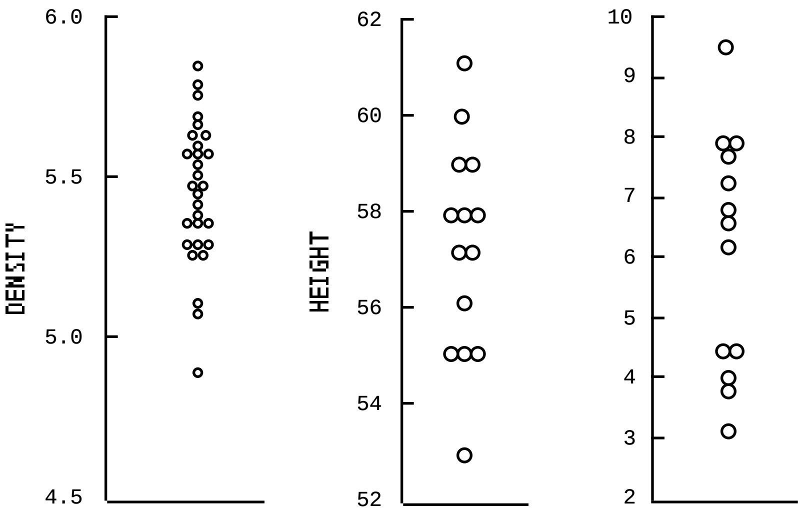
- 3 examples

Cavendish's 29 measurements of the earth's density | Heights (inches) of 14 11 year-old males from Alberta study | Half-life of Caffeine (hours) in n=13 healthy non-smokers



- Statistics

n	29	14	13
MIN	4.88	53.00	9.40
MAX	5.85	61.00	5.95
MEAN ($\bar{y}$)	**5.45**	**57.21**	**5.95**
VARIANCE s^2	0.0488	4.9506	3.9460
SD (**s**)	**0.22**	**2.22**	**1.99**

- Least Squares Estimate of μ

$\Sigma(y - \bar{y})^2$ is smaller than Σ(y – any other measure of the centre)^2

That's why we can call the statistic $\bar{y}$ the **Least Squares** estimator of μ.
(see applet on best location to wait for elevator in Ch 1 Resources for 607, and 'elevator article' in Ch 1 of Course 697; see also applets in Ch 10 for 607)

- Statistical Model

$$y = \mu + \varepsilon$$

$$\varepsilon \sim ?(0,\sigma)$$

- Note about shorthand

The shorthand ~ ?(0,σ) is used for some distribution with mean 0 and standard deviation σ. Some authors use variance rather than sd: notation ~ ?(0,σ^2).

- Note about "Minimum Requirements" for Least Squares Estimation

There is no requirement that ε ~ N(0,σ) i.e that the ε's be "Normal" i.e. Gaussian. Later, for statistical inferences about the parameters being estimated, the inferences may be somewhat inaccurate if n is small and the distribution of the ε's is not N(0,σ) or if the ε's are not independent of each other.

- Fitting (i.e. calculating the parameter estimates of) the model for height

By calculator or means procedure: $\bar{y} = \frac{\Sigma y}{n} = 57.21$

$$\mathbf{s}^2 = \frac{\Sigma(y - \bar{y})^2}{n-1} = \frac{64.357}{13} = 4.95 \; (\mathbf{s} = 2.2)$$

By Mystat/SYSTAT regression program
MODEL HEIGHT = CONSTANT
ESTIMATE

Output from Systat Regression Program:

```
DEP VAR: HEIGHT    N:14  MULTIPLE R: 0.0    SQUARED MULTIPLE R: 0.0
ADJUSTED SQUARED MULTIPLE R: 0.0   STANDARD ERROR OF ESTIMATE: 2.22

VARIABLE   COEFFICIENT  STD ERROR   STD COEF TOLERANCE  T    P(2 TAIL)
CONSTANT      57.21       0.59465   0.00000     .        96.0  0.00
```

- Finding parameter estimates on output of statistical packages

If you compare with the calculations above, you will readily identify the estimate $\bar{y}$ = 57.21 for the μ parameter. But what is the estimate of the σ^2 or σ parameter? We know from our calculator that s = 2.22. . In the SYSTAT output (SAS output later!), this estimate is given under the not-very-informative term STANDARD ERROR OF ESTIMATE. i.e.

$$\hat{\sigma}^2 = \frac{\Sigma(y - \bar{y})^2}{n-1} \; ; \; \hat{\sigma} = \text{STANDARD ERROR OF ESTIMATE} = = 2.22$$

(SPSS uses this SEE terminology too!)

You can think of each (y – $\bar{y}$) as the 'residual' variation from the mean, and you can therefore call $\Sigma(y - \bar{y})^2$ the Sum of Squares of the Residuals, or Residual Sum of Squares for short.

• What of the other items output by the regression program?

What is `STD ERROR = 0.59465`? It is the SE of CONSTANT i.e. of $\bar{y}$. It is what we called the SEM in Chapter 7, where it was given by the formula

Standard Error of Mean = SEM = SE($\bar{y}$) = s / √n = 2.22 / √ 14 = 0.59.

What is T = 96? (actually, it was t = .96E+02 before I translated it to the more friendly 0.96 x 100 = 96) it is the test statistic corresponding to the test of whether the underlying parameter (μ in our case) is ZERO i.e. of the H_0: μ=0. Of course, the silly computer programmer doesn't know what μ refers to, or that the mean height of 11 year old boys in Alberta cannot be zero. Since we might have a case where there was genuine interest in the H_0 that μ=0, we will show where t = 96 came from: remember from earlier the 1-sample t-test and the formula

t = (ybar-0)/SE(ybar)= 57.21/0.59 = 96 (if don't fuss about rounding errors)

What is `P(2 TAIL) = 0.00`? (it was `P(2 TAIL)` = 0.00000 before I truncated it)

It is the P-value obtained by calculating the probability that an observation from the t distribution with n–1 = 13 df would exceed 96 in absolute value.

What are `STD COEF and TOLERANCE`? Lets not worry about them for now!

• Fitting the beginning of all regression models using SAS

```
proc reg data=sasuser.alberta;
     where ( I_Female = 0 and age =11 ) ;
 model height = ;
run;
```

(I discovered this way of calculating ybar by accident—I forgot to put terms on the R hand side of a regression equation! It works the same way in INSIGHT)

The model is simply

$$y = \mu + \varepsilon$$

but it can be thought of as

$$y = \mu.1 + \varepsilon$$

$$y = \mu.x_0 + \varepsilon$$

where $x_0 \equiv 1$ (a constant); it is as though we have set it up so that the "predictor variable" x_0 in the regression equation is always 1. Then μ is the parameter to be estimated.

Some software programs insist that you specify the constant; others assume it unless told otherwise.

• Output from SAS PROC REG (see "fitting μ via regression" -- under resources for Ch 10)

Dependent Variable: HEIGHT

Analysis of Variance*

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	0	0.000	.	.	.
Error	13	64.357	4.95		
C Total	13	64.357			

Root MSE	2.22	R-square	0.0000
Dep Mean	57.21	Adj R-sq	0.0000
C.V.	3.89		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	57.21	0.59	96.2	0.0001

Notice SAS uses the word "INTERCEP" rather than CONSTANT ... and because all names before SAS version 8 were restricted to 8 letters, INTERCEPT gets shortened to "INTERCEP".

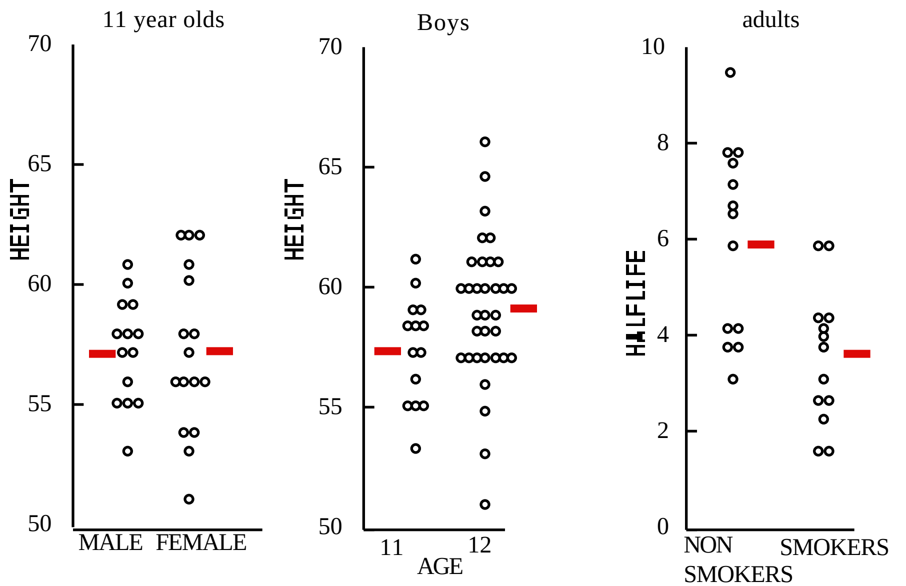
More importantly, note the name SAS gives to the square root of the average of the squared residuals.. ROOT MEAN SQUARE ERROR, shortened to ROOT MSE ie.

average squared deviation = 64.357/13 = 4.95; √4.95 = 2.22
here they are less confusing than SPSS and SYSTAT (to be fair, SEE is used a lot in measurement and psychophysics for variation of measurements on individuals [ie no √n involved], rather than of statistics)

• (*) The ANOVA TABLE

Usually, we are not interested in the overall mean μ of Y but rather in the 'effects' of variables x1, x2, x3 on the mean Y at each X configuration. In such situations, the 'remaining' y variation is measured from the fitted mean for each configuration of x's; here we have no explanatory variables x1 x2 x3. We cannot "explain" the variation in height reflected in the Error Sum of Squares 64.357 or the Error Mean Square = 64.357/13 = 4.95 or the Root Mean Square Error (RMSE) =sqrt[4.95] = 2.22. In analyses where there are explanatory variables x1 x2 x3... (rather than the constant x_0 we used here) the Anova Table will use the overall $\Sigma(y - \bar{y})^2$, which SAS calls the "Corrected Total Sum of Squares" as the bottom line SStotal, and R-Square will be the Model SS as a proportion of this SStotal. If we add variables x1, x2, x3... to the regression above, then the 64.35 will become the SStotal to be further partitioned into SSmodel and SSerror $\Sigma(y_{x1,x2,x3} - \hat{\mu}_{x1,x2,x3})^2$.

• For " $\hat{\mu}$ via regression" for density and caffeine 1/2 life, see resources for Ch 10.



n	14	16		14	33		13	13	
			$\bar{y} - \bar{y}$			$\bar{y} - \bar{y}$			$\bar{y} - \bar{y}$
$\bar{y}$	57.21	57.25	0.04	57.21	59.00	1.79	5.95	3.53	-2.42
s	*2.22*	*3.41*		*2.22*	*3.05*		*1.99*	*1.43*	

INDEPENDENT SAMPLES T-TEST based on $\bar{y} - \bar{y}$

VAR†	t	DF	PROB	t	DF	PROB	t	DF	PROB
S	0.03	26.0	0.9729	2.24	33.4	0.0319	-3.56	21.8	0.0018
P	0.03	28	0.9736	1.97	45	0.0547	-3.56	24	0.0016

VAR† S=SEPARATE VARIANCES T-TEST   P=POOLED VARIANCES* T-TEST

* (for later)

first panel (heights of males vs females)

$$\text{Pooled variance} = \frac{13(2.22)^2 + 15(3.41)^2}{13 + 15} = 8.5$$

- Statistical Model for difference in ave. male vs. ave. female height example (see M&M p 663)

Males $y = \mu_{MALE} + \varepsilon$ Females: $y = \mu_{FEMALE} + \varepsilon$

$\varepsilon \sim N(0,\sigma)$

All: $y = \mu_{MALE} + (\mu_{FEMALE} - \mu_{MALE})$ If Female $+ \varepsilon$; $\varepsilon \sim N(0,\sigma)$

Writing $\Delta = \mu_{FEMALE} - \mu_{MALE}$

$y =$	μ_{MALE}	$+ \varepsilon$	$= \mu_{MALE} + \Delta \cdot 0$	$+ \varepsilon$	Males
$y =$	μ_{MALE}	$+ \varepsilon$	$= \mu_{MALE} + \Delta \cdot 0$	$+ \varepsilon$	
...					
$y =$	μ_{MALE}	$+ \varepsilon$	$= \mu_{MALE} + \Delta \cdot 0$	$+ \varepsilon$	
$y =$	μ_{MALE}	$+ \varepsilon$	$= \mu_{MALE} + \Delta \cdot 0$	$+ \varepsilon$	
$y =$	$\mu_{MALE} + \Delta$	$+ \varepsilon$	$= \mu_{MALE} + \Delta \cdot 1$	$+ \varepsilon$	Females
$y =$	$\mu_{MALE} + \Delta$	$+ \varepsilon$	$= \mu_{MALE} + \Delta \cdot 1$	$+ \varepsilon$	
...					
$y =$	$\mu_{MALE} + \Delta$	$+ \varepsilon$	$= \mu_{MALE} + \Delta \cdot 1$	$+ \varepsilon$	
$y =$	$\mu_{MALE} + \Delta$	$+ \varepsilon$	$= \mu_{MALE} + \Delta \cdot 1$	$+ \varepsilon$	
$y =$			$= \mu_{MALE} + \Delta \cdot$ **I**	$+ \varepsilon$	**ALL**

I = "Indicator" of Female = 0 if Male; = 1 if Female

Or, in more conventional Greek letters (β's rather than μ and Δ) used in regression:

$$y = \beta_0 + \beta_1 \cdot \textbf{Indicator\_of\_Female} + \varepsilon$$

$$y = \beta_0 + \beta_1 \cdot \text{"}\mathbf{x}\text{"} + \varepsilon$$

- Fitting (i.e. calculating the parameter estimates of) the model

By calculator $\hat{\beta}_1 = b_1 = \text{"slope"} = r_{xy} \dfrac{SD(y)}{SD(x)}$

$\hat{\beta}_0 = b_0 = \text{"intercept"} = \bar{y} - b_1 \bar{x}$

$\hat{\sigma}^2 = \text{MSE} = \text{average squared residual} = \dfrac{\Sigma(y - \hat{y})^2}{n - 2}$.

- Fitting (i.e. calculating the parameter estimates of) the model

By SYSTAT computer package

```
MODEL HEIGHT = CONSTANT + I_FEMALE
ESTIMATE
```

OUTPUT (I've put the parameter estimates in *italics*)

N:30 MULTIPLE R:0.006 SQUARED MULTIPLE R: 0.000
ADJU. SQUARED MULTIPLE R: 0.00 STANDARD ERROR OF ESTIMATE: *2.92*

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOL.	T	P(2 TAIL)
CONSTANT	*57.21*	0.78	0.0000	.	.73E+02	0.0000
I_FEMALE	*0.04*	1.07	0.0063	1.00	0.03338	0.9736

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	0.00952	1	0.0095	0.00111	0.9736
RESIDUAL	239.35714	28	*8.5485* *		

Translation of OUTPUT ("matching up" parameter estimates)

$\hat{\beta}_1$ (COEFFICIENT for **I_FEMALE**) = *0.04*

$\hat{\beta}_0$ (COEFFICIENT for **CONSTANT**) = *57.21*

$\hat{\sigma}^2$ (MEAN-SQUARE RESIDUAL) = *8.5485*

$\hat{\sigma}$ (ROOT MEAN-SQUARE RESIDUAL) = $\sqrt{8.5485}$ = *2.92*

Remember what the Greek letters stood for in our statistical model:

$\hat{\beta}_0 = \hat{\mu}_{MALE} = 57.21$

$\hat{\beta}_1 = \hat{\Delta} = \hat{\mu}_{FEMALE} - \hat{\mu}_{MALE} = 0.04$

So $\hat{\mu}_{FEMALE} = 57.21 + 0.04 = 57.25$

* Residuals are calculated by squaring the deviation of each y from the estimated (fitted) mean for persons with the same "X" value—in this case those of the same sex, summing them to give 239.357, and dividing the sum by 28 to get 8.5485. This is the same procedure used in Ch 7 to calculate a pooled variance! (If I do the pooled variance calculations without rounding, I get the same 8.5485.

So the regression model 'recovers' the original means and pooled variance!

- What of the other items output by the regression program?

  - STD ERROR($\hat{\beta}_1$) = 1.070 is the SE of $\hat{\Delta} = \hat{\mu}_{FEMALE} - \hat{\mu}_{MALE}$

    In Chapter 7, we would have calculated it by the formula

    SE of difference in Means = $SE(\bar{y}_{FEMALE} - \bar{y}_{MALE})$

    $= \sqrt{s^2[1/n_1 + 1/n_2]} = s\sqrt{1/n_1 + 1/n_2}$

    if use pooled variances.

    You can check that pooled variance = 8.5485 so that

    s = sqrt[8.5485] =2.92

    [it is no accident that the regression gives the same values, since the residuals are the variation of the individuals from the mean in their own gender group... exactly the same procedure as is used in 'pooling' the variances for the t-test]

    Thus $SE(\hat{\mu}_{FEMALE} - \hat{\mu}_{MALE}) = 2.92\sqrt{1/14 + 1/16} = 1.07$

  - T = `0.03338` in the **I_FEMALE** row is the test statistic corresponding to the test of whether the underlying parameter

    $\beta_1 = \Delta = \mu_{FEMALE} - \mu_{MALE}$ in our case

    is ZERO.

    It is formed by taking the ratio of the parameter estimate to its SE, namely

    $$t = \frac{\hat{\beta}_1}{\text{STANDARD ERROR}[\hat{\beta}_1]} = \frac{0.04}{1.08} = 0.0355$$

  - `P(2 TAIL)` = `0.9736` is the P-value obtained by calculating the probability that an observation from the "central" or "null" t distribution with 14+16–2 = 28 df would exceed 0.0335 in absolute value.

100(1 - α)% Confidence Interval for β_1 (and other β's)

Use $\hat{\beta}_1$ and SE[$\hat{\beta}_1$], together with 100(α/2)% and 100(1–α/2)% percentiles of the t_{28} distribution, to form 100(1-α)% Confidence Interval (CI) for β_1. Most modern software packages print CI's—or will, if user requests them! (see examples under Resources for Chapter 10)

• Other items output by the regression program ... continued

• The Analysis of Variance (ANOVA) TABLE Since we are not usually interested in the overall mean μ of the two genders, but rather in their difference -- represented by the parameter $\beta_1 = \Delta = \mu_{FEMALE} - \mu_{MALE}$, the regression program uses the overall mean as a starting point, and calculates the overall variation of the 30 observations from the mean of the 30 observations; If you had taken a calculator and calculated the variance of the 30 observations, you would have to calculate

$$s^2 = \frac{\Sigma(y - \bar{y})^2}{30 - 1} = \frac{239.36666}{29} = \frac{SS_{total}}{29}$$

It then partitions the SStotal, based on 29 df, into

```
REGRESSION SS =   0.00952 based on  1 df
+
RESIDUAL   SS = 239.35714 based on 28 df
----------------------------------------
TOTAL      SS = 239.36666 based on 29 df
```

The regression has 1 term x whose coefficient represents the height variation across gender and that's why the df for the regression is 1.

As explained above, the Error Sum of Squares is calculated as

$$\text{Error Sum of Squares} = \frac{\Sigma(y - \hat{y})^2}{30 - 2}$$

where $\hat{y} = \bar{y}_{MALE} = 57.21$ in the case of males and $\hat{y} = \bar{y}_{FEMALE} = 57.25$ in the case of females. So in effect

$$\text{Mean Square Error} = \frac{\Sigma(y - \bar{y}_{MALE})^2 + \Sigma(y - \bar{y}_{FEMALE})^2}{\{14 - 1\} + \{16 - 1\}}$$

• Fitting the above regressions using PROC GLM in SAS:

see full analysis in separate document under Resources in Chapter 10.

```
PROC REG DATA=sasuser.alberta;
     where (age =11);
 MODEL height = I_Female;
run;
```

```
Dependent Variable: HEIGHT

                    Analysis of Variance

                     Sum of         Mean
Source       DF     Squares       Square    F Value     Prob>F

Model         1     0.00952      0.00952      0.001     0.9736
Error        28   239.35714      8.54847
C Total      29   239.36667

Root MSE      2.92     R-square    0.0000
Dep Mean     57.23     Adj R-sq   -0.0357
C.V.          5.10

                    Parameter Estimates

                Parameter     Standard    T for H0:
Variable   DF    Estimate        Error   Parameter=0    Prob > |T|

INTERCEP    1    57.21        0.781412      73.219        0.0001
I_FEMALE    1     0.04        1.069992       0.033        0.9736
```

Note the identical P Values from pooled variances t-test of the difference in two means, and the test of whether the regression parameter which represents this difference is zero.

I can't get SAS to directly print CI to accompany each estimate, but could use parameter estimate $\pm Z_{\alpha/2}$ × Standard Error[parameter estimate] for 100(1-α)% CI in this instance, since df for t are large enough (28) that t can be approximated by Z.

• **Height vs Age**: Fitting the regression using PROC REG in SAS

```
PROC REG DATA=sasuser.alberta;
          WHERE (I_Female = 0 and age < 13);
 MODEL height = age ; RUN;
```

Dependent Variable: HEIGHT

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	31.34498	31.34498	3.893	0.0547
Error	45	362.35714	8.05238		
C Total	46	393.70213			

Root MSE	2.83767	R-square	0.0796
Dep Mean	58.46809	Adj R-sq	0.0592
C.V.	4.85337		

Parameter Estimates

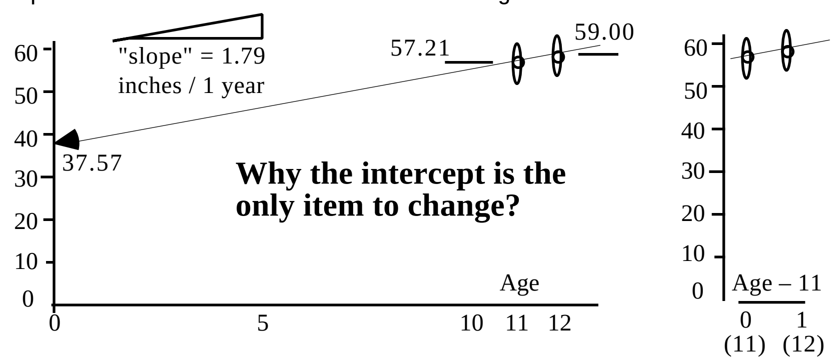
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	37.57	10.59	3.545	0.0009
AGE	1	1.78	0.90	1.973	0.0547

```
DATA from11;
 SET sasuser.alberta; /* create a new 'age' variable   */
 after11 = age - 11;  /* age 11 --> 0                  */
PROC REG DATA = from11; WHERE(I_Female = 0 and age < 13);
 MODEL height = after11 ; RUN;
```

• Output from SAS program   Same as above, except...

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	57.21	0.75	75.441	0.0001
AFTER11	1	1.78	0.90	1.973	0.0547

Note the much smaller SE for the INTERCEPT -- which now has interpretation:- the estimated mean at age 11.



• **Halflife of Caffeine in Smokers and Non Smokers** :

via 2 different SAS PROCedures: **GLM** {General Linear Model) and **REG**

GLM often used when some variables are categorical. with several (k > 2) levels, and user too lazy to create k-1 indicator variables. With k=2 categories, any 2 numerical values will suffice, as long as user remembers how far apart the 2 values are!

```
data a;  infile 'halflife.dat';
  input halflife smoking; smoking=1 if smoker, 0 if not.
  PROC GLM ; model halflife  = smoking; run;
```

• Output from SAS PROC GLM program

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	37.92	37.92	12.65	0.0016
Error	24	71.92	2.99		
Corrected Total	25	109.84			

R-Square	C.V.	Root MSE	HALFLIFE Mean
0.34	36.47 (%)	1.73	4.74

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate*
INTERCEPT	5.95	12.40	0.0001	0.48
SMOKING	-2.41	-3.56	0.0016	0.67

```
data a; infile 'halflife.dat';
input halflife smoking;
proc REG ;
  model halflife  = smoking; run;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	37.92	37.92	12.654	0.001
Error	24	71.92	2.99		
C Total	25	109.84			

Root MSE *1.73*   R-square 0.3452   Dep Mean 4.74615

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	*5.95*	0.48	12.401	0.0001
SMOKING	1	*-2.41*	0.67	-3.557	0.0016

* **Std Error of Estimate** Cf REG. SE's for parameter. estimate.
Do not confuse with same term, used by some, for the RMSE
Even within SAS Institute, different teams use different terms!

DO NOT USE AS MANY DECIMAL PLACES AS SAS REPORTS !!!
In most packages, can specify # of decimals; if not, TRUNCATE!