2/16/2015

Monday, February 16, 2015
9:30 AM

SOLDER.CSV
Exercise #4 – 1A

- Your first obligation is to make sense of the data
- Models that are simpler are better

lm0 = lm(skips~Solder+Mask+Opening+Solder:Opening+Opening:Mask, data=solder)
summary(lm0)
Residual standard error: 4.704 on 883 degrees of freedom
Multiple R-squared: 0.6954, Adjusted R-squared: 0.6898
F-statistic: 126 on 16 and 883 DF, p-value: < 2.2e-16
anova(lm0)

Analysis of Variance Table

Response: skips

	Df	Sum Sq	Mean	Sq F value
Solder	1	6204.2	6204.2	280.339
Mask	4	13825.0	3456.3	156.173
Opening	2	15967.0	7983.5	360.738
Solder:Opening	2	3345.5	1672.7	75.584
Mask:Opening	7	5262.9	751.8	33.972
Residuals	883	19541.7	22.1	

1. Take out the variables with the lowest sum squared
    a. This process is subjective because:
        i. if you add any variable, it's going to fit a little better to the data
2. Repeated the process
3. Checking the Multiple R-Squared after each adjustment to the formula
    a. This will tell you how well the data fits to your model

LIFEEXPECTANCY.CSV
Exercise #4 – 1B

- Dividing data up into subsets by "Group" will help the data fit better
    - Use dummy variables
- Life Expectancy eventually does plateau, but the model does not

Main Effects & Interaction Variables Discussion

- $y_i = \beta_0 + \beta_1 x_i + \beta_2 \mathbf{1}_{\{Group2\}}$

  Two Main Effects
  Same Slope
  Different Intercept

- $y_i = \beta_0 + \beta_1 x_i + \beta_2 \mathbf{1}_{\{Group2\}} + \beta_3 \mathbf{1}_{\{Group2\}X_i}$

  Adding Interaction Models
  Different Slopes

- $Life\ Expectancy = \beta_0 + \beta_1 \log(GDP)$

  Monotonic model ~ Keeps getting larger

How much can we trust a model?

How much will our answer change if data was a little different from the previous sample?

Analogy: Policeman and Criminal

The policeman asks the criminal where he was the night of the crime. The first time the criminal says that he was at home. The second time he says that he was at his friend's house.

The criminal's answer changes each time. Therefore he is not stable or trustworthy.

Definitions

- **Sampling Distribution**: how a statistic changes from sample to sample
  - Repeated samples from a population
- **Statistic**: Anything you get from data (mean, standard deviation, etc)
- **Unbiased**: Sample distribution is centered at the truth
  - Truth is measured by observing how the least squared estimate changes from sample to sample
- **Monte Carlo**: How you find the estimated sample distribution
  - A computer is used to simulate a sample distribution

SIMDATA_POP.CSV
Exercise #4 – 2

```
my_sim = do(1000)*{
        this_sample = sample)simdata_pop, 50)
        lmsamp = lm(y~x, data=this_sample
        coef(lmsamp))
}
```

- **Standard Error**: Standard Deviation of the sampling distribution
    - Sample distribution of size 50 conveys a lot more than one of size 10

Universe Analogy

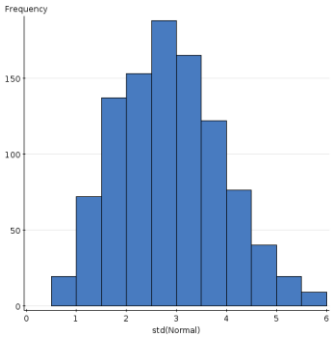
Universe 1 → $\{x_i, y_i\}$ → $\hat{\beta}_1^{(i)}$

.
.
.
.

Universe 10 → $\{x_i, y_i\}$ → $\hat{\beta}_1^{(i)}$

.
.
.
.

Universe 1000 → $\{x_i, y_i\}$ → $\hat{\beta}_1^{(i)}$



- **Bootstrapping**:
    - Used if you're interested in seeing how the estimator changes from sample to sample in a population
    - The sample acts as if it were the wider population
        - Observe how the estimator changes
    - Taking multiple samples WITH replacement (catch and release)
    - You should get a different slope and intercept each time
        - Like Monte Carlo
    - If it is a random sample and large, bootstrapping theory should hold true
- **Bootstrap Error**: Standard Deviation of the bootstrapped sample
- **Confidence Interval**: an interval estimate of a population parameter; "choppy chop"
    - The "cofint" function can be used to solve this in RStudio