STAT 513 fa 2020 Lec 07

Simple linear regression

Karl B. Gregory

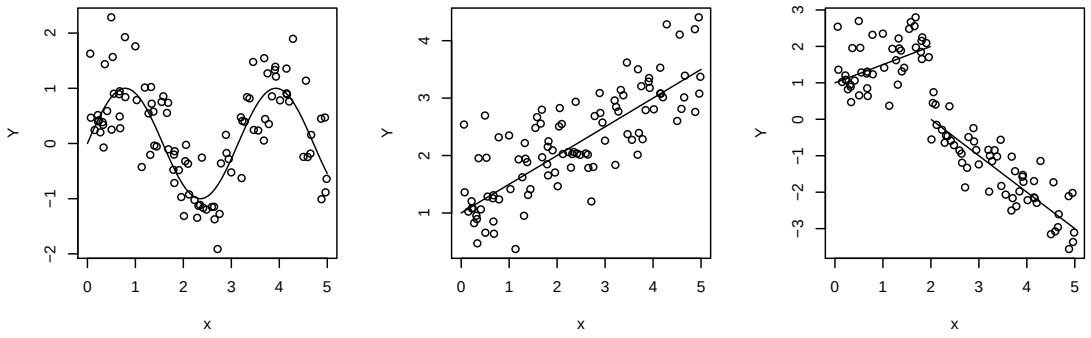
Regression

- **Regression model:** Let $x_1, \ldots, x_n$ be fixed real numbers and let $Y_1, \ldots, Y_n$ be independent random variables such that

$$Y_i = f(x_i) + \varepsilon_i, \qquad \text{for } i = 1, \ldots, n, \tag{1}$$

  for some $f : \mathbb{R} \to \mathbb{R}$, where $\varepsilon_1, \ldots, \varepsilon_n$ are independent identically distributed random variables such that $\mathbb{E}\varepsilon_i = 0$ and $\operatorname{Var}\varepsilon_i = \sigma^2$ for $i = 1, \ldots, n$.

- The idea is that we observe a function plus some random perturbations or "noise". For example, we might observe the points in the plots below, but we do not see the curves/lines around which they are scattered.



- The random variables $\varepsilon_1, \ldots, \varepsilon_n$ are often called *error terms*. These random quantities obscure the true function from us.

- The values $x_1, \ldots, x_n$ are the values of a *covariate*, a variable which varies randomly or of which the values are fixed by an experimenter. In our treatment of regression, we will regard the covariate as fixed (and we lose nothing by doing so).

- Since we assume $\mathbb{E}\varepsilon_i = 0$, the height of the function f at x_i represents the expected value of Y_i. That is, we have $\mathbb{E}Y_i = f(x_i)$.

- There is a vast body of literature about ways to estimate the function f. One generally begins by assuming that f belongs to some class or space of functions, where the space to which f belongs places restrictions on how "wiggly" f may be (to avoid getting too much into technicalities). We focus on the easiest case, in which we assume that the function f is a linear function—not wiggly at all.

- **Simple linear regression model:** Let $x_1, \ldots, x_n$ be fixed real numbers and let $Y_1, \ldots, Y_n$ be independent random variables such that

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \text{for } i = 1, \ldots, n, \tag{2}$$

  where $\varepsilon_1, \ldots, \varepsilon_n$ are independent identically distributed random variables such that $\mathbb{E}\varepsilon_i = 0$ and $\operatorname{Var} \varepsilon_i = \sigma^2$ for $i = 1, \ldots, n$.

- "Simple" refers to there being only a single covariate. We will cover models with multiple covariates later on.

- There are three unknown parameters in the simple linear regression model: the intercept β_0, the slope β_1, and the variance σ^2 of the error terms.

- We use the words *regression coefficients* in reference to β_0 and β_1.

- We will discuss the following in these notes:

  1. Estimation of β_0, β_1, and σ^2 as well as of $\beta_0 + \beta_1 x_{\text{new}}$, where in the latter case we are interested in estimating the expected value of the random variable Y_{new} of a "new" observation $(x_{\text{new}}, Y_{\text{new}})$.
  2. Inference about β_0 and β_1, focusing in particular on testing H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$. We will also discuss the construction of confidence intervals for $\beta_0 + \beta_1 x_{\text{new}}$.
  3. Prediction of the value of Y_{new} of a "new" observation $(x_{\text{new}}, Y_{\text{new}})$ by constructing an interval into which Y_{new} will fall with some desired probability.

  To do items 2 and 3, we will assume that $\varepsilon_1, \ldots, \varepsilon_n$ are independent $\text{Normal}(0, \sigma^2)$ random variables. To do item 1, we do not assume any particular distribution for the error terms.

- Sometimes the covariate is regarded as random, as in the following setup.

- **Alternate setup with random covariate:** Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be pairs of random variables such that

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

  where $\varepsilon_i, \ldots, \varepsilon_n$ are independent identically distributed random variables such that $\mathbb{E}[\varepsilon_i | X_i] = 0$ and $\operatorname{Var}[\varepsilon_i | X_i] = \sigma^2$ for all $i = 1, \ldots, n$. Then we have $\mathbb{E}[Y_i | X_i] = f(X_i)$. In this setup, we typically carry out all our analyses conditional on the values of $X_1, \ldots, X_n$, so that in practice it makes no difference whether we regard the covariate as random or fixed.

Least-squares estimation

- Under Model (1), one way to estimate the function f is to choose a space $\mathcal{G}$ of functions inside of which to search for the function that best fits the data in terms of some criterion. The least-squares criterion gives the estimator $\hat{f}$ of f defined by

$$\hat{f} = \text{argmin}_{g \in \mathcal{G}} \sum_{i=1}^{n} [Y_i - g(x_i)]^2,$$

  so that for any candidate function $g \in \mathcal{G}$, we consider the sum of the squared differences $[Y_i - g(x_i)]^2$, for $i = 1, \ldots, n$, and choose as our estimator $\hat{f}$ the function for which the sum of these squared differences is the smallest.

- Under the simple linear regression model, in which the function f takes the form $f(x) = \beta_0 + \beta_1 x$, we search over the space of all linear functions to find our estimator. That is, we define our estimator to be $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$, where

$$(\hat{\beta}_0, \hat{\beta}_1) = \text{argmin}_{(\beta_0, \beta_1) \in \mathbb{R}^2} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2.$$

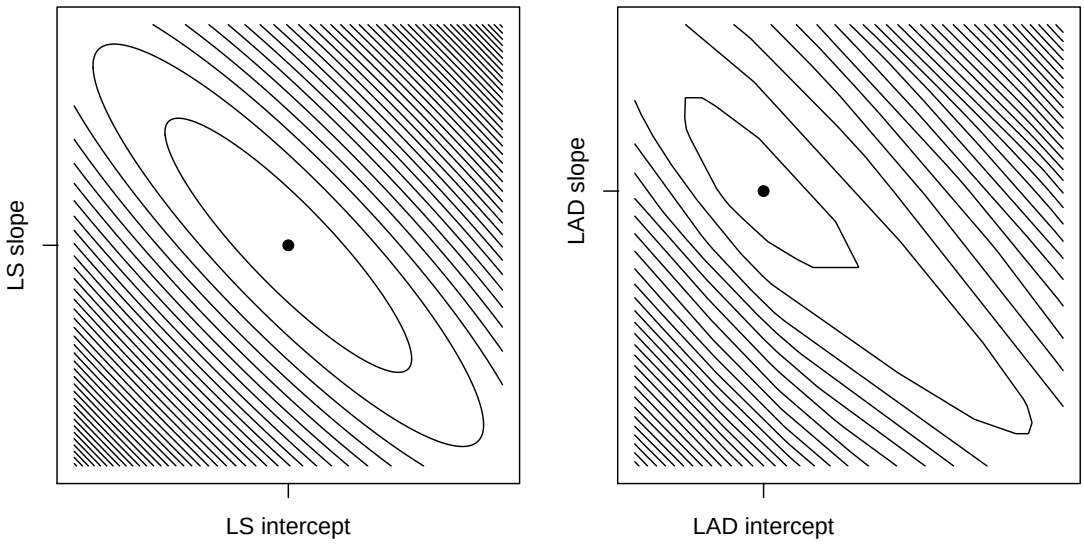
  The function $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ is the line through the data points for which the sum of the squared vertical distances between the points $Y_1, \ldots, Y_n$ and the line is minimized.

- We may think of it as the line of "best fit", but really, this is the line of best fit only according to the least-squares criterion; there are alternative criteria. For example, we could compute

$$(\check{\beta}_0, \check{\beta}_1) = \text{argmin}_{(\beta_0, \beta_1) \in \mathbb{R}^2} \sum_{i=1}^{n} |Y_i - (\beta_0 + \beta_1 x_i)|,$$

  where instead of minimizing a sum of squared differences, we minimize a sum of absolute values of differences. This criterion is sometimes called the least-absolute-deviations (LAD) criterion. The line given by $\check{f}(x) = \check{\beta}_0 + \check{\beta}_1 x$ is called the median regression line, and $\check{f}(x_i)$ is used as an estimator of the median of Y_i rather than of the expected value of Y_i.

- The plots below show, for a single simulated dataset, contours of the least-squares and least-absolute-deviations criteria with β_0 on the horizontal axis and β_1 on the vertical axis. We see that the two criteria result in different estimators. Note also that the contours of the least-absolute-deviations criterion are not smooth, which reflects the fact that the gradient of the function is not defined everywhere. This makes the least-absolute-deviations criterion more difficult to maximize, since we cannot simply set its derivative equal to zero and solve for β_0 and β_1. The contours of the least-squares criterion are, in contrast, smooth ellipses, and the gradient is defined everywhere, allowing us to use simple calculus methods to maximize it.



Least-squares estimators of simple linear regression coefficients

- We obtain $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing the function

$$Q_n(\beta_0, \beta_1) := \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2.$$

We will call the function $Q_n(\beta_0, \beta_1)$ the *objective function*. When our estimators are defined as the minimizers (or maximizers) of some function, that function is often referred to as the *objective function*).

- The pair of values $(\hat{\beta}_0, \hat{\beta}_1)$ minimizes $Q_n(\beta_0, \beta_1)$ if and only if

$$\frac{\partial}{\partial \beta_0} Q_n(\beta_0, \beta_1)\Big|_{(\beta_0,\beta_1)=(\hat{\beta}_0,\hat{\beta}_1)} = 0 \quad \text{and} \quad \frac{\partial}{\partial \beta_1} Q_n(\beta_0, \beta_1)\Big|_{(\beta_0,\beta_1)=(\hat{\beta}_0,\hat{\beta}_1)} = 0.$$

So to find expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$, we set the partial derivatives of $Q_n(\beta_0, \beta_1)$ with respect to β_0 and β_1 equal to zero and solve the resulting system of equations. We have

$$\frac{\partial}{\partial \beta_0} Q_n(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)] = 0$$
$$\frac{\partial}{\partial \beta_1} Q_n(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} x_i [Y_i - (\beta_0 + \beta_1 x_i)] = 0.$$

The first equation gives

$$\sum_{i=1}^{n} Y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0$$
$$\iff n\bar{Y} - n\beta_0 - n\beta_1 \bar{x}_n = 0$$
$$\iff \beta_0 = \bar{Y}_n - \beta_1 \bar{x}_n.$$

The second equation gives

$$\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i(\beta_0 + \beta_1 x_i) = 0,$$

which, when we plug in $\beta_0 = \bar{Y}_n - \beta_1 \bar{x}_n$, gives

$$\begin{aligned}
&\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i(\bar{Y}_n - \beta_1 \bar{x}_n + \beta_1 x_i) = 0 \\
\iff &\sum_{i=1}^{n} x_i Y_i - n\bar{x}_n \bar{Y}_n + \beta_1 n \bar{x}_n^2 - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \\
\iff &\sum_{i=1}^{n} x_i Y_i - n\bar{x}_n \bar{Y}_n = \beta_1 \sum_{i=1}^{n} x_i^2 - \beta_1 n \bar{x}_n^2 \\
\iff &\sum_{i=1}^{n} (x_i - \bar{x}_n)(Y_i - \bar{Y}_n) = \beta_1 \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \\
\iff &\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2},
\end{aligned}$$

where we have used the relations

$$\sum_{i=1}^{n} (x_i - \bar{x}_n)(Y_i - \bar{Y}_n) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}_n \bar{Y}_n \quad \text{and} \quad \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2.$$

From here we have

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{x}_n \tag{3}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}. \tag{4}$$

- **Remark:** If $x_1, \ldots, x_n$ all take the same value, we cannot compute the least-squares estimator of $\hat{\beta}_1$, since in this case we would have $\sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = 0$. But indeed, of what use is a covariate which does not vary? This case should not arise in simple linear regression, but in multiple linear regression (when there is more than one covariate) a situation analogous to this, but more subtle, may arise and cause problems.

- Define

$$S_{xY} = \sum_{i=1}^{n} (x_i - \bar{x}_n)(Y_i - \bar{Y}_n) \quad \text{and} \quad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \quad \text{and} \quad S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2,$$

as well as

$$r_{xY} = \frac{S_{xY}}{\sqrt{S_{xx} S_{YY}}},$$

so that r_{xY} is Pearson's correlation coefficient between $x_1, \ldots, x_n$ and $Y_1, \ldots, Y_n$. Moreover, let

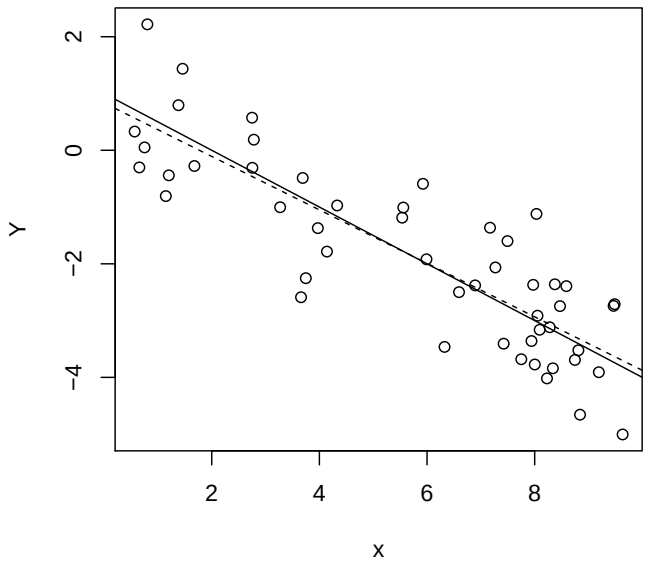
$$s_Y = \frac{S_{YY}}{n-1} \quad \text{and} \quad s_X = \frac{S_{xx}}{n-1}$$

denote the sample standard deviations of $Y_1, \ldots, Y_n$ and $x_1, \ldots, x_n$, respectively. Equipped with this notation, we may express $\hat{\beta}_1$ as

$$\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}} \quad \text{or} \quad \hat{\beta}_1 = r_{xY}\left(\frac{S_{YY}}{S_{xx}}\right)^{1/2} \quad \text{or} \quad \hat{\beta}_1 = r_{xY}(s_Y/s_x).$$

- **Example:** The following R code generates some values $x_1, \ldots, x_n$ from the Uniform$(0, 10)$ distribution, then generates $\varepsilon_1, \ldots, \varepsilon_n$ as a random sample from the Normal$(0, 1)$ distribution, and then sets $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $i = 1, \ldots, n$ under some choices of β_0 and β_1. After generating some data, the least-squares regression coefficients are computed and corresponding line is overlaid on a scatterplot of the points along with the true function.

```
n <- 50
x <- runif(n,0,10)
e <- rnorm(n,0,1)
beta0 <- 1
beta1 <- -.5
Y <- beta0 + beta1 * x + e

plot(Y~x)
abline(beta0,beta1) # true function

beta1.hat <- cor(x,Y)*sd(Y)/sd(x)
beta0.hat <- mean(Y) - beta1.hat*mean(x)
abline(beta0.hat,beta1.hat,lty=2) # estimated function
```



- Regression parlance: We often refer to the computation of the least-squares coefficients as "fitting" the least-squares line, and we call the line the "least-squares fit"; it is a line "fitted" to the data.

- After the least-squares line has been fit to the data, we refer to the differences between the values $Y_1, \ldots, Y_n$ and the heights of the least-squares line at the points $x_1, \ldots, x_n$ as the *residuals*, and we denote them by

$$\hat{\varepsilon}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \quad i = 1, \ldots, n.$$

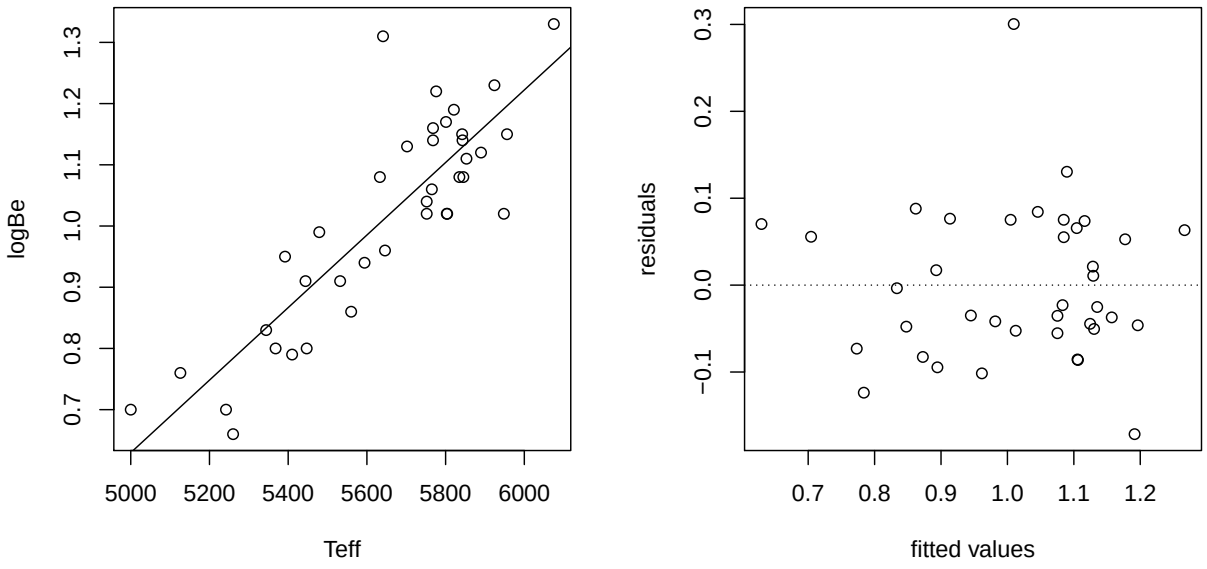
- In addition, we often refer to the heights of the fitted regression line at the points $x_1, \ldots, x_n$ as *fitted values*, denoting them by

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \ldots, n.$$

  The residuals can then be written as $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$, $i = 1, \ldots, n$.

- **Data example:** The data for this example are taken from [1], which studied the abundance of beryllium in stars which host orbiting planets versus stars without orbiting planets. The right-hand plot below shows the beryllium abundance `logBe` versus temperature `Teff` for 38 stars. The least-squares regression line $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ is overlaid. the left-hand plot shows the values of the residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_{38}$ versus the fitted values $\hat{Y}_1, \ldots, \hat{Y}_{38}$. Plotting the residuals versus the fitted values is one way to check whether the linear regression model is a good fit to the data; if we see a random scatter of points, it suggests that the linear regression model is appropriate.

```
# read the data in from PennState Center for Astrostatistics website
data <- read.table(file="https://astrostatistics.psu.edu/datasets/censor.dat",
                   header=TRUE,sep="")

# remove some censored data points and some outliers, similar to the paper:
rows.rm <- c(which(is.na(data$sig_Be) | data$Teff >= 6100 ),51,44,21)
beryllium <- data[-rows.rm,]

x <- beryllium$Teff # temperature
Y <- beryllium$logN_Be # log of beryllium abundance

beta1.hat <- cor(x,Y)*sd(Y)/sd(x)
beta0.hat <- mean(Y) - beta1.hat*mean(x)

Y.hat <- beta0.hat + beta1.hat * x
e.hat <- Y - Y.hat

par(mfrow=c(1,2))

plot(Y ~ x , xlab="Teff",ylab = "logBe")
abline(beta0.hat,beta1.hat)

plot(e.hat ~ Y.hat, ylab = "residuals",xlab="fitted values")
abline(h=0,lty=3)
```



Mean and variance of least-squares estimators in simple linear regression

- In this section we will show that the least-squares regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1. We will also get expressions for $\operatorname{Var}\hat{\beta}_0$, $\operatorname{Var}\hat{\beta}_1$, and $\operatorname{Cov}(\hat{\beta}_0, \hat{\beta}_1)$. As a first step, we will get expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ in terms of their target values β_0 and β_1 and a weighted sum of the error terms $\varepsilon_1, \ldots, \varepsilon_n$.

- **Result:** We have

$$\hat{\beta}_0 = \beta_0 + \frac{\bar{x}_n}{S_{xx}} \sum_{i=1}^{n} \left[\frac{S_{xx}}{n\bar{x}_n} - (x_i - \bar{x}_n) \right] \varepsilon_i \tag{5}$$

$$\hat{\beta}_1 = \beta_1 + \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}_n)\varepsilon_i. \tag{6}$$

Derivation:

We get these expressions by replacing Y_i, $\bar{Y}_n$, and $Y_i - \bar{Y}_n$ in (3) and (4) with

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \\ \bar{Y}_n &= \beta_0 + \beta_1 \bar{x}_n + \bar{\varepsilon}_n \\ Y_i - \bar{Y}_n &= \beta_1 (x_i - \bar{x}_n) + \varepsilon_i - \bar{\varepsilon}_n, \end{aligned}$$

where $\bar{\varepsilon}_n = n^{-1} \sum_{i=1}^{n} \varepsilon_i$.

For $\hat{\beta}_1$ we have

$$\begin{aligned} \hat{\beta}_1 &= S_{xx}^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)(Y_i - \bar{Y}_n) \\ &= S_{xx}^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)[\beta_1 (x_i - \bar{x}_n) + \varepsilon_i - \bar{\varepsilon}_n] \\ &= \beta_1 + S_{xx}^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)\varepsilon_i - S_{xx}^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)\bar{\varepsilon}_n \\ &= \beta_1 + S_{xx}^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)\varepsilon_i. \end{aligned}$$

For $\hat{\beta}_0$ we have

$$\begin{aligned}
\hat{\beta}_0 &= \bar{Y}_n - \hat{\beta}_1 \bar{x}_n \\
&= \beta_0 + \beta_1 \bar{x}_n + \bar{\varepsilon}_n - \hat{\beta}_1 \bar{x}_n \\
&= \beta_0 + \bar{\varepsilon}_n - (\hat{\beta}_1 - \beta_1)\bar{x}_n \\
&= \beta_0 + \bar{\varepsilon}_n - \bar{x}_n S_{xx}^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)\varepsilon_i \\
&= \beta_0 + \bar{x}_n S_{xx}^{-1} \sum_{i=1}^{n} \frac{S_{xx}}{n\bar{x}_n}\varepsilon_i - \bar{x}_n S_{xx}^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)\varepsilon_i \\
&= \beta_0 + \frac{\bar{x}_n}{S_{xx}} \sum_{i=1}^{n} \left[\frac{S_{xx}}{n\bar{x}_n} - (x_i - \bar{x}_n) \right] \varepsilon_i.
\end{aligned}$$

- **Unbiasedness of least-squares regression coefficients:** We are now in a position to easily show that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1. Applying the fact that $\mathbb{E}\varepsilon_i = 0$ for all $i = 1, \ldots, n$ to expressions (5) and (6) gives

$$\mathbb{E}\hat{\beta}_0 = \beta_0 + \frac{\bar{x}_n}{S_{xx}} \sum_{i=1}^{n} \left[\frac{S_{xx}}{n\bar{x}_n} - (x_i - \bar{x}_n) \right] \mathbb{E}\varepsilon_i = \beta_0$$

and

$$\mathbb{E}\hat{\beta}_1 = \beta_1 + S_{xx}^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)\mathbb{E}\varepsilon_i = \beta_1.$$

- **Variance and covariance of least-squares regression coefficients:** We have

$$\operatorname{Var} \hat{\beta}_0 = (n^{-1} + \bar{x}_n^2 S_{xx}^{-1})\sigma^2 \tag{7}$$

$$\operatorname{Var} \hat{\beta}_1 = S_{xx}^{-1}\sigma^2 \tag{8}$$

$$\operatorname{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x}_n S_{xx}^{-1}\sigma^2. \tag{9}$$

Derivations: For $\hat{\beta}_0$ we have

$$\begin{aligned}
\operatorname{Var}\hat{\beta}_0 &= \operatorname{Var}\left(\beta_0 + \frac{\bar{x}_n}{S_{xx}}\sum_{i=1}^{n}\left[\frac{S_{xx}}{n\bar{x}_n} - (x_i - \bar{x}_n)\right]\varepsilon_i\right) \\
&= \frac{\bar{x}_n^2}{S_{xx}^2}\operatorname{Var}\left(\sum_{i=1}^{n}\left[\frac{S_{xx}}{n\bar{x}_n} - (x_i - \bar{x}_n)\right]\varepsilon_i\right) \\
&= \frac{\bar{x}_n^2}{S_{xx}^2}\sum_{i=1}^{n}\left[\frac{S_{xx}}{n\bar{x}_n} - (x_i - \bar{x}_n)\right]^2 \operatorname{Var}\varepsilon_i \quad \text{(by independence of } \varepsilon_1, \ldots, \varepsilon_n) \\
&= \frac{\bar{x}_n^2}{S_{xx}^2}\left[\frac{S_{xx}^2}{n\bar{x}_n^2} - 2\sum_{i=1}^{n}(x_i - \bar{x}_n)\frac{S_{xx}}{n\bar{x}_n} + \sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right]\sigma^2 \\
&= \left(\frac{1}{n} + \frac{\bar{x}_n^2}{S_{xx}}\right)\sigma^2.
\end{aligned}$$

For $\hat{\beta}_1$ we have

$$\begin{aligned}
\operatorname{Var}\hat{\beta}_1 &= \operatorname{Var}\left(\beta_1 + S_{xx}^{-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)\varepsilon_i\right) \\
&= S_{xx}^{-2}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \operatorname{Var}\varepsilon_i \quad \text{(by independence of } \varepsilon_1, \ldots, \varepsilon_n) \\
&= S_{xx}^{-1}\sigma^2.
\end{aligned}$$

For $\operatorname{Cov}(\hat{\beta}_0, \hat{\beta}_1)$ we have

$$\begin{aligned}
\operatorname{Cov}(\hat{\beta}_0, \hat{\beta}_1) &= \mathbb{E}[(\hat{\beta}_0 - \mathbb{E}\hat{\beta}_0)(\hat{\beta}_1 - \mathbb{E}\hat{\beta}_1)] \\
&= \mathbb{E}[(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1)] \\
&= \mathbb{E}\left(\frac{\bar{x}_n}{S_{xx}}\sum_{i=1}^{n}\left[\frac{S_{xx}}{n\bar{x}_n} - (x_i - \bar{x}_n)\right]\varepsilon_i \cdot S_{xx}^{-1}\sum_{j=1}^{n}(x_j - \bar{x}_n)\varepsilon_j\right) \\
&= \frac{\bar{x}_n}{S_{xx}^2}\mathbb{E}\left[\sum_{i=1}^{n}\left[\frac{S_{xx}}{n\bar{x}_n} - (x_i - \bar{x}_n)\right](x_i - \bar{x}_n)\varepsilon_i^2\right. \\
&\qquad\qquad \left. + \sum_{i\neq j}\left[\frac{S_{xx}}{n\bar{x}_n} - (x_i - \bar{x}_n)\right](x_j - \bar{x}_n)\varepsilon_i\varepsilon_j\right] \\
&= \frac{\bar{x}_n}{S_{xx}^2}\mathbb{E}\sum_{i=1}^{n}\left[\frac{S_{xx}}{n\bar{x}_n}(x_i - \bar{x}_n) - (x_i - \bar{x}_n)^2\right]\varepsilon_i^2 \quad \text{(by independence of } \varepsilon_1, \ldots, \varepsilon_n) \\
&= -\frac{\bar{x}_n}{S_{xx}^2}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \operatorname{Var}\varepsilon_i \\
&= -\frac{\bar{x}_n}{S_{xx}}\sigma^2.
\end{aligned}$$

- **Mean and variance of estimated function at a point:** We will estimate the value of the regression function f at a "new" point x_{new}, given by $f(x_{\text{new}}) = \beta_0 + \beta_1 x_{\text{new}}$, with $\hat{f}(x_{\text{new}}) = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}$. We have

$$\mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}) = \beta_0 + \beta_1 x_{\text{new}} \tag{10}$$

$$\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}) = \left[\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x}_n)^2}{S_{xx}}\right]\sigma^2. \tag{11}$$

Derivations: For the expectation, we have

$$\mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}) = \mathbb{E}\hat{\beta}_0 + \mathbb{E}\hat{\beta}_1 x_{\text{new}} = \beta_0 + \beta_1 x_{\text{new}}.$$

For the variance, we use the fact that for any random variables U and V and constants a and b, we have

$$\text{Var}(aU + bV) = a^2 \,\text{Var}\, U + b^2 \,\text{Var}\, V + 2ab \,\text{Cov}(U, V).$$

From (7), (8), and (9), we have

$$\begin{aligned}
\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}) &= \text{Var}\,\hat{\beta}_0 + x_{\text{new}}^2 \,\text{Var}\,\hat{\beta}_1 + 2x_{\text{new}} \,\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\
&= (\bar{x}_n^2 S_{xx}^{-1} + n^{-1})\sigma^2 + x_{\text{new}}^2 S_{xx}^{-1}\sigma^2 - 2x_{\text{new}}\bar{x}_n S_{xx}^{-1}\sigma^2 \\
&= [n^{-1} + S_{xx}^{-1}(x_{\text{new}} - \bar{x}_n)^2]\sigma^2.
\end{aligned}$$

- Lastly in this section we give an estimator for the variance σ^2 of the error terms $\varepsilon_1, \ldots, \varepsilon_n$ and claim that it is unbiased.

Result: For

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

we have $\mathbb{E}\hat{\sigma}^2 = \sigma^2$.

Proof: We begin by rewriting $\hat{\sigma}^2$ as

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-2}\sum_{i=1}^{n}\left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right]^2 \\
&= \frac{1}{n-2}\sum_{i=1}^{n}\left[\varepsilon_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)x_i\right]^2 \\
&= \frac{1}{n-2}\sum_{i=1}^{n}\Big[\varepsilon_i^2 + (\hat{\beta}_0 - \beta_0)^2 + (\hat{\beta}_1 - \beta_1)^2 x_i^2 \\
&\qquad -2\varepsilon_i(\hat{\beta}_0 - \beta_0) - 2\varepsilon_i(\hat{\beta}_1 - \beta_1)x_i - 2(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1)x_i\Big].
\end{aligned}$$

From here we see that

$$\mathbb{E}\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left[\mathbb{E}\varepsilon_i^2 + \text{Var}\,\hat{\beta}_0 + x_i \,\text{Var}\,\hat{\beta}_1 - 2\mathbb{E}\varepsilon_i(\hat{\beta}_0 - \beta_0) - 2\mathbb{E}\varepsilon_i(\hat{\beta}_1 - \beta_1)x_i - 2x_i \,\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)\right],$$

where

$$\mathbb{E}\varepsilon_i(\hat{\beta}_0 - \beta_0) = \mathbb{E}\varepsilon_i \frac{\bar{x}_n}{S_{xx}} \sum_{j=1}^{n} \left[\frac{S_{xx}}{n\bar{x}_n} - (x_j - \bar{x}_n) \right] \varepsilon_j = \left[\frac{1}{n} + \frac{\bar{x}_n(x_i - \bar{x}_n)}{S_{xx}} \right] \sigma^2$$

$$\mathbb{E}\varepsilon_i(\hat{\beta}_1 - \beta_1)x_i = \mathbb{E}\varepsilon_i \frac{1}{S_{xx}} \sum_{j=1}^{n} (x_j - \bar{x}_n)\varepsilon_j x_i = \frac{(x_i - \bar{x}_n)x_i}{S_{xx}} \sigma^2,$$

by (5) and (6). Substituting these expressions as well as those in (7), (8), and (9) and simplifying gives the result.

Inference in simple linear regression

- In this section we assume that the values $Y_1, \ldots, Y_n$ are Normally distributed around the values of the regression function. That is, we assume that the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are independent Normal$(0, \sigma^2)$ random variables. This assumption gives us many useful results which enable inference, that is the construction of confidence intervals and tests of hypotheses.
- **Sampling distribution results:** If $\varepsilon_1, \ldots, \varepsilon_n$ are independent Normal$(0, \sigma^2)$ random variables, then

$$\hat{\beta}_0 \sim \text{Normal}(\beta_0, (n^{-1} + \bar{x}_n^2 S_{xx}^{-1})\sigma^2) \tag{12}$$

$$\hat{\beta}_1 \sim \text{Normal}(\beta_1, S_{xx}^{-1}\sigma^2) \tag{13}$$

$$\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \sim \text{Normal}(\beta_0 + \beta_1 x_{\text{new}}, [n^{-1} + S_{xx}^{-1}(x_{\text{new}} - \bar{x}_n)^2]\sigma^2) \tag{14}$$

and

$$(n-2)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}. \tag{15}$$

Results (12) and (13) follow from the fact that $\hat{\beta}_0$ and $\hat{\beta}_1$ can each be written as a constant plus a linear combination of the Normal$(0, \sigma^2)$ error terms (recall that in STAT 511 we used moment generating functions to show that linear combinations of Normal random variables are Normal). Result (14) follows from the Normality of $\hat{\beta}_0$ and $\hat{\beta}_1$ and from (11). The proof of (15) is beyond the scope of this course, but do note its similarity to results we have seen earlier!

In addition, we have

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{n^{-1} + \bar{x}_n^2 S_{xx}^{-1}}} \sim t_{n-2} \tag{16}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}\sqrt{S_{xx}^{-1}}} \sim t_{n-2} \tag{17}$$

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} - (\beta_0 + \beta_1 x_{\text{new}})}{\hat{\sigma}\sqrt{n^{-1} + S_{xx}^{-1}(x_{\text{new}} - \bar{x}_n)^2}} \sim t_{n-2}. \tag{18}$$

These results follow from (12)–(15) and from the fact that under Normal error terms, $\hat{\sigma}^2$ is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$.

- These distributional results can be immediately put to use in the ways that follow.
- **Confidence intervals:**
  - From (16) we see that a $(1-\alpha)100\%$ confidence interval for β_0 is given by

$$\hat{\beta}_0 \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{n^{-1}+\bar{x}_n^2 S_{xx}^{-1}}. \tag{19}$$

  - From (17) we see that $(1-\alpha)100\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{S_{xx}^{-1}}. \tag{20}$$

  - From (18) we see that a $(1-\alpha)100\%$ confidence interval for $\beta_0+\beta_1 x_{\text{new}}$ is given by

$$(\hat{\beta}_0+\hat{\beta}_1 x_{\text{new}}) \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{n^{-1}+S_{xx}^{-1}(x_{\text{new}}-\bar{x}_n)^2}. \tag{21}$$

- **Tests of hypotheses for simple linear regression coefficients:**
  - Consider testing hypotheses about β_0 with respect to a null value β_0^*, and define

$$T_{0,n} = \frac{\hat{\beta}_0 - \beta_0^*}{\hat{\sigma}\sqrt{n^{-1}+\bar{x}_n^2 S_{xx}^{-1}}}.$$

    We have the following:

H_0	H_1	Reject H_0 at α iff	p-value
$\beta_0 \le \beta_0^*$	$\beta_0 > \beta_0^*$	$T_{0,n} > t_{n-2,\alpha}$	$1 - F_{t_{n-2}}(T_{0,n})$
$\beta_0 \ge \beta_0^*$	$\beta_0 < \beta_0^*$	$T_{0,n} < -t_{n-2,\alpha}$	$F_{t_{n-2}}(T_{0,n})$
$\beta_0 = \beta_0^*$	$\beta_0 \ne \beta_0^*$	$\lvert T_{0,n}\rvert > t_{n-2,\alpha/2}$	$2[1 - F_{t_{n-2}}(\lvert T_{0,n}\rvert)]$

  - Consider testing hypotheses about β_1 with respect to a null value β_1^*, and define

$$T_{1,n} = \frac{\hat{\beta}_1 - \beta_1^*}{\hat{\sigma}\sqrt{S_{xx}^{-1}}}. \tag{22}$$

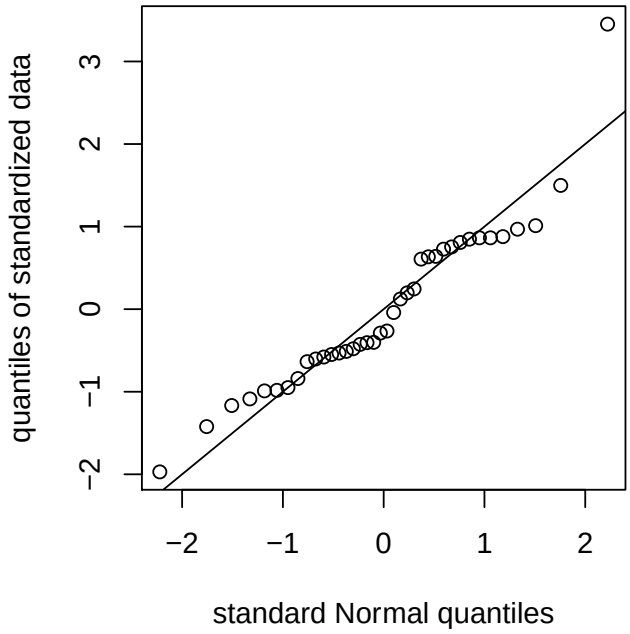
We have the following:

H_0	H_1	Reject H_0 at α iff	p-value
$\beta_1 \leq \beta_1^*$	$\beta_1 > \beta_1^*$	$T_{1,n} > t_{n-2,\alpha}$	$1 - F_{t_{n-2}}(T_{1,n})$
$\beta_1 \geq \beta_1^*$	$\beta_1 < \beta_1^*$	$T_{1,n} < -t_{n-2,\alpha}$	$F_{t_{n-2}}(T_{1,n})$
$\beta_1 = \beta_1^*$	$\beta_1 \neq \beta_1^*$	$\lvert T_{1,n}\rvert > t_{n-2,\alpha/2}$	$2[1 - F_{t_{n-2}}(\lvert T_{1,n}\rvert)]$

In the above tables, $F_{t_{n-2}}$ represents the cdf of the t_{n-2} distribution.

- **Data example:** For the beryllium abundance versus temperature of stars data, we can build 95% confidence intervals for the linear regression coefficients β_0 and β_1 as well as test hypotheses about them using the results in this section provided the values of $Y_1, \ldots, Y_n$ are Normally distributed around the regression function. To check whether we can assume this, we typically look at a Normal quantile-quantile plot of the residuals. The following R code generates this plot:

```
qqnorm(scale(e.hat),main="",
       xlab="standard Normal quantiles",
       ylab="quantiles of standardized data")
abline(0,1)
```



Recall that if the points arrange themselves along a straight line, we may assume that they are realizations from a Normal distribution. Besides one outlier, there do not appear to be huge

deviations from Normality, so we will proceed under the assumption that the error terms $\varepsilon_1, \ldots, \varepsilon_{38}$ are Normally distributed for the beryllium data.

The following R code computes the upper and lower bounds of the confidence intervals given in expressions (19) and (20).

```
n <- length(Y)
sigma.hat <- sqrt( sum(e.hat^2)/(n-2) )
x.bar <- mean(x)
Sxx <- sum( (x - x.bar)^2 )

alpha <- 0.05
tval <- qt(1-alpha/2,n-2)

# compute lower and upper limit of confidence interval for beta.0
se.beta.0.hat <- sigma.hat * sqrt( 1/n + x.bar^2 / Sxx  )
beta0.hat - tval * se.beta.0.hat
beta0.hat + tval * se.beta.0.hat

# compute lower and upper limit of confidence interval for beta.1
se.beta.1.hat <- sigma.hat * sqrt(1/Sxx)
beta1.hat - tval * se.beta.1.hat
beta1.hat + tval * se.beta.1.hat
```

The 95% confidence interval

- for β_0 is $(-2.998309, -1.668342)$.
- for β_1 is $(0.0004749841, 0.0007102168)$.

The following R code computes the p-value based on the test statistic in (22) for testing H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$.

```
T1 <- beta1.hat/ (sigma.hat*sqrt(1/Sxx))
pval <- 2*(1 - pt(abs(T1),n-2))
```

The value of $T_{1,n}$ is 10.21839 and the associated p-value is 3.477219×10^{-12}, so we would conclude at very small values of the significance level α that the slope parameter is nonzero and, moreover, positive, since the sign of $\hat{\beta}_1$ is positive, indicating a significant positive linear relationship between the beryllium levels and temperature of stars.

Prediction interval for new observation

- For a given x_{new}, we would like to construct an interval in which the as-yet-unobserved value of Y_{new} will fall with probability $1-\alpha$.

  We begin by finding the distribution of the residual corresponding to the pair $(x_{\text{new}}, Y_{\text{new}})$, which we define as

  $$\hat{\varepsilon}_{\text{new}} = Y_{\text{new}} - (\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}).$$

  We may rewrite this as

  $$\begin{aligned}\hat{\varepsilon}_{\text{new}} &= Y_{\text{new}} - (\beta_0 + \beta_1 x_{\text{new}}) + (\beta_0 + \beta_1 x_{\text{new}}) - (\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}) \\ &= \varepsilon_{\text{new}} - [(\hat{\beta}_0 - \beta_0) + x_{\text{new}}(\hat{\beta}_1 - \beta_1)],\end{aligned}$$

  where $\varepsilon_{\text{new}} = Y_{\text{new}} - (\beta_0 + \beta_1 x_{\text{new}})$ is the deviation of Y_{new} from the height of the true regression line $\beta_0 + \beta_1 x_{\text{new}}$, which is a $\text{Normal}(0, \sigma^2)$ random variable independent of $\hat{\beta}_0$ and $\hat{\beta}_1$. From here we have

  $$\begin{aligned}\mathbb{E}\hat{\varepsilon}_{\text{new}} &= \mathbb{E}\varepsilon_{\text{new}} - [\mathbb{E}(\hat{\beta}_0 - \beta_0) + x_{\text{new}}\mathbb{E}(\hat{\beta}_1 - \beta_1)] = 0 \\ \operatorname{Var}\hat{\varepsilon}_{\text{new}} &= \operatorname{Var}\varepsilon_{\text{new}} + \operatorname{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}) = [1 + n^{-1} + S_{xx}^{-1}(x_{\text{new}} - \bar{x}_n)^2]\sigma^2.\end{aligned}$$

  Moreover, $\hat{\varepsilon}_{\text{new}}$ has a Normal distribution, since it is a linear combination of Normally distributed random variables. Thus

  $$\hat{\varepsilon}_{\text{new}} \sim \text{Normal}(0, [1 + n^{-1} + S_{xx}^{-1}(x_{\text{new}} - \bar{x}_n)^2]\sigma^2).$$

  Combining this result with (15) gives

  $$\frac{\hat{\varepsilon}_{\text{new}}}{\hat{\sigma}\sqrt{1 + n^{-1} + S_{xx}^{-1}(x_{\text{new}} - \bar{x}_n)^2}} \sim t_{n-2}. \tag{23}$$

  This allows us to write

  $$P\left(-t_{n-2,\alpha/2} < \frac{Y_{\text{new}} - (\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}})}{\hat{\sigma}\sqrt{1 + n^{-1} + S_{xx}^{-1}(x_{\text{new}} - \bar{x}_n)^2}} < t_{n-2,\alpha/2}\right) = 1 - \alpha$$

  for any $\alpha \in (0, 1)$, which is equivalent to

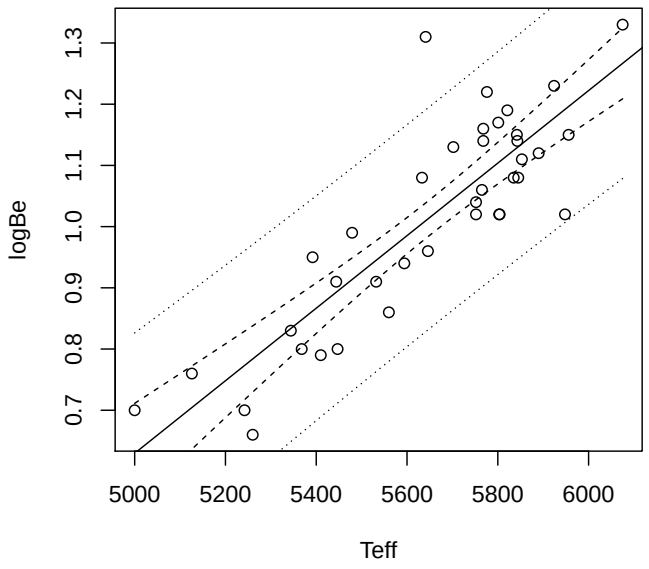
  $$\begin{aligned}P\Big(&\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} - t_{n-2,\alpha/2}\hat{\sigma}\sqrt{1 + n^{-1} + S_{xx}^{-1}(x_{\text{new}} - \bar{x}_n)^2} \\ &< Y_{\text{new}} < \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} + t_{n-2,\alpha/2}\hat{\sigma}\sqrt{1 + n^{-1} + S_{xx}^{-1}(x_{\text{new}} - \bar{x}_n)^2}\Big) = 1 - \alpha.\end{aligned}$$

  Therefore, if we wish to construct an interval around the fitted regression line at x_{new} which will contain the as-yet-unobserved Y_{new} with probability $1-\alpha$ for any $\alpha \in (0, 1)$, the above suggests the interval given by

  $$\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{1 + n^{-1} + S_{xx}^{-1}(x_{\text{new}} - \bar{x}_n)^2}.$$

- **Data example:** We now compute for the beryllium data 95% confidence intervals for $\beta_0 + \beta_1 x_{\text{new}}$ as well as 95% prediction intervals for Y_{new} of a new observation $(Y_{\text{new}}, x_{\text{new}})$ for a sequence of values of x_{new} within the range of the observed values of the covariate. The following R code computes and plots for each value x_{new} the upper and lower bounds of two intervals.

```
plot(Y ~ x , xlab="Teff",ylab = "logBe")
abline(beta0.hat,beta1.hat)

alpha <- .05
tval <- qt(1-alpha/2,n-2)

x.seq <- seq(min(x),max(x),length=99)
se.Y.hat.new <- sigma.hat * sqrt( 1/n + (x.seq - x.bar)^2 / Sxx )
loconf <- beta0.hat+beta1.hat*x.seq - tval * se.Y.hat.new
upconf <- beta0.hat+beta1.hat*x.seq + tval * se.Y.hat.new

lines(loconf~x.seq,lty=2)
lines(upconf~x.seq,lty=2)

sd.e.hat.new <- sigma.hat *sqrt( 1 + 1/n + (x.seq - x.bar)^2 / Sxx )
lopred <- beta0.hat + beta1.hat * x.seq - tval * sd.e.hat.new
uppred <- beta0.hat + beta1.hat * x.seq + tval * sd.e.hat.new

lines(lopred~x.seq,lty=3)
lines(uppred~x.seq,lty=3)
```



We see that the prediction intervals for new observations are wider than the confidence intervals for the height of the regression line. Moreover, we see that the intervals are narrower for values of x_{new} closer to $\bar{x}_n$.

Least-squares estimators as MLEs under Normal error terms

- If we assume

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent $\text{Normal}(0, \sigma^2)$ random variables, it is the same as

$$Y_i \sim \text{Normal}(\beta_1 + \beta_1 x_i, \sigma^2), \quad i = 1, \ldots, n,$$

with $Y_1, \ldots, Y_n$ independent.

- In this case the likelihood function based on the data $Y_1, \ldots, Y_n$ and $x_1, \ldots, x_n$ is given by

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2; Y_1, \ldots, Y_n, x_1, \ldots, x_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{1}{2\sigma^2} [Y_i - (\beta_0 + \beta_1 x_i)]^2 \right) \\ &= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2 \right) \end{aligned}$$

and the log-likelihood is

$$\ell(\beta_0, \beta_1, \sigma^2; Y_1, \ldots, Y_n, x_1, \ldots, x_n) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2.$$

We see that the pair $(\hat{\beta}_0, \hat{\beta}_1)$ which maximizes the likelihood function is the pair which minimizes the least-squares objective function

$$Q_n(\beta_0, \beta_1) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2.$$

So in the simple linear regression model with Normal error terms, the maximum likelihood estimators of β_0 and β_1 are the same as the least-squares estimators.

Likelihood ratio test for slope parameter

- In this section we find the form of the likelihood ratio test for H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$ and show that when it is calibrated to have size $\alpha \in (0, 1)$, it is equivalent to the test

$$\text{Reject } H_0 \text{ iff } \frac{|\hat{\beta}_1|}{\hat{\sigma} \sqrt{S_{xx}^{-1}}} > t_{n-2,\alpha/2}, \tag{24}$$

which is the test introduced previously.

- Obtaining an expression for the likelihood ratio will require finding the value of the triplet $(\beta_0, \beta_1, \sigma^2)$ which maximizes the likelihood function over i) the space dictated for the parameters by the null hypotheses, which is $\{(\beta_0, \beta_1, \sigma^2) : \beta_0 \in \mathbb{R}, \beta_1 = 0, \sigma^2 \geq 0\}$ and ii) over the entire parameter space, which is $\{(\beta_0, \beta_1, \sigma^2) : (\beta_0, \beta_1) \in \mathbb{R}^2, \sigma^2 \geq 0\}$. Let $(\hat{\beta}_0^*, \hat{\beta}_1^*, \hat{\sigma}^{*2})$ be the triplet which maximizes the likelihood over the null space. Then we have

$$\begin{aligned}
(\hat{\beta}_0^*, \hat{\beta}_1^*, \hat{\sigma}^{*2}) &= \underset{\{(\beta_0,\beta_1,\sigma^2):\beta_0\in\mathbb{R},\beta_1=0,\sigma^2\geq 0\}}{\operatorname{argmax}} -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 x_i)]^2 \\
&= \underset{\{(\beta_0,\beta_1,\sigma^2):\beta_0\in\mathbb{R},\beta_1=0,\sigma^2\geq 0\}}{\operatorname{argmax}} -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0)^2 \\
&= \left(\bar{Y}_n, 0, n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2\right),
\end{aligned}$$

where the second equality is obtained by substituting $\beta_1 = 0$ and where the third equality can be established with calculus methods. The triplet which maximizes the likelihood over the entire parameter space is $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2_{\text{mle}})$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least-squares estimators of β_0 and β_1 and where

$$\hat{\sigma}^2_{\text{mle}} = \frac{1}{n}\sum_{i=1}^{n}[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

We recognize that the maximum likelihood estimator $\hat{\sigma}^2_{\text{mle}}$ is biased, since we have discussed earlier the unbiased estimator $\hat{\sigma}^2$ of σ^2 which has the factor $(n-2)^{-1}$ instead of n^{-1} in front.

- The likelihood ratio is thus given by

$$\begin{aligned}
\text{LR}(Y_1, \ldots, Y_n, x_1, \ldots, x_n) &= \frac{L(\hat{\beta}_0^*, \hat{\beta}_1^*, \hat{\sigma}^{*2}; Y_1, \ldots, Y_n, x_1, \ldots, x_n)}{L(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2_{\text{mle}}; Y_1, \ldots, Y_n, x_1, \ldots, x_n)} \\
&= \frac{(2\pi)^{-n/2}\left[n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2\right]^{-n/2}\exp(-n/2)}{(2\pi)^{-n/2}\left[n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right]^{-n/2}\exp(-n/2)} \\
&= \left[\frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}\right]^{-n/2},
\end{aligned}$$

and the likelihood ratio test is of the form

$$\text{Reject } H_0 \text{ iff } \left[\frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}\right]^{-n/2} < c \tag{25}$$

for some $c \in [0, 1]$.

- Let $\bar{\hat{\varepsilon}}_n = n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_i$ denote the mean of the residuals, and note that this is equal to zero, since

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i = \frac{1}{n}\sum_{i=1}^{n}Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = \bar{Y}_n - (\hat{\beta}_0 + \hat{\beta}_1\bar{x}_n) = \bar{Y}_n - [(\bar{Y}_n - \hat{\beta}_1\bar{x}_n) + \hat{\beta}_1\bar{x}_n] = 0.$$

Now define the quantity $S_{\hat{\varepsilon}\hat{\varepsilon}} = \sum_{i=1}^n (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}}_n)^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2$ and recall the notation $S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$. Then the rejection criterion of the likelihood ratio test may be rewritten as

$$\begin{aligned}
\left[\frac{S_{YY}}{S_{\hat{\varepsilon}\hat{\varepsilon}}}\right]^{-n/2} &< c \\
\iff \frac{S_{YY}}{S_{\hat{\varepsilon}\hat{\varepsilon}}} &> c^{-2/n} \\
\iff \frac{S_{\hat{\varepsilon}\hat{\varepsilon}} + S_{YY} - S_{\hat{\varepsilon}\hat{\varepsilon}}}{S_{\hat{\varepsilon}\hat{\varepsilon}}} &> c^{-2/n} \\
\iff \frac{S_{YY} - S_{\hat{\varepsilon}\hat{\varepsilon}}}{S_{\hat{\varepsilon}\hat{\varepsilon}}} &> c^{-2/n} - 1.
\end{aligned}$$

At this point we will use the fact that

$$S_{YY} - S_{\hat{\varepsilon}\hat{\varepsilon}} = \hat{\beta}_1^2 S_{xx}, \tag{26}$$

which we will establish later. We now have that the rejection criterion of the likelihood ratio test is equivalent to

$$\begin{aligned}
\frac{\hat{\beta}_1^2 S_{xx}}{S_{\hat{\varepsilon}\hat{\varepsilon}}} &> c^{-2/n} - 1 \\
\iff \frac{\hat{\beta}_1^2 S_{xx}}{S_{\hat{\varepsilon}\hat{\varepsilon}}/(n-2)} &> (c^{-2/n} - 1)(n-2) \\
\iff \frac{\hat{\beta}_1^2}{\hat{\sigma}^2/S_{xx}} &> (c^{-2/n} - 1)(n-2) \\
\iff \frac{|\hat{\beta}_1|}{\hat{\sigma}\sqrt{S_{xx}^{-1}}} &> [(c^{-2/n} - 1)(n-2)]^{1/2},
\end{aligned}$$

where we have used the fact that

$$\hat{\sigma}^2 = S_{\hat{\varepsilon}\hat{\varepsilon}}/(n-2).$$

We see that the test in (24) and the likelihood ratio test in (25) are equivalent when

$$t_{n-2,\alpha/2} = [(c^{-2/n} - 1)(n-2)]^{1/2} \iff c = [t_{n-2,\alpha/2}^2/(n-2) + 1]^{-n/2}.$$

- We now show that $S_{YY} - S_{\hat{\varepsilon}\hat{\varepsilon}} = \hat{\beta}_1^2 S_{xx}$ from (26). We have

$$\begin{aligned}
S_{YY} - S_{\hat{\varepsilon}\hat{\varepsilon}} &= \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 - \sum_{i=1}^{n}(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}}_n)^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 - \sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) - [\bar{Y}_n - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_n)])^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 - \sum_{i=1}^{n}[(Y_i - \bar{Y}_n) - \hat{\beta}_1(x_i - \bar{x}_n)]^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 - \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 + 2\hat{\beta}_1 \sum_{i=1}^{n}(Y_i - \bar{Y}_n)(x_i - \bar{x}_n) - \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \\
&= 2\hat{\beta}_1 S_{xx} S_{xx}^{-1} \sum_{i=1}^{n}(Y_i - \bar{Y}_n)(x_i - \bar{x}_n) - \hat{\beta}_1^2 S_{xx} \\
&= 2\hat{\beta}_1^2 S_{xx} - \hat{\beta}_1^2 S_{xx} \\
&= \hat{\beta}_1^2 S_{xx}.
\end{aligned}$$

References

[1] Nuno C Santos, G Israelian, RJ García López, M Mayor, R Rebolo, S Randich, A Ecuvillon, and C Domínguez Cerdeña. Are beryllium abundances anomalous in stars with giant planets? *Astronomy & Astrophysics*, 427(3):1085–1096, 2004.