

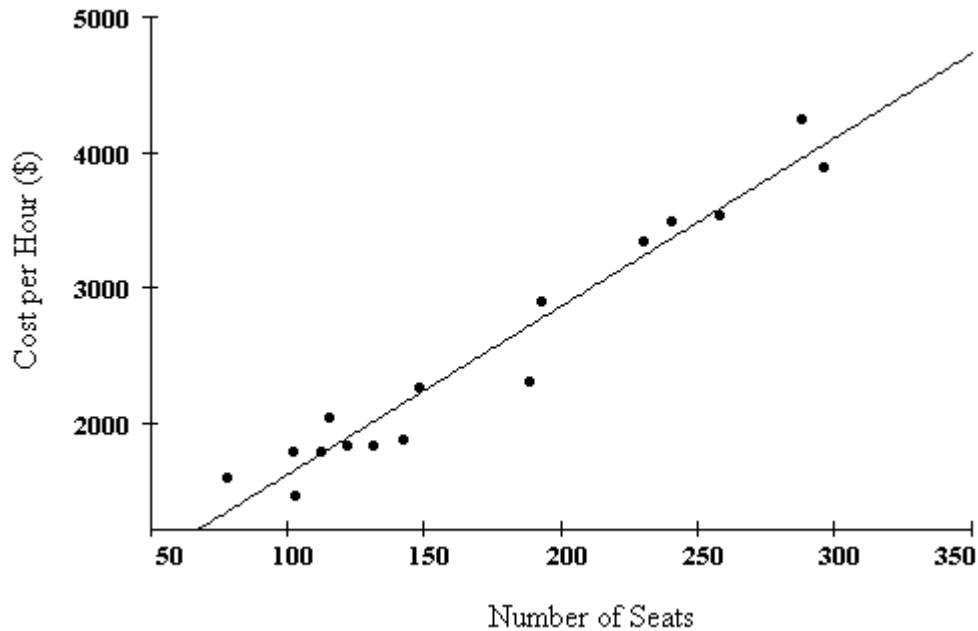
That ANOVA table at the bottom of Regression Output

Suppose we fit a linear model to the data below. The data give the average number of passenger seats in different airplanes flying in the US and the average cost per hour in flight for that airplane. We believe a linear model is appropriate, and we want to be able to predict the cost from the number of seats.

Plane	L-1011	DC-10	A300	A310	B767-300	B767-200	B757	B727-200
Seats	296	288	258	240	230	193	188	148
Cost	3885	4236	3526	3484	3334	2887	2301	2247

Plane	MD-80	B737-300	DC-9-50	B727-100	B737-200	F-100	DC-9-30	DC-9-10
Seats	142	131	122	115	112	103	102	78
Cost	1861	1826	1830	2031	1772	1456	1778	1588

1992 World Almanac and Book of Facts and <http://www.air-transport.org>



Along with the regression equation

$$\text{Cost per Hour} = 362.363 + 12.4706 \text{ Number of Seats}$$

you get a lot of additional information. One piece of information is the Analysis of Variance table. Just where do those numbers come from?

The output below is fairly typical of statistical packages.

Cost per Hour By Number of Seats

Linear Fit

$$\text{Cost per Hour} = 362.363 + 12.4706 \text{ Number of Seats}$$

Summary of Fit

RSquare	0.951609
RSquare Adj	0.948152
Root Mean Square Error	207.0796
Mean of Response	2502.625
Observations (or Sum Wgts)	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	11805736	11805736	275.3076
Error	14	600348	42881.98	Prob>F
C Total	15	12406084		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	362.36339	138.9916	2.61	0.0207
Number of Seats	12.47057	0.751583	16.59	<.0001

Number of Seats		Cost per Hour	
Mean	171.6250	Mean	2502.625
Std Dev	71.1401	Std Dev	909.435
Std Error Mean	17.7850	Std Error Mean	227.359
N	16.0000	N	16.000

Plane	L-1011	DC-10	A300	A310	B767-300	B767-200	B757	B727-200
Seats	296	288	258	240	230	193	188	148
Cost	3885	4236	3526	3484	3334	2887	2301	2247
Fit	4053.7	3953.9	3579.8	3355.3	3230.6	2769.2	2706.8	2208.0
Residual	-168.7	282.1	-53.8	128.7	103.4	117.8	-405.9	39.0

Plane	MD-80	B737-300	DC-9-50	B727-100	B737-200	F-100	DC-9-30	DC-9-10
Seats	142	131	122	115	112	103	102	78
Cost	1861	1826	1830	2031	1772	1456	1778	1588
Fit	2133.2	1996.0	1883.8	1796.5	1759.1	1646.8	1634.4	1335.1
Residual	-272.2	-170.0	-53.8	234.5	12.9	-190.8	143.6	252.9

Just what is it that has a sum of squares 11,805,736 (called the Model Sum of Squares), of 600,348 (called the Error Sum of Squares), and of 12,406,084 (called the Total Sum of Squares)?

For simplicity, we will call Cost per Hour, y , and Number of Seats, x .

The Total Sum of Squares is just the sum of the squares of the differences in the y -values and the average y -value, 2502.625. Symbolically, this is $\sum (y_i - \bar{y})^2$. From our data, we have

$$(3885 - 2502.625)^2 + (4236 - 2502.625)^2 + \dots + (1588 - 2502.625)^2 = 12,406,084.$$

This sum is the numerator in the variance of y . If we square the standard deviation of y (909.435) and multiply by 15, which is $(n-1)$, we should get this total sum of squares, $909.435^2(15) = 12,406,080$, except for some round-off error.

The Model Sum of Squares is the sum of the squares of the differences in the predicted y -values and the average y -value 2502.625. Symbolically, this is $\sum(\hat{y}_i - \bar{y})^2$. From our data we have

$$(4053.7 - 2502.625)^2 + (3953.9 - 2502.625)^2 + \dots + (1335.1 - 2502.625)^2 = 11,805,736.$$

Finally, the Error Sums of Squares is just the sum of squares of the residuals, or

$\sum(y_i - \hat{y}_i)^2$. From our data we have

$$(3885 - 4053.7)^2 + (3953.9 - 2502.625)^2 + \dots + (1335.1 - 2502.625)^2 = 600,348.$$

Notice that $R^2 = \frac{11805736}{12406080} = 0.0516$ and that $600,348 + 11,805,736 = 12,406,084$.

Further, $\frac{600348}{14} + \frac{11805736}{14} = \frac{12406084}{14}$ and the expression on the right side of this equation is the variance of y . So, on the left side of the equation, we have partitioned the variance of y into two groups, one group that is associated with the fitted values (explained by x) and the other with the residuals (unexplained by x). Dividing both sides of this equation by $\frac{12406084}{14}$ gives the standard interpretation of R^2 , the proportion of the variance of y attributed to or explained by x .