ANOVA and Simple Linear Regression[1]

November 29, 2017

[1] HMS, 2017, v1.0

Chapter References

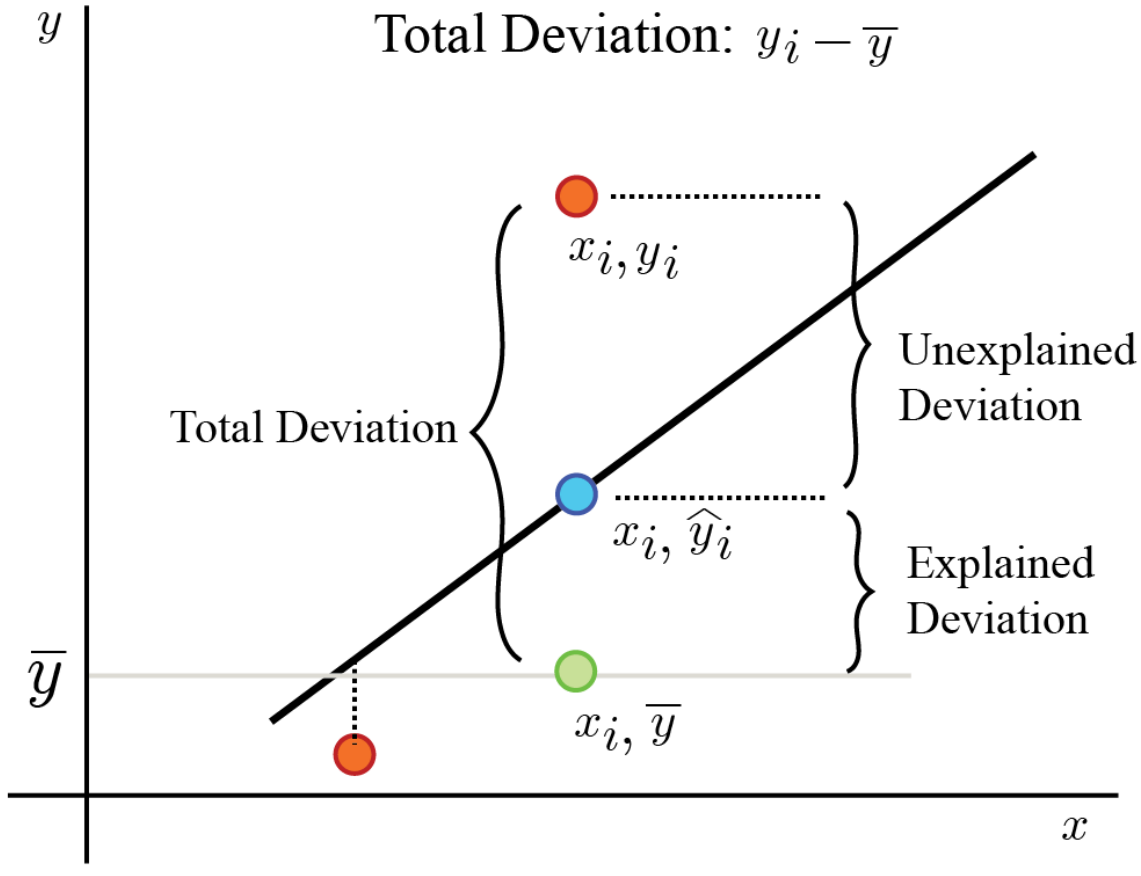
- Navidi, Chapter 7, not great coverage but these slides should be sufficient.

Splitting up the Variation

Explained Deviation: $\widehat{y}_i - \overline{y}$

Unexplained Deviation: $y_i - \widehat{y}_i$ $+$

Total Deviation: $y_i - \overline{y}$



Splitting up the Variation

$$\text{Total Deviation} = \text{Explained Deviation} + \text{Unexplained Deviation}$$

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

We can square and sum both sides, noting that the cross term equals zero (see slides 14 to 21 in the analysis of variance stack), such that:

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

$$\text{Total Sum of Squares} = \text{Regression Sum of Squares} + \text{Error Sum of Squares}$$

$$\text{TSS} = \text{SSR} + \text{SSE}$$

Splitting up the Variation

$$\text{TSS} = \text{SSR} + \text{SSE}$$

If TSS = SSE then SSR is zero. That means everything we see in the fit is due to errors in the data and there is no actual signal. The smaller is SSE relative to TSS the more likely there is a relationship between y and x.

The ratio, SSR/SSE, or something similar, could tell us how good the fit is. The higher the ratio the better the fit. Note if there is no error in the data then TSS = SSR and SSR/SSE tends to infinity.

If we convert SSE and SSR to variances we could do one better and examine the ratio using an F test and decide whether a given ratio came about by chance or is actually likely to due to a real signal in the data.

Splitting up the Variation

Summary table including degrees of freedom and variances:

Source of Variation	Deg Freedom	Sum of Squ	Mean Square	
Due to Regression	1	SSR	SSR/1	MSR
Due to Errors	$n-2$	SSE	SSE/$(n-2)$	MSE
Total Error	$n-1$	TSS		

To determine whether there is a good fit or not we can compute the F ratio:

$$F = \frac{\text{MSR}}{\text{MSE}}$$

In Summary

$H_o : \hat{\beta}_1 = 0$ and $H_1 : \hat{\beta}_1 \neq 0$

Test Statistic:

$$F = \frac{\text{MSR}}{\text{MSE}}$$

Rejection Region: $F \geq F_{\alpha,1,n-2}$

p-Value: $P(F_{1,n-2} \geq F)$

ANOVA Table for Linear Regression

Source	df	SS	MS	F	p-value
Reg	1	SSR $= \sum(\hat{y} - \bar{y})^2$	MSR $= \frac{SSR}{1}$	$F = \frac{\text{MSR}}{\text{MSE}}$	
Errors	$n-2$	SSE $= \sum(y_i - \hat{y})^2$	MSE $= \frac{\text{SSE}}{n-2}$		
Total	$n-1$	TSS $= \sum(y_i - \bar{y})^2$			

Example from StatDisk

```
x = [1,2,3,4,5]; y=[1,1,2,2,4]
Sample size, n:     5
Degrees of freedom: 3

Correlation Results:
Correlation coeff, r: 0.9036961
Critical r:           0.8783393
P-value (two-tailed): 0.03535

Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      -0.1
Slope, b1:            0.7

Total Variation:       6
Explained Variation:   4.9
Unexplained Variation: 1.1
Standard Error:        0.6055301 = Sqrt(MSE)
Coeff of Det, R^2:     0.8166667
```

Example from Excel

	A	B	C	D	E	F	G	H
1								
2	1	1		Slope	0.7	-0.1	Intercept	
3	2	1		Error in slope	0.191485422	0.635085296	Error in intercept	
4	3	2		R Squared	0.816666667	0.605530071	sqrt(MSE)	
5	4	2		F	13.36363636	3	df	
6	5	4		SSR	4.9	1.1	SSE	
7								
8								

Hypothesis testing on $\hat{\beta}_1$

Output from Python:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.817
Model:                            OLS   Adj. R-squared:                  0.756
Method:                 Least Squares   F-statistic:                     13.36
Date:                Sat, 25 Nov 2017   Prob (F-statistic):             0.0354
Time:                        12:01:23   Log-Likelihood:                -3.3094
No. Observations:                   5   AIC:                             10.62
Df Residuals:                       3   BIC:                             9.838
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.1000      0.635     -0.157      0.885      -2.121       1.921
x1             0.7000      0.191      3.656      0.035       0.091       1.309
==============================================================================
Omnibus:                        nan   Durbin-Watson:                   2.509
Prob(Omnibus):                  nan   Jarque-Bera (JB):                0.396
Skew:                        -0.174   Prob(JB):                        0.821
Kurtosis:                     1.667   Cond. No.                         8.37
==============================================================================
```

Hypothesis testing on $\hat{\beta}_1$

x = [1,2,3,4,5]; y= [3,3.5,2.5,3.,2.8]
Output from Python:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.152
Model:	OLS	Adj. R-squared:	-0.130
Method:	Least Squares	F-statistic:	0.5388
Date:	Sun, 26 Nov 2017	Prob (F-statistic):	0.516
Time:	16:55:18	Log-Likelihood:	-1.0804
No. Observations:	5	AIC:	6.161
Df Residuals:	3	BIC:	5.380
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.2300	0.407	7.943	0.004	1.936	4.524
x1	-0.0900	0.123	-0.734	0.516	-0.480	0.300

Omnibus:	nan	Durbin-Watson:	3.407
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.162
Skew:	-0.050	Prob(JB):	0.922
Kurtosis:	2.125	Cond. No.	8.37