

Coleman Report data / matrix approach review

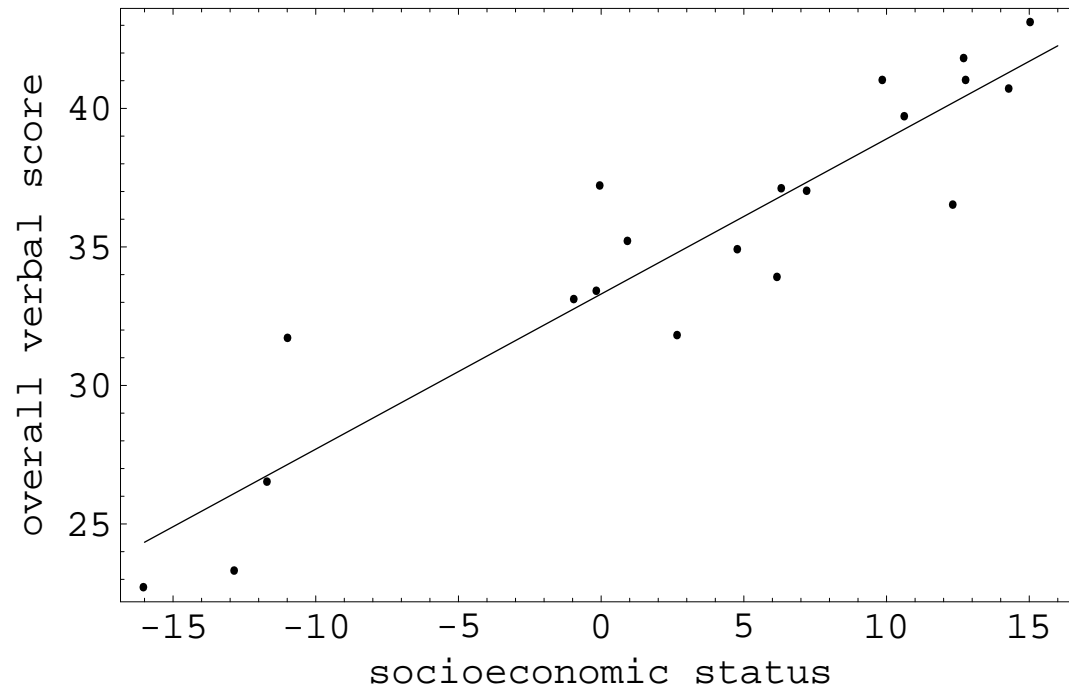


Figure 1: Coleman Report data with fitted line $\widehat{E}(Y) = 33.3 + 0.56SES$.

Matrix notation

Recall the model

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

where $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$. Both ϵ_i and e_i are used to denote regression deviations in statistics.

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

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

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 \mathbf{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

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

Quick review: The model is

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

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

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

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

Maximum likelihood estimation

There are three unknown population parameters in the probability model: $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$. Since all n observations are independent (but *not iid!*), the likelihood is the product of the n marginal densities:

$$\begin{aligned} f(y_1, \dots, y_n | \beta_0, \beta_1, \sigma^2) &= \mathcal{L}(\beta_0, \beta_1, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2} \left(\frac{y_i - [\beta_0 + \beta_1 x_i]}{\sigma} \right)^2 \right\}. \end{aligned}$$

Through some matrix manipulations (not terribly hard, but not illuminating either), it can be shown that the maximum likelihood estimators for (β_0, β_1) are given by

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

This is a function of the matrix \mathbf{X} (called the *design* matrix) and the data \mathbf{Y} only. The maximum likelihood estimator of σ^2 can be written in terms of $\hat{\beta}_0$ and $\hat{\beta}_1$ as

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_i])^2.$$

Back to the Coleman Report data. The design matrix and data vector are:

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

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 MLE's 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. The estimated slope is positive. Revisit Figure 1 to see that this line gives a nice “fit” to the data.

Chapter 9: hypothesis testing

This chapter goes through the theory of statistical hypothesis testing in some detail, including the Neyman-Pearson paradigm. In what follows we will essentially follow the Neyman-Pearson setup, but introduce ideas as they come up rather than present formal mathematical details first.

I encourage you to read through this chapter, especially 9.1, 9.2, 9.3, and 9.4. However, we will cover aspects of these sections in the context of specific models. You may actually prefer the formal presentation in the book. They should compliment each other.

Hypothesis testing and confidence intervals for *iid* data

It is often of scientific interest to test the validity of claims involving population parameters in a probability model. For example:

1. In an experiment $n = 139$ French skiers were given vitamin C over a two week period. Let $Y_i = 1$ if skier i developed a cold and $Y_i = 0$ if skier i did not develop a cold over the two weeks. Let

$$Y_1, \dots, Y_{139} \stackrel{iid}{\sim} \text{Bernoulli}(\theta).$$

We may want to show that $H_1 : \theta < 0.1$, i.e. less than 10% of the population of all French skiers develop colds while taking vitamin C.

2. The survival times (in weeks) from time of diagnosis of $n = 33$ leukemia patients Y_1, \dots, Y_{33} are recorded in a study. Let

$$Y_1, \dots, Y_{33} \stackrel{iid}{\sim} \exp(\theta).$$

We may want to show $H_1 : \theta > 52$, i.e. that the mean survival time θ is *less than* 52 weeks (one year). $\theta > 52$ implies $E(Y_i) = 1/\theta < 52$ weeks.

3. The calorie content Y_i of $n = 17$ poultry hot dogs is recorded.

$$Y_1, \dots, Y_{17} \stackrel{iid}{\sim} N(\mu, \sigma^2).$$

We may want to show $H_1 : \mu > 100$, i.e. the average calorie count for poultry brands is more than 100 calories.

The classical hypothesis test is a proof by contradiction. The null hypothesis H_0 is typically specified to be what you *do not* want to show and is a statement about the population parameter θ .

The null (which is opposite of what you really want to show) is assumed to be true, then evidence that the observed data is highly unlikely when H_0 is true is presented. The implication is that H_0 cannot be true given what we saw and thus the *alternative* hypothesis H_1 must be true.

Hopefully this will be made clear below.

Three steps in performing a hypothesis test:

1. State what you want to *disprove* in the null hypothesis H_0 . This is a statement about the population parameter θ .
2. Collect data Y_1, \dots, Y_n from your assumed probability model $f(y_1, \dots, y_n | \theta)$ or $p(y_1, \dots, y_n | \theta)$.
3. Compute some measure of how bizarre the observed data $(Y_1, \dots, Y_n) = (y_1, \dots, y_n)$ are given that H_0 is true. If this measure indicates seeing what you saw $(Y_1, \dots, Y_n) = (y_1, \dots, y_n)$ is highly unlikely, reject H_0 as false. The measure we discuss below is the p -value; another measure we will not discuss is called the “Bayes factor.”

We formalize these ideas for normal data.

Testing the mean in normal data

A very important probability model for data is the normal distribution. Most statistics books spend a great deal of time discussing this case:

$$Y_1, \dots, Y_n \stackrel{iid}{\sim} N(\mu, \sigma^2).$$

Recall two important facts about normal data:

1. The MLE's of μ and σ^2 are $\hat{\mu} = \bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$, the sample mean, and $\hat{\sigma}^2 = S_n^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$, the sample variance.
2. $\bar{Y}_n \sim N(\mu, \sigma^2/n)$ exactly because the data are normal, but μ and σ usually are not known.

Case I: σ known

When σ is known – which is almost never the case – then the exact distribution of \bar{Y}_n is known: $\bar{Y}_n \sim N(\mu, \sigma^2/n)$.

Although σ is almost never exactly known (how could you know σ exactly but not μ ?), when n is large σ is estimated by its MLE $\sigma \approx \hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2}$. Recall that $S_n^2 \xrightarrow{P} \sigma^2$, i.e. when n is large S_n^2 is close to σ^2 with high probability.

We typically want to show one of the following null hypotheses is false: $H_0 : \mu = \mu_0$, $H_0 : \mu \leq \mu_0$, or $H_0 : \mu \geq \mu_0$ for some known hypothesized μ_0 .

Example:

The calorie content in $n = 17$ brands of poultry hot dogs is recorded off the backs of packages. The data are $y_1, \dots, y_{17} = 129, 132, 102, 106, 94, 102, 87, 99, 107, 113, 135, 142, 86, 143, 152, 146, 144$. We assume the probability model

$$Y_1, \dots, Y_{17} \stackrel{iid}{\sim} N(\mu, \sigma^2).$$

Our estimate of μ is $\bar{y}_{17} = (129 + 132 + \dots + 144)/17 = 118.8$. Our estimate of σ^2 is

$$s_{17}^2 = [(129 - 118.8)^2 + (132 - 118.8)^2 + \dots + (144 - 118.8)^2]/17 = 478.7$$

and so σ is estimated by $\sqrt{s_{17}^2} = \sqrt{478.7} = 21.9$.

We wish to provide evidence that $H_1 : \mu > 100$ is true. So $\mu_0 = 100$ here, our hypothesized value. Our null is then $H_0 : \mu \leq 100$

1. $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. Seeing \bar{y}_n (which estimates the unknown μ) that is very different from μ_0 supports the alternative $H_1 : \mu \neq \mu_0$, but *how different* is \bar{y}_n from μ_0 ? A useful measure of how unusual the observed \bar{y}_n is given that $H_0 : \mu = \mu_0$ is true is given by the p -value:

$$p\text{-value} = P(|\bar{Y}_n - \mu_0| > |\bar{y}_n - \mu_0|) \text{ given that } \mu = \mu_0 \text{ is true.}$$

The p -value is the probability of seeing a sample mean \bar{Y}_n that is further away from μ_0 given $H_0 : \mu = \mu_0$ is true. If the p -value is small, we have evidence that $H_0 : \mu = \mu_0$ is false and therefore conclude $H_1 : \mu \neq \mu_0$ is true.

That is, we assume $H_0 : \mu = \mu_0$ and (hopefully) report that the observed data (y_1, \dots, y_n) as summarized by \bar{y}_n are so unlikely that what we assumed ($H_0 : \mu = \mu_0$) cannot possibly be true.

Note that the p -value is computed

$$\begin{aligned} p\text{-value} &= P(|\bar{Y}_n - \mu_0| > |\bar{y}_n - \mu_0|) \text{ given that } \mu = \mu_0 \text{ is true} \\ &= P\left(\left|\frac{\bar{Y}_n - \mu_0}{\sqrt{\sigma^2/n}}\right| > \left|\frac{\bar{y}_n - \mu_0}{\sqrt{\sigma^2/n}}\right|\right) \\ &= P(|Z| > |z_0|) \end{aligned}$$

where the *test statistic* $z_0 = (\bar{y}_n - \mu_0)/\sqrt{\sigma^2/n}$ is readily computed from known \bar{y}_n , μ_0 , and σ . The p -value is easily computed from the table for $\Phi(z)$ in the back of your book, or in SAS or R.

Before data are collected, the test statistic has a known, simple distribution assuming the null hypothesis is true:

$$Z_0 = \frac{\bar{Y}_n - \mu_0}{\sqrt{\sigma^2/n}} \sim N(0, 1).$$

2. $H_0 : \mu \geq \mu_0$ versus $H_1 : \mu < \mu_0$.

p -value = $P(\bar{Y} < \bar{y})$ given that $\mu = \mu_0$ is true.

That is, the p -value is the probability of seeing a sample mean \bar{Y}_n even smaller than what we saw \bar{y}_n given that $\mu = \mu_0$.

3. $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$.

p -value = $P(\bar{Y}_n > \bar{y}_n)$ given that $\mu = \mu_0$ is true.

That is, the p -value is the probability of seeing a sample mean \bar{Y}_n (our estimate of μ !) even larger than what we saw \bar{y}_n given that $\mu = \mu_0$. This is the setup for the poultry hot dog example.

Example: poultry hot dog data continued

The following R code loads the calorie data and makes a histogram.

```
> cal <- c(129,132,102,106,94,102,87,99,107,113,135,142,86,143,152,146,144)
> hist(cal)
```

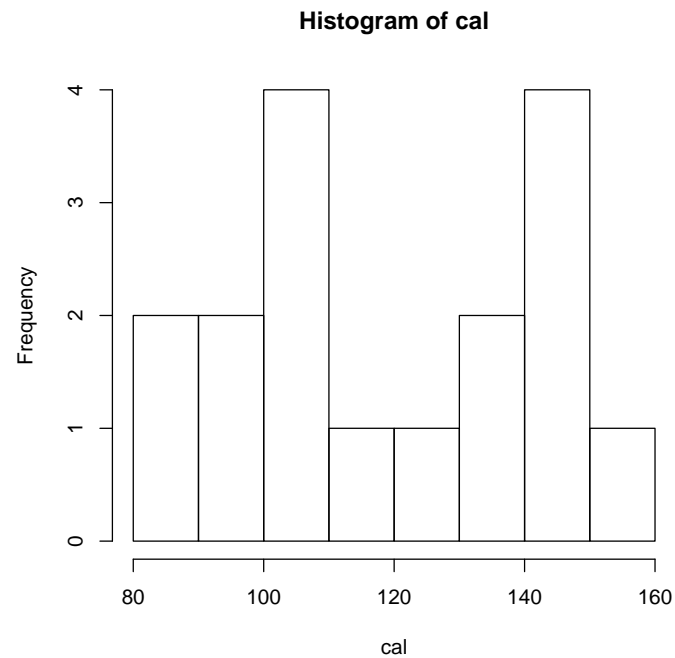


Figure 2: Histogram of $n = 17$ calorie counts.

The histogram shows some evidence of non-normality. We can perform the Anderson-Darling test for normality in R by first loading the package `nortest` then using the `ad.test()` function.

Under the **Packages** menu select **Install package(s)...**, pick a U.S. mirror and install the `nortest` package. Then under **Packages** select **Load package...**, pick `nortest` and click .

```
> ad.test(cal)
```

```
Anderson-Darling normality test
```

```
data: cal A = 0.6214, p-value = 0.088
```

We don't reject normality at the 5% level (more on this later), but it's a close call. Let's proceed anyway.

We wish to show that the average calorie content of poultry hot dogs is over 100: $H_1 : \mu > 100$. Our null is then $H_0 : \mu \leq 100$; here $\mu_0 = 100$, the hypothesized value. We compute $\bar{y}_{17} = 118.8$ and $\hat{\sigma}^2 = 478.7$. Let's assume that $\sigma^2 = 478.7$ and proceed. The test statistic is

$$z_0 = \frac{\bar{y}_{17} - \mu_0}{\sqrt{\hat{\sigma}^2/17}} = \frac{118.8 - 100}{\sqrt{478.7/17}} = 3.54.$$

The p -value is

$$p\text{-value} = P(Z > 3.54) = 0.00020.$$

If $H_0 : \mu \leq 100$ is true, then the probability of seeing \bar{Y}_{17} even further away from $\mu_0 = 100$ than $\bar{y}_{17} = 118.8$ is about 0.0002. This is highly unlikely and we should reject H_0 and accept $H_1 : \mu > 100$ as true. We discuss how to *formally* decide between H_0 and H_1 next time.

Note several key points:

1. When H_0 is assumed to be true, a *test statistic* Z_0 with a known distribution is formed.
2. A p -value is computed which measures how unlikely it is to see the observed data or data even more bizarre given that H_0 is true.
3. If the p -value is small, the evidence (data) dictates that H_1 must be true.