prepared by Professor Jenny Baglivo

4 MATH448001 Notebook 4 **3**

4.1 Simple Linear Model . . . 3

4.1.1 Least Squares Estimation . . . 3

4.1.2 Equivalent Model, Predicted Responses, Residuals . . . 7

4.1.3 Partitioning Sum of Squared Deviations . . . 8

4.1.4 Coefficient of Determination . . . 9

4.1.5 Relationship Between Least Squares Estimator of Slope and Correlation 10

4.1.6 Gauss-Markov Theorem . . . 12

4.1.7 Permutation Confidence Interval Procedure for the Slope Parameter . . 13

4.1.8 Diagnostics: Residual, Leverage, Influence . . . 16

4.2 Simple Linear Regression . . . 21

4.2.1 Least Squares Estimation . . . 21

4.2.2 Diagnostic Plots . . . 22

4.2.3 Analysis of Variance . . . 23

4.2.4 Inference for Slope and Intercept Parameters . . . 25

4.2.5 Inference for Mean Response Parameters . . . 26

4.2.6 *Mathematica* Example: Rod Weights Study . . . 27

4.3 Linear Model . . . 27

4.4 Linear Regression . . . 28

4.4.1 Least Squares Estimation . . . 28

4.4.2 Analysis of Variance . . . 32

4.4.3 Coefficient of Determination . . . 34

4.4.4 Inference for β Parameters . . . 35

4.4.5 Inference for Mean Response Parameters 36
4.4.6 *Mathematica* Example: Birthweight Study 37
4.4.7 *Mathematica* Example: Timber Yield Study 38
4.5 Additional Topics 39
4.5.1 Unconditional Least Squares Analysis and the Bootstrap 39
4.5.2 Locally Weighted Least Squares and the Bootstrap 41
4.6 Linear Algebra Review 45
4.6.1 Linear Algebra: Solving Linear Systems 45
4.6.2 Linear Algebra: Coefficient Matrices 46
4.6.3 Linear Algebra: Inner Product and Orthogonality 47
4.6.4 Linear Algebra: Orthogonal Decomposition Theorem 48
4.6.5 Linear Algebra: Fundamental Theorem of Linear Algebra 48
4.6.6 Linear Algebra: Orthogonal Projections and Least Squares Analysis . . 49

4 MATH448001 Notebook 4

This notebook introduces methods for analyzing linear models using Legendre's method of least squares. Classical methods assume that error distributions are normally distributed, and are carried out conditional on the observed values of the predictor variables. The notes include topics from Chapter 14 (linear least squares) of the Rice textbook.

4.1 Simple Linear Model

Let Y be a *response* variable. A *simple linear model* is a model of the form

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

where X (the *predictor*) and ϵ (the *measurement error*) are independent random variables and the distribution of ϵ has mean 0 and standard deviation σ.

The objective is to estimate the parameters in the conditional mean formula

$$E(Y|X = x) = \alpha + \beta x$$

using a list of paired observations:

$$(x_1, y_1),\ (x_2, y_2)\ , \ldots ,\ (x_N, y_N).$$

The observed pairs are assumed to be either

- the values of a random sample from the joint (X, Y) distribution, or
- a collection of independent responses made at pre-determined levels of the predictor.

  A typical example is a dose-response study, where a certain number of subjects are assigned to each of several dosage levels.

Analysis is done conditional on the observed values of the predictor variable.

4.1.1 Least Squares Estimation

The data are assumed to be the values of N independent random variables,

$$Y_i = \alpha + \beta x_i + \epsilon_i \text{ for } i = 1, 2, \ldots, N, \text{ where}$$

1. The means of the random variables are $E(Y_i) = \alpha + \beta x_i$,
2. The collection of error terms $\{\epsilon_i\}$ is a random sample from a distribution with mean 0 and standard deviation σ, and
3. All parameters (α, β, and σ) are assumed to be unknown.

Least squares is a general estimation method introduced by Legendre in the early 1800's. In the simple linear case, the *least squares (LS)* estimators of α and β are obtained by minimizing the following sum of squared deviations of observed from expected responses:

$$S(\alpha, \beta) = \sum_{i=1}^{N} (Y_i - (\alpha + \beta x_i))^2 .$$

For a given set of data, multivariable calculus can be used to find estimates for α and β.

For example, consider the following 8 data pairs.

x	10.98	18.50	14.36	11.51	13.48	11.44	17.32	16.00
y	38.71	60.98	46.31	46.37	42.10	36.47	52.34	49.53

The observed sum of squared deviations is the following quadratic function:

$$S(\alpha, \beta) = 8\alpha^2 + 227.18\beta\alpha - 745.62\alpha + 1670.07\beta^2 - 10871.3\beta + 17807.$$

The graph of $z = S(\alpha, \beta)$ is an elliptic paraboloid with a unique minimum at $(11.3233, 2.4846)$.

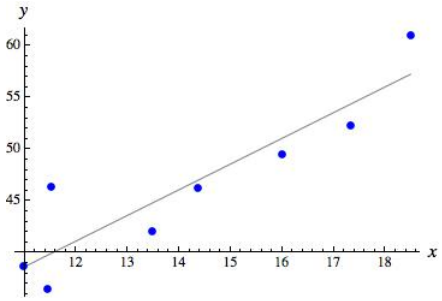
Numerical and graphical summaries for these data are given below:

X Mean, Standard deviation: $\overline{x} = 14.20,\ s_x = 2.86$

Y Mean, Standard deviation: $\overline{y} = 46.60,\ s_y = 7.87$

Sample correlation: $r = 0.9026$

Least squares fit line: $y = 11.3233 + 2.4846\ x$



Point-slope form. In general, the linear least squares fit line contains the point $(\overline{x}, \overline{y})$. This implies that the line with equation $y = \widehat{\alpha} + \widehat{\beta}x$ can be rewritten in point-slope form:

$$y = \widehat{\alpha} + \widehat{\beta}x = \overline{y} + \widehat{\beta}(x - \overline{x}).$$

For the small data set above, the line can be written as

$$y = 11.3233 + 2.4846\ x = 46.60 + 2.4846(x - 14.20).$$

Multivariable calculus can be used to prove the following general estimation theorem:

Theorem (Parameter Estimation). Given the conditions of this section, the following are LS estimators of the parameters:

1. Slope: $\widehat{\beta} = \sum_{i=1}^{N} \left[\frac{(x_i - \overline{x})}{S_{xx}} \right] Y_i$
2. Intercept: $\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{x} = \sum_{i=1}^{N} \left[\frac{1}{N} - \frac{(x_i - \overline{x})\overline{x}}{S_{xx}} \right] Y_i$
3. Error terms: $\widehat{\epsilon_i} = Y_i - \widehat{Y_i} = Y_i - (\widehat{\alpha} + \widehat{\beta} x_i)$

In these formulas, $\overline{x}$ is the sample mean of the predictors, $\overline{Y}$ is the sample mean of the responses, and $S_{xx} = \sum_{i=1}^{N} (x_i - \overline{x})^2$ is the sum of squared deviations of observed predictors from their sample mean.

Each estimator is unbiased. Further, the statistic

$$S^2 = \frac{1}{N-2} \sum_{i=1}^{N} (\widehat{\epsilon_i})^2 = \frac{1}{N-2} \sum_{i=1}^{N} \Big(Y_i - (\widehat{\alpha} + \widehat{\beta} x_i) \Big)^2 ,$$

is an unbiased estimator of the common variance σ^2.

Exercise. Demonstrate that $\widehat{\beta}$ is an unbiased estimator of β.

Exercise. Demonstrate that $\widehat{\beta}$ can be written in the following equivalent form:

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}}, \quad \text{where } S_{xx} = \sum_{i=1}^{N}(x_i - \overline{x})^2 \text{ and } S_{xy} = \sum_{i=1}^{N}(x_i - \overline{x})(Y_i - \overline{Y}).$$

Properties of expectation can be used to prove the following summary theorem:

> ***Theorem (Variance-Covariance).*** Let $\widehat{\beta}$ and $\widehat{\alpha}$ be the LS estimators of the slope and intercept of a simple linear model. Then
>
> $$Var(\widehat{\beta}) = \frac{\sigma^2}{S_{xx}}, \quad Var(\widehat{\alpha}) = \frac{\left(\sum_{i=1}^{N} x_i^2\right)\sigma^2}{NS_{xx}}, \quad Cov(\widehat{\alpha}, \widehat{\beta}) = -\frac{\sigma^2\,\overline{x}}{S_{xx}},$$
>
> where $\overline{x}$ is the mean of the predictor variables, and $S_{xx} = \sum_{i=1}^{N}(x_i - \overline{x})^2$.

Note that if $\overline{x} = 0$, then the LS estimators are uncorrelated.

Exercise. Demonstrate that $Var(\widehat{\beta}) = \dfrac{\sigma^2}{S_{xx}}$.

4.1.2 Equivalent Model, Predicted Responses, Residuals

The conditional model can be written in the following equivalent form:

$$Y_i = \mu + \beta(x_i - \overline{x}) + \epsilon_i, \quad i = 1, 2, \ldots, N,$$

where

1. $\mu = E(\overline{Y}) = E\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) = \frac{1}{N}\sum_{i=1}^{N} E(Y_i) = \frac{1}{N}\sum_{i=1}^{N}(\alpha + \beta x_i) = \alpha + \beta\overline{x}$
   is the *overall mean*. That is, μ is the expected value of the mean of the response variables.

2. $\beta(x_i - \overline{x})$ is the i^{th} *differential effect*. That is, $\beta(x_i - \overline{x})$ is the deviation of the i^{th} observed mean from the overall mean.

   Note that the sum of the differential effects is zero: $\sum_{i=1}^{N} \beta(x_i - \overline{x}) = 0$.

Note on Point-Slope Form:
The reparametrization of the conditional model has the effect of rewriting the linear least squares fit line in point-slope form, as illustrated with the small example on page 4.

Predicted responses, residuals. For each i,

1. $\widehat{Y_i} = \widehat{\alpha} + \widehat{\beta}x_i = \widehat{\mu} + \widehat{\beta}(x_i - \overline{x}) = \overline{Y} + \widehat{\beta}(x_i - \overline{x})$ is called the i^{th} *predicted response*.

2. $\widehat{\epsilon_i} = Y_i - \widehat{Y_i} = (Y_i - \overline{Y}) - \widehat{\beta}(x_i - \overline{x})$ is the i^{th} *estimated error* (or i^{th} *residual*).

4.1.3 Partitioning Sum of Squared Deviations

The goal of this section is to consider sums of squared deviations analogous to those we studied when doing analysis of variance.

Sum of squared deviations. The sum of squared deviations of the response variables (Y_i) from the estimated overall mean ($\overline{Y}$) can be partitioned as follows:

$$\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{N}(\widehat{Y}_i - \overline{Y})^2.$$

Another way to write this formula is $SS_t = SS_e + SS_m$, where

1. $SS_t = \sum_{i=1}^{N}(Y_i - \overline{Y})^2$ is the *total sum of squares.* That is, SS_t is the sum of squares of deviations of responses from the estimated overall mean.

   This quantity is also written as S_{yy}.

2. $SS_e = \sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2$ is the *error sum of squares.* That is, SS_e, is the sum of squared deviations of response variables from predicted responses.

   Note that the error sum of squares is used to estimate σ^2.

3. $SS_m = \sum_{i=1}^{N}(\widehat{Y}_i - \overline{Y})^2$ is the *model sum of squares.* That is, SS_m is the sum of squared deviations of predicted responses from the estimated overall mean.

   Since $\widehat{Y}_i - \overline{Y} = \widehat{\beta}(x_i - \overline{x})$, we know that $SS_m =$ _____.

Note on Working with the Squared Deviations:

If we have chosen the correct model for the conditional mean, then we expect the model sum of squares to be large and the error sum of squares to be small.

Equivalently, if we have chosen the correct model, then

1. We expect the ratio $\dfrac{SS_m}{SS_t}$ to be close to 1, and

2. We expect the ratio $\dfrac{SS_e}{SS_t}$ to be close to 0.

4.1.4 Coefficient of Determination

The ratio of the model sum of squares (SS_m) to the total sum of squares (SS_t) is known as the *coefficient of determination*. It is the proportion of the total variation in responses that is "explained" by the model for conditional means.

The coefficient of determination can be computed in many different ways, including the one shown below:

$$\frac{SS_m}{SS_t} = \frac{S_{xy}^2}{S_{xx}S_{yy}}.$$

Exercise. Use the work you did on page 6 to demonstrate that the coefficient of determination can be computed using the form shown on the right above.

Exercise. For the data pictured on page 4,

1. $s_{xx} = \sum_{i=1}^{8}(x_i - \overline{x})^2 = 57.2305$,
2. $s_{yy} = \sum_{i=1}^{8}(y_i - \overline{y})^2 = 433.573$, and
3. $s_{xy} = \sum_{i=1}^{8}(x_i - \overline{x})(y_i - \overline{y}) = 142.194$.

Find and interpret the coefficient of determination.

4.1.5 Relationship Between Least Squares Estimator of Slope and Correlation

Let (X, Y) be a random pair with finite means and variances, and finite correlation. If the conditional mean $E(Y|X = x)$ is a linear function of x, then it must have a specific form.

> ***Theorem (Linear Conditional Means).*** Assume that the joint (X, Y) distribution has finite summaries:
>
> $$\mu_x = E(X),\ \sigma_x = SD(X),\ \mu_y = E(Y),\ \sigma_y = SD(Y),\ \rho = Corr(X, Y),$$
>
> and suppose that $E(Y|X = x)$ is a linear function of x.
>
> Then the conditional mean can be written in the following form:
>
> $$E(Y|X = x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x).$$

Thus, if sampling is done from a joint distribution with a linear conditional mean,

1. The slope parameter can be written as $\beta = \rho\ (\sigma_y/\sigma_x)$.
2. The coefficient of determination is the square of the sample correlation:
$$R = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\left(\sum_{i=1}^{N}(X_i - \overline{X})^2\right)\left(\sum_{i=1}^{N}(Y_i - \overline{Y})^2\right)}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \quad \Longrightarrow \quad R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}.$$
3. Further, if S_x and S_y is the sample standard deviations of the X and Y samples, respectively, and R be the sample correlation statistic, then the LS estimator of β can be written in the following equivalent form: $\widehat{\beta} = R\ (S_y/S_x)$.

To see the third item (please complete):

Example (Rice textbook, Chapter 7). As part of a study of cancer mortality rates, public health officials gathered information on a simple random sample of 301 counties in the southeastern United States for the years 1950 through 1960.

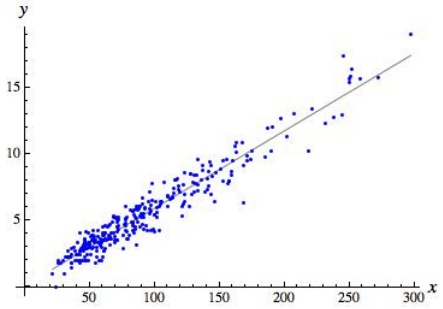
Their information included the following variables:

1. *Total Female Population:* The total number of women living in the county in 1960.
2. *Breast Cancer Deaths:* The total number of women who died of breast cancer in the county for the 10-year period from the beginning of 1950 to the end of 1959.

The goal of this example is to use total population size to predict the number of breast cancer deaths. Since the variability of the number of breast cancer deaths depends on total population size, we will instead consider transformed values using the *square root* transformation.

Let X be the *square root* of the total number of women living a county, and Y be the *square root* of the total number of women who die of breast cancer during a 10-year period, and assume that the transformed counts for the 301 counties are the values of a random sample from the joint (X, Y) distribution.

X Mean, Standard deviation:	$\overline{x} = 92.85$, $s_x = 51.73$
Y Mean, Standard deviation:	$\overline{y} = 5.49$, $s_y = 3.13$
Sample correlation:	$r = 0.9650$
Least squares fit line:	$y = 0.0664 + 0.0584\ x$

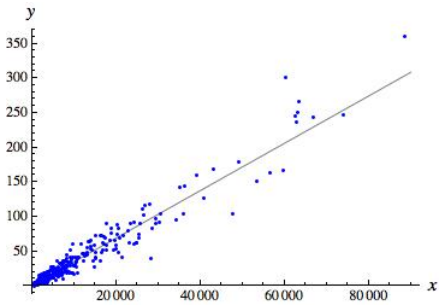


The coefficient of determination is $(0.965)^2 = 0.931225$. Thus, the model explains about 93.1% of the total variation in the response (the square root of the number of breast cancer deaths).

Note that the estimated slope can be computed as $\widehat{\beta} = 0.9650\left(\frac{3.13}{51.73}\right) = 0.0584$.

The plot on the right shows the original population-mortality pairs. The curve (which looks almost linear) is the transformed linear least squares fit line.

The formula for this curve is (please complete)



4.1.6 Gauss-Markov Theorem

This section considers estimators that can be written as linear combinations of response variables. Specifically, let Y_1, Y_2, ..., Y_N be N independent random variables,

1. *Linear Estimator:*

   A linear estimator of a parameter θ based on Y_1, ..., Y_N is an estimator of the form

   $$W = \sum_{i=1}^{N} d_i Y_i, \quad \text{where the } d_i\text{'s are known constants.}$$

2. *Best Linear Unbiased Estimator (BLUE):*

   The linear estimator W is said to be a best linear unbiased estimator of θ if it satisfies

   $$E(W) = E\left(\sum_{i=1}^{N} d_i Y_i\right) = \theta \text{ and } Var(W) = Var\left(\sum_{i=1}^{N} d_i Y_i\right) \text{ is minimum possible.}$$

For the conditional model we are studying, the best linear unbiased estimators of α and β are the least squares estimators.

> ***Gauss-Markov Theorem.***[1] If the random variables
>
> $$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, N,$$
>
> are independent with means $E(Y_i) = \alpha + \beta x_i$, and the collection $\{\epsilon_i\}$ is a random sample of size N from a distribution with mean 0 and standard deviation σ, then the least squares estimators obtained earlier are the best linear unbiased estimators of the slope β and the intercept α.

Notes:

1. Although least squares estimators are best among the unbiased estimators of the form $W = d_1 Y_1 + d_2 Y_2 + \cdots + d_N Y_N$, they may not be best over all.
2. If the error distribution is normal, however, then the least squares estimators are *both* the best linear unbiased estimators and the maximum likelihood estimators.
3. The proof of the Gauss-Markov theorem uses Lagrange multipliers.

[1] *Historical note:* Gauss provided a proof of part of the theorem in 1822, Markov used these ideas to complete the proof in 1900, and Graybill provided a modern version of the theorem in 1976.

Footnote on using Lagrange multipliers to estimate β. It is instructive to provide the outline of the Lagrange multipliers method for estimating β. Here, W must satisfy

1. $E(W) = \sum_{i=1}^{N} d_i E(Y_i) = \sum_{i=1}^{N} d_i(\alpha + \beta x_i) = \alpha \left(\sum_{i=1}^{N} d_i\right) + \beta \left(\sum_{i=1}^{N} d_i x_i\right) = \beta,$
2. $Var(W) = \sum_{i=1}^{N} d_i^2 Var(Y_i) = \sum_{i=1}^{N} d_i^2 \sigma^2 = \sigma^2 \left(\sum_{i=1}^{N} d_i^2\right)$ is minimum possible.

Thus, we need to minimize the N variable function

$$f(d_1, d_2, \ldots, d_N) = d_1^2 + d_2^2 + \cdots + d_N^2,$$

subject to two constraints:

$$\begin{aligned} &\text{[Constraint 1:]} \quad g_1(d_1, d_2, \ldots, d_N) = d_1 + d_2 + \cdots + d_N = 0 \text{ and} \\ &\text{[Constraint 2:]} \quad g_2(d_1, d_2, \ldots, d_N) = d_1 x_1 + d_2 x_2 + \cdots + d_N x_N = 1. \end{aligned}$$

The method of Lagrange multipliers then implies the unique solution is

$$d_i = \frac{(x_i - \overline{x})}{S_{xx}}, \quad \text{for } i = 1, 2, \ldots, N.$$

These are the coefficients used in least squares estimation.

4.1.7 Permutation Confidence Interval Procedure for the Slope Parameter

Let $Y = \alpha + \beta X + \epsilon$, where the predictor X and error ϵ are *independent* continuous random variables and where the ϵ distribution is symmetric around zero with finite standard deviation.

Permutation methods can be used to construct confidence intervals for the slope parameter. Since

$$X \text{ and } Y - \beta X = \alpha + \epsilon$$

are *independent*, and since $\beta = \rho(\sigma_y/\sigma_x)$, the idea follows the general strategy of permutation methods for the correlation coefficient ρ.

Outline for a $100(1-\gamma)\%$ confidence interval for β is as follows:

1. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ be a random sample from the (X, Y) distribution.

   A value β_0 will be in the $100(1-\gamma)\%$ confidence interval for β if and only if you would accept $\beta = \beta_0$ when testing

   $$\beta = \beta_0 \text{ versus } \beta \neq \beta_0 \text{ at the } 100\gamma\% \text{ level}$$

   using permutation methods.

Equivalently, a value β_0 will be in the $100(1-\gamma)\%$ confidence interval for β if and only if you would accept $\rho = 0$ when testing

$$\rho = 0 \text{ versus } \rho \neq 0 \text{ at the } 100\gamma\% \text{ level}$$

using permutation methods with the following samples:

$$\begin{array}{rl} \textit{First Coordinates:} & X_1, X_2, \ldots, X_N \\ \textit{Second Coordinates:} & (Y_1 - \beta_0 X_1), (Y_2 - \beta_0 X_2), \ldots, (Y_N - \beta_0 X_N) \end{array}$$

2. Working with the $(X_i, Y_i - \beta_0 X_i)$ pairs, let

$$R_{\beta_0} = \frac{S_{xy_0}}{\sqrt{S_{xx} S_{y_0 y_0}}}$$

be the sample correlation, where the subscript "y_0" indicates that the shifted $(Y_i - \beta_0 X_i)$ values are used instead of the original Y_i values.

3. Let π represent a permutation of the indices, and let

$$R_{\beta_0}(\pi) = \frac{S_{x\pi(y_0)}}{\sqrt{S_{xx} S_{\pi(y_0)\pi(y_0)}}}$$

be the permuted sample correlation, where the subscript "$\pi(y_0)$" indicates that the permuted shifted values $(Y_{\pi(i)} - \beta_0 X_{\pi(i)})$ are used instead of the original Y_i values.

Using these definitions, we would accept $\rho = 0$ if and only if

$$\frac{\#(|R_{\beta_0}(\pi)| \geq |R_{\beta_0}|)}{N!} > \gamma.$$

4. The last inequality can be used to find elementary estimates of β.

There are a total of $N! - 1$ elementary estimates of the form

$$B_\pi = \frac{\sum_{i=1}^N (X_i - \overline{X})(Y_{\pi(i)} - Y_i)}{\sum_{i=1}^N (X_i - \overline{X})(X_{\pi(i)} - X_i)}, \quad \text{where } \pi \text{ is a permutation of the indices;}$$

the identity permutation must be excluded since you get zero in the denominator.

[*Note* that these elementary estimates represent locations along the real line where the counts needed to find a permutation p would change. This strategy is similar to the strategy we used when we worked with the Smirnov two sample test.]

5. The ordered $N! - 1$ elementary estimates of β,

$$B_{(1)} < B_{(2)} < \cdots < B_{(N!-2)} < B_{(N!-1)},$$

divide the real line into $N!$ subintervals; β lies in each subinterval with equal probability. That is, if we let $B_{(0)} = -\infty$ and $B_{(N!)} = \infty$ for convenience, then

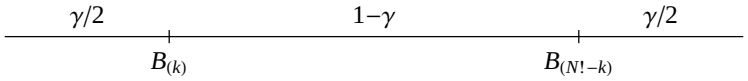
$$P(B_{(k)} < \beta < B_{(k+1)}) = \frac{1}{N!} \text{ for all } k.$$

These ideas can be used to prove the following theorem:

Slope Parameter Confidence Interval Theorem. Under the assumptions of this section, if k is chosen so that

$$\frac{k}{N!} = \frac{\gamma}{2},$$

then $\left[B_{(k)}, B_{(N!-k)}\right]$ is a $100(1-\gamma)\%$ confidence interval for β.



Note on implementing this method. In most practical situations, it is impossible to generate all $N!-1$ elementary estimates. Instead, we use the computer to generate a large number (M) of random permutations, generate an elementary estimate of β from each permutation,

$$b_1,\ b_2,\ \ldots,\ b_M,$$

and report the sample $\frac{\gamma}{2}$ and $1 - \frac{\gamma}{2}$ quantiles of the M elementary estimates.

Example from page 11, continued. Recall that the least squares fit line relating population (on the square root scale) to cancer mortality (on the square root scale) was $y = 0.0664 + 0.0584x$.

Since $N = 301$ is quite large, a complete enumeration is impossible and Monte Carlo analysis will be used instead. Using 5000 random permutations, an approximate 95% confidence interval for the slope parameter is $[0.0565, 0.0602]$.

Example (Mendenhall et al, 2002). The following data are cylinder volumes in cubic inches (x) and fuel efficiency in miles per gallon (y) for a simple random sample of 9 subcompact four-cylinder cars tested by the Environmental Protection Agency:

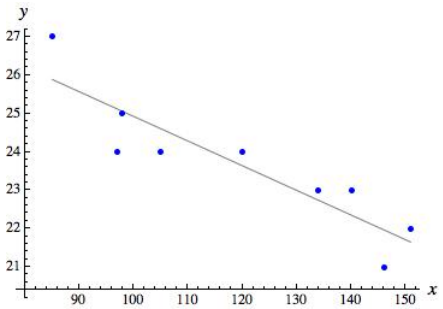
x (in^3)	85	97	98	105	120	134	140	146	151
y (MPG)	27	24	25	24	24	23	23	21	22

X Mean, Standard deviation: $\overline{x} = 119.56$, $s_x = 24.22$

Y Mean, Standard deviation: $\overline{y} = 23.67$, $s_y = 1.73$

Sample correlation: $r = -0.9007$

Least squares fit line: $y = 31.3667 - 0.0644\ x$



Assume the pairs are the values of a random sample from a bivariate distribution, and that we are interested in using permutation methods to find a 95% confidence interval for β.

Since $N = 9$ in this case, a complete enumeration is possible. Further, since $0.025(9!) = 9072$, the 95% confidence interval is

$$[b_{(9072)}, b_{(353808)}] = [-0.0929, -0.0386].$$

Thus, with 95% confidence, we believe that fuel efficiency will decrease by between about 0.039 and 0.093 miles per gallon with each cubic inch increase in cylinder volume.

4.1.8 Diagnostics: Residual, Leverage, Influence

This section is concerned with methods for identifying observations that are particularly influential in a linear least squares analysis.

Predicted responses, residuals. To begin, recall that the i^{th} predicted response is

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta} x_i = \overline{Y} + \widehat{\beta}(x_i - \overline{x}) = \sum_{j=1}^{N} \left[\frac{1}{N} + \frac{(x_j - \overline{x})(x_i - \overline{x})}{S_{xx}} \right] Y_j = \sum_{j=1}^{N} h_{ij} Y_j,$$

and that the i^{th} estimated error (residual) is

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i, \quad \text{for } i = 1, 2, \ldots, N.$$

Observations with unusual residuals may be important in our analysis.

Leverage. The last expression for the predicted response emphasizes that each estimated mean is a linear combination of all the responses. Based on this expression, we define the leverage of the i^{th} response to be the coefficient of Y_i in the linear combination:

$$h_i \equiv h_{ii} = \frac{1}{N} + \frac{(x_i - \overline{x})^2}{S_{xx}}.$$

Leverages satisfy the following properties:

1. For each i, $0 \leq h_i \leq 1$.
2. $\sum_{i=1}^{N} h_i = 2$.

Ideally, leverages should all be approximately the average, $2/N$. However, it is clear from the formula that leverage increases as the distance between x_i and $\overline{x}$ increases.

Footnote on rule-of-thumb for leverages:

Huber (1981)[a] suggests that cases with leverage greater than 0.50 (i.e., with $h_i > 0.50$) are potential problems in least squares analyses.

[a]Peter J. Huber, *Robust Statistics*, Wiley, New York, 1981.

Footnote on using matrix methods:

The *hat matrix*, $\boldsymbol{H}$, is the N-by-N matrix whose i^{th} row,

$$h_{ij} = \left[\frac{1}{N} + \frac{(x_j - \overline{x})(x_i - \overline{x})}{S_{xx}}\right], \quad j = 1, 2, \ldots, N,$$

is the vector of coefficients for the i^{th} predicted response.

If we let $\underline{Y}$ represent the N-by-1 vector of responses and $\underline{\widehat{Y}}$ represent the N-by-1 vector of predicted responses, then we can use notation from linear algebra to write

$$\boldsymbol{H}\,\underline{Y} = \underline{\widehat{Y}}.$$

($\boldsymbol{H}$ puts a "hat" on the vector of responses.) The trace of the hat matrix is 2.

Example (Hoaglin, 1988).[2] The data for this example are on public affairs activities in 12 US government agencies in 1976: x is the total number of fulltime employees working to provide press information or to prepare advertising, exhibits, films, publications, and speeches for release to the public; y is the total operating cost (in millions of dollars) of these activities.

i	Agency	x_i	y_i	i	Agency	x_i	y_i
1	Defense	1486	24.5	7	Energy R&D	128	5.2
2	HEW	388	21.0	8	NASA	208	4.5
3	Agriculture	650	11.5	9	Transportation	117	2.9
4	Treasury	202	5.8	10	HUD	69	2.5
5	Congress	446	5.7	11	White House	85	2.3
6	Commerce	164	5.7	12	Veterans Adm.	47	1.3

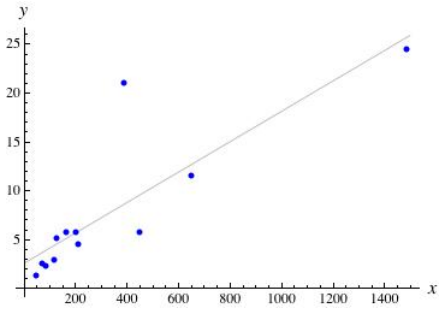
Numerical and graphical summaries for these data are given below:

X Mean, Standard deviation: $\overline{x} = 332.50$, $s_x = 405.50$

Y Mean, Standard deviation: $\overline{y} = 7.74$, $s_y = 7.52$

Sample correlation: $r = 0.8396$

Least squares fit line: $y = 12.5651 + 0.0156\ x$



The residuals (estimated errors) and leverages are as follows:

i	1	2	3	4	5	6	7	8	9	10	11	12
$\widehat{\epsilon_i}$	−1.20	12.39	−1.18	0.09	−3.81	0.58	0.64	−1.30	−1.49	−1.14	−1.59	−2.00
h_i	0.82	0.09	0.14	0.09	0.09	0.10	0.11	0.09	0.11	0.12	0.12	0.13

The first two observations (Defense and HEW) are the most unusual:

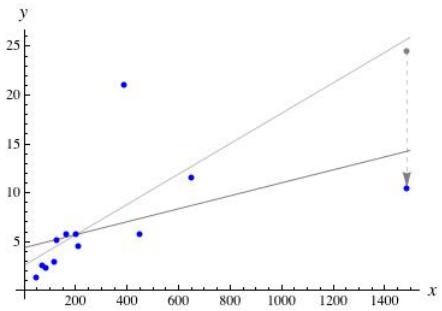
- Defense has a moderate residual but a large leverage;
- HEW has a large residual but moderate leverage.

[2] David C. Hoaglin, Using Leverage and Influence to Introduce Regression Diagnostics, *The College Mathematics Journal*, 19(5):387-401, 1988.

Impact of high leverage: An observation with high leverage can have high impact on a least squares analysis. To illustrate this, I changed the y-coordinate of the information for the Defense department from $y = 24.5$ million dollars to $y = 10.5$ million dollars.

The change acts like a "lever", pulling the least squares line close to the new point.

Since the Defense department has many workers devoted to public affairs, its weight has the potential of changing the results significantly.



Influence, standardized influence. The next two concepts combine both the leverage of an observation and the estimated error for that observation.

Influence: The influence of the i^{th} observation is the change in prediction if we delete the i^{th} observation from the data set. Influence can be written as $\widehat{Y}_i - \widehat{Y}_i(i)$, where

1. $\widehat{Y}_i = \overline{Y} + \widehat{\beta}(x_i - \overline{x})$ is the predicted response based on all N cases and
2. $\widehat{Y}_i(i) = \overline{Y}(i) + \widehat{\beta}(i)(x_i - \overline{x}(i))$ is the prediction for the "new" predictor x_i, based on least squares analysis using the remaining $N - 1$ cases.

   In this formula,

   - $\overline{x}(i)$ is the mean value of predictors,
   - $\overline{Y}(i)$ is the mean value of responses and
   - $\widehat{\beta}(i)$ is the estimate of slope

   for the $N - 1$ cases with the i^{th} case removed.

Since summaries for $N - 1$ cases are closely related to summaries for all N cases, we can demonstrate the following:

$$Var(\widehat{Y}_i(i)) = \sigma^2 h_i \quad \text{and} \quad \widehat{Y}_i - \widehat{Y}_i(i) = \frac{h_i\ \widehat{\epsilon}_i}{1 - h_i}.$$

Standardized influence: The standardized influence of the i^{th} observation is defined to be the influence divided by its standard error. Using the information above, we get

$$\frac{\widehat{Y}_i - \widehat{Y}_i(i)}{SD(\widehat{Y}_i(i))} = \frac{\widehat{\epsilon}_i\ h_i/(1 - h_i)}{\sqrt{\sigma^2\ h_i}}.$$

Estimated standardized influence: The estimated standardized influence of the i^{th} observation, denoted by δ_i, is the estimate of the standardized influence where the sample variance based on the $N-1$ cases with the i^{th} case removed is substituted for σ^2. That is,

$$\delta_i = \frac{\widehat{Y}_i - \widehat{Y}_i(i)}{\sqrt{S^2(i)\ h_i}} = \frac{\widehat{\epsilon}_i h_i/(1-h_i)}{\sqrt{S^2(i)\ h_i}} = \frac{\widehat{\epsilon}_i\sqrt{h_i}}{S(i)(1-h_i)}.$$

Observations with large absolute δ_i values are the most unusual.

Interestingly, the formula for $S^2(i)$ reduces to

$$S^2(i) \ = \ \frac{1}{N-3}\left((N-2)S^2 - \frac{\widehat{\epsilon}_i^{\,2}}{1-h_i}\right).$$

Thus, each δ_i can be computed using information available when all N cases are analyzed.

Footnote on the rule-of-them for standardized influences:

Belsley, Kuh & Welch (1980)[a] suggest that cases with standardized influences satisfying

$$|\delta_i| > 2\sqrt{2/N}$$

are highly influential in least squares analyses. They use the notation DFFITS_i for δ_i.

[a] D.A. Belsley, E. Kuh, and R.E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York, 1980.

For the public affairs data from page 18, estimated standardized influences are as follows:

i	1	2	3	4	5	6	7	8	9	10	11	12
δ_i	−1.36	3.01	−0.11	0.01	−0.29	0.05	0.05	−0.10	−0.12	−0.10	−0.14	−0.18

Since $N = 12$, the cutoff for influence is $2\sqrt{2/N} \approx 0.816$.

Thus, both HEW and the Defense Department are highly influential.

4.2 Simple Linear Regression

In simple linear regression, we assume that measurement errors are normally distributed.

4.2.1 Least Squares Estimation

The data are assumed to be the values of N independent random variables,

$$Y_i = \alpha + \beta x_i + \epsilon_i, \text{ for } i = 1, 2, \ldots, N, \text{ where}$$

1. The means are $E(Y_i) = \alpha + \beta x_i$,
2. the collection of errors $\{\epsilon_i\}$ is a random sample from a *normal distribution* with mean 0 and standard deviation σ, and
3. All parameters (α, β, σ) are assumed to be unknown.

Maximum likelihood estimation. Maximum likelihood (ML) estimators for the parameters of the linear model are given in the following theorem.

> ***Theorem (Parameter Estimation).*** Given the assumptions above, the following are ML estimators of the parameters of the linear model:
>
> 1. Slope: $\widehat{\beta} = \sum_{i=1}^{N} \left[\frac{(x_i - \overline{x})}{S_{xx}} \right] Y_i$
> 2. Intercept: $\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{x} = \sum_{i=1}^{N} \left[\frac{1}{N} - \frac{(x_i - \overline{x})\overline{x}}{S_{xx}} \right] Y_i$
> 3. Error terms: $\widehat{\epsilon_i} = Y_i - (\widehat{\alpha} + \widehat{\beta} x_i)$
>
> $$= Y_i - (\overline{Y} + \widehat{\beta}(x_i - \overline{x})) = Y_i - \sum_{j=1}^{N} \left[\frac{1}{N} + \frac{(x_j - \overline{x})(x_i - \overline{x})}{S_{xx}} \right] Y_j \quad \text{for all } i.$$
>
> Each estimator is unbiased, and each estimator has a normal distribution.
>
> Further, the statistic
>
> $$S^2 = \frac{1}{N-2} \sum_{i=1}^{N} (Y_i - (\widehat{\alpha} + \widehat{\beta} x_i))^2$$
>
> is an unbiased estimator of the common variance σ^2.

Properties of expectation can be used to prove the following summary theorem:

Theorem (Variance-Covariance). Under the assumptions above,

1. Variances for slope, intercept and measurement error estimators are as follows:

$$Var(\widehat{\beta}) = \frac{\sigma^2}{S_{xx}}, \quad Var(\widehat{\alpha}) = \frac{\left(\sum_{i=1}^{N} x_i^2\right)\sigma^2}{NS_{xx}}, \text{ and}$$

$$Var(\widehat{\epsilon_i}) = \sigma^2\left[\left(1 - \tfrac{1}{N} - \tfrac{(x_i-\overline{x})^2}{S_{xx}}\right)^2 + \textstyle\sum_{j\neq i}\left(\tfrac{1}{N} + \tfrac{(x_j-\overline{x})(x_i-\overline{x})}{S_{xx}}\right)^2\right] = \sigma^2\ c_i.$$

2. Covariances among the estimators are as follows:

$$Cov(\widehat{\alpha},\widehat{\beta}) = -\frac{\sigma^2\ \overline{x}}{S_{xx}}, \text{ and } Cov(\widehat{\epsilon_i},\widehat{\alpha}) = Cov(\widehat{\epsilon_i},\widehat{\beta}) = 0.$$

<u>Notice</u> that the variances of the $\widehat{\epsilon_i}$'s are not all equal.

Further, since the estimators for the error terms are uncorrelated with the estimators for the slope and intercept, we know that $\widehat{Y_i}$ and $\widehat{\epsilon_i}$ are uncorrelated for each i.

4.2.2 Diagnostic Plots

The summary theorem above allows us to develop two diagnostic plots:

1. *Mean-Residual Plot:*

   A plot of observed mean-residual pairs

$$(\widehat{y_i}, \widehat{\epsilon_i}), \text{ for } i = 1, 2, \ldots, N,$$

   should exhibit no relationship between estimated means and estimated errors.

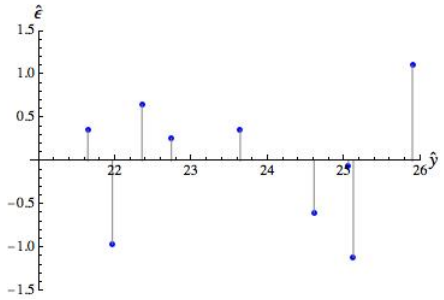
2. *Normal Probability Plot:*

   A normal probability plot of standardized residuals,

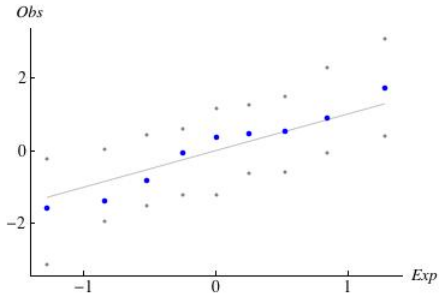
$$r_i = \frac{\widehat{\epsilon_i}}{\sqrt{S^2\ c_i}}, \quad \text{for } i = 1, 2, \ldots, N,$$

   where S^2 is the estimate of the common variance and c_i is the constant in brackets in the variance formula for i^{th} residual, should suggest that standardized errors are approximately normally distributed.

Consider again the fuel efficiency study (page 16), where x represents cylinder volume in cubic inches and y represents fuel efficiency in MPG of 9 subcompact four-cylinder cars tested by the EPA. For these data, the estimated common variance is $s^2 = 0.8043^2 = 0.6469$, and diagnostic plots are as follows:



Mean-Residual Plot



Normal Probability Plot of Standardized Residuals

4.2.3 Analysis of Variance

The first step in a linear regression analysis is to determine if the proposed predictors have any predictive value. In the simple case, this is equivalent to testing the null hypothesis that the slope parameter $\beta = 0$ versus the alternative hypothesis that $\beta \neq 0$.

We organize several computations into an initial *analysis of variance table:*

	df	*Sum of Squares*	*Mean Square*
Model:	1	$SS_m = \sum_{i=1}^{N}(\widehat{Y}_i - \overline{Y})^2$	$MS_m = SS_m$
Error:	$N-2$	$SS_e = \sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2$	$MS_e = \frac{1}{N-2}SS_e$
Total:	$N-1$	$SS_t = \sum_{i=1}^{N}(Y_i - \overline{Y})^2$	

In this table,

1. The total sum of squares (SS_t) is equal to the sum of the model sum of squares (SS_m) and the error sum of squares (SS_e):

$$SS_t = SS_m + SS_e.$$

2. The error mean square (MS_e) is the same as the estimator of the common variance. Thus,

$$E(MS_e) = E\left(\frac{1}{N-2}SS_e\right) = \sigma^2.$$

3. The model mean square (MS_m) is the same as the model sum of squares (SS_m). Since $MS_m = SS_m = \widehat{\beta}^2 S_{xx}$,

$$E(MS_m) = E(\widehat{\beta}^2 S_{xx}) = \sigma^2 + S_{xx}\beta^2.$$

Note that if $\beta = 0$, then $E(MS_m) = \sigma^2$; otherwise, $E(MS_m)$ is larger than σ^2.

Question: Can you see why $E(MS_m) = \sigma^2 + S_{xx}\beta^2$?

To determine if the proposed predictor has some predictive value, we work with the ratio of the model mean square to the error mean square, $F = MS_m/MS_e$. Large values of F favor the alternative hypothesis that the proposed predictor has some predictive value. Under the null hypothesis of no predictive value, F has an f ratio distribution.

Theorem (Distribution Theorem). Under the general assumptions of this section, if $\beta = 0$ then the ratio

$$F = MS_m/MS_e$$

has an f ratio distribution with 1 and $(N-2)$ degrees of freedom.

Example from page 16, continued. Assume that the data from the fuel efficiency study are the values of independent random variables satisfying the simple linear regression model and that we are interested in testing the null hypothesis that cylinder volume has no predictive value versus the alternative that it has some predictive value at the 5% significance level.

The following is an analysis of variance table for the test:

	df	*SS*	*MS*	*F*	*p value*
Model:	1	19.4719	19.4719	30.1017	0.0009
Error:	7	4.5281	0.6469		
Total:	8	24.0000			

Since the p value is less than 0.05, the results suggest that cylinder volume is a significant predictor of fuel efficiency. In fact, since the estimated slope and correlation are negative, the results further suggest that these variables are negatively associated.

4.2.4 Inference for Slope and Intercept Parameters

Let $\widehat{\alpha}$ and $\widehat{\beta}$ be the ML estimators of slope and intercept. Then

> ***Theorem (Distribution Theorem).*** Under the general conditions of this section, the statistics
>
> $$T_\alpha = \frac{(\widehat{\alpha} - \alpha)}{\sqrt{(S^2 \sum_{i=1}^N x_i^2)/(N S_{xx})}} \qquad \text{and} \qquad T_\beta = \frac{(\widehat{\beta} - \beta)}{\sqrt{S^2/S_{xx}}}$$
>
> have Student t distributions with $(N-2)$ df, where S^2 is the estimator of the common variance and $S_{xx} = \sum_{i=1}^N (x_i - \overline{x})^2$.

Example from page 16, continued. For the fuel efficiency study with $N = 9$,

$$\widehat{\beta} = -0.0644, \quad s^2 = 0.6469, \quad s_{xx} = 4694.22.$$

Assume that the data from the fuel efficiency study are the values of independent random variables satisfying the simple linear regression model.

Use this information to construct and interpret a 95% confidence interval for β.

4.2.5 Inference for Mean Response Parameters

Let (x_o, Y_o) be a new predictor-response pair, and let $E(Y_o) = \alpha + \beta x_o$ be the mean response at level x_o. Then, $E(Y_o)$ can be estimated using the following statistic:

$$\widehat{\alpha} + \widehat{\beta} x_o = \overline{Y} + \widehat{\beta}(x_o - \overline{x}) = \sum_{i=1}^{N} \left[\frac{1}{N} + \frac{(x_o - \overline{x})(x_i - \overline{x})}{S_{xx}} \right] Y_i.$$

This estimator is a normal random variable with mean $\alpha + \beta x_o$ and

$$Var(\widehat{\alpha} + \widehat{\beta} x_o) = \sigma^2 \left(\frac{1}{N} + \frac{(x_o - \overline{x})^2}{S_{xx}} \right).$$

These facts give us the following distribution theorem:

> ***Theorem (Distribution Theorem).*** Under the general conditions of this section, and with (x_o, Y_o) defined as above, the statistic
>
> $$T = \frac{(\widehat{\alpha} + \widehat{\beta} x_o) - (\alpha + \beta x_o)}{\sqrt{S^2 \left(\frac{1}{N} + \frac{(x_o - \overline{x})^2}{S_{xx}} \right)}}$$
>
> has a Student t distribution with with $(N-2)$ df, where S^2 is the estimator of the common variance and $S_{xx} = \sum_{i=1}^{N} (x_i - \overline{x})^2$.

Example from page 16, cntd. Use the fuel efficiency data to construct and interpret a 95% CI for the mean efficiency for a compact car whose cylinder volume is 130 cubic inches.

4.2.6 *Mathematica* Example: Rod Weights Study

The handout entitled "`SLMExample.pdf`" introduces *Mathematica*'s `LinearModelFit` function and other tools for implementing linear least squares and linear regression analyses. Data from an industrial experiment are used to illustrate the methods.

4.3 Linear Model

A *linear model* is a model of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \epsilon$$

where the distribution of each X_i is independent of ϵ, and the distribution of ϵ has mean 0 and standard deviation σ. Y is called the *response* variable, each X_i is a *predictor* variable, and ϵ represents the *measurement error*.

Generalizing the simple case, assume that

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_{p-1} x_{p-1,i} + \epsilon_i \quad \text{for } i = 1, 2, \ldots, N$$

are independent random variables with means

$$E(Y_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_{p-1} x_{p-1,i} \quad \text{for all } i,$$

that the collection of errors $\{\epsilon_i\}$ is a random sample from a distribution with mean 0 and standard deviation σ, and that all parameters (including σ) are unknown.

The system of N equations can be written in matrix form as follows:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{p-1,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{p-1,2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1,N} & x_{2,N} & \cdots & x_{p-1,N} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

Or, simply as $\underline{Y} = \boldsymbol{X}\underline{\beta} + \underline{\epsilon}$, where

- the N-by-1 vector $\underline{Y}$ is known as the *response vector*,
- the N-by-p matrix $\boldsymbol{X}$ is known as the *design matrix*,
- the p-by-1 vector $\underline{\beta}$ is known as the *parameter vector*, and
- the N-by-1 vector $\underline{\epsilon}$ is known as the *error vector*.

Approximate solutions to the inconsistent system $\boldsymbol{X}\underline{\beta} = \underline{Y}$ can be found using least squares.

Example from page 16, continued. To illustrate the the matrix computations needed to find least squares estimates of slope and intercept, consider again the fuel efficiency data.

x (in^3)	85	97	98	105	120	134	140	146	151
y (MPG)	27	24	25	24	24	23	23	21	22

The system of 9 equations in two unknowns (shown on the *left*) is inconsistent.

The system is then rewritten as the matrix equation $\boldsymbol{X}\underline{\beta} = \underline{Y}$ (shown in the *middle*), and the normal equation $\boldsymbol{X}^T\boldsymbol{X}\underline{\beta} = \boldsymbol{X}^T\underline{Y}$ (shown on the *right*) is formed.

$$\begin{array}{rcl} \alpha + 85\beta & = & 27 \\ \alpha + 97\beta & = & 24 \\ \alpha + 98\beta & = & 25 \\ \alpha + 105\beta & = & 24 \\ \alpha + 120\beta & = & 24 \\ \alpha + 134\beta & = & 23 \\ \alpha + 140\beta & = & 23 \\ \alpha + 146\beta & = & 21 \\ \alpha + 151\beta & = & 22 \end{array} \quad \Rightarrow \quad \begin{bmatrix} 1 & 85 \\ 1 & 97 \\ 1 & 98 \\ 1 & 105 \\ 1 & 120 \\ 1 & 134 \\ 1 & 140 \\ 1 & 146 \\ 1 & 151 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 27 \\ 24 \\ 25 \\ 24 \\ 24 \\ 23 \\ 23 \\ 21 \\ 22 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} 9 & 1076 \\ 1076 & 133336 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 213 \\ 25163 \end{bmatrix}$$

The least squares theorem of linear algebra tells us that the least squares solutions to the original inconsistent system are the same as the solutions to the consistent normal equation. For these data, the solution to the normal equation is $\widehat{\alpha} = 31.3667$ and $\widehat{\beta} = -0.0644$, as before.

4.4 Linear Regression

In linear regression, we assume that error distributions are normal.

4.4.1 Least Squares Estimation

We assume the data are the values of N independent random variables,

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_{p-1} x_{p-1,i} + \epsilon_i \quad \text{for } i = 1, 2, \ldots, N, \text{ where}$$

1. The means of the random variables are $E(Y_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_{p-1} x_{p-1,i}$,
2. The collection of errors $\{\epsilon_i\}$ is a random sample from a *normal distribution* with mean 0 and standard deviation σ, and
3. All parameters (including σ) are unknown.

The conditional model can be written in matrix form as follows

$$\underline{Y} = \boldsymbol{X}\underline{\beta} + \underline{\epsilon},$$

where $\underline{Y}$ is the N-by-1 response vector, $\boldsymbol{X}$ is the N-by-p design matrix, $\underline{\beta}$ is the p-by-1 vector of unknowns, and $\underline{\epsilon}$ is the N-by-1 error vector.

Maximum likelihood estimation. Maximum likelihood (ML) estimators for the parameters of the linear model are given in the following theorem.

> ***Theorem (Parameter Estimation).*** Given the assumptions and definitions above, the following are vectors of ML estimators:
>
> 1. Coefficients:
> $$\underline{\widehat{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\underline{Y}.$$
> 2. Error terms:
> $$\underline{\widehat{\epsilon}} = \underline{Y} - \boldsymbol{X}\underline{\widehat{\beta}} = (\boldsymbol{I} - \boldsymbol{H})\,\underline{Y},$$
>
> where $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$ and $\boldsymbol{I}$ is the N-by-N identity matrix.
>
> Each estimator is a normal random variable, and each is unbiased.
>
> Further, the statistic
> $$S^2 = \frac{1}{N-p}\sum_{i=1}^{N}\left(Y_i - \widehat{Y}_i\right)^2$$
> is an unbiased estimator of σ^2. In this formula, $\widehat{Y}_i$ is the i^{th} estimated mean.

Notes:

1. *Solving Normal Equation:*
   The estimate of $\underline{\beta}$ is the solution to the normal equation:
   $$\boldsymbol{X}^T\boldsymbol{X}\underline{\beta} = \boldsymbol{X}^T\underline{Y} \;\Rightarrow\; \underline{\widehat{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\underline{Y}.$$

2. *Predicted Responses:*
   The i^{th} estimated mean (predicted response) is the random variable
   $$\widehat{Y}_i = \underline{\widehat{\beta}} \cdot \underline{x}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1,i} + \cdots + \widehat{\beta_{p-1}}x_{p-1,i} \text{ for } i = 1, 2, \ldots, N.$$
   Each estimator is a linear combination of the responses. Thus, each predicted response is a linear combination of the responses.

3. *Hat Matrix:*
   The hat matrix $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$ is a projection matrix.
   The hat matrix is the matrix that transforms the responses to the predicted responses:
   $$\underline{\widehat{Y}} = \boldsymbol{X}\,\underline{\widehat{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\,\underline{Y} = \boldsymbol{H}\,\underline{Y}$$
   (The original responses are projected onto the column space of the design matrix.)

Focus on variances and covariances. If $\underline{V}$ is an m-by-1 vector of random variables and $\underline{W}$ is an n-by-1 vector of random variables, then the *covariance matrix*, $Cov(\underline{V}, \underline{W})$, is the m-by-n matrix whose (i, j) term is $Cov(V_i, W_j)$.

Properties of expectation can be used to prove the following summary theorem:

> ***Theorem (Covariance Matrix).*** Under the assumptions of this section, the covariance matrices for the coefficient estimators, for the error estimators, and for associations between error estimators and predicted responses are as follows:
>
> 1. *Coefficient Estimators:* $Cov\left(\underline{\widehat{\beta}}, \underline{\widehat{\beta}}\right) = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$
> 2. *Error Estimators:* $Cov\left(\underline{\widehat{\epsilon}}, \underline{\widehat{\epsilon}}\right) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$
> 3. *Error Estimators & Predicted Responses:* $Cov\left(\underline{\widehat{\epsilon}}, \underline{\widehat{Y}}\right) = \boldsymbol{O}$
>
> In these formulas, $\boldsymbol{H}$ is the hat matrix, $\boldsymbol{I}$ is the N-by-N identity matrix, and $\boldsymbol{O}$ is the N-by-N matrix of zeros.

To illustrate the formula for coefficient estimators, consider again the simple linear case:

$$Y_i = \alpha + \beta x_i + \epsilon_i \quad \text{for } i = 1, 2, \ldots, N.$$

(1) To see that the 2-by-2 matrix $\boldsymbol{X}^T \boldsymbol{X}$ has determinant $N S_{xx} = N \sum_{i=1}^{N} (x_i - \overline{x})^2$,

$$\boldsymbol{X}^T \boldsymbol{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_N \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} = \begin{bmatrix} \_\_\_\_\_ & \_\_\_\_\_ \\ & \\ \_\_\_\_\_ & \_\_\_\_\_ \end{bmatrix}$$

Now (please complete),

(2) Using the general formula for the inverse of a 2-by-2 matrix, the covariance matrix of the coefficient estimators now becomes

$$\begin{bmatrix} Cov(\widehat{\alpha},\widehat{\alpha}) & Cov(\widehat{\alpha},\widehat{\beta}) \\ Cov(\widehat{\beta},\widehat{\alpha}) & Cov(\widehat{\beta},\widehat{\beta}) \end{bmatrix}$$

$$= \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} = \frac{\sigma^2}{N S_{xx}} \begin{bmatrix} \sum_{i=1}^N x_i^2 & -\sum_{i=1}^N x_i \\ -\sum_{i=1}^N x_i & N \end{bmatrix} = \begin{bmatrix} \frac{\sigma^2 \left(\sum_{i=1}^N x_i^2\right)}{(N S_{xx})} & \frac{-\sigma^2 \overline{x}}{S_{xx}} \\ \frac{-\sigma^2 \overline{x}}{S_{xx}} & \frac{\sigma^2}{S_{xx}} \end{bmatrix}$$

The terms of the last matrix correspond to the variance and covariance formulas from page 6.

Example (Fine & Bosch, 2000). As part of a study to assess the adverse effects of a proposed treatment for tuberculosis, ten female rats were given the drug for a period of 14 days at each of five dosage levels (in 100 milligrams per kilogram per day):

0, 1, 2, 5 and 7.5 hundred mg/kg/day.

The response variable of interest was weight change in grams (WC), defined as the weight at the end of the period minus the weight at the beginning of the period.

Assume that dose-WC satisfy the following linear model:

$$\text{WC} = \beta_0 + \beta_1 \text{ dose} + \beta_2 \text{ dose}^2 + \epsilon.$$

Note that the model is linear in the unknown parameters and quadratic in the dosage level.

The following are numerical and graphical summaries for the 50 observations:

Dosage Summaries (100 mg/kg/day):

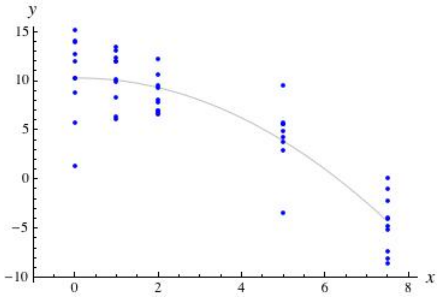
$\overline{x} = 3.1$, $s_x = 2.79$

Weight Change Summaries (grams):

$\overline{y} = 5.828$, $s_y = 6.391$

Least squares fit equation:

$y = 10.2475 + 0.053421x - 0.2658x^2$



The estimated common variance is $s^2 = 3.0757^2$.

4.4.2 Analysis of Variance

Generalizing the simple linear case, we can reparametrize the conditional model as follows:

$$Y_i = \mu + \sum_{j=1}^{p-1} \beta_j (x_{j,i} - \overline{x}_{j,\bullet}) + \epsilon_i \text{ for } i = 1, 2, \ldots, N$$

where

1. *Overall Mean:* μ is the overall mean,

$$\mu = E(\overline{Y}) = E\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) = \beta_0 + \beta_1 \overline{x}_{1,\bullet} + \beta_2 \overline{x}_{2,\bullet} + \cdots + \beta_{p-1}\overline{x}_{p-1,\bullet}$$

   where $\overline{x}_{j,\bullet}$ is the mean of the j^{th} predictor, for each j.

2. *Differential Effects:* The quantity $\sum_{j=1}^{p-1} \beta_j (x_{j,i} - \overline{x}_{j,\bullet})$ is the deviation of the i^{th} observation's mean from the overall mean, or the i^{th} differential effect.

   The sum of the differential effects is zero: $\sum_{i=1}^{N} \sum_{j=1}^{p-1} \beta_j (x_{j,i} - \overline{x}_{j\bullet}) = 0$.

Question: Can you see why $\sum_{i=1}^{N} \sum_{j=1}^{p-1} \beta_j (x_{j,i} - \overline{x}_{j\bullet}) = 0$?

Predictive value. The first step in a regression analysis is to determine if the proposed collection of predictors have any predictive value. This is equivalent to testing the null hypothesis

$$\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

versus the general alternative that not all β's are zero.

We organize several computations into an initial *analysis of variance table*:

	df	*Sum of Squares*	*Mean Square*
Model:	$p-1$	$SS_m = \sum_{i=1}^{N}(\widehat{Y}_i - \overline{Y})^2$	$MS_m = \frac{1}{p-1}SS_m$
Error:	$N-p$	$SS_e = \sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2$	$MS_e = \frac{1}{N-p}SS_e$
Total:	$N-1$	$SS_t = \sum_{i=1}^{N}(Y_i - \overline{Y})^2$	

In this table,

1. The total sum of squares (SS_t) is equal to the sum of the model sum of squares (SS_m) and the error sum of squares (SS_e):

$$SS_t = SS_m + SS_e.$$

2. The error mean square (MS_e) is the same as the estimator of the common variance. Thus,

$$E(MS_e) = E\left(\frac{1}{N-p}SS_e\right) = \sigma^2.$$

3. The model mean square can be written in terms of the estimated β-coefficients:

$$MS_m = \frac{1}{p-1}\sum_{i=1}^{N}(\widehat{Y}_i - \overline{Y})^2 = \frac{1}{p-1}\sum_{i=1}^{N}\left(\sum_{j=1}^{p-1}\widehat{\beta}_j(x_{j,i} - \overline{x}_{j,\cdot})\right)^2.$$

   Under the null hypothesis, $E(MS_m) = \sigma^2$; otherwise, $E(MS_m)$ is larger than σ^2.

Note: In fact, $E(MS_m)$ equals σ^2 plus a "TERM," where the "TERM" can be shown to be positive if at least $\beta_i \neq 0$, and 0 otherwise.

To determine if the proposed predictors have some predictive value, we work with the ratio of the model mean square to the error mean square,

$$F = MS_m/MS_e.$$

Large values of F favor the alternative hypothesis that the proposed predictors have some predictive value. Under the null hypothesis of no predictive value, F has an f ratio distribution.

Theorem (Distribution Theorem). Under the general assumptions of this section, if $\beta_i = 0$ for $i = 1, \ldots, p-1$, then the ratio

$$F = MS_m/MS_e$$

has an f ratio distribution with $(p-1)$ and $(N-p)$ degrees of freedom.

Example from page 31, continued. Assume that the data from the toxicology study are the values of independent random variables satisfying the linear regression model given in the problem, and that we are interested in testing the null hypothesis that dosage and square-dosage have no predictive value in predicting weight change versus the alternative that they have some predictive value at the 5% significance level.

The following is an analysis of variance table for the test:

	df	*SS*	*MS*	*F*	*p value*
Model:	2	1557.05	778.527	82.299	≈ 0
Error:	47	444.61	9.460		
Total:	49	2001.66			

Since the p value is virtually zero, the results suggest that dosage and square-dosage have predictive value.

4.4.3 Coefficient of Determination

As in the simple linear case, the *coefficient of determination* is the ratio of the model sum of squares to the total sum of squares:

$$R^2 = \frac{SS_m}{SS_t}.$$

The coefficient of determination is the proportion of the total variation in responses that is "explained" by the model for conditional means.

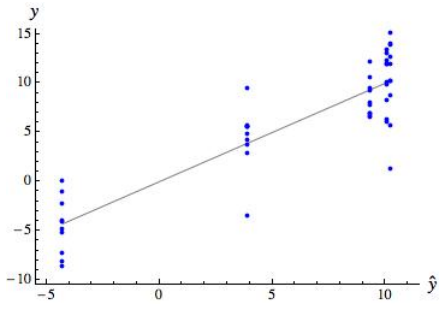
In the multiple predictor setting, the coefficient of determination is equivalent to the square of the correlation between predicted and observed responses.

Example from page 31, continued. The scatter plot below shows response pairs $(\widehat{y}, y)$ for the data from the toxicology study.

For these data,

(1) the correlation between the predicted and observed responses is 0.882, and

(2) The coefficient of determination is 0.778.



Thus, the estimated linear model explains about 77.8% of the variation in weight change.

4.4.4 Inference for β Parameters

Let $\widehat{\beta}_j$, $j = 0, \ldots, p-1$, be the ML estimators of the coefficients in the linear model.

The Covariance Matrix theorem (page 30) implies that the diagonal elements of

$$Cov\left(\underline{\widehat{\beta}}, \underline{\widehat{\beta}}\right) = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \quad \text{are the variances of the } \widehat{\beta}_j\text{'s.}$$

For simplicity, let

$$\sigma^2 v_0,\ \sigma^2 v_1,\ \ \ldots,\ \sigma^2 v_{p-1} \quad \text{be these diagonal elements.}$$

Then

> ***Theorem (Distribution Theorem).*** Under the general conditions of this section, the statistics
>
> $$T_j = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{S^2\ v_j}}, \quad j = 0, \ldots, p-1,$$
>
> have Student t distributions with $(N - p)$ degrees of freedom. In these formulas, S^2 is the unbiased estimator of σ^2 from page 29, and the v_j's are defined above.

Example from page 31, continued. For the toxicology study,

$$(\boldsymbol{X}^T \boldsymbol{X})^{-1} = \left(\begin{bmatrix} 50.00 & 155.00 & 862.50 \\ 155.00 & 862.50 & 5558.75 \\ 862.50 & 5558.75 & 38060.60 \end{bmatrix} \right)^{-1} = \begin{bmatrix} 0.072399 & -0.041502 & 0.004421 \\ -0.041502 & 0.043536 & -0.005418 \\ 0.004421 & -0.005418 & 0.000717 \end{bmatrix},$$

and the variances of the β-coefficients are

$$Var(\widehat{\beta}_0) = 0.072399\sigma^2, \quad Var(\widehat{\beta}_1) = 0.043536\sigma^2, \quad Var(\widehat{\beta}_2) = 0.000717\sigma^2.$$

Further, the following table gives estimated parameters and estimated standard errors, as well as observed t statistics and observed significance levels for two-sided tests of the null hypotheses that $\beta_i = 0$, for $i = 0, 1, 2$.

	Estimate	*Estimated SE*	*T Statistic*	*P Value*
β_0	10.2475	0.872572	12.3826	≈ 0
β_1	0.05342	0.641747	0.0832	0.934012
β_2	−0.2658	0.082378	−3.2266	0.002284

Any observations?

4.4.5 Inference for Mean Response Parameters

Let

$$\underline{x}_o = \begin{bmatrix} 1 & x_{1,o} & x_{2,o} & \cdots & x_{p-1,o} \end{bmatrix}^T \text{ be a new predictors vector,}$$

$(\underline{x}_o, Y_o)$ be a new predictors-response case, and

$$E(Y_o) = \beta_0 + \beta_1 x_{1,o} + \beta_2 x_{2,o} + \cdots + \beta_{p-1} x_{p-1,o} = \underline{\beta} \cdot \underline{x}_o$$

be the mean response at $\underline{x}_o$. Then $E(Y_o)$ can be estimated using the following statistic:

$$\widehat{Y}_o = \widehat{\underline{\beta}} \cdot \underline{x}_o = \underline{x}_o^T \widehat{\underline{\beta}} = \underline{x}_o^T \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \underline{Y}.$$

This estimator is a normal random variable with mean $E(Y_o)$ and variance

$$Var(\widehat{Y}_o) = \underline{x}_o^T \; Cov(\widehat{\underline{\beta}}, \widehat{\underline{\beta}}) \; \underline{x}_o = \sigma^2 \left(\underline{x}_o^T \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \underline{x}_o \right) = \sigma^2 \; v_o.$$

These facts can be used to prove the following distribution theorem:

> ***Theorem (Distribution Theorem).*** Under the general assumptions of this section, the statistic
>
> $$T = \frac{\widehat{\underline{\beta}} \cdot \underline{x}_o - \underline{\beta} \cdot \underline{x}_o}{\sqrt{S^2 \; v_o}}$$
>
> has a Student t distribution with $(N - p)$ degrees of freedom, where S^2 is the unbiased estimator of σ^2 from page 29, and v_o is as defined above.

Example from page 31, continued. Suppose we would like to construct a 95% confidence interval for the mean WC when the dosage level is 4 hundred mg/kg/day.

1. The predicted response is

$$\widehat{y}_o = 10.2475 + 0.05342(4) - 0.2658(16) = 6.20833.$$

2. The value of v_o in the formula for $Var(\widehat{Y}_o)$ is

$$v_o = \begin{bmatrix} 1 & 4 & 16 \end{bmatrix} \begin{bmatrix} 0.072399 & -0.041502 & 0.004421 \\ -0.041502 & 0.043536 & -0.005418 \\ 0.004421 & -0.005418 & 0.000717 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 16 \end{bmatrix} = 0.0685755.$$

3. Since $t_{47}(0.025) = 2.012$, the 95% confidence interval for the mean WC is

$$6.20833 \pm (2.012)\sqrt{3.0757^2(0.0685755)} \;\Rightarrow\; [4.5878, 7.8289].$$

   Thus, with 95% confidence, we believe that the mean WC for rats given a dosage of 4 hundred mg/kg/day for 2 weeks is between about 4.6 and 7.8 grams.

4.4.6 *Mathematica* Example: Birthweight Study

This example uses data on 56 normal births (births with no complications) at a Wellington, New Zealand, hospital (*Source:* Cook & Weisberg, 1994). The example will be discussed further in the handout `LMExample1.pdf`.

The response variable of interest is the baby's birth weight in pounds (BWT), the predictors are the mother's age in years, and the term (gestational age) in weeks of the newborn, and the hypothesized linear model is

$$\text{BWT} = \beta_o + \beta_1 \text{ age} + \beta_2 \text{ term} + \epsilon.$$

The following table gives summaries of the three variables:

Mother's Age (yrs):	*Term (wks):*	*Birthweight (lbs):*
$\overline{x}_1 = 26.29,\ s_1 = 4.60$	$\overline{x}_2 = 39.57,\ s_2 = 1.71$	$\overline{y} = 7.52,\ s_y = 1.22$

The following table gives correlations between each pair of variables:

Age-Birthweight	*Term-Birthweight*	*Age-Term*
$r = 0.4308$	$r = 0.4634$	$r = 0.0619$

Linear least squares fit equation:

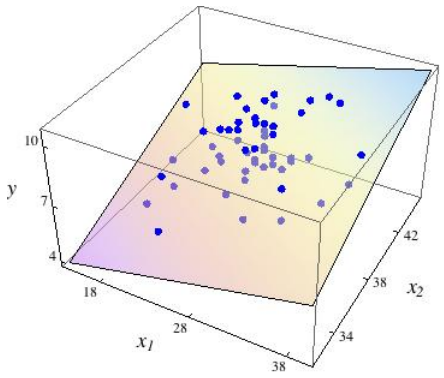
$$y = -7.5946 + 0.1068x_1 + 0.3111x_2$$

Estimated common variance:

$$s^2 = 0.9783^2$$

Coefficient of determination:

$$r^2 = 0.37712$$



4.4.7 *Mathematica* Example: Timber Yield Study

This example uses data on 31 black cherry trees (*Source:* Hand et al, 1994). The example will be discussed further in the handout `LMExample2.pdf`.

As part of a study to estimate the volume of a tree (and therefore its yield) given its diameter and height, data were collected on the volume (in cubic feet), diameter at 54 inches above the ground (in inches), and height (in feet) of 31 black cherry trees in the Allegheny National Forest. Since a multiplicative relationship is expected, the hypothesized linear model is

$$\text{log-Volume} = \beta_0 + \beta_1 \text{ log-Diameter} + \beta_2 \text{ log-Height} + \epsilon.$$

The following table gives summaries of the three variables on the original scale:

Diameter (in):	*Height (ft):*	*Volume (cubic-ft):*
$\overline{d} = 13.25$, $s_d = 3.14$	$\overline{h} = 76.0$, $s_h = 6.37$	$\overline{v} = 30.17$, $s_v = 16.44$

The following table gives correlations on the log-scale:

LogDiameter-LogVolume:	*LogHeight-LogVolume:*	*LogDiameter-LogHeight:*
$r = 0.9767$	$r = 0.6486$	$r = 0.5302$

Linear least squares fit equation:

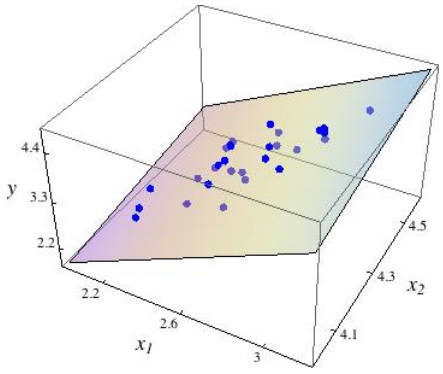
$$y = -6.632 + 1.983x_1 + 1.117x_2$$

Estimated common variance:

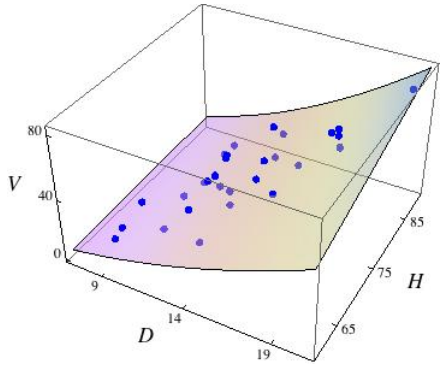
$$s^2 = 0.0814^2$$

Coefficient of determination:

$$r^2 = 0.9777$$



Equation on the original scale:



4.5 Additional Topics

The topics we have seen in this notebook can be generalized in many ways.

This section gives a brief introduction to two ideas where bootstrap analysis is particularly well-suited. In each case, we will focus on *simple* (one predictor) models.

4.5.1 Unconditional Least Squares Analysis and the Bootstrap

Consider the simple linear model

$$Y = \alpha + \beta X + \epsilon,$$

where X and ϵ are independent random variables, and where the distribution of ϵ has mean 0 and standard deviation σ.

Given a random sample from the joint (X, Y) distribution, and conditional on the observed X's, least squares estimators of slope and intercept are:

$$\widehat{\beta} = \sum_{i=1}^{N} \left(\frac{(x_i - \overline{x})}{S_{xx}} \right) Y_i \quad \text{and} \quad \widehat{\alpha} = \overline{Y} - \widehat{\beta}\, \overline{x}, \quad \text{where } S_{xx} = \textstyle\sum_{i=1}^{N} (x_i - \overline{x})^2.$$

The variability of the X_i's are ignored in this conditional analysis.

In unconditional least squares analysis, we attempt to find statistical procedures that take the variability of the X_i's into account. Using the work we did at the beginning of this notebook for the simple linear model, we know that

$$E(\widehat{\beta} | \underline{X} = \underline{x}) = \beta \quad \text{and} \quad Var(\widehat{\beta} | \underline{X} = \underline{x}) = \frac{\sigma^2}{S_{xx}},$$

where I have added "conditional on $\underline{X} = \underline{x}$" to emphasize that the results are conditional on the observed values of the predictor variable.

Using properties of expectation, it is possible to show the following unconditional results:

$$E(\widehat{\beta}) = \beta \quad \text{and} \quad Var(\widehat{\beta}) = \sigma^2 E\left(\frac{1}{\sum_{i=1}^{N} (X_i - \overline{X})^2} \right).$$

- Since the conditional expectation of $\widehat{\beta}$ does not depend on the observed x's, the unconditional expectation does not depend on them either.

- Since the conditional variance of $\widehat{\beta}$ depends on the observed x's, the unconditional variance has an additional factor that may be hard to estimate.

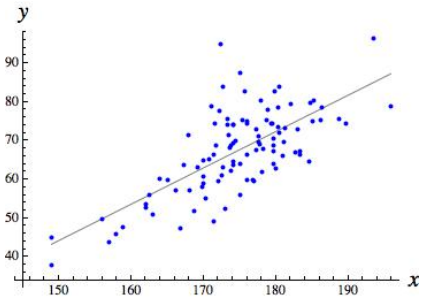
Bootstrap Application: Slope Analysis. Let $(X_1, Y_1), \ldots, (X_N, Y_N)$ be a random sample from a joint distribution satisfying the simple linear model of this section

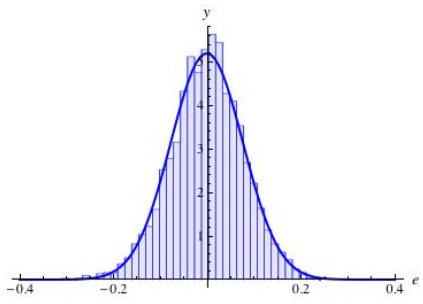
$$Y = \alpha + \beta X + \epsilon,$$

and let $\widehat{\beta}$ be the least squares estimator of β. The bootstrap can be used to study properties of $\widehat{\beta}$, and to construct confidence intervals for β, in the unconditional setting. In this method, random resampling is done from the observed distribution of the pairs.

Example (Cook & Weisberg, 1994). The data summarized below are the heights in centimeters (x) and the weights in kilograms (y) of 100 female Australian athletes.

Heights (cm):	*Weights (kg):*
$\overline{x} = 174.59,\ s_x = 8.24$	$\overline{y} = 67.34,\ s_y = 10.92$





1. *Left Plot:* The left plot is a scatter plot of height-weight pairs. The left plot includes the least squares fit line $y = -96.5332 + 0.9386x$.

2. *Right Plot:* The right plot shows an estimated error distribution, based on 5000 random resamples from the observed distribution. A normal density curve with mean -0.0011 (the bootstrap estimate of bias) and standard deviation 0.0768 (the bootstrap estimate of standard error) is superimposed.

   For this analysis, approximate 95% confidence intervals for β are as follows:

95% Normal Approximation Interval:	*95% Basic Bootstrap Interval:*
$[0.7891, 1.0903]$	$[0.7899, 1.0953]$

A 95% improved bootstrap interval, based 5000 random resamples, is $[0.7836, 1.0823]$.

An interpretation of the confidence interval is as follows (please complete):

4.5.2 Locally Weighted Least Squares and the Bootstrap

Consider the simple model $\boxed{Y = g(X) + \epsilon}$, where

- g is a differentiable function of unknown form,
- X and ϵ are independent random variables, and
- the distribution of ϵ has mean 0 and standard deviation σ.

Since differentiable functions are locally linear, Cleveland (1979)[3] proposed using a technique known as locally weighted least squares to estimate the form of g from sample data.

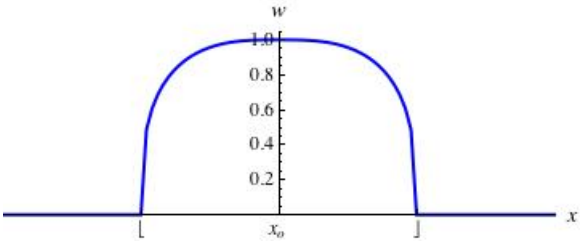
For a given x_o, data pairs whose x-values are close to x_o are used to estimate the tangent line at x_o

$$y = \alpha + \beta x = g(x_o) + \beta(x - x_o)$$

and the value at x_o is returned. The process is repeated for each observed x.

100p% lowess smooth curve. To implement Cleveland's technique, the user chooses a proportion p. Then, for a given value of the predictor variable, say x_o,

1. The Np data pairs whose x-coordinates are closest to x_o are assigned weights that smoothly decrease from 1 to 0 as you move further from x_o in either direction.
2. The remaining $N(1-p)$ pairs are assigned weights of 0.



3. The slope and intercept of the tangent line to the curve $y = g(x)$ when $x = x_o$ are estimated by minimizing the following quadratic function of two variables:

$$S(\alpha, \beta) = \sum_{i=1}^{N} w_i(y_i - (\alpha + \beta x_i))^2,$$

   where the collection $\{w_i\}$ are chosen as described in steps 1 and 2.

4. $g(x_o)$ is estimated by $\widehat{\alpha} + \widehat{\beta} x_o$, where $\widehat{\alpha}$ and $\widehat{\beta}$ are the estimates obtained in step 3.

The process is repeated for each observed value of the predictor, and successive points are connected by line segments. The estimated curve will vary depending on the choice of p.

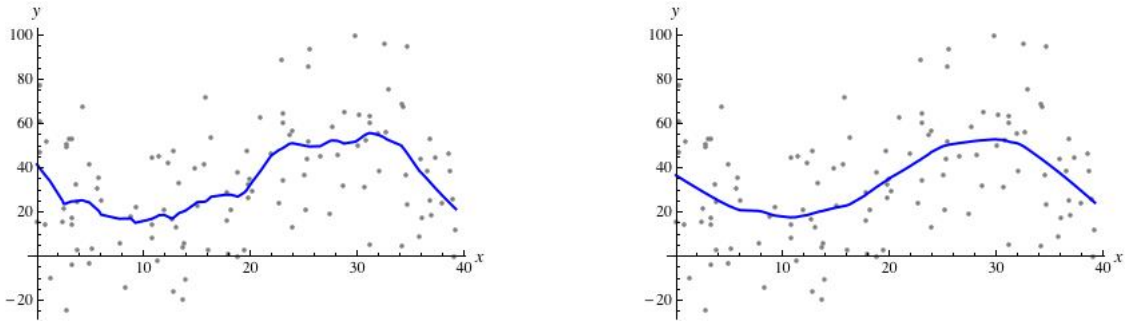
[3] William S. Cleveland, "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 1979, 74:829-836.

Example. I used the computer to generate 120 pseudo-observations from the following conditional model,

$$Y_i = 55 - 9x_i + 0.6x_i^2 - 0.01x_i^3 + \epsilon_i, \quad i = 1, 2, \ldots, 120.$$

The x_i's were chosen uniformly from the interval $(0, 40)$, and the ϵ_i's were chosen from a normal distribution with mean 0 and standard deviation 25.

The *left plot* below shows a 20% lowess smooth curve superimposed on a scatter plot of the simulated pairs, and the *right plot* shows a 35% lowess smooth curve.



In the left plot, each estimated $g(x_i)$ is based on 24 points (20% of 120). In the right plot, each estimated $g(x_i)$ is based on 42 points (35% of 120). Both curves find the cubic pattern; the curve on the right is smoother.

Footnote: Weighted Least Squares Analysis. In weighted least squares analysis, the user chooses a collection of weights $\{w_i\}$ to be used to find α and β by minimizing

$$S(\alpha, \beta) = \sum_{i=1}^{N} w_i(y_i - (\alpha + \beta x_i))^2.$$

Researchers use this technique for a variety of reasons. For example,

1. *Robust Least Squares Analysis:* In robust least squares analysis, "typical" observations are given more weight, and "non-typical" observations (that is, outliers) are given less weight. In so doing, you down weight the potential influence that a point whose x-value is far from the center might have on the analysis.

2. *Locally Weighted Least Squares Analysis:* In locally weighted least squares analysis, non-zero weights are used in small windows around each point to estimate a nonlinear curve by estimating tangent lines at each point, and weights are allowed to decrease smoothly as you move to the endpoints of each window.

Example (Efron & Tibshirani, 1993). High levels of serum cholesterol, a blood fat, have been associated with increased risk of coronary artery disease.

As part of a study of the effectiveness of a drug designed to lower serum cholesterol levels, 164 men between the ages of 35 and 59 and with initial serum cholesterol of 265 milligrams per deciliter (mg/dL) or more were assigned to receive the treatment. After a fixed period of time, serum cholesterol was measured again and the cholesterol reduction,

$$\text{CR} = \text{cholesterol before treatment} - \text{cholesterol after treatment},$$

was recorded.

The men were supposed to take six packets of the drug per day, but many of them actually took much less. Let x be the percentage of drug actually taken over the study period (the subject's *compliance* with the treatment protocol).

Assume that cholesterol reduction and compliance are related as follows:

$$\text{CR} = g(X) + \epsilon,$$

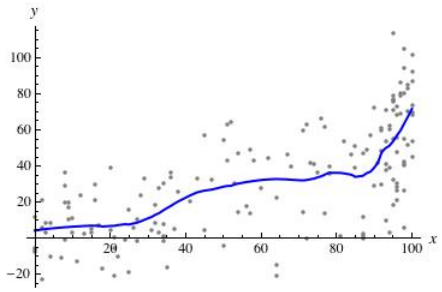
where the compliance and error distributions are independent, the error distribution has mean 0 and finite standard deviation, and g is a differentiable function.

The plot below shows a 30% lowess smooth curve superimposed on the (x,CR) pairs.

For each observed x, the pair

$$(x, a_x + b_x x)$$

is plotted, where the estimates of intercept and slope at x are obtained using a weighted least squares analysis based on 49 (30% of 164) data pairs.



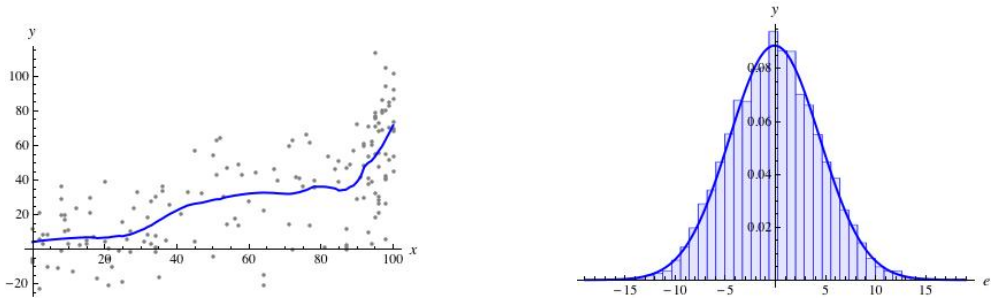
The plot suggests (please complete)

Bootstrap Application: Mean Response Analysis. Let $(X_1, Y_1), \ldots, (X_N, Y_N)$ be a random sample from a joint distribution satisfying the simple model of this section,

$$Y = g(X) + \epsilon,$$

let θ be the mean response at a fixed value of the predictor variable, and let $\widehat{\theta}$ be the estimator of θ based on a $100p\%$ lowess smooth. The bootstrap can be used to study the properties of $\widehat{\theta}$ and to construct confidence intervals for θ.

Continuing with the 30% lowess smooth of the compliance-CR data from page 43, suppose we are interested in constructing a 95% confidence interval for the mean cholesterol reduction for men with 90% compliance.



1. *Left Plot:* The left plot is a scatter plot of compliance-CR pairs, including a 30% lowess smooth curve. For $x = 90$ (that is, for 90% compliance), the estimated mean response is a reduction in cholesterol of 39.0012 mg/dL.

2. *Right Plot:* The right plot shows an estimated error distribution, based on 5000 random resamples from the observed distribution. A normal density curve with mean -0.0767 (the bootstrap estimate of bias) and standard deviation 4.4963 (the bootstrap estimate of standard error) is superimposed.

   For this analysis, approximate 95% confidence intervals for θ are as follows:

95% Normal Approximation Interval:	*95% Basic Bootstrap Interval:*
$[30.2653, 47.8905]$	$[30.3457, 47.6664]$

A 95% improved bootstrap interval, based 5000 random resamples, is $[30.5979, 47.5502]$.

An interpretation of the confidence interval is as follows (please complete):

4.6 Linear Algebra Review

This brief review uses the standard notation from linear algebra.

4.6.1 Linear Algebra: Solving Linear Systems

In linear algebra, we work with matrix equations of the form $A\boldsymbol{x} = \boldsymbol{b}$, where

- A is an m-by-n *coefficient matrix*,
- $\boldsymbol{x}$ is an n-by-1 *vector of unknowns*, and
- $\boldsymbol{b}$ is an m-by-1 *vector of values*.

Matrix equations arise naturally when working with systems of linear equations.

For example, the 3-by-3 system of linear equations

$$\begin{array}{rrrcr} x_1 & -2x_2 & +\,x_3 & = & 0 \\ & 2x_2 & -8x_3 & = & 8 \\ -4x_1 & +5x_2 & +9x_3 & = & -9 \end{array}$$

corresponds to the matrix equation $A\boldsymbol{x} = \boldsymbol{b}$ where

$$A = \begin{bmatrix} 1 & -2 & 1 \\ 0 & 2 & -8 \\ -4 & 5 & 9 \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \text{and } \boldsymbol{b} = \begin{bmatrix} 0 \\ 8 \\ -9 \end{bmatrix}.$$

The *solution set* of a linear system is the set of all possible solutions to the system. Two linear systems are said to be *equivalent* if each has the same solution set.

Solutions can be obtained by using a sequence of elementary row operations on the augmented matrix. Each elementary operation produces an equivalent system with the same solution set.

For example, the following equivalent augmented matrices imply that the unique solution to the system above is the vector whose coordinates are 29, 16, 3:

$$\left[\begin{array}{c|c} A & \boldsymbol{b} \end{array}\right] = \left[\begin{array}{ccc|c} 1 & -2 & 1 & 0 \\ 0 & 2 & -8 & 8 \\ -4 & 5 & 9 & -9 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & -2 & 1 & 0 \\ 0 & 2 & -8 & 8 \\ 0 & -3 & 13 & -9 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & -2 & 1 & 0 \\ 0 & 1 & -4 & 4 \\ 0 & 0 & 1 & 3 \end{array}\right]$$

$$\sim \left[\begin{array}{ccc|c} 1 & -2 & 0 & -3 \\ 0 & 1 & 0 & 16 \\ 0 & 0 & 1 & 3 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 0 & 0 & 29 \\ 0 & 1 & 0 & 16 \\ 0 & 0 & 1 & 3 \end{array}\right] \quad \Rightarrow \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 29 \\ 16 \\ 3 \end{bmatrix}$$

4.6.2 Linear Algebra: Coefficient Matrices

Let $A = \begin{bmatrix} \boldsymbol{a_1} & \boldsymbol{a_2} & \cdots & \boldsymbol{a_n} \end{bmatrix}$ be an m-by-n coefficient matrix.

1. *Column Space:* The column space of the matrix A is the collection of all $\boldsymbol{b}$ with the property that $A\boldsymbol{x} = \boldsymbol{b}$ is a consistent equation:

$$Col(A) = \{\boldsymbol{b} \ : \ \boldsymbol{b} = A\boldsymbol{x} \text{ for some } \boldsymbol{x} \in \mathbf{R}^n\} = \text{Span}\{\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_n}\} \subseteq \mathbf{R}^m.$$

   Each solution can be thought of as a linear combination of the columns of A.

2. *Null Space:* The null space of the matrix A is the collection of all $\boldsymbol{x}$ satisfying the homogeneous system $A\boldsymbol{x} = \boldsymbol{O}$:

$$Null(A) = \{\boldsymbol{x} \ : \ A\boldsymbol{x} = \boldsymbol{O}\} \subseteq \mathbf{R}^n.$$

   In these equations, $\boldsymbol{O}$ is the vector of all zeroes.

3. *Rank:* The rank of the matrix A is the dimension of the column space of A:

$$rank(A) = dim(Col(A)).$$

   Since the pivot columns of A form a basis for the column space, the rank of A equals the number of pivot columns of A.

4. *Nullity:* The nullity of the matrix A is the dimension of the null space of A:

$$nullity(A) = dim(Null(A)).$$

   The dimension of the null space is the number of non-pivot columns of A.

These facts imply that $rank(A) + nullity(A) = n$.

For example, consider the following 3-by-5 matrix A and its equivalent reduced echelon form:

$$A = \begin{bmatrix} \boldsymbol{a_1} & \boldsymbol{a_2} & \boldsymbol{a_3} & \boldsymbol{a_4} & \boldsymbol{a_5} \end{bmatrix} = \begin{bmatrix} 0 & 3 & -6 & 6 & 4 \\ 3 & -7 & 8 & -5 & 8 \\ 3 & -9 & 12 & -9 & 6 \end{bmatrix} \sim \cdots \sim \begin{bmatrix} 1 & 0 & -2 & 3 & 0 \\ 0 & 1 & -2 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

1. Columns 1, 2 and 5 are pivot columns. Thus, $\{\boldsymbol{a_1}, \boldsymbol{a_2}, \boldsymbol{a_5}\}$ is a basis for $Col(A)$.

2. Since

$$A\boldsymbol{x} = \boldsymbol{O} \ \Rightarrow \ \begin{array}{c} x_1 = 2x_3 - 3x_4 \\ x_2 = 2x_3 - 2x_4 \\ x_3 \text{ is free} \\ x_4 \text{ is free} \\ x_5 = 0 \end{array} \ \Rightarrow \ \boldsymbol{s} = x_3 \begin{bmatrix} 2 \\ 2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} -3 \\ -2 \\ 0 \\ 1 \\ 0 \end{bmatrix} = x_3\boldsymbol{w_1} + x_4\boldsymbol{w_2}$$

   is a solution to the homogeneous system, $\{\boldsymbol{w_1}, \boldsymbol{w_2}\}$ is a basis for $Null(A)$.

3. Thus, $rank(A) + nullity(A) = 3 + 2 = 5$.

4.6.3 Linear Algebra: Inner Product and Orthogonality

Let $\boldsymbol{v}$ and $\boldsymbol{w}$ be vectors in $\mathbf{R}^k$, and let $V \subseteq \mathbf{R}^k$ be a subspace.

1. *Inner Product:* The inner product (or dot product) of $\boldsymbol{v}$ and $\boldsymbol{w}$ is the number

$$\boldsymbol{v}\cdot\boldsymbol{w} = \boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{w} = \begin{bmatrix} v_1 & v_2 & \cdots & v_k \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix} = v_1w_1 + v_2w_2 + \cdots + v_kw_k.$$

2. *Length:* The length of $\boldsymbol{v}$ is the number $\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v}\cdot\boldsymbol{v}} = \sqrt{v_1^2 + v_2^2 + \cdots + v_k^2}$.

3. *Distance:* The distance between $\boldsymbol{v}$ and $\boldsymbol{w}$ is the length of the difference vector $\boldsymbol{v} - \boldsymbol{w}$:

$$dist(\boldsymbol{v}, \boldsymbol{w}) = \|\boldsymbol{v} - \boldsymbol{w}\| = \sqrt{(v_1 - w_1)^2 + (v_2 - w_2)^2 + \cdots + (v_k - w_k)^2}.$$

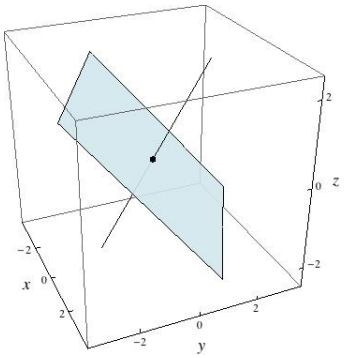
4. *Orthogonal:* The vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ are said to be orthogonal if their inner product is zero.

5. *Orthogonal Complement:* The orthogonal complement of V, denoted by $V^\perp$ ("V-perp"), is the collection of all vectors orthogonal to V:

$$V^\perp = \{\boldsymbol{w} \ : \ \boldsymbol{w} \text{ is orthogonal to each } \boldsymbol{v} \in V\}.$$

   The orthogonal complement of V is itself a subspace, and $dim(V) + dim(V^\perp) = k$.

For example, let $V = \text{Span}\{\boldsymbol{v}\} = \text{Span}\left\{\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}\right\} \subset \mathbf{R}^3$.
(V is represented as the line in the plot to the right.)



Given $\boldsymbol{w} \in V^\perp$, we know that

$$\boldsymbol{w}\cdot\boldsymbol{v} = w_1 + 2w_2 + 3w_3 = 0.$$

Since $w_1 = -2w_2 - 3w_3$, elements of the orthogonal complement of V must have the following form:

$$\boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} -2w_2 - 3w_3 \\ w_2 \\ w_3 \end{bmatrix} = w_2 \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} + w_3 \begin{bmatrix} -3 \\ 0 \\ 1 \end{bmatrix}$$

where w_2 and w_3 are free variables. Thus,

$$V^\perp = \text{Span}\left\{\begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -3 \\ 0 \\ 1 \end{bmatrix}\right\}.$$

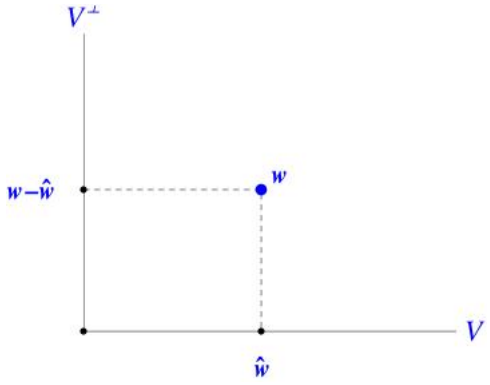
($V^\perp$ is represented as the plane in the plot to the right above. The combined set of basis elements of V and $V^\perp$ form a basis for $\mathbf{R}^3$.)

4.6.4 Linear Algebra: Orthogonal Decomposition Theorem

Theorem. Let $V \subset \mathbf{R}^k$ be a subspace of $\mathbf{R}^k$, and let $\boldsymbol{w} \in \mathbf{R}^k$ be a vector. Then there exists a unique vector $\widehat{\boldsymbol{w}} \in V$ (called the *orthogonal projection* of $\boldsymbol{w}$ onto V) satisfying the following properties:

1. The vector $\widehat{\boldsymbol{w}}$ is the *closest* point in V to $\boldsymbol{w}$.
2. The difference $(\boldsymbol{w} - \widehat{\boldsymbol{w}})$ is a vector in $V^{\perp}$.

The theorem is illustrated to the right.



<u>*Note:*</u> $\widehat{\boldsymbol{w}}$ can be found using matrix multiplication: $\widehat{\boldsymbol{w}} = P\boldsymbol{w}$, for some projection matrix P. I will show you how to find P in the linear least squares setting.

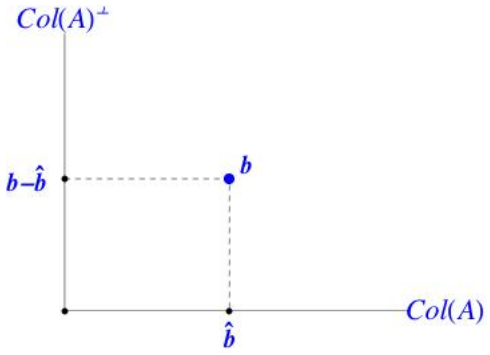
4.6.5 Linear Algebra: Fundamental Theorem of Linear Algebra

Theorem. Let A be an $m \times n$ matrix. Then

1. $Null(A)$ and $Col(A^T)$ are orthogonal complements in $\mathbf{R}^n$.
2. $Null(A^T)$ and $Col(A)$ are orthogonal complements in $\mathbf{R}^m$.
3. Matrices A and A^T have equal ranks: $rank(A) = rank(A^T)$.

<u>*Application to least squares analysis.*</u> In the application of the Fundamental Theorem of Linear Algebra to least squares analysis:

1. The system $A\boldsymbol{x} = \boldsymbol{b}$ is inconsistent. Thus, $\boldsymbol{b}$ is not in the column space of A.
2. Let $\widehat{b} = P\boldsymbol{b}$ be the orthogonal projection of $\boldsymbol{b}$ onto the column space of A.
3. The solutions to the system $A\boldsymbol{x} = \widehat{\boldsymbol{b}}$ are the least squares solutions.
4. The vector $\boldsymbol{e} = \boldsymbol{b} - \widehat{\boldsymbol{b}}$ contains the residuals.



The form of the projection matrix P can be found quickly using the fact that the orthogonal complement of the column space of A is $Null(A^T)$.

4.6.6 Linear Algebra: Orthogonal Projections and Least Squares Analysis

Let A be an $m \times n$ coefficient matrix and assume that $A\boldsymbol{x} = \boldsymbol{b}$ is inconsistent. We propose to find *approximate* solutions to the system as follows:

(1) Find the projection of $\boldsymbol{b}$ onto the column space of A, $\widehat{\boldsymbol{b}}$, and
(2) Report solutions to the consistent system $A\boldsymbol{x} = \widehat{\boldsymbol{b}}$.

Two important observations are the following:

1. *Observation 1*: Since $\widehat{\boldsymbol{b}}$ is as close to $\boldsymbol{b}$ as possible, each approximate solution $\boldsymbol{x}$ satisfies

$$\|\boldsymbol{b} - \widehat{\boldsymbol{b}}\| = \|\boldsymbol{b} - A\boldsymbol{x}\| \text{ is as } \textit{small} \text{ as possible.}$$

The difference vector is

$$\boldsymbol{b} - A\boldsymbol{x} = \begin{bmatrix} b_1 - (a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n) \\ b_2 - (a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n) \\ \vdots \\ b_m - (a_{m,1}x_1 + a_{m,2}x_2 + \cdots + a_{m,n}x_n) \end{bmatrix}$$

and the square of the length of the difference vector is

$$\sum_{i=1}^{m} \left(b_i - (a_{i,1}x_1 + a_{i,2}x_2 + \cdots + a_{i,n}x_n)\right)^2.$$

Each approximate solution $\boldsymbol{x}$ will minimize the above sum of squared differences. For this reason the approximate solutions are called *least squares solutions.*

2. *Observation 2:* Since the difference vector is in the orthogonal complement of $Col(A)$,

$$(\boldsymbol{b} - \widehat{\boldsymbol{b}}) = (\boldsymbol{b} - A\boldsymbol{x}) \in (Col(A))^{\perp},$$

and since, by the Fundamental Theorem of Linear Algebra, the orthogonal complement of the column space is the null space of A^T, we know that

$$A^T(\boldsymbol{b} - \widehat{\boldsymbol{b}}) = A^T(\boldsymbol{b} - A\boldsymbol{x}) = \boldsymbol{O}.$$

Further,

$$A^T(\boldsymbol{b} - A\boldsymbol{x}) = \boldsymbol{O} \quad \Leftrightarrow \quad A^T\boldsymbol{b} - A^TA\boldsymbol{x} = \boldsymbol{O} \quad \Leftrightarrow \quad A^TA\boldsymbol{x} = A^T\boldsymbol{b}.$$

Thus, least squares solutions can be found by solving the consistent system on the right (called the *normal equation* of the system).

Note that, by using the normal equation, we do not need to explicitly find the projection of $\boldsymbol{b}$ onto the column space of A.

The following theorem gives the properties of this process:

The Least Squares Theorem. Under the conditions above,

1. $\boldsymbol{x}$ is a least squares solution to $A\boldsymbol{x} = \boldsymbol{b}$ iff $\boldsymbol{x}$ is a solution to $A^TA\boldsymbol{x} = A^T\boldsymbol{b}$.
2. A^TA is invertible iff the columns of A are linearly independent. Thus, there is a unique least squares solution iff the columns of A are linearly independent.

For example, consider the inconsistent system $A\boldsymbol{x} = \boldsymbol{b}$ where $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$ and $\boldsymbol{b} = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}$. Then

$$\left[\begin{array}{c|c} A^TA & A^T\boldsymbol{b} \end{array}\right] = \left[\begin{array}{cc|c} 3 & 3 & 6 \\ 3 & 5 & 0 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 2 & -6 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 0 & 5 \\ 0 & 1 & -3 \end{array}\right] \Rightarrow \boldsymbol{x} = \begin{bmatrix} 5 \\ -3 \end{bmatrix}$$

is the unique least squares solution to the inconsistent system above.

Projection matrix. If the columns of A are linearly independent, and $\widehat{\boldsymbol{x}}$ is a least squares solution, then

1. Since A^TA is invertible and $A^TA\widehat{\boldsymbol{x}} = A^T\boldsymbol{b}$, we know that $\widehat{\boldsymbol{x}} = (A^TA)^{-1}A^T\boldsymbol{b}$.
2. Since $\widehat{\boldsymbol{b}} = A\widehat{\boldsymbol{x}} = A(A^TA)^{-1}A^T\boldsymbol{b}$, we know that $P = A(A^TA)^{-1}A^T$.

Thus, the projection matrix is $P = A(A^TA)^{-1}A^T$.

<u>*For example,*</u> the projection matrix for the example above can be computed as follows:

(i) Compute A^TA and its inverse:

$$A^TA = \begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix} \quad \Rightarrow \quad \left(A^TA\right)^{-1} = \left(\frac{1}{6}\right)\begin{bmatrix} 5 & -3 \\ -3 & 3 \end{bmatrix}$$

(ii) Use the answer to (i) to find the projection matrix:

$$P = A(A^TA)^{-1}A^T = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}\left(\left(\frac{1}{6}\right)\begin{bmatrix} 5 & -3 \\ -3 & 3 \end{bmatrix}\right)\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} = \left(\frac{1}{6}\right)\begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix}$$

<u>*Final note:*</u>

In linear least squares analysis, $A = \boldsymbol{X}$ is the design matrix, $\boldsymbol{b} = \underline{Y}$ is the response vector, $P = \boldsymbol{H}$ is the hat matrix, and $\widehat{\boldsymbol{b}} = P\boldsymbol{b} = \underline{\widehat{Y}}$ is the vector of predicted responses.