Data modeling: CSCI E-106

Applied Linear Statistical Models

Chapter 2

Summary

The linear regression function with one predictor variable:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, n$$

- Y_i: the value of response variable in the ith trial
- β_0, β_1: parameters
- X_i: a known constant; the value of the predictor variable in the ith trial
- ε_i: random error term; $E\{\varepsilon_i\} = 0$; $\sigma^2(\varepsilon_i) = \sigma^2$; uncorrelated ($\sigma\{\varepsilon i, \varepsilon j\} = 0$, i≠j)

How to obtain the estimators b_0, b_1 ?

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \quad \Rightarrow \quad b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) = \bar{Y} - b_1 \bar{X}$$

1. The sum of the residuals is zero:

$$\sum_{i=1}^{n} e_i = 0 \qquad (1.17)$$

Table 1.2, column 4, illustrates this property for the Toluca Company example. Rounding errors may, of course, be present in any particular case, resulting in a sum of the residuals that does not equal zero exactly.

2. The sum of the squared residuals, $\sum e_i^2$, is a minimum. This was the requirement to be satisfied in deriving the least squares estimators of the regression parameters since the criterion Q in (1.8) to be minimized equals $\sum e_i^2$ when the least squares estimators b_0 and b_1 are used for estimating β_0 and β_1.

3. The sum of the observed values Y_i equals the sum of the fitted values $\hat{Y}_i$:

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i \qquad (1.18)$$

This property is illustrated in Table 1.2, columns 2 and 3, for the Toluca Company example. It follows that the mean of the fitted values $\hat{Y}_i$ is the same as the mean of the observed values Y_i, namely, $\bar{Y}$.

4. The sum of the weighted residuals is zero when the residual in the ith trial is weighted by the level of the predictor variable in the ith trial:

$$\sum_{i=1}^{n} X_i e_i = 0 \qquad (1.19)$$

5. A consequence of properties (1.17) and (1.19) is that the sum of the weighted residuals is zero when the residual in the ith trial is weighted by the fitted value of the response variable for the ith trial:

$$\sum_{i=1}^{n} \hat{Y}_i e_i = 0 \qquad (1.20)$$

6. The regression line always goes through the point $(\bar{X}, \bar{Y})$.

Chapter 2: Inferences in Regression and Correlation Analysis

Inferences concerning:

- the regression parameters β_0 and β_1
- interval estimation of β_0 and β_1 tests about them
- interval estimation of E (Y) of the probability distribution of Y for given X
- prediction intervals of a new observation Y
- confidence bands for the regression line

Normal Error Regression Model

Assume that the normal error regression model is applicable:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

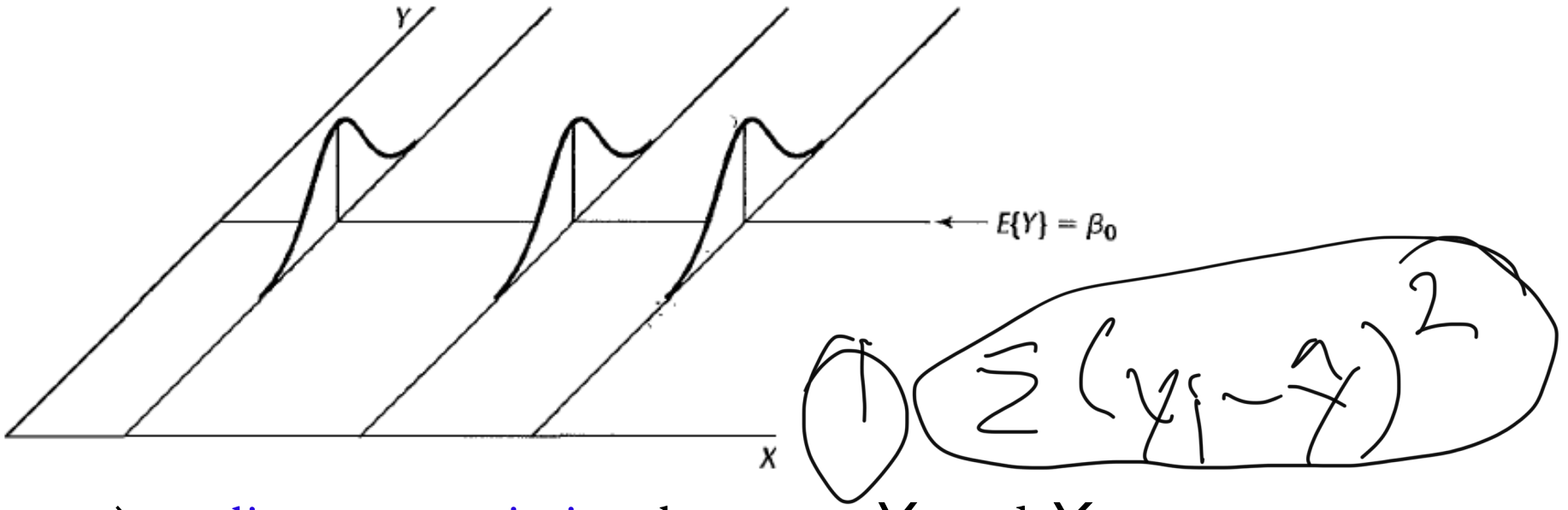
- β_0 and β_1 are parameters
- X_i are known constants
- $\varepsilon_i \sim N(0, \sigma^2)$: are independent
- Y_i: independently, normally distributed
- A linear combination of independent normal random variables is normally distributed

Inferences about the slope β_1 of the regression line

- Market research analyst studying the relation between sales *(Y)* and advertising expenditures *(X)* may wish to obtain an interval estimate of β_1, because it will provide information as to how many additional sales dollars, on the average, are generated by an additional dollar of advertising expenditure.
- At times, tests concerning β_1 are of interest, particularly one of the form:

$$H_0 : \beta_1 = 0;$$
$$H_a : \beta_1 \neq 0.$$

Inferences about the slope β_1 of the regression line, cont'd



- $\beta_1 = 0 \Rightarrow$ no linear association between Y and X
- The regression line is horizontal. The means of Y: $E\{Y\} = \beta_0$.
- The probability distribution of Y are identical at all levels of X

Sampling Distribution of b_1

- The sample distribution of b_1 refers to the different values of b_1 that would be obtained with repeated sampling when the levels of the predictor variable X are held constant from sample to sample.

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

- For normal error regression model, the sampling distribution of b_1 is normal, with mean and variance

$$E\{b_1\} = \beta_1$$

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$

Sampling Distribution of b_1, *cont'd*

- b_1 as Linear Combination of the Y_i • It can be shown that b_1 can be expressed as follows:

$$b_1 = \sum k_i Y_i$$

where:

$$k_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2}$$

The coefficients k_i have a number of interesting properties that will be used later:

$$\sum k_i = 0 \qquad (2.5)$$

$$\sum k_i X_i = 1 \qquad (2.6)$$

$$\sum k_i^2 = \frac{1}{\sum (X_i - \bar{X})^2} \qquad (2.7)$$

Sampling Distribution of b_1 , *cont'd*

- Mean. The unbiasedness of the point estimator b_1 stated earlier in the Gauss-Markov theorem (1.11), is easy to show:

$$E\{b_1\} = E\left\{\sum k_i Y_i\right\} = \sum k_i E\{Y_i\} = \sum k_i(\beta_0 + \beta_1 X_i)$$
$$= \beta_0 \sum k_i + \beta_1 \sum k_i X_i$$

By the equations 2.5 and 2.6 in the previous slide, we then obtain $E\{b_1\} = \beta_1$.

- Variance. The variance of b_1 can be derived readily. We need to remember that the Y_i are independent random variables, each with variance σ^2 , and that the k_i are constants. Hence, we obtain:

$$\sigma^2\{b_1\} = \sigma^2\left\{\sum k_i Y_i\right\} = \sum k_i^2 \sigma^2\{Y_i\}$$
$$= \sum k_i^2 \sigma^2 = \sigma^2 \sum k_i^2$$
$$= \sigma^2 \frac{1}{\sum (X_i - \bar{X})^2}$$

Sampling Distribution of b_1 , *cont'd*

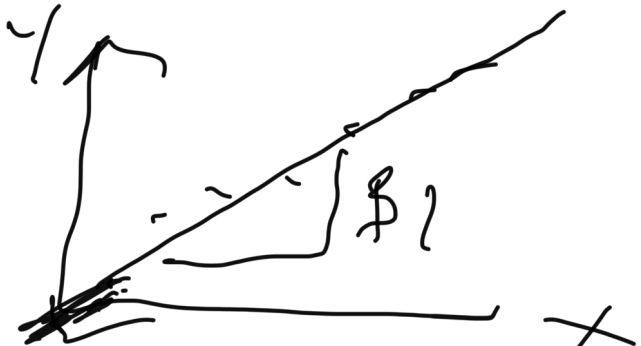
- Estimated Variance. We can estimate the variance of the sampling distribution of b_1 *:*

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$

- by replacing the parameter σ^2 with *MSE,* the unbiased estimator of σ^2*:*

$$s^2\{b_1\} = \frac{MSE}{\sum(X_i - \bar{X})^2}$$

- The point estimator $s^2\{b_1\}$ is an unbiased estimator of $\sigma^2\{b_1\}$. Taking the positive square root, we obtain $s\{b_1\}$, the point estimator of $\sigma\{b_1\}$.



Theorem 1:

- The estimator b_1 has minimum variance among all unbiased linear estimators of:

$$\widehat{\beta_1} = \sum c_i Y_i$$

Where c_i are arbitrary constants which holds $\sum c_i = 0$; $\sum c_i X_i = 1$

We now prove this. Since b_1 is required to be unbiased, the following must hold:

$$E\{\hat{\beta}_1\}=E\{\sum c_i Y_i\}=\sum c_i E(Yi)=\beta_1 \text{ then}$$
$$\sum c_i E(Yi) = \sum c_i E(\beta_0 + \beta_1 X_i)= \beta_0 \sum c_i+\beta_1 \sum c_i X_i$$

By using conditions above $\sum c_i = 0$; $\sum c_i X_i = 1$, it follows that

$$E\{\hat{\beta}_1\}= \beta_1$$

$$\sigma^2\{\hat{\beta}_1\} = \sum c_i^2 \sigma^2\{Y_i\} = \sigma^2 \sum c_i^2$$

Theorem 1, cont'd

- We now prove that b_1 has minimum variance among all unbiased linear estimators of:

$$\widehat{\beta_1} = \sum c_i Y_i$$

under the conditions that $\sum c_i = 0$; $\sum c_i X_i = 1$

$$\sigma^2\{\hat{\beta}_1\} = \sigma^2\{\sum c_i Y_i\} = \sum c_i^2 \sigma^2(Yi)$$
$$= \sum c_i^2{}_i \sigma^2(Yi) = \sum c_i^2 \sigma^2 = \sigma^2 \sum c_i^2$$

Let us define $c_i = d_i + k_i$, where k_i are the least squares constants (remember from slide 8 that $\sigma^2\{b_1\} = \sigma^2 \sum k_i^2$), now we can write that

$$\sigma^2\{\hat{\beta}_1\} = \sigma^2 \sum c_i^2 = \sigma^2 \sum (di + ki)2 = \sigma^2(\sum d_i^2 + \sum k_i^2 + \sum d_i\, k_i) = \sigma^2 \sum d_i^2 + \sigma^2 \sum k_i^2 + \sum d_i\, k_i$$
$$= \sigma^2 \sum d_i^2 + \sigma^2\{b_1\} + \sum d_i\, k_i$$

Theorem 1, cont'd

- We will show that $\sum d_i\, k_i = 0$

$$\begin{aligned}\sum k_i d_i &= \sum k_i (c_i - k_i) \\ &= \sum c_i k_i - \sum k_i^2 \\ &= \sum c_i \left[\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2}\right] - \frac{1}{\sum (X_i - \bar{X})^2} \\ &= \frac{\sum c_i X_i - \bar{X} \sum c_i}{\sum (X_i - \bar{X})^2} - \frac{1}{\sum (X_i - \bar{X})^2} = 0\end{aligned}$$

- $\sigma^2\{\hat{\beta}_1\} = \sigma^2 \sum d_i^2 + \sigma^2\{b_1\} + \sum d_i\, k_i = \sigma^2 \sum d_i^2 + \sigma^2\{b_1\}$

- Note that the smallest values of $\sum d_i^2$ is 0, it occurs if all d_i=0, which implies that $c_i = k_i$. Thus, this proves the theorem.

Sampling Distribution of $(b_1 - \beta_1)/s\{b_1\}$

- Since b_1, is normally distributed, we know that the standardized statistic $\frac{b_1-\beta_1}{\sigma(b_1)}$ is a standard normal variable.
- $\sigma(b_1)$ is estimated by $s(b_1)$.

$\frac{b_1-\beta_1}{s(b_1)}$ is distributed as t(n-2), t distribution with n-2 degree of freedom.

Proof:

Reminder:

Z_i are independent identical distributed (iid) with ~N(0,1):

- $\sum_{i=1}^{n} Z_i$ ~ N(0,n)
- $Z_i^2 \sim \chi^2(1)$ and $\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$

The pdf of a Normal distribution N(μ, σ^2)

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

The pdf of a Normal distribution N(0,1)

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

Sampling Distribution of $(b_1 - \beta_1)/s\{b_1\}$, cont'd

Chi-square Distribution

$$f(x|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}, \quad 0 < x < \infty,$$

Where Γ is the gamma function.

$t(v) = \frac{Z}{\sqrt{\frac{\chi^2(v)}{v}}}$ is the t-distribution with v degree of freedom, where Z is the standard normal distribution and $\chi^2(v)$ is the chi square distribution with v degree of freedom.

Sampling Distribution of $(b_1 - \beta_1)/s\{b_1\}$, cont'd

- First we *need to rewrite* $\frac{b_1-\beta_1}{s(b_1)}$
- $\frac{b_1-\beta_1}{s(b_1)} * \frac{\sigma(b_1)}{\sigma(b_1)} = \frac{b_1-\beta_1}{\sigma(b_1)} * \frac{\sigma(b_1)}{s(b_1)} = \frac{b_1-\beta_1}{\sigma(b_1)} \div \frac{s(b_1)}{\sigma(b_1)} = Z \div \frac{s(b_1)}{\sigma(b_1)}$

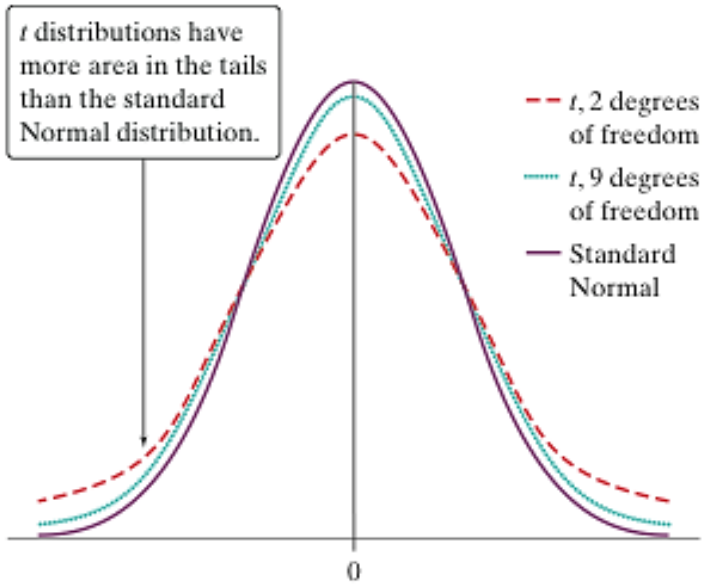
$$\frac{s^2\{b_1\}}{\sigma^2\{b_1\}} = \frac{\frac{MSE}{\sum(X_i-\bar{X})^2}}{\frac{\sigma^2}{\sum(X_i-\bar{X})^2}} = \frac{MSE}{\sigma^2} = \frac{\frac{SSE}{n-2}}{\sigma^2}$$

$$= \frac{SSE}{\sigma^2(n-2)} \sim \frac{\chi^2(n-2)}{n-2}$$

- *Then,* $\frac{b_1-\beta_1}{s\{b_1\}} \sim \frac{z}{\sqrt{\frac{\chi^2(n-2)}{n-2}}}$ *is distributed with t distribution with n-2 degree of freedom*

The t Distributions

When comparing the density curves of the standard Normal distribution and *t* distributions, several facts are apparent:



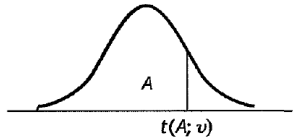
- ✓ The density curves of the *t* distributions are similar in shape to the standard Normal curve.
- ✓ The spread of the *t* distributions is a bit larger than that of the standard Normal distribution.
- ✓ The *t* distributions have more probability in the tails and less in the center than does the standard Normal.
- ✓ As the degrees of freedom increase, the *t* density curve becomes ever closer to the standard Normal curve.

We can use Table B.2 in the Appendix of the book to determine critical values *t** for *t* distributions with different degrees of freedom.

T TABLE

TABLE B.2
Percentiles of the *t* Distribution.

Entry is $t(A;\nu)$ where $P\{t(\nu) \le t(A;\nu)\} = A$



	A						
ν	.60	.70	.80	.85	.90	.95	.975
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.537	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

TABLE B.2
(concluded)
Percentiles of the *t* Distribution.

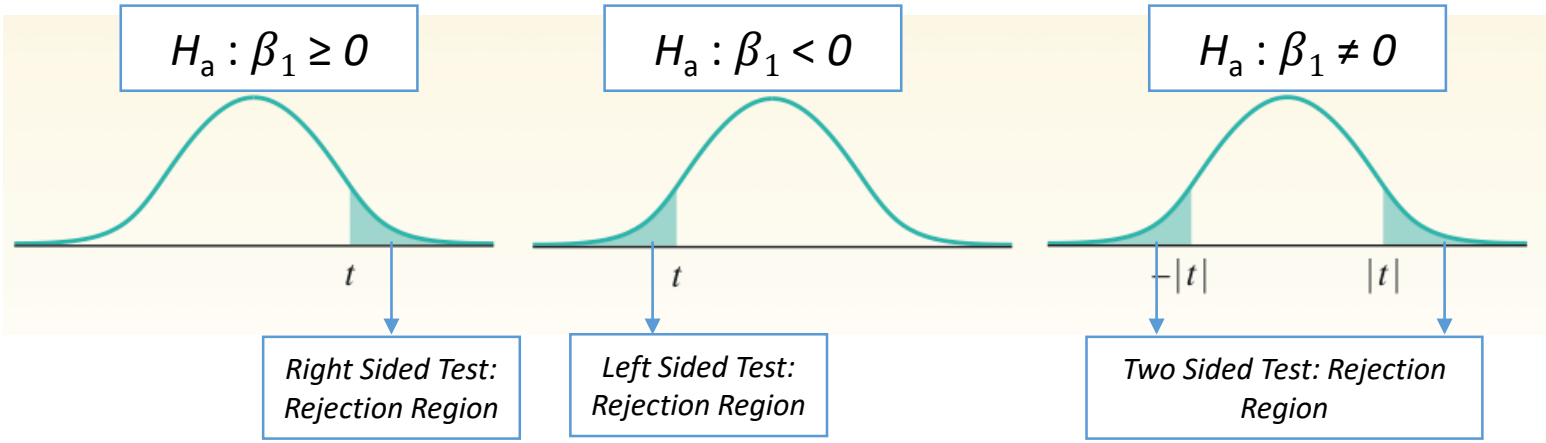
	A						
ν	.98	.985	.99	.9925	.995	.9975	.9995
1	15.895	21.205	31.821	42.434	63.657	127.322	636.590
2	4.849	5.643	6.965	8.073	9.925	14.089	31.598
3	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	2.197	2.336	2.528	2.661	2.845	3.153	3.849
21	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	2.158	2.291	2.473	2.598	2.771	3.057	3.690
28	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	2.150	2.282	2.462	2.586	2.756	3.038	3.659
30	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	2.123	2.250	2.423	2.542	2.704	2.971	3.551
60	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	2.054	2.170	2.326	2.432	2.576	2.807	3.291

The One-Sample *t* Test

To test the hypothesis $H_0 : \beta_1 = 0$, compute the one-sample *t* statistic:

$$t = \frac{b_1}{s(b_1)}$$

Find the *P*-value by calculating the probability (at degree of freedom = $n - 2$) of getting a *t* statistic this large or larger *in the direction specified by the alternative hypothesis* H_a.



Hypothesis testing for β_1

(1-$\alpha/2$)% confidence interval for β_1 is

$$b_1 \pm t(1-\alpha/2;n-2)s(b_1) = [b_1 - t(1-\alpha/2;n-2)s(b_1),\ b_1 + t(1-\alpha/2;n-2)s(b_1)]$$

Example: Consider the Toluca Company example of Chapter 1.

$n = 25$	$\bar{X} = 70.00$
$b_0 = 62.37$	$b_1 = 3.5702$
$\hat{Y} = 62.37 + 3.5702X$	$SSE = 54{,}825$
$\sum(X_i - \bar{X})^2 = 19{,}800$	$MSE = 2{,}384$
$\sum(Y_i - \bar{Y})^2 = 307{,}203$	

The regression equation is
Y = 62.4 + 3.57 X

Predictor	Coef	Stdev	t-ratio	p
Constant	62.37	26.18	2.38	0.026
X	3.5702	0.3470	10.29	0.000

s = 48.82 R-sq = 82.2% R-sq(adj) = 81.4%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	252378	252378	105.88	0.000
Error	23	54825	2384		
Total	24	307203			

Inferences about the slope β_0 of the regression line

- There are only infrequent occasions when we wish to make inferences concerning β_0, the intercept of the regression line. These occur when the scope of the model includes X = 0 times.
- The point estimator b_0 was shown in Chapter as follow:

$$b_0 = \bar{Y} - b_1 \bar{X}$$

- The sampling distribution of b_0 is normal, with mean and variance:

$$E\{b_0\} = \beta_0$$

$$\sigma^2\{b_0\} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right]$$

Inferences about the slope β_0 of the regression line, cont'd

- The normality of the sampling distribution of b_0 follows because b_0 , like b_1 , is a linear combination of the observations Yi
- The results for the mean and variance of the sampling distribution of b_0 can be obtained in similar fashion as those for b_1
- An estimator of a $\sigma^2(b_0)$ is obtained by replacing a σ^2 by its point estimator MSE:

$$s^2\{b_0\} = MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]$$

- The positive square root, $s(b_o)$, is an estimator of $\sigma(b_0)$

Sampling Distribution of $(b_o - \beta_0)/s(b_o)$

- Similar to b_1, the sampling distributions follows a t distribution with n-2 degree of freedom.

$$\frac{b_o - \beta_0}{s(b_0)} \sim t(n-2)$$

- $(1-\alpha/2)\%$ confidence interval for $\boldsymbol{\beta_0}$ is

$$b_0 \pm t(1 - \alpha/2; n - 2)s(b_0) = [b_0 - t(1 - \alpha/2; n - 2)s(b_0) , b_0 + t(1 - \alpha/2; n - 2)s(b_0)]$$

Example: Toluca Example

Long Way

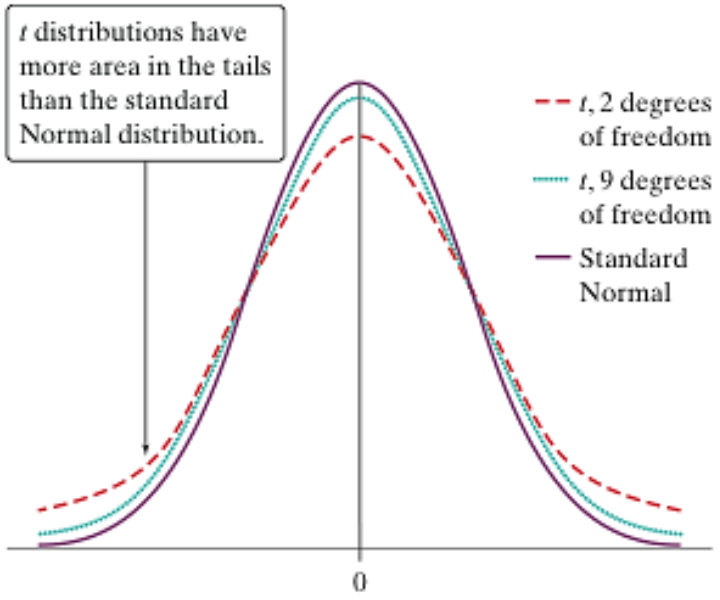
- n<-length(lotsize)
- alpha<-0.05
- tdf<-qt(1-alpha/2,n-2)
- Sxx<-sum((lotsize-mean(lotsize))^2)
- b1<-sum((lotsize-mean(lotsize))*(workhrs-mean(workhrs)))/Sxx
- b0<-mean(workhrs)-b1*mean(lotsize)
- SSE<-sum((workhrs-(b0+b1*lotsize))^2)
- MSE<-SSE/(n-2)
- sb1<-sqrt(MSE/sum((lotsize-mean(lotsize))^2)) #s{b1}
- conf_b1<-c(b1-tdf*sb1,b1+tdf*sb1)
- conf_b1
- sb0<-sqrt(MSE*(1/n+mean(lotsize)^2/sum((lotsize-mean(lotsize))^2)))
- conf_bo<-c(b0-tdf*sb0,b0+tdf*sb0)
- conf_bo

Using build in Function

- toluca.reg<-lm(workhrs~ lotsize)
- summary(toluca.reg)
- confint(toluca.reg)

Departures from Normality

- Even if the distribution of Y are far from normal, the estimators b_0 and b_1 generally have the property of asymptotic normality-their distributions approach normality under very general conditions as the sample size increases. Thus, with sufficiently large samples, the confidence intervals and decision rules given earlier still apply even if the probability distributions of Y depart far from normality. For large samples, the t value is, of course, replaced by the z value for the standard normal distribution.



Power of the test

When we draw a conclusion from a significance test, we hope our conclusion will be correct. But sometimes it will be wrong. There are two types of mistakes we can make.

If we reject H_0 when H_0 is true, we have committed a **Type I error.**

If we fail to reject H_0 when H_0 is false, we have committed a **Type II error.**

		Truth about the population	
		H_0 true	H_0 false (H_a true)
Conclusion based on sample	Reject H_0	**Type I error**	*Correct conclusion*
	Fail to reject H_0	*Correct conclusion*	**Type II error**

Power of the test, cont'd

Type I Error

The probability of a Type I error is the probability of rejecting H_0 when it is really true. This is exactly the significance level of the test.

The significance level α of any fixed-level test is the probability of a Type I error. That is, α is the probability that the test will reject the null hypothesis H_0 when H_0 is in fact true. Consider the consequences of a Type I error before choosing a significance level.

Type II Error and Power

A significance test makes a Type II error when it fails to reject a null hypothesis that really is false. There are many values of the parameter satisfying the alternative hypothesis. We can calculate the probability that a test *does* reject H_0 when any *specific* alternative is true. This probability is called the **power** of the test against that specific alternative.

The power of a test against any alternative is 1 minus the probability of a Type II error for that alternative; that is, power = $1 - \beta$.

The One-Sample *t* Test

To test the hypothesis

$$H_0 : \beta_1 = \beta_{10}$$

$$H_0 : \beta_1 \neq \beta_{10}$$

The test statistics becomes

$t = \frac{b_1 - \beta_{10}}{s(b_1)}$ ~ non-central t-distribution with $\delta = \left|\frac{b_1 - \beta_{10}}{\sigma(b_1)}\right|$ and n-2 degree of freedom (df)

Where δ is the non-central measure i.e., a measure of how far the true value of b_1 is from β_{10}

Table B.5 (on page 1327) presents the power of the two-sided *t* test for α = .05 and α = .01, for various df

Power Values

TABLE B.5
Power Values for Two-Sided *t* Test.

	$\alpha = .05$								
	δ								
df	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0
1	.07	.13	.19	.25	.31	.36	.42	.47	.52
2	.10	.22	.39	.56	.72	.84	.91	.96	.98
3	.11	.29	.53	.75	.90	.97	.99	1.00	1.00
4	.12	.34	.62	.84	.95	.99	1.00	1.00	1.00
5	.13	.37	.67	.89	.98	1.00	1.00	1.00	1.00
6	.14	.39	.71	.91	.98	1.00	1.00	1.00	1.00
7	.14	.41	.73	.93	.99	1.00	1.00	1.00	1.00
8	.14	.42	.75	.94	.99	1.00	1.00	1.00	1.00
9	.15	.43	.76	.94	.99	1.00	1.00	1.00	1.00
10	.15	.44	.77	.95	.99	1.00	1.00	1.00	1.00
11	.15	.45	.78	.95	.99	1.00	1.00	1.00	1.00
12	.15	.45	.79	.96	1.00	1.00	1.00	1.00	1.00
13	.15	.46	.79	.96	1.00	1.00	1.00	1.00	1.00
14	.15	.46	.80	.96	1.00	1.00	1.00	1.00	1.00
15	.16	.46	.80	.96	1.00	1.00	1.00	1.00	1.00
16	.16	.47	.80	.96	1.00	1.00	1.00	1.00	1.00
17	.16	.47	.81	.96	1.00	1.00	1.00	1.00	1.00
18	.16	.47	.81	.97	1.00	1.00	1.00	1.00	1.00
19	.16	.48	.81	.97	1.00	1.00	1.00	1.00	1.00
20	.16	.48	.81	.97	1.00	1.00	1.00	1.00	1.00
21	.16	.48	.82	.97	1.00	1.00	1.00	1.00	1.00
22	.16	.48	.82	.97	1.00	1.00	1.00	1.00	1.00
23	.16	.48	.82	.97	1.00	1.00	1.00	1.00	1.00
24	.16	.48	.82	.97	1.00	1.00	1.00	1.00	1.00
25	.16	.49	.82	.97	1.00	1.00	1.00	1.00	1.00
26	.16	.49	.82	.97	1.00	1.00	1.00	1.00	1.00
27	.16	.49	.82	.97	1.00	1.00	1.00	1.00	1.00
28	.16	.49	.83	.97	1.00	1.00	1.00	1.00	1.00
29	.16	.49	.83	.97	1.00	1.00	1.00	1.00	1.00
30	.16	.49	.83	.97	1.00	1.00	1.00	1.00	1.00
40	.16	.50	.83	.97	1.00	1.00	1.00	1.00	1.00
50	.17	.50	.84	.98	1.00	1.00	1.00	1.00	1.00
60	.17	.50	.84	.98	1.00	1.00	1.00	1.00	1.00
100	.17	.51	.84	.98	1.00	1.00	1.00	1.00	1.00
120	.17	.51	.85	.98	1.00	1.00	1.00	1.00	1.00
∞	.17	.52	.85	.98	1.00	1.00	1.00	1.00	1.00

Example –Toluca Company

Suppose we want to calculate the power of the test for β_1=1.5

$H_0 : \beta_1 = \beta_{10} = 0$

$H_0 : \beta_1 \neq \beta_{10} \neq 0$

From the previous slides, $\sigma^2(b_1)=\frac{\sigma^2}{\sum(X_i-\bar{X})^2}=\frac{2,500}{19,800}$=0.1263 and $\sigma(b_1)$=0.3553

Then, $\delta = |\frac{b_1-\beta_{10}}{\sigma(b_1)}|=\frac{|1.5-0|}{0.3553}$=4.22

From the table by using α = .05 and 23 df and interpolate linearly between

$\delta = 4$ and δ =5, we obtain

$$.97 + \frac{4.22 - 4.00}{5.00 - 4.00}(1.00 - .97) = .9766$$

- Thus, if β_1 = 1.5, the probability would be about .98 that we would be led to conclude *Ha* ($\beta_1 \neq 0$). In other words, if β_1 = i .5, we would be almost certain to conclude that there is a linear relation between work hours and lot size.

Interval Estimation of $E\{Y_h\}$

- Let X_h denote the level of X for which we wish to estimate the mean response.
- X_h may be a value which occurred in the sample, or it may be some other value of the predictor variable within the scope of the model.
- The mean response when X = X_h is denoted by $E\{Y_h\}$.
- Formula gives us the point estimator Y_h of $E\{Y_h\}$:
  - $Y_h = b_o + b_1 X_h$
- What is the sampling distribution of Y_h?

Sampling Distribution of $\hat{Y}_h$

- For normal error regression model, the sampling distribution of $\hat{Y}_h$ is normal, with mean and variance:

$$E\{\hat{Y}_h\} = E\{Y_h\}$$

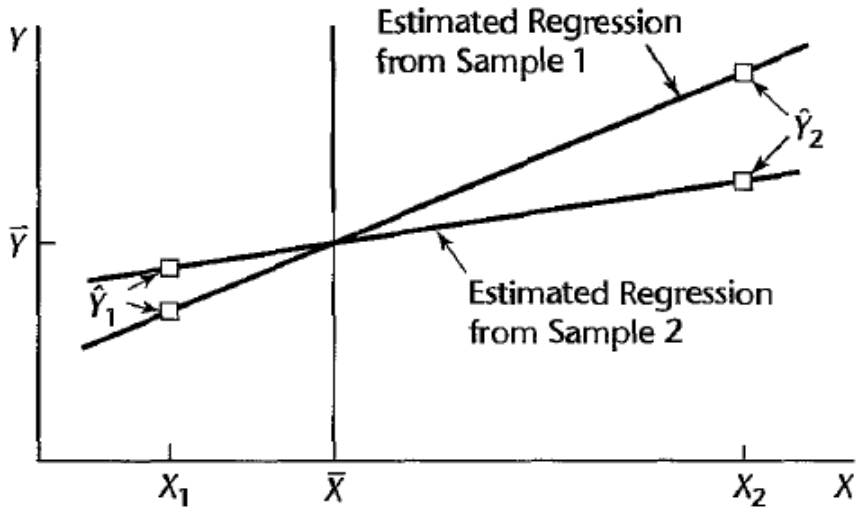
$$\sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

- Normality: The normality of the sampling distribution of Y_h follows directly from the fact that $\hat{Y}_h$, like b_o and b_1 , is a linear combination of the observations Y_i

- Mean. Note from that Y_h is an unbiased estimator of $E\{Y_h\}$. To prove this, we proceed as follows:

$$E\{\hat{Y}_h\} = E\{b_0 + b_1 X_h\} = E\{b_0\} + X_h E\{b_1\} = \beta_0 + \beta_1 X_h$$

Sampling Distribution of $\hat{Y}h$, cont'd



- The two regression lines are assumed to go through the same $(\bar{X}, \bar{Y})$ point to isolate the effect of interest, namely, the effect of variation in the estimated slope b_1 from sample to sample.
- Note that at X_1 near $\bar{X}$, the fitted values $\hat{Y}_1$ for the two sample regression lines are close to each other.
- At X_2 which is far from $\bar{X}$, the fitted values $\hat{Y}_2$ differ substantially

- Thus, variation in the slope b_1 from sample to sample has a much more pronounced effect on $\hat{Y}_h$ for X levels far from the mean X than for X levels near X.
- Hence, the variation in the $\hat{Y}_h$ values from sample to sample will be greater when X_h is far from the mean than when X_h is near the mean.

Sampling Distribution of $\hat{Y}h$, cont'd

- When *MSE* is substituted for σ^2, we obtain $s^2\{\hat{Y}_h\}$, the estimated variance of $\hat{Y}_h$:

$$s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

- The estimated standard deviation of $\hat{Y}_h$is then $s\{\hat{Y}_h\}$, the positive square root of $s^2\{\hat{Y}_h\}$.

Sampling Distribution of $\hat{Y}h$, cont'd

Derivation of $s^2\{\hat{Y}_h\}$

- $Y_h = b_o + b_1 X_h$ *and* $b_o = \bar{Y} - b_1 \bar{X}$ *and then*
- $Y_h = \bar{Y} - b_1 \bar{X} + b_1 X_h = \bar{Y} + b_1(X_h - \bar{X})$
- $\sigma^2(Y_h) = \sigma^2(\bar{Y} + b_1(X_h - \bar{X})) = \sigma^2(\bar{Y}) + \sigma^2(b_1)(X_h - \bar{X})^2 + 2(Xh\_\bar{X})\,^2\sigma^2(\bar{Y}, b_1)$
- $= \frac{\sigma^2}{n} + (X_h - \bar{X})^2 \frac{\sigma^2}{\sum(X_i - \bar{X})^2} + 0$
- $= \sigma^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)$

Sampling Distribution of $(\hat{Y}_h - E(\hat{Y}_h))/s(\hat{Y}_h)$,

$\frac{\hat{Y}_h - E(\hat{Y}h)}{s(\hat{Y}_h)}$ ~ t distribution with n-2 df

(1-α/2)% Confidence Interval for *E{*$\hat{Y}_h$*}*

$\hat{Y}_h \pm t(1 - \alpha/2;n - 2)s(\hat{Y}_h) = [\hat{Y}_h - t(1 - \alpha/2;n - 2)s(\hat{Y}_h)\ ,\hat{Y}_h + t(1 - \alpha/2;n - 2)s(\hat{Y}_h)]$

Example

- Returning to the Toluca Company example, let us find a 90 percent confidence interval for E $\{Y_h\}$ when the lot size is $X_h = 65$ units.

$$\hat{Y}_h = 62.37 + 3.5702(65) = 294.4$$

$$s^2\{\hat{Y}_h\} = 2{,}384\left[\frac{1}{25} + \frac{(65 - 70.00)^2}{19{,}800}\right] = 98.37$$

$$s\{\hat{Y}_h\} = 9.918$$

- For a 90 percent confidence coefficient, we require t(.95; 23) = 1.714. Hence, our confidence interval with confidence coefficient .90 is

$$294.4 - 1.714(9.918) \leq E\{Y_h\} \leq 294.4 + 1.714(9.918)$$

$$277.4 \leq E\{Y_h\} \leq 311.4$$

- We conclude with confidence coefficient .90 that the mean number of work hours required when lots of 65 units are produced is somewhere between 277.4 and 311.4 hours.

Example

- Let us find a 90 percent confidence interval for E $\{Y_h\}$ when the lot size is X_h = 100 units.

$$\hat{Y}_h = 62.37 + 3.5702(100) = 419.4$$

$$s^2\{\hat{Y}_h\} = 2{,}384\left[\frac{1}{25} + \frac{(100 - 70.00)^2}{19{,}800}\right] = 203.72$$

$$s\{\hat{Y}_h\} = 14.27$$

$$t(.95; 23) = 1.714$$

- our confidence interval with confidence coefficient .90 is

$$419.4 - 1.714(14.27) \leq E\{Y_h\} \leq 419.4 + 1.714(14.27)$$

$$394.9 \leq E\{Y_h\} \leq 443.9$$

- Note that this confidence interval is somewhat wider than that from the previous example, since the X_h level here (X_h = 100) is substantially farther from the mean $\bar{X}$ = 70.0 than the X_h level from the previous example (X_h = 65).

Prediction of New Observation

- The new observation on Y to be predicted is viewed as the result of a new trial, independent of the trials on which the regression analysis is based.
- We denote the level of X for the new trial as X_h and the new observation on Y as $Y_{h(new)}$.
- **It is assumed that the underlying regression model applicable for the basic sample data continues to be appropriate for the new observation !**
- We predict an individual outcome drawn from the distribution of Y.

Example: College Admissions

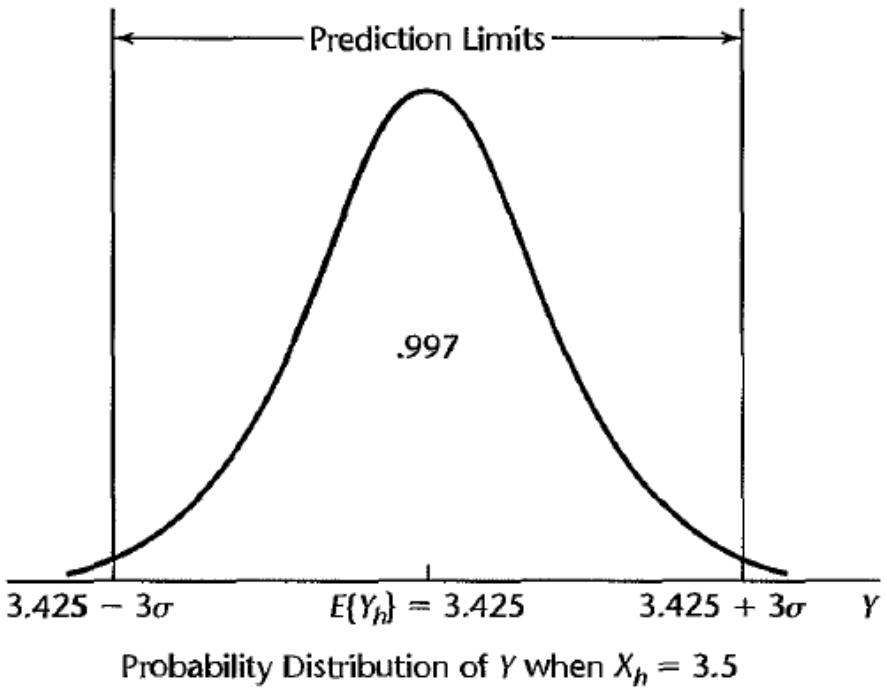
- Suppose that in the college admissions example the relevant parameters of the regression model are known to be

$$\beta_0 = .10 \qquad \beta_1 = .95$$

$$E\{Y\} = .10 + .95X$$

$$\sigma = .12$$

- The admissions officer is considering an applicant whose high school GPA is $X_h = 3.5$.
- The mean college GPA for students whose high school average is 3.5 is:
- $E\{Y_h\} = .10 + .95(3.5) = 3.425$
- $E\{Y_h\} \pm 3\sigma$:
  - $3.425 \pm 3(0.12) \Rightarrow 3.065 \leq Y_{h(new)} \leq 3.785$



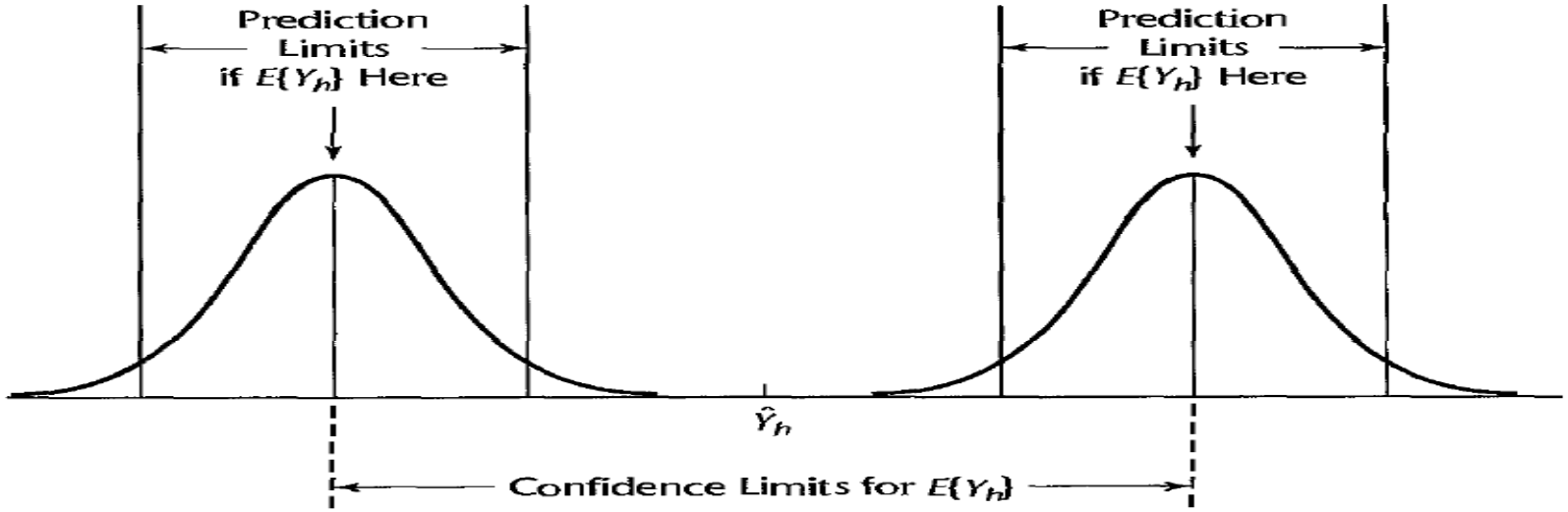
Probability Distribution of Y when $X_h = 3.5$

Prediction of New Observation, cont'd

- The basic idea of a prediction interval is to choose a range in the distribution of Y wherein most of the observations will fall, and then to declare that the next observation will fall in this range.
- When the regression parameters of normal error regression model are known, the 1 – α prediction limits for $Y_{h(new)}$ are:

$$E\{Y_h\} \pm z(1-\alpha/2)\sigma$$

Prediction Interval for $Y_{h(new)}$ when Parameters Unknown



- Figure shows the prediction limits for each of the two probability distributions of *Y* presented there. Since we cannot be certain of the location of the distribution of *Y,* prediction limits for $Y_{h(new)}$ must take account of two elements

    1. Variation in possible location of the distribution of *Y* .
    2. Variation within the probability distribution of *Y.*

Prediction Interval for $Y_{h(new)}$ when Parameters Unknown, cont'd

- Prediction limits for a new observation $Y_{h(new)}$ at a given level *Xh* are obtained by means of the following theorem:

- $\frac{Y_{h(new)} - \hat{Y}_h}{S(pred)}$ ~ *t(n-2) distribution*

- The 1 – α prediction limits for $Y_{h(new)}$:
  - $\hat{Y}h \pm$ t(1 – α/2; n – 2)s{pred}

- The difference may be viewed as the prediction error, with $\hat{Y}_h$ serving as the best point estimate of the value of the new observation $Y_{h(new)}$
- σ^2{pred}: the variance of the prediction error
  - $\sigma^2\{\text{pred}\} = \sigma^2\{Y_{h(new)} - \hat{Y}_h\} = \sigma^2 + \sigma^2\{\hat{Y}_h\}$

Prediction Interval for $Y_{h(new)}$ when Parameters Unknown, cont'd

- $\sigma^2\{pred\}$ has two components:
  - The variance of the distribution of Y at $X = X_h$; σ^2
  - The variance of the sampling distribution of $\hat{Y}_h$; $\sigma^2\{\hat{Y}_h\}$

- An unbiased estimator of $\sigma^2\{pred\}$

$$s^2(pred) = MSE + s^2(\hat{Y}_h) = MSE + MSE\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)$$

$$= MSE\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

Example: Toluca Company

- Toluca Company: $X_h = 100$
- 90 percent prediction interval: $t(0.95; 23) = 1.714$

  - $\hat{Y}_h = 419.4 \qquad s^2\{\hat{Y}_h\} = 203.72 \qquad MSE = 2,384$
  - $\Rightarrow s^2(\text{pred}) = 2,384 + 203.72 = 2,587.72$
  - $s(\text{pred}) = 50.87$

- The 90 percent prediction interval for $Y_{h(new)}$:

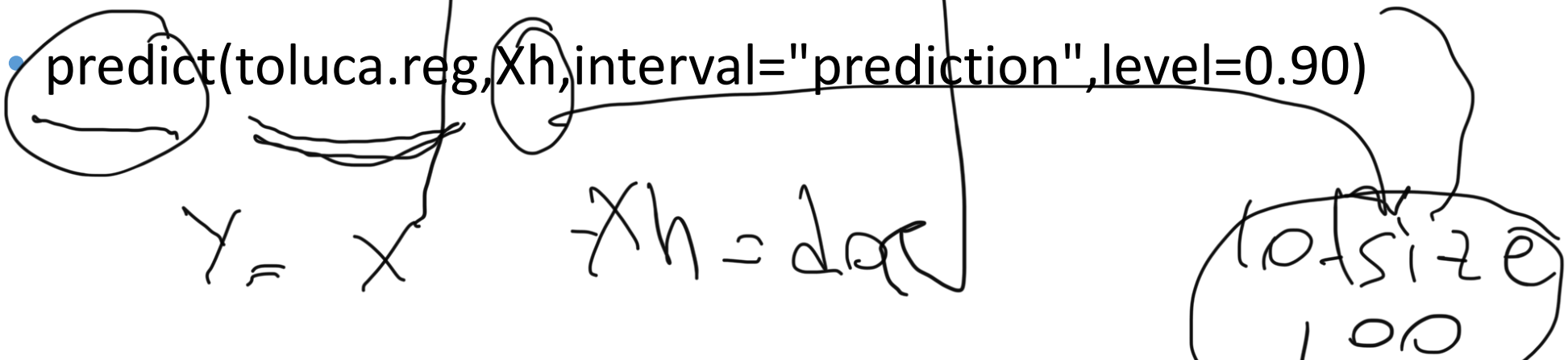
  - $332.2 \leq Y_{h(new)} \leq 506.6$

Example: Toluca Company - Rcode

- toluca.reg <-lm(workhrs ~ lotsize , data= toluca_data)
- Xh<-data.frame(lotsize=100)
- fitnew<-predict(toluca.reg,Xh,se.fit=T, interval="prediction", level=0.9)
- s2pred<-fitnew$se.fit^2+fitnew$residual.scale^2
- c(s2pred,sqrt(s2pred))

OR

- predict(toluca.reg,Xh,interval="prediction",level=0.90)



Example: Toluca Company, cont'd

- This prediction interval is rather wide and may not be useful for planning worker requirements for the next lot.
- The interval can still be useful for control purposes.
  - If the actual work hours fall outside the prediction limits ⇒ some alerts may have occurred a change in the production process

- Toluca Company: The C.I. for $Y_{h(new)}$ is wider than the C.I. for $E\{Y_h\}$:
- predicting the work hours required for a new lot
- encounter the variability in $\hat{Y}_h$ from sample to sample as well as the lot-to-lot variation within the probability distribution of Y
- The prediction interval is wider the further X_h is from $\bar{X}$

Prediction of Mean of m New Observations for Given X_h

- Predict the mean of *m* new observations on Y for a given X_h
- Y: the mean of the new observations to be predicted as $\bar{Y}_{h(new)}$
- The new Y observations are independent
- The appropriate 1 − α prediction limits:
  - $\hat{Y}_h \pm t(1 - \alpha/2; n - 2)s\{predmean\}$

$$s^2(predmean) = \frac{MSE}{m} + s^2(\hat{Y}_h) = MSE\left[\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

1. The variance of the mean of m observations from the probability distribution of Y at x= X_h
2. The variance of the sampling distribution of Y_h

Example: Toluca Example

- Toluca Company: $X_h = 100$
- %90 prediction interval for the mean number of work hours $\bar{Y}_{h(new)}$ in three new production runs
- Previous results:
- $\hat{Y}_h = 419.4$; $s^2\{Y_h\} = 203.72$
- MSE = 2, 384; t(0.95; 23) = 1.714;

- $s^2\{predmean\} = \frac{2384}{3} + 203.72 = 998.4$
- $s\{predmean\} = 31.60$
- $419.4 - 1.714(31.60) \leq \bar{Y}_{h(new)} \leq 419.4 + 1.714(31.60)$
- $365.2 \leq \bar{Y}_{h(new)} \leq 473.6$

- The prediction interval for the total number of work hours for the three lots by multiplying the prediction limits for $\bar{Y}_{h(new)}$ by 3:

- $3(365.2) = 1{,}095.6 \leq$ Total work hours $\leq 3(473.6) = 1{,}420.8$

- Thus, it can be predicted with 90% confidence that between 1,096 and 1,421 work hours will be needed to fill the contract for three lots of 100 units each.

Confidence-Band for Regression Line

- At times we would like to obtain a confidence band for the entire regression line E{Y} =$\beta_0 + \beta_1 X$
- This band enables us to see the region in which the entire regression line lies.
- The Working-Hotelling $1 - \alpha$ confidence band for the regression line for regression model has the following two boundary values at any level X_h :

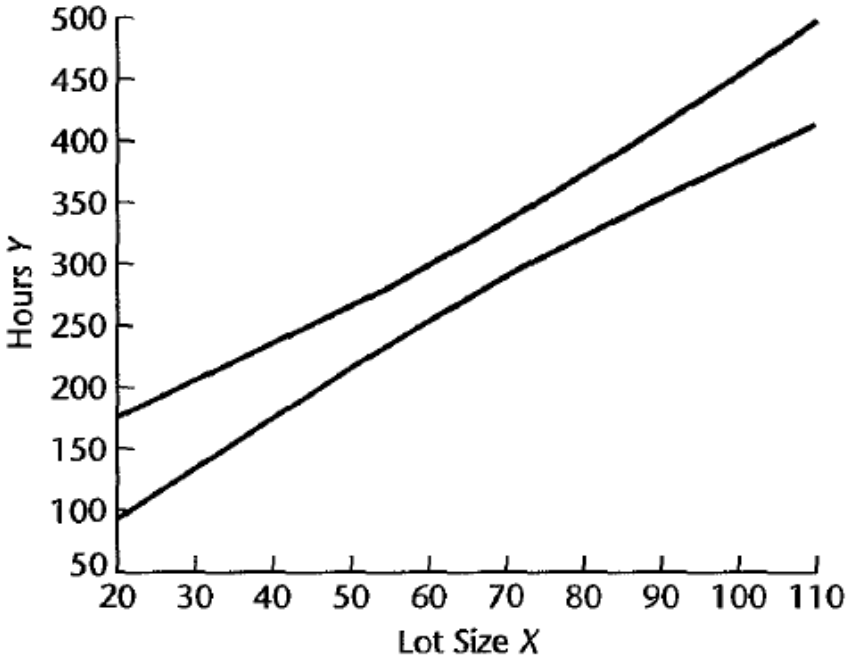
$$\hat{Y}_h \pm W\, s(\hat{Y}_h)$$

- Where $W^2 = 2F(1 - \alpha; 2, n - 2)$ and F is the F distribution with 2 and n-2 degree of freedom.

Example: Toluca Data

- Toluca Company: X_h = 100
- $\hat{Y}_h$ = 419.4; $s^2\{Y_h\}$ = 203.72; MSE = 2, 384;

- $W^2 = 2F(1 - \alpha; 2, n - 2) = 2F(.90; 2, 23) = 2(2.549) = 5.098$
- W = 2.258
- Hence, the boundary values of the confidence band for the regression line at X_h = 100 are
- 419.4 ± 2.258(14.27), and the confidence band is:
- $387.2 \leq \beta_0 + \beta_1 Xh \leq 451.6$ for X_h = 100



Confidence Band for the regression line

Partition of Total Sum of Squares

- The analysis of variance approach is based on the partitioning of sums of squares and degrees of freedom associated with Y .

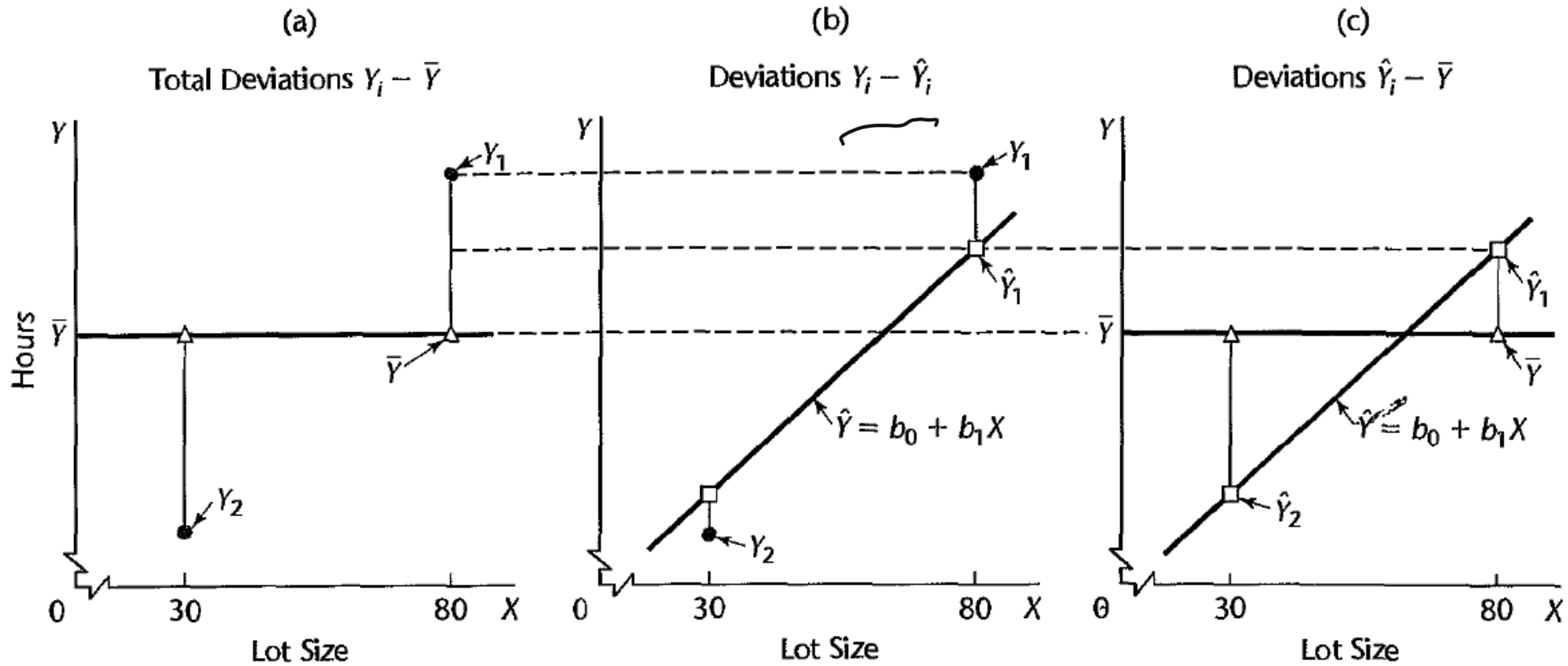
- The variation is measured: the deviations of the Y_i around their mean $\bar{Y}$:

$$Y_i - \bar{Y}$$

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

$$SST = \sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2 + 2\sum e_i \hat{Y}_i - 2\sum e_i \bar{Y}$$

Partition of Total Sum of Squares, cont'd



Partition of Total Sum of Squares, cont'd

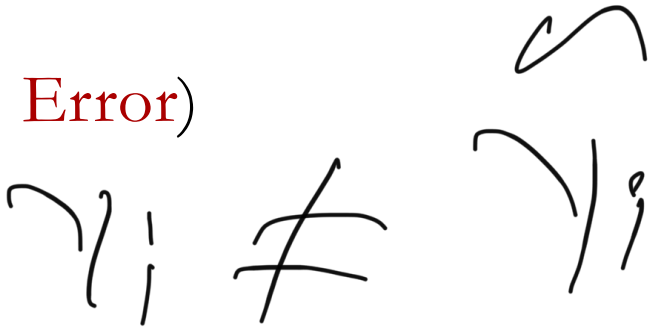
- Total variation: (SSTO or SST: total sum of squares)

$$\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

- Y_i are the same ⇒ SST = 0
- The greater the variation among the Y_i, the larger is SST.

- Total Error: (*SSE*: Sum of Squares of Error)

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

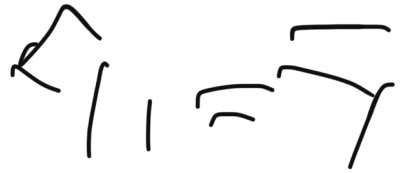


- Y_i fall on the fitted regression line ⇒ SSE = 0
- The greater the variation of the Y_i around the fitted regression line, the larger is SSE .

Partition of Total Sum of Squares, cont'd

- Total Regression: (*SSR:* Regression Sum of Squares)

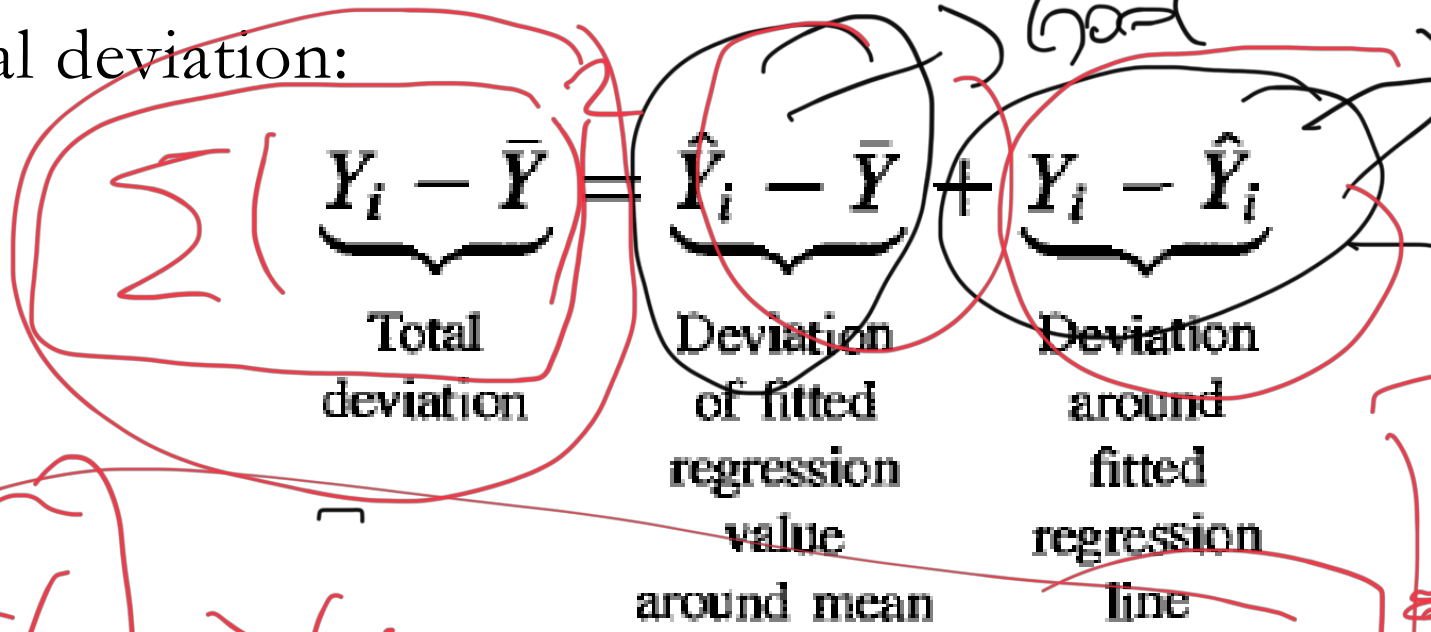
$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$



- The regression line is horizontal ⇒ SSR = 0, otherwise SSR > 0
- a measure associated with the regression line
- The larger SSR is in relation to SST, the greater is the effect of the regression relation in accounting for the total variation in the Yi observations.

Partition of Total Sum of Squares, cont'd

- The total deviation:



- Two components:
  1. The deviation of the fitted value $\hat{Y}_i$ around the mean $\bar{Y}$.
  2. The deviation of the observation Y_i around the fitted regression line.

Partition of Total Sum of Squares, cont'd

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

$$SSTO \qquad SSR \qquad SSE$$

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

$$SSR = b_1^2 \sum (X_i - \bar{X})^2$$

Source of Variation	Degree of Freedom (df)	Explanation
SSR	1	$\beta_0 \, and \, \beta_1$ are estimated, due to constraint we lose 1 df.
SSE	n – 2	$\beta_0 \, and \, \beta_1$ are estimated, as such we lose 2 df
SST	n – 1	$\sum_{i=1}^{n} (Y_i - \bar{Y}) = 0$, as such we lose 1 df

MEAN SQUARES

$$Var = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

- Mean Square (MS) is Sum of Square (SS) divided by the degree of freedom (df)

SS	df	MS
SSR	1	$MSR = \frac{SSR}{1}$ (regression mean square)
SSE	$n-2$	$MSE = \frac{SSE}{n-2}$ (error mean square)

Analysis of Variance (ANOVA) Table

Source of Variation	SS	df	MS	$E\{MS\}$
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$\sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	σ^2
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$		

ANOVA TABLE, cont'd

$$SST = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2$$

- In the modified ANOVA table, the total uncorrected sum of squares, denoted by SSTOU, is defined as:

$$SSTOU = \sum Y_i^2$$

- SS(correction for mean) = $n\bar{Y}^2$

Modified ANOVA Table

Source of Variation	SS	df	MS
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$	
Correction for mean	$SS(\text{correctionformean}) = n\bar{Y}^2$	1	
Total, uncorrected	$SSTOU = \sum Y_i^2$	n	

Expected Mean Squares

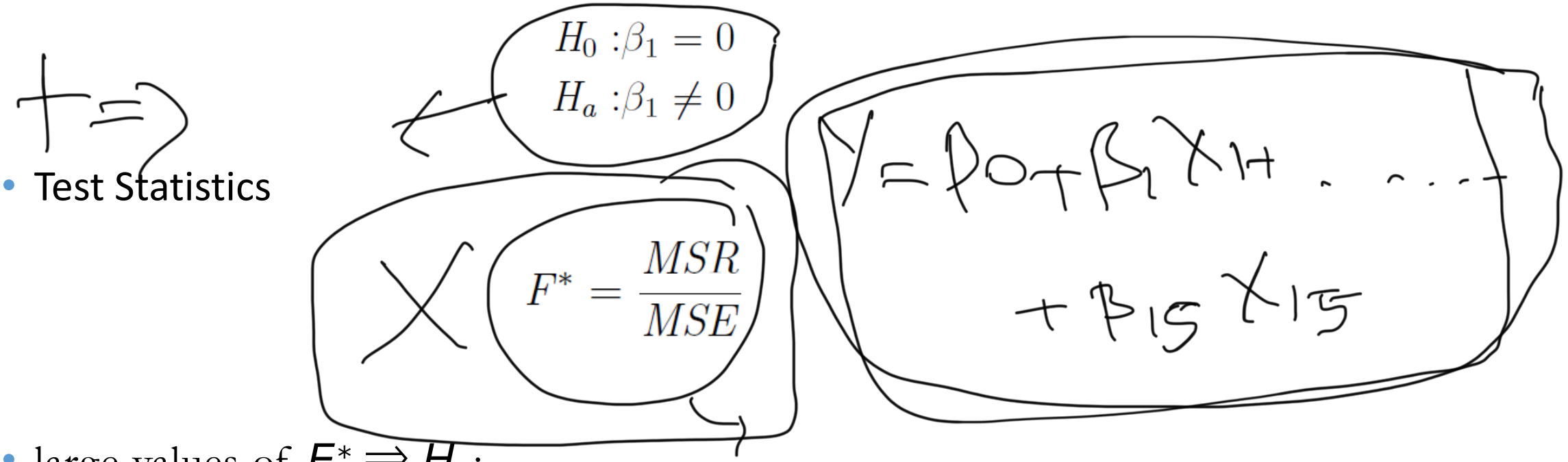
$$E\{MSE\} = \sigma^2$$

$$E\{MSR\} = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$$

- MSE is an unbiased estimator of σ^2. Implication:
  - The mean of the sampling distribution of MSE is σ^2;
  - when $\beta_1 = 0$, the mean of the sampling distribution of MSR is σ^2;
  - when $\beta_1 \neq 0$, $E\{MSR\} > E\{MSE\}$

F test of $\beta_1 = 0$ vs. $\beta_1 \neq 0$

- The analysis of variance approach provides us with a battery highly useful tests for regression models.
- For the simple linear regression case, the ANOVA provides us with a test:



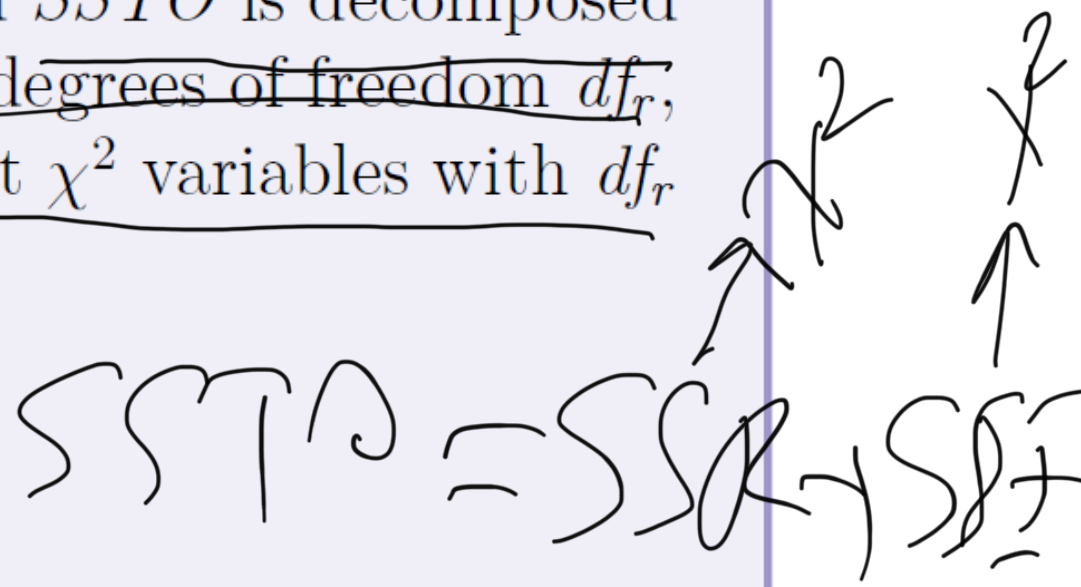
- Test Statistics
- large values of $F^* \Rightarrow H_a$;
- values of F^* near $1 \Rightarrow H_0$;

F test of $\beta_1 = 0$ vs. $\beta_1 \neq 0$, cont'd

Cochran's theorem

If all n observations Y_i come from the same normal distribution with mean μ and variance σ^2, and $SSTO$ is decomposed into k sums of squares SS_r, each with degrees of freedom df_r, then the SS_r/σ^2 terms are independent χ^2 variables with df_r degrees of freedom if

$$\sum_{r=1}^{k} df_r = n - 1$$



F test of $\beta_1 = 0$ vs. $\beta_1 \neq 0$

Property

If $\beta_1 = 0$ so that all Y_i have the same mean $\mu = \beta_0$ and the same variance σ^2, SSE/σ^2 and SSR/σ^2 are independent χ^2 variables.

- When H_0 holds:

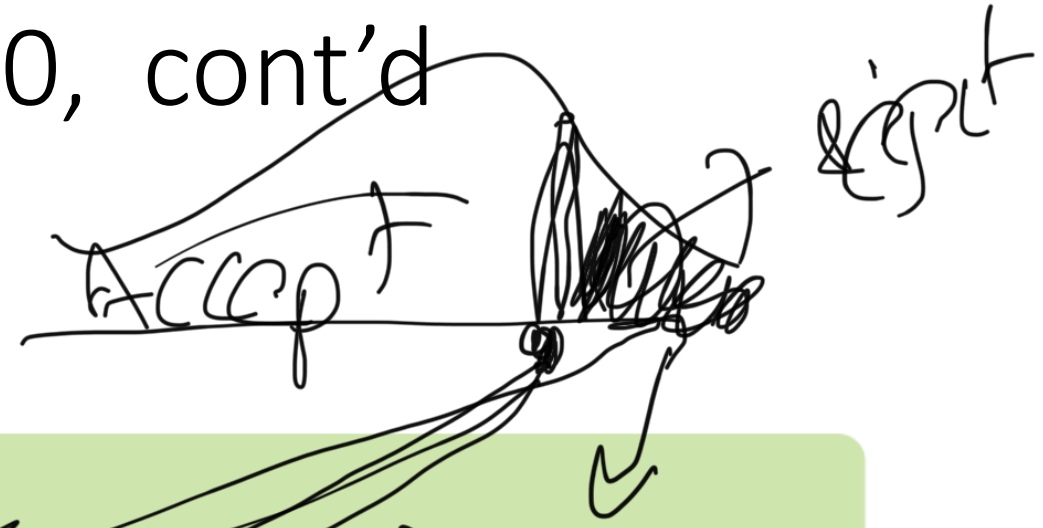
$$F^* = \frac{\frac{SSR}{\sigma^2}}{1} \div \frac{\frac{SSE}{\sigma^2}}{n-2} = \frac{MSR}{MSE} \sim \frac{\chi^2(1)}{1} \div \frac{\chi^2(n-2)}{n-2} \sim F(1, n-2)$$

- When H_a holds, F^* follows the noncentral F distribution.

F test of $\beta_1 = 0$ vs. $\beta_1 \neq 0$, cont'd

$F^* \sim F(1, n-2)$

The decision rule: α=Type I error



Decision

If $F^* \leq F(1-\alpha; 1, n-2)$, conclude H_0;

If $F^* > F(1-\alpha; 1, n-2)$, conclude H_a;

where $F(1-\alpha; 1, n-2)$ is the $(1-\alpha)100$ percentile of the approximate F distribution.

Example: Toluca Data

Two-Sided Test

$$H_0 : \beta_1 = \beta_{10}$$

$$\text{If } |t^*| = \frac{b_1 - \beta_{10}}{s\{b_1\}} \leq t(1 - \alpha/2; n - 2), \text{ conclude } H_0$$

$$\text{If } |t^*| \frac{b_1 - \beta_{10}}{s\{b_1\}} > t(1 - \alpha/2; n - 2), \text{ conclude } H_a$$

Example: Toluca Data, cont'd

Using the F test

$$H_0 : \beta_1 = \beta_{10} = 0$$

$$H_a : \beta_1 \neq \beta_{10} = 0$$

$\alpha = 0.05;\ n = 26;\ F(0.95; 1, 23) = 4.28$

$$\text{If } F^* \leq 4.28, \text{ conclude } H_0$$

We have

$$F^* = \frac{MSR}{MSE} = \frac{252,378}{2,384} = 105.9$$

What is the conclusion?

Equivalence of F test and t Test

For a given α level, the F test of

$$H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0$$

is equivalence algebraically to the two-tailed t test.

Decision

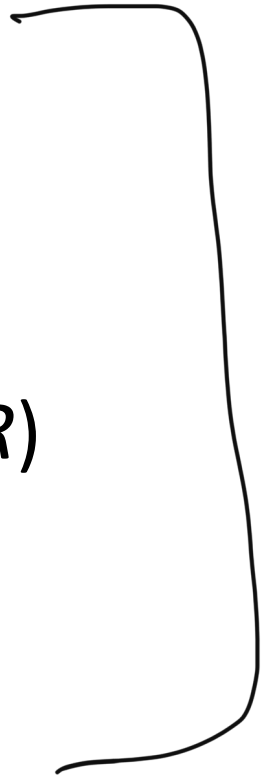
$$F^* = \frac{b_1^2}{s^2\{b_1\}} = \left(\frac{b_1}{s\{b_1\}}\right)^2 = (t^*)^2$$

$$[t(1-\alpha/2; n-2)]^2 = F(1-\alpha; 1, n-2)$$

t test: two-tailed; F test: one-tailed;

General Linear Test Approach

1. Fit the full model and $SSE(F)$
2. Fit the reduced model under H_0 and $SSE(R)$
3. Use test statistic and decision rule



Full Model

The *full* or *unrestricted model*:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$SSE(F)$:

$$SSE(F) = \sum \left[Y_i - (b_0 + b_1 X_i) \right]^2 = \sum (Y_i - \hat{Y}_i)^2 = SSE$$

Reduced Model

Hypothesis:

$$H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0$$

The *reduced* or *restricted model* when H_0 holds:

$$Y_i = \beta_0 + \varepsilon_i$$

$SSE(R)$:

$$SSE(R) = \sum \left[Y_i - (b_0) \right]^2 = \sum (Y_i - \bar{Y})^2 = SSTO$$

Test Statistics

$$SSE(F) \leq SSE(R)$$

SSE SSR

$SST = SSR \downarrow + SSE \uparrow$

- The more parameters are in the model, the better one can fit the data and the smaller are the deviations around the fitted regression function.

Test Statistics, cont'd

- When $SSE(F)$ is not much less than $SSE(R)$, using the full model does not account for much more of the variability of the Y_i than does the reduced model.

- $\Rightarrow$ Suggest that the reduced model is adequate i.e., H_0 holds.

- When $SSE(F)$ is close to $SSE(R)$, the variation of the observations around the fitted regression function for the full model is almost as great as the variation around the fitted regression function for the reduced model.

Test Statistics, cont'd

- A small difference $SSE(R) - SSE(F)$ suggests that H_0 holds.
- $\Leftrightarrow$ A large difference suggests that H_a holds.
- Test Statistic: a function of $SSE(R) - SSE(F)$:

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F} \sim F\text{distribution}$$

- when H_0 holds. Decision rule:

- If $F^* \leq F(1 - \alpha; df_R - df_F, df_F)$, conclude H_0
- If $F^* > F(1 - \alpha; df_R - df_F, df_F)$, conclude H_a

Test Statistics, cont'd

For testing whether or not $\beta_1 = 0$, we have

$$SSE(R) = SSTO \qquad SSE(F) = SSE$$

$$df_R = n - 1 \qquad df_F = n - 2$$

$$\Rightarrow F^* = \frac{MSR}{MSE}$$

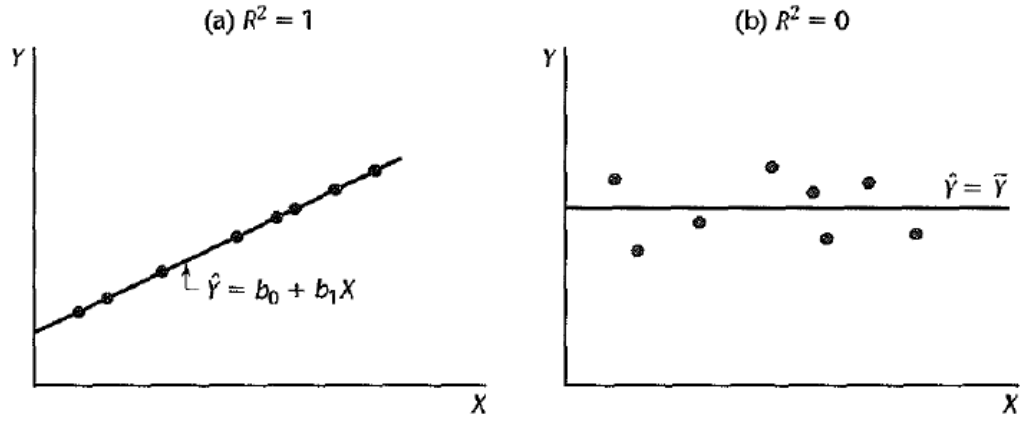
Coefficient of Determination

- SST (or SSTO) is a measure of the uncertainty in predicting Y when X is not considered.
- SSE measures the variation in the Y_i when a regression model utilizing the predictor variable X is employed.
- A natural measure of the effect of X in reducing the variation in Y is to express the reduction in variation (SST − SSE = SSR) as a proportion of the total variation:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Coefficient of Determination, cont'd

- $R^2=\frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ and given $SST \geq SSR \Rightarrow 0 \leq R^2 \leq 1$



$R^2 = 0$: There is no linear association between *X* and *Y* in the sample data, and the predictor variable *X* is of no help in reducing the variation in Y_i with linear regression.

$R^2 = 1$: The closer it is to 1, the greater is said to be the degree of linear association between *X* and *Y*.

Example: Toluca Company, cont'd

- fitreg<-lm(Hrs~Size,data=toluca)
- anova(fitreg)
- summary(fitreg)

```
> anova(fitreg)
Analysis of Variance Table

Response: Hrs
          Df Sum Sq Mean Sq F value    Pr(>F)
Size       1 252378  252378  105.88 4.449e-10 ***
Residuals 23  54825    2384
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Toluca Company, cont'd

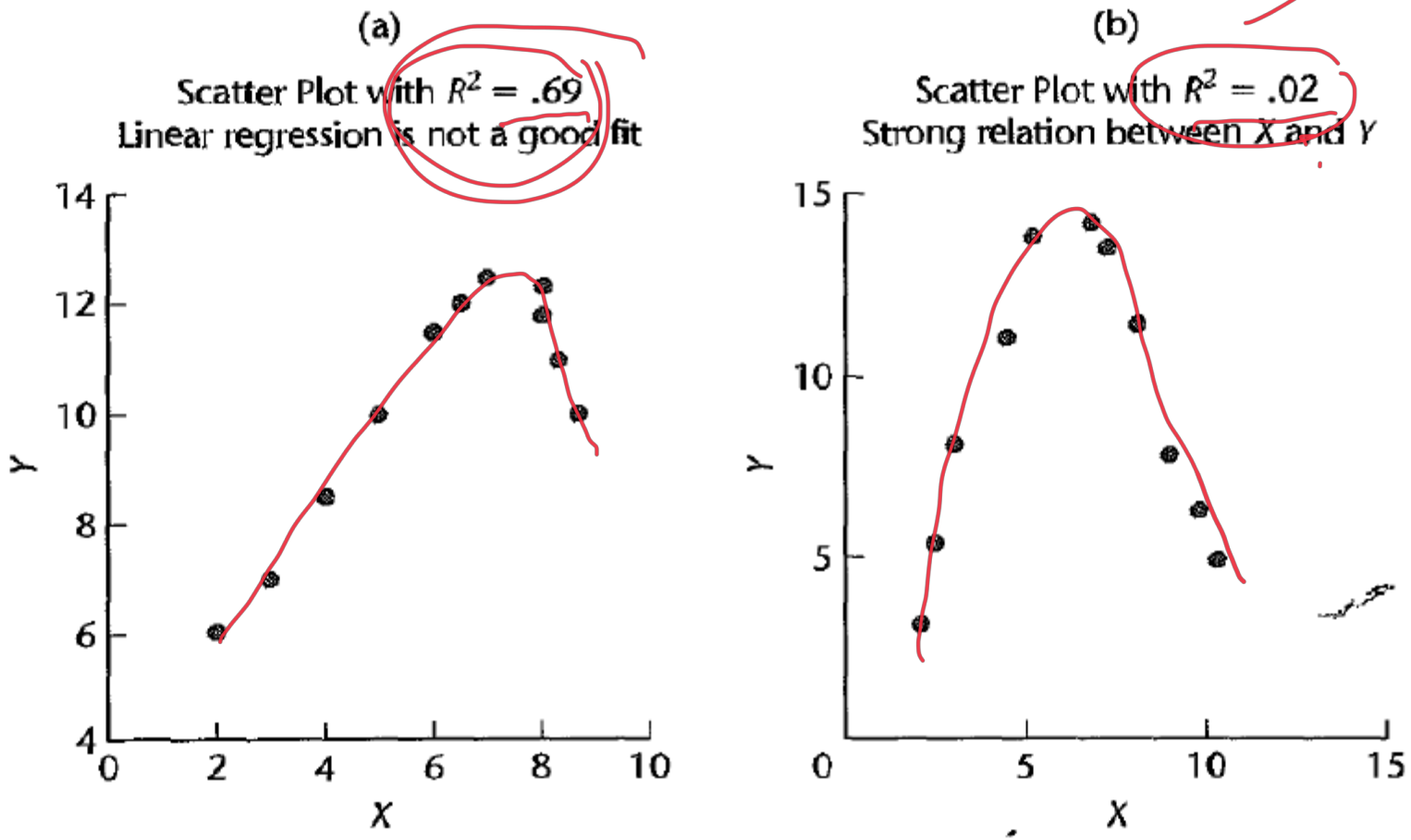
```
> summary(fitreg)

Call:
lm(formula = Hrs ~ Size, data = toluca)

Residuals:
    Min      1Q  Median      3Q     Max
-83.876 -34.088  -5.982  38.826 103.528

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.366     26.177   2.382   0.0259 *
Size           3.570      0.347  10.290 4.45e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 48.82 on 23 degrees of freedom
Multiple R-squared: 0.8215,     Adjusted R-squared: 0.8138
F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

- The variation in work hours is reduced by 82.2% when lot size is considered.

Coefficient of Determination, cont'd

- R^2 is widely used for describing the usefulness of a regression model.
- Serious misunderstanding:
  - A high R^2 indicates that useful predictions can be made. (Not necessarily correct. Ex: $X_h = 100$)
  - A high R^2 indicated that the estimated regression line is a good fit. (Not necessarily correct. Curvilinear)
  - A R^2 near 0 indicated that X and Y are not related. (Not necessarily correct. Curvilinear)

Coefficient of Determination, cont'd



Coefficient of Correlation

- Measure of linear association between Y and X when both Y and X are random is the coefficient of correlation. This measure is the signed square root of R^2:

$$r = \pm\sqrt{R^2}$$

correlation

- A plus or minus sign is attached to this measure according to whether the slope of the fitted regression line is positive or negative. Thus, the range of r is: $-1 \leq r \leq 1$.

Example:

- For the Toluca Company example, we obtained $R^2 = .822$.
- Treating X as a random variable, the correlation coefficient here is:
- $r = \pm\sqrt{.822} = .907$
- The plus sign is affixed since b1 is positive.

Normal Correlation Models

Assume that the X values are known constants?

The confidence coefficients and risks of errors refer to repeated sampling when X values are kept the same from sample to sample.

Frequently, it may not be appropriate to consider the X values as known constants.

- cannot control daily temperatures
- "height of person" vs. "weight of person": using correlation model

the normal correlation model

Bivariate Normal Distribution

Two variables Y_1, Y_2: bivariate normal distribution.

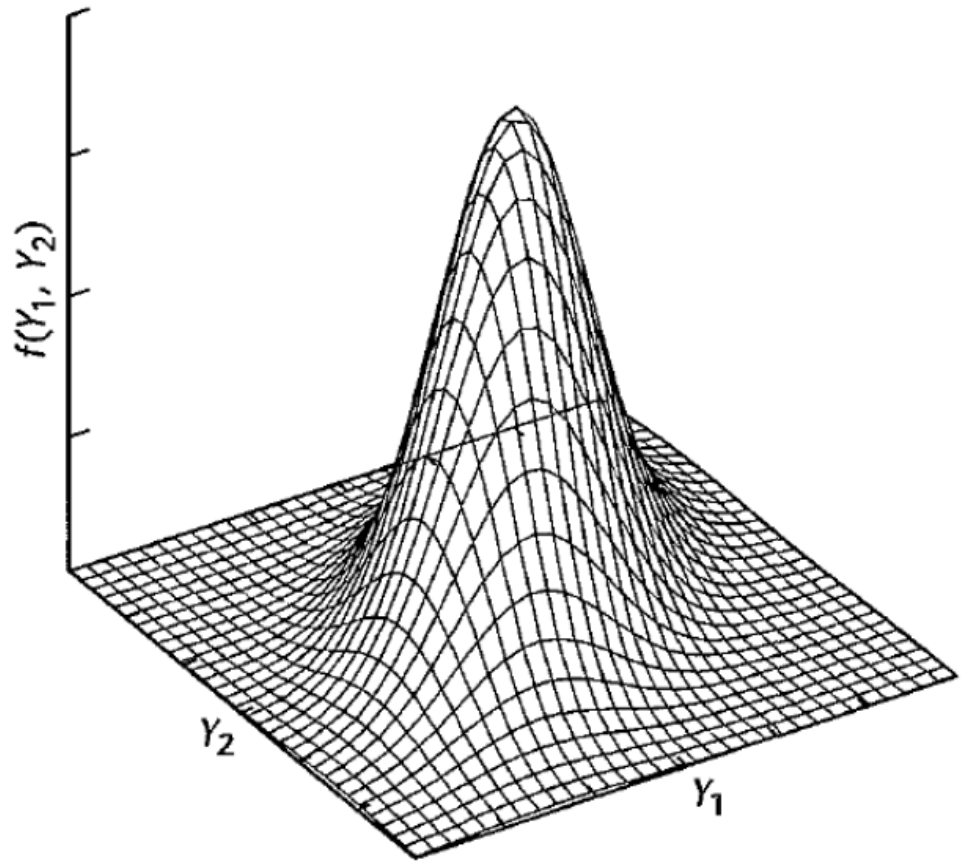
Density Function

The density function of the bivariate normal distribution:

$$f(Y_1, Y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}}} \exp\left\{ -\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{Y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho_{12}\left(\frac{Y_1-\mu_1}{\sigma_1}\right)\left(\frac{Y_2-\mu_2}{\sigma_2}\right) + \left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

Five parameters: $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_{12}$

Bivariate Normal Distribution, cont’d



Bivariate Normal Distribution, cont'd

Marginal Distribution

$Y_1, Y_2 \sim N_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_{12})$:

$$Y_1 \sim N(\mu_1, \sigma_1^2) \Rightarrow f_1(Y_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{Y_1 - \mu_1}{\sigma_1}\right)^2\right]$$

$$Y_2 \sim N(\mu_2, \sigma_2^2) \Rightarrow f_2(Y_2) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{1}{2}\left(\frac{Y_2 - \mu_2}{\sigma_2}\right)^2\right]$$

- When $Y_1, Y_2 \sim N_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_{12}) \Rightarrow$ $Y_1 \sim N(\mu_1, \sigma_1^2)$; $Y_2 \sim N(\mu_2, \sigma_2^2)$
- The converse is not generally true.

Bivariate Normal Distribution, cont'd

ρ_{12}: the coefficient of correlation between Y_1, Y_2

$$\rho_{12} = \frac{\sigma\{Y_1, Y_2\}}{\sigma\{Y_1\}\sigma\{Y_2\}} = \frac{\sigma_{12}}{\sigma_1\sigma_2}$$

$Y_1 \perp Y_2 \Rightarrow \sigma_{12} = 0 \Rightarrow \rho_{12} = 0$

If Y_1 and Y_2 are positively related$\Rightarrow$ σ_{12} and ρ_{12} are positive.

$-1 \leq \rho_{12} \leq 1$

Conditional Probability Distribution of Y_1

$Y_1, Y_2 \sim N_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_{12})$; and
$Y_1 \sim N(\mu_1, \sigma_1^2)$; $Y_2 \sim N(\mu_2, \sigma_2^2)$;

The density function of the conditional probability of Y_1 for given value of Y_2: $(Y_1|Y_2 \sim N(\alpha_{1|2} + \beta_{12} Y_2, \sigma_{1|2}))$

$$f(Y_1|Y_2) = \frac{f(Y_1, Y_2)}{f_2(Y_2)} = \frac{1}{\sqrt{2\pi}\sigma_{1|2}} \exp\left[-\frac{1}{2}\left(\frac{Y_1 - \alpha_{1|2} - \beta_{12} Y_2}{\sigma_{1|2}}\right)^2\right]$$

$$\alpha_{1|2} = \mu_1 - \mu_2 \rho_{12} \frac{\sigma_1}{\sigma_2}$$

$$\beta_{12} = \rho_{12} \frac{\sigma_1}{\sigma_2}$$

$$\sigma_{1|2}^2 = \sigma_1^2(1 - \rho_{12}^2)$$

Conditional Probability Distribution of Y_1 , cont'd

$\alpha_{1|2}$: the intercept of the line of regression of Y_1 and Y_2

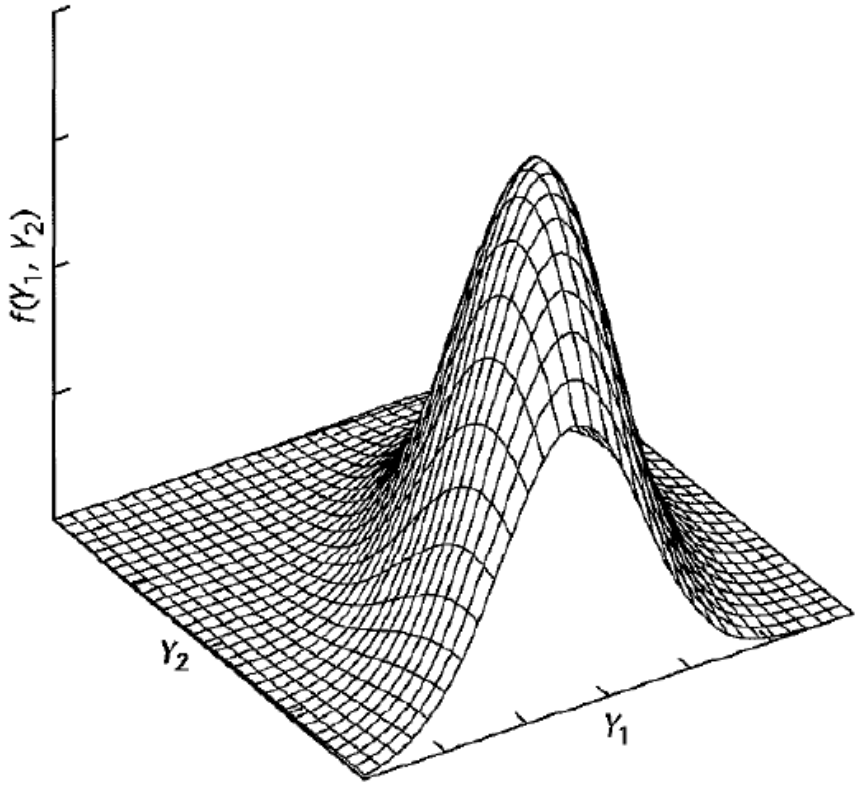
β_{12}: the slope of this line

The conditional distribution of Y_1, given Y_2, is equivalent to the normal error regression model (1.24).

Conditional Probability Distribution of Y_1 , cont'd

Three important characteristics of the conditional distributions of Y_1:

1. normal: slice a bivariate normal distribution vertically; scaled its area;



Conditional Probability Distribution of Y_1 , cont'd

2. The means of the conditional probability distributions of Y_1 fall on a straight line:

$$E\{Y_1 | Y_2\} = \alpha_{1|2} + \beta_{12} Y_2$$

3. All conditional probability distribution have the same standard deviation $\sigma_{1|2}$.

Equivalence to Normal Error Regression Model.

Inferences on Correlation Coefficients

To study the relationship between two variables: ρ_{12}

MLE of ρ_{12}:

$$r_{12} = \frac{\sum(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{[\sum(Y_{i1} - \bar{Y}_1)^2 \sum(Y_{i2} - \bar{Y}_2)^2]^{1/2}}$$

r_{12} is a biased estimator of ρ_{12}.

$-1 \leq r_{12} \leq 1$

Inferences on Correlation Coefficients, cont'd

Test

The population is bivariate normal

$$H_0 : \rho_{12} = 0 \quad \text{vs.} \quad H_a : \rho_{12} \neq 0$$

$$(\rho_{12} = 0 \Rightarrow Y_1 \perp Y_2)$$

$$\iff H_0 : \beta_{12} = 0 \quad \text{vs.} \quad H_a : \beta_{12} \neq 0$$

$$\iff H_0 : \beta_{21} = 0 \quad \text{vs.} \quad H_a : \beta_{21} \neq 0$$

Test statistics:

$$t^* = \frac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}} \sim t(n-2)$$

Inferences on Correlation Coefficients, cont'd

This test statistics is identical to the regression t^* test statistics $\frac{b_1}{s\{b_1\}}$. $(\because r = \sqrt{\frac{SSR}{SSTO}} = b_1\left(\frac{\sum(X_i-\bar{X})^2}{SSTO}\right)^{1/2})$

The appropriate decision rule to control the Type I error α:

$$\begin{cases} \text{If } |t^*| \leq t(1-\alpha/2; n-2), \text{ conlude } H_0 \\ \text{If } |t^*| > t(1-\alpha/2; n-2), \text{ conlude } H_a \end{cases}$$

Inferences on Correlation Coefficients, cont'd

- A national oil company: service station gasoline sales vs. sales of auxiliary product
- 23 of its service stations
- average monthly sales data: Y_1 =gasoline sales vs. Y_2 =auxiliary products and services
- $r_{12} = 0.52$; $\alpha = 0.05$
- To test whether or not the association was positive

$$H_0: \rho_{12} \leq 0$$

$$H_a: \rho_{12} > 0$$

- $t(.95;\ 21) = 1.721$

$$t^* = \frac{.52\sqrt{21}}{\sqrt{1-(.52)^2}} = 2.79$$

- We would conclude H_a, that $\rho_{12} > 0$. The P-value for this test is .006

Interval Estimation of ρ_{12}

- Fisher's transformation, is as follows:

$$z' = \frac{1}{2}\log_e\left(\frac{1+r_{12}}{1-r_{12}}\right)$$

- When n is large (25 or more is a useful rule of thumb), the distribution of z' is approximately normal with approximate mean and variance:

$$E\{z'\} = \zeta = \frac{1}{2}\log_e\left(\frac{1+\rho_{12}}{1-\rho_{12}}\right)$$

$$\sigma^2\{z'\} = \frac{1}{n-3}$$

- (Z', ξ): Table B.8 (on page 1332 shows the transformation correlation coefficient

Interval Estimation of ρ_{12}, *con't*

Interval estimate: $(n \geq 25)$

$$\frac{z' - \varsigma}{\sigma\{z'\}} \sim N(0,1)$$

$$\Rightarrow z' \pm z(1-\alpha/2)\sigma\{z'\}$$

A C.I. for ρ_{12} can be employed to test whether or not ρ_{12} has a specified value. (ex. 0.5)

$0 \leq \rho_{12}^2 \leq 1$: measures the relative reduction in the variability of Y_2 associated with the use of variable Y_1.

$$\rho_{12}^2 = \frac{\sigma_1^2 - \sigma_{1|2}^2}{\sigma_1^2}$$

$$\rho_{12}^2 = \frac{\sigma_2^2 - \sigma_{2|1}^2}{\sigma_2^2}$$

Example:

- An economist investigated food purchasing patterns by households in a midwestern city. Two hundred households with family incomes between $40,000 and $60,000 were selected to ascertain, among other things, the proportions of the food budget expended for beef and poultry, respectively. The economist expected these to be negatively related, and wished to estimate the coefficient of correlation with a 95 percent confidence interval.
- The point estimate of ρ_{12} was $r_{12} = -.61$. From table B.8, $Z' = -0.7089$ when $r_{12} = -.61$

$$\sigma\{z'\} = \frac{1}{\sqrt{200 - 3}} = .07125$$

$$z(.975) = 1.960$$

- Hence, the confidence limits for ξ are $-.7089 \pm 1.960(.07125)$, and the approximate 95 percent confidence interval is
  - $-.849 \leq \xi \leq -.569$
- Using Table B.8 to transform back to ρ_{12} , we obtain:
  - $-.69 \leq \rho_{12} \leq -.51$
- This confidence interval was sufficiently precise to be useful to the economist, confirming

the negative relation and indicating that the degree of linear association is moderately high.

Spearman Rand Correlation Coefficient

- When no appropriate transformations can be found, a nonparametric *rank correlation* procedure may be useful for making inferences about the association between Y_1 and Y_2.
- The ordinal Pearson product-moment correlation coefficient:

$$r_s = \frac{\sum(R_{i1} - \bar{R}_1)(R_{i2} - \bar{R}_2)}{[\sum(R_{i1} - \bar{R}_1)^2 \sum(R_{i2} - \bar{R}_2)^2]^{1/2}}$$

- Test: (two-sided)

H_0 : There is no association between Y_1 and Y_2

H_a : There is an association between Y_1 and Y_2

$$\Rightarrow t^* = \frac{r_s\sqrt{n-2}}{1-r_s^2} \sim t(n-2)$$

Example:

- A market researcher wished to examine whether an association exists between population size *(Y_1)* and per capita expenditures for a new food product *(Y_2).*
- The data for a random sample of 12 test markets are given below. Because the distributions of the variables do not appear to be approximately normal, a nonparametric test of association is desired.

	(1)	(2)	(3)	(4)
Test Market i	Population (in thousands) Y_{i1}	Per Capita Expenditure (dollars) Y_{i2}	R_{i1}	R_{i2}
1	29	127	1	2
2	435	214	8	11
3	86	133	3	4
4	1,090	208	11	10
5	219	153	7	6
6	503	184	9	8
7	47	130	2	3
8	3,524	217	12	12
9	185	141	6	5
10	98	154	5	7
11	952	194	10	9
12	89	103	4	1

- r_s is 0.985

$$t^* = \frac{.895\sqrt{12-2}}{\sqrt{1-(.895)^2}} = 6.34$$

- t(.995; 10) = 3.169. Since $|t^*| = 6.34 > 3.169$, we conclude Ha , that there is an association between population size and per capita expenditures for the food product.

Example:

Test Market i	(1) Population (in thousands) Y_{i1}	(2) Per Capita Expenditure (dollars) Y_{i2}	(3) R_{i1}	(4) R_{i2}
1	29	127	1	2
2	435	214	8	11
3	86	133	3	4
4	1,090	208	11	10
5	219	153	7	6
6	503	184	9	8
7	47	130	2	3
8	3,524	217	12	12
9	185	141	6	5
10	98	154	5	7
11	952	194	10	9
12	89	103	4	1

- r_s is 0.895
- t(.995; 10) = 3.169. Since |t*| = 6.34 > 3.169, we conclude Ha , that there is an association between population size and per capita expenditures for the food product.

$$t^* = \frac{.895\sqrt{12-2}}{\sqrt{1-(.895)^2}} = 6.34$$

- cor(expenditures$Y1,expenditures$Y2,method="spearman")
- cor.test(expenditures$Y1,expenditures$Y2,method="spearman")