Notes on Linear Regression

April 1, 2019

0.1 Least square Estimates

The basic assumption in linear regression is

$$\mathbb{E}(Y \mid X = x) = \alpha + \beta x. \tag{1}$$

This assumption essentially says that a linear relationship between X and Y, except for some error, is a reasonable assumption. We therefore consider the following model for further analysis

$$Y = \alpha + \beta x + \varepsilon, \tag{2}$$

where $\mathbb{E}(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. Note that under Model(2), $\mathbb{E}(Y \mid X = x) = \alpha + \beta x$ and $Var(Y \mid X = x) = \sigma^2$. When we observe a random sample of pairs (X_i, Y_i), $i = 1, 2, \cdots, n$, we have

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \tag{3}$$

where we now assume that ε_i are n independent copies of ε. Using the method of least square, we should find estimates of α and β such that

$$f(a, b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

is minimized. As we learned in class after differentiation and solving the resulting equations, the so-called *least squares normal equations*, we obtain,

$$\begin{cases} \frac{\partial f}{\partial a} & = \sum_{i=1}^{n} -2\left[y_i - (a + bx_i)\right] = 0 \\ \frac{\partial f}{\partial b} & = \sum_{i=1}^{n} -2x_i\left[y_i - (a + bx_i)\right] = 0 \end{cases}$$

which, after some simplification, lead us to the following system of equations

$$\begin{cases} \overline{y} & = a + b\overline{x} \\ \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2. \end{cases}$$

Define

$$S_{XY} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \qquad S_{XX} = \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Some simplification results in

$$S_{XY} = \sum_{i=1}^{n} x_i y_i - n\overline{xy} \qquad S_{XX} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2.$$

Then

$$\begin{cases} \overline{y} & = a + b\overline{x} \\ \sum_{i=1}^{n} x_i y_i = an\overline{x} + b\left(S_{XX} + n\overline{x}^2\right) = n\overline{x}(a + b\overline{x}) + bS_{XX}. \end{cases}$$

Using the first equation in the above system, the second equation is simplified to

$$\sum_{i=1}^{n} x_i y_i = n\overline{xy} + bS_{XX},$$

and hence

$$S_{XY} = \sum_{i=1}^{n} x_i y_i - n\overline{xy} = bS_{XX}$$

which give us

$$a = \overline{y} - b\overline{x} \qquad b = \frac{S_{XY}}{S_{XX}}. \tag{4}$$

Following we show that b is an unbiased estimator of β. First note that

$$\sum_{i=1}^{n} \left(x_i - \overline{x}\right) \overline{y} = 0,$$

and hence

$$S_{XY} = \sum_{i=1}^{n} \left(x_i - \overline{x}\right) y_i.$$

Thus we have

$$\mathbb{E}\left(b | x_1, \cdots, x_n\right) = \frac{1}{S_{XX}} \sum_{i=1}^{n} \left(x_i - \overline{x}\right) \mathbb{E}\left(Y_i | x_i\right).$$

On the other hand,

$$\mathbb{E}\left(Y_i | x_i\right) = \alpha + \beta x_i,$$

and therefore

$$\begin{aligned}\mathbb{E}\left(b|x_i,\cdots,x_n\right) &= \frac{1}{S_{XX}}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(\alpha+\beta x_i\right)\\ &= \frac{1}{S_{XX}}\left[\alpha\sum_{i=1}^{n}(x_i-\overline{x})+\beta\sum_{i=1}^{n}(x_i-\overline{x})x_i\right].\end{aligned}$$

Note that the first summation in the square brackets is zero. The second summation is equal S_{XX}. Thus

$$\mathbb{E}\left(b|x_1,\cdots,x_n\right) = \frac{1}{S_{XX}}\left[\beta S_{XX}\right] = \beta, \tag{5}$$

which establishes unbiasedness of b as an estimator of β. To find the variance of b, we notice once again that $S_{XY}=\sum_{i=1}^{n}(x_i-\overline{x})y_i$. Then

$$Var(b|x_1,\cdots,x_n) = \frac{1}{S_{XX}^2}\sum_{i=1}^{n}(x_i-\overline{x})^2 Var(Y_i|x_i) = \frac{S_{XX}\sigma^2}{S_{XX}^2} = \frac{\sigma^2}{S_{XX}}. \tag{6}$$

Note that

$$b=\sum_{i=1}^{n}\frac{x_i-\overline{x}}{S_{XX}}Y_i=\sum_{i=1}^{n}\frac{x_i-\overline{x}}{\sum_{j=1}^{n}(x_j-\overline{x})^2}Y_i=\sum_{i=1}^{n}\gamma_{n,i}Y_i, \tag{7}$$

where

$$\gamma_{n,i}=\frac{x_i-\overline{x}}{\sum_{j=1}^{n}(x_j-\overline{x})^2},$$

and hence b is a linear combination of Y_i.

Assumption 1: $S_{XX}/n\to\kappa^2<\infty$ as $n\to\infty$.

Suppose Assumption 1 holds. Then using Chebyshev's inequality and (5) we have, for any $\epsilon>0$,

$$P(|b-\beta|>\epsilon|x_1,\cdots,x_n)\le\frac{Var(b|x_1,\cdots,x_n)}{\epsilon^2}\overset{\text{using (6)}}{=}\frac{\sigma^2}{\epsilon^2 n\frac{S_{XX}}{n}}\to 0\quad\text{as}\quad n\to\infty,$$

which means that b is a consistent estimator of β.

Using (7) and CLT (a more general form than the one we cover in Math323 and 324) we can establish that

$$\sqrt{n}(b-\beta)\overset{W}{\to}N\left(0,\frac{\sigma^2}{\kappa^2}\right)\quad\text{as}\quad n\to\infty.$$

which means that for large enough n,

$$b \overset{app}{\sim} N(\beta, \frac{\sigma^2}{S_{XX}}), \tag{8}$$

Using this result we can make the following $(1-\lambda) \times 100\%$ confidence interval for β,

$$b \overset{+}{-} \mathfrak{z}_{\lambda/2} \frac{\sigma}{\sqrt{S_{XX}}},$$

where $P(Z > \mathfrak{z}_{\lambda/2}) = \lambda/2$ and $Z \sim N(0,1)$. To apply the above confidence interval, we need to estimate σ^2. We discuss estimating σ^2 later in this section.

Similarly we can study a. We first notice that $a = \overline{y} - b\overline{x}$. Then

$$\begin{aligned} \mathbb{E}(a|x_1, \cdots, x_n) &= \mathbb{E}(\overline{Y}|x_1, \cdots, x_n) - \overline{x}\ \mathbb{E}(b|x_1, \cdots, x_n) \\ &= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} Y_i|x_1, \cdots, x_n\right) - \beta\overline{x} \\ &= \frac{1}{n}\sum_{i=1}^{n}[\alpha + \beta x_i] - \beta\overline{x} = \alpha. \end{aligned} \tag{9}$$

Next, we find $Var(a)$. We need first show that $Cov(\overline{y}, b) = 0$. Note once again that $S_{XY} = \sum_{i=1}^{n} x_i(y_i - \overline{y})$. On the other hand,

$$\begin{aligned} Cov(\overline{Y}, Y_i - \overline{Y}|x_1, \cdots, x_n) &= Cov(\overline{Y}, Y_i|x_1, \cdots, x_n) - Var(\overline{Y}|x_1, \cdots, x_n) \\ &= \frac{1}{n}Var(Y_i|x_i) - \frac{1}{n^2}\sum_{j=1}^{n} Var(Y_j|x_j) \\ &= \frac{\sigma^2}{n} - \frac{n\sigma^2}{n^2} = 0. \end{aligned}$$

Thus using Theorem 5.12(c) from Chapter 5 we have

$$\begin{aligned} Cov\left(\overline{Y}, b|x_1, \cdots, x_n\right) &= Cov\left(\overline{Y}, \frac{\sum_{i=1}^{n} x_i(Y_i - \overline{Y})}{S_{XX}}|x_1, \cdots, x_n\right) \\ &= \sum_{i=1}^{n} \frac{x_i}{S_{XX}} Cov\left(\overline{Y}, Y_i - \overline{Y}|x_1, \cdots, x_n\right) = 0 \end{aligned}$$

To calculate $Var(a)$ we then have

$$\begin{aligned} Var(a|x_1,\cdots,x_n) &= Var(\overline{Y} \mid x_1,\cdots,x_n) + \overline{x}^2 Var(b \mid x_1,\cdots,x_n) \\ &= \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{S_{XX}} = \frac{\sigma^2}{n}(1 + \frac{\overline{x}^2}{S_{XX}/n}). \end{aligned} \tag{10}$$

Suppose Assumption 1 holds. Then using Chebyshev's inequality and (9) we have, as $n \to \infty$

$$P(|a-\alpha| > \epsilon|x_1,\cdots,x_n) \le \frac{Var(a|x_1,\cdots,x_n)}{\epsilon^2} \overset{\text{using (10)}}{=} \frac{\sigma^2}{\epsilon^2 n\left(1+\frac{\overline{x}}{S_{XX}/n}\right)} \to 0.$$

This establishes consistency of a.

To find asymptotic distribution of a, we first note that

$$a = \overline{y} - b\overline{x} = \sum_{i=1}^{n} \delta_{n,i} Y_i, \tag{11}$$

where $\delta_{n,i} = 1/n - \overline{x}\gamma_{n,i}$. Using (11) and CLT (a more general form than the one we cover in Math323 and 324) we can establish that

$$\sqrt{n}(a-\alpha) \overset{W}{\to} N\left(0, \sigma^2\left[1 + \frac{\overline{x}^2}{\kappa^2}\right]\right) \quad \text{as} \quad n \to \infty.$$

which means that for large enough n,

$$a \overset{app}{\sim} N\left(\alpha, \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right]\right), \tag{12}$$

Using this result we can make the following $(1-\lambda)\times 100\%$ confidence interval for α,

$$a \overset{+}{-} \mathfrak{z}_{\lambda/2}\sigma\sqrt{\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right]},$$

where $P(Z > \mathfrak{z}_{\lambda/2}) = \lambda/2$ and $Z \sim N(0,1)$.

For practical purposes we need to estimate σ^2. A natural estimate of σ^2 which is the variance of the error term, should be based on the squared of residuals, i.e.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left[y_i - (a + bx_i)\right]^2.$$

As discussed in class an unbiased consistent estimator of σ^2 is

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}, \tag{13}$$

where $\hat{y}_i = a + bx_i$, is the predicted value for y_i using the regression line. Thus $e_i = y_i - \hat{y}_i$ shows the deviance between the observed and the predicted values for each response. This can therefore mimic ε_i. That is why, s^2 is a good candidate for estimating σ^2. Using (8), (12) and (13) we can make confidence intervals for α and β, and also test a statistical hypothesis, such as $H_0 : \beta = 0$.

The reason for having $n-2$, not $n-1$ or n, on the bottom of (13) is as follows. First notice that each $y_i - \hat{y}_i$ can be considered as a vector, note that y_i's are random variables and can therefore be considered as vectors. Now $\{y_i - \hat{y}_i \mid i = 1, 2, \cdots, n\}$ is a set of n vectors. These vectors are, however, linearly dependent. Indeed,

$$\sum_{i=1}^n (y_i - \hat{y}_i) = 0 \quad \text{and} \quad \sum_{i=1}^n (y_i - \hat{y}_i)x_i = 0.$$

Therefore we have only $n-2$ linearly independent vectors among the n vectors $\{y_i - \hat{y}_i \mid i = 1, 2, \cdots, n\}$. To prove above identities, note that

$$\begin{aligned} y_i - \hat{y}_i &= y_i - (a + bx_i) = y_i - (\overline{y} - b\overline{x} + bx_i) \\ &= (y_i - \overline{y}) - b(x_i - \overline{x}) \end{aligned}$$

Then $\sum_{i=1}^n (y_i - \hat{y}_i) = 0$ follows from $\sum_{i=1}^n (y_i - \overline{y}) = 0$ and $\sum_{i=1}^n (x_i - \overline{x}) = 0$. As for the second identity we note that

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{y}_i)x_i &= \sum_{i=1}^n (y_i - \overline{y})x_i - b\sum_{i=1}^n (x_i - \overline{x})x_i \\ &= S_{XY} - bS_{XX} = S_{XY} - \frac{S_{XY}}{S_{XX}} S_{XX} = 0 \end{aligned}$$

Remark 1: Note that if $\varepsilon \sim N(0, \sigma^2)$, then a and b, the least square estimates, are also the maximum likelihood estimates. However, the maximum likelihood estimate of σ^2 is not s^2. In fact,

$$\hat{\sigma}^2_{ML} = s^2(1 - \frac{2}{n}).$$

Remark 2: One can also use the following argument to arrive at a consistent estimator of σ^2.

$$Y_i - \overline{Y} = \beta(x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon}),$$

raising both sides to power 2 and summing up over $i = 1, \cdots, n$, we have

$$S_{YY}^2 = \beta^2 S_{XX}^2 + S_{\varepsilon}^2 + \beta S_{X\varepsilon}.$$

Assuming that we have a random sample of pairs (X, Y), ε is independent of X, we have using the Law of Large Numbers (LLN), after dividing by n of course,

$$Var(Y) = \beta^2 Var(X) + \sigma^2.$$

Thus σ^2 can be consistently estimated by $[S_{YY}^2 - b^2 S_{XX}^2]/n$.

In many applications we are interested in predicting the response value or estimation of mean response for a given value of X, say x^*. Clearly

$$\hat{y}(x^*) = a + bx^*$$

is an estimate of the mean Y at $X = x^*$. Note once again that $\mathbb{E}(Y \mid X = x^*) = \alpha + \beta x^*$. In view of (4), (8), (12) and using the fact that $Cov(\overline{y}, b) = 0$, we have

$$\begin{aligned} Var(a + bx^*) &= Var[(\overline{y} - b\overline{x}) + bx^*] = Var[\overline{y} + b(x^* - \overline{x})] \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2 (x^* - \overline{x})^2}{S_{XX}} = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}} \right]. \end{aligned} \tag{14}$$

It is also possible to show that

$$\hat{y}(x^*) \overset{app}{\sim} N\left(\alpha + \beta x^*, \; \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}} \right] \right). \tag{15}$$

When n is large this approximation can be used to make a confidence interval for the mean response, $\hat{y}(x^*)$,

$$\hat{y}(x^*) \overset{+}{-} \mathfrak{z}_{\lambda/2} s \sqrt{\left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}} \right]},$$

where $P(Z > \mathfrak{z}_{\lambda/2}) = \lambda/2$ and $Z \sim N(0, 1)$.

Obviously the statistic $\hat{y}(x^*)$, the point on the regression line at $X = x^*$, serves the dual purpose as the estimate of mean response and the predicted value. The variance given by (14) is used in constructing a confidence interval on the mean response. It is not, however, appropriate for establishing any form of inference on a future single observation. In many cases we are interested in some type of bound on a single response observation at x^*. Consider a single observation at x^*, denoted by y^*. Note that $Y(x^*) = \alpha + \beta x^* + \varepsilon$. Thus if y^* is a predicted value for $Y(x^*)$ we have two sources of variabilities, one is due to estimating $\alpha + \beta x^*$ by $a + bx^*$ and the other is due to ε. We therefore have

$$\begin{aligned} Var(y^*) &= Var[\hat{y}(x^*)] + Var(\varepsilon) \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}} \right) \quad . \end{aligned} \tag{16}$$

Using CLT and (16) we can show that

$$y^* \overset{app}{\sim} N \left(\alpha + \beta x^*, \; \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}} \right] \right) . \tag{17}$$

When n is large Equation (17) can be used to make a confidence interval for the predicted value corresponding to a given value x^*,

$$y^* \overset{+}{-} \mathfrak{z}_{\lambda/2} s \sqrt{\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}} \right]},$$

where $P(Z > \mathfrak{z}_{\lambda/2}) = \lambda/2$ and $Z \sim N(0,1)$.

Remark: When $\varepsilon \sim N(0, \sigma^2)$, we can replace σ^2 by s^2 and obtain a T-distribution with (n-2) degrees of freedom in the following four cases, and hence use the T-distribution with (n-2) d.f. to make confidence intervals for β, α, $\hat{y}(x^*)$, and y^* respectively.

$$\frac{b - \beta}{\frac{s}{\sqrt{S_{XX}}}} \sim t_{(n-2)}$$

$$\frac{a - \alpha}{s \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}}} \sim t_{(n-2)}$$

$$\frac{\hat{y}(x^*) - (\alpha + \beta x^*)}{s \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}}}} \sim t_{(n-2)}$$

$$\frac{y^* - (\alpha + \beta x^*)}{s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}}}} \sim t_{(n-2)}$$

0.2 Analysis of Variance and R^2

Under the normality assumption for the error term, ε, we can devise test statistics to check quality of the fitted model, as well as a test for $H_0 : \beta = 0$. We first notice that total variability can be partitioned as follows. Using the following simple equation

$$(y_i - \overline{y}) = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i),$$

and the fact that

$$\sum_{i=1}^{n} (\hat{y}_i - \overline{y})(y_i - \hat{y}_i) = 0 \tag{18}$$

we have

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \overline{y})^2 &= \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ SS_{Total} &= SS_{Reg} + SS_{Res}, \end{aligned}$$

which is

$$\begin{pmatrix}\text{Total variability} \\ \text{in response}\end{pmatrix} = \begin{pmatrix}\text{Variability explained} \\ \text{by the model}\end{pmatrix} + \begin{pmatrix}\text{Variability} \\ \text{unexplained}\end{pmatrix} \quad .$$

To prove (18) we first notice that

$$(\hat{y}_i - \overline{y}) = [(a + bx_i) - \overline{y}] = [(a + bx_i) - (a + b\overline{x})] = b(x_i - \overline{x}),$$

and hence

$$\sum_{i=1}^{n} (\hat{y}_i - \overline{y})(y_i - \hat{y}_i) = b \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \hat{y}_i) = b \sum_{i=1}^{n} x_i (y_i - \hat{y}_i) = 0.$$

Under normality assumption, we can show that SS_{Total}, SS_{Reg} and SS_{Res} are all distributed according to χ^2-distribution respectively with (n-1), 1, (n-2) degrees of freedom. As I explained in class, the justification for having only 1 d.f. for SS_{Reg} is that

$$(\hat{y}_i - \overline{y}) = [(a + bx_i) - \overline{y}] \overset{\text{using (4)}}{=} b(x_i - \overline{x}), \tag{19}$$

and thus the set $\{\hat{y}_i - \overline{y} \mid i = 1, 2, \cdots, n\}$ is completely specified when we know b. In other words, all the vectors are on one line, hence the reason for having only 1 d.f. We therefore have, under normality assumption,

$$F = \frac{SS_{Reg}/1}{SS_{Res}/(n-2)} \sim F_{1,(n-2)}. \tag{20}$$

A. Testing $H_0 : \beta = 0$

We now devise a test for $H_0 : \beta = 0$ using the above decomposition of variability. To justify our test we need to calculate $\mathbb{E}(SS_{Reg})$. We have

$$\begin{aligned} \mathbb{E}(SS_{Reg}) &= \sum_{i=1}^{n} \mathbb{E}(\hat{y}_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{x}_i - \overline{x})^2 \mathbb{E}(b^2) \\ &= \sum_{i=1}^{n} (\hat{x}_i - \overline{x})^2 \{Var(b) + [\mathbb{E}(b)]^2\} \\ &= S_{XX}\{\frac{\sigma^2}{S_{XX}} + \beta^2\} = \sigma^2 + \beta^2 S_{XX}. \end{aligned} \tag{21}$$

Noting that

$$(y_i - \overline{y}) = \beta(x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon}),$$

and $\mathbb{E}(\varepsilon_i - \overline{\varepsilon}) = 0$, we have

$$\mathbb{E}\left[\sum_{i=1}^{n} (x_i - \overline{x})(\varepsilon - \overline{\varepsilon}) \mid x_1, \cdots, x_n\right] = 0.$$

We therefore obtain

$$\begin{aligned} \mathbb{E}\left(SS_{Total} \mid x_1, \cdots, x_n\right) &= \mathbb{E}\left(\sum_{i=1}^{n} (Y_i - \overline{Y})^2 \mid x_1, \cdots, x_n\right) \\ &= \beta^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 + \mathbb{E}\left[\sum_{i=1}^{n} (\varepsilon_i - \overline{\varepsilon})^2 \mid x_1, \cdots, x_n\right], \end{aligned}$$

and hence

$$\mathbb{E}\left(SS_{Total} \mid x_1, \cdots, x_n\right) = \beta^2 S_{XX} + (n-1)\sigma^2. \tag{22}$$

Note that

$$
\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n}(\varepsilon_i-\overline{\varepsilon})^2 \mid x_1,\cdots,x_n\right] &= \sum_{i=1}^{n}\mathbb{E}\left[(\varepsilon_i-\overline{\varepsilon})^2\right]\\
&= \sum_{i=1}^{n}\left[\mathbb{E}(\epsilon_i^2)+\mathbb{E}(\overline{\varepsilon}^2)-2\mathbb{E}(\varepsilon_i\overline{\varepsilon})\right]\\
\text{and since } \mathbb{E}(\varepsilon)=0 \text{ we have} \quad &= \sum_{i=1}^{n}\left[Var(\varepsilon_i)+Var(\overline{\varepsilon})-\frac{2}{n}\mathbb{E}\left(\varepsilon_i^2+\sum_{j\neq i}\varepsilon_i\varepsilon_j\right)\right]\\
\text{note that } \mathbb{E}(\varepsilon_i\varepsilon_j)=0 \text{ for } i\neq j \quad &= \sum_{i=1}^{n}\left[\sigma^2+\frac{\sigma^2}{n}-\frac{2}{n}\sigma^2\right]\\
&= n\left[\sigma^2-\frac{\sigma^2}{n}\right]=(n-1)\sigma^2.
\end{aligned}
$$

Using equations (21) and (22), we have

$$
\begin{aligned}
\mathbb{E}(SS_{Res}) &= \mathbb{E}(SS_{Total})-\mathbb{E}(SS_{Reg})\\
&= \beta^2S_{XX}+(n-1)\sigma^2-(\beta^2S_{XX}+\sigma^2)\\
&= (n-2)\sigma^2.
\end{aligned}
$$

This then implies that

$$
\mathbb{E}(s^2)=\mathbb{E}(\frac{SS_{Res}}{n-2})=\sigma^2 \qquad (23)
$$

A comparison between (21) and (23) indicates that detection of a slope significantly different from zero through analysis of variance is essentially equivalent to detecting a statistically significant value of β^2S_{XX} over the mere experimental error, σ^2. The F-ratio defined by equation (20) can therefor serve as a test statistic for testing $H_0:\beta=0$. In fact, we calculate the following p-value

$$
P(F>F_{obs}\mid H_0:\beta=0),
$$

where F_{obs} is the observed value of F and $F\sim F_{1,(n-2)}$.

Following is a typical analysis of variance table:

Source	Sume of Squares (SS)	df	Mean Square (MS)	F
Regression	SS_{Reg}	1	$SS_{Reg}/1$	$F=MS_{Reg}/s^2$
Residual	SS_{Res}	n-2	s^2	
Total	SS_{Total}	n-1		

Example: (*Wood Density Data*)

The F-ratio is 110.243. The p-value for this exmaple is

$$P(F_{1,(30-2)} > 110.243) \approx 0.0001$$

which strongly suggests that $H_0 : \beta = 0$ cannot be true.

B. Checking quality of the fitted model

When fitting a regression line two important questions arise naturally, one is on *quality of the fit*, and the second one is on *prediction*;

(1) Does the data fit the model well?

(2) Can the model predict the response values well enough?

We confine ourselves to answering only the first question. A criterion usually used to measure quality of the fitted model is the so-called *coefficient of determination*, denoted by R^2 and defined as follows:

$$R^2 = \frac{SS_{Reg}}{SS_{Total}} = \frac{SS_{Reg}}{SS_{Reg} + SS_{Res}} \tag{24}$$

The coefficient of determination, R^2, is very easy to interpret. It simply represents *the proportion of the variability in the response that is explained by the model.* Clearly the larger the value of R^2, the better the fit.

Using the law of large numbers (LLN)

$$R^2 = \frac{b^2 \sum_{i=1}^{n}(x_i - \overline{x}_n)^2}{b^2 \sum_{i=1}^{n}(x_i - \overline{x}_n)^2 + SS_{Res}} \xrightarrow{P} \frac{\beta^2 \kappa^2}{\beta^2 \kappa^2 + Var(\varepsilon)} \qquad \text{as} \quad n \to \infty$$

Having noted that κ^2 is essentially $\sigma_X^2 = Var(X)$, we have

$$R^2 \xrightarrow{P} \frac{\beta^2}{\beta^2 + \sigma_\varepsilon^2/\sigma_X^2} \qquad \text{as} \quad n \to \infty.$$

Some remarks are now in order:

- Having noticed that

$$b = r_{X,Y}\sqrt{\frac{S_{YY}}{S_{XX}}},$$

where $r_{X,Y}$ is the sample correlation between X and Y, we have

$$R^2 = \frac{b^2}{b^2 + (SS_{Res}/S_{XX})} = \frac{r_{X,Y}^2}{r_{X,Y}^2 + (SS_{Res}/S_{YY})} \tag{25}$$

Now we notice that

$$y_i - \hat{y}_i = (y_i - \overline{y}) - b(x_i - \overline{x})$$

and hence

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = S_{YY} + b^2 S_{XX} - 2bS_{XY} = S_{YY} + b^2 S_{XX} - 2b^2 S_{XX} = S_{YY} - b^2 S_{XX}.$$

We therefore have

$$\frac{SS_{Res}}{S_{YY}} = 1 - \frac{b^2 S_{XX}}{S_{YY}} = 1 - \frac{S_{XY}}{S_{XX} S_{YY}} = 1 - r^2_{X,Y}.$$

which using (25) implies $R^2 = r^2_{X,Y}$.

• It should be noted that b-value is heavily driven by variation in the response and the covariate. For example, large variation in the response values, relative to the variation in the covariate values, even when the actual correlation between the response and the covariate is small. Or it can mask a considerable relationship between the response and the covariate when the variation in the response is considerable small compare to the variation in the covariate values. Let

$$\tilde{y}_i = \frac{y_i - \overline{y}_n}{\sqrt{S_{YY}/n}} \qquad \text{and} \qquad \tilde{x}_i = \frac{x_i - \overline{x}_n}{\sqrt{S_{XX}/n}}.$$

The least squares estimates are therefore $\tilde{a} = 0$ and $\tilde{b} = r_{X,Y}$. Thus

$$\tilde{R}^2 = \tilde{b}^2 = r^2_{X,Y} = R^2.$$

This means that standardization does not have any effect on the value of R^2 since it is location-scale invariant.

• Using the postulated model $\mathbb{E}(Y|X = x) = \alpha + \beta x$, we have

$$\begin{aligned}\mathbb{C}\text{ov}(X,Y) &= \mathbb{E}(YX) - \mathbb{E}(Y)\mathbb{E}(X) = \mathbb{E}\left[X\mathbb{E}\left(Y|X\right)\right] - \mathbb{E}\left[\mathbb{E}\left(Y|X\right)\right]\mathbb{E}(X) \\ &= \mathbb{E}\left[X\left(\alpha + \beta X\right)\right] - \mathbb{E}\left(\alpha + \beta X\right)\mathbb{E}(X) \\ &= \alpha\mathbb{E}\left(X\right) + \beta\mathbb{E}\left(X^2\right) - \alpha\mathbb{E}\left(X\right) - \beta\left[\mathbb{E}\left(X\right)\right]^2 \\ &= \beta\mathbb{V}\text{ar}\left(X\right),\end{aligned}$$

and hence

$$\rho(X,Y) = \beta\frac{\sigma_X}{\sigma_Y}. \tag{26}$$

We also notice that

$$\begin{aligned}\sigma_Y^2 = \mathbb{V}\mathrm{ar}(Y) &= \mathbb{E}\left[\mathbb{V}\mathrm{ar}\left(Y|X\right)\right] + \mathbb{V}\mathrm{ar}\left[\mathbb{E}\left(Y|X\right)\right] = \sigma_\varepsilon^2 + \mathbb{V}\mathrm{ar}\left[\alpha + \beta X\right] \\ &= \sigma_\varepsilon^2 + \beta^2 \mathbb{V}\mathrm{ar}(X) = \sigma_\varepsilon^2 + \beta^2 \sigma_X^2.\end{aligned}$$

Using (26) we have $\sigma_Y^2 = \sigma_\varepsilon^2 + \rho^2 \sigma_Y^2$ and hence

$$\frac{\sigma_\varepsilon^2}{\sigma_Y^2} = 1 - \rho^2.$$

- $R^2 = 0$ if and only if $\beta = 0$.
- $R^2 = 1$ if and only $\sigma_\varepsilon^2 = 0$, i.e. Y and X are perfectly linearly correlated.

It is not hard to see that

$$R^2 = \frac{1}{1 + (n-2)F^{-1}}$$

where F is the F-ratio given by (20). As mentioned above, under $H_0 : \beta = 0$ and the normality assumption, $F \sim F_{1,(n-2)}$. Thus under these assumptions, one can show, using the transformation method discussed in Math323, that

$$R^2 \sim \mathcal{B}(\frac{1}{2}, \frac{n-2}{2}) \tag{27}$$

where $\mathcal{B}(\alpha, \beta)$ is the $\mathcal{B}$ distribution with parameters α and β (see Math323 notes). Thus R^2 can be used to test H_0. The following probability,

$$p_{fit} = P(R^2 > R_{obs}^2),$$

where R_{obs}^2 is the observed value of R^2, can serve as a p-value using which one can accept or reject the assumption of a linear relationship between X and Y. The small values of p_{fit} can indicate that the data do not support $H_0 : \beta = 0$ while the large values of p_{fit} indicate that the data do not adequately fit a linear model. Note that small values of p_{fit} correspond to the case that R_{obs}^2, and hence $r_{X,Y}^2$, is large meaning that there is a strong correlation, i.e. strong linear relationship, between X and Y.

Using (27) we can find the rate at which $R^2 \to 0$ when H_0 is true. Note that

$$\mathbb{E}(R^2) = \frac{1/2}{1/2 + (n-2)/2} = \frac{1}{n-1} \to 0 \quad \text{as} \quad n \to \infty,$$

and hence using Markov's inequality we find

$$R^2 = O_p(\frac{1}{n}),$$

which, of course, implies that

$$R^2 \xrightarrow{P} 0.$$

- Hawkins (1989, *The American Statistician*, Vol. 43, No. 4, pp 235-237) shows that $f(r) = \text{arctanh}(r) = 1/2 \ln(\frac{1+r}{1-r})$ is asymptotically normally distributed, regardless of the joint distribution of X and Y. Using this result one can find confidence intervals for R^2 and hence decides on the goodness of fit when n, the sample size, is large enough. This essentially means that

$$\sqrt{n}(f(r) - f(\rho)) \overset{\text{app}}{\sim} N(0, \tau_F^2),$$

where

$$\tau_F^2 = (1-\rho^2)^{-2} 1/4 \{(m_{40} + 2m_{22} + m_{04})\rho^2 - 4(m_{31} + m_{13})\rho + 4m_{22}\},$$

$m_{rs} = m_{rs}(F) = \mathbb{E}\,(X^r Y^s)$, and F is the joint distribution of X and Y. Thus

$$f(r) \overset{+}{-} \mathfrak{z}_{\lambda/2} \frac{\tau_F}{\sqrt{n}}$$

is a 100(1-λ)% confidence interval for $f(\rho)$. Given that arctanh is a monotone increasing function, we can easily convert the above confidence interval to a confidence interval for ρ.