

Mahler's Guide to Regression

Copyright ©2005 by Howard C. Mahler.

This study guide covers the material on the regression portion of the VEE-Applied Statistical Methods Exam.¹ **This study guide does not cover the material on time series, that is also on this exam.**²

Information in bold or sections whose title is in bold are more important for passing the exam. Larger bold type indicates it is extremely important. Information presented in italics (or sections whose title is in italics) should not be needed to directly answer exam questions and should be skipped on first reading. It is provided to aid the reader's overall understanding of the subject, and to be useful in practical applications.

For those who have trouble getting through the material, concentrate on the sections in bold.

Highly Recommended problems (about 1/6 of the total) are double underlined.

Recommended problems (about 1/6 of the total) are underlined.

Do at least the Highly Recommended problems your first time through.

It is important that you **do problems when learning a subject and then some more problems a few weeks later.**

The points assigned to each problem are based on 100 points for a four hour exam.

1 point problems are shorter than typical exam questions.

2 and 3 point problems are similar in length to typical exam questions.

4 point problems are longer than typical exam questions.

Solutions to problems are given at the end.³

The following tables will be provided to the candidate with the exam:

Normal Distribution, Chi-square Distribution, t-Distribution, and F-Distribution.⁴

I expect the questions on your exam to be on average somewhat more straightforward and basic than those on the old Course 4 Exam.

Feel free to send me any questions:

Howard Mahler, Email: hmahler@mac.com

¹ Econometric Models and Economic Forecasts, by Pindyck and Rubinfeld.

Chapters 1, 3, 4, 5, 6 (excluding Appendix 6.1), and Sections 8.1, 8.2, 10.1.

Sections 8.1, 8.2, and 10.1, covered in my Sections 29-31 and 38, were added to the syllabus in 2005.

² Econometric Models and Economic Forecasts, by Pindyck and Rubinfeld.

Chapters 15, 16 (excluding Appendix 16.1), 17 (excluding Appendix 17.1), and 18, cover time series.

³ Note that problems include both some written by me and some from past exams. The latter are copyright by the CAS and SOA, and are reproduced here solely to aid students in studying for exams. The solutions and comments are solely the responsibility of the author; the CAS and SOA bear no responsibility for their accuracy. While some of the comments may seem critical of certain questions, this is intended solely to aid you in studying and in no way is intended as a criticism of the many volunteers who work extremely long and hard to produce quality exams. In some cases I've rewritten these questions in order to match the notation in the current Syllabus.

⁴ <http://www.casact.org/admissions/syllabus/2005/webnotes.htm>

	Section #	Pages	Section Name
A	1	4 - 15	Fitting a Straight Line with No Intercept
	2	16 - 23	Fitting a Straight Line with an Intercept
	3	24 - 26	Residuals
	4	27 - 34	Analysis of Variance
B	5	35 - 46	R-Squared
	6	47 - 53	Normal Distribution
	7	54 - 55	Assumptions of Linear Regression
	8	56 - 59	Properties of Estimators
	9	60 - 66	Variances and Covariances
C	10	67 - 72	t-Distribution
	11	73 - 79	t-test
	12	80 - 83	Confidence Intervals for Estimated Parameters
	13	84 - 96	F Distribution
	14	97 - 105	Testing the Slope, Two Variable Model
D	15	106 - 113	Three Variable Regression Model
	16	114 - 123	Matrix Form of Multiple Regression
	17	124 - 144	Tests of Slopes, Multiple Regression
E	18	145 - 161	Additional Models
	19	162 - 172	Dummy Variables
	20	173 - 175	Piecewise Linear Regression
F	21	176 - 184	Weighted Regression
	22	185 - 189	Heteroscedasticity
	23	190 - 195	Tests for Heteroscedasticity
	24	196 - 206	Correcting for Heteroscedasticity
G	25	207 - 215	Serial Correlation
	26	216 - 225	Durbin-Watson Statistic
	27	226 - 232	Correcting for Serial Correlation
	28	233 - 235	Multicollinearity
H	29	236 - 248	Forecasting
	30	249 - 256	Testing Forecasts
	31	257 - 266	Forecasting with Serial Correlation
I	32	267 - 272	Standardized Coefficients
	33	273 - 276	Elasticity
	34	277 - 282	Partial Correlation Coefficients
	35	283	<i>Stepwise Regression</i>
	36	284	<i>Stochastic Explanatory Variables</i>
	37	285 - 288	<i>Generalized Least Squares</i>
	38	289 - 295	Nonlinear Estimation
39	296 - 308	Important Ideas and Formulas	
J		309 - 338	Solutions to Problems, Sections 1-14
K		339 - 370	Solutions to Problems, Sections 15-20
L		371 - 406	Solutions to Problems, Sections 21-38

Course 4 Exam Questions by Section of this Study Aid

Section	Sample	5/00	11/00	5/01	11/01	11/02	11/03	11/04
1		16	35				29	
2								
3								
4								
5	29	31				5 30		
6								
7								
8					35			
9				40				35
10								
11								
12			5		5	38		
13								
14		1						
15		35			13			
16							36	3
17	12 35	9	21		21	27	20	19
18				5		20		
19				24			5 9	
20								
21						7		
22								
23								
24			31	21	28			23
25								
26	30	24						
27			12	33				
28								
29				25				
30								
31								
32			37	13				
33								27
34	5					12		11
35								
36								
37								
38								

The CAS/SOA did not release the 5/02, 5/03, and 5/04 exams.

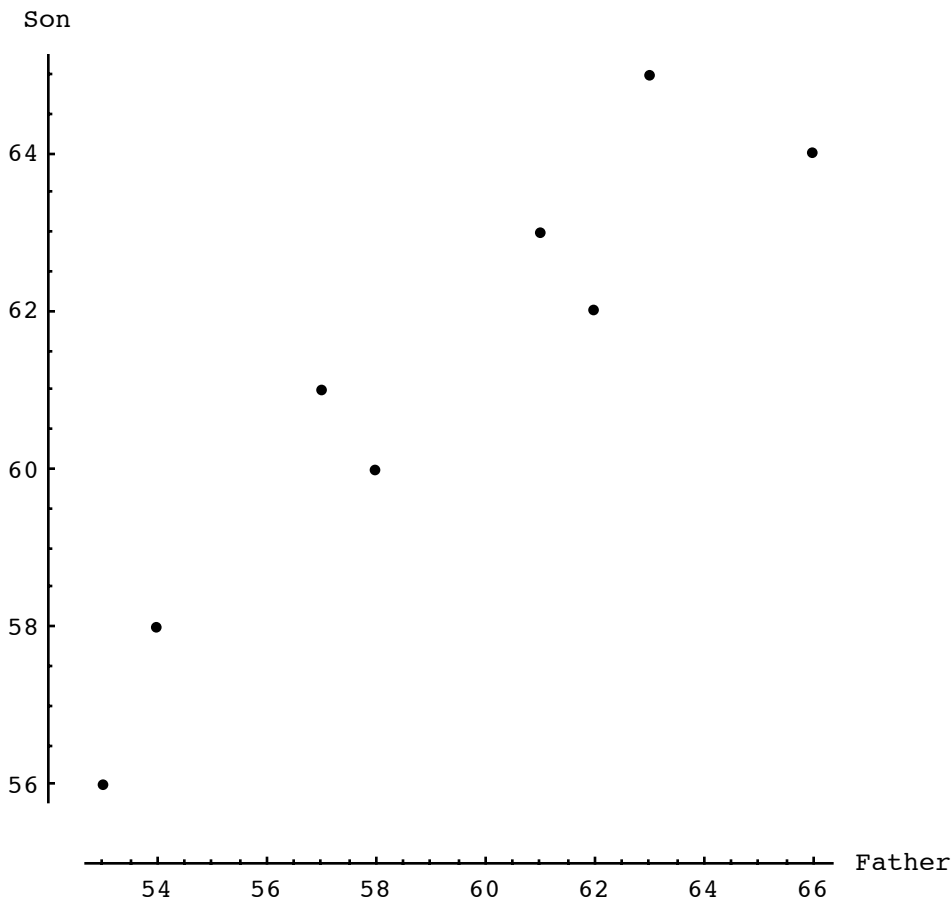
Sample Exam Q.40: statements C, D, and E are from chapter 7 of Pindyck & Rubinfeld, no longer on the Syllabus. 5/00, Q.16 and 11/00 Q.35 can be answered using ideas specifically discussed in chapter 7 of Pindyck & Rubinfeld, no longer on the syllabus. However, they can also be answered from first principles.

Section 1, Fitting a Straight Line with No Intercept

Assume we have the following heights of eight fathers and their adult sons (in inches):⁵

<u>Father</u>	<u>Son</u>
53	56
54	58
57	61
58	60
61	63
62	62
63	65
66	64

Here is a graph of this data:



There appears to be a relationship between the height of the father, X, and the height of his son, Y. A taller father seems to be more likely to have a taller son.

Straight Line with No Intercept:

⁵ There are only 8 pairs of observations solely in order to keep things simple.

Straight Line with No Intercept:

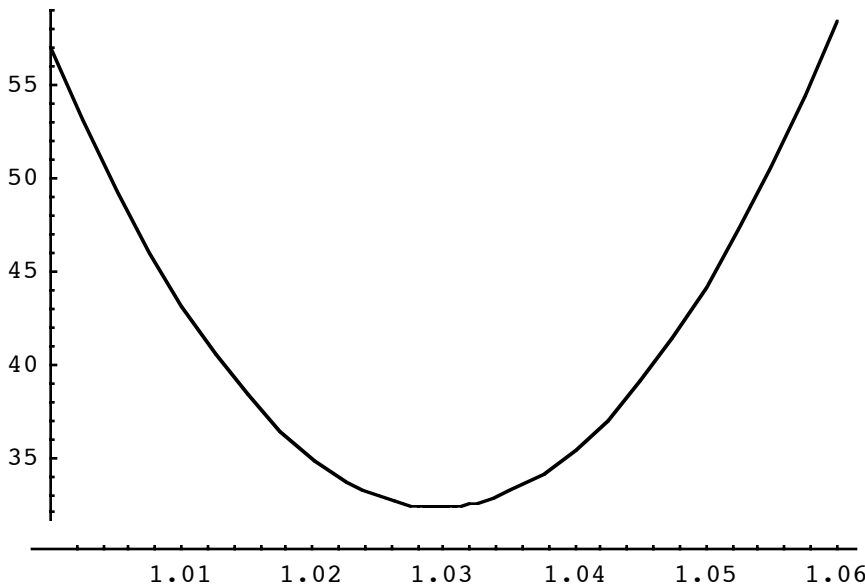
Let us assume $Y = \beta X$. We want to determine the “best” value of β . The most common way to do so is to **minimize the sum of the squared differences** between the height of each son estimated by our equation βX_i , and the actual height of that son Y_i .

Sum of Squared Errors = $\sum(Y_i - \beta X_i)^2$.

Exercise: If $\beta = 1.01$, what is the sum of squared errors?

[Solution: $\sum(Y_i - \beta X_i)^2 = (56 - 53.53)^2 + (58 - 54.54)^2 + (61 - 57.57)^2 + (60 - 58.58)^2 + (63 - 61.61)^2 + (62 - 62.62)^2 + (65 - 63.63)^2 + (64 - 66.66)^2 = 43.12$.]

Here is a graph of the sum of squared errors, as a function of β :



The smallest sum of squared errors corresponds to $\beta \approx 1.03$. We refer to 1.03 as the least squares estimate of the slope, β .

We would determine the least squares estimate of β algebraically, by setting equal to zero the partial derivative with respect to β of the sum of squared errors.⁶

$$0 = \partial \sum(Y_i - \beta X_i)^2 / \partial \beta = -2 \sum(Y_i - \beta X_i) X_i \Rightarrow 0 = \sum X_i Y_i - \sum \beta X_i X_i \Rightarrow \beta \sum X_i^2 = \sum X_i Y_i \Rightarrow \beta = \sum X_i Y_i / \sum X_i^2.$$

Exercise: Use the above equation in order to determine the least squares estimate of β .

[Solution: $\sum X_i Y_i = (53)(56) + \dots + (66)(64) = 29063$. $\sum X_i^2 = 53^2 + \dots + 66^2 = 28228$. estimate of $\beta = \sum X_i Y_i / \sum X_i^2 = 29063 / 28228 = 1.02958 \approx 1.03$.]

⁶ We treat β as the only variable.

The estimated value of beta is usually written with a ^ over it, $\hat{\beta}$. In this case $\hat{\beta} = 1.03$. Estimated values of other quantities are written in a similar manner.

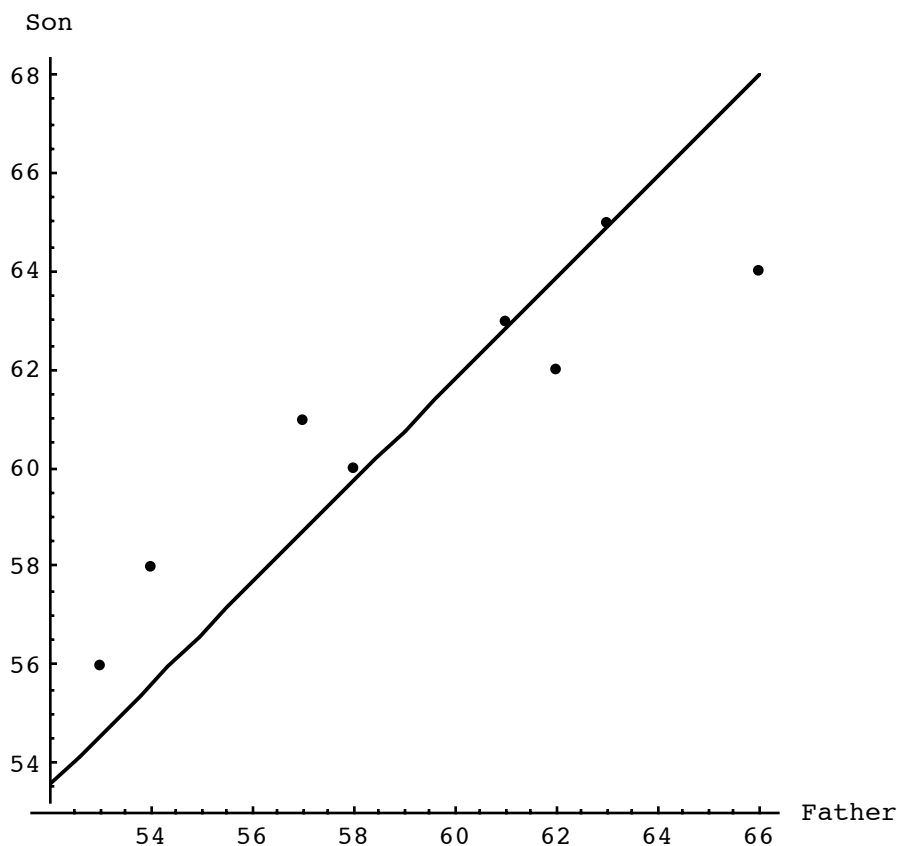
We do not expect any model to exactly predict the height of a son from the height of his father, therefore we include an error term in the model. The model we have been using is usually written $Y = \beta X + \epsilon$, or $Y_i = \beta X_i + \epsilon_i$, where ϵ_i is an error term.

In general, for the **least squares fit to the linear model with no intercept**,

$$Y = \beta X + \epsilon:$$

$$\hat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2}.$$

Here is a graph of the least squares line fit to the data on heights, with $\hat{\beta} = 1.03$:



Residuals:

The estimated height of the sons is written as \hat{Y} . $\hat{Y}_i = 1.03X_i$. The difference between each son's height and his height estimated by the model is the error, referred to as the residual.

Residual = actual - estimated.

The residual for son i is written as $\hat{\varepsilon}_i$.

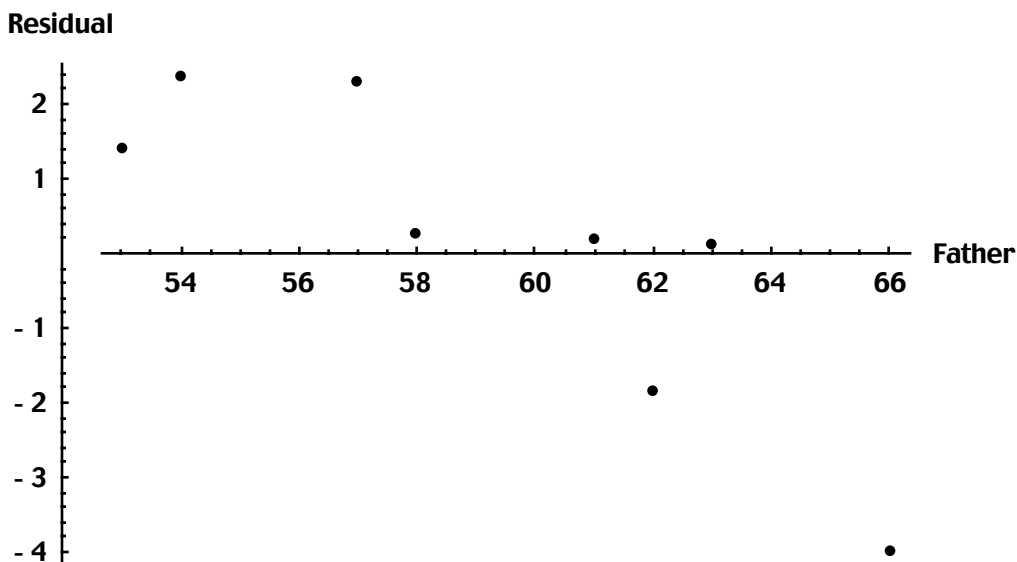
$$\hat{\varepsilon}_i \equiv Y_i - \hat{Y}_i.$$

Exercise: What are the residuals for the fitted model $\hat{Y}_i = 1.03X_i$?

[Solution: $\hat{\varepsilon}_i = 56 - 54.59, 58 - 55.62, 61 - 58.71, 60 - 59.74, 63 - 62.83, 62 - 63.86, 65 - 64.89, 64 - 67.98 = 1.41, 2.38, 2.29, .26, .17, -1.86, .11, -3.98$.

Comment: Note that these residuals do not sum to zero.⁷]

Here is a plot of these residuals:



Exercise: For the fitted model $\hat{Y}_i = 1.03X_i$, what is the sum of squared errors?

[Solution: $\sum \hat{\varepsilon}_i^2 = 1.41^2 + 2.38^2 + 2.29^2 + .26^2 + .17^2 + (-1.86)^2 + .11^2 + (-3.98)^2 = 32.3$.

Comment: This matches the result shown previously in a graph.]

The sum of squared errors is referred to the Error Sum of Squares or ESS.⁸

$$ESS = \sum \hat{\varepsilon}_i^2 = \sum (Y_i - \hat{Y}_i)^2.$$

In this case, ESS = 32.3.

⁷ As will be discussed later, when there is an intercept, the residuals sum to zero.

⁸ ESS is the sum of squared errors for a fitted model, as opposed to the sum of squared errors for any value of β .

Unbiased Estimator:

For the one variable linear regression model with no intercept, $Y_i = \beta X_i + \varepsilon_i$:

We assume $E[\varepsilon_i]$. Each error term has mean of zero.

Then $E[Y_i] = E[\beta X_i + \varepsilon_i] = \beta X_i$.

$$\hat{\beta} = \sum X_i Y_i / \sum X_i^2.$$

$$E[\hat{\beta}] = \sum X_i E[Y_i] / \sum X_i^2 = \sum X_i \beta X_i / \sum X_i^2 = \beta \sum X_i^2 / \sum X_i^2 = \beta.$$

Thus, $\hat{\beta}$ is an unbiased estimator of the slope β .

Expected Value of Residuals:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta} X_i.$$

$$E[\hat{\varepsilon}_i] = E[Y_i] - E[\hat{\beta}] X_i = \beta X_i - \beta X_i = 0.$$

Thus the expected value of each residual is zero.

However, it is important to note that the observed residuals will usually be nonzero. One is interested in the variance of each residual around its expected value.

Variances of Residuals:⁹

Assume we have:

i	X_i	$\text{Var}(\varepsilon_i)$
1	1	1
2	3	5
3	8	10

Exercise: Fit the model $Y_i = \beta X_i + \varepsilon_i$.

[Solution: $\hat{\beta} = \sum X_i Y_i / \sum X_i^2 = (Y_1 + 3Y_2 + 8Y_3)/74$.]

$$\hat{\varepsilon}_1 = Y_1 - X_1 \hat{\beta} = Y_1 - (1)(Y_1 + 3Y_2 + 8Y_3)/74 = (73Y_1 - 3Y_2 - 8Y_3)/74.$$

⁹ See 4, 11/03, Q.29.

Assuming ε_i and ε_j are independent,¹⁰ then Y_i and Y_j are independent.

$$\text{Var}[\hat{\varepsilon}_1] = \text{Var}[(73Y_1 - 3Y_2 - 8Y_3)/74] = (73^2\text{Var}[\varepsilon_1] + 3^2\text{Var}[\varepsilon_2] + 8^2\text{Var}[\varepsilon_3])/74^2 = (73^2(1) + 3^2(5) + 8^2(10))/74^2 = 1.098.$$

Note that since $E[\hat{\varepsilon}_1] = 0$, $E[\hat{\varepsilon}_1^2] = \text{Var}[\hat{\varepsilon}_1] = 1.098$.

Exercise: What is $\text{Var}[\hat{\varepsilon}_2]$?

[Solution: $\hat{\varepsilon}_2 = Y_2 - X_2\hat{\beta} = Y_2 - (3)(Y_1 + 3Y_2 + 8Y_3)/74 = (65Y_2 - 3Y_1 - 24Y_3)/74$.

$$\text{Var}[\hat{\varepsilon}_2] = \text{Var}[(65Y_2 - 3Y_1 - 24Y_3)/74] = (65^2\text{Var}[\varepsilon_2] + 3^2\text{Var}[\varepsilon_1] + 24^2\text{Var}[\varepsilon_3])/74^2 = (65^2(5) + 3^2(1) + 24^2(10))/74^2 = 4.9111.]$$

Formula for the Variance of the Residuals:

One can derive a general formula for $\text{Var}[\hat{\varepsilon}_i]$ as follows.

$$E[Y_i^2] = \text{Var}[Y_i] + E[Y_i]^2 = \text{Var}[\varepsilon_i] + \beta^2X_i^2.$$

$$Y_i \text{ and } Y_j \text{ are independent} \Rightarrow E[Y_iY_j] = \text{Cov}[Y_i, Y_j] + E[Y_i]E[Y_j] = 0 + \beta X_i\beta X_j = \beta^2X_iX_j, \quad i \neq j.$$

$$\hat{\beta} = \sum X_iY_i / \sum X_i^2.$$

$$E[\hat{\beta}^2] = E[\sum \sum X_iY_i X_jY_j / (\sum X_i^2)^2] = \sum X_i^2\text{Var}[\varepsilon_i] / (\sum X_i^2)^2 + \sum \sum \beta^2 X_i^2 X_j^2 / (\sum X_i^2)^2 = \sum X_i^2\text{Var}[\varepsilon_i] / (\sum X_i^2)^2 + \beta^2.$$

$$E[Y_j\hat{\beta}] = E[Y_j \sum X_iY_i / \sum X_i^2] = X_j\text{Var}[\varepsilon_j] / \sum X_i^2 + \beta^2 \sum X_jX_i^2 / \sum X_i^2 = X_j\text{Var}[\varepsilon_j] / \sum X_i^2 + X_j\beta^2.$$

$$\hat{\varepsilon}_i = Y_i - \hat{\beta}X_i.$$

$$E[\hat{\varepsilon}_i^2] = E[Y_i^2] + X_i^2E[\hat{\beta}^2] - 2X_iE[Y_i\hat{\beta}] =$$

$$\text{Var}[\varepsilon_i] + \beta^2X_i^2 + X_i^2\sum X_j^2\text{Var}[\varepsilon_j] / (\sum X_j^2)^2 + X_i^2\beta^2 - 2X_i^2\text{Var}[\varepsilon_i] / \sum X_j^2 - 2X_i^2\beta^2.$$

$$\text{Var}[\hat{\varepsilon}_i] = E[\hat{\varepsilon}_i^2] = \text{Var}[\varepsilon_i] + X_i^2\sum X_j^2\text{Var}[\varepsilon_j] / (\sum X_j^2)^2 - 2X_i^2\text{Var}[\varepsilon_i] / \sum X_j^2.$$

¹⁰ In the absence of serial correlation, we assume that the error terms are independent. Serial correlation will be discussed in a subsequent section.

Exercise: What is $\text{Var}[\hat{\epsilon}_3]$?

[Solution: $\text{Var}[\hat{\epsilon}_3] = \text{Var}[\epsilon_3] + X_3^2 \sum X_j^2 \text{Var}[\epsilon_j] / \{\sum X_j^2\}^2 - 2X_3^2 \text{Var}[\epsilon_3] / \sum X_j^2 =$
 $10 + 8^2\{(1^2)(1) + (3^2)(5) + (8^2)(10)\}/74^2 - (2)(8^2)(10)/74 = .720.$

Alternately, $\hat{\epsilon}_3 = Y_3 - X_3\hat{\beta} = Y_3 - (8)(Y_1 + 3Y_2 + 8Y_3)/74 = (10Y_3 - 8Y_1 - 24Y_2)/74.$

$\text{Var}[\hat{\epsilon}_3] = \text{Var}[(10Y_3 - 8Y_1 - 24Y_2)/74] = (10^2\text{Var}[\epsilon_3] + 8^2\text{Var}[\epsilon_1] + 24^2\text{Var}[\epsilon_2])/74^2 =$
 $(10^2(10) + 8^2(1) + 24^2(5))/74^2 = .720.]$

If all of the $\text{Var}[\epsilon_i]$ are equal, $\text{Var}[\epsilon_i] = \sigma^2$, then:¹¹

$\text{Var}[\hat{\epsilon}_i] = \text{Var}[\epsilon_i] + X_i^2 \sum_{j \neq i} X_j^2 \text{Var}[\epsilon_j] / \{\sum X_j^2\}^2 - 2X_i^2 \text{Var}[\epsilon_i] / \sum X_j^2 =$
 $\sigma^2 + X_i^2 \sum_{j \neq i} X_j^2 \sigma^2 / \{\sum X_j^2\}^2 - 2X_i^2 \sigma^2 / \sum X_j^2 = \sigma^2(1 - X_i^2 / \sum X_j^2) = \sigma^2 \sum_{j \neq i} X_j^2 / \sum X_j^2.$

$E[\hat{\epsilon}_i^2] = \text{Var}[\hat{\epsilon}_i] = \sigma^2(1 - X_i^2 / \sum X_j^2).$

$E[\text{ESS}] = E[\sum \hat{\epsilon}_i^2] = \sum E[\hat{\epsilon}_i^2] = \sum \sigma^2(1 - X_i^2 / \sum X_j^2) = \sigma^2(N - 1).$

Thus $\text{ESS}/(N-1)$ is an unbiased estimator of σ^2 .

Covariances of Residuals:

$E[\hat{\epsilon}_1 \hat{\epsilon}_2] = E[(Y_1 - \hat{\beta}X_1)(Y_2 - \hat{\beta}X_2)] = E[Y_1Y_2] + X_1X_2E[\hat{\beta}^2] - X_2E[Y_1\hat{\beta}] - X_1E[Y_2\hat{\beta}] =$
 $\beta^2X_1X_2 + X_1X_2\sum X_i^2\text{Var}[\epsilon_i] / \{\sum X_i^2\}^2 + X_1X_2\beta^2 - X_2X_1\text{Var}[\epsilon_1] / \sum X_i^2 - X_1X_2\text{Var}[\epsilon_2] / \sum X_i^2 - 2X_1X_2\beta^2$
 $= X_1X_2\sum X_i^2\text{Var}[\epsilon_i] / \{\sum X_i^2\}^2 - X_1X_2(\text{Var}[\epsilon_1] + \text{Var}[\epsilon_2]) / \sum X_i^2.$

$\text{Cov}[\hat{\epsilon}_1, \hat{\epsilon}_2] = E[\hat{\epsilon}_1 \hat{\epsilon}_2] - E[\hat{\epsilon}_1]E[\hat{\epsilon}_2] = X_1X_2\sum X_i^2\text{Var}[\epsilon_i] / \{\sum X_i^2\}^2 - X_1X_2(\text{Var}[\epsilon_1] + \text{Var}[\epsilon_2]) / \sum X_i^2.$

In the example, $\text{Cov}[\hat{\epsilon}_1, \hat{\epsilon}_2] = (1)(3)(686)/74^2 - (1)(3)(1 + 5)/74 = .1326.$

$\text{Corr}[\hat{\epsilon}_1, \hat{\epsilon}_2] = .1326 / \sqrt{((1.098)(4.911))} = .057.$ ¹²

$\text{Cov}[\hat{\epsilon}_i, \hat{\epsilon}_j] = X_iX_j\{\sum X_k^2\text{Var}[\epsilon_k] / \sum X_k^2 - \text{Var}[\epsilon_i] - \text{Var}[\epsilon_j]\} / \sum X_k^2.$

¹¹ Homoscedasticity is the term used for the situation in which all of the error terms have the same variance. Homoscedasticity and heteroscedasticity will be discussed in a subsequent section.

¹² While ϵ_1 and ϵ_2 are independent, the same is not true of the observed residuals.

Exercise: What is $\text{Corr}[\hat{\epsilon}_1, \hat{\epsilon}_3]$?

[Solution: $\text{Cov}[\hat{\epsilon}_1, \hat{\epsilon}_3] = (1)(8)\{686/74 - 1 - 10\}/74 = -.1870$.

$\text{Corr}[\hat{\epsilon}_1, \hat{\epsilon}_3] = -.1870/\sqrt{((1.098)(.720))} = -.210$.]

Exercise: What is $\text{Corr}[\hat{\epsilon}_2, \hat{\epsilon}_3]$?

[Solution: $\text{Cov}[\hat{\epsilon}_2, \hat{\epsilon}_3] = (3)(8)\{686/74 - 5 - 10\}/74 = -1.8583$.

$\text{Corr}[\hat{\epsilon}_2, \hat{\epsilon}_3] = -1.8583/\sqrt{((4.911)(.720))} = -.988$.]

For this example, the variance-covariance matrix of the residuals is:

$$\begin{pmatrix} 1.098 & .133 & -.187 \\ .133 & 4.911 & -1.858 \\ -.187 & -1.858 & .720 \end{pmatrix}$$

If all of the $\text{Var}[\epsilon_i]$ are equal, $\text{Var}[\epsilon_i] = \sigma^2$, then:

$$\text{Cov}[\hat{\epsilon}_i, \hat{\epsilon}_j] = X_i X_j \{ \sum_{k \neq i} X_k^2 \sigma^2 / \sum X_k^2 - \sigma^2 - \sigma^2 \} / \sum X_k^2 = -\sigma^2 X_i X_j / \sum X_k^2.$$

$$\text{Corr}[\hat{\epsilon}_i, \hat{\epsilon}_j] = -X_i X_j / \sqrt{(\sum_{k \neq i} X_k^2)(\sum_{k \neq j} X_k^2)}.$$

Simulation:

Assume we have $Y_i = 2X_i + \epsilon_i$, with ϵ_i independent and Normal with mean zero, and:

i	X_i	$\text{Var}(\epsilon_i)$
1	1	1
2	3	5
3	8	10

We can simulate this situation as follows:

1. Simulate ϵ_1, ϵ_2 , and ϵ_3 .
2. Calculate $Y_i = 2X_i + \epsilon_i$.
3. Fit a regression, $\hat{\beta} = \sum X_i Y_i / \sum X_i^2$.
4. Calculate $\hat{Y}_i = \hat{\beta} X_i$.
5. Calculate $\hat{\epsilon}_i = Y_i - \hat{Y}_i$.

For example, let -1.272, -.620, and .574, be 3 independent random Standard Normals.

$$\varepsilon_1 = -1.272\sqrt{1} = -1.272. \quad \varepsilon_2 = -.620\sqrt{5} = -1.386. \quad \varepsilon_3 = .574\sqrt{10} = 1.815.$$

$$Y_1 = 2X_1 + \varepsilon_1 = (2)(1) - 1.272 = .728. \quad Y_2 = (2)(3) - 1.386 = 4.614. \quad Y_3 = (2)(8) + 1.815 = 17.815.$$

$$\hat{\beta} = \sum X_i Y_i / \sum X_i^2 = 157.1/74 = 2.123.$$

$$\hat{Y}_1 = (2.123)(1) = 2.123. \quad \hat{Y}_2 = (2.123)(3) = 6.369. \quad \hat{Y}_3 = (2.123)(8) = 16.984.$$

$$\hat{\varepsilon}_1 = .728 - 2.123 = -1.395. \quad \hat{\varepsilon}_2 = 4.614 - 6.369 = -1.755. \quad \hat{\varepsilon}_3 = 17.815 - 16.984 = .831.$$

Exercise: Let 2.388, -.849, and -2.315, be 3 independent random Standard Normals. Simulate the above situation and determine the residuals.

$$[\text{Solution: } \varepsilon_1 = 2.388\sqrt{1} = 2.388. \quad \varepsilon_2 = -.849\sqrt{5} = -1.898. \quad \varepsilon_3 = -2.315\sqrt{10} = -7.321.]$$

$$Y_1 = (2)(1) + 2.388 = 4.388. \quad Y_2 = (2)(3) - 1.898 = 4.102. \quad Y_3 = (2)(8) - 7.321 = 8.679.$$

$$\hat{\beta} = \sum X_i Y_i / \sum X_i^2 = 86.126/74 = 1.164.$$

$$\hat{Y}_1 = (1.164)(1) = 1.164. \quad \hat{Y}_2 = (1.164)(3) = 3.492. \quad \hat{Y}_3 = (1.164)(8) = 9.312.$$

$$\hat{\varepsilon}_1 = 4.388 - 1.164 = 3.224. \quad \hat{\varepsilon}_2 = 4.102 - 3.492 = .610. \quad \hat{\varepsilon}_3 = 8.679 - 9.312 = -.633.]$$

Notice that each time we perform this simulation we get a different set of Y_i s, a different fitted slope, and a different set of residuals. If we ran this simulation 1000 times, we would get a set of 1000 different values for $\hat{\varepsilon}_1$. $\text{Var}[\hat{\varepsilon}_1]$ measures the variance of $\hat{\varepsilon}_1$ around its expected value of zero.

If we ran this simulation 1000 times, we would get a set of 1000 different values for $\hat{\beta}$.

$\text{Var}[\hat{\beta}]$ measures the variance of $\hat{\beta}$ around its expected value of $\beta = 2$.¹³

¹³ The variance of fitted regression parameters will be discussed subsequently.

Problems:

Use the following 4 observations for the next 3 questions:

X: 4 7 13 19
Y: 5 15 22 35

1.1 (1 point) Via least squares, fit to the above observations the following model $Y = \beta X + \varepsilon$.

What is the fitted value of β ?

(A) 1.4 (B) 1.5 (C) 1.6 (D) 1.7 (E) 1.8

1.2 (2 points) For the model fit in the previous question, what is the Error Sum of Squares?

(A) 11 (B) 12 (C) 13 (D) 14 (E) 15

1.3 (2 points) For the model $Y = 2X$, what is the sum of squared errors?

(A) 30 (B) 35 (C) 40 (D) 45 (E) 50

1.4 (2 points) You are given:

(i) The model is $Y_i = \beta X_i + \varepsilon_i$, $i = 1, 2, 3$.

(ii)

i	X_i	$\text{Var}(\varepsilon_i)$
1	1	1
2	5	2
3	10	4

(iii) The ordinary least squares residuals are $\hat{\varepsilon}_i = Y_i - \hat{\beta}X_i$, $i = 1, 2, 3$.

Determine $E(\hat{\varepsilon}_2^2 | X_1, X_2, X_3)$.

(A) 1.7 (B) 1.8 (C) 1.9 (D) 2.0 (E) 2.1

1.5 (1 point) Via ordinary least squares, the model $Y = \beta X + \varepsilon$ is fit to the following data:

X: 1 5 10 25
Y: 5 15 50 100

Determine $\hat{\beta}$.

(A) 3.9 (B) 4.0 (C) 4.1 (D) 4.2 (E) 4.3

1.6 (2 points) You are given the following data on the appraised values and sale prices of six homes, in thousands of dollars:

Appraised Value: 170 213 68 66 96 137
Sale Price: 180 245 85 88 132 156

Fit a least squares line with no intercept.

What is the estimated sale price of a home appraised at 300?

(A) 340 (B) 342 (C) 344 (D) 346 (E) 348

1.7 (2 points) Fit a least squares line with no intercept to the following data:

X	-2	-1	0	1	2	3	4	5
Y	-12	-7	0	6	14	21	24	31

What is the slope of the fitted line?

- (A) 6.1 (B) 6.2 (C) 6.3 (D) 6.4 (E) 6.5

1.8 (3 points) You are given the following information on the SAT scores for 10 students.

English:	630	700	540	610	580	670	710	630	580	760
Math:	570	710	570	580	610	640	660	640	670	720

Fit via least squares the model: Math Score = β (English Score).

What is the fitted value of β ?

- A. 0.98 B. 0.99 C. 1.00 D. 1.01 E. 1.02

1.9 (2, 5/85, Q. 19) (1.5 points) For the data $(x_1, y_1) = (1, 2)$ and $(x_2, y_2) = (5, 3)$ and the model $E(Y) = \beta x$, the least squares estimate of β is:

- A. 1/4 B. 17/26 C. 17/13 D. 17/6 E. 4

1.10 (4, 5/00, Q.16) (2.5 points) You are given:

(i) $x_1 = -2$ $x_2 = -1$ $x_3 = 0$ $x_4 = 1$ $x_5 = 2$

(ii) The true model for the data is $y = 10x + 3x^2 + \varepsilon$.

(iii) The model fitted to the data is $y = \beta^*x + \varepsilon^*$.

Determine the expected value of the least-squares estimator of β^* .

- (A) 6 (B) 7 (C) 8 (D) 9 (E) 10

1.11 (4, 11/00, Q.35) (2.5 points)

You are analyzing a large set of observations from a population.

The true underlying model is: $y = 0.1t - z + \varepsilon$.

You fit a two-variable model to the observations, obtaining: $y = 0.3t + \varepsilon^*$.

You are given: $\sum t = 0$. $\sum t^2 = 16$. $\sum z = 0$. $\sum z^2 = 9$.

Estimate the correlation coefficient between z and t .

- (A) -0.7 (B) -0.6 (C) -0.5 (D) -0.4 (E) -0.3

1.12 (4, 11/03, Q.29) (2.5 points) You are given:

(i) The model is $Y_i = \beta X_i + \varepsilon_i$, $i = 1, 2, 3$.

(ii)

i	X_i	$\text{Var}(\varepsilon_i)$
1	1	1
2	2	9
3	3	16

(iii) The ordinary least squares residuals are $\hat{\varepsilon}_i = Y_i - \hat{\beta}X_i$, $i = 1, 2, 3$.

Determine $E(\hat{\varepsilon}_1^2 \mid X_1, X_2, X_3)$.

- (A) 1.0 (B) 1.8 (C) 2.7 (D) 3.7 (E) 7.6

Section 2, Fitting a Straight Line with an Intercept

In the previous section we fit a straight line with no intercept, to the heights of fathers and their sons. In this section we will include an intercept in the model.

Let us assume $Y = \alpha + \beta X + \varepsilon$, where X is the height of the father and Y is the height of his son. This model with one independent variable and one intercept, is called the **two-variable regression model**. We want to determine the best values of α and β , those that **minimize the sum of the squared differences** between the height of each son estimated by our equation $\alpha + \beta X_i$, and the actual height of that son Y_i . **This is called the ordinary least squares regression.**¹⁴

$$\text{Sum of Squared Errors} = \sum (Y_i - \alpha - \beta X_i)^2.$$

We would determine the least squares estimates of α and β algebraically, by setting equal to zero the partial derivatives with respect to α and β of the sum of squared errors.

$$0 = \partial \sum (Y_i - \alpha - \beta X_i)^2 / \partial \alpha = -2 \sum (Y_i - \alpha - \beta X_i). \Rightarrow 0 = \sum Y_i - \sum \alpha - \sum \beta X_i. \Rightarrow$$

$$\alpha N + \beta \sum X_i = \sum Y_i, \text{ where } N \text{ is the number of observations.}$$

$$0 = \partial \sum (Y_i - \alpha - \beta X_i)^2 / \partial \beta = -2 \sum (Y_i - \alpha - \beta X_i) X_i. \Rightarrow 0 = \sum X_i Y_i - \alpha \sum X_i - \sum \beta X_i X_i. \Rightarrow$$

$$\alpha \sum X_i + \beta \sum X_i^2 = \sum X_i Y_i.$$

Exercise: Use the above equations in order to determine the least squares estimates of α and β for the fathers and sons example.

[Solution: $\sum X_i Y_i = (53)(56) + \dots + (66)(64) = 29063$.

$$\sum X_i^2 = 53^2 + \dots + 66^2 = 28228.$$

$$N = \text{number of observations} = 8.$$

$$\sum X_i = 53 + \dots + 66 = 474.$$

$$\sum Y_i = 56 + \dots + 64 = 489.$$

$$\text{Therefore, } 8\alpha + 474\beta = 489 \text{ and } 474\alpha + 28228\beta = 29063.$$

$$\text{Therefore, } \hat{\alpha} = \{(489)(28228) - (29063)(474)\} / \{(8)(28228) - 474^2\} = 27630 / 11148 = 24.07, \text{ and}$$

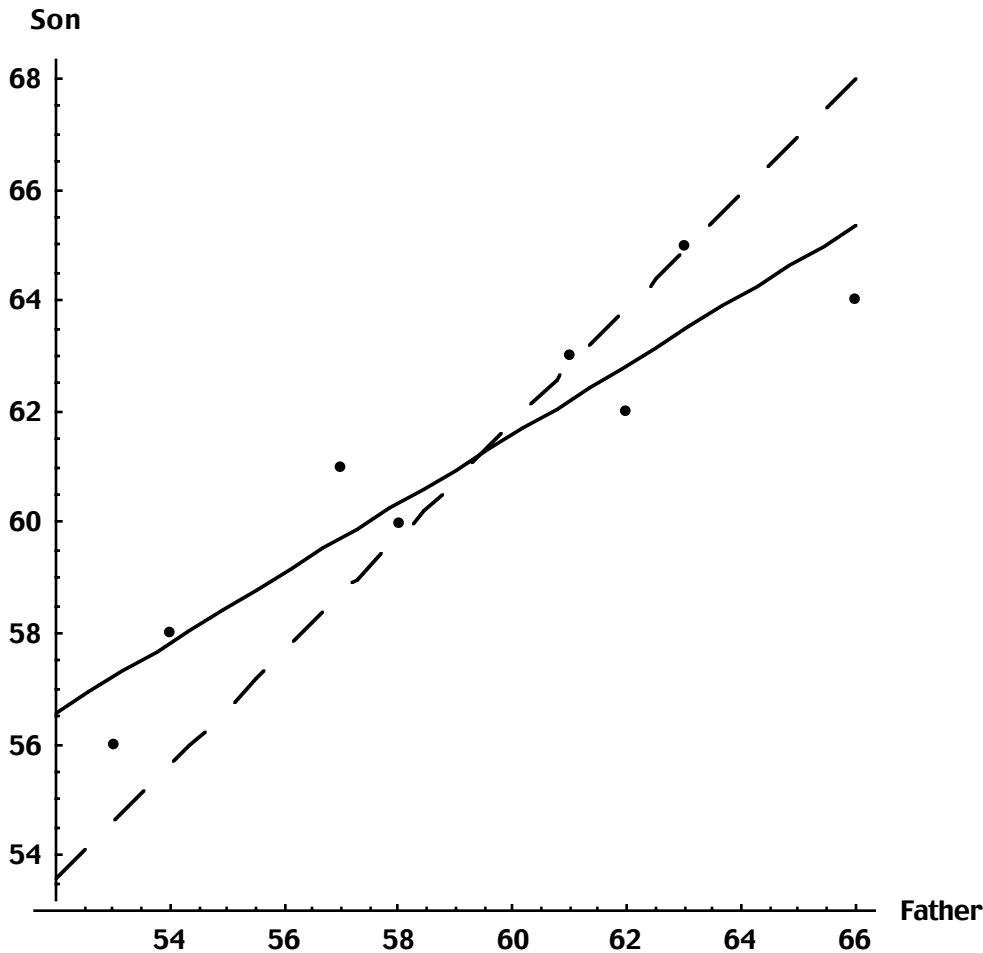
$$\hat{\beta} = \{(29063)(8) - (489)(474)\} / \{(8)(28228) - 474^2\} = 718 / 11148 = .6254.]$$

Thus the result of this regression is: $\hat{Y}_i = 24.07 + .6254X_i$.

For example, the fitted height of the first son is: $\hat{Y}_1 = 24.07 + .6254X_1 = 24.07 + (.6254)(53) = 57.216$. This of course differs somewhat from the actual height of the first son which is 56.

¹⁴ The term "regression" was introduced by Francis Galton in the 1880s, referring to his analysis of the heights of adult children versus the heights of their parents.

Here is a graph of the least squares line with intercept (solid) and that without intercept (dashed), each fit to the same data on heights:



The line with intercept (solid) seems to fit better than that without intercept (dashed). However, this will always be the case, since the line with no intercept is just a special case of that with intercept, with $\hat{\alpha} = 0$. How to determine whether the line with intercept is a significantly better fit, will be discussed subsequently.

We obtained two equations in two unknowns:

$$\alpha N + \beta \sum X_i = \sum Y_i, \text{ where } N \text{ is the number of observations.}$$

$$\alpha \sum X_i + \beta \sum X_i^2 = \sum X_i Y_i.$$

The solution is:

$$\hat{\alpha} = \{\sum Y_i \sum X_i^2 - \sum X_i \sum X_i Y_i\} / \{N \sum X_i^2 - (\sum X_i)^2\}, \text{ or } \hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}.$$

$$\hat{\beta} = \{N \sum X_i Y_i - \sum X_i \sum Y_i\} / \{N \sum X_i^2 - (\sum X_i)^2\}.$$

While these are perfectly valid solutions, most people find it easier to work with the variables in deviations form.¹⁵

Deviations Form:

Exercise: What is the mean height of the fathers?

[Solution: $(53 + 54 + 57 + 58 + 61 + 62 + 63 + 66)/8 = 474/8 = 59.25$.]

Exercise: What is the mean height of the sons?

[Solution: $(56 + 58 + 61 + 60 + 63 + 62 + 65 + 64)/8 = 489/8 = 61.125$.]

The mean of a variable is written as that variable with a bar over it. Mean of X is \bar{X} .

Mean height of fathers = $\bar{X} = 59.25$.

Mean height of sons = $\bar{Y} = 61.125$.

To convert a variable to deviations form, one subtracts its mean.

A variable in deviations form is written with a small rather than capital letter.

$$x_i = X_i - \bar{X}.$$

Exercise: What are x_i and y_i ?

[Solutions: $x_i = X_i - \bar{X} = (53, 54, 57, 58, 61, 62, 63, 66) - 59.25 = (-6.25, -5.25, -2.25, -1.25, 1.75, 2.75, 3.75, 6.75)$.

$y_i = Y_i - \bar{Y} = (56, 58, 61, 60, 63, 62, 65, 64) - 61.125 =$

$(-5.125, -3.125, -.125, -1.125, 1.875, .875, 3.875, 2.875)$.]

$$\sum x_i = \sum X_i - N\bar{X} = N\bar{X} - N\bar{X} = 0.$$

Verify that in this case both x_i and y_i sum to zero. In general, the sum of any variable in deviations form is zero. Therefore, its mean is also zero.

Variables in deviations always have a mean of zero.

¹⁵ In any case, you will be expected to know deviations form in order to answer exam questions.

Least Squares Regression in Deviations Form:

We have assumed the model, $Y_i = \alpha + \beta X_i + \varepsilon_i$, $i = 1, 2, \dots, N$.

Then adding up the N equations and dividing by N we get:

$$\bar{Y} = \alpha + \beta \bar{X} + \sum \varepsilon_i / N.$$

We have no reason to believe the average error is positive or negative. Let's assume it is zero.¹⁶ Then we would expect that: $\bar{Y} = \hat{\alpha} + \hat{\beta} \bar{X} \Rightarrow \hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}$.¹⁷

One could verify that this is true in general for the solutions given previously, for $\hat{\alpha}$ and $\hat{\beta}$. In any case, when we set the partial derivative of the squared error with respect to α equal to zero we got: $\hat{\alpha} N + \hat{\beta} \sum X_i = \sum Y_i \Rightarrow \bar{Y} = \hat{\alpha} + \hat{\beta} \bar{X} \Rightarrow \hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}$.

Exercise: For the regression fit to heights, verify that $\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}$.

[Solution: $\hat{\alpha} = 24.07$. $\hat{\beta} = .6254$. $\bar{X} = 59.25$. $\bar{Y} = 61.125$.
 $24.07 = 61.125 - (.6254)(59.25)$.]

We can take the original model and convert it to deviations form:

$$Y_i = \alpha + \beta X_i + \varepsilon_i = \bar{Y} - \hat{\beta} \bar{X} + \beta X_i + \varepsilon_i \Rightarrow Y_i - \bar{Y} = \beta(X_i - \bar{X}) + \varepsilon_i \Rightarrow y_i = \beta x_i + \varepsilon_i.$$

In deviations we get the same equation, except with no intercept. Based on the previous section, the least squares fit is: $\hat{\beta} = \sum x_i y_i / \sum x_i^2$.

In deviations form, the least squares regression to the two-variable (linear) regression model, $Y_i = \alpha + \beta X_i + \varepsilon_i$, has solution:

$$\hat{\beta} = \sum x_i y_i / \sum x_i^2$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}.$$

Exercise: Using deviations form, fit the least squares regression to the data on heights.

[Solution: $x_i = (-6.25, -5.25, -2.25, -1.25, 1.75, 2.75, 3.75, 6.75)$.

$y_i = (-5.125, -3.125, -.125, -1.125, 1.875, .875, 3.875, 2.875)$.

$$\sum x_i^2 = 143.5. \quad \sum x_i y_i = 89.75. \quad \hat{\beta} = \sum x_i y_i / \sum x_i^2 = 89.75 / 143.5 = .625.$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} = 61.125 - (.625)(59.25) = 24.1.$$

Comment: This matches the result obtained previously.]

¹⁶ Assumptions behind least squares regression will be discussed subsequently.

¹⁷ This is a good way to remember this formula.

A Shortcut when using Deviations Form:

$$\sum x_i y_i = \sum x_i (Y_i - \bar{Y}) = \sum x_i Y_i - \bar{Y} \sum x_i = \sum x_i Y_i - \bar{Y} 0 = \sum x_i Y_i.$$

$$\text{Therefore, } \hat{\beta} = \sum x_i Y_i / \sum x_i^2.$$

This can save some time on an exam, by avoiding having to calculate $y_i = Y_i - \bar{Y}$.

One would still have to calculate \bar{Y} , in order to calculate $\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}$.

Relation of Fitted Slope to Covariances or Correlations:

The sample variance of X is: $s_X^2 = \sum (X_i - \bar{X})^2 / (N - 1) = \sum x_i^2 / (N - 1)$.

The sample covariance of X and Y is: $\text{Cov}[X, Y] = \sum (X_i - \bar{X})(Y_i - \bar{Y}) / (N - 1) = \sum x_i y_i / (N - 1)$.

Therefore, $\hat{\beta} = \sum x_i y_i / \sum x_i^2 = \text{Cov}[X, Y] / \text{Var}[X]$.

$$\hat{\beta} = \text{Cov}[X, Y] / \text{Var}[X].$$

Exercise: The sample variance of X is 125. The sample covariance of X and Y is 167.

What is $\hat{\beta}$ in a two variable linear regression?

[Solution: $\hat{\beta} = \text{Cov}[X, Y] / \text{Var}[X] = 167/125 = 1.336$.]

Note that the sample correlation coefficient is: $r = \text{Cov}[X, Y] / (s_X s_Y) =$

$$\{\sum (X_i - \bar{X})(Y_i - \bar{Y}) / (N - 1)\} / \sqrt{\{\sum (X_i - \bar{X})^2 / (N - 1)\} \{\sum (Y_i - \bar{Y})^2 / (N - 1)\}} = \sum x_i y_i / \sqrt{(\sum x_i^2 \sum y_i^2)}.$$

Therefore, $\hat{\beta} = \sum x_i y_i / \sum x_i^2 = r \sqrt{(\sum y_i^2 / \sum x_i^2)} = r s_Y / s_X$.

$$\hat{\beta} = r s_Y / s_X.$$

Exercise: The sample correlation of X and Y is -.4. The sample standard deviation of X is 5.

The sample standard deviation of Y is 10. What is $\hat{\beta}$ in a two variable linear regression?

[Solution: $\hat{\beta} = r s_Y / s_X = -.4(10/5) = -.8$.]

Problems:

Use the following 4 observations for the next 2 questions:

X: 0 4 8 12
 Y: 834 889 916 950

2.1 (2 points) Via least squares, fit to the above observations the following model

$Y = \alpha + \beta X + \varepsilon$. What is the fitted value of β ?

- (A) 9.0 (B) 9.2 (C) 9.4 (D) 9.6 (E) 9.8

2.2 (1 point) Via least squares, fit to the above observations the following model

$Y = \alpha + \beta X + \varepsilon$. What is the fitted value of α ?

- (A) 800 (B) 810 (C) 820 (D) 830 (E) 840

2.3 (1 point) The sample covariance of X and Y is -413. The sample variance of X is 512.

What is $\hat{\beta}$ in a two variable linear regression?

- (A) -0.8 (B) -0.7 (C) -0.6 (D) -0.5 (E) -0.4

2.4 (3 points) You fit a two-variable linear regression to the following 5 observations:

X: 1 2 3 4 5
 Y: 202 321 404 480 507

What is the predicted value of Y, when $X = 7$?

- (A) 650 (B) 670 (C) 690 (D) 710 (E) 730

2.5 (1 point) The sample correlation of X and Y is 0.6. The sample variance of X is 36. The

sample variance of Y is 64. What is $\hat{\beta}$ in a two variable linear regression?

- (A) 0.6 (B) 0.8 (C) 1.0 (D) 1.2 (E) 1.4

2.6 (2 points) Use the following 4 observations:

X: -1 1 3 5
 Y: 3 4 7 6

Fit a least squares straight line and use it to estimate y for $x = 6$.

- (A) 7.0 (B) 7.2 (C) 7.4 (D) 7.6 (E) 7.8

2.7 (3 points) Use the following information:

<u>Year (t)</u>	<u>Loss Ratio (Y)</u>
1	82
2	78
3	80
4	73
5	77

You fit the following model: $Y = \alpha + \beta t + \varepsilon$.

What is the estimated Loss Ratio for year 7?

- (A) 71 (B) 72 (C) 73 (D) 74 (E) 75

2.8 (2 points) For each of five policy years an actuary has estimated the ultimate losses based on the information available at the end of that policy year.

<u>Policy Year</u>	<u>Estimated</u>	<u>Actual Ultimate</u>
1991	45	43
1992	50	58
1993	55	63
1994	60	76
1995	65	78

Let X_t be the actuary's estimate and Y_t be the actual ultimate.

Fit the ordinary least squares model, $Y_t = \alpha + \beta X_t$.

2.9 (2 points) You are given the following data on the number of exams and the salaries of seven actuaries in the land of Elbonia:

Number of Exams:	2	3	3	4	2	4	3
Salaries:	50	63	56	66	60	82	71

Fit a least squares line with intercept.

What is the estimated salary of an actuary with 5 exams?

- (A) 83 (B) 84 (C) 85 (D) 86 (E) 87

2.10 (3 points) You are given the following data for 10 taxi drivers. For each driver you are given the number of moving traffic violations during three years and the sum of their basic limit losses for Bodily Injury Liability Insurance (in \$1000) during the following three years.

Violations:	0	0	0	0	1	1	1	2	3	5
Losses:	10	0	43	0	35	0	80	0	58	64

Fit a least squares line with intercept.

What are the estimated losses for a taxi driver with 4 moving violations?

- (A) 49 (B) 51 (C) 53 (D) 55 (E) 57

2.11 (2 points). Use the following information:

<u>Year (t)</u>	<u>Claim Frequency (Y)</u>
1	3.18%
2	3.12%
3	3.30%
4	3.39%
5	3.41%

You fit via least squares the following model: $Y = \alpha + \beta t$.

What is the fitted claim frequency for year 7?

- (A) 3.51% (B) 3.53% (C) 3.55% (D) 3.57% (E) 3.59%

2.12 (2 points) For each of 10 insureds, you are given the number of claims in year 1 and the number of claims in year 2.

Insured:	1	2	3	4	5	6	7	8	9	10
Year 1:	0	0	0	0	0	1	1	1	1	1
Year 2:	0	0	0	1	1	0	0	1	1	1

Fit a least squares line with intercept, using the number of claims in year 1 as the independent variable and the number of claims in year 2 as the dependent variable.

What is the estimated future claim frequency for an insured with one claim in the most recent year?

- (A) 45% (B) 50% (C) 55% (D) 60% (E) 65%

2.13 (Course 120 Sample Exam #2, Q.1) (2 points) You fit the model $Y_i = \alpha + \beta X_i + \epsilon_i$ to the following data:

i	1	2	3
X_i	1	3	4
Y_i	2	Y_2	5

You determine that $\hat{\alpha} = 5/7$. Calculate Y_2 .

- (A) 0 (B) 1 (C) 2 (D) 3 (E) 4

2.14 (Course 120 Sample Exam #2, Q.7) (2 points) You are given the following information about a simple linear regression fit to 52 observations:

$$\sum_{i=1}^{10} X_i = 20. \quad \sum_{i=1}^{10} Y_i = 100. \quad \sum_{i=1}^{10} (X_i - \bar{X})^2/9 = 4. \quad \sum_{i=1}^{10} (Y_i - \bar{Y})^2/9 = 64.$$

You are also given that the simple correlation coefficient $r = -0.98$.

Determine the predicted value of Y when $X = 5$.

- (A) -10 (B) -2 (C) 11 (D) 30 (E) 37

Section 3, Residuals

Continuing the example from the previous section, the fitted height of a son is:

$$\hat{Y}_i = 24.07 + .6254X_i, \text{ where } X_i \text{ is the height of his father.}$$

As discussed previously, the difference between each son's height and his height estimated by the model is the residual.

$$\text{Residual} = \text{actual} - \text{estimated. } \hat{\varepsilon}_i \equiv Y_i - \hat{Y}_i.$$

Exercise: What are the residuals for the fitted model $\hat{Y}_i = 24.07 + .6254X_i$?

[Solution: $\hat{\varepsilon}_i = 56 - 57.216, 58 - 57.842, 61 - 59.718, 60 - 60.343, 63 - 62.219, 62 - 62.845, 65 - 63.470, 64 - 65.346 = -1.216, .158, 1.282, -.343, .781, -.845, 1.530, -1.346.$]

For the two variable linear regression model with an intercept:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta}X_i = Y_i - (\bar{Y} - \hat{\beta}\bar{X}) - \hat{\beta}X_i = y_i - \hat{\beta}x_i.$$

$$\sum \hat{\varepsilon}_i = \sum (y_i - \hat{\beta}x_i) = \sum y_i - \hat{\beta}\sum x_i = 0 - \hat{\beta}0 = 0.$$

For the linear regression model with an intercept, the sum of the residuals is always zero.¹⁸

This provides a good check of your work.

For the current example, $\sum \hat{\varepsilon}_i = -1.216 + .158 + 1.282 - .343 + .781 - .845 + 1.530 - 1.346 = 0.001$, zero subject to rounding.

Error Sum of Squares:

Exercise: For the fitted model $\hat{Y}_i = 24.07 + .6254X_i$, what is the sum of squared errors?

[Solution: $\sum \hat{\varepsilon}_i^2 = 1.216^2 + .158^2 + 1.282^2 - .343^2 + .781^2 - .845^2 + 1.53^2 + 1.346^2 = 8.741.$]

The sum of squared errors = Error Sum of Squares = ESS = $\sum \hat{\varepsilon}_i^2 = \sum (Y_i - \hat{Y}_i)^2$.

In this case, ESS = 8.741.

The error sum of squares will be discussed further in the section on Analysis of Variance.

¹⁸ This is not necessarily true for a model with no intercept.

Other Properties of Residuals:¹⁹

One can prove that the residuals are uncorrelated with X.

$\text{Corr}[\hat{\varepsilon}, X] = \text{Cov}[\hat{\varepsilon}, X] / \sqrt{(\text{Var}[\hat{\varepsilon}]\text{Var}[X])}$. $\text{Cov}[\hat{\varepsilon}, X] = E[\hat{\varepsilon}X] - E[\hat{\varepsilon}]E[X]$. Since the mean of the residuals is always zero, the numerator of $\text{Corr}[\hat{\varepsilon}, X]$ is:

$$\text{Cov}[\hat{\varepsilon}, X] = E[\hat{\varepsilon}X] = \sum \hat{\varepsilon}_i(X_i - \bar{X}) = \sum \hat{\varepsilon}_i x_i.$$

In the current example, $\sum \hat{\varepsilon}_i x_i$ is: $(-6.25)(-1.216) + (-5.25)(.158) + (-2.25)(1.282) + (-1.25)(-.343) + (1.75)(.781) + (2.75)(-.845) + (3.75)(1.53) + (6.75)(-1.346) = .01$, or zero subject to rounding.

In general, $\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta}X_i = Y_i - (\bar{Y} - \hat{\beta}\bar{X}) - \hat{\beta}X_i = y_i - \hat{\beta}x_i$.

$$\sum \hat{\varepsilon}_i x_i = \sum (y_i - \hat{\beta}x_i)x_i = \sum x_i y_i - \hat{\beta} \sum x_i^2 = 0, \text{ since } \hat{\beta} = \sum x_i y_i / \sum x_i^2.$$

Therefore, **$\text{Corr}[\hat{\varepsilon}, X] = 0$** .

As will be seen when we discuss analysis of variance, the difference between the fitted Y and the mean of Y, $\hat{Y}_i - \bar{Y}$, is also of interest. $\hat{Y}_i - \bar{Y} = \hat{\alpha} + \hat{\beta}X_i - \bar{Y} = \bar{Y} - \hat{\beta}\bar{X} + \hat{\beta}X_i - \bar{Y} = \hat{\beta}x_i$.

$$\sum \hat{\varepsilon}_i (\hat{Y}_i - \bar{Y}) = \sum \hat{\varepsilon}_i \hat{\beta}x_i = \hat{\beta} \sum \hat{\varepsilon}_i x_i = \hat{\beta}(0) = 0.$$

Thus, **$\hat{Y} - \bar{Y}$ and $\hat{\varepsilon}$ are uncorrelated.**

Exercise: In the current example, compute $\sum \hat{\varepsilon}_i (\hat{Y}_i - \bar{Y})$.

[Solution: $\hat{Y}_i - \bar{Y} = 57.22 - 61.125, 57.84 - 61.125, 59.72 - 61.125, 60.34 - 61.125, 62.22 - 61.125, 62.84 - 61.125, 63.47 - 61.125, 65.35 - 61.125 = -3.91, -3.28, -1.41, -.78, 1.09, 1.72, 2.34, 4.22$.

$$\sum \hat{\varepsilon}_i (\hat{Y}_i - \bar{Y}) = (-3.91)(-1.216) + (-3.28)(.158) + (-1.41)(1.282) + (-.78)(-.343) + (1.09)(.781) + (1.72)(-.845) + (2.34)(1.53) + (4.22)(-1.346) = -.006, \text{ or zero subject to rounding.}]$$

¹⁹ See Appendix 3.2 of Pindyck and Rubinfeld.

Problems:

3.1 (1 point) A regression is fit to 5 observations. The first four residuals are: 12, -4, -9, and 6. What is the error sum of squares?

- A. 220 B. 240 C. 260 D. 280 E. 300

3.2 (3 points) A two-variable regression is fit to the following 4 observations.

t	1	2	3	4
Y	30	40	55	60

What is the error sum of squares?

- (A) Less than 16
 (B) At least 16, but less than 17
 (C) At least 17, but less than 18
 (D) At least 18, but less than 19
 (E) At least 19

3.3 (2 points) A two-variable regression is fit to 5 observations.

The first four values of the independent variable X and the residuals are as follows:

i	1	2	3	4
X_i	7	12	15	21
$\hat{\epsilon}_i$	1.017	0.409	-0.557	-2.487

What is X_5 ?

- A. 29 B. 30 C. 31 D. 32 E. 33

3.4 (3 points) A two-variable regression is fit to 5 observations. The first 4 values of the

dependent variable Y and the corresponding fitted values \hat{Y} are as follows:

i	1	2	3	4
Y_i	13	25	36	40
\hat{Y}_i	18.036	22.989	30.419	40.325

What is Y_5 ?

- A. 48 B. 49 C. 50 D. 51 E. 52

Section 4, Analysis of Variance

The extremely important idea of **Analysis of Variance (ANOVA)** applies to regression analysis, as well as other subjects such as Buhlmann Credibility. As will be discussed, one can divide the Total Sum of Squares (TSS) into two pieces: the Regression Sum of Squares (RSS) and Error Sum of Squares (ESS).

Sample Variance:

Exercise: X_1 and X_2 are two independent, identically distributed variables, with mean μ and variance σ^2 . $\bar{X} = (X_1 + X_2)/2$. What is the expected value of: $(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2$?

[Solution: $(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 = (X_1/2 - X_2/2)^2 + (X_2/2 - X_1/2)^2 = 2(X_1 - X_2)^2/4 = X_1^2/2 + X_2^2/2 - X_1X_2$. $E[(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2] = E[X_1^2/2 + X_2^2/2 - X_1X_2] = (\sigma^2 + \mu^2)/2 + (\sigma^2 + \mu^2)/2 - \mu^2 = \sigma^2$.]

Thus $\{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2\}/(2 - 1) = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2$ is an unbiased estimator of σ^2 .

In general, with N independent, identically distributed variables X_i , $\Sigma(X_i - \bar{X})^2/(N - 1)$ is an unbiased estimator of the variance.

$\Sigma(X_i - \bar{X})^2/(N - 1)$ is called the sample variance of X.²⁰

The sample variance has in its numerator the sum of squared differences between each element and the mean. **The denominator of the sample variance is the number of elements minus one.**²¹ With this denominator, **the sample variance is an unbiased estimator of the underlying variance, when the underlying mean is unknown.**²²

Exercise: The heights of the eight sons were: 56, 58, 61, 60, 63, 62, 65, and 64.

What is the sample variance of heights of these sons?

[Solution: $\bar{Y} = 61.125$. Sample Variance $\equiv \Sigma(Y_i - \bar{Y})^2/(N - 1) =$

$\{(56 - 61.125)^2 + (58 - 61.125)^2 + (61 - 61.125)^2 + (60 - 61.125)^2 + (63 - 61.125)^2 + (62 - 61.125)^2 + (65 - 61.125)^2 + (64 - 61.125)^2\} / (8 - 1) = 64.875/7 = 9.27$.]

²⁰ The sample variance is used extensively in Empirical Bayesian Credibility.

²¹ As will be discussed subsequently, the number of degrees of freedom associated with the sum of squares in the numerator is N - 1.

²² The (non-sample) variance, $\Sigma(Y_i - \bar{Y})^2/N = 2\text{nd moment} - \text{square of mean}$, is a biased estimator of the true underlying variance.

Total Sum of Squares:

The **Total Sum of Squares or TSS** is defined as the sum of squared differences between Y_i and \bar{Y} .

$$\text{TSS} = \sum (Y_i - \bar{Y})^2 = \sum y_i^2.$$

Note that while both TSS and ESS involve squared differences from the observations of the dependent variable, Y_i , in the case of the total sum of squares we subtract the mean, \bar{Y} , while in the case of the error sum of squares we subtract the estimated height, \hat{Y}_i .

TSS is just the numerator of the sample variance of Y.

In this example, $\text{TSS} = (56 - 61.125)^2 + (58 - 61.125)^2 + (61 - 61.125)^2 + (60 - 61.125)^2 + (63 - 61.125)^2 + (62 - 61.125)^2 + (65 - 61.125)^2 + (64 - 61.125)^2 = 64.875$.

The TSS quantifies the total variation in the observations of the dependent variable. In the case of a series of experiments, TSS would measure the total variation in outcomes.

Error Sum of Squares:

Recall that the Error Sum of Squares or ESS is:²³

$$\text{ESS} = \sum \hat{\varepsilon}_i^2 = \sum (Y_i - \hat{Y}_i)^2.$$

As computed previously, for this example, $\text{ESS} = 8.741$.

Since $\sum \hat{\varepsilon}_i = 0$, ESS is the numerator of the variance of $\hat{\varepsilon}_i$.

Other Ways to write ESS for the two-variable model:

Since, $\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta}X_i = Y_i - (\bar{Y} - \hat{\beta}\bar{X}) - \hat{\beta}X_i = y_i - \hat{\beta}x_i$. $\hat{\beta} = \sum x_i y_i / \sum x_i^2$

$$\text{ESS} = \sum \hat{\varepsilon}_i^2 = \sum (y_i - \hat{\beta}x_i)^2 = \sum y_i^2 + \hat{\beta}^2 \sum x_i^2 - 2\hat{\beta} \sum x_i y_i =$$

$$\sum y_i^2 + (\sum x_i y_i / \sum x_i^2)^2 \sum x_i^2 - 2(\sum x_i y_i / \sum x_i^2) \sum y_i x_i = \sum y_i^2 - (\sum x_i y_i)^2 / \sum x_i^2 = \sum y_i^2 - \hat{\beta} \sum x_i y_i.$$

$$\sum \hat{\varepsilon}_i \hat{Y}_i = \sum \hat{\varepsilon}_i (\hat{\alpha} + \hat{\beta}X_i + \hat{\varepsilon}_i) = \hat{\alpha} \sum \hat{\varepsilon}_i + \hat{\beta} \sum \hat{\varepsilon}_i X_i + \sum \hat{\varepsilon}_i^2 = \hat{\alpha} 0 + \hat{\beta} 0 + \sum \hat{\varepsilon}_i^2 = \sum \hat{\varepsilon}_i^2 = \text{ESS}.$$

²³ The Error Sum of Squares, ESS, is also sometimes called the residual sum of squares.

Regression Sum of Squares:

There is a third sum of squared differences that is of importance.

The **Regression Sum of Squares or RSS**²⁴ is defined as the sum of squared differences between the fitted values and the mean of Y.

$$\text{RSS} = \sum (\hat{Y}_i - \bar{Y})^2.$$

Exercise: For the fitted model of heights, $\hat{Y}_i = 24.07 + .6254X_i$, what is the RSS?

[Solution: $\hat{Y}_i - \bar{Y} = 57.216 - 61.125, 57.842 - 61.125, 59.718 - 61.125, 60.343 - 61.125, 62.219 - 61.125, 62.845 - 61.125, 63.470 - 61.125, 65.346 - 61.125 = -3.909, -3.283, -1.407, -0.782, 1.094, 1.720, 2.345, 4.221.$

$$\text{RSS} = 3.909^2 + 3.283^2 + 1.407^2 + .782^2 + 1.094^2 + 1.720^2 + 2.345^2 + 4.221^2 = 56.121.]$$

In this example, $\text{RSS} = 56.121$.

$$\sum \hat{\varepsilon}_i = 0 \Rightarrow \sum Y_i - \hat{Y}_i \Rightarrow \sum Y_i = \sum \hat{Y}_i \Rightarrow \text{mean of } \hat{Y}_i \text{ is } \bar{Y}.$$

Exercise: For this example, verify that the mean of \hat{Y}_i is $61.125 = \bar{Y}$.

[Solution: $\sum \hat{Y}_i = 57.216 + 57.842 + 59.718 + 60.343 + 62.219 + 62.845 + 63.470 + 65.346 = 488.999. 488.999/8 = 61.125.$]

Therefore, $\text{RSS} = \sum (\hat{Y}_i - \bar{Y})^2$ is the numerator of the variance of \hat{Y}_i .

$$\begin{aligned} \text{RSS} &= \sum (\hat{Y}_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})(Y_i - \bar{Y} - \hat{\varepsilon}_i) = \sum (\hat{Y}_i - \bar{Y})(Y_i - \bar{Y}) - \sum (\hat{Y}_i - \bar{Y})\hat{\varepsilon}_i = \\ &\sum (\hat{Y}_i - \bar{Y})(Y_i - \bar{Y}), \text{ since } \hat{Y}_i - \bar{Y} \text{ and } \hat{\varepsilon}_i \text{ are uncorrelated.} \end{aligned}$$

Therefore, $\text{RSS} = \sum (\hat{Y}_i - \bar{Y})(Y_i - \bar{Y})$ = the numerator of the correlation of \hat{Y} and Y.

For this example, one can verify that $\sum (\hat{Y}_i - \bar{Y})(Y_i - \bar{Y}) = 56.121 = \text{RSS}$.

²⁴ The RSS is also sometimes called the sum of squares associated with the model as opposed to the error.

Other Ways to write RSS for the two-variable model:

$$\hat{Y}_i - \bar{Y} = \hat{\alpha} + \hat{\beta}X_i - \bar{Y} = (\bar{Y} - \hat{\beta}\bar{X}) + \hat{\beta}X_i - \bar{Y} = \hat{\beta}x_i.$$

$$RSS = \sum(\hat{Y}_i - \bar{Y})^2 = \hat{\beta}^2\sum x_i^2 = (\sum x_i y_i / \sum x_i^2)^2 \sum x_i^2 = (\sum x_i y_i)^2 / \sum x_i^2 = \hat{\beta} \sum x_i y_i.$$

TSS = RSS + ESS:

For this example, TSS = 64.875, RSS = 56.121, and ESS = 8.741.
 Note that RSS + ESS = 64.862, equal to TSS subject to rounding.

In general for a regression model with an intercept, the Total Sum of Squares is equal to the Regression Sum of Squares plus the Error Sum of Squares.

TSS = RSS + ESS.

The total variation has been broken into two pieces: that explained by the regression model, RSS, and that unexplained by the regression model, ESS.

This very important result holds for any linear regression model with an intercept, whether it is the two-variable model such as in this example, or a multivariable regression model to be discussed subsequently.

Proof of TSS = RSS + ESS:

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

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

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

It has been shown previously that $\hat{\varepsilon}_i$ and $\hat{Y}_i - \bar{Y}$ have a correlation of zero and $\sum \hat{\varepsilon}_i(\hat{Y}_i - \bar{Y}) = 0$. Thus the final term drops out and:

$$TSS = \sum \hat{\varepsilon}_i^2 + \sum (\hat{Y}_i - \bar{Y})^2 = ESS + RSS.$$

Note that the final term dropping out followed from a result that was proven for a regression model with an intercept. Analysis of Variance is not generally applied to a model without an intercept.

Alternately, for the two-variable model:

$$RSS + ESS = \hat{\beta} \sum x_i y_i + \sum y_i^2 - \hat{\beta} \sum x_i y_i = \sum y_i^2 = TSS.$$

Degrees of Freedom:

Degrees of Freedom:

Each of these sums of squares has a number of “**Degrees of Freedom**” associated with it. The number of degrees of freedom will be needed in order to perform t-tests and F-tests.

Exercise: There were four observations. In deviations form, $y_1 = -6$, $y_2 = -3$, and $y_3 = 2$. The value of y_4 is unreadable because a coworker spilled coffee on the report.

What is TSS?

[Solution: In deviations form, the sum of y_i is zero. Therefore, the missing y_4 must be: 7.

$TSS = \sum y_i^2 = 6^2 + 3^2 + 2^2 + 7^2 = 98.$]

In this exercise, we can compute TSS only knowing three out of the four y_i . In that sense, TSS only depends on 3 pieces of information. Therefore, we say TSS has 3 degrees of freedom. Another way to look at the same thing, is that TSS has 4 squared terms, but there is one linear constraint on the y_i : $\sum y_i = 0$. This linear constraint results in a loss of one degree of freedom, and therefore we have: $4 - 1 = 3$ degrees of freedom.

In any case, in general, if we have N points, **TSS has N - 1 degrees of freedom.**

Now $RSS = \sum (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}^2 \sum x_i^2$. Treating the x_i as known, we need only $\hat{\beta}$, one piece of information depending on the y_i , the outcomes of the experiment. Therefore, **RSS has 1 degree of freedom, for the two-variable model.**

Since $TSS = RSS + ESS$,

(number of d.f. for TSS) = (number of d.f. for RSS) + (number of d.f. for ESS).

Therefore, **ESS has N - 2 degrees of freedom, for the two-variable model.**

The number of degrees of freedom for ESS is the number of points minus the number of fitted parameters (including the fitted intercept.)

When we subsequently discuss the multivariable regression model, the following more general formulas will hold:

<u>Source of Variation</u>	<u>Sum of Squares</u>	<u>Degrees of Freedom</u>
Model	RSS	k - 1
Error	ESS	N - k
Total	TSS	N - 1

Where N is the number of points, and k is the number of variables including the intercept (k = 2 for the two-variable model with one slope and an intercept.)

Note that $TSS = RSS + ESS$, while $N - 1 = (k - 1) + (N - k)$.

Exercise: For the model fit to heights, $\hat{Y}_i = 24.07 + .6254X_i$, what are the degrees of freedom?
 [Solution: There are 8 points, $N = 8$. There are two variables, including the intercept, $k = 2$.
 RSS has $k - 1 = 2 - 1 = 1$ degree of freedom. ESS has $N - k = 8 - 2 = 6$ degrees of freedom.
 TSS has $N - 1 = 8 - 1 = 7$ degrees of freedom. Note $7 = 1 + 6$.]

ANOVA Table:

When you run a regression program on a computer, it will usually print out an Analysis of Variance (ANOVA) Table.²⁵

For example, for the two-variable model ($k = 2$) fit to heights, with eight observations ($N = 8$), the ANOVA Table might look like:²⁶

Source of Variation	Sum of Squares ²⁷	Degrees of Freedom	Mean Square
Model	56.13	1	56.13
Error	8.74	6	1.46
Total	64.87	7	9.27

Note that: $RSS + TSS = 56.13 + 8.74 = 64.87 = TSS$. $1 + 6 = 7$.
 $8.74/6 = 1.46$. $64.87/7 = 9.27 =$ sample variance of Y .

This ANOVA table was for a two-variable regression model. For a multivariable regression model, the ANOVA table would look similar, with of course the appropriate degrees of freedom.

²⁵ Those who have not done so, will probably benefit from running such a program a few times. Most such programs will print out many values related to items on the Syllabus, such as residuals, ESS, RSS, TSS, t-statistics, F-Statistics, Durbin-Watson Statistics, variance-covariance matrices, etc.

²⁶ Different computer programs may arrange things slightly differently. Also some additional information is probably shown relating to items we have yet to discuss. This ANOVA table was produced by Mathematica.

²⁷ The values for the sums of squares differ slightly from those shown previously, due to the lack of intermediate rounding in the calculations underlying what is shown here.

Problems:

4.1 (1 point) For a two variable model (slope and intercept) fit to 25 points, what are the degrees of freedom associated with the three sums of squares?

Use the following information for the next two questions:

For a multivariable regression, you have the following ANOVA Table, with certain items left blank:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Model	1020		255
Error			7
Total	1230		

4.2 (1 point) How many observations were there?

- (A) 30 or less (B) 35 (C) 40 (D) 45 (E) 50 or more

4.3 (1 point) How many variables were there in the regression, including the intercept?

- (A) 2 (B) 3 (C) 4 (D) 5 (E) 6 or more

4.4 (1 point) A regression model with 4 variables (3 slopes and one intercept) has been fit to 50 observations. What are the degrees of freedom associated with the Total Sum of Squares, Regression Sum of Squares, and Error Sum of Squares?

4.5 (10 points) You are given the following 17 observations:

X	0	25	50	75	100	125	150	175	200	225	250	275
Y	4.90	7.41	6.19	5.57	5.17	6.89	7.05	7.11	6.19	8.28	4.84	8.29

X	300	325	350	375	395
Y	8.91	8.54	11.79	12.12	11.02

Fit a two-variable linear regression.

Graph the data and the fitted line.

Graph the residuals.

Put together the ANOVA Table, showing the sum of squares and the degrees of freedom.

(You may use a computer, but do not use a regression software package.)

After completing your work, you may then check it using a regression software package.)

4.6 (1 point) A linear regression has been fit to 10 points, (X_i, Y_i) .

The fitted intercept is $\hat{\alpha}$. The fitted slope is $\hat{\beta}$.

$\sum(\hat{\alpha} + \hat{\beta}X_i - \bar{Y})^2 = 49$. The sample variance of Y is 8. Determine $\sum(\hat{\alpha} + \hat{\beta}X_i - Y_i)^2$.

- (A) 19 (B) 20 (C) 21 (D) 22 (E) 23

4.7 (165, 11/88, Q.2) (1.7 points) You are given the following table:

X_i	$E[Y_i]$	e_i
0	2.0	1.0
1	3.5	1.5
2	5.0	-2.0
3	6.5	0.5

where:

- (i) $E[Y_i]$ is the sequence of true values to be estimated.
- (ii) e_i are particular realizations of the error random variables ε_i .
- (iii) Y_i are the corresponding particular observations.
- (iv) \hat{Y}_i are obtained by linear regression of Y_i on X_i , including an intercept.

Determine $\sum_{i=0}^3 (Y_i - \hat{Y}_i)^2$.

- (A) 4 (B) 6 (C) 8 (D) 10 (E) 12

Note: The original exam question has been rewritten.