Simple Linear Regression: Partitioning Variance, Quality of Fit, the F-test

BIOS 6611

CU Anschutz

Week 7

Partitioning the Total Variability

Partitioning the Variability

We can examine the fit of the regression line by partitioning the **total variability** of Y into two components:

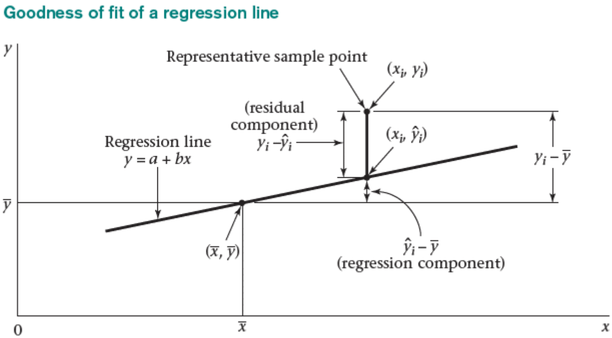
Regression component: The variability in Y due to the regression of Y on X. The regression component is the difference between the predicted Y and the mean of the Y's:

$$\hat{Y}_i - \bar{Y}$$

Residual component (error): The variability in Y "left-over" after the regression of Y on X. The residual component is the difference between the observed Y and predicted Y:

$$Y_i - \hat{Y}_i$$

Partitioning the Variability



Source: Rosner 7th Ed., pg. 435

The simplest regression estimate for Y_i is $\bar{Y}$ (an intercept-only model). The difference between the observed Y's and the mean of the Y's, $Y_i - \bar{Y}$, is the **total error**. The total error can be broken down further as the sum of the **regression component** and the **residual component**:

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

The Fundamental Equation of Regression Analysis

This partitioning of the variability leads to the **fundamental equation of regression analysis**:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$\text{SS}_{\text{Total}} = \text{SS}_{\text{Model}} + \text{SS}_{\text{Error}}$$

Total Sums of Squares (SS_{Total})

The total sum of squares is the sum of squares of the deviations of the individual sample points from the sample mean (note the relationship between SS_{Total} and the variance of Y, $\hat{\sigma}^2_Y$):

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2; \; \hat{\sigma}^2_Y = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \frac{SS_{Total}}{n-1}$$

The Fundamental Equation of Regression Analysis

Error Sums of Squares (SS_{Error})

The error sum of squares is the sum of squares of the residual components (note the relationship between SS_{Error} and the variance of Y given X, $\hat{\sigma}^2_{Y|X}$):

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2; \; \hat{\sigma}^2_{Y|X} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{SS_{Error}}{n-2}$$

Model Sums of Squares (SS_{Model})

The model sum of squares is the sum of squares of the regression components:

$$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = SS_{Model} = SS_{Total} - SS_{Error}$$

Measuring Goodness of Fit

Quality of the Fit

Once the least-squares line is determined, we may wish to know how well the least-squares regression line 'fits' the data.

- Does the fitted line help us predict Y? That is, is least-squares line better than no line at all for predicting Y?
- And if so, to what extent?

Measuring the **goodness of fit** involves quantifying how much scatter there is around the regression line.

We know that the SS_{Error} represents the variation in the data after fitting our regression line (i.e., the "left-over" variation), where large values indicate a lot of left-over variation. Leveraging the partitioning of the variability, we can describe this variability.

Coefficient of Determination (R^2)

The "R-squared" value, also known as the **coefficient of determination**, is the proportion of total variation in the data (about the average $\bar{Y}$) that is removed by fitting the regression line.

In other words, it is the proportion of the variance of Y that can be explained by the variable X. It is calculated as

$$R^2 = \frac{SS_{Total} - SS_{Error}}{SS_{Total}} = \frac{SS_{Model}}{SS_{Total}}$$

R^2 is often multiplied by 100 and is interpreted as the percent of the total variation in the dependent variable Y that is explained by the independent variable X (*using a linear model*).

Properties of R^2

- R^2 can only be between 0 and 1: $0 \leq R^2 \leq 1$
- If $R^2 = 0$, then the regression line explains *none* of the variability in Y and the regression line is no better than using $\bar{Y}$ as our predictor of Y.
- If $R^2 = 1$, then there is a perfect fit and the regression line explains all of the variability. In this case, every data point falls exactly on the regression line and there is no residual variation
- R^2 does **not** measure the magnitude of the slope or measure the appropriateness of the straight-line model (i.e., a large R^2 does not necessarily imply an "adequate" model).

R^2 Example

```r
fev <- read.csv('FEV_rosner.csv', header=T)
summary( lm(fev ~ age, data=fev))
```

```
##
## Call:
## lm(formula = fev ~ age, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57539 -0.34567 -0.04989  0.32124  2.12786
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.431648   0.077895   5.541 4.36e-08 ***
## age         0.222041   0.007518  29.533  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5675 on 652 degrees of freedom
## Multiple R-squared:  0.5722, Adjusted R-squared:  0.5716
## F-statistic: 872.2 on 1 and 652 DF,  p-value: < 2.2e-16
```

FEV (outcome) and age (predictor) have

$$R^2 = \frac{SS_{Model}}{SS_{Total}} = \frac{280.91916}{490.91984} = 0.5722$$

Interpretation: 57.22% of the variability in FEV is explained by the linear relationship with age.

ANOVA Table and F-test

The ANOVA Table

The analysis of variance (ANOVA) table is typically used to summarize regression results, where n is the sample size and p is the number of predictors included in the model:

Source	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio (F)	p-value
Model	SS_{Model}	p	MS_{Model}	$\text{F}=\frac{MS_{Model}}{MS_{Error}}$	$\Pr(F_{p,n-p-1} > \text{F})$
Error	SS_{Error}	$n-p-1$	MS_{Error}		
Total	SS_{Total}	$n-1$			

Where $MS_{Model} = \frac{SS_{Model}}{p}$ and $MS_{Error} = \frac{SS_{Error}}{n-p-1}$.

`PROC REG` in `SAS` will produce this table automatically. In `R` we have to do a little more work to get our results into this format.

F-Test for Simple Linear Regression

From our ANOVA table we saw that the **model mean square** is the regression (model) sum of squares divded by the number of predictor variables, p, in the model ($p = 1$ for SLR). Theoretically, the expectation of our MS_{Model} is

$$E(MS_{Model}) = \sigma^2_{Y|X} + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

The **residual mean square** was the residual sum of squares divided by its degrees of freedom ($n - 2$ for SLR). Its expectation is

$$E(MS_{Error}) = E(s^2_{Y|X}) = \sigma^2_{Y|X}$$

F-Test for Simple Linear Regression

It can be shown that the ratio of two variances follows an F distribution under the null hypothesis that the two variances are equal ($\sigma_1^2 = \sigma_2^2$):

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$$

In the context of regression, under the null hypothesis that the true slope of the regression line is zero ($H_0 : \beta_1 = 0$), both MS_{Model} and MS_{Error} are independent estimates of $\sigma^2_{Y|X}$. Thus, the ratio of the regression mean square to the residual mean square will have an F distribution with p and $n - p - 1$ degrees of freedom:

$$F = \frac{MS_{Model}}{MS_{Error}} \sim F_{p,n-p-1}$$

F-Test for Simple Linear Regression

The F test is used to test if the model including covariate(s) results in a significant reduction of the residual sum of squares compared to a model containing only an intercept.

If the null hypothesis is true, then the expected value of the F ratio should be 1. If the null hypothesis is false, then the expected value of the F ratio is greater than 1.

The t-test and the F-test are equivalent for testing $H_0 : \beta_1 = 0$ in simple linear regression:

- If $X \sim t_n$, then $X^2 \sim F_{1,n}$.
- Recall, $t = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)}$, where $t \sim t_{n-p-1}$ under H_0.

F-Test Example

```r
fev <- read.csv('FEV_rosner.csv', header=T)
summary( lm(fev ~ age, data=fev))
```

```
##
## Call:
## lm(formula = fev ~ age, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57539 -0.34567 -0.04989  0.32124  2.12786
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.431648   0.077895   5.541 4.36e-08 ***
## age         0.222041   0.007518  29.533  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5675 on 652 degrees of freedom
## Multiple R-squared:  0.5722, Adjusted R-squared:  0.5716
## F-statistic: 872.2 on 1 and 652 DF,  p-value: < 2.2e-16
```

$H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

$F = 872.18$, $Pr(F_{1,652} > 872.18) < 0.0001$ (can use `pf(872.2, df1=1, df2=652, lower.tail=F)`)

Conclusion: Reject the null hypothesis that $\beta_1 = 0$ and conclude that there is a significant association between age and FEV (p < 0.0001).