Chapter 5

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

Chapter 3: Diagnostics

- Section 3.1: Outlying x-values can be found via boxplot (or a scatterplot!) Useful for assessing extrapolation. More advanced method in Sec. 10.3 for multiple predictors.
- Section 3.2: Recall $e_i = Y_i - \hat{Y}_i$ is ith residual. The (externally) studentized residual t_i (defined on p. 396) has a t_{n-3} distribution.
- Section 3.3: Plots to consider
  1. Plot of e_i vs. $\hat{Y}_i$ or e_i vs. x_i: nonlinearity means line $\beta_0 + \beta_1 x_i$ inappropriate. Nonconstant variance means $\text{var}(\epsilon_i) = \sigma^2$ not appropriate. Outlying observations (very large or small residuals) can indicate several potential problems (later).
  2. Histogram or boxplot of e_i, normal probability plot of e_i to check normality. Expect one outlier out of 150 observations for truly normal data in boxplot. There are also formal tests for normality (later).

5.1 Matrices

A matrix is a rectangular array of numbers. Here's an example:

$$\mathbf{A} = \begin{bmatrix} 2.3 & -1.4 & 17 \\ -22.5 & 0 & \sqrt{2} \end{bmatrix}.$$

This matrix has dimensions 2×3. The number of rows is first, then the number of columns.

We can write the $n \times p$ matrix $\mathbf{X}$ abstractly as

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix}.$$

Other notation

Another notation that is common is $\mathbf{A} = [a_{ij}]_{n \times m}$ for an $n \times m$ matrix $\mathbf{A}$ with element a_{ij} in the i^{th} row and j^{th} column.

The matrix $\mathbf{X}$ on the previous page would then be written $\mathbf{X} = [x_{ij}]_{n \times p}$.

5.1 (cont'd) Transpose

The transpose of a matrix $\mathbf{A}'$ takes the matrix $\mathbf{A}$ and makes the rows the columns and the columns the rows. Precisely, if $\mathbf{A} = [a_{ij}]_{n \times m}$ then $\mathbf{A}'$ is the $m \times n$ matrix with elements $a'_{ij} = a_{ji}$. For example:

$$\text{If } \mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}, \text{ then } \mathbf{A}' = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}.$$

Question: what is $(\mathbf{A}')'$?

5.2 Matrix addition

If two matrices $\mathbf{A} = [a_{ij}]_{n\times m}$ and $\mathbf{B} = [b_{ij}]_{n\times m}$ have the same dimensions, you can add them together, element by element, to get a new matrix $\mathbf{C} = [c_{ij}]_{n\times m}$. That is, $\mathbf{C} = \mathbf{A} + \mathbf{B}$ is the matrix with elements $c_{ij} = a_{ij} + b_{ij}$. For example,

$$\begin{bmatrix} -1 & -2 \\ 5 & 7 \\ -10 & 20 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} -1+1 & -2+2 \\ 5+3 & 7+4 \\ -10+1 & 20+2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 8 & 11 \\ -9 & 22 \end{bmatrix}.$$

5.3 Multiplying a matrix times a number

Multiplying a matrix $\mathbf{A} = [a_{ij}]$ by a number b yields the matrix $\mathbf{C} = \mathbf{A}b$ with elements $c_{ij} = a_{ij}b$. For example,

$$(-2)\begin{bmatrix} -1 & -2 \\ 5 & 7 \\ -10 & 20 \end{bmatrix} = \begin{bmatrix} -1(-2) & -2(-2) \\ 5(-2) & 7(-2) \\ -10(-2) & 20(-2) \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ -10 & -14 \\ 20 & -40 \end{bmatrix}.$$

Vectors

A vector is a matrix with only one column or row, called a "column vector" or "row vector" respectively. Here's an example of each:

$$\mathbf{x} = \begin{bmatrix} 1 \\ -1 \\ 14 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 & -1 & 14 \end{bmatrix}.$$

Note that for these vectors, $\mathbf{x}' = \mathbf{y}$ and $\mathbf{y}' = \mathbf{x}$.
The **product** of an $1 \times n$ row vector and a $n \times 1$ column vector is the sum of the pairwise products of elements. So if $\mathbf{x} = [x_i]_{1\times n}$ and $\mathbf{y} = [y_i]_{n\times 1}$ then $\mathbf{xy} = \sum_{i=1}^{n} x_i y_i$.

Inner product of two vectors

For example, if $\mathbf{x} = \begin{bmatrix} -1 & 2 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} 10 \\ -5 \end{bmatrix}$ then

$$\mathbf{xy} = \begin{bmatrix} -1 & 2 \end{bmatrix} \begin{bmatrix} 10 \\ -5 \end{bmatrix} = -1(10) + 2(-5) = -20.$$

The *inner product* of two $n \times 1$ column vectors $\mathbf{x}$ and $\mathbf{y}$ is the product

$$\mathbf{x}'\mathbf{y} = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i.$$

Euclidean distance

Note that if $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ is a point in the plane $\mathbb{R}^2$, then $\mathbf{x}'\mathbf{x} = x_1^2 + x_2^2$ is the square of the length of $\mathbf{x}$. That is, $||\mathbf{x}|| = \sqrt{\mathbf{x}'\mathbf{x}}$.

In general, for any point in $\mathbb{R}^n$, $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}$, $\mathbf{x}'\mathbf{x} = ||\mathbf{x}||^2$.

5.3 (cont'd) Matrix multiplication

We are now ready to define general matrix multiplication. The product of an $n \times p$ matrix $\mathbf{A}$ and a $p \times m$ matrix $\mathbf{B}$ is the $n \times m$ matrix $\mathbf{C}$ with elements $c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$. Let $\mathbf{A}$ be comprised of n $1 \times p$ row vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$ and let $\mathbf{B}$ be comprised of m $p \times 1$ column vectors $\mathbf{b}_1, \ldots, \mathbf{b}_m$ like

$$\mathbf{A} = \begin{bmatrix} \cdots \mathbf{a}_1 \cdots \\ \cdots \mathbf{a}_2 \cdots \\ \vdots \\ \cdots \mathbf{a}_n \cdots \end{bmatrix}_{n \times p} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \vdots & \vdots & & \vdots \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_m \\ \vdots & \vdots & & \vdots \end{bmatrix}_{p \times m}.$$

Matrix multiplication

Then $c_{ij} = \mathbf{a}_i \mathbf{b}_j$

$$\mathbf{C} = \begin{bmatrix} \mathbf{a}_1\mathbf{b}_1 & \mathbf{a}_1\mathbf{b}_2 & \cdots & \mathbf{a}_1\mathbf{b}_m \\ \mathbf{a}_2\mathbf{b}_1 & \mathbf{a}_2\mathbf{b}_2 & \cdots & \mathbf{a}_2\mathbf{b}_m \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_n\mathbf{b}_1 & \mathbf{a}_n\mathbf{b}_2 & \cdots & \mathbf{a}_n\mathbf{b}_m \end{bmatrix}.$$

For example, let $\mathbf{A} = \begin{bmatrix} 1 & -1 & -2 \\ -3 & -1 & 5 \end{bmatrix}$ and

$\mathbf{B} = \begin{bmatrix} 2 & 0 & -2 \\ 0 & -5 & 7 \\ 1 & 0.5 & -4 \end{bmatrix}$. Then

$$\begin{aligned} \mathbf{AB} &= \begin{bmatrix} 1(2) - 1(0) - 2(1) & 1(0) - 1(-5) - 2(0.5) & 1(-2) - 1(7) - 2(-4) \\ -3(2) - 1(0) + 5(1) & -3(0) - 1(-5) + 5(0.5) & -3(-2) - 1(7) + 5(-4) \end{bmatrix} \\ &= \begin{bmatrix} 0 & 4 & 0 \\ -1 & 8.5 & -21 \end{bmatrix}. \end{aligned}$$

5.4 Identity matrix

On the previous slide, does $\mathbf{BA}$ make sense? No. The rows of the first matrix must be the same length as the columns of the second. Note that, in general, $\mathbf{AB} \neq \mathbf{BA}$.

Define $\mathbf{I}_{n\times n}$ as

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Then

$$\mathbf{A}_{n\times p}\mathbf{I}_{p\times p} = \mathbf{A}_{n\times p} \text{ and } \mathbf{I}_{n\times n}\mathbf{A}_{n\times p} = \mathbf{A}_{n\times p},$$

for any $\mathbf{A}_{n\times p}$. The matrix $\mathbf{I}_{n\times n}$ is called the $n \times n$ *identity matrix*.

5.4 Identity matrix

For example,

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 & -2 \\ -3 & -1 & 5 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -2 \\ -3 & -1 & 5 \end{bmatrix}$$

and

$$\begin{bmatrix} 1 & -1 & -2 \\ -3 & -1 & 5 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -2 \\ -3 & -1 & 5 \end{bmatrix}.$$

5.6 Matrix inverse

The *inverse* of a square $(p \times p)$ matrix $\mathbf{A}$ is the $p \times p$ matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_{p\times p}$. For example, if $\mathbf{A} = \begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix}$, then

$$\begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

and so $\mathbf{A}^{-1} = \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix}$. Note that we must have

$$\begin{bmatrix} 1 & \frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

as well.

Finding a 2×2 inverse

There is an algorithm for finding the inverse of any size matrix but it is very computationally intensive, except for 2×2 matrices. Let

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

Then

$$\mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}.$$

That is, switch the diagonal entries, multiply the off-diagonals by -1, and divide the works by $a_{11}a_{22} - a_{12}a_{21}$.

We can show that this is the inverse in class. Try it out on $\mathbf{A}$ on the previous slide.

Singular matrix

Not every square matrix has an inverse. For example

$$\mathbf{A} = \begin{bmatrix} -1 & 2 \\ 2 & -4 \end{bmatrix},$$

does not. Try the formula on the previous slide out on this matrix. What happens?

Square matrices that do not have an inverse are said to be *singular*.

The two sample normal model in terms of matrices

Recall the two-sample normal model with equal variances:

$$Y_{11}, Y_{12}, \ldots, Y_{1n_1} \overset{iid}{\sim} N(\mu_1, \sigma^2),$$

$$Y_{21}, Y_{22}, \ldots, Y_{2n_2} \overset{iid}{\sim} N(\mu_2, \sigma^2).$$

We can rewrite this as

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

where

$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2),$$

where $i = 1, 2$ indexes the group (1 or 2) and $j = 1, \ldots, n_i$ is the observation within the group.

Two sample normal model

Each piece of data Y_{ij} follows:

$$
\begin{aligned}
Y_{11} &= \mu_1 + \epsilon_{11} \\
Y_{12} &= \mu_1 + \epsilon_{12} \\
Y_{13} &= \mu_1 + \epsilon_{13} \\
&\vdots \quad \vdots \quad \vdots \\
Y_{1n_1} &= \mu_1 + \epsilon_{1n_1} \\
Y_{21} &= \mu_2 + \epsilon_{21} \\
Y_{22} &= \mu_2 + \epsilon_{22} \\
Y_{23} &= \mu_2 + \epsilon_{23} \\
&\vdots \quad \vdots \quad \vdots \\
Y_{2n_2} &= \mu_2 + \epsilon_{2n_2}
\end{aligned}
$$

Two sample normal model

Define the following vectors and matrices:

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_2} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \text{ and } \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1n_2} \\ \epsilon_{21} \\ \epsilon_{22} \\ \vdots \\ \epsilon_{2n_2} \end{bmatrix}.$$

Two sample normal model

Then we can write the model succinctly as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

We'll show this on the board for $n_1 = n_2 = 3$.
It turns out the the least squares estimators (and MLE's!) for μ_1 and μ_2 are $\hat{\mu}_1 = n_1^{-1} \sum_{j=1}^{n_1} y_{1j} = \bar{y}_1$ and $\hat{\mu}_2 = n_2^{-1} \sum_{j=1}^{n_2} y_{2j} = \bar{y}_2$ are obtained in matrix terms as

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Two sample normal model

We'll show part of this:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix}.$$

Two sample normal model

And so

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{bmatrix}.$$

Also,

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_2} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{j=1}^{n_2} y_{2j} \end{bmatrix}.$$

Two sample normal model

Then

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{bmatrix} \begin{bmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{j=1}^{n_2} y_{2j} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \end{bmatrix} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix},$$

as promised. The unbiased estimate of σ^2 in terms of matrices is

$$\text{MSE} = (\mathbf{y}-\mathbf{X}\hat{\beta})'(\mathbf{y}-\mathbf{X}\hat{\beta})/(n_1+n_2-2) = ||\mathbf{y}-\mathbf{X}\hat{\beta}||^2/(n_1+n_2-2).$$

What is the point? Although the two-sample normal model is fairly simple, very complex models with multiple predictors, both categorical and continuous, can be written as

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

including the simple linear regression model, multiple regression models, oneway and multiway ANOVA models, and ANCOVA models.

The linear model

In

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\mathbf{Y}$ is the $n \times 1$ *data vector*.
- $\mathbf{X}$ is the $n \times p$ *design matrix*. Often the i^{th} row of $\mathbf{X}$ is comprised of $p-1$ measurements taken on the i^{th} subject in a study, e.g. the i^{th} row of an Excel spreadsheet, and an intercept term.
- $\boldsymbol{\beta}$ is the $p \times 1$ *coefficient vector*. For the two-sample model, $p = 2$ and $\boldsymbol{\beta} = (\mu_1, \mu_2)$.
- $\boldsymbol{\epsilon}$ is the $n \times 1$ error vector. All the elements of $\boldsymbol{\epsilon}$ are assumed to be *iid* $N(0, \sigma^2)$.

Example: Celts versus modern Englishmen

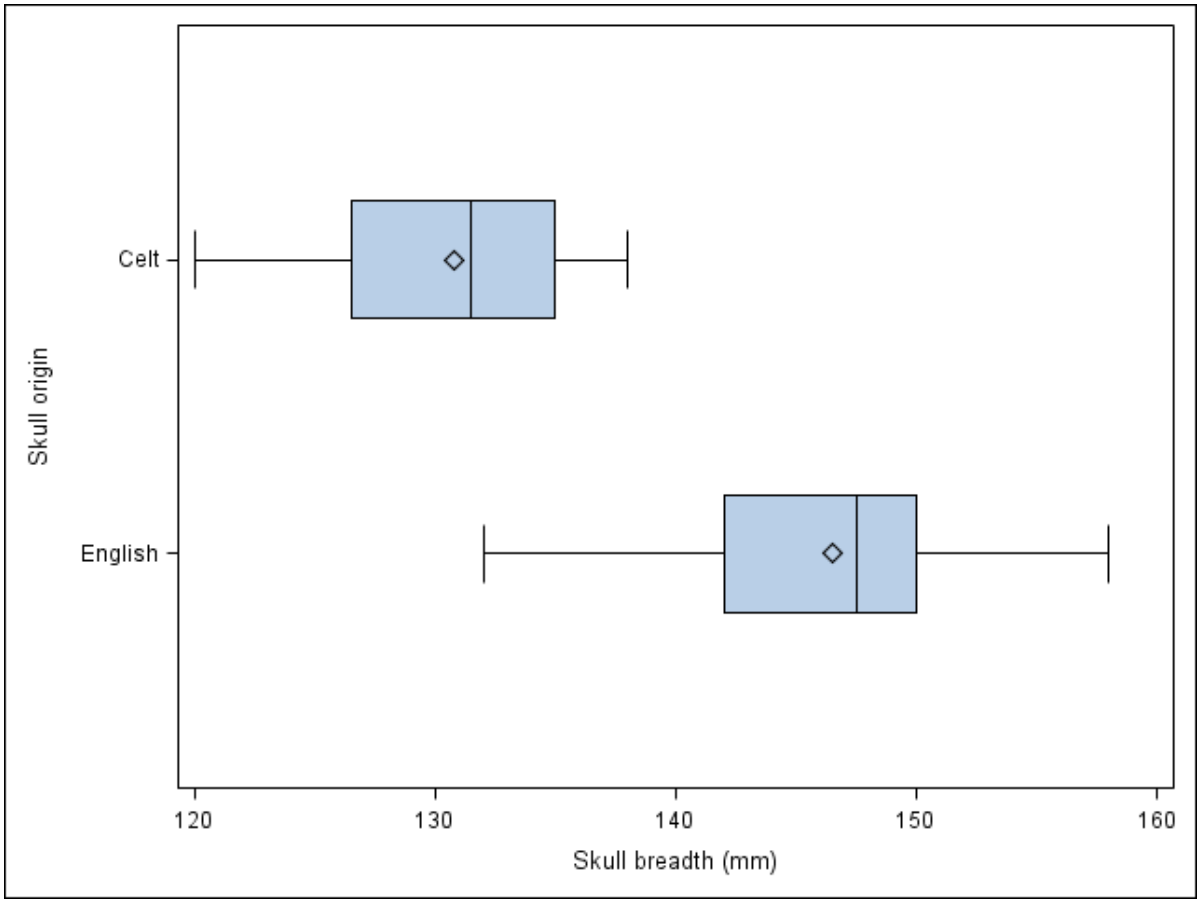
The Celts were a vigorous race of people who once populated parts of England. It is not entirely clear whether they simply died out or merged with other people who were the ancestors of those who live in England today.

The maximum head breadths (mm) were measured on $n_2 = 16$ unearthed Celtic skulls and on $n_1 = 18$ modern-day Englishmen skulls. It is of interest to determine and quantify differences in skull size between the two populations.

Look at the data...

```
data headbreadth;
input breadth group$ @@;
label breadth='Skull breadth (mm)' group='Skull origin';
datalines;
141 English 148 English 132 English 138 English 154 English 142 English 150 English
146 English 155 English 158 English 150 English 140 English 147 English 148 English
144 English 150 English 149 English 145 English 133 Celt    138 Celt    130 Celt
138 Celt    134 Celt    127 Celt    128 Celt    138 Celt    136 Celt    131 Celt
126 Celt    120 Celt    124 Celt    132 Celt    132 Celt    125 Celt
;
proc sgplot data=headbreadth;
 hbox breadth / category=group; run;
proc glm plots=diagnostics;
 class group;
 model breadth=group / noint solution;
 estimate "English-Celt" group -1 1; run;
```

Side-by-side boxplots



Spread is roughly the same, we'll fit a normal-errors model with common variance

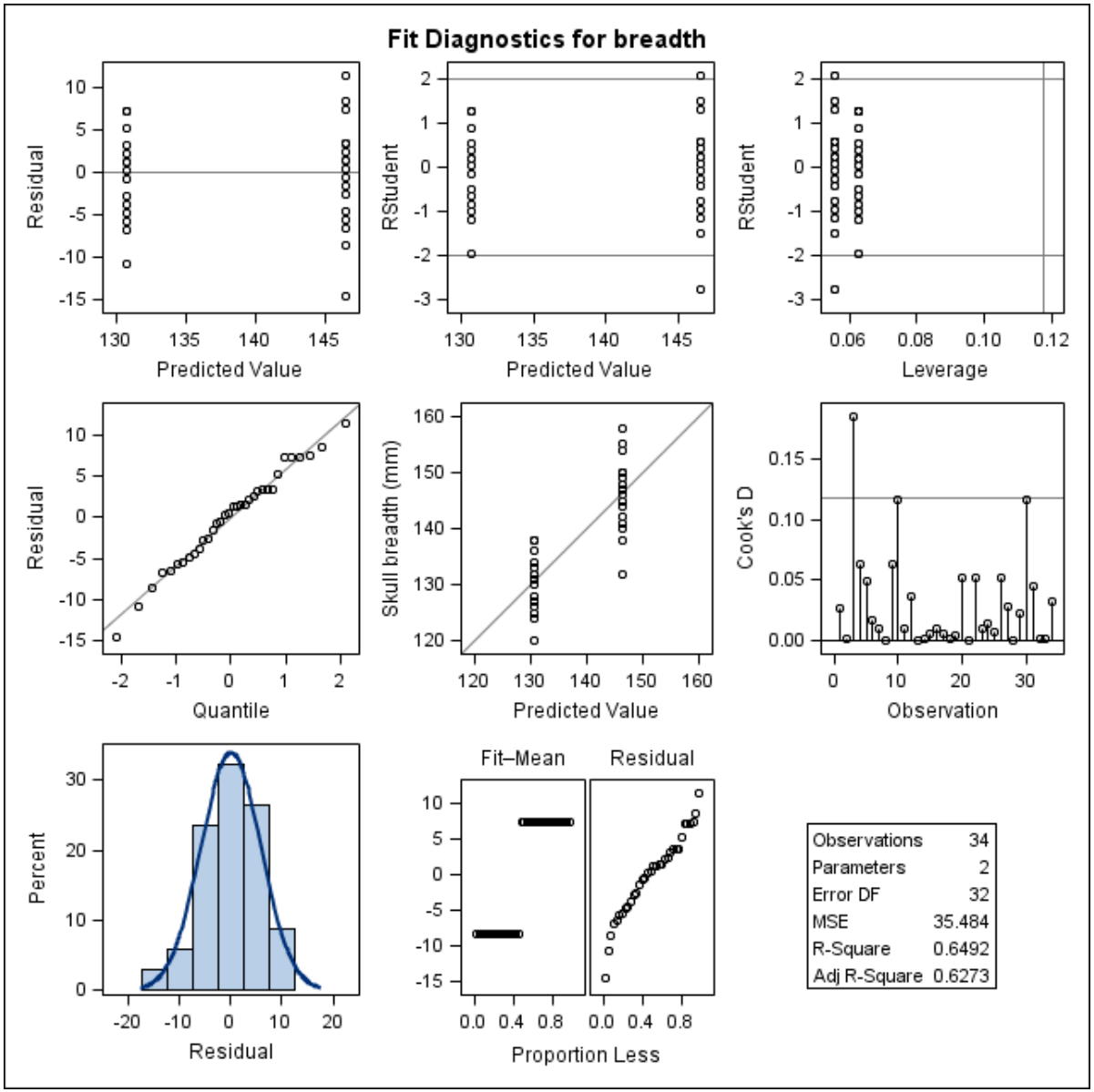
$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

SAS output

```
                                        Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        2     659849.5000     329924.7500    9297.75    <.0001
Error                       32       1135.5000         35.4844
Uncorrected Total           34     660985.0000

R-Square     Coeff Var      Root MSE    breadth Mean
0.649184      4.282804      5.956876        139.0882

                                                 Standard
Parameter                      Estimate             Error    t Value    Pr > |t|
English-Celt                15.7500000       2.04673584       7.70      <.0001

                                               Standard
Parameter                    Estimate            Error    t Value    Pr > |t|
group      Celt          130.7500000       1.48921907      87.80      <.0001
group      English       146.5000000       1.40404920     104.34      <.0001
```

The `English-Celt` row is from the `estimate` command and gives an estimate of $\mu_2 - \mu_1$ (English vs. Celt) and a test of $H_0 : \mu_2 - \mu_1 = 0$.

Diagnostics



5.9 Simple linear regression in matrix terms

Recall the model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

Place the data $(Y_1, \ldots, Y_n)$, predictors $(x_1, \ldots, x_n)$, and trend parameters (β_0, β_1) into vectors and matrices and write the data and model as

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 x_n + \epsilon_n \end{aligned}$$

Simple linear regression

or equivalently in vector/matrix terms

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 + \epsilon_1 \\ \beta_0 + \beta_1 x_2 + \epsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_n + \epsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Define the vectors $\mathbf{Y}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$, and the matrix X as

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

The model is succinctly written

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Simple linear regression

The model is

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \ldots, n\,,$$

where the ϵ_i are *iid* $N(0, \sigma^2)$. In matrix terms the model can be written

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times 2} \boldsymbol{\beta}_{2\times 1} + \boldsymbol{\epsilon}_{n\times 1}$$

5.10 Least squares estimation (pp. 199–200)

As before, it can be shown through some matrix manipulations (not terribly hard, but not illuminating either) that the least squares estimators for (β_0, β_1) are given by

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

This is a function of the matrix $\mathbf{X}$ (called the *design* matrix) and the data $\mathbf{Y}$ only. The unbiased estimator of σ^2 can be written in terms of b_0 and b_1 as

$$\text{MSE} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - [b_0 + b_1 x_i])^2 = \frac{1}{n-2}||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2.$$

Example: Coleman Report Data

Mosteller and Tukey (1977) and Christensen (1996) considered data collected from $n = 20$ schools in the New England and Mid-Atlantic states of the USA.

There are two variables: Y_i, the overall verbal test score for sixth graders and x_i, a composite measure of socioeconomic status. The data are presented in the following table.

We wish to predict Y based on x and test whether there is a relationship between socioeconomic status and verbal test scores.

Coleman report data

School	y	x	School	y	x
1	37.01	7.20	11	23.30	−12.86
2	26.51	−11.71	12	35.20	0.92
3	36.51	12.32	13	34.90	4.77
4	40.70	14.28	14	33.10	−0.96
5	37.10	6.31	15	22.70	−16.04
6	33.90	6.16	16	39.70	10.62
7	41.80	12.70	17	31.80	2.66
8	33.40	−0.17	18	31.70	−10.99
9	41.01	9.85	19	43.10	15.03
10	37.20	−0.05	20	41.01	12.77

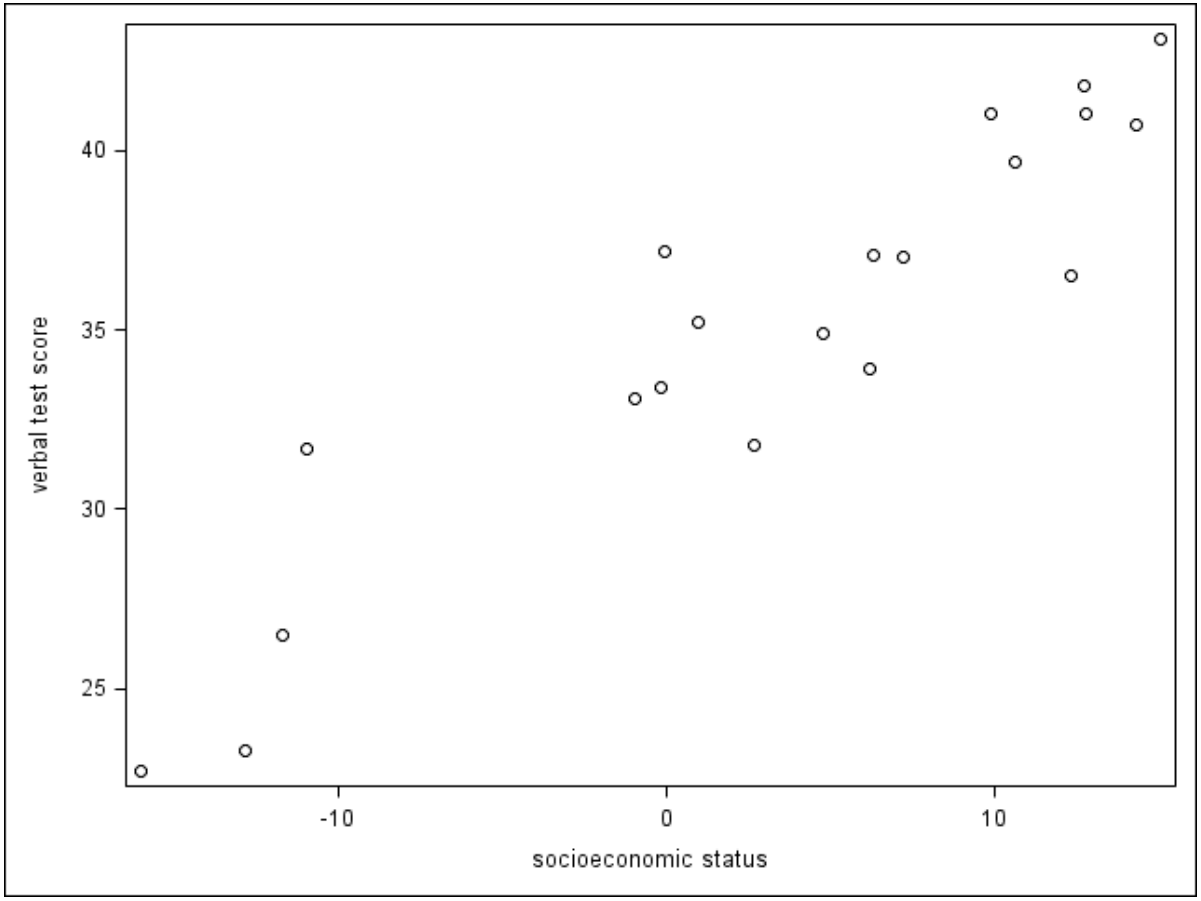
Look at the data...

```
data coleman;
input verbal ses @@;
label verbal='verbal test score' ses='socioeconomic status';
datalines;
 37.01   7.20 23.30 -12.86 26.51 -11.71 35.20   0.92
 36.51  12.32 34.90   4.77 40.70  14.28 33.10  -0.96
 37.10   6.31 22.70 -16.04 33.90   6.16 39.70  10.62
 41.80  12.70 31.80   2.66 33.40  -0.17 31.70 -10.99
 41.01   9.85 43.10  15.03 37.20  -0.05 41.01  12.77
;
proc sgscatter data=coleman;
 plot verbal*ses; run;
 options nocenter;
proc glm plots=diagnostics;
 model verbal=ses / solution;
run;
```

Linear increasing trend; roughly constant variance, we'll fit

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Scatterplot



SAS output

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	552.6756109	552.6756109	110.23	<.0001
Error	18	90.2487641	5.0138202		
Corrected Total	19	642.9243750			

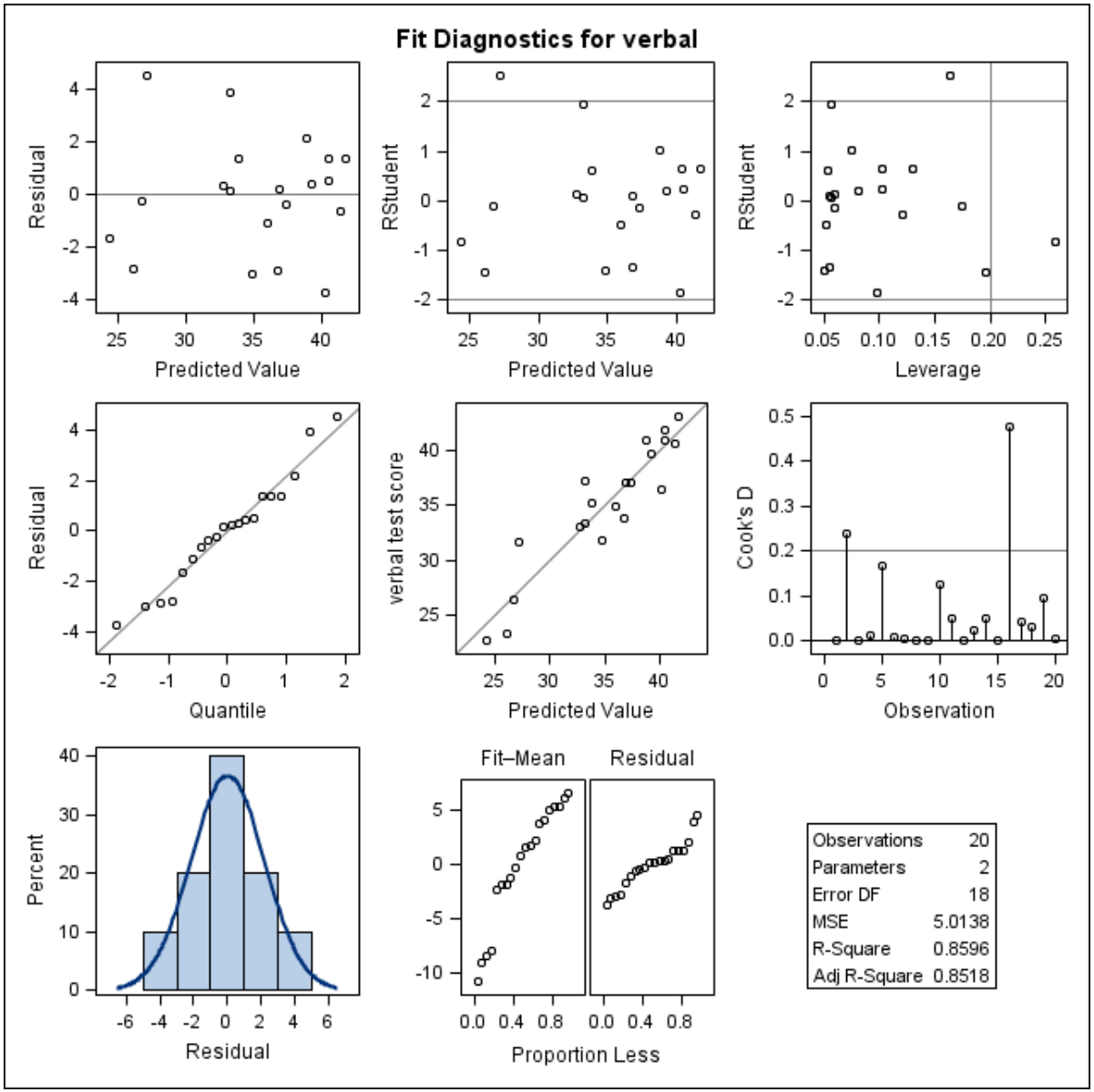
R-Square	Coeff Var	Root MSE	verbal Mean
0.859628	6.382544	2.239156	35.08250

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	33.32279787	0.52799870	63.11	<.0001
ses	0.56032547	0.05336906	10.50	<.0001

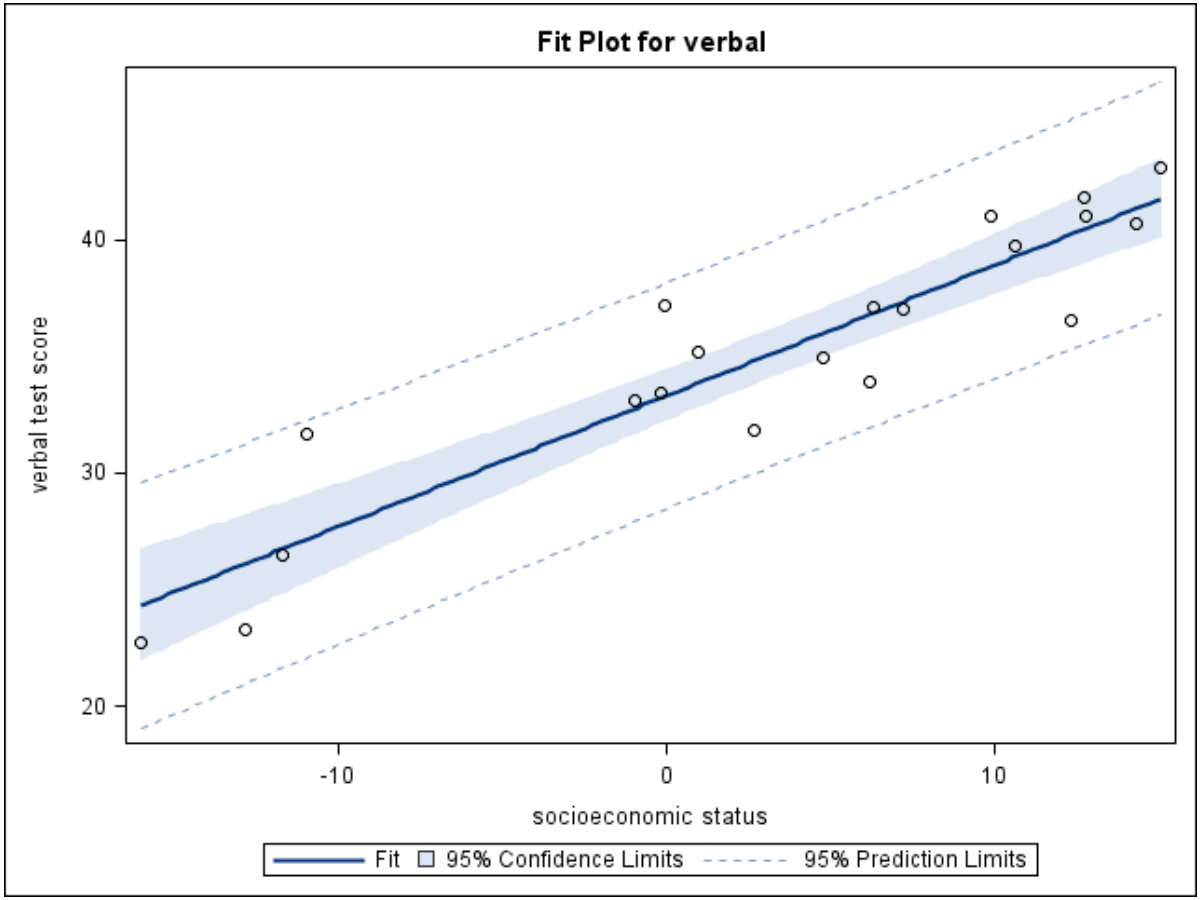
Fitted line is

$$\widehat{\text{verbal}} = 33.3 + 0.56 \text{ ses}.$$

Diagnostics



Line fit


Fit Plot for verbal
verbal test score
40
30
20
-10
0
10
socioeconomic status
Fit
95% Confidence Limits
95% Prediction Limits

Coleman Report data in matrix terms

$$\mathbf{X} = \begin{bmatrix} 1 & 7.2 \\ 1 & -11.71 \\ 1 & 12.32 \\ 1 & 14.28 \\ 1 & 6.31 \\ 1 & 6.16 \\ 1 & 12.7 \\ 1 & -0.17 \\ 1 & 9.85 \\ 1 & -0.05 \\ 1 & -12.86 \\ 1 & 0.92 \\ 1 & 4.77 \\ 1 & -0.96 \\ 1 & -16.04 \\ 1 & 10.62 \\ 1 & 2.66 \\ 1 & -10.99 \\ 1 & 15.03 \\ 1 & 12.77 \end{bmatrix}, \text{ and } \mathbf{y} = \begin{bmatrix} 37.01 \\ 26.51 \\ 36.51 \\ 40.7 \\ 37.1 \\ 33.9 \\ 41.8 \\ 33.4 \\ 41.01 \\ 37.2 \\ 23.3 \\ 35.2 \\ 34.9 \\ 33.1 \\ 22.7 \\ 39.7 \\ 31.8 \\ 31.7 \\ 43.1 \\ 41.01 \end{bmatrix}.$$

Matrix computations by hand...

This yields

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 20 & 62.81 \\ 62.81 & 1957.57 \end{bmatrix}, (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.05560 & -0.00178 \\ -0.00178 & 0.000568 \end{bmatrix}, \text{ and } \mathbf{X}'\mathbf{y} = \begin{bmatrix} 701.65 \\ 3189.9 \end{bmatrix}.$$

So the least squares estimates are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 0.05560 & -0.00178 \\ -0.00178 & 0.000568 \end{bmatrix} \begin{bmatrix} 701.65 \\ 3189.9 \end{bmatrix} = \begin{bmatrix} 33.3 \\ 0.56 \end{bmatrix}.$$

So our best guess of the unknown $\boldsymbol{\beta} = (\beta_0, \beta_1)$ is given by $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1) = (33.3, 0.56)$. So our best guess for the overall population *trend* is

$$\widehat{E(Y)} = 33.3 + 0.56x.$$

For every unit increase in socioeconomic status, we see *on average* an increase of 0.56 in the overall verbal test scores.