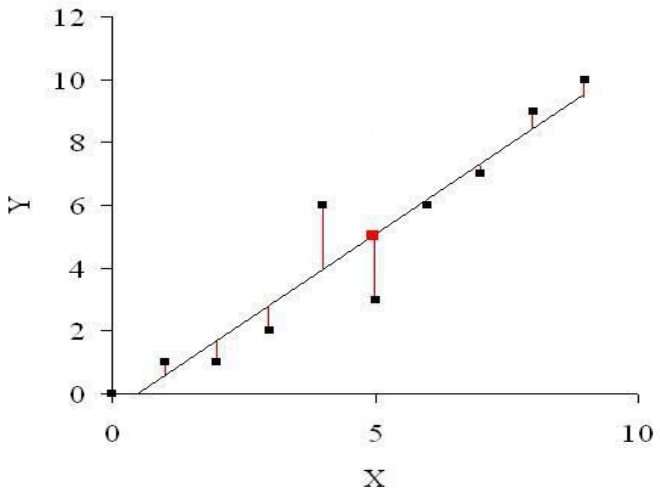
Simple Linear Regression

Gather data points (X, Y) where x is considered a predictor or explanatory variable, and Y is the response or dependent variable. The goal is to find a "best fit" line that satisfies the equation $Y = B_o + B_1X + e$ and is used to predict other Y values.

Methods of Best Fit

Several methods exist to estimate the line. For our purposes we will consider the Least Squares method where the objective is to minimize the sum of squares between the line and the data points. For instance, look at the following graph of X, Y data points:



The sum of squares equals $(Y_1 - \hat{Y}_1)^2 + (Y_2 - \hat{Y}_2)^2 + \ldots = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - [b_o + b_1 X_i])^2$. The line is represented by $\hat{Y}_i = b_o + b_1 X_i$ where b_o and b_1 are the quantities which minimize the previous equation. Through calculus, this minimization turns out to be:

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \frac{S_{xy}}{S_{xx}}$$ where $S_{xy} = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}$ and S_{xx} is the variation of X.

$b_o = \overline{Y} - b_1\overline{X}$

Notice that the last equation can be written as $\overline{Y} = b_o + b_1\overline{X}$ meaning that the point $(\overline{X}, \overline{Y})$ lies on the regression line.

Least Squares Estimation of B_o and B_1

The simple, linear regression model is given by:

$Y_i = B_0 + B_1X_i + e_i \qquad i = 1\ldots N$

where Y_i = value of response variable for the ith person
B_0, B_1 = population parameters intercept and slope, respectively
X_i = value of fixed variable X for the ith person
E_i = random error term with mean = 0

We want to calculated values from a sample which will estimate B_o and B_1 in the model, such that the sum of the squared residuals, or errors of prediction, is minimized.

$$\text{Let } S = \sum E_i^2 = \sum \left(Y_i - \hat{Y}_i\right)^2 = \sum \left(Y_i - \left(B_o + B_1 X_i\right)\right)^2 \qquad (1)$$

Then the estimates b_o and b_1 are called the **least squares** estimates of B_o and B_1. To find these estimates:

Step 1: Find the partial derivative of (1) with respect to B_0 and the partial derivative of (1) with respect to B_1.

First, expand the right side of (1) and distribute the summation sign:

$$\sum \left(Y_i - \left(B_o + B_1 X_i\right)\right)^2 = \sum \left(Y_i^2 - 2Y_i\left(B_o + B_1 X_i\right) + \left(B_o + B_1 X_i\right)^2\right)$$
$$\sum \left(Y_i^2 - 2B_o Y_i - 2B_1 X_i Y_i + B_o^2 + 2B_o B_1 X_i + B_i^2 X_i^2\right)$$
$$\Sigma Y_i^2 - 2B_o \Sigma Y_i - 2B_1 \Sigma X_i Y_i + NB_o^2 + 2B_o B_1 \Sigma X_i + B_i^2 \Sigma X_i^2$$

From calculus the partial derivatives are:

$$\frac{\partial S}{\partial B_o} = 0 - 2\Sigma Y_i - 0 + 2NB_o + 2B_1 \Sigma X_i + 0 = -2\Sigma Y_i + 2NB_o + 2B_1 \Sigma X_i$$
$$\frac{\partial S}{\partial B_1} = 0 - 0 - 2\Sigma X_i Y_i + 0 + 2B_o \Sigma X_i + 2B_1 \Sigma X_1^2 = -2\Sigma X_i Y_i + 2B_o \Sigma X_i + 2B_1 \Sigma X_1^2$$

Step 2: Rearrange terms, set the two partial derivatives equal to 0

$$2NB_o + 2B_1 \Sigma X_i - 2\Sigma Y_i = 0$$
$$2B_o \Sigma X_i + 2B_1 \Sigma X_1^2 - 2\Sigma X_i Y_i = 0$$

Since we are now going to solve for the values of our sample estimates b_o and b_1 and n, replace B_o, B_1 and N in the two above simultaneous equations, and dividing by 2 yields the two **normal equations.**

$$nb_o + b_1 \Sigma X_i - \Sigma Y_i = 0$$
$$b_o \Sigma X_i + b_1 \Sigma X_1^2 - \Sigma X_i Y_i = 0$$

NOTE: $\sum Y_i = nb_o + b_1\sum X_i = \sum b_o + \sum b_1X_i = \sum(b_o + b_1X_i) = \Sigma\hat{Y}_i$ so $\sum Y_i = \Sigma\hat{Y}_i$ or in words, the sum of the observed Y values equals the sum of the fitted values which is one of the properties of a fitted linear regression line.

Step 3: Solve the simultaneous normal equations to give the estimates that will minimize S.

First multiply both sides of the first equation by $\sum X_i$ and the second equation by n.

$$nb_o\Sigma X_i + b_1(\Sigma X_i)^2 - \Sigma X_i\Sigma Y_i = 0$$
$$nb_o\Sigma X_i + nb_1\Sigma X_1^2 - n\Sigma X_iY_i = 0$$

Subtract the first equation from the second yielding,

$$nb_1\Sigma X_i^2 - n\Sigma X_iY_i - b_1(\Sigma X_i)^2 + \Sigma X_i\Sigma Y_i = 0$$

Factor b_1 from the two terms involving it,

$$b_1(n\Sigma X_i^2 - (\Sigma X_i)^2) - n\Sigma X_iY_i + \Sigma X_i\Sigma Y_i = 0$$

Solve for b_1,

$$b_1 = \frac{n\Sigma X_iY_i - \Sigma X_i\Sigma Y_i}{(n\Sigma X_i^2 - (\Sigma X_i)^2)}$$

By dividing both numerator and denominator by n, this can be expressed as:

$$b_1 = \frac{\Sigma X_iY_i - \dfrac{\Sigma X_i\Sigma Y_i}{n}}{(\Sigma X_i^2 - \dfrac{(\Sigma X_i)^2}{n})} \text{ which is equivalent to } \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma(X_i - \overline{X})^2}$$

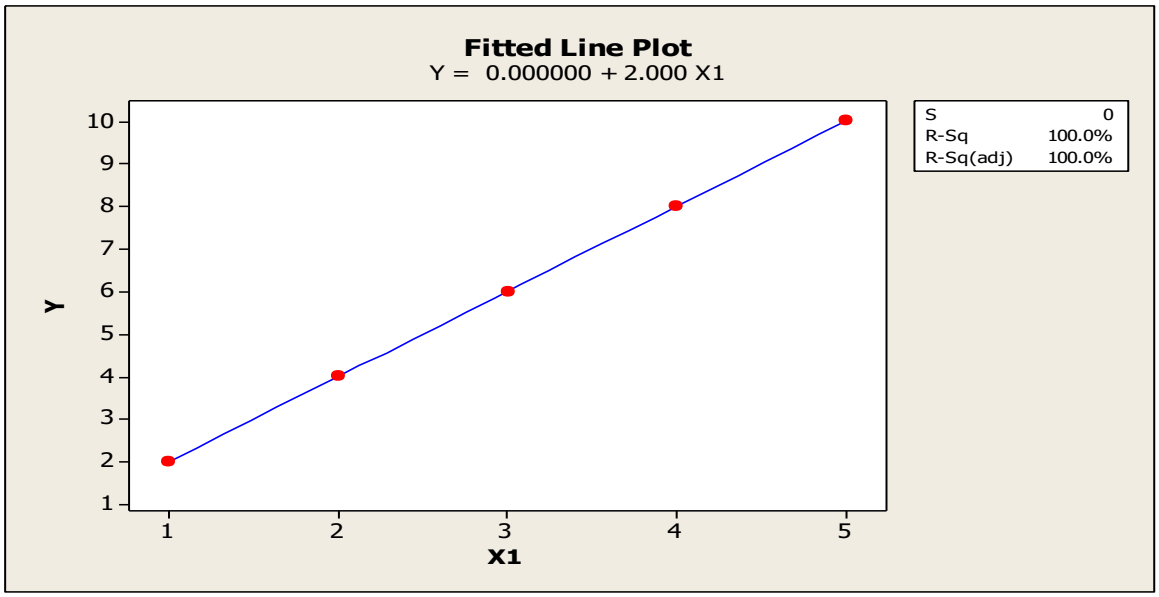
Either one, then, can be used to estimate the slope b_1

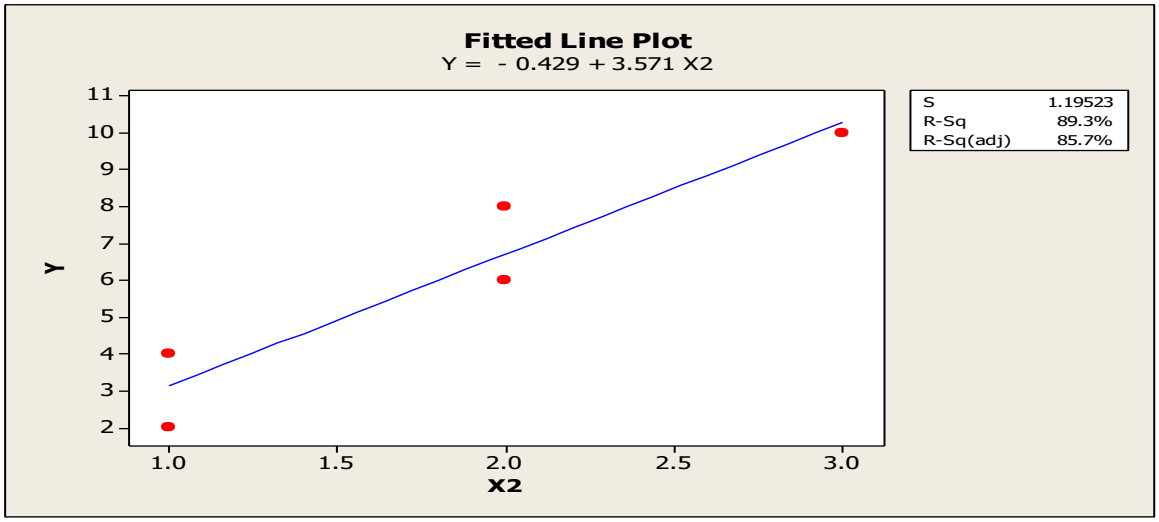
Having found b_1 either of the normal equations found in Step 2 can be used to find b_o. For example using the first one,

$$nb_o + b_1\Sigma X_i - \Sigma Y_i = 0 \text{ leads to } b_o = \frac{\Sigma Y_i}{n} - b_1\frac{\Sigma X_i}{n}$$

Thus $b_o = \overline{Y} - b_1\overline{X}$

This result illustrates another property of the fitted regression line: the line passes through the point $(\overline{Y}, \overline{X})$





Y	X1	X2	YX1	YX2	RESI1	FITS1	RESI2	FITS2
2	1	1	2	2	0	2	-1.14286	3.1429
4	2	1	8	4	0	4	0.85714	3.1429
6	3	2	18	12	0	6	-0.71429	6.7143
8	4	2	32	16	0	8	1.28571	6.7143
10	5	3	50	30	0	10	-0.28571	10.2857

Variable	N	Mean	StDev	Variance	Sum	Sum Squares
Y	5	6.00	3.16	10.00	30.00	220.00
X1	5	3.000	1.581	2.500	15.000	55.000
X2	5	1.800	0.837	0.700	9.000	19.000

ΣYX1 = 110 ΣYX2 = 64 Sxx(x1) = (n-1)*varX = 4*2.5 = 10; Sxx(x1)= 4*0.7 = 2.8

For Y regressed on X1: b1 = (110 – 5*6*3)/10 = 20/10 = 2; bo = 6 – 2*3 = 0
For Y regressed on X2: b1 = (64 – 5*6*1.8)/2.8 = 10/2.8 = 3.571; bo = 6 – 3.571*1.8 = –0.428

Important Properties of Fitted Regression line

1. The sum of the residuals, e_i, equals zero
2. The sum of the squared residuals is a minimum – needed to satisfy requirement of LS regression
3. The sum of the observed Y_i equals the sum of the fitted values, $\hat{Y}_i$.
4. The regression line always goes through the point $(\overline{X}, \overline{Y})$

Distribution Assumptions in Linear Regression

1. The error terms are independent random variables that follow a normal distribution with a mean of 0 and a constant variance σ^2 (homoscedasticity)
2. There exists a linear relationship between X and Y.

$$\sigma\{e_i, e_j\} = \begin{Bmatrix} \sigma^2 & \text{if i = j} \\ \sigma & \text{if i } \neq \text{ j} \end{Bmatrix} and\ e_i \overset{iid}{\sim} N(0, \sigma^2)$$

Recall:

$$E(b_o) = \beta_o \quad \sigma^2\{b_o\} = \frac{\sigma^2 \sum x_i^2}{nS_{xx}}$$

$$E(b_1) = \beta_1 \quad \sigma^2\{b_1\} = \frac{\sigma^2}{S_{xx}}$$

$$\hat{\sigma}^2 = MSE = \frac{\sum e_i^2}{n-2} \text{ or } \frac{\sum y_i^2 - b_o \sum y_i - b_1 \sum x_i y_i}{n-2}$$

Concerning B_1:

Consider the Toluca Company data from Chapter 1: Work Hours is dependent variable and Lot Size is the independent, or explanatory, variable. We want to answer the question, "Does Lot Size (X) contribute to the prediction of Work Hours (Y)?" If the answer is yes, then E(Y) = $B_o + B_1$X. To answer, we test: H_o: $B_1 = 0$ vs. H_1: $B_1 \neq 0$

Remember that since B_1 is a linear combination of Y and Y is normally distributed, the B_1 is distributed $\sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$. By standardizing: $\frac{b_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$ but we do not know σ^2.

Therefore, $\frac{b_1 - \beta_1}{\sqrt{S^2(b_1)}} \overset{H_o}{\sim} t_{n-2}$ where $\frac{MSE}{S_{xx}} = S^2\{b_1\}$.

Proof that $\frac{b_1 - \beta_1}{S\{b_1\}} \overset{H_o}{\sim} t_{n-2}$:

$\frac{b_1-\beta_1}{S\{b_1\}} = \frac{\frac{b_1-\beta}{\sigma\{b_1\}}}{\frac{S\{b_1\}}{\sigma\{b_1\}}}$ where numerator is ~ N(0,1). The denominator is:

$\frac{S\{b_1\}}{\sigma\{b_1\}} = \frac{S^2\{b_1\}}{\sigma^2\{b_1\}} = \frac{\frac{MSE}{S_{xx}}}{\frac{\sigma^2}{S_{xx}}} = \frac{MSE}{\sigma^2} = \frac{\frac{\sum e_i^2}{n-2}}{\sigma^2} = \frac{1}{n-2}\chi^2_{n-2}$. This produces a ratio consisting of a $\frac{Z}{\sqrt{\frac{\chi^2_{n-2}}{n-2}}}$ which, recall from our first lecture, results in a random variable ~t_{n-2}

The Toluca Example, **Example 1 on page 47:**

Testing H_o: $B_1 = 0$ vs. H_1: $B_1 \neq 0$ at $\alpha = 0.05$ we get –
$t = \frac{b_1-0}{S\{b_1\}} = \frac{3.57-0}{0.347} = 10.29$ with p-value = 0.000. **[NOTE: from Table B.2 using DF = 23, the p-value would be p < 2*(1 – 0.9995) = p < 0.001]** Since $p < \alpha$, we reject H_o and conclude that there is a linear association between lot size and work hours.

For a 1-sided test, i.e. H_1: $B_1 > 0$ with $\alpha = 0.05$, we would simply divide the p-value in half, p-value ≈ 0.00, and still reject H_o

Another way to test the 2-sided hypothesis for SLR is to use the F-test since the square of the t-statistic for SLR = $F_{1,n-2}$. From our example here, if we square 10.29 we get 105.88which is same as the F-statistic in the ANOVA table for our example. {NOTE: they may not be quite exact due to rounding error in calculating the t-statistic.}

<u>Confidence Intervals</u>

$$P\left(t^{n-2}_{\alpha/2} \le \frac{b_1-\beta_1}{S\{b_1\}} \le t^{n-2}_{1-\alpha/2}\right) = 1-\alpha$$

$\frac{b_1-\beta_1}{S\{b_1\}} \sim t_{n-2}$ $P\left(b_1 - S\{b_1\}t^{n-2}_{1-\alpha/2} \le \beta_1 \le b_1 + S\{b_1\}t^{n-2}_{1-\alpha/2}\right) = 1-\alpha$

$\therefore$ a (1- α)100% confidence interval for $\beta_1 = b_1 \pm S\{b_1\}t^{n-2}_{1-\alpha/2}$

From our example, calculating a 95% confidence interval would be 3.57 ± (0.347)(2.069) = $2.85 \le B_1 \le 4.29$.

As a test of $B_1 = 0$, since the interval does not contain 0, we can reject Ho at the 5% level.

Concerning B_o – Example Page 49:

$t = \frac{b_o - 0}{S\{b_o\}} = \frac{62.37 - 0}{26.18} = 2.38$ with p-value = 0.026 **[NOTE: from Table B.2 using DF = 23, the p-value would be 2*(1 – 0.99) < p 2*(1 – 0.985) = 0.02 < p < 0.03]**.

SPECIAL NOTE: Keep in mind that a test of $B_o = 0$ may NOT be relevant if the value of X = 0 has no relevant meaning. In this example, considering Lot Size = 0 may not be noteworthy.

Interval Estimation of $E(Y_h)$

$E(Y_h) = B_o + B_1X_h$ where X_h = level of X for which we want to estimate mean response $\hat{Y}_h = b_o + b_1 X_h$ and is unbiased since b_o and b_1 are unbiased and ~N and is also a point estimate of Y_h

$\sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{S_{xx}}\right]$ and its variance will be large when $X_h - \overline{X}$ is large and will be smallest when $X_h = \overline{X}$. Also, $\hat{\sigma}^2\{\hat{Y}_h\} = S^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{S_{xx}}\right]$

Confidence Interval of $E(Y_h)$

$\frac{\hat{Y}_h - E(\hat{Y}_h)}{S\{\hat{Y}_h\}} \sim t^{n-2}$ so a (1- α)*100% confidence interval for $E(Y_h)$ is: $\hat{Y}_h \pm t^{n-2}_{1-\alpha/2} S\{\hat{Y}_h\}$

Example 1 Page 55:

90% C.I. for the mean Work Hours when Lot Size = 100 units.
$X_h = 100$ n = 25 df= 25 – 2 = 23
$\hat{Y}_h = 62.4 + 0.357X_h = 62.4 + 0.357(100) = 419.4$
$t^{23}_{0.95} = 1.714$ MSE = 2384 $\overline{X}$ =70 S_{xx} = 19,800

$S^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{S_{xx}}\right] = (2384)\left[\frac{1}{25} + \frac{(100 - 70.0)^2}{19,800}\right] = 203.72$

$\mathbf{S\{\hat{Y}_h\}} = \sqrt{\mathbf{S^2\{\hat{Y}_h\}}} = \sqrt{203.72} = 14.27$

∴ 90% CI = 419.4 ± (1.714)(14.27) = 394.9 ≤ $E(Y_h)$ ≤ 443.9

Prediction Interval for $Y_{h(new)}$ – Parameters Unknown: Example Page 59

$Y_{h(new)} - \hat{Y}_h$ = prediction error

$$\sigma^2\{Y_{h(new)} - \hat{Y}_h\} = \sigma^2\{Y_{h(new)}\} + \sigma^2\{\hat{Y}_h\} = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{S_{xx}}\right)$$

$$= \sigma^2\left(1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{S_{xx}}\right) = MSE\left(1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{S_{xx}}\right) = S^2\{pred\}$$

Thus a (1- α)*100% prediction interval for $Y_{h(new)}$ is: $\hat{Y}_h \pm t^{n-2}_{1-\alpha/2} S\{pred\}$

EX for $Y_{h(new)} = 4$:

$$S^2\{pred\} = MSE\left(1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{S_{xx}}\right) = 2384\left(1 + \frac{1}{25} + \frac{(100 - 70.0)^2}{19,800}\right) = 2587.72$$

Alternately, we could have used: $S^2\{pred\} = MSE + S^2\{\hat{Y}_h\} = 2384 + 203.72 = 2587.72$

$$S\{pred\} = \sqrt{S^2\{pred\}} = \sqrt{2587.72} = 50.87$$

90% PI = $419.4 \pm (1.714)(50.87) = 332.2 \le Y_{h(new)} \le 506.6$

Note how the 90% prediction interval is **wider** than the 90% confidence interval obtained in the previous example.

NOW USE MINITAB!!!

Analysis of Variance (ANOVA)

Sum of Squares Regression (SSR) = $\sum(\hat{Y}_i - \overline{Y})^2$

Sum of Squares Error (SSE) = $\sum(Y_i - \hat{Y})^2$

Sum of Squares Total (SST) = $\sum(Y_i - \overline{Y})^2$

SST = SSR + SSE

SSE measures variation of Y's around fitted regression line (i.e. residuals) and if all points fall on the regression line then SSE = 0. All of the variation in Y is accounted for by X.

SSR measures the variation of the fitted values from the mean of the response values. If the regression line ignores X and equals simply $\overline{Y}$, then SSR = 0 and the line is horizontal. None of the variation in Y is accounted for by X.

So if SSR or SSE is 0, then SST = whichever is not 0. Usually this is not the case and neither SSR nor SSE equals 0. The goal of linear regression is to find a set of variables that maximizes SSR or minimizes SSE, while at the same time developing as simple a model acceptable.

Proof SST = SSR + SSE

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

$$\Sigma(Y_i - \bar{Y})^2 = \Sigma(\hat{Y}_i - \bar{Y})^2 + \Sigma(Y_i - \hat{Y}_i)^2 = \sum[(\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)]^2$$

$$= \sum(\hat{Y}_i - \bar{Y})^2 + 2\sum(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) + \sum(Y_i - \hat{Y}_i)^2, \text{ the middle expression:}$$

$$2\sum(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = 2\sum[\hat{Y}_i(Y_i - \hat{Y}_i) - \bar{Y}(Y_i - \hat{Y}_i)] = 2\sum\hat{Y}e_i - 2\bar{Y}\sum e_i = 0$$

This leaves: $\sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2 = \text{SSR} + \text{SSE}$

SSR = $\sum(\hat{Y}_i - \bar{Y})^2$, where $\hat{Y}_i = b_o + b_1X_i$ $b_o = \bar{Y} - b_1\bar{X}$ so $\hat{Y}_i = \bar{Y} - b_1\bar{X} + b_1X_i$

$$= \sum(\bar{Y} + b_1(X_i - \bar{X}) - \bar{Y})^2 = \sum[b_1(X_i - \bar{X})^2 = b_1^2\sum(X_i - \bar{X})^2$$

$$SSR = b_1^2 S_{xx}$$

SST has n – 1 DF where SST/(n – 1) ~ χ^2_{n-1}

SSE has n – 2 DF where SSE/(n – 2) = $\hat{\sigma}^2$ ~ χ^2_{n-2}

SSR has 1 DF

Mean Square Regression (MSR) = SSR/1
Mean Square Error (MSE) = SSE/(n – 2)

ANOVA TABLE - Table 2.2 on Page 67

Source	SS	DF	MS	Var Ratio (F)
Regression	SSR = $\sum(\hat{Y}_i - \bar{Y})^2$	1	SSR/1	MSR/MSE
Error	SSE = $\sum(Y_i - \hat{Y})^2$	n – 2	SSE/(n – 2)	
Total	SST = $\sum(Y_i - \bar{Y})^2$	n – 1		

If $B_1 \neq 0$, then MSR > MSE, with larger values of F = MSR/MSE supporting H_A that $B_1 \neq 0$ where MSR/MSE ~ $F_{1, n-2}$ and is only used for two-sided tests of B_1, and for 1-sided use t-tests.

Example using Toluca Company Page 71

SSR = $\sum(\hat{Y}_i - \bar{Y})^2 = 252378$

SSE = $\sum(Y_i - \hat{Y})^2 = 54825$

SST = SSR + SSE = 252378 + 54825 = 307203

```
Analysis of Variance
Source           DF      SS      MS        F      P
Regression        1  252378  252378  105.88  0.000
Residual Error   23   54825    2384
Total            24  307203
```

The critical value for $F_{1,23}$ for alpha = 0.05 is 4.28 which we can either interpolate using Table **B.4** or use the relationship that the F distribution with numerator degrees of freedom of 1 is equal to the square of the t-distribution. Using this relationship and $t_{0.975,\,23} = 2.069$ we get $2.069^2 = 4.28$. From table **B.4** using F* = 105.88 we see that for either denominator DF of 20 or 24 that the p-value is < (1 – 0.999) = p < 0.001 which is precisely what we got when testing $B_1 = 0$!

Remember:

F = MSR/MSE or F = $\frac{b_1^2 S_{xx}}{MSE}$

General Linear Test Approach

Full Model (F): $Y_i = B_o + B_1 X_i + e_i$

Reduced Model (R): $Y_i = B_o + e_i$

SSE(F) = $\sum\left(y_i - b_o - b_1 x_i\right)^2 = \sum(Y_i - \hat{Y})^2$ = SSE

SSE(R) = $\sum(Y_i - B_o)^2 = \sum(Y_i - \bar{Y})^2$ = SST

and SSE(R) > SSE(F)

IF SSE(R) – SSE(F) is "small" then reduced model is adequate

Q? How small is "small"?

A - $F^* = \frac{[SSE(R) - SSE(F)]/(df_R - df_F)}{SSE(F)/df_F} \sim F_{df_R - df_F,\, df_F}$ and we reject H_o if the p-value associated with F* < some stated alpha.

Ex: Toluca Company

Test H_o: $B_1 = 0$ versus H_A: $B_1 \neq 0$ at alpha = 0.05 using General Linear Test Approach

SSE(R) = SST = 307203 SSE(F) = SSE = 54825 $df_R = n - 1 = 24$ $df_F = n - 2 = 23$

$F_{stat} = \frac{[307203 - 54825]/(24-23)}{54825/23} = 105.88$. So for a simple linear regression model the General Linear Test approach is simply the ANOVA approach of F = MSR/MSE

Measures of Association between X and Y

SST = SSR + SSE

R^2 = Coefficient of Determination = SSR/SST or 1 – SSE/SST and measures the percent of variability in Y that is explained by using X to predict Y. Therefore, if SSR is large, then X has an influence on Y, but if SSR is small then $\bar{Y}$ is a good estimate. Thus if the line is a good fit, SSE <<<< SSR

$0\% \leq R^2 \leq 100\%$

Correlation, r, is equal to the square root of R^2 (only when there exists one predictor variable in the model, i.e. SLR) and is ± depending on the sign of the slope. That is: sign(r) = sign(b_1) and correlation ranges from $-1 \leq r \leq 1$

Example: Toluca Company Page 75

R^2 = SSR/SST = $\frac{252,378}{307,203} = 0.822$ or 82.2% of the variation in Work Hours is reduced when Lot Size is considered. Consequently, the correlation, *r*, between Work Hours and Lot Size is r = $\sqrt{R^2} = \sqrt{0.822} = 0.907$ and is positive since the slope is positive.

Also, $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ where $S_{xy} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$ -- $r = b_1\sqrt{\frac{S_{xx}}{S_{yy}}} = 3.57\sqrt{\frac{19,800}{307,203}} = 0.906$

Bivariate Normal

In many regression problems, X values are fixed and Y values are random. However both X and Y might be random. One model for this is the Bivariate Normal Distribution. Parameters are: $u_1, u_2, \sigma_1, \sigma_2,$ *and* ρ_{12}, where ρ_{12} is the population correlation coefficient and follow a density function as depicted by equation **2.74** on page 79 of the text. The explanation of these parameters is as follows:

u_1 *and* σ_1: the population mean and standard deviation of variable one (e.g. Y1 or X1)
u_2, *and* σ_2: the population mean and standard deviation of variable two (e.g. Y2 or X2)
ρ_{12}: population correlation coefficient between the two random variables (e.g Y1,Y2 or X1, X2)

NOTE: the variable designations are unimportant, but to conform to the text we will use Y1 and Y2.

If the Y2 values are fixed, the values of Y1 given Y2 are approximately linear and have a conditional mean of:

$$E[Y_1 \mid Y_2] = u_1 + \rho_{12}\frac{\sigma_1}{\sigma_2} = \left(\mu_1 - \rho_{12}\frac{\sigma_1}{\sigma_2}\mu_2\right) + \rho_{12}\frac{\sigma_1}{\sigma_2}Y_2$$

$$\Downarrow \qquad\qquad \Downarrow$$

$$\beta_o \qquad\qquad \beta_1$$

For a sample (X_1, Y_2), (X_2, Y_2), ….., (X_n, Y_n) if we calculate $\bar{X}, \bar{Y}, S_x, S_y, r$ and we use these to estimate $u_1, u_2, \sigma_1, \sigma_2,$ *and* ρ_{12} respectively, then to estimated E[Y|X] is:

$$E[\hat{Y}_i \mid X] = \bar{Y} + r\frac{S_y}{S_x}(X_i - \bar{X}).$$

Example Sales - Advertising

MONTH	ADVERTISING EXPENDITURE *x*, hundreds of dollars	SALES REVENUE *y*, thousands of dollars
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

$\bar{X} = 3,\ \bar{Y} = 2,\ S_y = 1.225,\ S_x = 1.581,\ r = 0.904$

$$E[\hat{Y} \mid X] = \bar{Y} + r\frac{S_y}{S_x}(X - \bar{X}) = 2 + (0.904)\frac{1.225}{1.581}(X - 3) = 2 + .7(X - 3)$$

$$\Downarrow$$

$$\beta_1$$

So to predict E[Y|X] when X is fixed at 4 results produces a predicted result of 2.7 just as before for the regression model when we predicted the mean response when $X_h = 4$.

NOTE: Since this is a correlation model you could just as well be interested in E[X|Y]. In such an instance you just edit the variables in the equations accordingly.

Tests and Confidence Intervals for ρ_{12}

Test of H_o: $\rho_{12} = 0$ is equivalent to H_o: $B_1 = 0$ since $B_1 = \rho_{12}\frac{\sigma_1}{\sigma_2}$. To test, we use:

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \overset{H_o}{\sim} t_{n-2}$$

Test of H_o: $\rho_{12} = \rho_o$ which is NOT equivalent to H_o: $B_1 = 0$, we use Fisher's Z transformation, z'

$$z' = \frac{1}{2}\ln\left(\frac{1+r_{12}}{1-r_{12}}\right) \sim N\left[\left(\frac{1}{2}\ln\frac{1+\rho}{1-\rho}\right), \sqrt{\frac{1}{n-3}}\right] \text{ for } n \geq 25 \text{ and can use Table } \mathbf{B.8}$$

instead of the formula. Therefore, for a large enough N, we can test H_o: $\rho_{12} = \rho_o$ by:

$$Z^* = \frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)}{\sqrt{\frac{1}{n-3}}} \overset{H_o}{\sim} N(0,1)$$ and this test statistic can be compared to the critical value from the standard normal table for $(1 - \alpha/2)$ for a 2-sided test or $(1 - \alpha)$ for a 1- sided test. Or, you can find the p-value of Z* by P(Z* < Z).

The $(1 - \alpha)100\%$ confidence interval for ρ_{12} is found by:

1. $z' \pm z_{(1-\alpha/2)}\sigma\{z'\} \Rightarrow z' \pm \frac{z_{(1-\alpha/2)}}{\sqrt{n-3}}$ which gives the CI for z', then to convert this to a CI for ρ_{12};
2. We go to Table **B.8** and locate the values under z' that are closest to the bounds for z' found in step1 to select the bounds for ρ_{12}, keeping the signs for z' the same

The interpretation of this interval follows the usual interval interpretations.

Example: Data on 30 individuals was taken including Age and Systolic Blood Pressure (SBP). The regression line is $AGE = 98.75 + 0.97SBP$ with $r = 0.66$. The regression line is $AGE = 98.75 + 0.97SBP$ with $r = 0.66$. Using the correlation model to test $B_1 = 0$ we get $t = \frac{0.66\sqrt{28}}{\sqrt{1-0.66^2}} = 4.62$ From MTB, $2P(t_{28} > 4.62) = 0.00008$, so reject Ho that $B_1 = 0$. To test H_o: $\rho_{12} = 0.85$ vs H_a: $\rho_{12} < 0.85$ we compute the Fisher Z by:

$$Z^* = \frac{\frac{1}{2}\ln\left(\frac{1+0.66}{1-0.66}\right) - \frac{1}{2}\ln\left(\frac{1+0.85}{1-0.85}\right)}{\sqrt{\frac{1}{30-3}}} = -2.41$$ and P(Z* < -2.41) = 0.008 so reject Ho.

A 99% confidence interval for $\rho_{12} = z' \pm \frac{z_{(1-\alpha/2)}}{\sqrt{n-3}}$ where $z' = \frac{1}{2}\ln\left(\frac{1+r_{12}}{1-r_{12}}\right)$ begins with z'= $\frac{1}{2}\ln\left(\frac{1+0.66}{1-0.66}\right) = 0.793$ >>> $0.793 \pm \frac{2.575}{\sqrt{27}} = 0.297 \leq z' \leq 1.289$ and converting z' to a

confidence interval for ρ_{12} use Table **B.8** to find the limits for ρ_{12}. From the table we get the values under z' closest to these limits which are 0.2986 and 1.2933 which gives the 99% CI for ρ_{12}: $0.29 \leq \rho_{12} \leq 0.86$