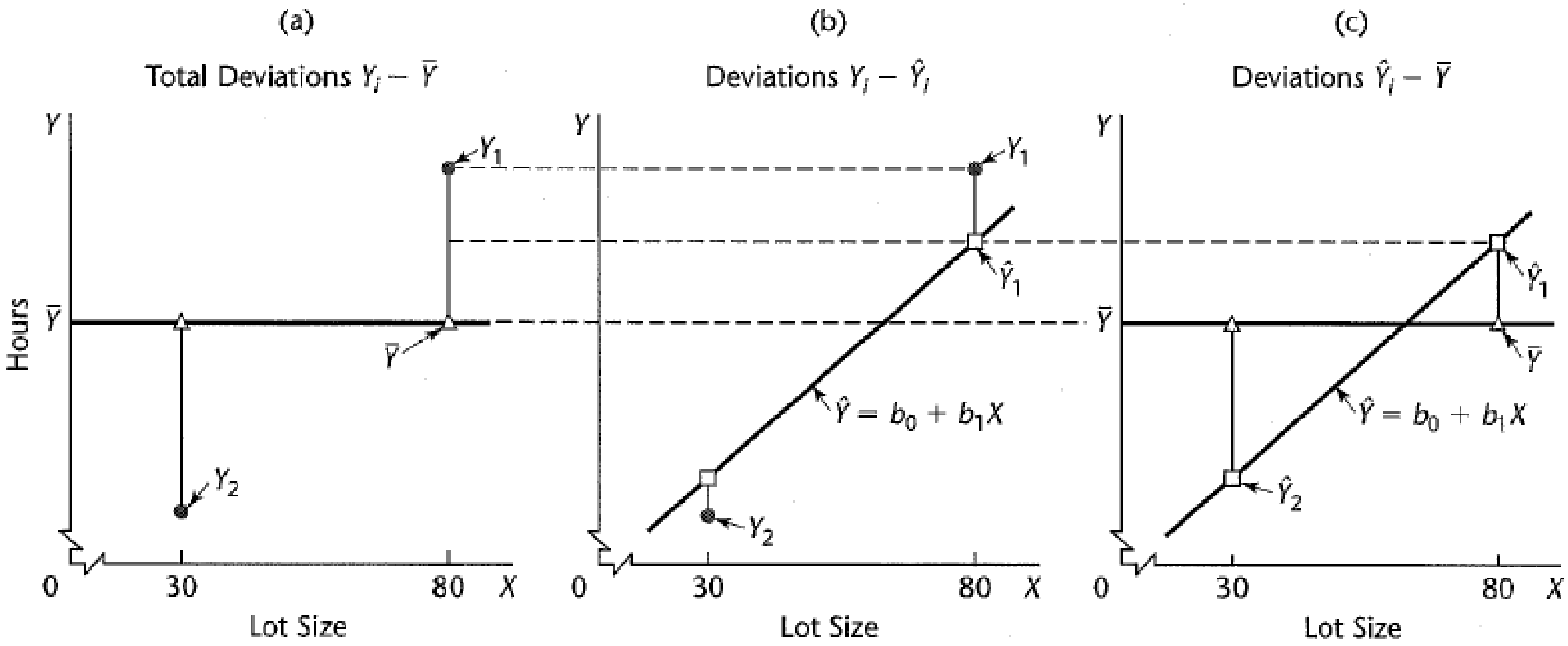
The ANOVA F test and the General Linear Test (GLT)

The Analysis of Variance Test (ANOVA)

- The ANOVA test is a hypothesis test to study different variances from different resource in the data
  - The most common type of ANOVA test is the Global F test, also known as the significance test of the model.
  - The test statistic follows a F distribution; therefore, it is a F test which sometimes can be replaced by a T-test.
- The General Linear Test (GLT) is a test to study different variances in different models defined in Ho (Reduced model) and Ha (Full model), respectively.
  - It uses a F test to analyze the variances, hence it is essentially an ANOVA test.
  - GLT test is usually used in model improvement.
  - It is different from the Generalized Linear Model (GLM).

Partitioning variance in the total sum of squares $Y_i - \bar{Y}$



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

$$SSTO = SSE + SSR$$

Also known as SSM (model)

"Total sum of squares" " error sum of squares" "regression sum of squares"

Partitioning Degree of freedom

$$\boldsymbol{\Sigma(Y_i - \bar{Y})^2 = \Sigma\left(Y_i - \hat{Y}_i\right)^2 + \Sigma\left(\hat{Y}_i - \bar{Y}\right)^2}$$

$$\boldsymbol{SSTO = SSE + SSR}$$

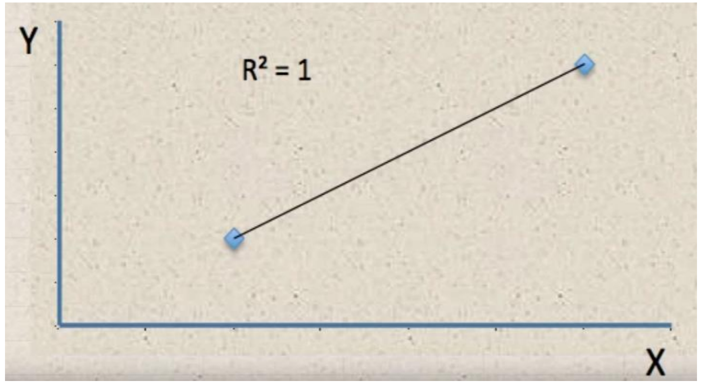
Degree of freedom $\quad n - 1 = n - 2 + 1$

Degree of freedom of error (an intuitive flavor)

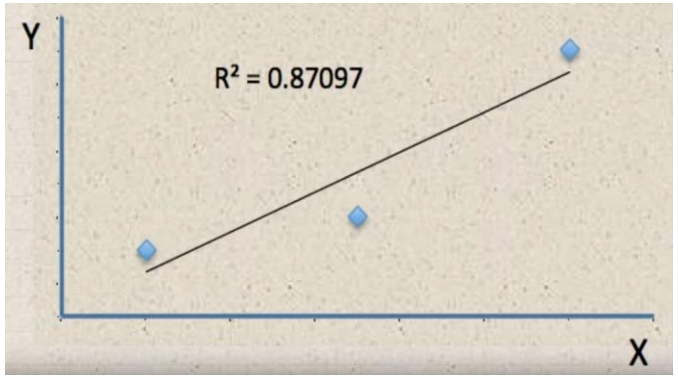
Q: what is the minimum requirement on data points to estimate this regression?

$Y = \beta_0 + \beta_1 X + \epsilon,$ and $\epsilon = Y - \beta_0 - \beta_1 X$

$n = 2, dfE = 0$



$n = 3, dfE = 1$



So, $dfE = n - 2 = n - p$

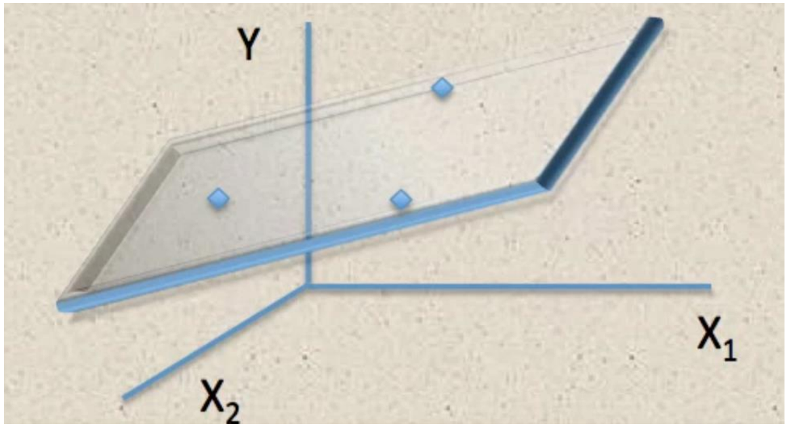
Where p is the number of parameters ($p = 2$ $in\ this\ case$)

Degree of freedom of error (an intuitive flavor)

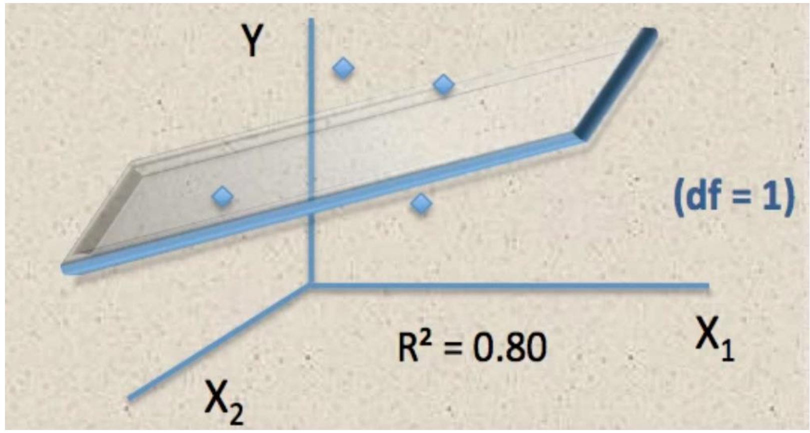
Q: what is the minimum requirement on data points to estimate this regression? What is the degree of freedom left?

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$ and $\epsilon = Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2$

$n = 3, dfE = 0$



$n = 4, dfE = 1$



So, $dfE = n - 3$

$= n - p$

Where p is the number of parameters ($p = 3$ *in this case*)

The F test of Ho: $\beta_1 = 0$ versus Ha: $\beta_1 \neq 0$

This F test is also known as the Significant test of a SLR model, or the significant linear impact of the independent variable.

Source of Variation	SS	df	MS	E{MS}
Regression	$SSR = \Sigma(\hat{Y}_i - \bar{Y})^2$ $= b_1^2\ \Sigma(X_i - \bar{X})^2$	1	MSR= $\frac{SSR}{1}$	$\sigma^2 + \beta_1^2(X_i - \bar{X})^2$
Error	$SSE = \Sigma(Y_i - \hat{Y}_i)^2$	$n-2$	MSE= $\frac{SSE}{n-2}$	σ^2
Total	$SSTO = \Sigma(Y_i - \bar{Y})^2$	$n-1$		

$$\textit{The test statistic is denoted by } F^* \textit{ or } F_s = \frac{MSR}{MSE} \sim F\,(1, n-2)$$

Reject Ho if $F^* > F\,(1-\alpha; 1, n-2)$

Example 1 Complete the hypothesis test Ho: $\beta_1 = 0$ versus Ha: $\beta_1 \neq 0$
On a (partial) given ANVOA table.

Source of Variation	SS	df	MS	F
Regression	252378	1	252378/1=252378	232378/2384=105.88
Error	54825	23	54825/23=2384	
Total	307203	24		

$$F_s = \frac{MSR}{MSE} = 105.88 \sim F\,(1, 23)$$

Reject Ho if
$F_s > F\,(0.95; 1, 23)$= 4.28

```
qf(0.95,1,23)
```

Conclude that X has a significant linear impact on Y, or the SLR model is statistically significant.

Example 2 Complete the hypothesis test Ho: $\beta_1 = 0$ versus Ha: $\beta_1 \neq 0$
On a model summary output.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -259.63      17.32  -14.99   <2e-16 ***
weight       3721.02      81.79   45.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.84 on 46 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9783,	Adjusted R-squared:  0.9778 
F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

$F_s = 2070 \sim$ **F (1, 46)**

Conclude that X has a significant linear impact on Y, or the SLR model is statistically significant.

Equivalence of a two-sided F test (ANOVA) and t test (SLR)

$$Ho: \beta_1 = 0 \qquad Ha: \beta_1 \neq 0$$

The T test statistic $t_s = \frac{b_1}{s\{b_1\}} \sim t(n-2)$

The F test statistic $F_s = \frac{MSR}{MSE} \sim F\,(1, n-2)$

$$F_s = \frac{MSR}{MSE} = \frac{b_1^2 \Sigma (X_i - \bar{X})^2}{MSE} = \frac{b_1^2}{s^2\{b_1\}} = t_s^2$$

$$Since\ s^2\{b_1\} = MSE/\Sigma(X_i - \bar{X})^2$$

The T test and F tests are equivalent in SLR $F_s = (t_s)^2$ for two sided test, the critical values:

$$t\left(1 - \frac{\alpha}{2}, n-2\right)^2 = F(1-\alpha; 1, n-2)$$

For example, at $\alpha = 0.05, dfe = 23$: $t(0.975; 23)^2 = (2.069)^2 = 4.28 = F(0.95, 1, 23)$

Example 3 The equivalence of the F and the T test in the SLR.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -259.63      17.32  -14.99   <2e-16 ***
weight        3721.02      81.79   45.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.84 on 46 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9783,    Adjusted R-squared:  0.9778 
F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

ts= $\frac{3721.02}{81.79} = 45.5 \sim$ t (46)

Fs= $2070 \sim$ F (1, 46)

The T-test and the F-test are the same because $45.5^2 = 2070$,
And they both have the same p-value.

F Test (ANOVA) and T test are **not always equivalent**

1. They are equivalent in simple linear regression (SLR) problem and will not be so for Multiple regression.

2. They are equivalent when $H_0 : \beta_1 = 0$.
   - Ho: $\beta_1 = \beta_1^* \ (\neq 0)$ can be tested with a *t*-test.
   - In Ho: $\beta_1 = \beta_1^* \ (\neq 0)$, the test statistic F^* has a *non-central F* distribution and require extra steps and not covered in the course.

3. In SLR, the T test is more flexible and more commonly used than the F test. We will continue to compare them in MLR.

The General Linear Test (GLT) approach

Ho: $\boldsymbol{\beta_1 = 0}$ versus Ha: $\boldsymbol{\beta_1 \neq 0}$

Full model: $Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$

Under Ha

$$SSE(F) = \Sigma\left(Y_i - \hat{Y}_i\right)^2 = SSE\,, \quad df_F = n - 2$$

Reduced model: $Y_i = \beta_0 + \epsilon_i$

Under Ho

$$SSE(R) = \Sigma(Y_i - \bar{Y}_i)^2 = SSTO, \quad df_R = n - 1$$

"Significant reduction in SSE?"

$$F^* = \frac{\dfrac{SSE(R) - SSE(F)}{df_R - df_F}}{SSE(F)/df_F} = \frac{MSR}{MSE} \sim F\,(1, n-2)$$

*The test statistic of the **general linear test** in **simple linear regression** is identical to the **ANOVA** test statistic.*

Example 4 The global F test in example 1 can convert to a GLT test

Source of Variation	SS	df	MS	F
Regression	252378	1	252378	105.88
Error	54825	23	2384	
Total	307203	24		

Full model: $Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$

Under Ha

$$SSE(F) = \Sigma\left(Y_i - \hat{Y}_i\right)^2 = SSE = 54825, \quad df_F = n - 2 = 25 - 2 = 23$$

Reduced model: $Y_i = \beta_0 + \epsilon_i$

Under Ho

$$SSE(R) = \Sigma(Y_i - \bar{Y}_i)^2 = SSTO = 307203, \quad df_R = n - 1 = 25 - 1 = 24$$

$$F_s = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{SSE(F)/df_F} = \frac{\frac{307203 - 54825}{24 - 23}}{\frac{54825}{23}} = \frac{\frac{252378}{1}}{2384} = 105.88, \textit{which is same as the test statistic in the Global F test}, F_s = \frac{MSR}{MSE}$$

General Linear Test can be extended to multiple parameters ($\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \ldots$)

Given the number of additional parameters in the the full (more complex) model compared to the reduced model, does the full model yield a larger reduction in SSE than we would expect to get by adding a similar number of unrelated (i.e., useless) predictor variables?

$$F^* = \frac{\dfrac{SSE(R) - SSE(F)}{df_R - df_F}}{SSE(F)/df_F} = \frac{MSR}{MSE} \sim F\ (df_R - df_F, df_F)$$

The GLT is a very general tool.

We will see it again in Multiple Linear Regression.

Summary

- The basic idea of ANOVA is to study the source and proportion of variance in data
- F test (ANOVA) and T test (Simple Linear model) are not always equivalent
- GLT can be used to compare two models that containing different X variables, and decide whether (dropping) some of the X variable affect the effectiveness of the linear model to explain the variance in Y.