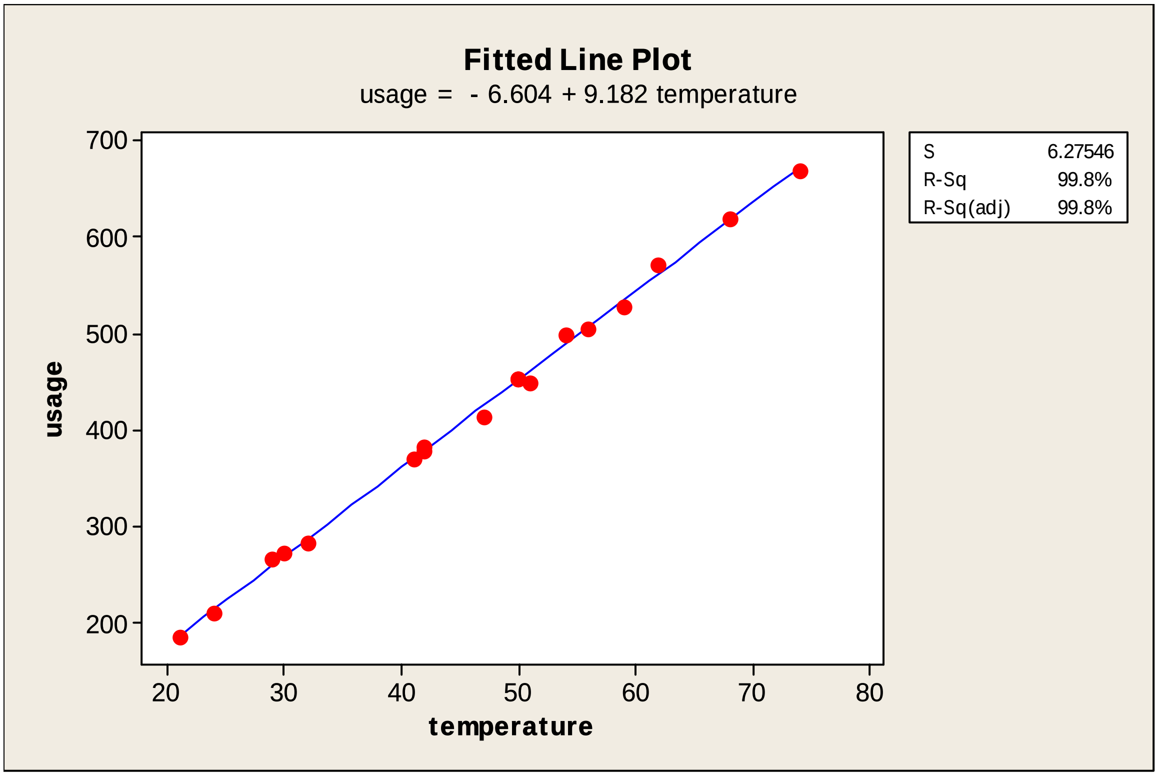
Minitab Output for the same problem



Regression Analysis: usage versus temperature

```
The regression equation is
usage = - 6.60 + 9.18 temperature


Predictor       Coef  SE Coef      T      P
Constant      -6.604    4.920  -1.34  0.198
temperature   9.1820   0.1015  90.45  0.000

S = 6.27546   R-Sq = 99.8%   R-Sq(adj) = 99.8%
```

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	322156	322156	8180.42	0.000
Residual Error	16	630	39		
Total	17	322786			

Section 7.3: Uncertainties in the Least-Squares Coefficients

Recall the formal model:

$$Y=\beta_0+\beta_1 x_i+\varepsilon_i$$

where β_0 and β_1 are unknown and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ for i=1,..,n.

Here the assumptions are:

i) The errors $\varepsilon_1,\ldots,\varepsilon_n$ are independent random variables.

ii) The errors have the same mean 0.

iii) The errors have the same variance σ^2

iv) The errors are normally distributed

We don't know the value of σ^2. So we estimate σ^2 by

$$s^2 = \frac{1}{n-2}\sum\left(y_i - \hat{y}_i\right)^2 = \frac{\sum e_i^2}{n-2} = \frac{SSE}{n-2}$$

s^2 is measure of the spread of the points of the residuals around the line.

The denominator (n-2) is the degrees of freedom (df)

Recall: $\hat{\beta}_0$, $\hat{\beta}_1$ are the least-squares estimates of the **unknown** regression coefficients β_0 and β_1 respectively

Inferences on the Slope

- β_1 is the true change in the mean of y with an increase of one unit in the value of x
- $\hat{\beta}_1$ estimates the true population slope β_1

- $\hat{\beta}_1$ is normally distributed r.v. with mean β_1
- Test: We often want to see if β_1 is significantly different from 0. If β_1 is significantly different from zero, then x is considered a good predictor of y.

How do you test this?

- CI for β_1
- Significance test for β_1
- ANOVA F test (we will discuss this later)

<u>Confidence Intervals and Hypothesis Test for Slope,*β_1*</u>

- Keep in mind $\hat{\beta}_1$ is a statistic, not exactly equal to true slope *β_1*. Every sample will have a different $\hat{\beta}_1$.
- The sampling distribution of $\hat{\beta}_1$ is

$$N\left(\beta_{1,}\ \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right)$$

- We estimate σ^2 with s^2
- $$\sqrt{\left(\frac{s^2}{\sum (x_i - \bar{x})^2}\right)} = s_{\hat{\beta}_1}$$

Note: Notice that the spread of the x values effect the value of $s_{\hat{\beta}_1}$. You want to have as much spread in your x values as possible as long as you don't go beyond the range where the linear model holds.

Parameter	Statistic	Std. Error	Sampling Distribution
		$\sqrt{\left(\frac{s^2}{\sum(x_i-\bar{x})^2}\right)} = s_{\hat{\beta}_1}$	t(n-2)

Confidence Interval for β_1

Hypothesis Test for β_1

Ho: ____________ H_1: ______________

Test Statistic: $\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} \sim t_{n-2}$ under H_0

(since s estimates σ and df for s is (n-2))

- You can follow standard t-procedure for the test

Note: You can also test that the Ho: $\beta_1 = \#$ versus Ha: $\beta_1 \neq \#$, this test to see if y changes by # for every one unit increase in x.

Ex:Refer to the previous example

Regression Analysis: usage versus temperature

```
The regression equation is
usage = - 6.60 + 9.18 temperature


Predictor       Coef  SE Coef      T      P
Constant      -6.604    4.920  -1.34  0.198
temperature   9.1820   0.1015  90.45  0.000

S = 6.27546   R-Sq = 99.8%   R-Sq(adj) = 99.8%
```

The Analysis of Variance (ANOVA) Table

Source	df	SS	MS	F	p-value
Model	1	SSR	MSR= SSR / 1	MSR/ MSE	
Error	n-2	SSE	MSE= SSE /(n-2)		
Total	n-1	SST			

Recall the ANOVA identity:

$$SST=SSR+SSE$$

- MSE – variability of points around line
- s^2= MSE = estimate of σ^2
- $s = \sqrt{MSE}$

- $R^2 = \dfrac{SSR}{SST}$

F Test for Slope

Ho: $\beta_1=0$ vs H$_1$: $\beta_1 \neq 0$

Test Statistic: $F = \dfrac{MSR}{MSE} \sim F(1,n-2)$

What is the F distribution?

$$f(x) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\left(\frac{\nu_1}{2}-1\right)}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(1+\frac{\nu_1 x}{\nu_2}\right)^{\left(\frac{\nu_1+\nu_2}{2}\right)}}, \quad x>0$$

ν_1 and ν_2 are the degrees of freedom for the F distribution
If X is said to have an F distribution with df ν_1 and ν_2, it can be denoted as $F(\upsilon_1,\upsilon_2)$.

The cdf Q of the F distribution follows the relationship: $Q_{\nu_1,\nu_2}(p) = \dfrac{1}{Q_{\nu_2,\nu_1}(1-p)}$

F distribution is used to compare two variances