Topic 4 - Analysis of Variance
Approach to Regression

STAT 525 - Fall 2013

Outline

• Partitioning sums of squares
• Degrees of freedom
• Expected mean squares
• General linear test
• $R^2$ and the coefficient of correlation
• What if $X$ random variable?

Partitioning Sums of Squares

• Organizes results arithmetically
• Total sums of squares in $Y$ is defined

$$\text{SSTO} = \sum (Y_i - \bar{Y})^2$$

• Can partition sum of squares into
  - Model (explained by regression)
  - Error (unexplained / residual)

• Rewrite the total sum of squares as

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2$$
$$= \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

Total Sum of Squares

• If we ignored $X_h$, the sample mean $\bar{Y}$ would be the best linear unbiased predictor

$$Y_i = \beta_0 + \varepsilon_i = \mu + \varepsilon_i$$

• SSTO is the sum of squared deviations for this predictor
• Sum of squares has $n - 1$ degrees of freedom because we replace $\beta_0$ with $\bar{Y}$
• The total mean square is SSTO/$(n - 1)$ and represents an unbiased estimate of $\sigma^2$ under the above model
**SAS & Total Sum of Squares**

- SAS uses “Corrected Total” for SSTO
- Uncorrected total sum of squares is $\sum Y_i^2$
- “Corrected” means that the sample mean has been subtracted off before squaring

**Regression Sum of Squares**

- SAS calls this *model* sum of squares
  \[ SSR = \sum (\hat{Y}_i - \bar{Y})^2 \]
- Degrees of freedom is 1 due to the addition of the slope
- SSR large when $\hat{Y}_i$’s are different from $\bar{Y}$
- This occurs when there is a linear trend
- Under regression model, can also express SSR as
  \[ SSR = b_1^2 \sum (X_i - \bar{X})^2 \]

**Error Sum of Squares**

- Error (or residual) sum of squares is equal to the sum of squared residuals
  \[ SSE = \sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2 \]
- Degrees of freedom is $n - 2$ due to using $(b_0, b_1)$ in place of $(\beta_0, \beta_1)$
- SSE large when |residuals| are large. This implies $Y_i$’s vary substantially around fitted line
- The MSE=SSE/(n−2) and represents an unbiased estimate of $\sigma^2$ when taking $X$ into account

**ANOVA Table**

- Table puts this all together

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression (Model)</td>
<td>1</td>
<td>$b_1^2 \sum (X_i - \bar{X})^2$</td>
<td>SSR/1</td>
</tr>
<tr>
<td>Error</td>
<td>$n - 2$</td>
<td>$\sum (Y_i - \hat{Y})^2$</td>
<td>SSE/(n−2)</td>
</tr>
<tr>
<td>Total</td>
<td>$n - 1$</td>
<td>$\sum (Y_i - \bar{Y})^2$</td>
<td></td>
</tr>
</tbody>
</table>

Topic 4
Expected Mean Squares

- All means squares are random variables
- Already showed $E(MSE) = \sigma^2$
- What about the MSR?
  $$E(\text{MSR}) = E(b_1^2 \sum (X_i - \bar{X})^2) = E(b_1^2) \sum (X_i - \bar{X})^2 = (\text{Var}(b_1) + E(b_1)^2) \sum (X_i - \bar{X})^2 = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$$
- If $\beta_1 = 0$, MSR unbiased estimate of $\sigma^2$

F test

- Can use this structure to test $H_0 : \beta_1 = 0$
- Consider
  $$F* = \frac{\text{MSR}}{\text{MSE}}$$
- If $\beta_1 = 0$ then $F*$ should be near one
- Need sampling distribution of $F*$ under $H_0$
- By Cochran’s Thm (pg 70)
  $$F* = \frac{\text{SSR}}{\sigma^2} \div \frac{\text{SSE}}{n-2} \sim \frac{\chi^2_1}{F_{1,n-2}}$$
- If $H_0$ is false, MSR > MSE
- P-value = $\Pr(F(1, n-2) > F*)$
- Reject when $F*$ large, P-value small
- Recall t-test for $H_0 : \beta_1 = 0$
- Can show $t^2_{n-2} \sim F_{1,n-2}$
- Obtain exactly the same result (P-value)

Example

```r
data a1;
    infile 'C:\Textdata\CH01TA01.txt';
    input size hours;
    proc reg data=a1;
        model hours=size;
        id size;
    run;
```
General Linear Test

- Reduced model $\rightarrow H_0 : \beta_1 = 0$

- Can be shown that $SSE(F) \leq SSE(R)$

- Idea: more parameters provide better fit

- If $SSE(F)$ not much smaller than $SSE(R)$, full model doesn’t better explain $Y$

$$F^* = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F}$$

$$= \frac{(SSTO - SSE)}{SSE/(n-2)}$$

- Same test as before but more general

Pearson Correlation

- Number between -1 and 1 which measures the strength of the linear relationship between two variables

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

$$= b_1 \sqrt{\frac{\sum (X_i - \bar{X})^2}{\sum (Y_i - \bar{Y})^2}}$$

- Test $H_0 : \beta_1 = 0$ similar to $H_0 : \rho = 0$
**Coefficient of Determination**

- Defined as the proportion of total variation explained by the model utilizing $X$

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

- $0 \leq R^2 \leq 1$

- Often multiplied by 100 and described as a percent

**Coefficient of Determination**

- Can show this is equal to $r^2$

$$r^2 = b_1^2 \left( \frac{\sum (X_i - \bar{X})^2}{\sum (Y_i - \bar{Y})^2} \right)$$

$$= \frac{b_1^2 \sum (X_i - \bar{X})^2}{\sum (Y_i - \bar{Y})^2}$$

$$= \frac{SSR}{SSTO}$$

- Relationship not true in multiple regression

- See page 75 for limitations of $R^2/r$

**Normal Correlation Model**

- So far, have assumed $X_i$’s are known constants

- In inference we’ve considered repeat sampling of error terms with the $X_i$’s remaining fixed ($Y_i$’s vary)

- What if this assumption is not appropriate?

- In other words, what if $X_i$’s are random?

- If interest still in relation between these two variables can use correlation model

- Normal correlation model assumes a bivariate normal distribution

**Bivariate Normal Distribution**

- Consider random variables $Y_1$ and $Y_2$

- Distribution requires five parameters
  - $\mu_1$ and $\sigma_1$ are the mean and std dev of $Y_1$
  - $\mu_2$ and $\sigma_2$ are the mean and std dev of $Y_2$
  - $\rho_{12}$ is the coefficient of correlation

- Bivariate normal density and marginal distributions given on page 79

- Marginal distributions are normal

- Conditional distributions are also normal
Conditional Distribution

• Consider the distribution of $Y_1$ given $Y_2$

1. Can show the distribution is normal
2. The mean can be expressed

$$\left( \mu_1 - \mu_2 \rho_{12} \frac{\sigma_1}{\sigma_2} \right) + \rho_{12} \frac{\sigma_1}{\sigma_2} Y_2 = \alpha_{1|2} + \beta_{12} Y_2$$

3. With constant variance $\sigma_1^2 (1 - \rho_{12}^2)$

• Similar properties of normal error regression model
• Can use regression to make inference about $Y_1$ given $Y_2$

So What if $X$ is Random?

• Suppose $X_i$’s are random samples from $g(X_i)$?
• Then the previous regression results hold if:
  - The conditional distributions of $Y_i$ given $X_i$ are normal and independent with conditional means $\beta_0 + \beta_1 X_i$ and conditional variance $\sigma^2$.
  - The $X_i$ are independent and $g(X_i)$ does not involve the parameters $\beta_0$, $\beta_1$, and $\sigma^2$.

Inference on $\rho_{12}$

• Point estimate using $Y = Y_1$ and $X = Y_2$ given on 4-15
• Interest in testing $H_0 : \rho_{12} = 0$
• Test statistic is

$$t^* = \frac{r_{12} \sqrt{n - 2}}{\sqrt{1 - r_{12}^2}}$$

• Same result as $H_0 : \beta = 0$
• Can also form CI using Fisher $z$ transformation or large sample approx (pg 85)
• If $X$ and $Y$ are nonnormal, can use Spearman correlation (pg 87)

Background Reading

• Appendix A
• KNNL Chapter 3