Regression STAT 415/615

Chapter 2

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$\varepsilon_i \sim N(0, \sigma^2)$ Assumption

We would like to make inference.

2.1. Inferences Concerning β_1

Ex. study relationship between sales Y and advertising expenditures X.

We would like to get an estimate of β_1

$\beta_1 \rightsquigarrow$ provides information as to how many additional sales dollars, on average, are generated by an additional dollar of advertising expenditure.

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

When $\beta_1 = 0$ there is no linear association between Y & X.

Before discussing inference concerning β_1 we need sampling distribution of b_1, the point estimator of β_1.

Sampling distribution of b_1

The sampling distribution of b_1 refers to the different values of b_1 that would be obtained with repeated sampling. "like we showed in R demonst."

b_1 is a linear combination of Y_i and each Y_i is normally distributed $\Rightarrow$ b_1 is normally distributed. "Gauss-Markov"

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$E\{b_1\} = \beta_1 \quad \text{and} \quad \sigma^2\{b_1\} = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Normality

b_1 is a linear combination of Y_i

Thus since Y_i are independently normally distributed then a linear combination of indep. normal random variables is normally distributed

Start

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_1 - \bar{X})^2}$$

$$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} (X_i - \bar{X}) Y_i - \underbrace{\sum_{i=1}^{n} (X_i - \bar{X})}_{0} \bar{Y} =$$

$$= \sum_{i=1}^{n} (X_i - \bar{X}) Y_i$$

Thus $b_1 = \sum_{i=1}^{n} k_i Y_i$ where $k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$

Mean.

$$E\{b_1\} = E\{\sum_{i=1}^{n} k_i Y_i\} = \sum_{i=1}^{n} k_i E\{Y_i\} = \sum_{i=1}^{n} k_i (\beta_0 + \beta_1 X_i)$$

$$= \beta_0 \sum k_i + \beta_1 \sum k_i X_i = \beta_1$$

but $\sum k_i = 0$ & $\sum k_i X_i = 1$

① $\sum k_i = \sum \left[\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right] = \frac{1}{\sum (X_i - \bar{X})^2} \cancel{\sum (X_i - \bar{X})}^{0} = 0.$

② $\sum k_i X_i = \sum \left[\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right] X_i = \cancel{\frac{1}{\sum (X_i - \bar{X})^2}} \cancel{\sum (X_i - \bar{X}) X_i} = 1.$

$$= \frac{1}{\underbrace{\sum (X_i - \bar{X})^2}} [\sum X_i^2 - \bar{X} \sum X_i]$$

$$\sum (X_i^2 - 2X_i\bar{X} + \bar{X}^2)$$

$$\sum X_i^2 - 2\bar{X} \sum X_i + n \cdot \bar{X} \cdot \bar{X}$$

$$\sum X_i^2 - 2\bar{X} \sum X_i + \cancel{n} \cdot \bar{X} \cdot \frac{\sum X_i}{\cancel{n}}$$

$$\sum X_i^2 - \bar{X} \sum X_i$$

Variance.

$$\sigma^2\{b_1\} = \sigma^2\left\{\sum_{i=1}^{n} k_i Y_i\right\} = \sum_{i=1}^{n} k_i^2 \cdot \sigma^2\{Y_i\} = \sum_{i=1}^{n} k_i^2 \cdot \sigma^2$$

$$= \sigma^2 \sum_{i=1}^{n} k_i^2 = \sigma^2 \frac{1}{\sum (X_i - \bar{X})^2} \qquad \left\{ k_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right.$$

$$\sum_{i=1}^{n} k_i^2 = \sum_{i=1}^{n} \left[\frac{X_i - \bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\right]^2 = \frac{1}{\left[\sum_{i=1}^{n} (X_i - \bar{X})^2\right]^{\cancel{2}}} \cdot \cancel{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Estimated Variance.

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum (X_i - \bar{X})^2} \xrightarrow[\text{estimator of } \sigma^2]{\text{unbiased}} s^2\{b_1\} = \frac{MSE}{\sum (X_i - \bar{X})^2}$$

Review of related distributions.

Let Y be a random variable that follows a normal distribution with $E\{Y\} = \mu$ and $\sigma^2\{Y\} = \sigma^2$

- The standard normal random variable is

$$\left\{ \sigma^2\left\{\frac{Y-\mu}{\sigma}\right\} = \frac{\sigma^2\{Y\} - 0}{\sigma^2} = 1 \right\} \quad Z = \frac{Y - \mu}{\sigma} \Rightarrow Z \sim N(0, 1)$$

you divide by σ^2.

- Let $Y_1, Y_2, \ldots, Y_n$ indep. normal $\Rightarrow a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$ is norm distributed with $\sum a_i E\{Y_i\}$ and variance $\sum a_i^2 \sigma^2\{Y_i\}$.

- Let $Z_1, Z_2, \ldots, Z_\nu$ be ν indep. standard normal.
A chi square random variable is defined as

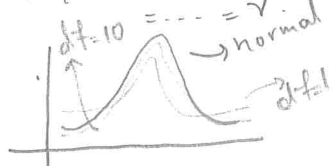
$$\chi^2(\nu) = Z_1^2 + Z_2^2 + \cdots + Z_\nu^2$$

ν is called degrees of freedom (df) $\rightarrow$ $E\{\chi^2(\nu)\} = \nu$

$\rightarrow$ for this proof start with $E\{\chi^2(\nu)\} = \int x \cdot f(x) dx = \cdots = \nu$.

- Let Z & $\chi^2(\nu)$ $\quad t(\nu) = \dfrac{Z}{\left[\dfrac{\chi^2(\nu)}{\nu}\right]^{1/2}}$ $\qquad E\{t(\nu)\} = 0$

Show here t distrib. vs normal ~~1~~



- For interval estimation we need t-distribution.

ex. Let $Y_1, \dots, Y_n$ observations of $Y \sim N(0,1)$

$$\Rightarrow \quad \bar{Y} = \frac{\sum X_i}{n} \quad \& \quad s = \left[\frac{\sum (Y_i - \bar{Y})^2}{n-1}\right]^{1/2}$$

$$\& \quad s\{\bar{Y}\} = \frac{s}{\sqrt{n}}.$$

We have that $\frac{\bar{Y} - \mu}{s\{\bar{Y}\}}$ is distributed as t with $n-1$ df.

The confidence limits for μ with conf coef. $1-\alpha$ are $\bar{Y} \pm t(1 - \frac{\alpha}{2}; n-1)\, s\{\bar{Y}\}$

Note: Similarly we have to work for conf interval of β_1.

Steps.

1. We need to find distribution of $\frac{\overset{\text{estimator.}}{b_1} - \overset{\text{parameter like } \mu.}{\beta_1}}{\underset{\text{estim. st. deviation.}}{s\{b_1\}}}$

Like previously if Y_i come from same normal population then $\frac{\bar{Y} - \mu}{s\{\bar{Y}\}}$ follows t distribution with $n-1$ degrees of freedom

df is $n-1$ because only one parameter needs to be estimated

For the regression model we need to estimate two parameters thus we have $df = n-2$ (since two df are lost).

In addition b_1 is a linear combination of Y_i therefore

$\frac{b_1 - \beta_1}{s\{b_1\}}$ is distributed as t with $n-2$ degrees of freedom.

2. Confidence interval.

Similar to $\bar{Y} \pm t(1-\frac{\alpha}{2}; n-1)\, s\{\bar{Y}\}$.

$$b_1 \pm t(1-\frac{\alpha}{2}; n-2)\, s\{b_1\}$$

3. Tests concerning β_1.

Test statistic (TS) for testing means often takes the form.

$$TS = \frac{EST - HYP}{SE}$$

estimate for parameter

hypothesized value of par

standard error

So

$$H_0: \beta_1 = \beta_{10} \quad \text{Vs} \quad H_1: \beta_1 \neq \beta_{10}.$$

We use test statistic.

$$t = \frac{b_1 - \beta_{10}}{\sqrt{s^2\{b_1\}}} = \frac{b_1 - \beta_{10}}{s\{b_1\}}.$$

t-distribution with $n-2$ df.

where $s^2\{b_1\} = \dfrac{MSE}{\sum(X_i - \bar{X})^2}$.

Ex (*) on page 6.

2.2. Inference concerning β_0.

Recall that $b_0 = \bar{Y} - b_1\bar{X}$

$$E\{b_0\} = \beta_0 \quad \text{and} \quad \sigma^2\{b_0\} = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]$$

Estimator of $\sigma^2\{b_0\}$ by $s^2\{b_0\} = MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]$.

Sampling distribution of $\frac{b_0 - \beta_0}{s\{b_0\}}$.

$$\frac{b_0 - \beta_0}{s\{b_0\}} \quad t \text{ distrib } df = n-2$$

Confidence interval for β_0

$$b_0 \pm t(1-\frac{\alpha}{2}; n-2)\, s\{b_0\}$$

Hypothesis tests.

$$H_0 : \beta_0 = \beta_{00} \qquad H_1 : \beta_0 \neq \beta_{00}.$$

The test statistic is

$$t = \frac{b_0 - \beta_{00}}{\sqrt{MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]}}$$

---

* Tests concerning β_1

Ex. Two sided-test.

$$H_0 : \beta_1 = \underset{\beta_{10}}{0} \qquad H_1 : \beta_1 \neq 0.$$

$$t = \frac{b_1 - \beta_{10}^{\,=0}}{s\{b_1\}} = \frac{b_1}{s\{b_1\}}.$$

The decision rule with this test statistic is

If $|t| \leq t(1-\frac{\alpha}{2}; n-2)$ conclude H_0

If $|t| > t(1-\frac{\alpha}{2}; n-2)$ conclude H_1

2.4. Interval Estimation of $E\{Y_h\}$.

Let X_h denote level of X for which we wish to estimate the mean response.

Point estimator $\hat{Y}_h$ of $E\{Y_h\}$ is given by

$$\hat{Y}_h = b_0 + b_1 X_h .$$

Normality.

The normality of the sampling distribution of $\hat{Y}_h$ follows directly from the fact that $\hat{Y}_h$, like is a linear combination of the observations Y_i.

Mean.

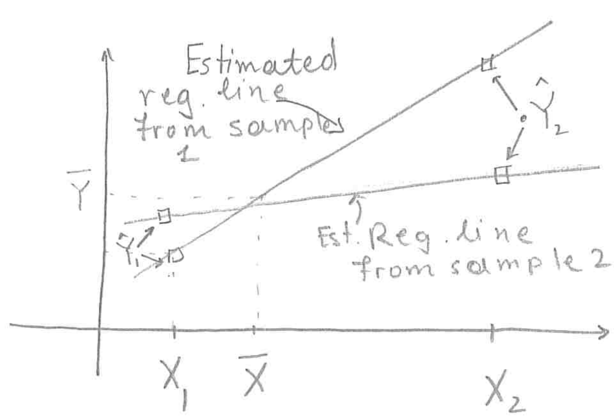
$$E\{\hat{Y}_h\} = E\{b_0 + b_1 X_h\} = \beta_0 + \beta_1 X_h .$$ $\hat{Y}_h$ is unbiased estim. of $E\{Y_h\}$.

Variance.

$$\sigma^2\{\hat{Y}_h\} = \sigma^2 \cdot \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right].$$

Note: The variability of the sampling distribution of $\hat{Y}_h$ is affected by how far X_h is from $\bar{X}$ through $(X_h - \bar{X})^2$.

Ex



When MSE is subtit. for σ^2 we get

$$s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

Confidence Interval.

We define $\dfrac{\hat{Y}_h - E\{Y_h\}}{s\{\hat{Y}_h\}}$ t-dist. with $n-2$ d.o.f.

$$\hat{Y}_h \pm t\left(1-\frac{\alpha}{2}; n-2\right) \cdot s\{\hat{Y}_h\}$$

Prediction Interval for New Observations.

Objective: Prediction of new observation Y corresponding to a given level X of the predictor variable.

The new observation on Y to be predicted is viewed as the result of a new trial independent of the trials on which the regression analysis is based.

Let X_h be the level of X for new trial and new observation on Y as $Y_{h(new)}$

Goal: Predict an individual outcome drawn from the distribution of Y.

Note: In the previous case we were estimating $E\{Y_h\}$ by $\hat{Y}_h$

Our best guess for a new observation is still $\hat{Y}_h$. The estimated mean is still the best prediction we can make.

The difference is in the amount of variability.

Thus. $\overset{var}{\sigma^2}\{Y_{h(new)} - \hat{Y}_h\} = \overset{var}{\sigma^2}\{Y_{h(new)}\} + \overset{var}{\sigma^2}\{\hat{Y}_h\} = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]$

$$\Rightarrow \sigma^2\{Y_{h(new)} - \hat{Y}_h\} = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]$$

$$\& \quad s^2\{Y_{h(new)} - \hat{Y}_h\} = MSE\left[1 + \frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right].$$

Mean response at X_h

Want to estimate $E\{Y_h\}$

Point estimator is $\hat{Y}_h$

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\Sigma(X_i-\bar{X})^2}\right]$$

$$s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\Sigma(X_i-\bar{X})^2}\right]$$

CI: $\hat{Y}_h \pm t\left(1-\frac{\alpha}{2}; n-2\right)\cdot s\{\hat{Y}_h\}$

New observation at X_h

Want to predict $Y_{h(new)}$ → r. variable drawn from Y

$$s^2\{\underbrace{Y_{h(new)} - \hat{Y}_h}_{\text{prediction.}}\} = \underbrace{s^2\{Y_{h(new)}\}}_{MSE} + \sigma^2\{\hat{Y}_h\}$$

CI: $\hat{Y}_h \pm t\left(1-\frac{\alpha}{2}; n-2\right)\cdot s\{\text{predict}\}$

Note: This will be wider.

It accounts for both the uncertainty in knowing the value of the population mean + $\underbrace{\text{data scatter}}_{\text{variability}}$.

Confidence Band for Regression Line

Obtain a confidence band for the entire regression line $E\{Y\} = \beta_0 + \beta_1 X$.

Why? It helps us to determine the appropriateness of a fitted regression function

$$\hat{Y}_h \pm W s\{\hat{Y}_h\}$$

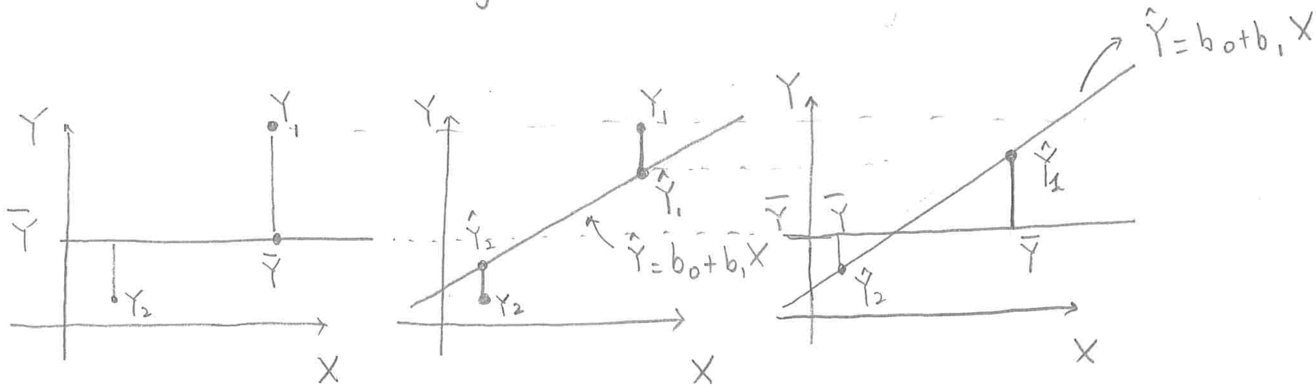
where $W^2 = 2F(1-\alpha; 2, n-2)$

In the case of a simple linear regression is equivalent to another test, the F test for the significance of the regression.

This equivalence is true only for simple linear regression

Let's start with $Y_i - \bar{Y}$ which measures the deviation of an observation from the sample mean

$$\underbrace{Y_i - \bar{Y}}_{\text{Total deviation}} = \underbrace{Y_i - \hat{Y}_i}_{\text{Deviation around fitted regression line.}} + \underbrace{\hat{Y}_i - \bar{Y}}_{\text{Deviation of fitted regression value around mean.}}$$



It can be shown that the sums of these squared deviations have the same relationship.

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

proof

$$\sum (Y_i - \bar{Y})^2 = \sum [(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})]^2 = \sum [(\hat{Y}_i - \bar{Y})^2 + 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) + (Y_i - \hat{Y}_i)^2]$$

$$= \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 + 2\sum (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)$$

but $2\sum (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = 2 \underbrace{\sum \hat{Y}_i (Y_i - \hat{Y}_i)}_{\substack{= 0 \\ \text{because} \\ \sum \hat{Y}_i e_i = 0}} - 2\bar{Y} \underbrace{\sum (Y_i - \hat{Y}_i)}_{\substack{= 0 \\ \text{because} \\ \sum e_i = 0.}}$

$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{} = \underbrace{\sum(Y_i-\hat{Y}_i)^2}_{} + \underbrace{\sum(\hat{Y}_i-\bar{Y})^2}_{}$$

$$\underbrace{SST}_{\text{Total sum of squares.}} = \underbrace{SSE}_{\text{sum of square errors}} + \underbrace{SSR}_{\text{regression sum of squares}}$$

ANOVA Table.

Source of Variation	Sum of Squares	Degrees of freedom	Mean Square	F
Regression	$\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$	$p-1 \to \overset{p=2}{\boxed{1}}$	$\frac{SSR}{1}=MSR$	$\frac{MSR}{MSE}$
Error	$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$	$n-p \Rightarrow \boxed{n-2}$	$MSE=\frac{SSE}{n-2}$	
Total	$\sum_{i=1}^{n}(Y_i-\bar{Y})^2$	$n-\cancel{p}+\cancel{p}-1=\boxed{n-1}$		

$$F=\frac{MSR}{MSE}=\frac{\frac{SSR}{1}}{\frac{SSE}{n-2}}=\frac{\frac{\sum(\hat{Y}_i-\bar{Y})^2}{1}}{\frac{\sum(Y_i-\hat{Y}_i)^2}{n-2}} \sim \underbrace{F_{1,n-2}}_{\text{F-distribution with deg. of freedom } 1 \,\&\, n-2}$$

Why $p-1$?: This corresponds to the fact that we specify a line by two points.

We will prove this in multiple regression.

Significance of the regression test.

Idea.

Comparison of MSR and MSE is useful for testing whether or not $\beta_1=0$. If MSR and MSE are of the same order of magnitude, this would suggest that $\beta_1=0$

If MSR is substantially greater than MSE, this would suggest that $\beta_1\neq 0$.

Note that: $E\{MSE\}=\sigma^2$

& $E\{MSR\}=\sigma^2+\beta_1^2\sum(X_i-\bar{X})^2$

If $\beta_1=0 \Rightarrow E\{MSE\}=E\{MSR\}$

OR $SSR=b_1^2\sum(X_i-\bar{X})^2$
If $b_1=0 \Rightarrow SSR=0$
$\Downarrow$
$MSR=0$

So it makes sense to compare them by

$$F^*=\frac{MSR}{MSE}$$

$H_0: \beta_1=0 \qquad H_a: \beta_1\neq 0$

If $F^*\leq F(1-\alpha;1,n-2)$ conclude H_0

If $F^*> F(1-\alpha;1,n-2)$ conclude H_a

F^* is always positive.

Ex. The time it takes to transmit a file always depends on the file size. Suppose you transmitted 30 files with the average size of 126 Kbytes and the st. deviation of 35 Kbytes. The average transmittance time was 0.04 seconds with s.d. of 0.01 second. The correlation coef. between the time & size was 0.86. In previous HWK we fit a reg. model that predicted the time it will take to transmit a 400 kbyte file.

We are given $n=30$, $s\{X\}=35$, $s\{Y\}=0.01$, and $r=0.86$.

(a) Compute the total, regression, and error sum of squares.

$$SST=\sum_{i=1}^{n}(Y_i-\bar{Y})^2=(n-1)\cdot s^2\{Y\}=29\cdot 0.01^2=0.0029$$

$$SSR=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 \ ?$$

$$r^2=\frac{[\sum(X_i-\bar{X})(Y_i-\bar{Y})]^2}{[\sum(X_i-\bar{X})\sum(Y_i-\bar{Y})]^2}=\frac{\frac{[\sum(X_i-\bar{X})(Y_i-\bar{Y})]^2}{\sum(X_i-\bar{X})^2}}{\sum(Y_i-\bar{Y})^2}=\frac{?}{SST}$$

$$SSR=\sum_{i=1}^{n}(\underbrace{\hat{Y}_i}-\bar{Y})^2=\sum_{i=1}^{n}(b_1(X_i-\bar{X}))^2=\sum_{i=1}^{n}b_1^2(X_i-\bar{X})^2$$

$$=b_1^2\sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$\cancel{\bar{Y}}+b_1(X_i-\bar{X})-\cancel{\bar{Y}}$$

normal equ

$$=b_1\cdot b_1\cdot\sum(X_i-\bar{X})^2$$

$$b_1=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}$$

$$=b_1\cdot\sum(X_i-\bar{X})(Y_i-\bar{Y})$$

$$=\frac{[\sum(X_i-\bar{X})(Y_i-\bar{Y})]^2}{\sum(X_i-\bar{X})^2}$$

$Y_i=b_0+b_1X_i+\varepsilon_i \Rightarrow$

$\Rightarrow Y_i=b_0+b_1X_i-b_1\bar{X}+b_1\bar{X}+\varepsilon_i$

$\Rightarrow Y_i=\underbrace{b_0+b_1\bar{X}}+b_1(X_i-\bar{X})+\varepsilon_i$

$\Rightarrow Y_i=b_0^*+b_1(X_i-\bar{X})+\varepsilon_i$

For point est we have $b_0^*=b_0+b_1\bar{X}\Rightarrow$

$\Rightarrow b_0^*=\bar{Y}-b_1\bar{X}+b_1\bar{X}=\bar{Y}$

$\Rightarrow \hat{Y}_i=\bar{Y}+b_1(X_i-\bar{X})$

Thus $r^2=\frac{SSR}{SST}$

$\Rightarrow SSR=r^2\cdot SST=0.86\cdot 0.0029$

$=0.00214$

$SSE = SST - SSR = 0.00076$

(b) Compute the ANOVA table.

Sum of squares	DF	Mean. Sq	F
SSR	1	0.00214	$\dfrac{MSR}{MSE} = 79.3$
SSE	$n-2=28$	0.000027	
SST	$n-1=29$		

(c) Use the F-statistic to test significance of our reg. model that relates transmittion time to the size of the file. State H_0 and H_1, and draw conclusion. For $1-\alpha = 0.95$

$$H_0 : b_1 \neq 0 \qquad H_a : b_1 \neq 0 .$$

We have $F^* = 79.3$ $\qquad F(1-\alpha; 1, 28) = 4.17 < F^*$

Reject H_0. The slope is significant. There is evid. of linear relation between X & Y.

(d) Coefficient of Determination.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

The coefficient of determination is interpreted as the proportion of observed variation in Y that can be explained by the simple linear regression model.

Here $R^2 = \dfrac{0.00214}{0.0029} = 0.738$.

It means that 73.8% of total variation of transmission times is explained solely by the file sizes.