EEOS 601
UMASS/Online
Introduction to Probability
& Applied Statistics
Handout 14, Week 11
Tu 8/9/11- M 8/15/11
Revised: 2/20/11

WEEK 11: CHAPTER 11, REGRESSION

TABLE OF CONTENTS

List of Figures

List of Tables

List of m.files

Assignment

REQUIRED READING

- Larsen, R. J. and M. L. Marx. 2006. An introduction to mathematical statistics and its applications, 4th edition. Prentice Hall, Upper Saddle River, NJ. 920 pp.
  - Read All of Chapter 11

Understanding by Design Templates

Understanding By Design Stage I — Desired Results Week 11

LM Chapter 11 Regression **8/9/11 Tu - 8/15/11 M**

G Established Goals

- Fit an ordinary least squares regression line with appropriate confidence limits
- Graphically assess the goodness of fit of a regression
- Use a standard curve with inverse regression to find expected values
- Find the Pearson, Spearman's rho and Kendall's tau correlations for paired data

U Understand

- The etymology of regression is based on regression to the mean, which gives rise to the regression fallacy, one of the most common error in statistics
- The magnitude of R^2 is a poor way to judge the adequacy of a regression model
- The distinction between independent and correlated variables

Q Essential Questions

- Why are children's heights, on average, closer to the mean than that of their parents, and if that is the case, why doesn't the range of heights contract?
- What are the assumptions of OLS regression?
- How should replicates be allocated for producing a standard curve?

K *Students will know how to define (in words or equations)*

- Allometric regression, Anscombe's quartet, confidence intervals for regression, **exponential regression**, fiducial bounds, Hotelling-Woking interval, **inverse regression**, **logarithmic regression**, **logistic regression**, **OLS regression**, **Kendall's τ**, **Pearson's r**, **R^2**, **logarithmic regression**, **prediction interval**, **regression**, **residual**, **residual plot**, **regression to the mean**, **Spearman's ρ**

S *Students will be able to*

- Write Matlab programs to fit the OLS linear, exponential, and logarithmic regression models with confidence limits
- Graphically assess the goodness of fit of a regression model
- Use inverse regression to find the expected value of an unknown with fiducial limits
- Assess the equality of two slopes with a *t* test with $n+m-4$ df
- Assess the null hypothesis that the correlation between variables is 0

Understanding by Design Stage II — Assessment Evidence Week 11 8/9 Tu - 8/15 M

Chapter 11 All (P 647-731) Regression

- **Post in the discussion section by 8/17/11 W**
    - Find an example in your own field or the popular press of a regression analysis and describe it with a link to the article.
- **HW 4 Problems due Wednesday 8/17/11 W 10 PM**
    - **Basic problems (4 problems 10 points)**
        - Problem 11.2.8 using data from 8.2.11. Use case study 11.2.1, 11.2,2 (or the more complicated 11.2.3) as a model
        - Problem 11.2.20 Radioactive gold clearance, p 671-672. Solve for the half life using case study 11.2.6 as a model.
        - Problem 11.3.2 using case study 11.3.1 or 11.3.2 as models [No need to consider weighted regression]
        - Run the Matlab correlation plot simulation & note the effects of outliers, post a 1-paragraph summary of your conclusions
            - **http://www.dartmouth.edu/~raj/intro-stats.html**
    - **Advanced problems (2.5 points each)**
        - **Problem** Analyze Moore's Law transistor count data and test whether transistors per chip have been doubling every two years through 2010. Use LMcs110204_4th.m as a model
            - **http://en.wikipedia.org/wiki/Transistor_count**
        - **Problem**
    - **Master problems (1 only, 5 points)**
        - Redo Case Study 11.2.2 with data on social security costs from 1937 to 2008, available at **http://en.wikipedia.org/wiki/Social_Security_%28United_States%29**
            - Fit a **logarithmic model** and predict Social Security costs for 2015

Introduction to Regression

Chapter 11 is one of my favorite chapters in **Larsen & Marx (2006)**, but it is not one of my favorite chapters on regression analysis. Matlab and regression are made for each other but the link is linear algebra and Larsen & Marx don't express linear regression using matrix notation. It is pretty simple.

Regression in matrix notation

This section is drawn from two sources: **Draper & Smith (1998)** and **Searle (1982)**. A regression equation with one explanatory variable can be expressed as:

$$y_i = b_0 x_{i0} + b_1 x_{i1} + e_i. \tag{1}$$

Now define the following matrices and vectors:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} x_{10} & x_{11} \\ x_{20} & x_{21} \\ \vdots & \vdots \\ x_{N0} & x_{N1} \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} b_0 \\ b_2 \end{bmatrix}, \text{ and } \boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}$$

Then, equation (1) can be written in matrix terms as:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{e}. \tag{2}$$

By convention, variables indicating vectors or matrices are usually **bolded**.

The first column of the **X** matrix is set to all ones in order to fit the Y intercept. In some Matlab regression functions, you must add the first column of 1's. Other Matlab regression functions add the ones for you.

Estimation by least squares

The sum of squared residuals for equation (2) is (**y**-**Xb**)'(**y**-**Xb**) =**e**'**e**. The apostrophes indicate the use of the matrix transpose (turning the matrix or vector on its side so that the rows become columns and vice versa). The matrix multiplication e'e will produce a single number since **e** is a column vector. The regression parameter estimates can be obtained using the following matrix equation:

$$\hat{\boldsymbol{b}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}. \tag{3}$$

In this equation, called **the normal equation**, (**X'X**)$^{-1}$ indicates the inverse of the square matrix **X'X**, which in the case of **ordinary least squares regression (OLS)** with a single explanatory variable will be a 2x2 matrix. It is possible to include many explanatory variables in a regression and the same equation (3) would be used to solve for the regression parameters. A regression analysis with 10 explanatory variables and a y intercept would have an 11 x 11 square matrix **X'X**.

Singular explanatory matrices

In order to solve for $\hat{b}$, it must be possible to calculate the inverse of X'X, called the sums of squares and cross products matrix. This isn't an issue with a single explantory variable, but it can be a problem if more than one explanatory variable is used if the explanatory variables are strongly correlated or are linear functions of other combinations of explanatory variables. The X'X matrix will then be called singular or not full rank, and it an inverse can not be calculated. If there are strong correlation patterns among the explanatory variables, the X'X matrix may be called ill conditioned.

Matlab solves the normal equation using the following matrix notation, with the backslash \ operator being specific to Matlab[1]:

$$\hat{\boldsymbol{b}} = \boldsymbol{X} \backslash \boldsymbol{y}. \tag{4}$$

The expected values for a regression equation can be obtained using:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{b}}. \tag{5}$$

The residuals can be obtained by subtraction of vectors:

$$Residuals = \boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}. \tag{6}$$

You can perform a regression in Matlab and plot the results is just 4-7 lines, with one of those lines being a sort function to make sure that the prediction line plots properly.

```
% assume x is a vector for a single explanatory variable and Y is the response variable
x=x(:); % converts x to a column vector (if it isn’t already a column vector);
r=length(x);          % needed to create the correct column vector of 1's
X=[repmat(1,r,1) x];  % repmat writes a column of all 1's
B=X\Y                 % B is a 2 x 1 vector containing the Y intercept and slope
Yest=X*B;
[Xsorted,i]=sort(X(:,2));
plot(X(:,2),Y,’sk’,Xsorted,Yest(i),’-r’);
Resid=Y-Yest;
plot(X,Resid,’o’);xlabel(‘X’), ylabel(‘Residual’)
```

The sort function is needed because the predicted line will be jagged for a non-monotonic x vector.

Case Study 11.2.1 Rod Weights

A manufacturer of air conditioners is having problems because too many rods used in the air conditioners are too heavy. They need a predictive model for the finished weight based on the rough weight. Using matrix algebra and Matlab (**RScs110201_4th.m**), it is straightforward to find the regression equation to predict finished weight from rough weight (see Figure 1).

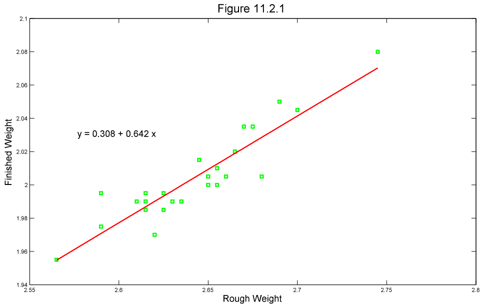


Figure 1 The regression of finished weight vs. Rough weight.

[1] At the Matlab prompt, type help mldivide for a description of Matlab’s backslash operator.

The residual plot (Figure 2) indicates no problems with the regression.

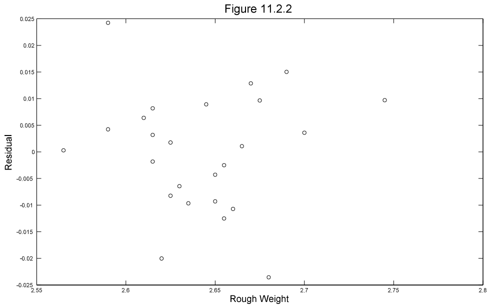


Figure 2 Residuals vs. Rough Weight.

Case Study 11.2.2 Social Security

Social security costs from selected years from 1965 to 1992 are plotted and analyzed with regression as shown in Figure 3.

The residual plot, shown in Figure 4, reveals a distinct concave-up pattern in the data.

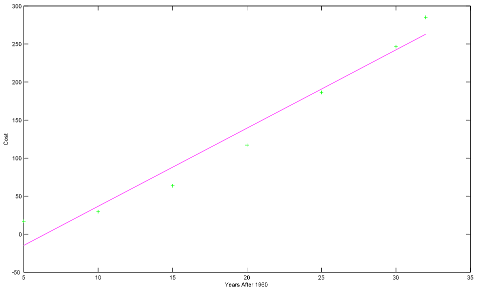


Figure 3 The regression of Social Security cosgts in $billions plotted vs. year since 1960.

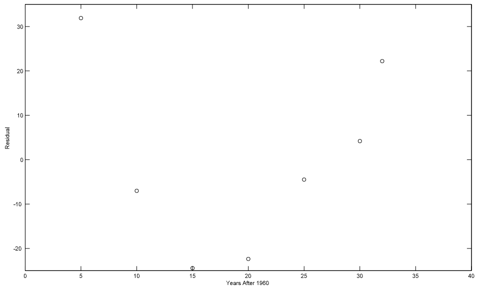


Figure 4 Residuals vs Year for Social Security costs.

Case Study 11.2.4 Chemists recovering CaO

This case study assesses whether the results from two chemists indicate more than random variation. Figure 5 shows the relationship betewen CaO recovered (in mg) and CaO present (in mg).

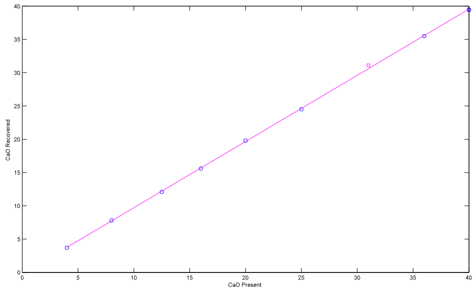


Figure 5 CaO recovered (mg) vs. CaO present (mg). A square symbol indicates the single sample analyzed by chemist B.

The residual plot shown in Figure 6 shows the residual for Chemist B. A convex hull surrounds the residuals for chemist A. Clearly chemist B's recovery of CaO exceeded that of chemist A.

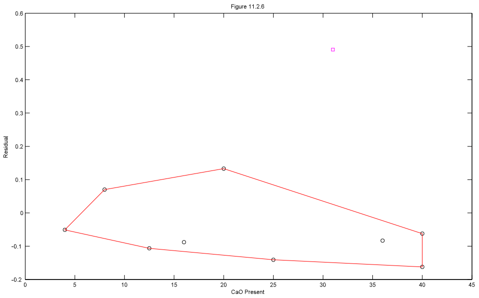


Figure 6 Residual plot indicating the 9 samples processed by chemist A, sorrounded by a convex hull, and the singl sample processed by chemist B.

Case Study 11.2.4

This case study, programmed as **LScs110204_4th.m**, is an example of an exponential regression. As shown in Figure 7, the transistors per chip has apparently increased exponentially with year since 1975. David Moore, one of the founders of Intel, predicted in 1975 that the number of transistors per chip should double every two years. **Larsen & Marx (2006, p. 664 & 666)** cite the widely misquoted estimate that chips should double every 18 months. According to Wikipedia's entry on Moore's law, Moore has stated that he never predicted an 18-month doubling.

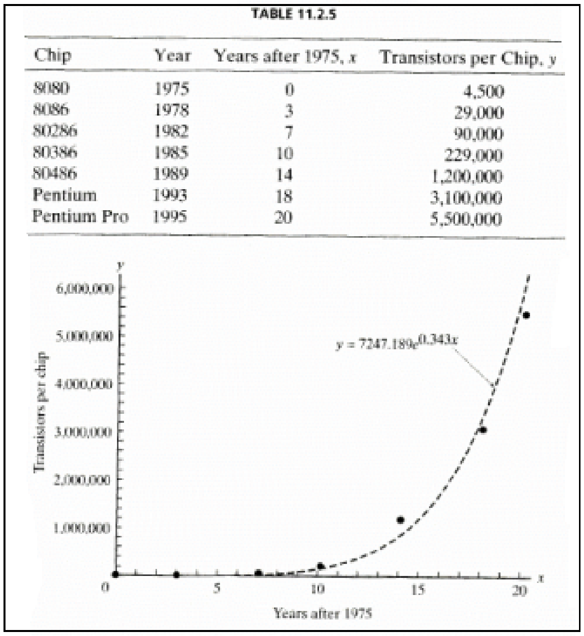


TABLE 11.2.5

Chip	Year	Years after 1975, x	Transistors per Chip, y
8080	1975	0	4,500
8086	1978	3	29,000
80286	1982	7	90,000
80386	1985	10	229,000
80486	1989	14	1,200,000
Pentium	1993	18	3,100,000
Pentium Pro	1995	20	5,500,000

Figure 7 The exponential regression between transistors per chip and years after 1975.

The equation describing Moore's relationship has the form $y = ae^{bx}$, so ln(y) and x should have a linear relationship. That relation is shown in Figure 8. The estimated slope is 0.343, which would translate into an doubling time = ln(2)/0.343 = 2.022 or a 2 year 8 day doubling time. The 95% confidence interval for the doubling time is 1.8 to 2.3 years.

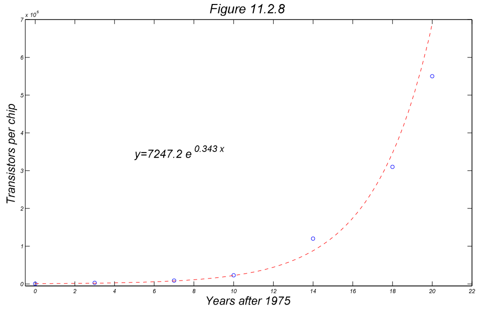


Figure 8 The exponential regression between transistors per chip and years after 1975.

Case Study 11.2.5

This case study is an example of **logarithmic regression**, in which both x and y are log-transformed. Any base logarithm could be used, but $\log_{10}$ and natural logs are by far the most common. In biology, especially ecology and physiology, logarithmic regression arises in the analysis of allometric relationships. The term **allometry** is derived from the Greek for 'different scales.' Allometric processes follow the **allometric equation**, Rate = αX^{β}. Beta (β) is the allometric exponent, and for a huge number of physiological rate processes, beta has a value of 0.75 when X is body weight.

In Case Study 11.25 data are provided, shown in Figure 9, on the first play date and first locomotion dates for 11 mammals. The asymptotic relationship is usually an indication that a log-log transform is required.

The fitted regression line is shown in Figure 10. As shown in the Matlab code below, a $\log_{10}$ transform was used for both x and y producing Y intercept and slope of 0.734 and 0.561. The leading coefficient, 5.42, in the equation shown in the figure is obtained by calculating $10^{0.734}$

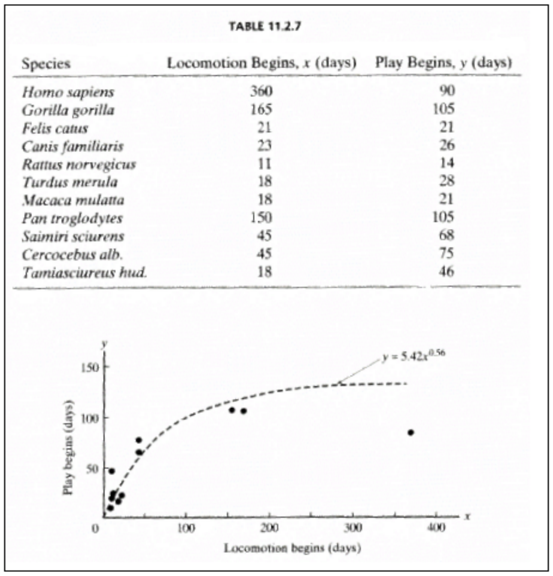


TABLE 11.2.7

Species	Locomotion Begins, *x* (days)	Play Begins, *y* (days)
Homo sapiens	360	90
Gorilla gorilla	165	105
Felis catus	21	21
Canis familiaris	23	26
Rattus norvegicus	11	14
Turdus merula	18	28
Macaca mulatta	18	21
Pan troglodytes	150	105
Saimiri sciurens	45	68
Cercocebus alb.	45	75
Tamiasciureus hud.	18	46

y
y = 5.42x^0.56
150
100
50
Play begins (days)
0
100
200
300
400
x
Locomotion begins (days)

Figure 9 Beginning of locomotion and playing in 11 mammals (days).

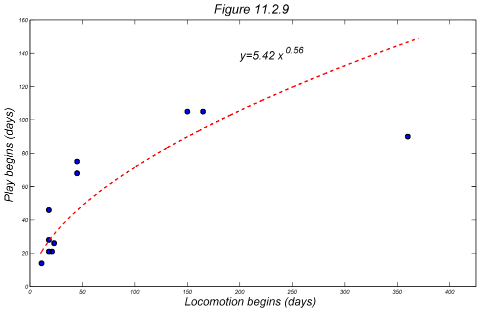


Figure 10 The exponential regression between transistors per chip and years after 1975.

Case 11.2.6

Case 11.2.6 is an example of logistic regression. In this case study, with data shown in Figure 11, graduation rate is regressed against SAT score. The book's use of logistic regression is the fit of the following relation to a set of data: ln({L-y}/y = a + bx. The approach is to examine the S-shaped pattern in the data and to estimate the asymptote, L in the logistic equation, by eye and then fit the equation. The fit is shown in Figure 12.

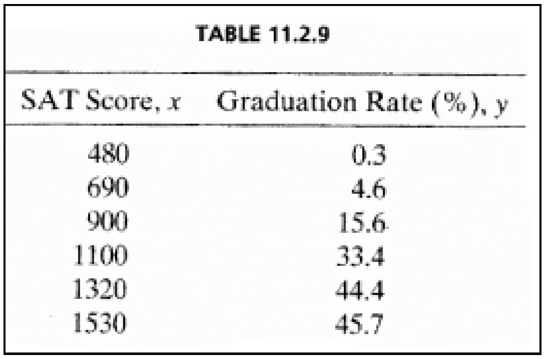


TABLE 11.2.9

SAT Score, x	Graduation Rate (%), y
480	0.3
690	4.6
900	15.6
1100	33.4
1320	44.4
1530	45.7

Figure 11 SAT scores and Graduation rate.

This isn't the standard use of logistic regression in statistics. Logistic regression differs from ordinary least squares regression, although both OLS regression and true logistic regression are subsets of generalized linear models. General linear models are usually fit by the method of maximum likelihood through a computer search algorithm for the optimal set of parameter estimates that maximizes the likelihood. Each form of generalized linear model has a link function *g*(u) linking the data to the linear regression equation equation. For OLS regression, g(u) = μ. If the assumptions of OLS regression are met, then the least squares estimates of the regression parameters are also the maximum likelihood estimates (Larsen & Marx 2006 Theorem 11.3.1).

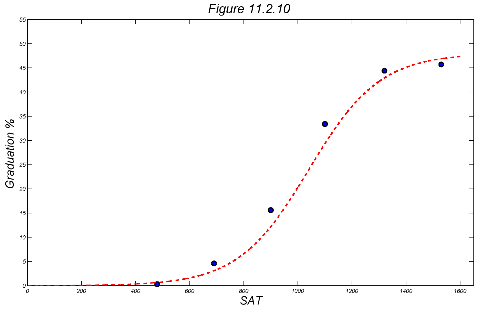


Figure 12 Graduation % vs. SAT scores using logistic regression. The asymptote, 48, was estimated by eye.

For logistic regression of proportions, π, the logit link function is used. *g*(π)=logit (π) = log (π/(1-π)). This logit link function is then used to fit the linear equation: logit(π) = $\beta_o + \beta_1 X_1 + ... + \beta_p X_p$. The logit link is also known as the log of the odds ratio, where π/(1-π) is the odds ratio. About two weeks of EEOS611 is devoted to the general linear model, including binary logistic regression, binomial logistic regression, Poisson regression and probit regression.

Case Study 11.3.1 Cigarettes & cancer

The coronary heart disease rates and cigarette consumption rates for 21 countries are shown in Figure 13. The data are analyzed with OLS regression shown in Figure 14.

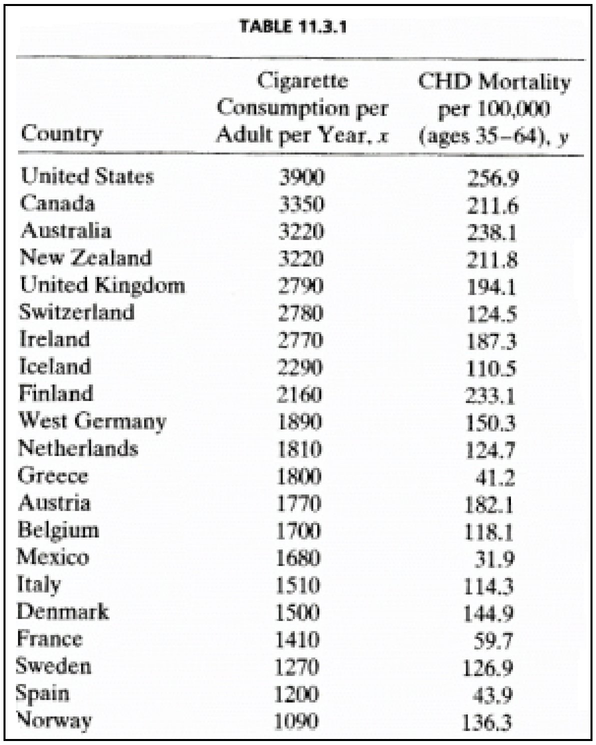


TABLE 11.3.1

Country	Cigarette Consumption per Adult per Year, x	CHD Mortality per 100,000 (ages 35–64), y
United States	3900	256.9
Canada	3350	211.6
Australia	3220	238.1
New Zealand	3220	211.8
United Kingdom	2790	194.1
Switzerland	2780	124.5
Ireland	2770	187.3
Iceland	2290	110.5
Finland	2160	233.1
West Germany	1890	150.3
Netherlands	1810	124.7
Greece	1800	41.2
Austria	1770	182.1
Belgium	1700	118.1
Mexico	1680	31.9
Italy	1510	114.3
Denmark	1500	144.9
France	1410	59.7
Sweden	1270	126.9
Spain	1200	43.9
Norway	1090	136.3

Figure 13 SAT scores and Graduation rate.

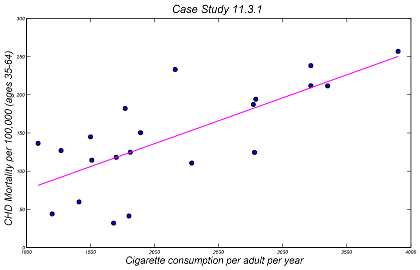


Figure 14 CHD mortality rate vs. cigarette consumption in 21 countries.

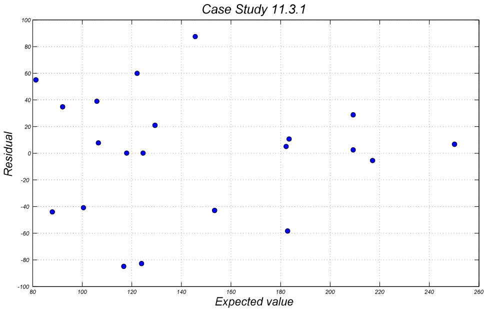


Figure 15 Residual plot for CHD mortality rate vs. cigarette consumption in 21 countries. There is a decreased spread with increased expected value.

This plot has a non-random pattern in the residuals, with increased spread at lower values of the residuals, as shown in Figure 15. This calls for a weighted least squares regression, which is a form of generalized least squares analysis. This is a topic beyond the scope of EEOS601 but covered next semester in EEOS611. One of the nice features about having a residual pattern like Figure 15 is that the parameter estimates are unbiased (Draper & Smith 1981). Unfortunately, the estimates of the statndard errors will be too large. This case study was an exercise in finding the significance of the slope parameter. Without the weighted regression the standard error for the regression parameter is obtained through Matlab's regress function:

[b,bint,r,rint,stats] = regress(Y,X,alpha);

The stats function provides the F statistic for the regression which contains in the following order, the R-square statistic, the F statistic, and p value for the full model, and an estimate of the error variance. In the text, the value of the t statistic is provided 4.64 and is found to be significant since it exceeds the 1-tailed critical value of 1.7291 (obtained in Matlab by tinv(0.95,19)). This F statistic provided by Matlab for the overall regression, 21.615, is the square of this *t* statistic for the regression slope and the *p* value for the full regression model, 1.75 x 10^{-4} is exactly that obtained from the p value for the t statistic for the regression coefficient ((1-tcdf(4.6492,19))*2). The standard error for the slope is 0.0129, shown as 46.708/3613.38 in the text on p. 688. The estimated slope and 95% CI for the slope with unweighted regression is 0.060 ± 0.027.

With generalized linear regression (performed with PASW/SPSS), the variance is weighted by {Variance of CHD}$^{-1.8}$, which downweights the contribution of the countries with lower cigarette consumption. The weighted regression estimate of the slope and 95% CI is 0.063 ± 0.023, which is quite similar to the unweighted regression's 0.060 ± 0.027. The standard error for the slope is 0.0108, which is 16% less than with the unweighted regression. The one-sided p value is 6 x 10^{-6}, which is lower than the 8.7 x 10^{-5} 1-tailed *p* value for the unweighted regression. The take-home message is that it is to your advantage to fit a weighted regression, but your estimated parameters will be close and your p values conservative if you use an unweighted regression.

Case Study 11.3.2

The expenses vs. sales regression is shown in Figure 16. The slope with 95% CI is obtained from Matlab's regress program and is shown in the figure and printed by the program:

```
The Y intercept = 25 with 95% CI = [-29 78]
The slope is 0.22 with 95% CI = [0.20 0.24]
R^2 = 99.1%;
F = 651.0 with p-value = 2.4e-007;
RMSE = s^2 = 138.8
P(t >= 25.5|Ho:slope = 0 & 6 df) = 2.4e-007
```

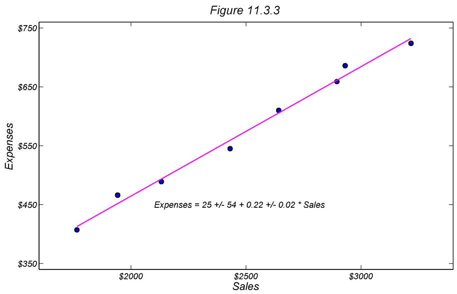


Figure 16 Expenses vs. sales regression for the Acme corporation.

In the output of the program, I used the term RMSE, which stands for root mean square error. This is the estimate of the variance around the regression line.

Example 11.3.2

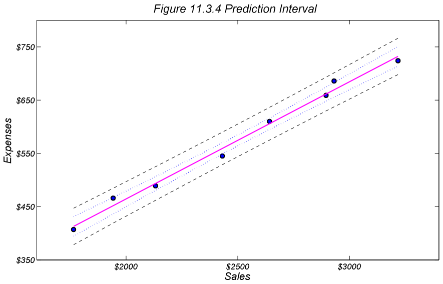


Figure 17 Figure 11.3.4 Larsen & Marx (2006) p 694-695

This example uses the Case Study 11.3.2 sales and expense data to introduce confidence intervals. There are two families of confidence intervals that can be used with regression. Figure 17 shows the two endpoints of the first class of confidence intervals based on the t multiplier. These are usually referred to as the 95% confidence interval for the mean and the 95% prediction interval. Both are based on t multipliers of the regression standard deviation. The tighter confidence limit is where for a given value of X, a mean based on a theoretically infinite sample size would be found in 95% of applications. The outer bound, called the prediction interval, would contain 95% of the observations drawn from a distribution having regression parameters equal to those estimated for the regression curve. If means at each value of x are based on just q replicates then a confidence interval specific for that sample size *q* should be used. The three standard errors are shown below. Equation 3.14 from Draper & Smith shows the standard error for samples based on very large n.

$$\mathrm{se}(\hat{Y}_0) = \mathrm{est.\,sd}(\hat{Y}_0) = s\left\{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right\}^{1/2}. \qquad (3.1.4)$$

The confidence limit for the prediction interval is shown in equation 3.1.6.

$$\hat{Y}_0 \pm t(\nu, 0.975)\left\{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right\}^{1/2} s, \qquad (3.1.6)$$

The following confidence limit is intermediate between these two intervals and is based on means calculated with q replicates:

$$\hat{Y}_0 \pm t(\nu, 0.975)\left[\frac{1}{q} + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right]^{1/2} s. \qquad (3.1.7)$$

There is another class of intervals based on Scheffé F multipliers called Hotelling-Working intervals, which are designed to contain 95% of the regression lines or to simultaneously provide coverage for estimates made simultaneously at different portions of the regression lines. For every interval based on the t statistic, there is an analogous Hotelling-Working interval obtained by replacing the t multiplier with a Scheffé F multiplier of the following form: $\{2\ F_{(2,n\text{-}2,1\text{-}\alpha)}\}^{1/2}$. This F multiplier, being slightly larger than the t multiplier, will produce broader confidence intervals. The Hotelling-Working intervals are not covered in Larsen & Marx (2006).

Example 11.3.4

This example, based on population genetics tests whether two slopes are the same. The data are graphed in Figure 18. I wrote a new m.file, **test2slopes**, to handle the problem of testing the slope of two different lines.

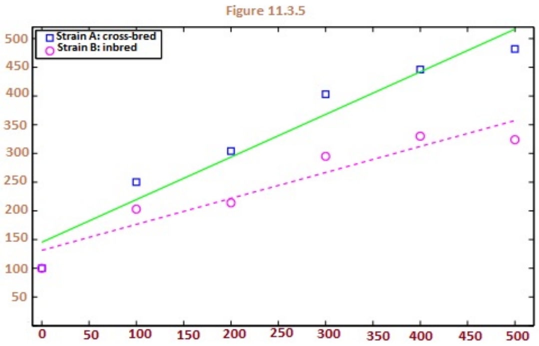


Figure 18 Figure 11.3.5 Larsen & Marx (2006) p 698

Annotated outline (with Matlab scripts) for Larsen & Marx Chapter 11

11 **Regression** (Week 11)

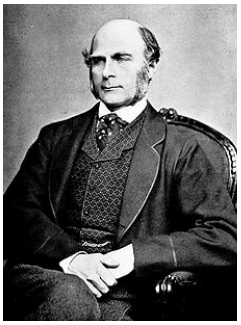


Figure 19
Francis Galton

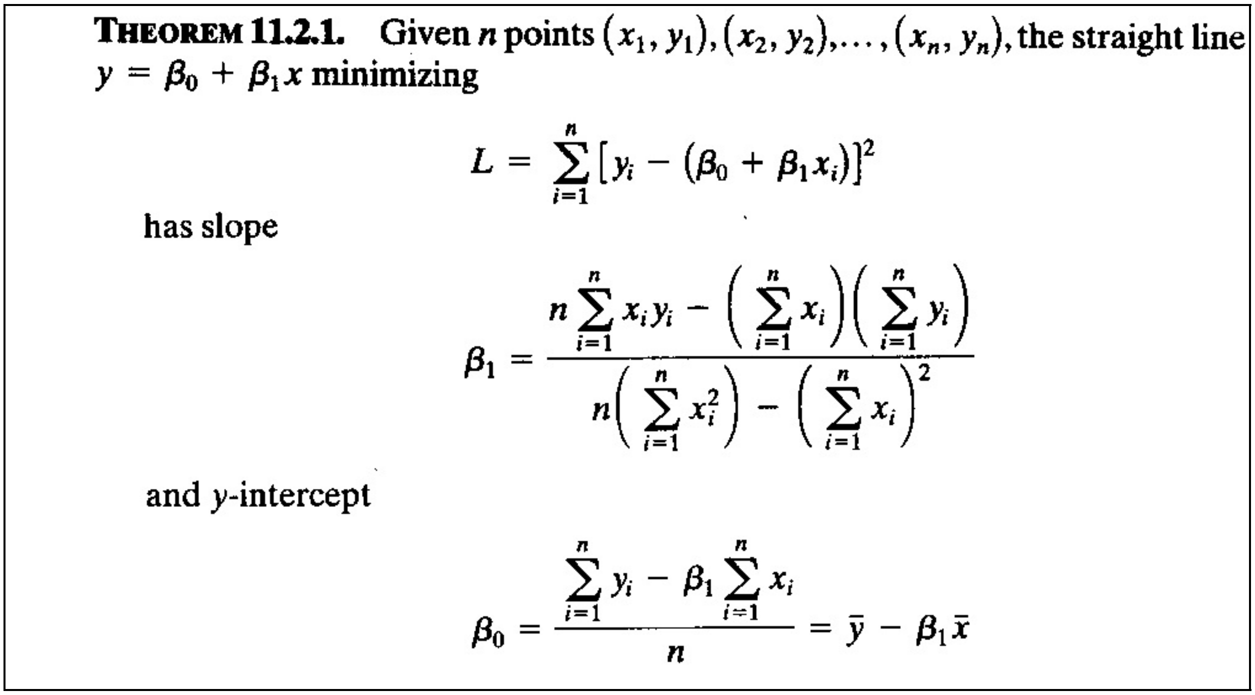
http://upload.wikimedia.org/wikipedia/commons/thumb/e/ec/Francis_Galton_1850s.jpg/225px-Francis_Galton_1850s.jpg

Francis Galton (1822-1911)

11.1 **INTRODUCTION**

11.2 **THE METHOD OF LEAST SQUARES**

Theorem 11.2.1



Theorem 11.2.1. Given n points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the straight line $y = \beta_0 + \beta_1 x$ minimizing

$$L = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

has slope

$$\beta_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and y-intercept

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - \beta_1 \bar{x}$$

Case Study 11.2.1

```
% LMcs110201_4th.m
% Case Study 11.2.1, page 648-650
% Larsen & Marx (2006) Introduction to Mathematical Statistics. 4th Edition
% Written 10/2010 by Eugene.Gallagher@umb.edu 2001, revised 2/18/11
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Regression of finished weight of rods (y) vs. rough casting weight (x)
DATA=[2.745 2.08
   2.7 2.045
   2.69 2.05
   2.68 2.005
   2.675 2.035
   2.670 2.035
   2.665 2.02
   2.66 2.005
   2.655 2.01
   2.655 2.0
   2.65 2
   2.65 2.005
   2.645 2.015
   2.635 1.99
   2.630 1.99
   2.625 1.995
   2.625 1.985
   2.62 1.97
   2.615 1.985
   2.615 1.99
   2.615 1.995
   2.61 1.99
   2.59 1.975
   2.59 1.995
   2.565 1.955];
[r,c]=size(DATA);
% To fit the Y intercept in Matlab, need to have a column vector of 1's:
X=[ones(r,1) DATA(:,1)]
Y=DATA(:,2);
B=X\Y;
   fprintf('\nThe Y intercept = %5.3f and slope = %5.3f\n',B(1),B(2));
   Yest=X*B;
   s=sprintf('y = %5.3f + %5.3f x',B);  % this writes a string variable
   %                              for plotting in the graph
   Resid=Y-Yest;
   % sort by the independent variable so the lines are smooth
   [Xs,k]=sort(X(:,2));
   % sort the estimated values using the index vector k:
   Ys=Yest(k);
   plot(DATA(:,1),DATA(:,2),'sg',Xs,Ys,'-r','LineWidth',2);
   xlabel('Rough Weight','FontSize',16),
   ylabel('Finished Weight','FontSize',16)
   text(2.5768,2.0309,s,'FontSize',16)
   title('Figure 11.2.1','FontSize',20)
   figure(gcf)
   pause
   % Not in text, but plot the residuals
   plot(X(:,2),Resid,'ok');
   xlabel('Rough Weight','FontSize',16)
   ylabel('Residuals','FontSize',16);
   title('Residuals for Figure 11.2.1','FontSize',20)
   figure(gcf)
```

11.2.1 Residuals

Definition 11.2.1 Let *a* and *b* be the least squares coefficients associated with the sample (x_1, y_1), (x_2, y_2), ... (x_n, y_n). For any value of *x*, the quantity $\hat{y}$= a + bx is known as the *predicted value* of y. For any given i, i=1, 2, ..., i, the difference y_i - $\hat{y}_i$= y_i - (a + bx_i) is called the ith **residual**. A graph of yi -$\hat{y}_i$ versus x_i for all *i* is called a **residual plot**.

11.2.2 Interpreting Residual Plots

Example 11.2.1 Residuals plotted as part of LMcs110201_4th.m

Case Study 11.2.2

```
% LMcs110202_4th.m
% Case Study 11.2.2
% Larsen & Marx (2006) Introduction to Mathematical Statistics. 4th Edition
% Written by Eugene.Gallagher@umb.edu 2001, revised 11/23/2010
% Regression of finished weight of rods (y) vs. rough casting weight (x)
Years=[5:5:30 32]';
Y=[17.1 29.6 63.6 117.1 186.4 246.5 285.1]';
% There is a typo in the 3rd and 4th editions. It is 246.5, not 346.5
X=[ones(length(Years),1) Years];
B=X\Y;
fprintf('\nThe Y instercept = %5.3f and slope = %5.3f\n',B(1),B(2));
Yest=X*B;
Resid=Y-Yest;
% sort by the independent variable so the lines are smooth
[Xs,k]=sort(X(:,2));
% sort the estimated values using the index vector k:
Ys=Yest(k);
plot(Years,Y,'+g',Xs,Ys,'-m');
xlabel('Years After 1960'),ylabel('Cost')
figure(gcf)
pause
plot(Years,Resid,'ok');axis([0 40 -25 35])
xlabel('Years After 1960'),ylabel('Residual');
figure(gcf)
```

Case Study 11.2.3

```
% LMcs110203_4th.m
% Case Study 11.2.3
% Larsen & Marx (2006) Introduction to Mathematical Statistics. 4th Edition
% Page 654-655
% Written by Eugene.Gallagher@umb.edu 2001, revised 11/23/2010
% Regression of finished weight of rods (y) vs. rough casting weight (x)
CAOPresent=[4 8 12.5 16 20 25 31 36 40 40]';
CAORecovered=[3.7 7.8 12.1 15.6 19.8 24.5 31.1 35.5 39.4 39.5]';
%
X=[ones(length(CAOPresent),1) CAOPresent];
Y=CAORecovered;
B=X\Y;
fprintf('\nThe Y instercept = %5.3f and slope = %5.3f\n',B(1),B(2));
Yest=X*B;
Resid=Y-Yest;
% sort by the independent variable so the lines are smooth
[Xs,k]=sort(X(:,2));
% sort the estimated values using the index vector k:
```

```
Ys=Yest(k);
chemist=[repmat('A',6,1);'B';repmat('A',3,1)];
i=find(chemist=='A');
j=find(chemist=='B');
plot(CAOPresent(i),Y(i),'ob',CAOPresent(j),Y(j),'sm',Xs,Ys,'-m');
xlabel('CaO Present'),ylabel('CaO Recovered')
figure(gcf)
pause
plot(CAOPresent(i),Resid(i),'ok',CAOPresent(j),Resid(j),'sm');
xlabel('CaO Present'),ylabel('Residual');
figure(gcf)
  x = CAOPresent(i);y=Resid(i);
     k = convhull(x,y);
     hold on,
     plot(x(k), y(k), '-r'), axis([0 45 -0.2 0.6]);hold off
```

Questions p. 556-662

11.2.3 Nonlinear Models

11.2.3.1 Exponential Regression

Case Study 11.2.4

```
% LMcs110204_4th.m
% Larsen & Marx Case Study 11.2.4, an example of exponential regression
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu; written 11/23/10; revised 11/23/10
DATA=[0 4.5e3;3 2.9e4;7 9e4;10 2.29e5;14 1.2e6;18 3.1e6;20 5.5e6];
[r,c]=size(DATA);
X=[ones(r,1) DATA(:,1)];Y=log(DATA(:,2));
B=X\Y;
fprintf('\nThe Y intercept = %8.3f and slope = %5.3f\n',B(1),B(2));
s=sprintf('y=%6.1f e^{%5.3f x}',exp(B(1)),B(2));
Yest=X*B;
Resid=Y-Yest;
% sort by the independent variable so the lines are smooth
[Xs,k]=sort(X(:,2));
% sort the estimated values using the index vector k:
Ys=Yest(k);
plot(X(:,2),Y,'ob',Xs,Ys,'-m');
xlabel('Years after 1975'),ylabel('Transistors per chip')
figure(gcf)
pause
% Generated a fine-scale estimated curve
xintrp=[0:0.1:20]';
[r,c]=size(xintrp);
xintrp=[ones(r,1) xintrp];
yestintrp=xintrp*B;
plot(DATA(:,1),DATA(:,2),'ob',xintrp(:,2),exp(yestintrp),'--r');
```

```
xlabel('Years after 1975'),ylabel('Transistors per chip')
axis([0 25 0 7e6])
text(11,3.5e6,s,'Fontsize',20);title('Figure 11.2.8')
figure(gcf)
pause
plot(Yest,Resid,'ok');
xlabel('Expected value'),ylabel('Residual');
grid;figure(gcf)

% Use the Matlab statistical toolbox's regress for confidence limits
[B,BINT] = regress(Y,X);
fprintf(...
'The doubling time for transistors per chip is %4.2f years\n',...
log(2)./B(2));
% Is Moore's Law's 2-year estimate within the 95% CI?
fprintf(...
'The 95%% confidence limits for the doubling time are: %3.2f and %3.2f years\n',...
fliplr(log(2)./BINT(2,:)));
```

11.2.4 Logarithmic regression

Case Study 11.2.5

```
% LMcs110205_4th.m
% Larsen & Marx Case Study 11.2.5, an example of logarithmic regression
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu; written 11/23/10; revised 2/19/11
DATA=[360 90;165 105;21 21;23 26;11 14;18 28;18 21;150 105;45 68;45 75;18 46];
[r,c]=size(DATA);
X=[ones(r,1) log10(DATA(:,1))];Y=log10(DATA(:,2));
B=X\Y;
fprintf('\nThe Y intercept = %8.3f and slope = %5.3f\n',B(1),B(2));
Yest=X*B;
Resid=Y-Yest;
% sort by the independent variable so the lines are smooth
[Xs,k]=sort(X(:,2));
% sort the estimated values using the index vector k:
Ys=Yest(k);
plot(X(:,2),Y,'o',Xs,Ys,'-m','LineWidth',2,...
                'MarkerEdgeColor','k',...
                'MarkerFaceColor','b',...
                'MarkerSize',10)
xlabel('Log_{10}(Locomotion begins (days))','FontSize',20),
ylabel('Log_{10}(Play begins (days))','FontSize',20)
title('Figure 11.2.9a','FontSize',22)
figure(gcf)
pause
% Generated a fine-scale estimated curve
```

```
xintrp=log10([10:0.1:370]');
[r,c]=size(xintrp);
xintrp=[ones(r,1) xintrp];
yestintrp=xintrp*B;
s=sprintf('y=%4.2f x^{%4.2f}',10^B(1),B(2))
plot(DATA(:,1),DATA(:,2),'o',10.^(xintrp(:,2)),10.^(yestintrp),'--r',...
   'LineWidth',2,...
                'MarkerEdgeColor','k',...
                'MarkerFaceColor','b',...
                'MarkerSize',8)
axis([0 425 0 160])
xlabel('Locomotion begins (days)','FontSize',20),
ylabel('Play begins (days)','FontSize',20)
title('Figure 11.2.9','FontSize',22)
text(200,140,s,'FontSize',20)
figure(gcf)
pause
loglog(log10(DATA(:,1)),log10(DATA(:,2)),'o',...
   xintrp(:,2),yestintrp,'--r',...
   'LineWidth',2,...
                'MarkerEdgeColor','k',...
                'MarkerFaceColor','b',...
                'MarkerSize',8)
axis(log10([10 150 10 400]))
xlabel('Locomotion begins (days)','FontSize',20),
ylabel('Play begins (days)','FontSize',20)
title('Figure 11.2.9b','FontSize',22)
text(200,140,s,'FontSize',20)
figure(gcf)
pause
plot(Yest,Resid,'ok');
xlabel('Expected value'),ylabel('Residual');
grid;figure(gcf)
```

11.2.5 **Logistic regression**

Case Study 11.2.6

```
% LMcs110206_4th.m
% Fitting a logistic regression
% Written 11/10 by Eugene.Gallagher@umb.edu
% Revised 2/20/2011.
DATA=[480 0.3;690 4.6;900 15.6;1100 33.4;1320 44.4;1530 45.7];
X=DATA(:,1);y=DATA(:,2)/100;
[b,dev,stats]=glmfit(X,y,'binomial')
% [yhat,dylo,dyhi] = glmval(b,X,'logit',stats)
yhat = glmval(b,X,'logit')
plot(X,y,'o',X,yhat,'-b','LineWidth',2);
```

```
xlabel('SAT Score');ylabel('Graduation Rate (%)')
figure(gcf);pause
% The fit looks really awful
% Regress ln (L-y/y) vs X as described in Larsen & Marx (2006) p 669
Y=log((48-DATA(:,2))./DATA(:,2));
[r,c]=size(Y);
X=[ones(r,1) DATA(:,1)];
B=X\Y;
fprintf('\nThe Y intercept = %8.3f and slope = %7.5f\n',B(1),B(2));
Yest=X*B;
Resid=Y-Yest;
% sort by the independent variable so the lines are smooth
[Xs,k]=sort(X(:,2));
% sort the estimated values using the index vector k:
Ys=Yest(k);
plot(X(:,2),Y,'o',Xs,Ys,'m','LineWidth',2,...
                'MarkerEdgeColor','k',...
                'MarkerFaceColor','b',...
                'MarkerSize',8)
xlabel('SAT','FontSize',20),ylabel('Graduation %','FontSize',20)
title('Figure 11.2.10a','FontSize',22);
figure(gcf)
pause
% Generated a fine-scale estimated curve
xintrp=[1:0.2:1600]';
[r,c]=size(xintrp);
xintrp=[ones(r,1) xintrp];
yestintrp=xintrp*B;
plot(DATA(:,1),DATA(:,2),'o',...
   xintrp(:,2),48./ (1+exp(B(1)+B(2).*xintrp(:,2) )),'--r',...
                 'LineWidth',2,...
                'MarkerEdgeColor','k',...
                'MarkerFaceColor','b',...
                'MarkerSize',8)
  xlabel('SAT','FontSize',20),ylabel('Graduation %','FontSize',20)
  title('Figure 11.2.10','FontSize',22)
axis([0 1650 0 55])
figure(gcf)
pause
plot(Yest,Resid,'ok');
xlabel('Expected value'),ylabel('Residual');
grid;figure(gcf)
```

11.2.6 Other Curvilinear models

Questions p. 671-676

11.3 THE LINEAR MODEL

Definition 11.3.1

Example 11.3.1

```
% LMex110301_4th.m
syms x y
int(sym('(x+y)/(x+1/2)'),y, 0,1)
```

11.3.1 A special case

11.3.2 Estimating the linear model parameters

Theorem 11.3.1. [maximum likelihood estimators of regression parameters are variables and are equal to the least squares estimates (numbers) if the assumptions are met]

11.3.3 Properties of linear model Estimators

Theorem 11.3.2 Properties of maximum likelihood estimators for linear model.

Theorem 11.3.3

Corollary

11.3.4 Estimating σ^2

Comment

11.3.5 Drawing inferences about β_1

Theorem 11.3.4 Testing β_1 with a t test

Theorem 11.3.5 Testing β_1 with a t test

Case Study 11.3.1 **Cigarette consumption**

```
% LMcs110301_4th.m
% Hypothesis testing with general linear model
% Larsen & Marx Case Study 11.3.1, Page 685-688 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu; written 11/23/10; revised 11/23/10
DATA=[3900 256.9;3350 211.6;3220 238.1;3220 211.8;2790 194.1;2780 124.5
    2770 187.3; 2290 110.5;2160 233.1;1890 150.3;1810 124.7;1800 41.2
    1770 182.1;1700 118.1;1680 31.9;1510 114.3;1500 144.9;1410 59.7
    1270 126.9;1200 43.9;1090 136.3];
[r,c]=size(DATA);
 X=[ones(r,1) DATA(:,1)];Y=DATA(:,2);
B=X\Y;
fprintf('\nThe Y intercept = %8.3f and slope = %5.3f\n',B(1),B(2));
Yest=X*B;
Resid=Y-Yest;
% sort by the independent variable so the lines are smooth
[Xs,k]=sort(X(:,2));
% sort the estimated values using the index vector k:
Ys=Yest(k);
plot(X(:,2),Y,'o',Xs,Ys,'m',...
                 'LineWidth',2,...
                'MarkerEdgeColor','k',...
                'MarkerFaceColor','b',...
                'MarkerSize',8)
```

```
xlabel('Cigarette consumption per adult per year','FontSize',20),
ylabel('CHD Mortality per 100,000 (ages 35-64)','FontSize',20)
title('Case Study 11.3.1','FontSize',22)
figure(gcf)
pause
plot(Yest,Resid,'o',...
                'MarkerEdgeColor','k',...
                'MarkerFaceColor','b',...
                'MarkerSize',8)
xlabel('Expected value','FontSize',20),ylabel('Residual','FontSize',20);
title('Case Study 11.3.1','FontSize',22)
grid;figure(gcf)
% Now use Matlab's regress
alpha=0.05;
[b,bint,r,rint,stats] = regress(Y,X,alpha);
% The help regress notation produces the following components of the
% stats variable:
% [B,BINT,R,RINT,STATS] = REGRESS(Y,X) returns a vector STATS containing, in
%    the following order, the R-square statistic, the F statistic and p value
%    for the full model, and an estimate of the error variance.
fprintf('R^2=%4.1f%%; F for overall regression =%4.1f with p-value=%7.2g;\n',...
   stats(1)*100,stats(2),stats(3));
df=length(Y)-2;
% Now, test whether the slope is different from 0. The F statistic
% produced by regress with 1 and df is the square of the t statistic
% in the t test for slope =1.
P=2*(1-tcdf(sqrt(stats(2)),df));
fprintf('t test for B_1=0 is %5.2f with % 3.0f df and p-value=%7.2g\n',...
   sqrt(stats(2)),df,P)
% Generate a 95% CI for s^2, also called Root Mean Square Error or RMSE:
% See Larsen & Marx page 691 for equation. Change alpha above for CI's
% other than 95% CI
L95CIRMSE=df*stats(4)/chi2inv(1-alpha/2,df);
U95CIRMSE=df*stats(4)/chi2inv(alpha/2,df);
fprintf('RMSE= %7.2f with %3.1f%% CI= [%7.2f
%7.2f]\n',stats(4),(1-alpha)*100,L95CIRMSE,...
   U95CIRMSE)
% Weighted least squares regression. PASW found weight of -1.8 power
[Beta,STDX,MSE,S] = lscov(X,Y,DATA(:,1).^1.8)




% Hypothesis testing with general linear model
% Larsen & Marx Case Study 11.3.1, Page 685-688 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu; written 11/23/10; revised 11/23/10
DATA=[3900 256.9;3350 211.6;3220 238.1;3220 211.8;2790 194.1;2780 124.5
    2770 187.3; 2290 110.5;2160 233.1;1890 150.3;1810 124.7;1800 41.2
    1770 182.1;1700 118.1;1680 31.9;1510 114.3;1500 144.9;1410 59.7
    1270 126.9;1200 43.9;1090 136.3];
[r,c]=size(DATA);
 X=[ones(r,1) DATA(:,1)];Y=DATA(:,2);
B=X\Y;
fprintf('\nThe Y intercept = %8.3f and slope = %5.3f\n',B(1),B(2));
Yest=X*B;
Resid=Y-Yest;
% sort by the independent variable so the lines are smooth
[Xs,k]=sort(X(:,2));
% sort the estimated values using the index vector k:
Ys=Yest(k);
plot(X(:,2),Y,'ob',Xs,Ys,'-m');
xlabel('Cigarette consumption per adult per year'),
ylabel('CHD Mortality per 100,000 (ages 35-64)')
figure(gcf)
pause
plot(Yest,Resid,'ok');
xlabel('Expected value'),ylabel('Residual');
grid;figure(gcf)
% Now use Matlab's regress
alpha=0.05;
[b,bint,r,rint,stats] = regress(Y,X,alpha);
% The help regress notation produces the following components of the
% stats variable:
% [B,BINT,R,RINT,STATS] = REGRESS(Y,X) returns a vector STATS containing, in
%    the following order, the R-square statistic, the F statistic and p value
%    for the full model, and an estimate of the error variance.
fprintf('R^2=%4.1f%%; F for overall regression =%4.1f with p-value=%7.2g;\n',...
   stats(1)*100,stats(2),stats(3));
df=length(Y)-2;
% Now, test whether the slope is different from 0. The F statistic
% produced by regress with 1 and df is the square of the t statistic
% in the t test for slope =1.
P=2*(1-tcdf(sqrt(stats(2)),df));
fprintf('t test for B_1=0 is %5.2f with % 3.0f df and p-value=%7.2g\n',...
   sqrt(stats(2)),df,P)
% Generate a 95% CI for s^2, also called Root Mean Square Error or RMSE:
% See Larsen & Marx page 691 for equation. Change alpha above for CI's
% other than 95% CI
L95CIRMSE=df*stats(4)/chi2inv(1-alpha/2,df);
```

```
U95CIRMSE=df*stats(4)/chi2inv(alpha/2,df);
fprintf('RMSE= %7.2f with %3.1f%% CI= [%7.2f
%7.2f]\n',stats(4),(1-alpha)*100,L95CIRMSE,...
   U95CIRMSE)
```

Theorem 11.3.6

Case Study 11.3.2

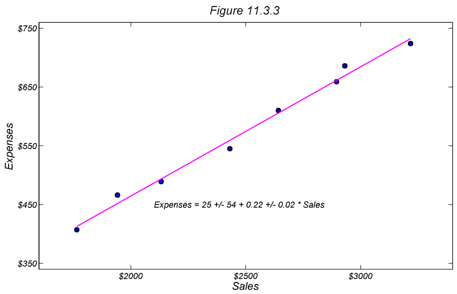


Figure 20 Figure 11.3.3

```
% LMcs110302_4th.m
% Hypothesis testing with general linear model
% Larsen & Marx Case Study 11.3.2
% Larsen & Marx (2006) Introduction to
Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu; written
11/23/10; revised 11/23/10
DATA=...
[1765 407;1942 466;2132 489;2431 545;2642
610;2895 659;2931 686;3217 724];
[r,c]=size(DATA);df=r-2;
 X=[ones(r,1) DATA(:,1)];Y=DATA(:,2);
B=X\Y;
fprintf('\nThe Y intercept = %8.3f and slope = %5.3f\n',B(1),B(2));
Yest=X*B;
Resid=Y-Yest;
% sort by the independent variable so the lines are smooth
[Xs,k]=sort(X(:,2));
% sort the estimated values using the index vector k:
Ys=Yest(k);
% Now use Matlab's regress to produce the same regression coefficients:
[b,bint,r,rint,stats] = regress(Y,X);
% Note for other problems, more significant figures will be needed for
% reporting the regression coefficients and 95%% CI's
fprintf('\nThe Y intercept = %2.0f with 95%% CI = [%2.0f %2.0f]\n',...
   b(1),bint(1,1),bint(1,2));
fprintf('The slope is %4.2f with 95%% CI = [%4.2f %4.2f]\n',...
   b(2),bint(2,1),bint(2,2));
fprintf(...
'R^2 = %4.1f%%; F = %5.1f with p-value = %7.2g; RMSE = s^2 = %4.1f\n',...
   stats(1)*100,stats(2),stats(3),stats(4))
P=2*(1-tcdf(sqrt(stats(2)),df));
fprintf('P(t >= %4.1f |Ho: slope = 0 & %3.0f df) = %7.2g\n',...
   sqrt(stats(2)),df,P)
s=sprintf('Expenses = %2.0f +/- %2.0f + %4.2f +/- %4.2f * Sales',...
   b(1),bint(1,2)-b(1),b(2),bint(2,2)-b(2));
plot(X(:,2),Y,'o',Xs,Ys,'m','LineWidth',2,...
                 'MarkerEdgeColor','k',...
                 'MarkerFaceColor','b',...
```

```
                    'MarkerSize',8)
axis([1600 3400 340 760]);
ax1=gca;
xlabel('Sales','Color','black','FontSize',20)
ylabel('Expenses','Color','black','FontSize',20)
text(2100,450,s,'FontSize',16)
 set(ax1,'xtick',2000:500:3000,'FontSize',16)
 set(ax1,'xticklabel',{'$2000', '$2500','$3000'},'FontSize',16)
 set(ax1,'ytick',350:100:750,'FontSize',16)
  set(ax1,'yticklabel',{'$350', '$450','$550','$650','$750'},...
      'FontSize',16)
  title('Figure 11.3.3','FontSize',22)
figure(gcf)
pause
hold off
plot(Yest,Resid,'ok');
xlabel('Expected value','FontSize',20),ylabel('Residual','FontSize',20);
grid;figure(gcf)
```

11.3.6 Drawing inferences about β_o

11.3.7 Drawing inferences about σ^2

Questions p. 691-693

11.3.8 Drawing inferences about E(Y | x)

Theorem 11.3.7

Example 11.3.2

```
% LMex110302_4th.m
% Example 11.3.2 pp. 692 in
% Larsen & Marx (2006)Introduction to Mathematical Statisticcs, 4th edition
% Written by Eugene D. Gallagher, Eugene.Gallagher@umb.edu
% Written November 2001, updated 1/20/2011
% 95% Confidence intervals
if exist('LMex110302_4th.out')==2;delete LMex110302_4th.out;end
diary LMex110302_4th.out
DATA=...
[1765 407;1942 466;2132 489;2431 545;2642 610;2895 659;2931 686;3217 724];
[r,c]=size(DATA);df=r-2;
 X=DATA(:,1);Y=DATA(:,2);
x=[ones(r,1) X];B=x\Y;
fprintf('\nThe Y intercept = %8.3f and slope = %5.3f\n',B(1),B(2));
Yest=x*B;
Resid=Y-Yest;
% sort by the independent variable so the lines are smooth
[Xs,k]=sort(X);
% sort the estimated values using the index vector k:
Ys=Yest(k);
```

```
% Now use Matlab's regress to produce the same regression coefficients:
[b,bint,r,rint,stats] = regress(Y,x);
% Note for other problems, more significant figures will be needed for
% reporting the regression coefficients and 95%% CI's
fprintf('\nThe Y intercept = %2.0f with 95%% CI = [%2.0f %2.0f]\n',...
   b(1),bint(1,1),bint(1,2));
fprintf('The slope is %4.2f with 95%% CI = [%4.2f %4.2f]\n',...
   b(2),bint(2,1),bint(2,2));
fprintf(...
'R^2 = %4.1f%%; F = %5.1f with p-value = %7.2g; RMSE = s^2 = %4.1f\n',...
   stats(1)*100,stats(2),stats(3),stats(4))
P=2*(1-tcdf(sqrt(stats(2)),df));
fprintf('P(t >= %4.1f |Ho: slope = 0 & %3.0f df) = %7.2g\n',...
   sqrt(stats(2)),df,P)
s=sprintf('Expenses = %2.0f +/- %2.0f + %4.2f +/- %4.2f * Sales',...
   b(1),bint(1,2)-b(1),b(2),bint(2,2)-b(2));
plot(X,Y,'o',Xs,Ys,'m','LineWidth',2,...
                   'MarkerEdgeColor','k',...
                   'MarkerFaceColor','b',...
                   'MarkerSize',8)
axis([1600 3400 340 760]);
ax1=gca;
xlabel('Sales','Color','black','FontSize',20)
ylabel('Expenses','Color','black','FontSize',20)
title('Figure 11.3.3','FontSize',22)
% text(2100,450,s,'FontSize',16)
set(ax1,'xtick',2000:500:3000,'FontSize',16)
set(ax1,'xticklabel',{'$2000', '$2500','$3000'},'FontSize',16)
set(ax1,'ytick',350:100:750,'FontSize',16)
set(ax1,'yticklabel',{'$350', '$450','$550','$650','$750'},...
     'FontSize',16)
figure(gcf)
pause

% This code based on Draper & Smith
disp('Example of least-squares regression from')
disp('Draper & Smith (1981, p. 9, p. 617) Applied Regression')
disp('Analysis, Jonh Wiley & Sons')
disp('Uses Gallagher''s leastsqu.m program')
[b,Yest,resid,SS,F,R2,Vb,VYest]=leastsqu(X,Y);
disp(' ');disp(' ');
disp('ANOVA TABLE');disp(' ');disp('     SS         df        MS');
disp(' ');
disp(SS)
disp('F statistics & p values');disp(' ');
disp('  MS-num    MS-den   df-num    df-den      F        prob');
disp(' ');
disp(F)
ResMS=F(1,2);
[r,c]=size(F);
if r>1
 LFMS=F(2,1);
 if F(2,6)>0.05
   s2=ResMS;
   fprintf('Lack of fit F-test was non-significant, p=%6.4g\n',F(2,6))
   fprintf('Error variances were pooled to estimate s^2=%7.5g\n\n',s2);
 else
    PureErMS=F(2,2);
    fprintf('Lack of fit was significant, p=%6.4g\n',F(2,6))
    s2=PureErMS;
    fprintf('Pure error MS (=%6.4g) used to estimate s^2.\n\n',s2);
  end
else
 s2=ResMS;
end;
s=sqrt(s2);
fprintf('The probability that the slope=0 is %8.4g\n',F(1,6))
% now calculate upper & lower 95% confindence intervals
alpha=0.05;
[r,c]=size(F);
if r>1 & F(r,6)<0.05
  errordf=F(r,4);
else
  errordf=F(1,4);
end
% Studentt=tfind(errordf,alpha);  % Unfortunately, this is slow
% Studentt=qt(1-alpha/2,errordf); % qt.m from stixbox toolbox
Studentt=tinv(1-alpha/2,errordf);
disp(' ');disp('Regression equation:');disp(' ');
if length(b)>1
     fprintf('Y = %7.5g + %7.5g * X\n\n',b(1),b(2))
else
     fprintf('Y = %7.5g * X\n\n',b(1))
end
fprintf('The regression equation has R-squared=%3.1f%%\n\n',R2*100)
fprintf('The slope=%7.5g, with standard error=%7.5g\n',b(2),sqrt(Vb(2,2)))
disp('The 95% confidence limits for the slope are:');
Int=Studentt*sqrt(Vb(2,2));disp([b(2)-Int b(2)+Int]);disp(' ')
fprintf('The 95%% confidence limits for y-intercept=%7.5g are:\n',b(1));
Int=Studentt*sqrt(Vb(1,1));disp([b(1)-Int b(1)+Int])
```

```
UL=Yest+Studentt*sqrt(VYest);
LL=Yest-Studentt*sqrt(VYest);
% sort by the independent variable so the lines are smooth
[Xs,k]=sort(X);
Ys=Yest(k);
UL=UL(k);
LL=LL(k);
diary off;
yl=sprintf('Y & 95%% confidence limits');
lt=sprintf('Larsen & Marx (2006) Example 11.3.2');
plot(X,Y,'o',Xs,Ys,'m',Xs,LL,':b',Xs,UL,':b','LineWidth',2,...
                'MarkerEdgeColor','k',...
                'MarkerFaceColor','b',...
                'MarkerSize',8)

set(ax1,'xtick',2000:500:3000,'FontSize',16)
set(ax1,'xticklabel',{'$2000', '$2500','$3000'},'FontSize',16)
set(ax1,'ytick',350:100:750,'FontSize',16)
set(ax1,'yticklabel',{'$350', '$450','$550','$650','$750'},...
     'FontSize',16)
xlabel('Sales','Color','black','FontSize',20)
ylabel('Expenses','Color','black','FontSize',20)
title('Figure 11.3.4 Mean Confidence Interval','FontSize',22)
figure(gcf)
pause
hold off
clc
fprintf('The 95%% confidence interval in the previous graph was for the\n')
fprintf('expected mean(Y) at each value of X.  It doesn''t have to\n')
fprintf('include the individual data points (see Draper & Smith, p. 30).\n\n')
fprintf('We can calculate the broader 95%% confidence interval to predict\n')
fprintf('where a single new observation at a given X might fall\n')
fprintf('95%% of the time.\n\n')
[rw,cl]=size(VYest);
newVYest=VYest+(ones(rw,cl)*s2);          % See eqaution 1.4.11
newUL=Yest+Studentt*sqrt(newVYest);
newLL=Yest-Studentt*sqrt(newVYest);
newUL=newUL(k);
newLL=newLL(k);
plot(Xs,newLL,'--k',Xs,newUL,'--k');
hold on
plot(X,Y,'o',Xs,Ys,'m',Xs,LL,':b',Xs,UL,':b','LineWidth',2,...
                'MarkerEdgeColor','k',...
                'MarkerFaceColor','b',...
                'MarkerSize',8)
```

```
set(ax1,'xtick',2000:500:3000,'FontSize',16)
set(ax1,'xticklabel',{'$2000', '$2500','$3000'},'FontSize',16)
set(ax1,'ytick',350:100:750,'FontSize',16)
set(ax1,'yticklabel',{'$350', '$450','$550','$650','$750'},...
    'FontSize',16)
xlabel('Sales','Color','black','FontSize',20)
ylabel('Expenses','Color','black','FontSize',20)
title('Figure 11.3.4 Prediction Interval','FontSize',22)
pause
hold off
clc

diary on;
disp('*************** Prediction limits for Y estimates *************')
disp(' ')
fprintf('We can use Equation 1.4.13 in Draper & Smith or Box 14.4 in\n')
fprintf('Sokal & Rolf''s Biometry to predict where the mean of q new\n')
fprintf('new observations might lie 95%% of the time.\n\n')
fprintf('These confidence intervals play a key role in inverse regression\n')
fprintf('(covered later).\n\n\n')
diary off
fprintf('Let''s calculate the 95%% confidence interval\n')
fprintf('for mean (Yest) based on 3 replicates (q=3) at each X.\n')
fprintf('If the regression model is correct, these intervals will cover\n')
fprintf('the true mean (Y) at each X in 95%% of cases over the long run.\n')
replicates=3;
newVYest=VYest+(ones(rw,cl)*s2/replicates);
% Kendall & Stuart Vol. 2, 4th ed. section 28.23 discuss this equation.
newUL=Yest+Studentt*sqrt(newVYest);newLL=Yest-Studentt*sqrt(newVYest);
newUL=newUL(k);
newLL=newLL(k);
plot(Xs,newLL,'--m',Xs,newUL,'--m',X,Y,'o',Xs,Ys,'r',...
   Xs,LL,':b',Xs,UL,':b','LineWidth',2,...
                'MarkerEdgeColor','k',...
                'MarkerFaceColor','b',...
                'MarkerSize',8)
set(ax1,'xtick',2000:500:3000,'FontSize',16)
set(ax1,'xticklabel',{'$2000', '$2500','$3000'},'FontSize',16)
set(ax1,'ytick',350:100:750,'FontSize',16)
set(ax1,'yticklabel',{'$350', '$450','$550','$650','$750'},...
    'FontSize',16)
xlabel('Sales','Color','black','FontSize',20)
ylabel('Expenses','Color','black','FontSize',20)
title('Figure 11.3.4, n=3','FontSize',22)
pause
```

```
hold off
diary off
```

11.3.9 Drawing inferences about future observations

Theorem 11.3.8

Example 11.3.3

Not programmed yet

11.3.10 **Testing the equality of two slopes**

Theorem 11.3.9

Example 11.3.4

```
% LMex110304_4th.m
% LMex110304_4th.m
% Testing the equality of two slopes
% Using Larsen & Marx Theorem 11.3.9
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu 11/27/10
X=[0:100:500]';
A=[100 250 304 403 446 482]';
B=[100 203 214 295 330 324]';
[rA,cA]=size(A);
BA=[ones(rA,1) X]\A;
Aest=[ones(rA,1) X]*BA;
BB=[ones(rA,1) X]\B;
Best=[ones(rA,1) X]*BB;
[rB,cB]=size(B);
plot(X,A,'s',X,B,'om',X,Aest,'g',...
   X,Best,'--m','LineWidth',2,...
                 'MarkerSize',12)
axis([-10 520 0 520]);
legend('Strain A: cross-bred', 'Strain B: inbred','Location','NorthWest')
xlabel('Day number','FontSize',16);
ylabel('Census count','FontSize',16)
title('Figure 11.3.5','FontSize',20)
figure(gcf)
pause
[T,pvalue,df,S,SE]=test2slopes(X,A,X,B);
CI95diffU=(0.742-0.452)+tinv(0.975,8)*S*sqrt(1/(2*(1.75e5)))
CI95diffU=(0.742-0.452)+tinv(0.975,8)*SE
CI95diffL=(0.742-0.452)-tinv(0.975,8)*S*sqrt(1/(2*(1.75e5)))
CI95diffL=(0.742-0.452)-tinv(0.975,8)*SE
half95CI=tinv(0.975,8)*S*sqrt(1/(2*(1.75e5)))
half95CI=tinv(0.975,8)*SE
fprintf('\nThe difference in slopes = %4.2g, with 95%%CI = [ %4.2g %4.2g ]\n',...
   BA(2)-BB(2),CI95diffL, CI95diffU)
```

```
fprintf('The pooled variance for the difference in slopes is %5.2f\n',S)
fprintf('Student''s t=%5.2f {%2.0f df} ',T,rA+rB-4)
fprintf('with 1-tailed p=%3.2g\n',pvalue/2)
```

```
function [T,pvalue,df,S,SE]=test2slopes(X1,A,X2,B,plot)
%  Format [T,pvalue,df,S,SE]=test2slopes(X1,A,X2,B,plot)
% Testing the equality of two slopes
% Using Larsen & Marx Theorem 11.3.9
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Input X1 & X2 Explanatory variables, no leading 1's, single column
%       A & B  Two response variables
%       plot=1 Scatterplot of data
% Output T=Student's t
%        pvalue= 2-tailed p value;
%        S=variance
%        df=degrees of freedom
%        Standard error for the difference in slopes
% Written by Eugene.Gallagher@umb.edu 11/27/10
[rA,cA]=size(X1);
[rB,cB]=size(X2);
if nargin>4 & plot==1
   scatter(X1,A,'b','*')
   hold on
   plot(X2,B,'mo')
   lsline
   figure(gcf)
   hold off
end
x1=[ones(rA,1) X1];
x2=[ones(rB,1) X2];
BA=regress(A,x1);
BB=regress(B,x2);
YestA=x1*BA;YestB=x2*BB;
ResidA=A-YestA;ResidB=B-YestB;
df=rA+rB-4;
S=sqrt((sum(ResidA.^2)+sum(ResidB.^2))/df);
Xdev2A=sum((X1-mean(X1)).^2);Xdev2B=sum((X2-mean(X2)).^2);
SE=S*sqrt(1/Xdev2A + 1/Xdev2B);
T=(BA(2)-BB(2))/(S*sqrt(1/Xdev2A + 1/Xdev2B));
if T<0
   pvalue=2*tcdf(T,df);
else
   pvalue=2*(1-tcdf(T,df));
end
```

Questions p 699-702

11.4 COVARIANCE & CORRELATION

11.4.1 Measuring the dependence between two random variables

Definition 11.4.1

Theorem 11.4.1

Example 11.4.1

```
% LMex110401_4th.m
% An interesting application of Matlab's symbolic math toolbox.
% Page 703 in:
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu, Revised 11/28/10
syms x y yint mux muy; int('8*x*y',y,0,x)
int('8*x*y',x,y,1)
mux=int('x * 4*x^3',x,0,1)
muy=int('y*(-4*y*(y^2 - 1))',y,0,1)
int('x*y*8*x*y',y,0,x)
int(int('x*y*8*x*y',y,0,x),x,0,1)-32/75
```

11.4.2 The relationship between the covariance and independence

Example 11.4.2

Not programmed

Questions p. 704-705

11.4.3 The role of the covariance in the variance of a sum (propagation of error)

Theorem 11.4.3

$$\mathrm{Var}(S) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{j<k} \mathrm{Cov}(X_j, X_k)$$

Example 11.4.3

Not programmed

11.4.4 The correlation coefficient

Definition 11.4.2

Theorem 11.4.4

Questions p. 707-708

11.4.5 Estimating ρ(X,Y): The sample correlation coefficient

Questions p. 709

11.4.6 Interpreting R

Case Study 11.4.1 SAT’s & Participation Rate

%LMcs110401_4th.m

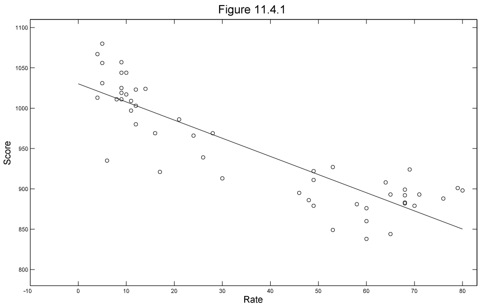


Figure 21 Figure 11.4.1 Larsen & Marx (2006) p 712

```
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Page 712
% Written by Eugene.Gallagher@umb.edu
X=[49 8 26 6 46 28 80 68 53 49 65 58 16 14 60 5 10 11 9 68 64 79 11 9 4 ...
 10 21 9 30 69 71 12 76 60 5 24 9 53 70 68 60 5 12 48 4 68 65 49 17 9 12]';
Y=[911 1011 939 935 895 969 898 892 849 879 844 881 969 1024 876 1080 ...
 1044 997 1011 883 908 901 1009 1057 1013 1017 986 1025 913 924 893 ...
 1003 888 860 1056 966 1019 927 879 882 838 1031 1023 886 1067 899 ...
 893 922 921 1044 980]';
plot(X,Y,'ok');xlabel('Rate');ylabel('Score')
lsline
axis([-10 83 780 1110])
figure(gcf)
corrcoef(X,Y)
[r,c]=size(X)
[B,BINT,R,RINT,STATS] = regress(Y,[ones(r,1) X]);
fprintf('\nThe R^2 = %5.1f%%\n',STATS(1)*100)
```

Case Study 11.4.2

```
5%LMcs110402_4th.m
% Larsen & Marx (2006) Introduction to
Mathematical Statistics, 4th edition
% Page 712
% Written by Eugene.Gallagher@umb.edu
X=[3508.9 5283.1 7557.1 11032.8 16009.5 24134.4
33785.7 44e3]';
Y=[12400 13500 13900 14300 14600 15300 16200
18000]';
plot(X,Y,'ok');xlabel('CellSubs','FontSize',16);
ylabel('Law$Rev','FontSize',16),Title('Figure 11.4.2','FontSize',20)
lsline
axis([-10 46000 11700 19100])
figure(gcf)
corrcoef(X,Y)
[r,c]=size(X)
[B,BINT,R,RINT,STATS] = regress(Y,[ones(r,1) X]);
fprintf('\nThe R^2 = %5.1f%%\n',STATS(1)*100)
```

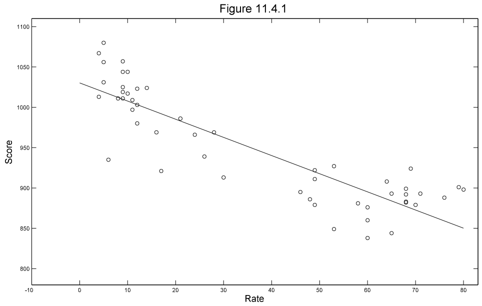


Figure 22 Figure 11.4.2 Larsen & Marx (2006) p 713

Questions p 714-716

11.5 THE BIVARIATE NORMAL DISTRIBUTION

11.5.1 Generalizing the univariate normal pdf

Definition 11.5.1

Comment

11.5.2 Properties of the bivariate normal distribution

Theorem 11.5.1

Comment Regression to mediocrity

Questions p. 720-721

11.5.3 Estimating parameters in the bivariate normal pdf

Theorem 11.5.2

11.5.4 Testing $H_o: \rho(X,Y)=0$

Theorem 11.5.3

Example 11.5.1 Butterfat vs. Temperature

```
% LMex110501_4th.m
% Page 722-723 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu 11/28/10; revised 12/12/10
% Butterfat content vs. temperature
X=[64 65 65 64 61 55 39 41 46 59 56 56 62 37 37 45 57 58 60 55]';
Y=[4.65 4.58 4.67 4.60 4.83 4.55 5.14 4.71 4.69 4.65 4.36 4.82 4.65 ...
  4.66 4.95 4.60 4.68 4.65 4.60 4.46]';
plot(X,Y,'ok');axis([50 70 4.3 4.9]), xlabel('Temperature')
ylabel('Butterfat Content');figure(gcf)
C=corrcoef(X,Y);
r=C(1,2);
[n,c]=size(X);
t=sqrt(n-2)*r/sqrt(1-r^2);
if t<0
   p=2*tcdf(t,n-2);
else
   p=2*(1-tcdf(t,n-2));
end
fprintf('\nPearson"s r=%4.2f, 2-tailed P(t>|%4.2f| |Ho)= %6.3g\n',C(1,2),t,p)
% Fisher's approach
z=1/2*log((1+r)/(1-r))/sqrt(1/(n-3));
if z<0
   P= 2*normcdf(z);
else
   p=2*(1-normcdf(z));
end
fprintf('\nThe two-tailed Fisher"s approximate p = % 6.3g\n',P)

[r,pval] = corr(X,Y,'type','Pearson','tail','both');
fprintf('\nPearson"s r=%4.2f with 2-tailed p = %6.3g\n',r,pval)
[rho,pvalrho] = corr(X,Y,'type','Spearman','tail','both');
fprintf('\nSpearman"s rho is %4.2f with 2-tailed p=%4.3f\n',rho, pvalrho)
[tau,pvaltau] = corr (X,Y,'type','Kendall','tail','both');
fprintf('\nKendall"s tau is %4.2f with 2-tailed p=%4.3f\n',tau, pvaltau);

% Call's Gallagher's Kendall's tau
[TAU,PROB,S,Z,E]=kendall(X,Y);
fprintf('\n Kendall"s tau is %4.2f with 2-tailed p=%4.3f\n',TAU, 2*PROB)
```

Questions p. 724

11.6 TAKING A SECOND LOOK AT STATISTICS (HOW NOT TO INTERPRET THE SAMPLE CORRELATION COEFFICIENT)

Appendix 11.A.1: Minitab applications

Appendix 11.A.2 A proof of theorem 11.3.3

References

Draper, N. R. and H. Smith. 1998. Applied Regression Analysis, 3rd Edition. John Wiley & Sons, New York. 706 p, with data diskette.*[This is a major revision of the tremendous 2nd edition, now with a greatly enhanced treatment of regression diagnostics. Includes a superb discussion of inverse regression]* {**5**}

Larsen, R. J. and M. L. Marx. 2006. An introduction to mathematical statistics and its applications, 4th edition. Prentice Hall, Upper Saddle River, NJ. 920 pp. {**5**, **9**}

Searle, S. R. 1982. Matrix algebra useful for statistics. John Wiley & Sons, New York. 438 pp.{**5**}

Index