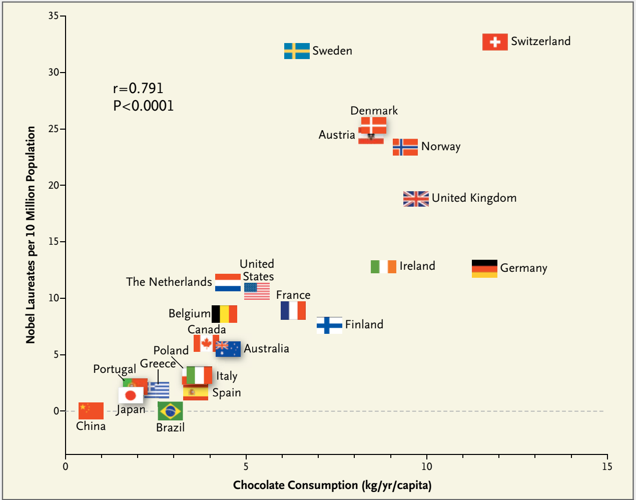


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Partitioning variability: Sums of squares

- Total sums of squares (TSS) quantifies the overall squared distance of the Y values from the overall mean of the response $\bar{Y}$

$$TSS = \sum (y - \bar{y})^2$$

- Regression sums of squares (RegSS) quantifies the squared distance from the fitted line to overall mean

$$RegSS = \sum (\hat{y} - \bar{y})^2$$

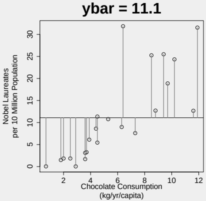
- Residual sums of squares (RSS) quantifies the squared distance from the Y values to the fitted line

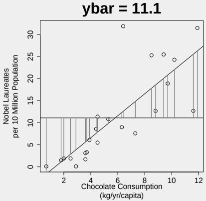
$$RSS = \sum (y - \hat{y})^2$$

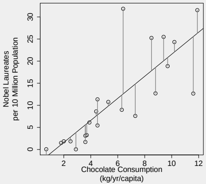
Intuition behind partitioning variability

SS = Sums of Squares

Total variability	=	Explained variability	+	Unexplained variability
Total SS	=	Regression SS	+	Residual SS
$\sum(y-\bar{y})^2$	=	$\sum(\hat{y}-\bar{y})^2$	+	$\sum(y-\hat{y})^2$
SS_{total}	=	SS_{reg}	+	SS_{error}







R^2 EXPLAINED

- R^2 (coefficient of determination)

$$R^2 = \frac{RegSS}{TSS} = 1 - \frac{RSS}{TSS}$$

  - Proportion of variation in the response y that is explained by the model
  - Stated as r^2 in simple linear regression
  - Square of the correlation coefficient r
  - $0 \leq R^2 \leq 1$
  - R^2 near 1 suggests a good fit to the data, if $R^2 = 1$, all points fall exactly on the line

Another way: Analysis of Variance (ANOVA)

- Sums of squares are summarized in an ANOVA table (Analysis of Variance)
- Ex: Price of clock at auction

```
> lm.full<-lm(clock$Price~clock$Age+clock$Bidders)
> anova(lm.full)
Analysis of Variance Table

Response: clock$Price
              Df  Sum Sq Mean Sq F value    Pr(>F)
clock$Age      1 2554859 2554859 144.136 8.957e-13 ***
clock$Bidders  1 1722301 1722301  97.166 9.135e-11 ***
Residuals     29  514035   17725
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> RegSS=sum((lm.full$fitted.values-mean(clock$Price))^2)
> RegSS
[1] 4277160
> RSS=sum((clock$Price-lm.full$fitted.values )^2)
> RSS
[1] 514034.5
> F=(RegSS/2)/(RSS/29)
> F
[1] 120.6511
> pf(F,2,29, lower.tail = FALSE)
[1] 8.769066e-15
```

Another way: Analysis of Variance (ANOVA)

- $R^2 = \frac{RegSS}{TSS} = \frac{27419.5}{27419.5+348.8} = 0.9874$
- 98.7% of the variation in the price of a clock is explained by the age and number of bidders

Wrap-up: What we've learned so far...

Week 1: Stats Intro

1. Review of statistics terms
2. Quantifying concepts: Types of data
3. Making inferences from data
   - Statistic vs. parameter
   - Sampling distribution, C.L.T.
   - Point estimate, confidence interval

Wrap-up: What We've learned so far...

Week 2: H_0 testing

- Wanted to understand if X causes Y
- We talked about 2 ways to think about this:
  - Compare two independent samples

Wrap-up: What we've learned so far...

Week 3: Intro to Regression

1. Estimate if two variables are dependent
   - Chi-squared test of independence
   - Standardized residuals
2. Correlations
3. Simple linear regression:
   - Assumptions
   - Estimation

Week 4: Bivariate regression

- Correlation inference
- Parameters
- Prediction

Looking ahead to next class

- **Next week:**
  - Problem set #2 due by Monday 23:59
  - Exam 1 in-person during lecture