

**An important model for data:**  $N(\mu, \sigma^2)$ :

$$X_1, X_2, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2).$$

May be interested in estimating  $\mu$ ,  $\sigma$ , or  $x_p$ , testing  $H_0 : \mu \leq 200$  calories,  $H_0 : x_{0.5} \geq 24$  months (more later).

- MOM estimators:  $\hat{\mu} = \bar{X}_n$  and  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ , the sample mean and variance.
- MLE estimators:  $\hat{\mu} = \bar{X}_n$  and  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ , the sample mean and variance.

Recall  $E(\bar{X}_n) = \mu$  and  $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$ .  $\bar{X}_n$  is unbiased for  $\mu$ .  $\hat{\sigma}^2$  is biased for  $\sigma^2$ , but the bias goes to zero as  $n \rightarrow \infty$ .

Also recall, for normal data  $X_1, \dots, X_n$ ,

$$\bar{X}_n \sim N(\mu, \sigma^2/n),$$

*exactly*. This is said to be the sampling distribution of  $\bar{X}_n$ . For normal  $X_1, \dots, X_n$  we know how  $\bar{X}_n$  is distributed in terms of the unknown  $(\mu, \sigma)$ .

What about  $\hat{\sigma}^2$ ? An unbiased estimate of the variance  $\sigma^2$  is

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Turns out that

$$S_n^2 \sim \text{gamma} \left( \frac{n-1}{2}, \frac{n-1}{2\sigma^2} \right).$$

Note then that  $E(S_n^2) = \sigma^2$  as required.

Also,

$$\hat{\sigma}^2 \sim \text{gamma} \left( \frac{n-1}{2}, \frac{n}{2\sigma^2} \right),$$

and so  $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$ . These sampling distributions for  $S_n^2$  and  $\hat{\sigma}^2$  are obtained using moment generating functions; see Section 6.3 (pp. 195–198) if interested.

Something that will come in useful later on:

**def'n:** Let  $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$ . Then

$$T = \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}} = \frac{\bar{X}_n - \mu}{\sqrt{\hat{\sigma}_n^2/(n-1)}} \sim t_{n-1},$$

a Student's  $t$  distribution with  $n - 1$  degrees of freedom.

## Showing $S_n^2$ is unbiased for $\sigma^2$ from scratch...

Showing this is doable using tools we've seen, but tedious. We might try it if we've got time, or else you can try it on your own. Note that

$E(\bar{X}_n^2) = \text{var}(\bar{X}_n) + [E(\bar{X}_n)]^2 = \sigma^2/n + \mu^2$  and similarly  
 $E(X_i^2) = \sigma^2 + \mu^2$ . What is  $E(X_i\bar{X}_n)$ ?

$$\begin{aligned} E(X_i\bar{X}_n) &= \frac{E(X_iX_1 + X_iX_2 + \cdots + X_iX_i + \cdots + X_iX_n)}{n} \\ &= \frac{E(X_iX_1) + E(X_iX_2) + \cdots + E(X_iX_i) + \cdots + E(X_iX_n)}{n} \\ &= \frac{\mu^2 + \mu^2 + \cdots + \sigma^2 + \mu^2 + \cdots + \mu^2}{n} \\ &= \mu^2 + \frac{\sigma^2}{n}. \text{ This can get you started...} \end{aligned}$$

**Example:** calories from  $n = 17$  randomly selected brands of poultry hot dogs.

```
> cal=c(129,132,102,106,94,102,87,99,107,113,135,142,86,143,152,146,144); n=length(cal)
> unbiased.var=sum((cal-mean(cal))^2)/(n-1)
> biased.var=sum((cal-mean(cal))^2)/n
> unbiased.var
[1] 508.5662
> biased.var
[1] 478.6505
> var(cal) # BUILT-IN R FUNCTION, WHICH ESTIMATE IS IT?
[1] 508.5662
> xbar=mean(cal); mle.std=sqrt(biased.var)
> grid=seq(xbar-3*mle.std,xbar+3*mle.std,1); est.dens=dnorm(grid,xbar,mle.std)
> hist(cal,freq=FALSE,xlim=c(60,180)); lines(grid,est.dens,lty=3)
> cal.sort=sort(cal)
> cal.sort[ceiling(0.95*n)]
[1] 152
> qnorm(0.95,xbar,mle.std)
[1] 154.7510
> cal.sort[ceiling(0.99*n)]
[1] 152
> qnorm(0.99,xbar,mle.std)
[1] 169.6607
```

If we are willing to *assume* normality for these data, we get to do things like estimate  $x_{0.99}$ , the 99<sup>th</sup> percentile of the data.

Of course we can do this nonparametrically using the sample estimate  $x_{(\lceil n0.99 \rceil)}$  instead, but what is happening here?

For no other reason than we can, let's find the MLE of the signal to noise ratio:

```
> xbar/mle.std  
[1] 5.428479
```

That is,  $\hat{\mu}/\hat{\sigma} = 5.43$ ; the mean is estimated to be over five times larger than the standard deviation.

Let's see if normality makes sense here; remember, there's only  $n = 17$  observations...

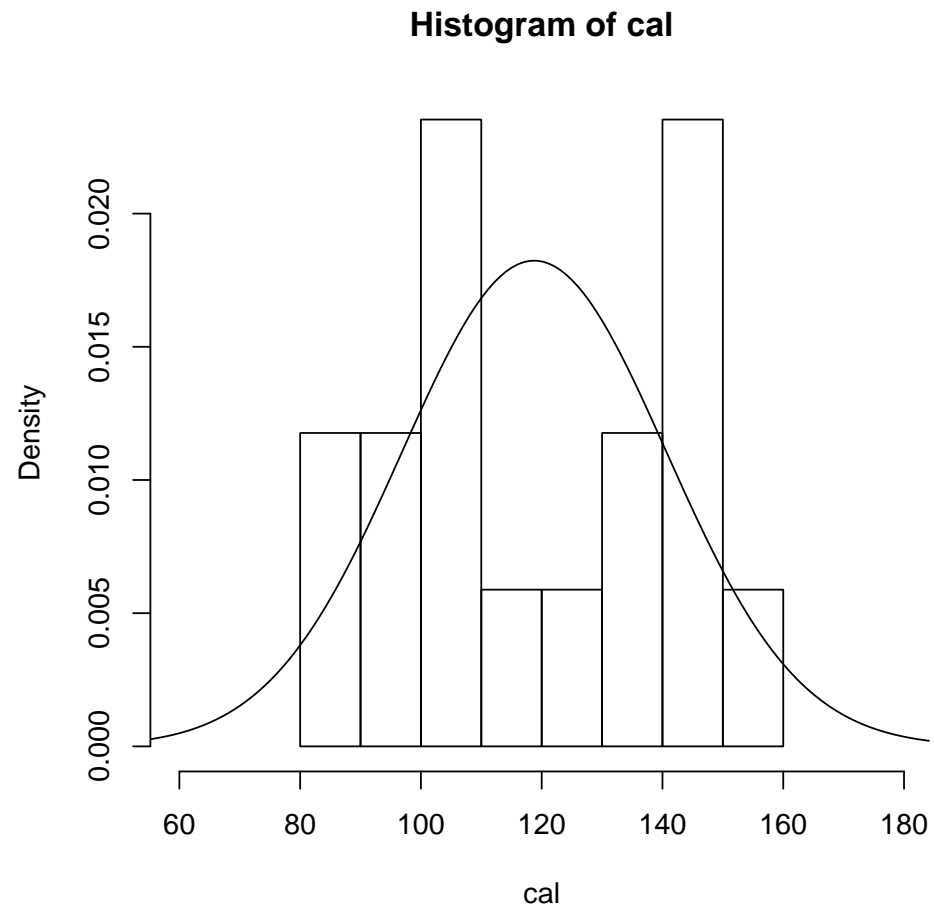


Figure 1: Default R histogram and best-fitting (MLE) normal density.

## Beyond *iid* data: two-sample and regression models

We've been exploring *iid* data

$$X_1, \dots, X_n \stackrel{iid}{\sim} f(x|\boldsymbol{\theta}).$$

We discussed two methods for estimating  $\boldsymbol{\theta}$ , namely MOM and MLE. Let's now explore some useful, commonly-used statistical models that do not have *iid* data:

- Two-sample data.
- Simple linear regression.
- Simple logistic regression.

## The two-sample normal model

We are often in a position to compare a continuous outcome measure across two groups. For example we may want to compare the hours of sleep obtained from two sleeping pills, survival time in months for two different cancer treatments, number of fleas on dogs versus cats, etc. The two-sample model with common variance stipulates:

$$X_{11}, X_{12}, \dots, X_{1n_1} \stackrel{iid}{\sim} N(\mu_1, \sigma^2),$$

independent of

$$X_{21}, X_{22}, \dots, X_{2n_2} \stackrel{iid}{\sim} N(\mu_2, \sigma^2).$$

There are  $n_1$  observations  $X_{11}, X_{12}, \dots, X_{1n_1}$  from the first sample and  $n_2$  observations  $X_{21}, X_{22}, \dots, X_{2n_2}$  from the second sample. We allow each group to have its own population mean  $\mu_1$  or  $\mu_2$ , but both groups have common variance (initially).

Note that the data are not *iid*, but rather independent. Within each group the data are *iid* from their own distributions. I'll draw a picture on the board.

**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.

The Englishmen skull measurements in *mm* are  $x_{1,1}, \dots, x_{1,18} =$   
141, 148, 132, 138, 154, 142, 150, 146, 155, 158, 150, 140, 147, 148, 144, 150,  
149, 145.

The Celtic skulls are  $x_{2,1}, \dots, x_{2,16} =$   
133, 138, 130, 138, 134, 127, 128, 138, 136, 131, 126, 120, 124, 132, 132, 125.

R code to read in the data and create side-by-side histograms on the same scale:

```
> english <- c(141,148,132,138,154,142,150,146,155,158,150,140,147,148,144,150,149,145)
> celt <- c(133,138,130,138,134,127,128,138,136,131,126,120,124,132,132,125)
> left <- min(min(celt),min(english))-1
> right <- max(max(celt),max(english))+1
> par(mfrow=c(1,2))
> hist(english,xlim=c(left,right))
> hist(celt,xlim=c(left,right))
```

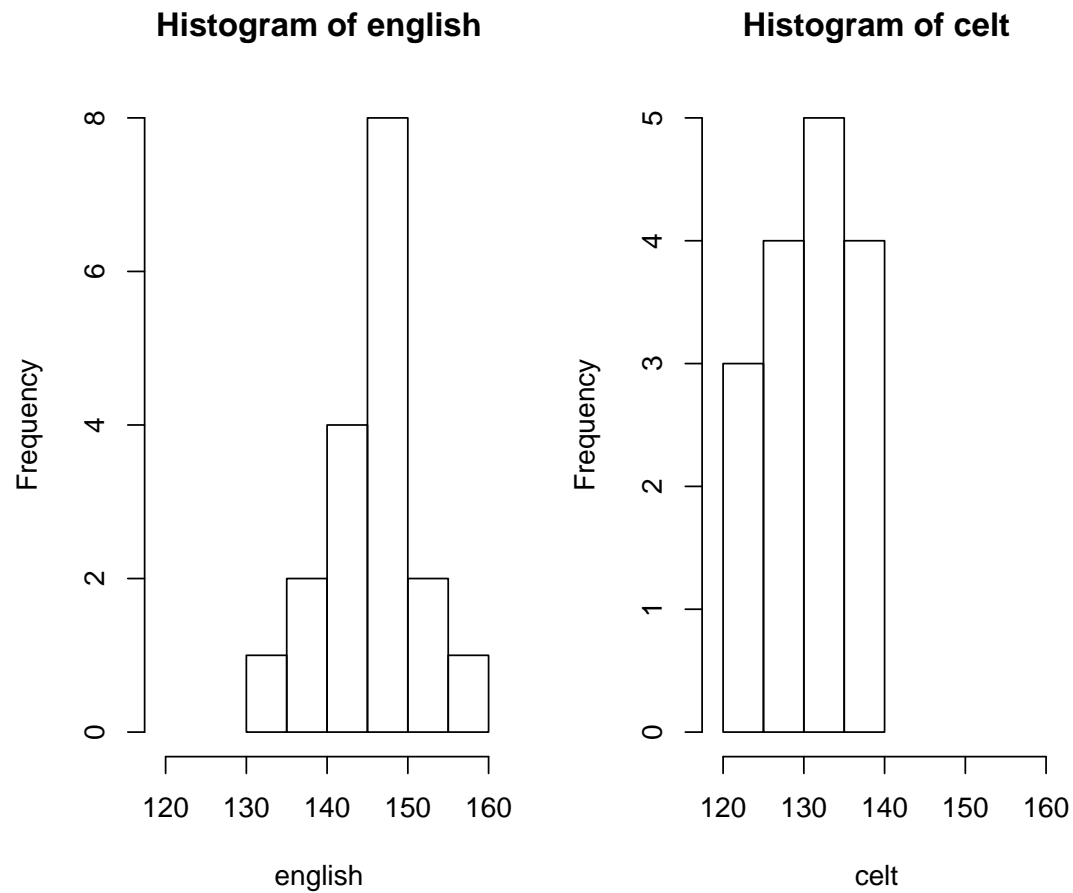


Figure 2: The normal model with equal variances seems okay.

Questions we might have include:

- Are the model assumptions reasonable here? Are the data approximately normal? Are the variances approximately equal?
- Can we reject that there are no population differences in skull size? That is, can we reject  $H_0 : \mu_1 = \mu_2$ ?
- What is an estimate and range of values that we're fairly certain the difference  $\mu_1 - \mu_2$  lies in?
- **Before collecting the data, we might ask:** If we expect that there's a difference  $\Delta = \mu_1 - \mu_2$  ahead of time and have on hand an estimate of  $\sigma$ , what sample sizes  $n_1 = n_2$  would allow us to reject  $H_0 : \mu_1 = \mu_2$  with high probability?

There are three model parameters  $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma)$ . The likelihood is given by

$$\begin{aligned}\mathcal{L}(\mu_1, \mu_2, \sigma) &= \prod_{i=1}^{n_1} f(x_{1i}|\mu_1, \sigma) \prod_{i=1}^{n_2} f(x_{2i}|\mu_2, \sigma) \\ &= \prod_{i=1}^{n_1} \frac{\exp\{-0.5(x_{1i} - \mu_1)^2/\sigma^2\}}{\sqrt{2\pi\sigma^2}} \times \\ &\quad \prod_{i=1}^{n_2} \frac{\exp\{-0.5(x_{2i} - \mu_2)^2/\sigma^2\}}{\sqrt{2\pi\sigma^2}}\end{aligned}$$

Through calculus we can show that the MLE's are

$$\begin{aligned}\hat{\mu}_1 &= \bar{x}_{1\bullet} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i}, \quad \hat{\mu}_2 = \bar{x}_{2\bullet} = \frac{1}{n_2} \sum_{i=1}^{n_2} x_{2i}, \quad \text{and} \\ \hat{\sigma}^2 &= \frac{1}{n_1+n_2} \left[ \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_{1\bullet})^2 + \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_{2\bullet})^2 \right].\end{aligned}$$

The MLE's for  $\mu_1$  and  $\mu_2$  are unbiased:  $E(\hat{\mu}_1) = \mu_1$  and  $E(\hat{\mu}_2) = \mu_2$ , but  $E(\hat{\sigma}^2) \neq \sigma^2$ . However, we know that as  $n_1$  and  $n_2$  get large that  $E(\hat{\sigma}^2) \approx \sigma^2$ . In R, estimates can be obtained as:

```
> mean(english)
[1] 146.5
> mean(celt)
[1] 130.75
> (sum((english-mean(english))^2)+sum((celt-mean(celt))^2))/(18+16)
[1] 33.39706
```

## Simple linear regression:

Say we have measurements  $Y_i$  which tend to increase or decrease linearly with covariates  $x_i$ . A model which allows us to test for a linear relationship and perhaps predict future values is the simple linear regression model:

$$Y_i \stackrel{ind.}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2), \quad i = 1, \dots, n.$$

Here the  $Y_1, \dots, Y_n$  are independent but not identically distributed. The mean of each  $Y_i$  changes linearly with  $x_i$ . I'll draw a picture on the board.

The *observed* pairs  $(x_i, y_i)$  can be plotted to check for an overall linear trend and constant variance.

Examples of  $(x_i, Y_i)$  include:

- (average cigarettes smoked per day, lifetime in years)
- (height of mother, length of baby)
- (age, salary in Euros)
- (MPG of automobile driven, resting heart rate)

There are three parameters  $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma)$  and via independence, the likelihood is written

$$\begin{aligned}\mathcal{L}(\beta_0, \beta_1, \sigma) &= \prod_{i=1}^n f(y_i | \beta_0, \beta_1, \sigma) \\ &= \prod_{i=1}^n \frac{\exp\{-0.5(y_i - [\beta_0 + \beta_1 x_i])^2 / \sigma^2\}}{\sqrt{2\pi\sigma^2}}.\end{aligned}$$

Through calculus, the likelihood can be maximized to find the MLE's  $\hat{\beta}_1 = \sum_{i=1}^n (x_i - \bar{x}_{\bullet})(y_i - \bar{y}_{\bullet}) / \sum_{i=1}^n (x_i - \bar{x}_{\bullet})^2$  and  $\hat{\beta}_0 = \bar{y}_{\bullet} - \hat{\beta}_1 \bar{x}_{\bullet}$ .

These estimates of the slope and intercept are unbiased,  $E(\hat{\beta}_1) = \beta_1$  and  $E(\hat{\beta}_0) = \beta_0$ . The MLE of  $\sigma^2$  is  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_i])^2$ , however, as in the two-sample model,  $E(\hat{\sigma}^2) \neq \sigma^2$ .

**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 perhaps eventually test that there is a relationship between socioeconomic status and verbal test scores.

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

The following R code reads in the data and makes a scatterplot. Notice an overall increasing trend for the verbal score to increase with socioeconomic status.

```
> score <- c(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)  
> ses <- c(7.2,-11.71,12.32,14.28,6.31,6.16,12.7,-0.17,9.85,-0.05,  
+ -12.86,0.92,4.77,-0.96,-16.04,10.62,2.66,-10.99,15.03,12.77)  
> plot(ses,score)
```

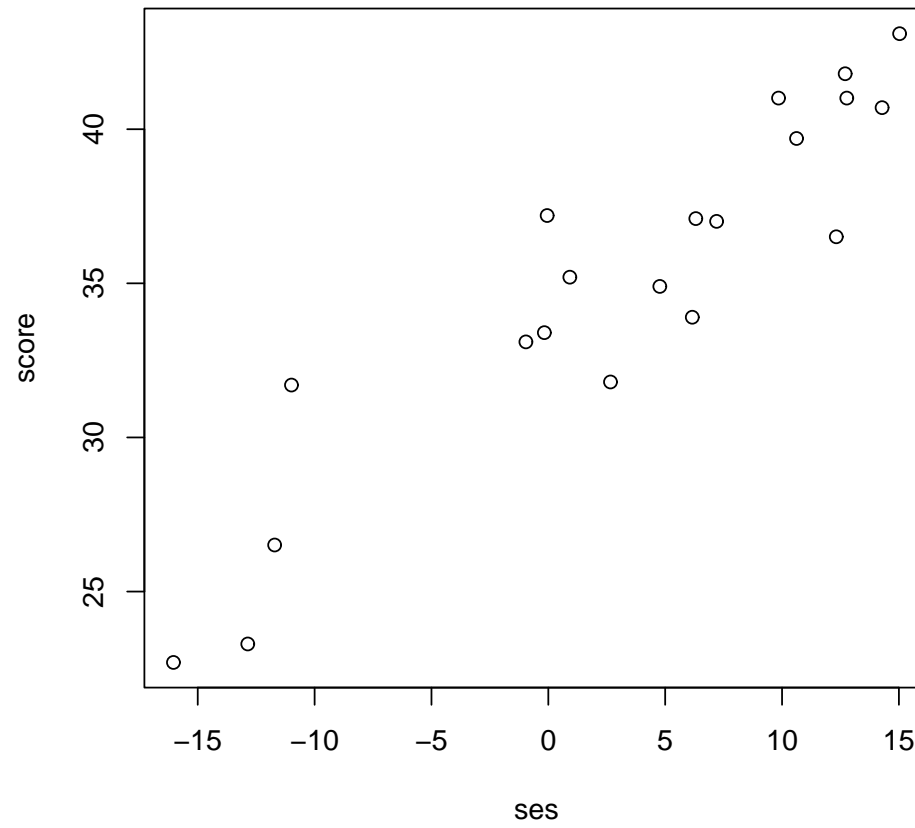


Figure 3: Scatterplot of  $(x_i, y_i)$  for Coleman Report data.

R code to obtain the MLE's looks like:

```
> beta.1 <- sum((ses-mean(ses))*(score-mean(score)))/sum((ses-mean(ses))^2)
> beta.0 <- mean(score)-beta.1*mean(ses)
> sigma2 <- sum((score-beta.0-beta.1*ses)^2)/20
> beta.0
[1] 33.3228
> beta.1
[1] 0.5603255
> sigma2
[1] 4.512438
```

The estimated mean score given SES is given by

$$\widehat{E}(Y) = 33.32 + 0.56 \text{ SES.}$$

Of course, R can fit the the simple linear regression model automatically using the “linear model” function. Try and pick out the slope and intercept estimates from the following output.

```
> model.fit <- lm(score~ses)
> summary(model.fit)
```

```
Call: lm(formula = score ~ ses)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-3.716	-1.230	0.207	1.357	4.535

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	33.32280	0.52800	63.11	< 2e-16 ***
ses	0.56033	0.05337	10.50	4.2e-09 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.239 on 18 degrees of freedom
```

```
Multiple R-Squared: 0.8596,    Adjusted R-squared: 0.8518
```

```
F-statistic: 110.2 on 1 and 18 DF,  p-value: 4.199e-09
```

We have the data  $\{(x_i, Y_i) : i = 1, \dots, n\}$ . The pair  $(x_i, Y_i)$  are the  $i^{th}$  data point;  $x_i$  is the *predictor* variable, or *independent* variable for observation  $i$ , and  $Y_i$  is the response or *dependent* variable.

The model assumes the response follows a mean, overall trend that is a linear function of  $x_i$  plus individual-to-individual variability:

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

Written another way,

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

The  $\{(x_i, Y_i) : i = 1, \dots, n\}$  are a random sample from the population. Let  $(x, Y)$  be a predictor/response pair randomly sampled from the population. This is anyone, whether they are in our sample or not. For this randomly selected observation we have

$$Y \sim N(\beta_0 + \beta_1 x, \sigma^2),$$

and so

$$E(Y) = \beta_0 + \beta_1 x.$$

This is the overall trend. But each observation will deviate randomly about this trend, so

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where  $\epsilon$  represents a random “deviation” about the mean.

For example, we assumed the simple linear regression model for the Coleman report data. For each indicator of socioeconomic status  $x$ , a randomly selected school's overall verbal test score is given by

$$y = \beta_0 + \beta_1 x + \epsilon.$$

$\beta_0 + \beta_1 x$  is the mean overall test score for *all* schools with socioeconomic status  $x$ , and  $\epsilon$  represents *one school's* deviation about this mean.

All of  $\beta_0$ ,  $\beta_1$ , and  $\epsilon$  are unknown. If  $\beta_1 > 0$  then on average a school's overall verbal test scores will increase with socioeconomic status, if  $\beta_1 < 0$ , they will decrease. We may want to show that the hypothesis  $H_1 : \beta_1 \neq 0$  is likely true, or more specifically, that  $H_1 : \beta_1 > 0$ .

Two important, normal-based probability models that do not assume *iid* data:

- Two independent samples:

$$X_{11}, X_{12}, \dots, X_{1n_1} \stackrel{iid}{\sim} N(\mu_1, \sigma^2),$$

$$X_{21}, X_{22}, \dots, X_{2n_2} \stackrel{iid}{\sim} N(\mu_2, \sigma^2).$$

Reduces data to three parameters  $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma^2)$ . Main interest:  $H_0 : \mu_1 = \mu_2$  (more on this shortly).

- Simple (i.e. only one covariate) linear regression:

$$Y_i \stackrel{ind.}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2).$$

Here data are independent, and each individual has their *own* mean  $E(Y_i) = \beta_0 + \beta_1 x_i$ . The mean changes linearly with the  $x_i$ . Reduces data down to three parameters:  $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$ .