Math 362: Mathematical Statistics II

Le Chen
le.chen@emory.edu

Emory University
Atlanta, GA

Last updated on April 13, 2021

2021 Spring

Chapter 11. Regression

Recall For any two random variables X and Y, the regression curve of Y on X, namely,

$$f(x) = \mathbb{E}[Y|X = x].$$

minimizes the squared error

$$\mathbb{E}\left[(Y - f(X))^2\right]$$

Difficulties The regression curve $y = \mathbb{E}[Y|x]$ is complicated and hard to obtain.

Compromise Assume that $f(x) = a + bx$ (i.e., the first order approximation)

Def. **(Simple) linear model**:

1. $f_{Y|x}(y)$ is a normal pdf for any x given.

2. The standard deviation, σ, of $Y|x$ is the same for all x, i.e.,

$$\sigma^2 \equiv \mathbb{E}[Y^2|x] - \mathbb{E}[Y|x]^2.$$

3. The mean of $Y|x$ is collinear, i.e.,

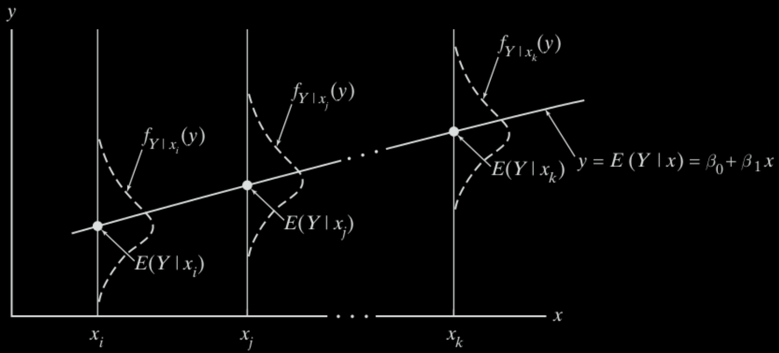
$$y = \mathbb{E}[Y|x] = \beta_0 + \beta_1 x.$$

4. All of the conditional distributions represnt indep. random variables.

Summary Let $Y_1, \cdots, Y_n$ be independent r.v.'s where $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ with x_i are known and β_0, β_1 and σ^2 are unknown.

$$\Updownarrow$$

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


y
$f_{Y|x_k}(y)$
$f_{Y|x_j}(y)$
$f_{Y|x_i}(y)$
$E(Y|x_k)$
$y = E(Y|x) = \beta_0 + \beta_1 x$
$E(Y|x_j)$
$E(Y|x_i)$
x
x_i
x_j
x_k

MLE for linear model

Thm. Let $(x_1, Y_1), \cdots, (x_n, Y_n)$ be a set of points satisfying the linear model, $\mathbb{E}[Y|x] = \beta_0 + \beta_1 x$.

($\iff$ let $Y_1, \cdots, Y_n$ be independent r.v.'s where $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ with x_i are known and β_0, β_1 and σ^2 are unknown.)

The maximum likelihood estimators for β_0, β_1 and σ^2 are given by

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^n x_i Y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n Y_i\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2}$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n x_i}{n} = \overline{Y} - \hat{\beta}_1 \overline{x}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \widehat{Y}_i\right)^2, \qquad \widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

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

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f_{Y_i|x_i}(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right).$$

Then take partial derivatives and set them to zero:

$$\frac{\partial \ln L}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial \ln L}{\partial \beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{(\sigma^2)^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = 0$$

Once β_0 and β_1 are solved from the first relations, then the third relation shows that

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

The first two relations give

$$\left(\sum_{i=1}^{n} y_i\right) - \beta_0 n - \beta_1\left(\sum_{i=1}^{n} x_i\right) = 0$$
$$\left(\sum_{i=1}^{n} x_i y_i\right) - \beta_0\left(\sum_{i=1}^{n} x_i\right) - \beta_1\left(\sum_{i=1}^{n} x_i^2\right) = 0$$

or

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}\begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix} = \begin{pmatrix}\sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i\end{pmatrix}$$

Hence,

$$\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

$$= \frac{1}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

$$= \frac{1}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i y_i\right) \\ -\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right) + n\left(\sum_{i=1}^n x_i y_i\right) \end{pmatrix}$$

$$\Downarrow$$

$$\beta_0 = \frac{\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i y_i\right)}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}$$

$$\beta_1 = \frac{-\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right) + n\left(\sum_{i=1}^n x_i y_i\right)}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}$$

Recall

$$\beta_1 = \frac{n\left(\sum_{i=1}^n x_i y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}$$

Let's simply β_0:

$$\beta_0 = \frac{\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i y_i\right)}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}$$

$$= \frac{\left[\left(\sum_{i=1}^n x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2\right]\left(\sum_{i=1}^n y_i\right)}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}$$

$$+ \frac{\frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2\left(\sum_{i=1}^n y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i y_i\right)}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}$$

$$= \frac{1}{n}\sum_{i=1}^n y_i + \frac{1}{n}\beta_1 \sum_{i=1}^n x_i$$

Finally, replacing β_0, β_1, σ^2 and y_i by $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$ and Y_i, respectively, proves the theorem. □

Properties of linear model estimators

Theorem:

1. $\hat{\beta}_0$ and $\hat{\beta}_1$ are both normally distributed.
2. $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased: $\mathbb{E}[\hat{\beta}_0] = \beta_0$ and $\mathbb{E}[\hat{\beta}_1] = \beta_1$.
3. Variances are eqal to

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x - \bar{x})^2}$$

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

4. $\hat{\beta}_1$, $\overline{Y}$ and $\hat{\sigma}^2$ are mutually independent.
5. $\frac{n\hat{\sigma}^2}{\sigma^2} \sim$ Chi Square with $n-2$ degrees of freedom. $\implies \mathbb{E}[\hat{\sigma}^2] = \frac{n-2}{n}\sigma^2$

Remark 1 Because

$$\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \overline{Y} - \overline{x}\hat{\beta}_1 + \hat{\beta}_1 x_i = \overline{Y} + (x_i - \overline{x})\hat{\beta}_1,$$

(4) implies that, for all $i = 1, \cdots, n$,

$$\widehat{Y}_i \perp \hat{\sigma}^2$$

Remark 2 By (5)

$$\mathbb{E}\left[\frac{n\hat{\sigma}^2}{\sigma^2}\right] = n-2 \iff \mathbb{E}[\hat{\sigma}^2] = \frac{n-2}{n}\sigma^2$$

$$\iff \mathbb{E}\left[\frac{n}{n-2}\hat{\sigma}^2\right] = \sigma^2$$

Or equivalently,

$\hat{\sigma}^2$ is a biased, but asymptotically unbiased, estimator for σ^2

$\dfrac{n}{n-2}\hat{\sigma}^2$ is an unbiased estimator for σ^2.

Proof. (1) Notice that both

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i Y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n}$$

are linear combinations for normal random variables, we see that both β_0 and β_1 are normal.

(2) Because $\mathbb{E}[Y|x] = \beta_0 + \beta_1 x$, we see that

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_1] &= \frac{n\sum_{i=1}^n x_i \mathbb{E}[Y_i] - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n \mathbb{E}[Y_i]\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2} \\
&= \frac{n\sum_{i=1}^n x_i(\beta_0 + \beta_1 x_i) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n (\beta_0 + \beta_1 x_i)\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2} \\
&= \frac{n\beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)\left(n\beta_0 + \beta_1 \sum_{i=1}^n x_i\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2} \\
&= \beta_1,
\end{aligned}$$

and then

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_0] &= \frac{\sum_{i=1}^n \mathbb{E}[Y_i] - \mathbb{E}[\hat{\beta}_1]\sum_{i=1}^n x_i}{n} \\
&= \frac{\sum_{i=1}^n (\beta_0 + \beta_1 x_i) - \beta_1 \sum_{i=1}^n x_i}{n} \\
&= \beta_0.
\end{aligned}$$

Hence, both $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators for β_0 and β_1, respectively.

(3) Notice that

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^n x_i Y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n Y_i\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2}$$

$$= \frac{\sum_{i=1}^n x_i Y_i - \overline{x}\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}$$

$$= \frac{\sum_{i=1}^n (x_i - \overline{x}) Y_i}{\sum_{i=1}^n x_i^2 - n\overline{x}^2} = \sum_{i=1}^n \frac{(x_i - \overline{x})}{\sum_{i=1}^n x_i^2 - n\overline{x}^2} Y_i$$

By independence of Y_i, we see that

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \sum_{i=1}^n \frac{(x_i - \overline{x})^2}{\left(\sum_{i=1}^n x_i^2 - n\overline{x}^2\right)^2} \mathrm{Var}\left(Y_i\right) = \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{\left(\sum_{i=1}^n x_i^2 - n\overline{x}^2\right)^2}\sigma^2$$

Because $\sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n x_i^2 - n\overline{x}^2$, we see that

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

- As for $\hat{\beta}_0$, notice that

$$\hat{\beta}_0 = \frac{\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n Y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i Y_i\right)}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}$$

$$= \frac{\left(\frac{1}{n}\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n Y_i\right) - \overline{x}\left(\sum_{i=1}^n x_i Y_i\right)}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}$$

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

Hence,

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \sum_{j=1}^n \left[\frac{\left(\frac{1}{n}\sum_{i=1}^n x_i^2\right) - \overline{x}x_j}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}\right]^2 \sigma^2$$

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \sum_{j=1}^{n} \left[\frac{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}x_j}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \right]^2 \sigma^2$$

$$= \sigma^2 \frac{\sum_{j=1}^{n} \left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}x_j\right]^2}{\left[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right]^2}$$

$$= \sigma^2 \frac{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right)^2 - \overline{x}^2 \sum_{j=1}^{n} x_j^2}{\left[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right]^2}$$

$$= \sigma^2 \frac{\frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right]^2 + 2\overline{x}^2\left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^4 - \overline{x}^2 \sum_{j=1}^{n} x_j^2}{\left[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right]^2}$$

$$= \sigma^2 \frac{\frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right]^2 + \overline{x}^2\left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^4}{\left[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right]^2}$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\right] = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]$$

(4) Since both $\hat{\beta}_1$ and $\overline{Y}$ are Gaussian, to show that they are independent, we need only to show that

$$\mathbb{E}[\hat{\beta}_1 \overline{Y}] = \mathbb{E}[\hat{\beta}_1]\mathbb{E}[\overline{Y}]$$

One can compute separately left- and right-hand sides and compare them. The computations are long and tedious but there is no fundamental difficulties.

The independence with $\hat{\sigma}^2$ is deeper and out of the scope of the book.

(5) See Appendix 11.A.1. $\square$

Estimating σ^2

1. MLE:

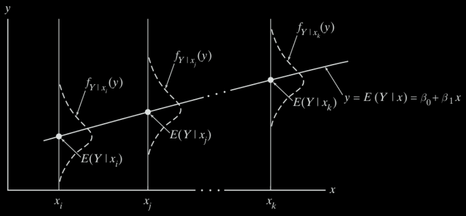
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2.$$

2. The unbiased estimator:

$$MSE = S^2 = \frac{n}{n-2}\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2.$$

Notation

Parameter	Estimator	Estimate
β_1	$\hat{\beta}_1$	β_{1e}
β_0	$\hat{\beta}_0$	β_{0e}
σ	S	s
σ^2	S^2	s^2
σ^2	$\hat{\sigma}^2$	σ_e^2
	$\overline{Y}$	$\bar{y}$
	$\widehat{Y}_i$	$\hat{y}_i = \beta_{0e} + \beta_{1e} x_i$



Drawing inferences on

1. the slope β_1
2. the intercept β_0
3. shape parameter σ^2
4. the regresion line itself
   $y = \mathbb{E}[Y|x] = \beta_0 + \beta_1 x$
5. the future observations
6. testing two slopes.

1. Drawing inferences on β_1

Thm. $T_{n-2} = \dfrac{\hat{\beta}_1 - \beta_1}{s \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim$ Student t distribution with df $= n - 2$.

1. Hypothesis test $H_0 : \beta_1 = \beta_1'$ vs. ....
2. C.I. for β_1: $\quad \beta_{1e} \pm t_{\alpha/2,n-2} \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})}}$

2. Drawing inferences on β_0

The GLRT procedure for assessing the credibility of $H_0 : \beta_0 = \beta_{0_o}$ is based on a Student t random variable with $n-2$ degrees of freedom:

$$T_{n-2} = \frac{(\hat{\beta}_0 - \beta_{0_o})\sqrt{n}\sqrt{\sum_{i=i}^{n}(x_i - \bar{x})^2}}{S\sqrt{\sum_{i=1}^{n} x_i^2}} = \frac{\hat{\beta}_0 - \beta_{0_o}}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_0)}} \qquad (11.3.6)$$

"Inverting" Equation 11.3.6 (recall the proof of Theorem 11.3.6) yields

$$\left[\hat{\beta}_0 - t_{\alpha/2,n-2} \cdot \frac{s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{n}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}},\ \hat{\beta}_0 + t_{\alpha/2,n-2} \cdot \frac{s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{n}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]$$

as the formula for a $100(1-\alpha)\%$ confidence interval for β_0.

3. Drawing inferences on σ^2

Since $(n-2)S^2/\sigma^2$ has a χ^2 pdf with $n-2$ df (if the n observations satisfy the stipulations implicit in the simple linear model), it follows that

$$P\left[\chi^2_{\alpha/2,n-2} \leq \frac{(n-2)S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2,n-2}\right] = 1-\alpha$$

Equivalently,

$$P\left[\frac{(n-2)S^2}{\chi^2_{1-\alpha/2,n-2}} \leq \sigma^2 \leq \frac{(n-2)S^2}{\chi^2_{\alpha/2,n-2}}\right] = 1-\alpha$$

in which case

$$\left[\frac{(n-2)s^2}{\chi^2_{1-\alpha/2,n-2}}, \frac{(n-2)s^2}{\chi^2_{\alpha/2,n-2}}\right]$$

becomes the *100(1 − α)% confidence interval for* σ^2 (recall Theorem 7.5.1). Testing $H_0: \sigma^2 = \sigma_o^2$ is done by calculating the ratio

$$\chi^2 = \frac{(n-2)s^2}{\sigma_o^2}$$

which has a χ^2 distribution with $n-2$ df when the null hypothesis is true. Except for the degrees of freedom ($n-2$ rather than $n-1$), the appropriate decision rules for one-sided and two-sided H_1's are similar to those given in Theorem 7.5.2.

4. Drawing inference on the regression line

Intuition tells us that a reasonable point estimator for $E(Y \mid x)$ is the height of the regression line at x—that is, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. By Theorem 11.3.2, the latter is unbiased:

$$E(\hat{Y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x) = E(\hat{\beta}_0) + x\,E(\hat{\beta}_1) = \beta_0 + \beta_1 x$$

Of course, to use $\hat{Y}$ in any inference procedure requires that we know its variance. But

$$\begin{aligned}
\mathrm{Var}(\hat{Y}) &= \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x) = \mathrm{Var}(\bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x) \\
&= \mathrm{Var}[\bar{Y} + \hat{\beta}_1 (x - \bar{x})] \\
&= \mathrm{Var}(\bar{Y}) + (x - \bar{x})^2 \mathrm{Var}(\hat{\beta}_1) \quad \text{(why?)} \\
&= \frac{1}{n}\sigma^2 + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sigma^2 \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]
\end{aligned}$$

An application of Definition 7.3.3, then, allows us to construct a Student t random variable based on $\hat{Y}$. Specifically,

$$T_{n-2} = \frac{\hat{Y} - (\beta_0 + \beta_1 x)}{\sigma \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \Bigg/ \sqrt{\frac{(n-2)S^2}{\frac{\sigma^2}{n-2}}} = \frac{\hat{Y} - (\beta_0 + \beta_1 x)}{S \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$

has a Student t distribution with $n-2$ degrees of freedom. Isolating $\beta_0 + \beta_1 x = E(Y \mid x)$ in the center of the inequalities $P(-t_{\alpha/2,n-2} \leq T_{n-2} \leq t_{\alpha/2,n-2}) = 1 - \alpha$ produces a $100(1-\alpha)\%$ confidence interval for $E(Y \mid x)$.

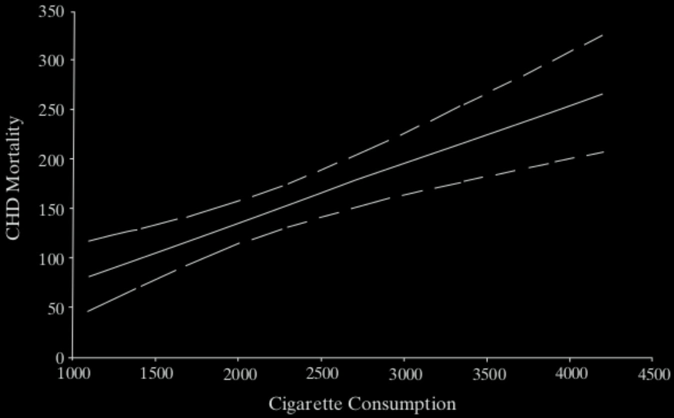

350
300
250
200
150
100
50
0
CHD Mortality
1000
1500
2000
2500
3000
3500
4000
4500
Cigarette Consumption

Figure 11.3.4

Let $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ be a set of n points that satisfy the assumptions of the simple linear model, and let (x, Y) be a hypothetical future observation, where Y is independent of the n Y_i's. A *prediction interval* is a range of numbers that contains Y with a specified probability.

Consider the difference $\hat{Y} - Y$. Clearly,

$$E(\hat{Y} - Y) = E(\hat{Y}) - E(Y) = (\beta_0 + \beta_1 x) - (\beta_0 + \beta_1 x) = 0$$

and

$$\begin{aligned}
\mathrm{Var}(\hat{Y} - Y) &= \mathrm{Var}(\hat{Y}) + \mathrm{Var}(Y) \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] + \sigma^2 \\
&= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]
\end{aligned}$$

Following exactly the same steps that were taken in the derivation of Theorem 11.3.7, a Student t random variable with $n-2$ degrees of freedom can be constructed from $\hat{Y} - Y$ (using Definition 7.3.3). Inverting the equation $P(-t_{\alpha/2,n-2} \leq T_{n-2} \leq t_{\alpha/2,n-2}) = 1 - \alpha$ will then yield the prediction interval $(\hat{y} - w, \hat{y} + w)$ given in Theorem 11.3.8.

Theorem 11.3.8 *Let $(x_1, Y_1), (x_2, Y_2), \ldots,$ and (x_n, Y_n) be a set of n points that satisfy the assumptions of the simple linear model. A* $100(1-\alpha)\%$ prediction interval for *Y at the fixed value x is given by $(\hat{y} - w, \hat{y} + w)$, where*

$$w = t_{\alpha/2,n-2} \cdot s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

and $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. □

E.g. 1 Does smoking contribute to coronary heat disease?

Table 11.3.1

Country	Cigarette Consumption per Adult per Year, x	CHD Mortality per 100,000 (ages 35–64), y
United States	3900	256.9
Canada	3350	211.6
Australia	3220	238.1
New Zealand	3220	211.8
United Kingdom	2790	194.1
Switzerland	2780	124.5
Ireland	2770	187.3
Iceland	2290	110.5
Finland	2160	233.1
West Germany	1890	150.3
Netherlands	1810	124.7
Greece	1800	41.2
Austria	1770	182.1
Belgium	1700	118.1
Mexico	1680	31.9
Italy	1510	114.3
Denmark	1500	144.9
France	1410	59.7
Sweden	1270	126.9
Spain	1200	43.9
Norway	1090	136.3

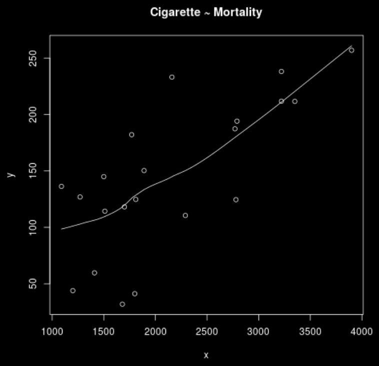
1) Test $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 > 0$ at $\alpha = 0.05$.

2) Find C.I. for β_1 with the same α.

Sol. http://r-statistics.co/Linear-Regression.html

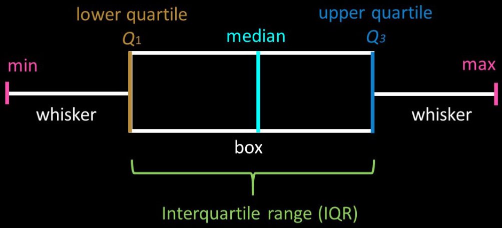
1. Let's first take of look of the data by scatter plot:

```
scatter.smooth(x=x, y=y, main="Cigarette ~
        Mortality")
```



Suggests a linearly increasing relationship between x and y.

2. Check outliers using boxplot.

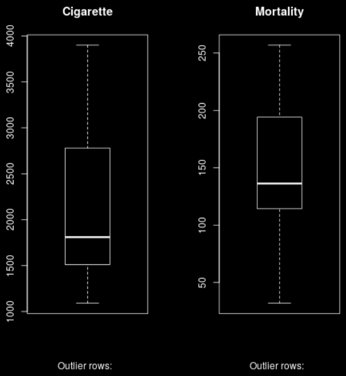


Any datapoint that lies outside the $r \times \text{IQR}$ is considered an outlier.

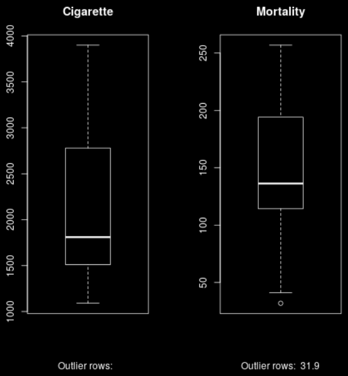
Generally, $r = 1.5$.

```r
r <- 1.5
par(mfrow=c(1, 2)) # divide graph area in 2 columns
boxplot(x, main="Cigarette", range=r, sub=paste("Outlier rows: ", boxplot.stats(x,
    coef=r)$out)) # box plot for 'Cigarette'
boxplot(y, main="Mortality", range=r, sub=paste("Outlier rows: ", boxplot.stats(y,
    coef=r)$out)) # box plot for 'Mortality'
```

$r = 1.5$

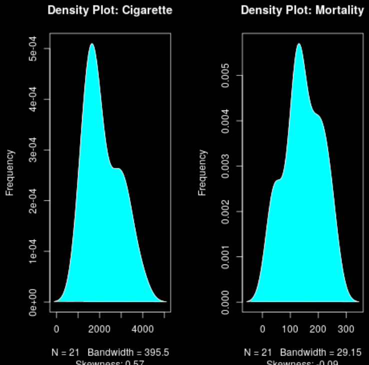


$r = 1$



3. Compute kernel density estimates

```r
library(e1071)
plot(density(x), main="Density Plot: Cigarette", ylab="Frequency",
     sub=paste("Skewness:", round(e1071::skewness(x), 2))) # density plot for 'Cigarette'
polygon(density(x), col="red")
plot(density(y), main="Density Plot: Mortality", ylab="Frequency",
     sub=paste("Skewness:", round(e1071::skewness(y), 2))) # density plot for 'Mortality'
polygon(density(y), col="red")
```



4. Compute correlation coeficient.

Correlation is a statistical measure with values in $[-1, 1]$ that suggests the level of linear dependence between two variables.

A value closer to 0 suggests a weak relationship between the variables. A low correlation $(-0.2, 0.2)$ probably suggests that much of variation of the response variable Y is unexplained by the predictor X, in which case, we should probably look for better explanatory variables.

```
> cor(x,y)
[1] 0.7295154
```

5. Compute linear regression.

```
> CigMort <- data.frame("Cigarette" = x, "Mortality" = y) # Build the data frame
> linearMod <- lm(Mortality ~ Cigarette, data=CigMort) # linear regression
> print(linearMod) # Print out the result

Call:
lm(formula = Mortality ~ Cigarette, data = CigMort)

Coefficients:
(Intercept)   Cigarette
    15.7711      0.0601
```

$$y = 15.7711 + 0.0601x$$

6. Check statistical significance of the linear model

```
> summary(linearMod)

Call:
lm(formula = Mortality ~ Cigarette, data = CigMort)

Residuals:
    Min      1Q  Median      3Q     Max
-84.835 -40.809  5.058  28.814  87.518

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.77115  29.57889   0.533 0.600085
Cigarette    0.06010   0.01293   4.649 0.000175 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 46.71 on 19 degrees of freedom
Multiple R-squared: 0.5322, Adjusted R-squared: 0.5076
F-statistic: 21.62 on 1 and 19 DF, p-value: 0.0001749
```

0.1 By default, p-values are computed for $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$, $i = 0, 1$.

0.2 The more stars by the variable's p-Value, the more significant the variable.

Testing $H_0 : \beta_1 = 0$ v.s.
$H_1 : \beta_1 \neq 0$

t-score is 4.4649.

p-value= 0.000175

Conclusion: reject at $\alpha = 0.05$.

95% C.I. for β_1:

Testing $H_0 : \beta_0 = 0$ v.s.
$H_1 : \beta_0 \neq 0$

t-score is 0.533.

p-value= 0.600

Conclusion: fail to reject at $\alpha = 0.05$.

95% C.I. for β_0:

```
> # 95% C.I. for slope parameter beta_1
> alpha <- 0.05
> for (i in c(1,0)) {
+  coef <- summary(linearMod)$coefficient
+  df <- linearMod$df.residual
+  lbd <- coef[i+1,1] - pt(1-alpha/2,df) * coef[i+1,2]
+  ubd <- coef[i+1,1] + pt(1-alpha/2,df) * coef[i+1,2]
+  print(paste("95% C.I. for the slope is beta_",i,
+                 " is (", round(lbd,3), ",", round(ubd,3),")"))
+ }
[1] "95% C.I. for the slope is beta_ 1 is ( 0.049 , 0.071 )"
[1] "95% C.I. for the slope is beta_ 0 is ( -8.753 , 40.295 )"
```

7. Compute R-Squared and the adjusted R-Squared.

$$R^2 = 1 - \frac{SSE}{SST} \quad \text{and} \quad R^2_{adj} = 1 - \frac{MSE}{MST}$$

```
> names(summary(linearMod))
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
> summary(linearMod)$r.squared
[1] 0.5321927
> summary(linearMod)$adj.r.squared
[1] 0.5075712
```

The large r^2 or r^2_{adj} the better, the more powerful or expressive is the L.M.

8. Residue standard error and F-statistic

$$\text{Residue standard error} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-q}}$$

$$F = \frac{MSR}{MSE} = \frac{SSR/(q-1)}{SSE/(n-q)} \sim \text{F-distribution } (df_1 = q-1, df_2 = n-q)$$

```
> names(summary(linearMod))
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
> summary(linearMod)$sigma
[1] 46.70826
> summary(linearMod)$fstatistic
   value    numdf   dendf
21.61501 1.00000 19.00000
> f <- summary(linearMod)$fstatistic
> pf(f[1], f[2], f[3], lower=FALSE)
       value
0.0001748805
```

9. Model selection:

Akaike's information criterion — AIC (Akaike, 1974)

$$AIC = -2\ln(\widehat{L}) + 2q$$

Bayesian information criterion — BIC (Schwarz, 1978)

$$BIC = -2\ln(\widehat{L}) + q\ln(n)$$

$\widehat{L}$: the maximum of likelihood.

q: the number of parameters in the model.

n: the sample size.

```
> AIC(linearMod)
[1] 224.9383
> BIC(linearMod)
[1] 228.0719
```

The lower the better!

10. Does L.M. fit our model?

Statistic	criterion	our case
R^2	Higher the better (>0.7)	0.53
R^2_{adj}	Higher the better	0.51
AIC	Lower the better	225
BIC	Lower the better	228
$\vdots$	$\vdots$	$\vdots$

11. Drawing inference on $\mathbb{E}(Y|x)$

Find 95% C.I. for Y at $x = 4200$.

Here, $n = 21$, $t_{.025,19} = 2.0930$, $\sum_{i=1}^{21}(x_i - \bar{x})^2 = 13{,}056{,}523.81$, $s = 46.707$, $\hat{\beta}_0 = 15.7661$, $\hat{\beta}_1 = 0.0601$, and $\bar{x} = 2148.095$. From Theorem 11.3.7, then,

$$\hat{y} = 15.7661 + 0.0601(4200) = 268.1861$$

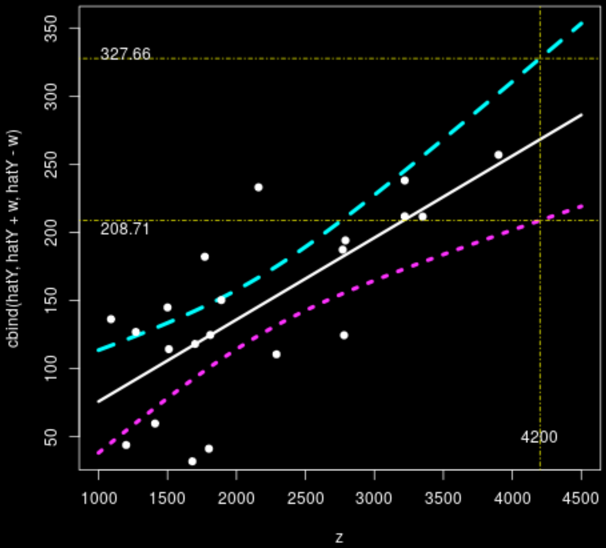
$$w = 2.0930(46.707)\sqrt{\frac{1}{21} + \frac{(4200 - 2148.095)^2}{13{,}056{,}523.81}} = 59.4714$$

and the 95% confidence interval for $E(Y|4200)$ is

$$(268.1861 - 59.4714,\ 268.1861 - 59.4714)$$

which rounded to two decimal places is

$$(208.71, 327.66)$$


327.66
208.71
4200
cbind(hatY, hatY + w, hatY - w)
50
100
150
200
250
300
350
1000
1500
2000
2500
3000
3500
4000
4500
z

```
s <- summary(linearMod)$sigma
beta <- linearMod$coefficients
z <- seq(1000,4500,1)
hatY <- beta[1]+beta[2]*z
w <- qt(0.975,19) * s * sqrt(1/21+(z-mean(x))^2/(sum((x-mean(x
    ))^2)))
matplot(z,cbind(hatY,hatY+w,hatY-w),type = c("l","l","l"),lwd=c
    (3,4,4))
points(x, y, pch = 19)
abline(v=4200,col = "blue", lty = 4)
abline(h=208.71,col = "blue", lty = 4)
abline(h=327.66,col = "blue", lty = 4)
text(4200,50,4200)
text(1200,203,208.71)
text(1200,331,327.66)
```

12. Drawing inference on future observations.

Find 95% prediction interval for Y at $x = 4200$.

When $x = 4200$, $\hat{y} = 268.1861$ for both intervals. From Theorem 11.3.8, the width of the 95% prediction interval for Y is:

$$w = 2.0930(46.707)\sqrt{1 + \frac{1}{21} + \frac{(4200 - 2148.095)^2}{13{,}056{,}523.81}} = 114.4725$$

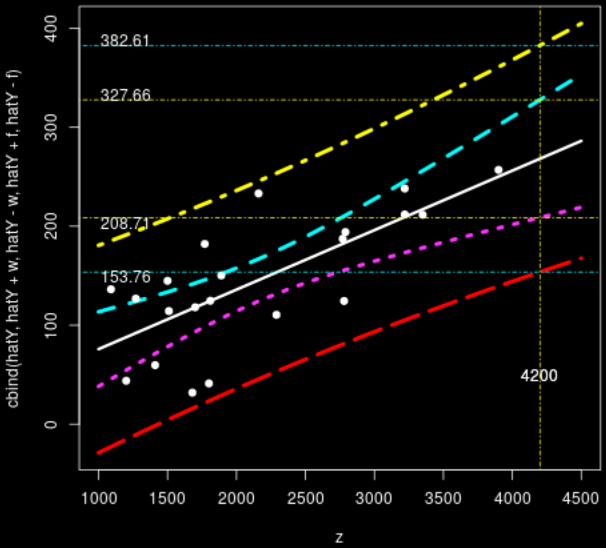
The 95% prediction interval, then, is

$$(268.1861 - 114.4725, 268.1861 + 114.4725)$$

which rounded to two decimal places is

$$(153.76, 382.61)$$

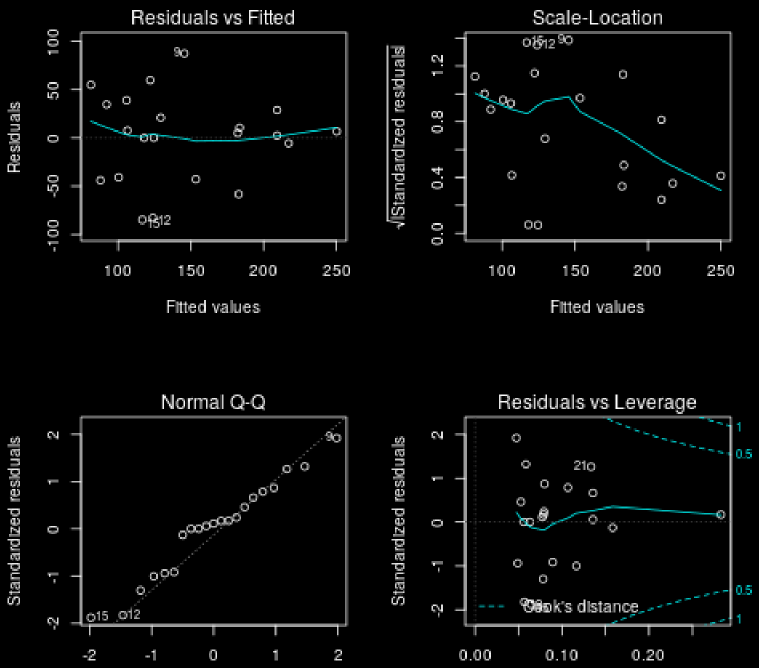
which makes it 92% wider than the 95% confidence interval for $E(Y|4200)$. ■


382.61
327.66
208.71
153.76
4200
400
300
200
100
0
1000
1500
2000
2500
3000
3500
4000
4500
z
cbind(hatY, hatY + w, hatY - w, hatY + f, hatY - f)

```r
s <- summary(linearMod)$sigma
beta <- linearMod$coefficients
z <- seq(1000,4500,1)
hatY <- beta[1]+beta[2]*z
w <- qt(0.975,19) * s * sqrt(1/21+(z-mean(x))^2/(sum((x-mean(x))^2)))
f <- qt(0.975,19) * s * sqrt(1+1/21+(z-mean(x))^2/(sum((x-mean(x))^2)))
matplot(z,cbind(hatY,hatY+w,hatY-w,hatY+f,hatY-f),
        type = c("l","l","l","l","l"),lwd=c(3,4,4,4,4))
points(x, y, pch = 19)
abline(v=4200,col = "blue", lty = 4)
abline(h=208.71,col = "blue", lty = 4)
abline(h=327.66,col = "blue", lty = 4)
text(4200,50,4200)
text(1200,208.71-5,208.71)
text(1200,327.66+5,327.66)
abline(h=153.76,col = "red", lty = 4)
abline(h=382.61,col = "red", lty = 4)
text(4200,50,4200)
text(1200,153.76-5,153.76)
text(1200,382.61+5,382.61)
```

13. More about diagnozing the linear model:

```
# diagnostic plots
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(linearMod)
```


Residuals vs Fitted
Residuals
Fitted values
Scale-Location
√|Standardized residuals|
Fitted values
Normal Q-Q
Standardized residuals
Residuals vs Leverage
Standardized residuals
Cook's distance

E.g. 2 Find 95% C.I. for the amount of increas year-by-year in the cost of Toyota Camry sedan.

Table 11.3.2

Year	Year after 2005	Suggested Retail Price ($)
2005	0	7,935
2006	1	8,495
2007	2	10,160
2008	3	10,817
2009	4	11,078
2010	5	11,967
2011	6	12,658
2012	7	13,844
2013	8	13,982
2014	9	14,629

Data from: kbb.com

Sol. We first find the regression:

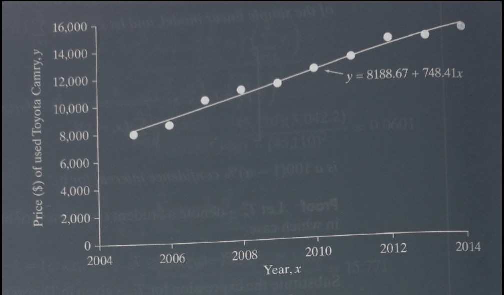


Figure 11.3.3

The slope of the line, $\hat{\beta}_1$, represents the amount of increase year-by-year in the cost of an older model. Often a range of values is better than a single estimate, so a good way to provide this is using a confidence interval for the true value β_1.

Here, $\sqrt{\sum_{i=0}^{9}(x_i - \bar{x})^2} = \sqrt{82.5} = 9.083$

and from Equation 11.3.5, $s^2 = \frac{1}{10-2}\left(\sum_{i=0}^{9} y_i^2 - \hat{\beta}_0 \sum_{i=0}^{9} y_i - \hat{\beta}_1 \sum_{i=0}^{9} x_i y_i\right)$

$$\frac{1}{8}[1{,}382{,}678{,}777 - (8188.67)(115{,}565) - (748.41)(581{,}786)] = 117{,}727.98$$

so $s = \sqrt{117{,}727.98} = 343.11$.

Using $t_{.025,8} = 2.3060$, the expression given in Theorem 11.3.6 reduces to $(748.41 - 2.3060\frac{343.11}{9.083}, 748.41 + 2.3060\frac{343.11}{9.083}) = (\$661.30, \$835.52)$

7. Testing the equality of two slopes

Table 11.3.3

Date	Day no., $x(=x^*)$	Strain A pop^n, y	Strain B pop^n, y^*
Feb 2	0	100	100
May 13	100	250	203
Aug 21	200	304	214
Nov 29	300	403	295
Mar 8	400	446	330
Jun 16	500	482	324

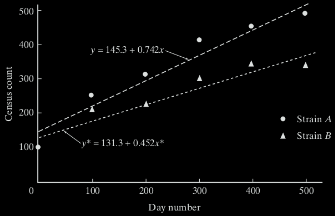


Figure 11.3.5

Do you believe that $\beta_1 = \beta_1^*$?

Or is $\beta_1 > \beta_1^*$ statisically significantly?

Theorem 11.3.9 *Let $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ and $(x_1^*, Y_1^*), (x_2^*, Y_2^*), \ldots, (x_m^*, Y_m^*)$ be two independent sets of points, each satisfying the assumptions of the simple linear model—that is, $E(Y \mid x) = \beta_0 + \beta_1 x$ and $E(Y^* \mid x^*) = \beta_0^* + \beta_1^* x^*$.*

a. *Let*

$$T = \frac{\hat{\beta}_1 - \hat{\beta}_1^* - (\beta_1 - \beta_1^*)}{S\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{1}{\sum_{i=1}^{m}(x_i^* - \bar{x}^*)^2}}}$$

where

$$S = \sqrt{\frac{\sum_{i=1}^{n}[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 + \sum_{i=1}^{m}[Y_i^* - (\hat{\beta}_0^* + \hat{\beta}_1^* x_i)]^2}{n+m-4}}$$

Then T has a Student t distribution with $n+m-4$ degrees of freedom.

b. *To test $H_0: \beta_1 = \beta_1^*$ versus $H_1: \beta_1 \neq \beta_1^*$ at the α level of significance, reject H_0 if t is either (1) $\leq -t_{\alpha/2, n+m-4}$ or (2) $\geq t_{\alpha/2, n+m-4}$, where*

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_1^*}{s\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{1}{\sum_{i=1}^{m}(x_i^* - \bar{x}^*)^2}}}$$

(One-sided tests are defined in the usual way by replacing $\pm t_{\alpha/2, n+m-4}$ with either $t_{\alpha, n+m-4}$ or $-t_{\alpha, n+m-4}$.)

$$S^2 = SSE \text{ and } q = 4.$$

Sol. Test

$$H_0 : \beta_1 = \beta_1^* \quad \text{v.s.} \quad H_1 : \beta_1 > \beta_1^*.$$

Long computations ... $t = 2.50$.

`http://math.emory.edu/~lchen41/teaching/2020_Spring/Ex_11-3-4.R`

Critical region: $t > t_{0.05,8} = 1.8595$.

Reject. □

```
> # Example 11.3.4
> # Read data first
> Input <- ("
+ x     yA    yB
+ 0     100   100
+ 100   250   203
+ 200   304   214
+ 300   403   295
+ 400   446   330
+ 500   482   324
+ ")
> Data = read.table(textConnection(Input),
+                   header=TRUE)
> Data
    x  yA  yB
1   0 100 100
2 100 250 203
3 200 304 214
4 300 403 295
5 400 446 330
6 500 482 324
```

```
> #fit the first model ...
> DataA <- data.frame(x = Data$x,yA = Data$yA)
> fitA <- lm(yA~x, DataA)
> summary(fitA)

Call:
lm(formula = yA ~ x, data = DataA)

Residuals:
      1       2       3       4      5       6
-45.333 30.467 10.267 35.067  3.867 -34.333

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 145.33333   26.86684   5.409  0.00566 **
x             0.74200    0.08874   8.362  0.00112 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.12 on 4 degrees of freedom
Multiple R-squared: 0.9459, Adjusted R-squared: 0.9324
F-statistic: 69.92 on 1 and 4 DF, p-value: 0.001119
```

```
> #fit the second model ...
> DataB <- data.frame(x = Data$x,yB = Data$yB)
> fitB <- lm(yB~x, DataB)
> summary(fitB)

Call:
lm(formula = yB ~ x, data = DataB)

Residuals:
      1       2       3       4       5       6
-31.333  26.467  -7.733  28.067  17.867 -33.333

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 131.33333  22.77255   5.767  0.00449 **
x             0.45200   0.07522   6.009  0.00386 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.46 on 4 degrees of freedom
Multiple R-squared: 0.9003, Adjusted R-squared: 0.8754
F-statistic: 36.11 on 1 and 4 DF, p-value: 0.00386
```

```
> # Now compute t-score and p-value
> sA <- summary(fitA)$coefficients
> sA
             Estimate Std. Error t value    Pr(>|t|)
(Intercept) 145.3333 26.86683800 5.409395 0.005656733
x             0.7420  0.08873825 8.361671 0.001118570
> sB <- summary(fitB)$coefficients
> sB
             Estimate Std. Error t value    Pr(>|t|)
(Intercept) 131.3333 22.77254682 5.767178 0.004486443
x             0.4520  0.07521525 6.009420 0.003860274
> db <- (sA[2,1]-sB[2,1]) # difference of beta_1's
> db
[1] 0.29
> sd <- sqrt(sB[2,2]^2+sA[2,2]^2) # standard deviation
> sd
[1] 0.1163263
> df <- (fitA$df.residual+fitB$df.residual) # degrees of freedom
> df
[1] 8
> td <- db/sd # t-score
> pv <- 2*pt(-abs(td), df) # two-sided p-value
> print(paste("t-score is ", round(td,3),
+             "and p-value is", round(pv,3)))
[1] "t-score is 2.493 and p-value is 0.037"
```

You should always visualize your data before any analysis

N = 157 ; X mean = 50.7333 ; X SD = 19.5661 ; Y mean = 46.495 ; Y SD = 27.2828 ; Pearson correlation = -0.1772