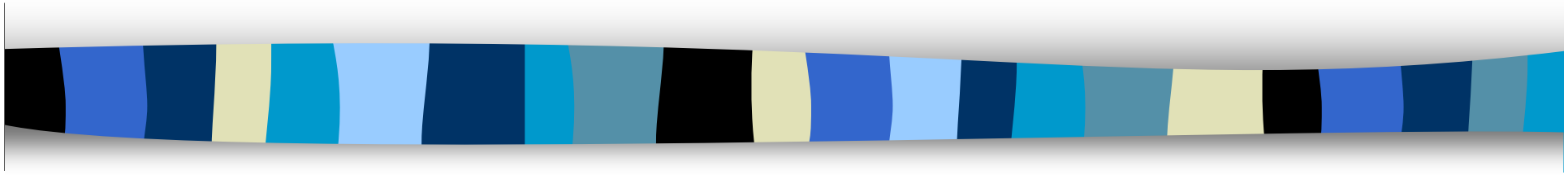
Graduate School

Social Science Statistics II

Gwilym Pryce
g.pryce@socsci.gla.ac.uk



Lecture 2: ANOVA, Prediction, Assumptions and Properties

Notices:

- Register

Aims and Objectives:

- Aim:
  - to complete our introduction to multiple regression
- Objectives:
  - by the end of this lecture students should be able to:
    - understand and apply ANOVA
    - understand how to use regression for prediction
    - understand the assumptions underlying regression and the properties of estimates if these assumptions are met

Last lecture:

- 1. Correlation Coefficients
- 2. Multiple Regression
  - OLS with more than one explanatory variable
- 3. Interpreting coefficients
  - b_k estimates how much $y\uparrow$ if $x_k\uparrow$ by one unit.
- 4. Inference
    - b_k only a sample estimate, thus distribution of b_k across lots of samples drawn from a given population
  - confidence intervals
  - hypothesis testing
- 5. Coefficient of Determination: R^2 and Adj R^2

Plan of today's lecture:

- 1. Prediction
- 2. ANOVA in regression
- 3. F-Test
- 4. Regression assumptions
- 5. Properties of OLS estimates

1. Prediction

- Given that the regression procedure provides estimates the values of coefficients, we can use these estimates to predict the value of *y* for given values of *x*:
  - e.g. Income, education & experience from L1:

Model		Unstandardized Coefficients B	Std. Error
1	(Constant)	-4.200	23.951
	X1	1.450	1.789
	X2	2.633	3.117

a. Dependent Variable: Y

where x1 = post-school education,

x2 = experience

  - Implies the following equation:

$$\hat{y} = -4.2 + 1.45\,x_1 + 2.63\,x_2$$

Predicting y for particular values of x_k

- We can use this equation to predict the value of y for particular values of x_k:
  - e.g. what is the predicted income of someone with 3 years of post-school education & 1 year experience?

$$\hat{y} = -4.2 + 1.45\, x_1 + 2.63\, x_2$$
$$= -4.2 + 1.45 \times (3) + 2.63\ (1) = \textbf{£2,780}$$

  - How does this compare with the predicted income of someone with 1 year of post-school education and 3 years work experience?

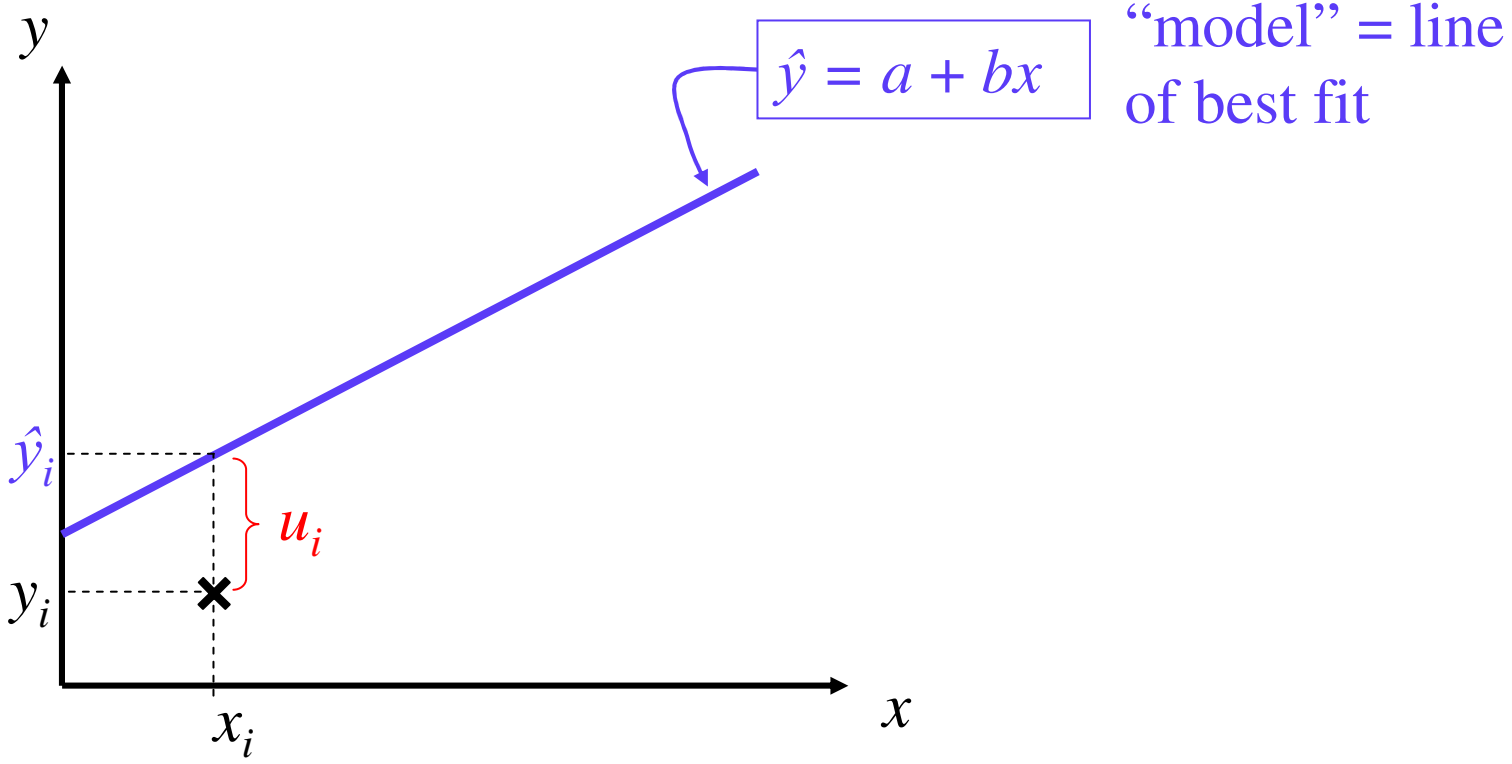
$$\hat{y} = -4.2 + 1.45\, x_1 + 2.63\, x_2$$
$$= -4.2 + 1.45 \times (1) + 2.63\ (3) = \textbf{£5,140}$$

Predicting y for each value of x_k in the data set:

$$\hat{y}_i = -4.2 + 1.45\, x_{1i} + 2.63\, x_{2i}$$

y (Salary £000)	x_1 (yrs of educ)	x_2 (yrs of exp.)	$\hat{y}$
35	5	10	29.35
22	2	9	22.37
31	7	10	32.25
21	3	9	23.82
42	9	13	43.04

Notice that there is a difference between observed values, y_i, and model predictions, $\hat{y}_i$



Where does this "residual" or "error" come from?

- u = unobserved factors not included in the regression.
  - Even if we run the regression on the **population**, these omitted variables would mean that there would be an unexplained component in our model of y, leading to discrepancies between what the model predicts and our observations on y:

$\hat{y} = a + bx$ (1) *line, or "model", we have estimated*

$y = \hat{y} + u$ (2) observed value of y = *model prediction plus* unexplained component, u.

Sub (1) in (2): $\Rightarrow$ $y = \overbrace{a + bx} + u$

Residuals, u_i = prediction error for obs i:

$$u_i = y_i - \hat{y}_i$$
$$= y_i - (a + b_1 x_1 + b_2 x_2)$$

where a, b_1, and b_2 are our sample estimates of the population coefficients: α, β_1, β_2

$$\hat{y} = -4.2 + 1.45\, x_{1i} + 2.63\, x_{2i} + u_i$$

Y (Salary £000)	X1	X2	Y*	u
35	5	10	29.35	5.65
22	2	9	22.37	-0.37
31	7	10	32.25	-1.25
21	3	9	23.82	-2.82
42	9	13	43.04	-1.04

Forecasting

- If the observations in the regression are not individuals, but time periods
  - e.g. observation 1 = 1970, observation 2 = 1971
- and if you know (or can guess) what the value of x_k will be in the next period, then you can use the estimated regression equation to predict what y will be next period.

2. ANOVA in regression

- The **variance** of *y* is calculated as the sum of squared deviations from the mean divided by the degrees of freedom:

$$\sigma_y = \frac{\sum_i (y_i - \bar{y})^2}{n-1}$$

- **Analysis of variance** (ANOVA) is about examining the proportion of this variance that is explained by the regression, and the proportion of the variance that cannot be explained by the regression (reflected in the random error term)

- This amounts to an analysis of the numerator in the variance equation – the sum of squared deviations of *y* from the mean.
  - the denominator is constant for all analysis on a particular sample
    - the error variance, for example, will have the same denominator as the variance of y.
  - the sum of squared deviations from the mean without dividing by (*n*-1) is called the "Total Sum of Squares"

$$TSS = \sum_i (y_i - \bar{y})^2$$

Regression Sum of Squares

- Measures how much the **predicted values** vary
- The variation in $\hat{y}$, the predicted values of y for the observed values of the explanatory variables in our sample, can be thought of as the explained variation in y,
  - If we square the deviations of $\hat{y}$ from the mean value of y, we get the ***explained sum of squares*****,** often called the ***Regression Sum of Squares***.
  - REGSS measures the sample variation in $\hat{y}$

$$REGSS = \sum_i (\hat{y}_i - \bar{y})^2$$

Residual sum of squares

- **Measures how much variation there is in *u***
- When a line of best fit is calculated, we get errors (unless the line fits perfectly) and this can be thought of as **un**explained variation in y
  - We calculate the residual or error for a particular observation i as the difference between our observed value of the dependent variable, y_i, and the value predicted by our model, $\hat{y}_i$:

    $u_i = y_i - \hat{y}_i$
  - if we square these errors – or *residuals* – before adding them up we get the ***residual sum of squares (RSS)***
  - RSS represents the **degree of unexplained variation in y**.

$$RSS = \sum_i (y_i - \hat{y}_i)^2$$

- **Total variation in y** is measured by the Total Sum of Squares (TSS)

- If the ***REGSS, the explained variation in y,*** is **large** relative to the total variation in y, then the regression line is doing a good job of explaining y
  - i.e. the model fits the data well

- If the ***REGSS, the explained variation in y,*** is **small** relative to the total variation in y then the regression model is not doing a good job of explaining y
  - i.e. the model fits the data poorly

$$R^2 = \frac{\mathrm{Re}gSS}{TSS}$$

- A useful measure that we have already come across is the proportion of the variation of *y* that can be explained by the model

TSS = REGSS + RSS

- The sum of squared deviations of *y* from the mean (i.e. the numerator in the variance of *y* equation) are called the

  TOTAL SUM OF SQUARES **(TSS)**

- The sum of squared deviations of residuals (error) *e* are called the

  RESIDUAL SUM OF SQUARES* **(RSS)**

  * sometimes called the "**error sum of squares**"

- The difference between TSS & RSS is called the

  REGRESSION SUM OF SQUARES# **(REGSS)**

  #the REGSS is sometimes called the "**explained sum of squares**" or "**model sum of squares**"

$$\Rightarrow \textbf{TSS} = \textbf{REGSS} + \textbf{RSS}$$

- R^2 is the proportion of the variation in *y* that is explained by the regression.

$$R^2 = REGSS/TSS$$

- Thus, the explained sum of squares is equal to R^2 times the total variation in *y*:

$$REGSS = R^2 \times TSS$$

- Given that RSS is the unexplained variation in *y* we can say that:

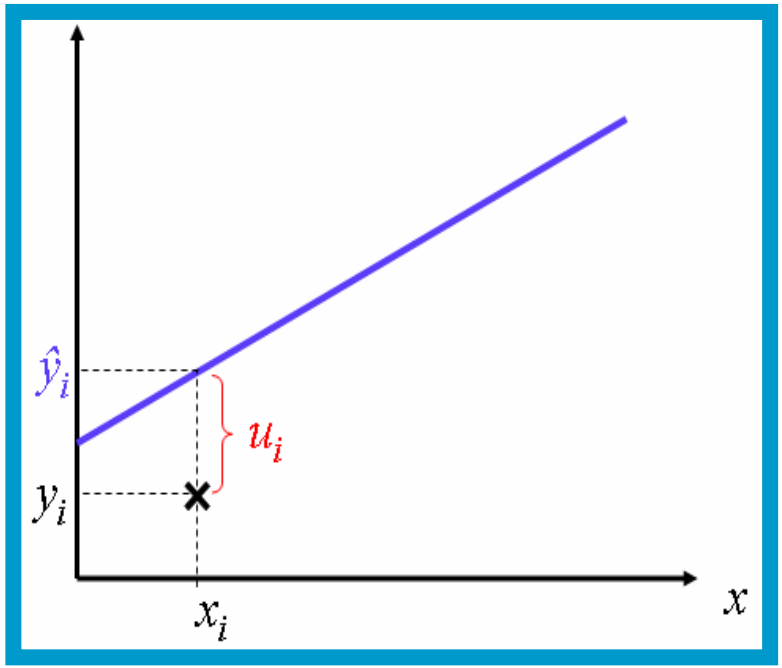
$$RSS = (1-R^2) \times TSS$$

Diagrammatic representation of TSS, REGSS and RSS:

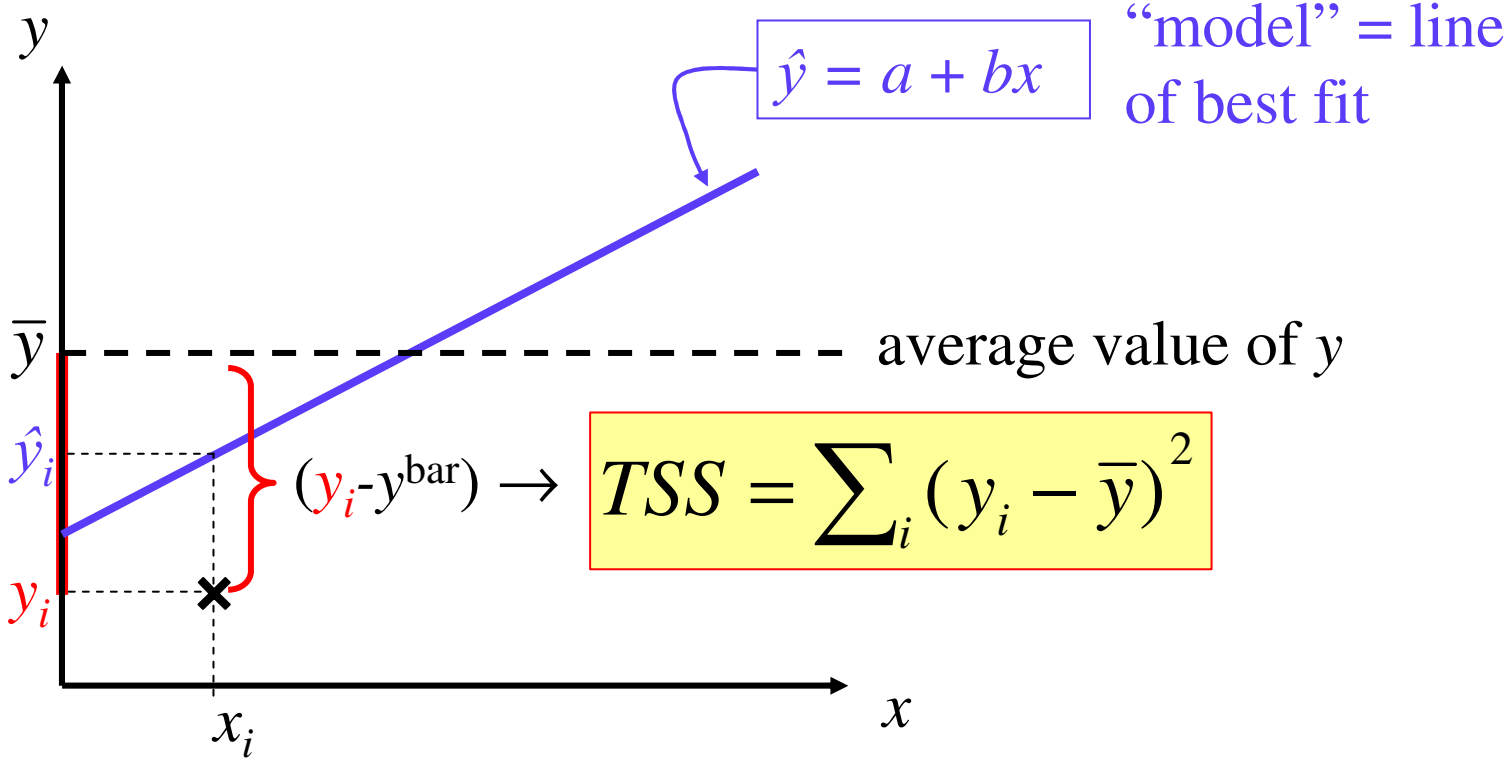
$$TSS = \sum_i (y_i - \bar{y})^2$$

$$REGSS = \sum_i (\hat{y}_i - \bar{y})^2$$

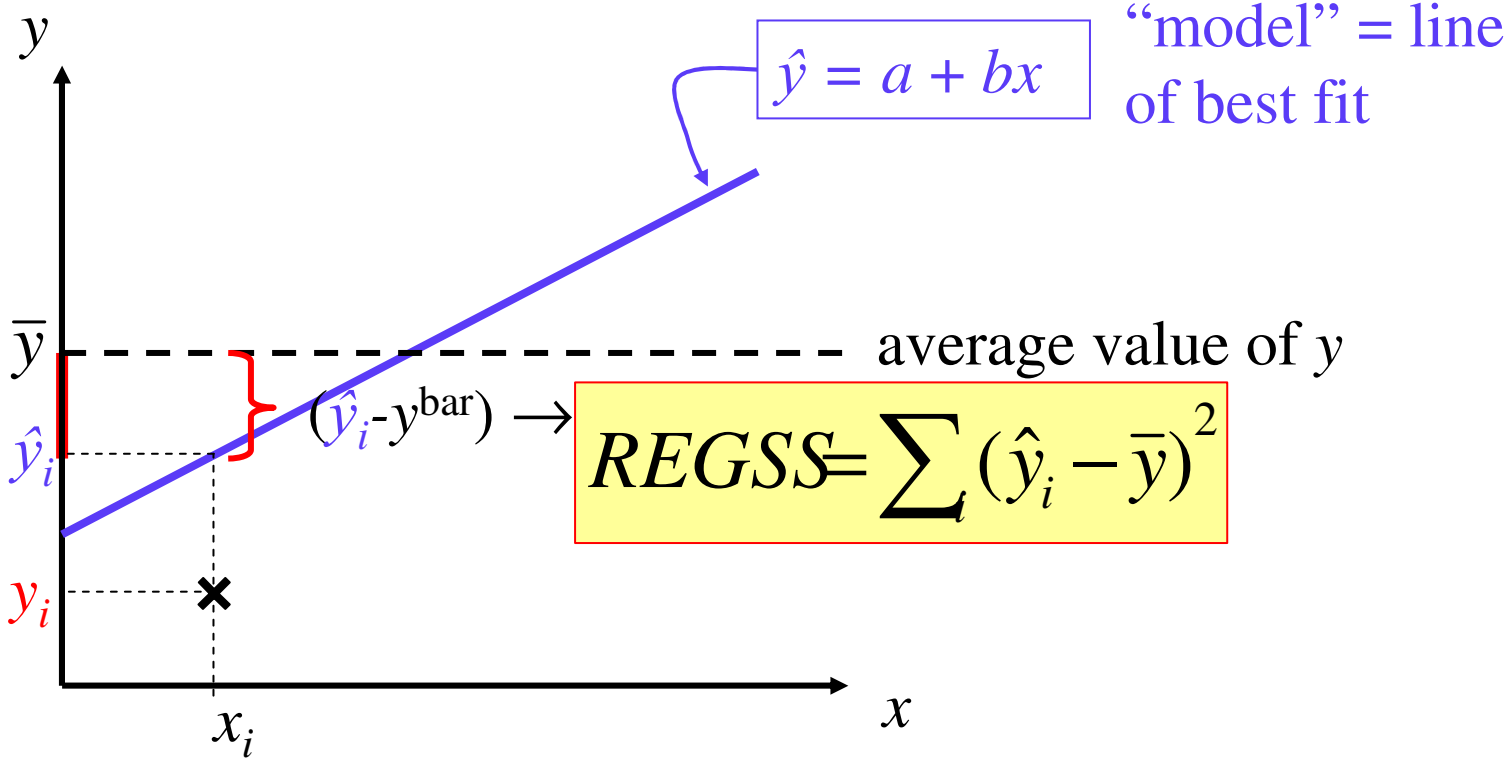
$$RSS = \sum_i (y_i - \hat{y}_i)^2$$



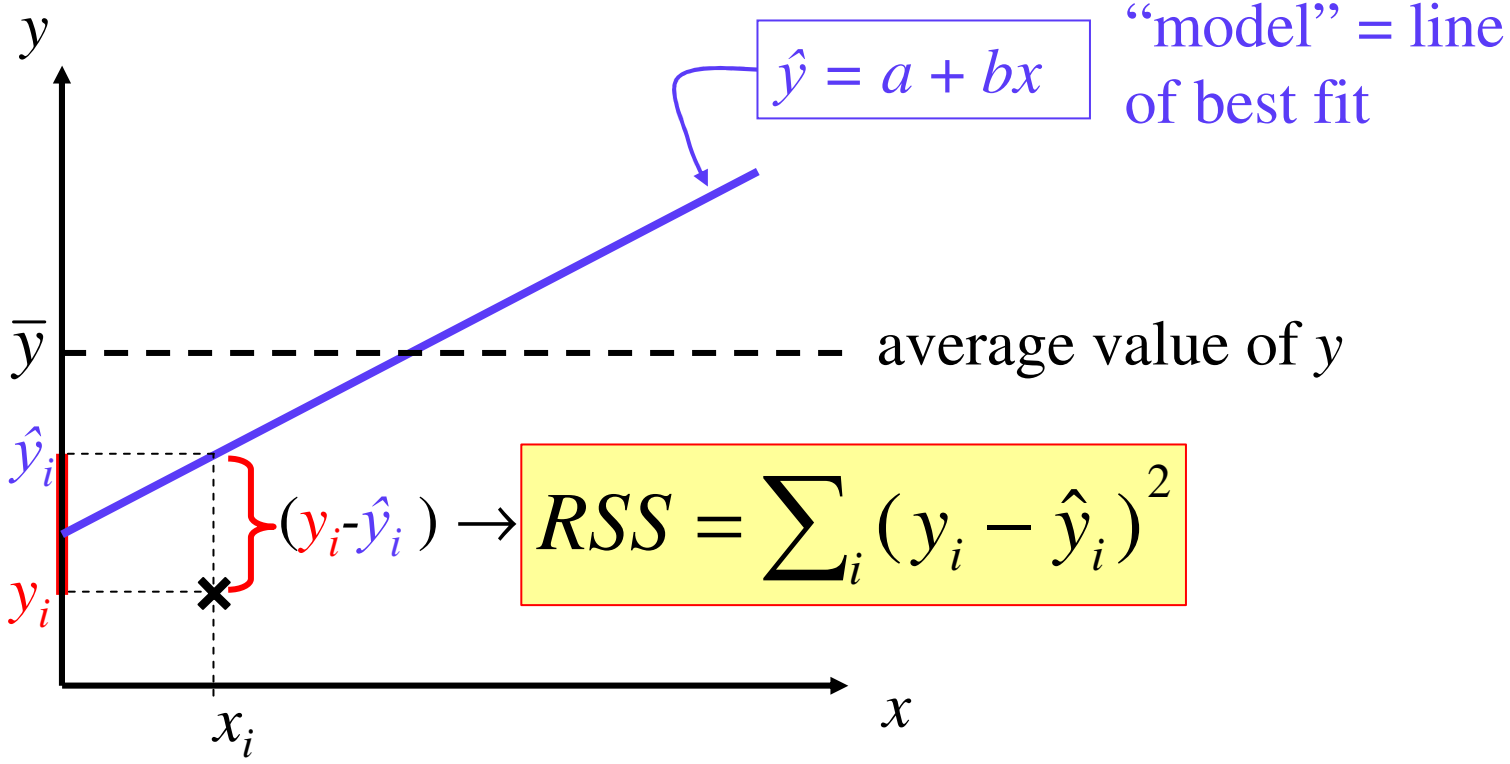
Diagrammatic representation of TSS for particular observation *i*:



Diagrammatic representation of REGSS for particular observation *i*:



Diagrammatic representation of RSS for particular observation *i*:



SPSS ANOVA table explained

Variation in y accounted for by your model

Mean Square = REGSS / df_{REGSS}

Model		Sum of Squares	df	Mean Square	F	Sig.
1	**Regression**	**11101.959**	**2**	**5550.980**	684.793	.000
	Residual	843.031	104	8.106		
	Total	11944.991	106			

Variation in y *not* accounted for by your model

Mean Square = RSS / df_{RSS}

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11101.959	2	5550.980	684.793	.000
	Residual	**843.031**	**104**	**8.106**		
	Total	11944.991	106			

Total variation in y:

TSS = REGSS + RSS

$F = MS_{REGSS} / RSS_{REGSS}$

Tests the H_0 that there is no relationship between y and any of the x variables in the model

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11101.959	2	5550.980	684.793	.000
	Residual	843.031	104	8.106		
	Total	**11944.991**	**106**			

A model with a large regression sum of squares in comparison to the residual sum of squares indicates that the model accounts for most of variation in the dependent variable.

whereas

Very high residual sum of squares indicate that the model fails to explain a lot of the variation in the dependent variable, and you may want to look for additional factors that help account for a higher proportion of the variation in the dependent variable.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	**Regression**	**11101.959**	2	5550.980	684.793	.000
	Residual	**843.031**	104	8.106		
	Total	11944.991	106			

The mean square is the sum of squares divided by the degrees of freedom.

The F statistic is the regression mean square (MSR) divided by the residual mean square (MSE).

The regression degrees of freedom is the numerator df and the residual degrees of freedom is the denominator df for the F statistic.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	**Regression**	11101.959	**2**	5550.980	**684.793**	.000
	Residual	843.031	**104**	8.106		
	Total	11944.991	106			

The total number of degrees of freedom is the number of cases minus 1.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11101.959	2	5550.980	684.793	.000
	Residual	843.031	104	8.106		
	Total	11944.991	**106**			

3. The F-Test

- These sums of squares, particularly the RSS, are useful for doing hypothesis tests about *groups* of coefficients.
- The test statistic used in such tests is the F distribution:

$$F = \frac{(RSS_R - RSS_U)/r}{RSS_U/(n-k-1)}$$

Where:

RSS_U = unrestricted residual sum of squares = RSS under H_1

RSS_R = unrestricted residual sum of squares = RSS under H_0

r = number of restrictions

Test for $\beta_k = 0 \; \forall k$

- The most common group coefficient test is that $\beta_k = 0 \; \forall \; k$. (NB $\forall$ means "for all")
  - i.e. there is no relationship between y and any of the explanatory variables.
  - The hypothesis test has 4 steps:

    (1) H_0: $\beta_k = 0 \; \forall \; k$

    H_1: $\beta_k \neq 0 \; \forall \; k$

    (2) $\alpha = 0.05$,

    $$F = \frac{(RSS_R - RSS_U)/r}{RSS_U/(n-k-1)}$$

    (3) Reject H_0 iff Prob($F > F_c$) < α

    (4) Calculate P = Prob($F > F_c$) and conclude.

    (P is the "Sig." value reported by SPSS in the ANOVA table)

- For this particular test:

RSS_U = RSS under H_1 = RSS

RSS_R = RSS under H_0 = TSS

(RSS_R = TSS under H_0 because if all coeffs were zero, the explained variation would be zero, and so error element would comprise 100% of the variation in TSS, I.e. RSS under H_0 = 100% TSS = TSS)

r = number of restrictions

= number of slope coefficients in the regression that we are restricting

= equals all slope coefficients = k

- For this particular test, the F statistic reduces to $(R^2/k)/((1-R^2)/(n-k-1))$ so it isn't telling us much more than the R^2

Proof of alternative F calculation:

$$F = \frac{(RSS_R - RSS_U)/r}{RSS_U/(n-k-1)}$$

$$= \frac{(TSS - RSS)/k}{RSS/(n-k-1)}$$

$$= \frac{(TSS - (1-R^2)TSS)/k}{(1-R^2)TSS/(n-k-1)} = \frac{-R^2TSS/(k+1)}{(TSS - R^2TSS)/(n-k-1)}$$

$$= \frac{R^2/(k+1)}{(1-R^2)/(n-k-1)}$$

If the significance value of the F statistic is small (smaller than say 0.05) then the independent variables do a good job explaining the variation in the dependent variable.

If the significance value of F is larger than say 0.05 then the independent variables do not explain the variation in the dependent variable.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11101.959	2	5550.980	684.793	.000
	Residual	843.031	104	8.106		
	Total	11944.991	106			

Source of Variation	*Sum of squares*	*Degrees of Freedom df*	*Average square = (sum of squares)/df*	*F*
Regression	R^2TSS	k	REGSS / k $= R^2$TSS / k	$F = \frac{REGSS/k}{RSS/(n-k-1)} = \frac{R^2TSS/(k)}{(1-R^2)TSS/n-k-1}$
Residual	(1-R^2)TSS	$n-k-1$	RSS /$(n-k-1)$ $=$ (1-R^2)TSS/$(n-k-1)$	
Total	TSS	$n-1$		

- Very simply, the ANOVA table F-test can be thought of as the ratio of the mean regression sum of squares and the mean residual sum of squares:

  F = regression mean squares / residual mean squares

  – if the line of best fit is good, F is large:
    - the improvement in prediction due the regression will be large (so regression mean squares is large)
    - the difference between the regression line and the observed data will be small (residual MS is small)

House Price Equation Example:

el		Unstandardized Coefficients B	Std. Error
	(Constant)	306.981	3174.456
	AGEDW_SQ	4.636E-02	.093
	TERRACE	-22538.1	2287.060
	FLOORARE	615.245	25.081

Dependent Variable: PURCHASE

R Square
.594

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.96E+11	3	1.655E+11	269.248	.000[a]
	Residual	3.39E+11	552	614589637.9		
	Total	8.36E+11	555			

a. Predictors: (Constant), FLOORARE, TERRACE, AGEDW_SQ

b. Dependent Variable: PURCHASE

4. Regression assumptions

For estimation of *a* and *b* and for regression inference to be correct:

- 1. Equation is correctly specified:
  - Linear in parameters (can still transform variables)
  - Contains all relevant variables
  - Contains no irrelevant variables
  - Contains no variables with measurement errors
- 2. Error Term has zero expected mean
- 3. Error Term has constant variance

- 4. Error Term is not autocorrelated
  - I.e. correlated with error term from previous time periods
- 5. Explanatory variables are fixed
  - observe normal distribution of *y* for repeated fixed values of *x*
- 6. No linear relationship between RHS variables
  - I.e. no “multicolinearity”

5. Properties of OLS estimates

- If the above assumptions are met, OLS estimates are said to be BLUE:
  - Best — I.e. most efficient = least variance
  - Linear — I.e. best amongst linear estimates
  - Unbiased — I.e. in repeated samples, mean of *b* $= \beta$
  - Estimates — I.e. estimates of the population parameters.

Summary

- 1. ANOVA in regression
- 2. Prediction
- 3. F-Test
- 4. Regression assumptions
- 5. Properties of OLS estimates

Reading:

- Chapter 2 of Pryce's notes on *Advanced Regression in SPSS*
- Chapters 1 and 2 of Kennedy *"A Guide to Econometrics"*
- Achen, Christopher H. *Interpreting and Using Regression* (London: Sage, 1982), sections 2 and 5.
- Chapter 4 of Andy Field, "*Discovering statistics using SPSS for Windows : advanced techniques for the beginner*".
- Chapters 1 & 2 of Wooldridge *Introductory Econometrics*