Simple linear regression

BIOST 515

January 8, 2004

Simple Linear Regression

Simple linear regression of response Y on predictor X

Begin with sample: $(X_1, Y_1), \ldots (X_N, Y_N)$

$$Y_i = E[Y_i|X_i] + \epsilon_i$$

where

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

and

$$E(\epsilon_i) = 0,\ var(\epsilon_i) = \sigma^2 \text{ and } cov(\epsilon_i, \epsilon_j) = 0.$$

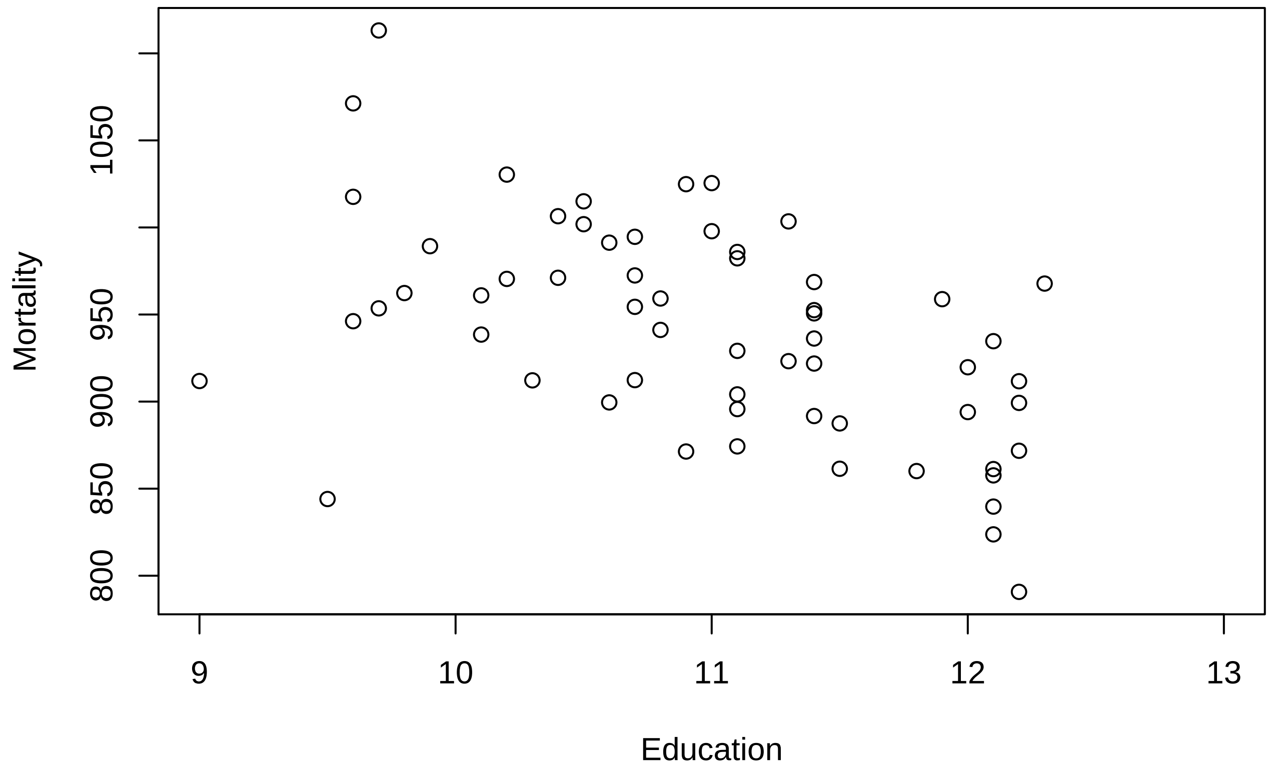
Simple linear regression: example

Trends in mortality with education level

Properties of 60 Standard Metropolitan Statistical Areas (a standard Census Bureau designation of the region around a city) in the United States, collected from a variety of sources.

- Outcome variable: Mortality
- Data collected on possible predictors: social and economic conditions, climate and indices of air pollution
- Question: How is mortality in an SMSA related to the median education level of the population in the SMSA?

Scatterplot of Mortality versus Education

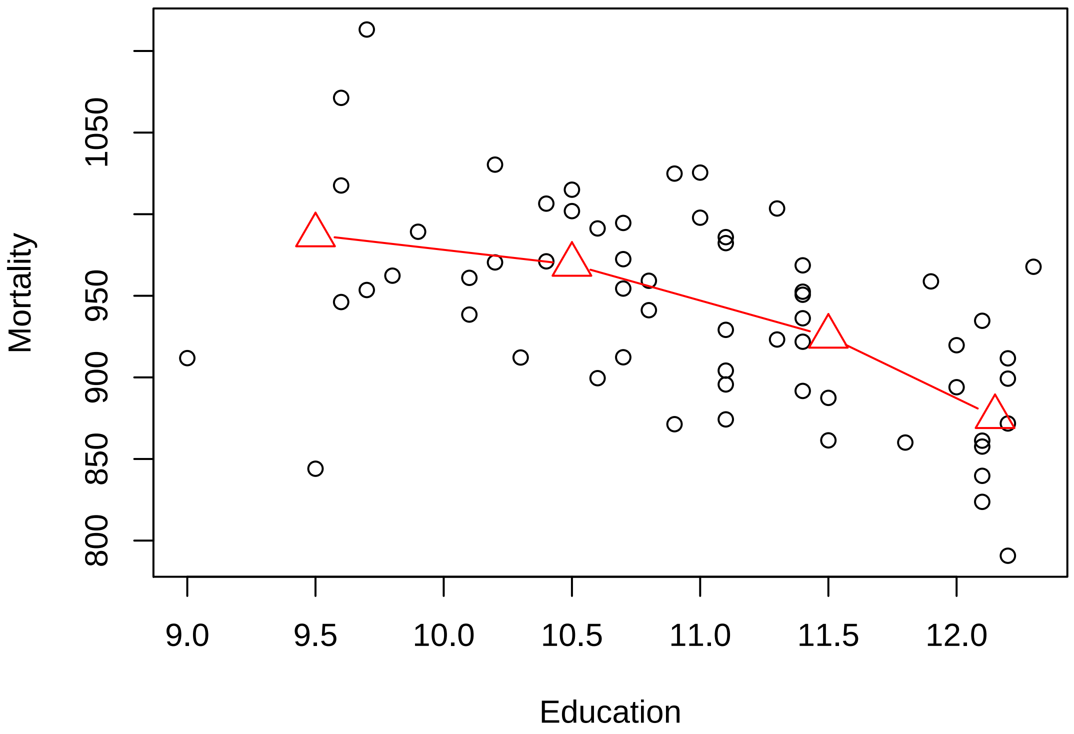


Descriptives for Mortality in Education Strata

Median years of education	Number in strata	Mean mortality	Standard deviation
8-10	9	978.81	81.27
10-11	21	969.13	44.48
11-12	20	925.08	41.93
12+	10	875.83	53.31

Plot of Mean mortality versus Yrs. Educ.

```
smsa <- read.table("smsa.dat",header=T)
plot(smsa$Education,smsa$Mortality, xlab="Education",  ylab="Mortality")
m1=tapply(smsa$Mortality,
cut(smsa$Education,breaks=c(8,seq(10,13,1))),mean)
points(c(9,10.5,11.5,12.5), m1, pch=2, cex=2, col=2, type="b")
```



Least Squares Estimation

How do we estimate the parameters in

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i?$$

We want to minimize the distance between the observed Y_is and their fitted values, $\beta_0 + \beta_1 X_i$.

For the ith observation, this distance is expressed as

$$(Y_i - (\beta_0 + \beta_1 X_i))^2.$$

But we want to determine this over all observations.

Obtaining least squares estimates

Minimize

$$S^2 = \sum_{i=1}^{N} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

Set the first derivatives equal to 0

$$\frac{\partial S^2}{\partial \beta_0} = -2 \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\frac{\partial S^2}{\partial \beta_1} = -2 \sum_{i=1}^{N} X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

And solve for β_0 and β_1.

Least squares estimates

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Using these results, we get estimates of the fitted value of the ith observation

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

and the ith residual

$$e_i = Y_i - \hat{Y}_i.$$

Using these results, we can make statements about the relationship of the predictor and the outcome (the mean), but we cannot say much else without more assumptions.

Estimation of Least Squares Line

```
lm1 <- lm(Mortality~Education, data=smsa)
summary(lm1)
```

```
Call:
lm(formula = Mortality ~ Education, data = smsa)

Residuals:
     Min       1Q   Median       3Q      Max
-151.724  -37.099    2.419   43.813  124.909

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1353.158     91.423  14.801  < 2e-16 ***
Education    -37.619      8.307  -4.529 3.01e-05 ***
```

Interpretation of Output

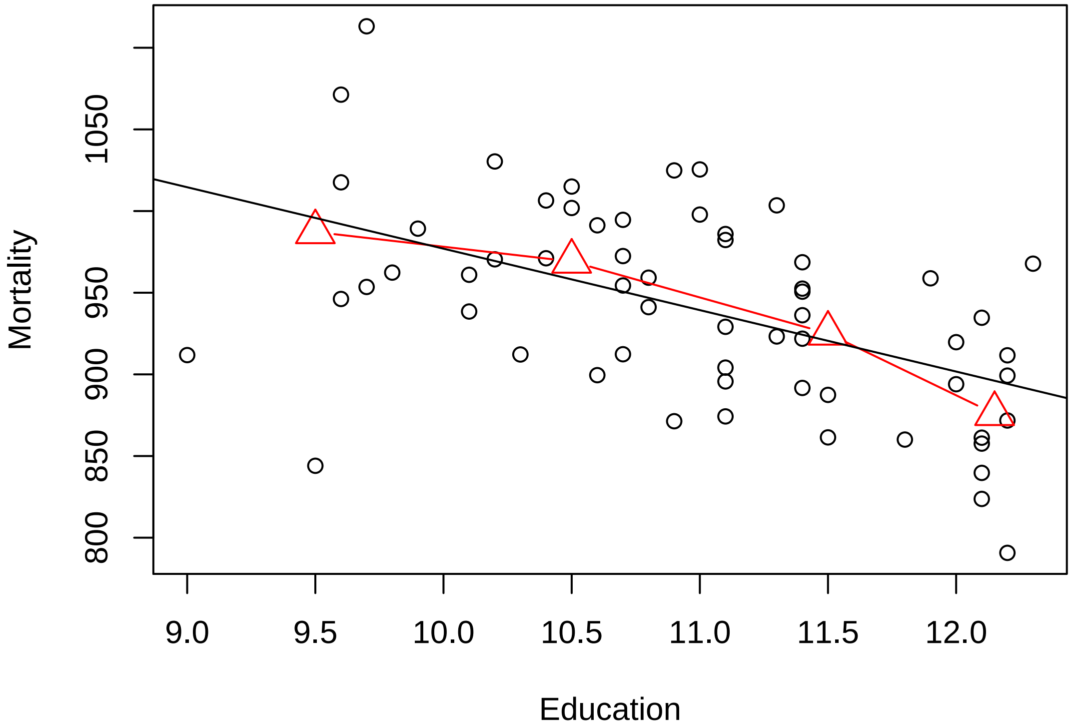
Estimates of regression parameters

- Intercept is labeled "(Intercept)"
  Estimated intercept: 1353.158

- The slope is labeled by its variable name: "Education"
  Estimated slope: -37.62

Superimposed Plot of Least Squares Line

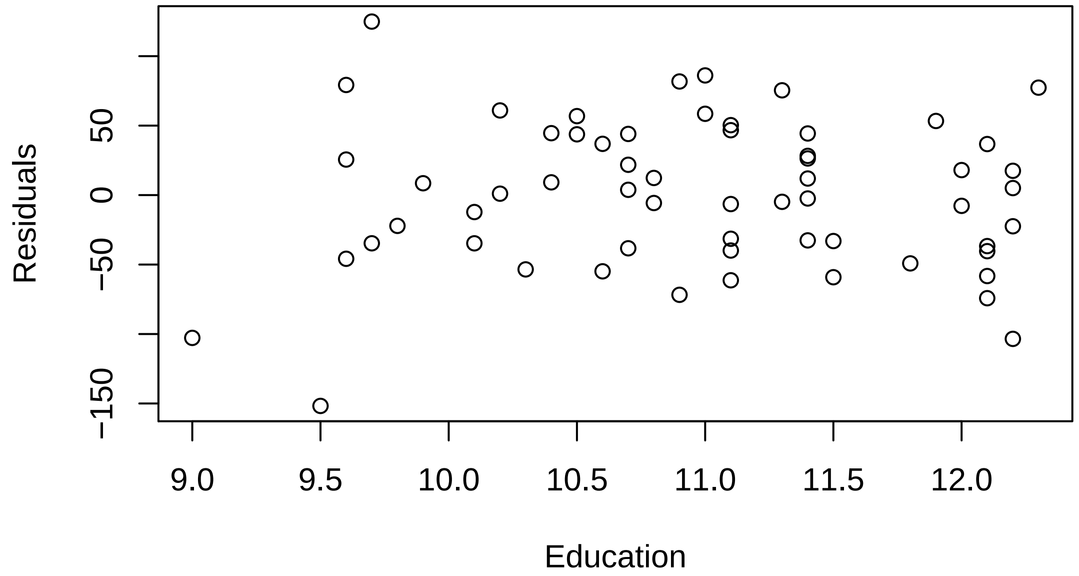
```
plot(smsa$Education,smsa$Mortality, xlab="Education", ylab="Mortality")
points(c(9.5,10.5,11.5,12.15),m1,pch=2,cex=2,col=2,type="b")
abline(coef(lm1))
```



Graphical examination of the model

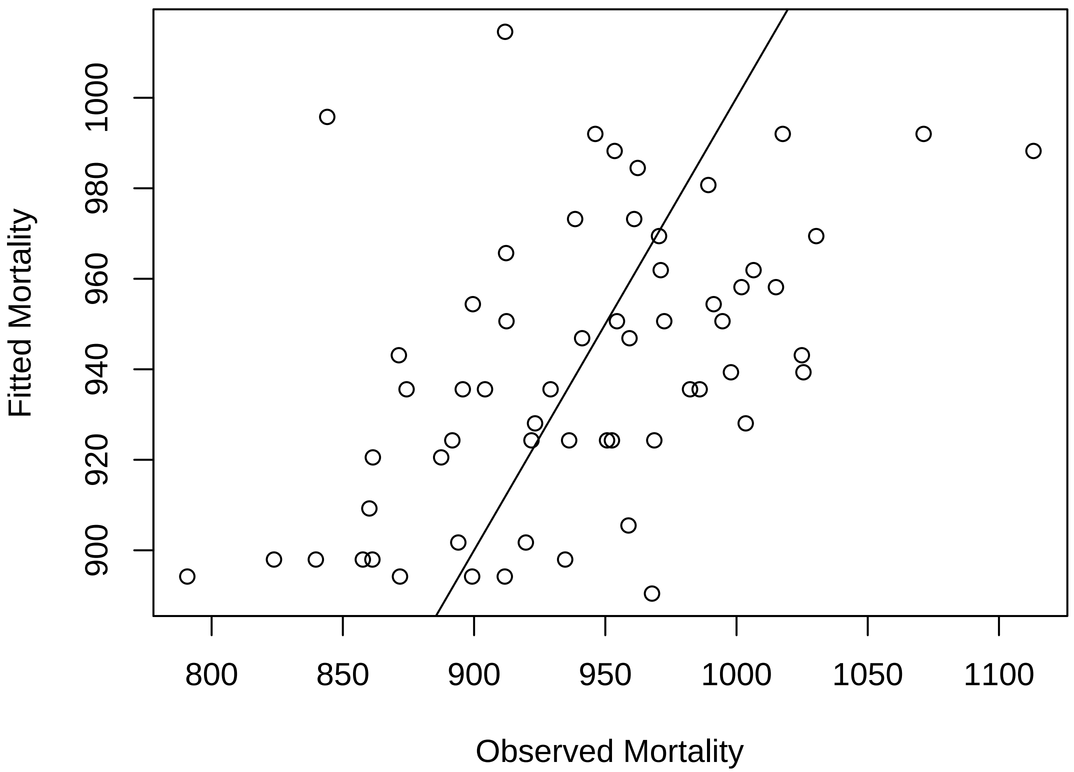
Plotting residuals against the predictor

```
resids=smsa$Mortality-fitted(lm1)
plot(smsa$Education,resids,xlab="Education", ylab="Residuals")
```



Plotting the fitted outcome against the observed outcome

```
plot(smsa$Mortality,fitted(lm1),xlab="Observed Mortality", ylab="Fitted Mortality")
```



Inference

In general, a point estimate is not very useful. We require a measure of the precision of the estimate.

The least squares estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$ may be expressed as

$$\hat{\beta}_0 = \sum_{i=1}^{N} l_i Y_i$$

and

$$\hat{\beta}_1 = \sum_{i=1}^{N} k_i Y_i,$$

where

$$l_i = \frac{1}{N} - \frac{\bar{x}(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

and

$$k_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}.$$

It is easily show that the least squares estimators are $unbiased$ since

$$E[\hat{\beta}_0] = \sum_{i=1}^{N} l_i E[Y_i] = \beta_0$$

and

$$E[\hat{\beta}_1] = \sum_{i=1}^{N} k_i E[Y_i] = \beta_1$$

where $\sum_i l_i = 1$, $\sum_i l_i x_i = 0$, $\sum_i k_i = 0$ and $\sum_i k_i x_i = 1$. Note that this dervation required no assumptions about the second moments of Y_i.

Variance of least squares estimators

Following the previous derivations we have

$$\text{var}(\hat{\beta}_0) = \sigma^2 \left\{ \frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^N (x_i - \bar{x})^2} \right\} = \sigma^2 c_0^2$$

$$\text{var}(\hat{\beta}_1) = \sigma^2 \left\{ \frac{1}{\sum_{i=1}^N (x_i - \bar{x})^2} \right\} = \sigma^2 c_1^2,$$

where $c_0^2 = \sum_i l_i^2$ and $c_1^2 = \sum_i k_i^2$.

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = \sigma^2 \left\{ -\frac{\bar{x}}{\sum_{i=1}^N (x_i - \bar{x})^2} \right\}$$

So far, we haven't made any distributional assumptions about ϵ_i. If we assume normality ($\epsilon_i \sim N(0, \sigma^2)$), then the least squares estimators are normally distributed.
Alternatively,

- If we have a large sample size, asymptotic normality may be assumed for the estimators.

- If asymptotic normality does not hold, bootstrap or Monte Carlo methods may be appropriate.

Confidence intervals

If β_0 and β_1 are normally distributed and σ^2 is known, we can construct the following $100(1-\alpha)\%$ *confidence intervals*

$$\hat{\beta}_j \pm Z_{1-\alpha/2} \times \sqrt{\mathsf{var}(\hat{\beta}_j)},\ j = 0, 1$$

In general, σ^2 is unknown. An unbiased estimate is given by

$$\hat{\sigma}^2 = \frac{1}{N-2} \sum_{i=1}^{N} e_i^2 = \frac{1}{N-2} \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{\mathsf{RSS}}{N-2},$$

where RSS is the residual sums of squares. $\hat{\sigma}^2$ is also known as MSE (mean sqare error).

It can be shown that

$$\frac{RSS}{\sigma^2} = \frac{(N-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-2}.$$

Confidence intervals for least squares estimates with unknown σ^2

The relevant $100(1-\alpha)\%$ confidence intervals are given by

$$\hat{\beta}_j \pm t_{N-2,1-\alpha/2} \times \text{s.}\hat{\text{e}}.(\hat{\beta}_j),\ \ j = 0, 1, \tag{1}$$

where $t^{N-2}(1-\alpha/2)$ denotes the $1-\alpha/2$ point of the standard t-distribution with $N-2$ degrees of freedom and $\text{s.}\hat{\text{e}}.(\hat{\beta}_j) = \hat{\sigma} \times c_j$.

From the SMSA example, we can now calculate a confidence interval for the estimates of the slope and intercept.

Parameter	Formula	95% CI
β_0	$1353.158 \pm 2.00 \times 91.423$	(1334.3, 1372.0)
β_1	$-37.619 \pm 2.00 \times 8.307$	(-54.2, -21.0)

Confidence interval for a point on the regression line

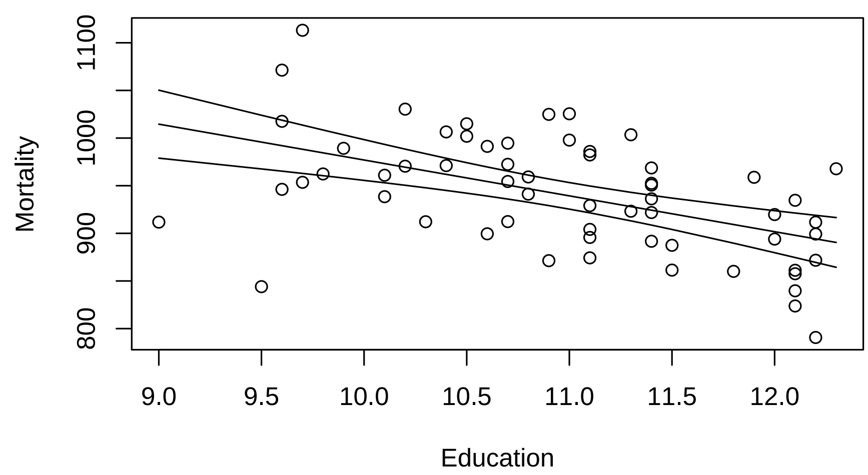
$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i \\ &= \bar{Y} + \hat{\beta}_1 (x_i - \bar{x}) \\ \text{var}(\hat{Y}_i) &= \text{var}(\bar{Y}) + (x_i - \bar{x})^2 \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \\ &= \sigma^2 \left[\frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right]\end{aligned}$$

The $100(1-\alpha)\%$ confidence interval for $\hat{Y}_i$ is

$$\hat{Y}_i \pm t_{N-2,1-\alpha/2} \hat{\sigma} \sqrt{\frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}.$$

For the SMSA example:

```
N=dim(smsa)[1]
SSXi=(smsa$Education-mean(smsa$Education))^2
SSX=sum(SSXi)
plot(smsa$Education,smsa$Mortality,xlab="Education", ylab="Mortality")
ord=order(smsa$Education)
lines(smsa$Education[ord],fitted(lm1)[ord])
for(i in c(-1,1))lines(smsa$Education[ord],(fitted(lm1)+i*qt(.025,N-2)*53.94*
        sqrt(1/N+SSXi/SSX))[ord])
```



Hypothesis Testing for least squares estimates

Similar to the approach for obtaining confidence intervals for β_j, we find that

$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\text{s.e.}}(\hat{\beta}_1)} \sim t^{N-2}. \tag{2}$$

Now we can construct hypothesis tests for the regression parameters. From the SMSA example:

Test: $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$

Under the null hypothesis,

$t_{obs} = \frac{\hat{\beta}_1 - \beta_1}{\hat{\text{s.e.}}(\hat{\beta}_1)} \sim t^{N-2} = -37.619/8.307 = -4.529$. To perform an $\alpha = .05$ level test we compare t_{obs} (our observed value of (2)) to $t^{N-2}(\alpha/2) = -2.00$ which is not as extreme as t_{obs}; therefore, we reject the null hypothesis.