A multiple linear regression that describes the variation in the variable y in terms of the k (explanatory) variables $x_1, x_2, \ldots, x_k$ has the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon.$$

- $\beta_0, \beta_1, \ldots, \beta_k$ are population parameters.
- Assumptions about the model include
  - (i) The error ε has a normal distribution with mean 0 for each fixed set of values $x_1 = \xi_1, x_2 = \xi_2, \ldots, x_k = \xi_k$.
  - (ii) The error is independent of the x_j s, so the standard deviation of ε is fixed. I.e., $\varepsilon \sim N(0, \sigma^2)$. (Homoskedasticity)
- These assumptions imply that

$$E(y|x_1 = \xi_1, x_2 = \xi_2, \ldots, x_k = \xi_k) = \beta_0 + \beta_1 \xi_1 + \beta_2 \xi_2 + \cdots + \beta_k \xi_k,$$

  where $E(y|x_1 = \xi_1, x_2 = \xi_2, \ldots, x_k = \xi_k)$ is the *expected (mean) y value* for all observations satisfying $x_1 = \xi_1, x_2 = \xi_2, \ldots, x_k = \xi_k$.

Using sample data, we compute the estimated regression equation

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k.$$

- The sample coefficients $b_0, b_1, \ldots, b_k$ are sample statistics (that estimate the population coefficients $\beta_0, \ldots, \beta_k$). As such, $b_0, \ldots b_k$ are all random variables, and the assumptions above imply that they all have *normal* distributions. In particular...

  ... for each i between 0 and k:

$$\frac{b_i - \beta_i}{SE(b_i)} \sim t\text{-distribution, with } n - (k+1)\ d.f.$$

  We can use this to test individual coefficients for statistical significance:

$$H_0 : \beta_i = 0 \quad \text{vs.} \quad H_a : \beta_i \neq 0.$$

  Rejecting H_0 means that there is significant linear correlation between x_i and y, and b_i is a reliable measure of the *marginal change* in y for a one-unit change in x_i, assuming that all other variables are held fixed.

We can also test the *overall significance* of the regression:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

$H_a : \beta_j \neq 0$, *for at least one* j *between* 1 *and* k.

The test statistic is

$$F^* = \frac{MSS_m}{MSS_e} = \frac{\left(\sum (\hat{y}_i - \overline{y})^2\right)/k}{\left(\sum (y_i - \hat{y}_i)^2\right)/(n-k+1)}$$

which follows the ***F-distribution with k numerator d.f. and*** $\boldsymbol{n-k+1}$ ***denominator d.f.***

H_0 is rejected at the α level of significance if $F^* > F_\alpha$ (the critical F-value). Alternatively, reject H_0 if the P-value, $Prob(F > F^*)$, is smaller than α.

Comments:

1. For a *simple regression* $\overline{y}(x) = \beta_0 + \beta_1 x$, testing $H_0 : \beta_1 = 0$ (t-test) yields the same conclusion as the F-test for overall significance.

2. Most software packages (including JMP) compute the t-scores for all of the regression coefficients and their p-values, as well as the F-score of overall significance and its p-value.

3. One can also test other hypotheses about individual coefficients, e.g.,

$$H_0 : \beta_3 = 2 \text{ vs. } H_a : \beta_3 < 2.$$

With this null hypothesis,

$$\frac{b_3 - 2}{SE(b_3)} \sim t\text{-distribution, with } n - (k+1) \; d.f.,$$

and the mechanics of the test are the same as any other t-test.

Example: Estimating weight with a measuring tape.

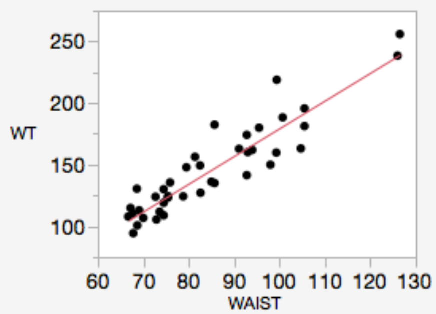
Response variable: weight (lbs). Effect variables: waist circumfrence (cm), height (inches), age (years), arm circumfrence (cm).

FHEALTH.JMP: Fit Least Squares

Response WT

Whole Model

Regression Plot



Summary of Fit

RSquare	0.848464
RSquare Adj	0.844476
Root Mean Square Error	14.83642
Mean of Response	146.22
Observations (or Sum Wgts)	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	46833.830	46833.8	212.7657
Error	38	8364.534	220.1	Prob > F
C. Total	39	55198.364		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	-45.1058	13.32477	-3.39	0.0017*	-72.08038	-18.13123
WAIST	2.2500315	0.154254	14.59	<.0001*	1.9377596	2.5623034

FHEALTH.JMP: Fit Least Squares

Response WT

Whole Model

Summary of Fit

RSquare	0.885628
RSquare Adj	0.879446
Root Mean Square Error	13.06238
Mean of Response	146.22
Observations (or Sum Wgts)	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	48885.206	24442.6	143.2526
Error	37	6313.158	170.6	Prob > F
C. Total	39	55198.364		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	-207.6542	48.32505	-4.30	0.0001*	-305.57	-109.7383
WAIST	2.1588224	0.138334	15.61	<.0001*	1.8785313	2.4391135
HT	2.6948985	0.777217	3.47	0.0013*	1.1201072	4.2696897

FHEALTH.JMP: Fit Least Squares

Response WT

Whole Model

Summary of Fit

RSquare	0.888886
RSquare Adj	0.879626
Root Mean Square Error	13.0526
Mean of Response	146.22
Observations (or Sum Wgts)	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	49065.035	16355.0	95.9969
Error	36	6133.329	170.4	Prob > F
C. Total	39	55198.364		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	-209.9287	48.33956	-4.34	0.0001*	-307.9658	-111.8915
WAIST	2.2814176	0.182611	12.49	<.0001*	1.9110662	2.6517691
HT	2.6865238	0.776677	3.46	0.0014*	1.1113492	4.2616984
AGE	-0.229371	0.223257	-1.03	0.3111	-0.682158	0.2234156

FHEALTH.JMP: Fit Least Squares

Response WT

Whole Model

Summary of Fit

RSquare	0.934142
RSquare Adj	0.928654
Root Mean Square Error	10.04885
Mean of Response	146.22
Observations (or Sum Wgts)	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	51563.103	17187.7	170.2098
Error	36	3635.261	101.0	Prob > F
C. Total	39	55198.364		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	-278.4536	39.63704	-7.03	<.0001*	-358.8413	-198.066
WAIST	1.05752	0.238874	4.43	<.0001*	0.5730619	1.5419781
HT	3.601471	0.623289	5.78	<.0001*	2.3373827	4.8655593
ARM	3.5531821	0.689981	5.15	<.0001*	2.1538359	4.9525283