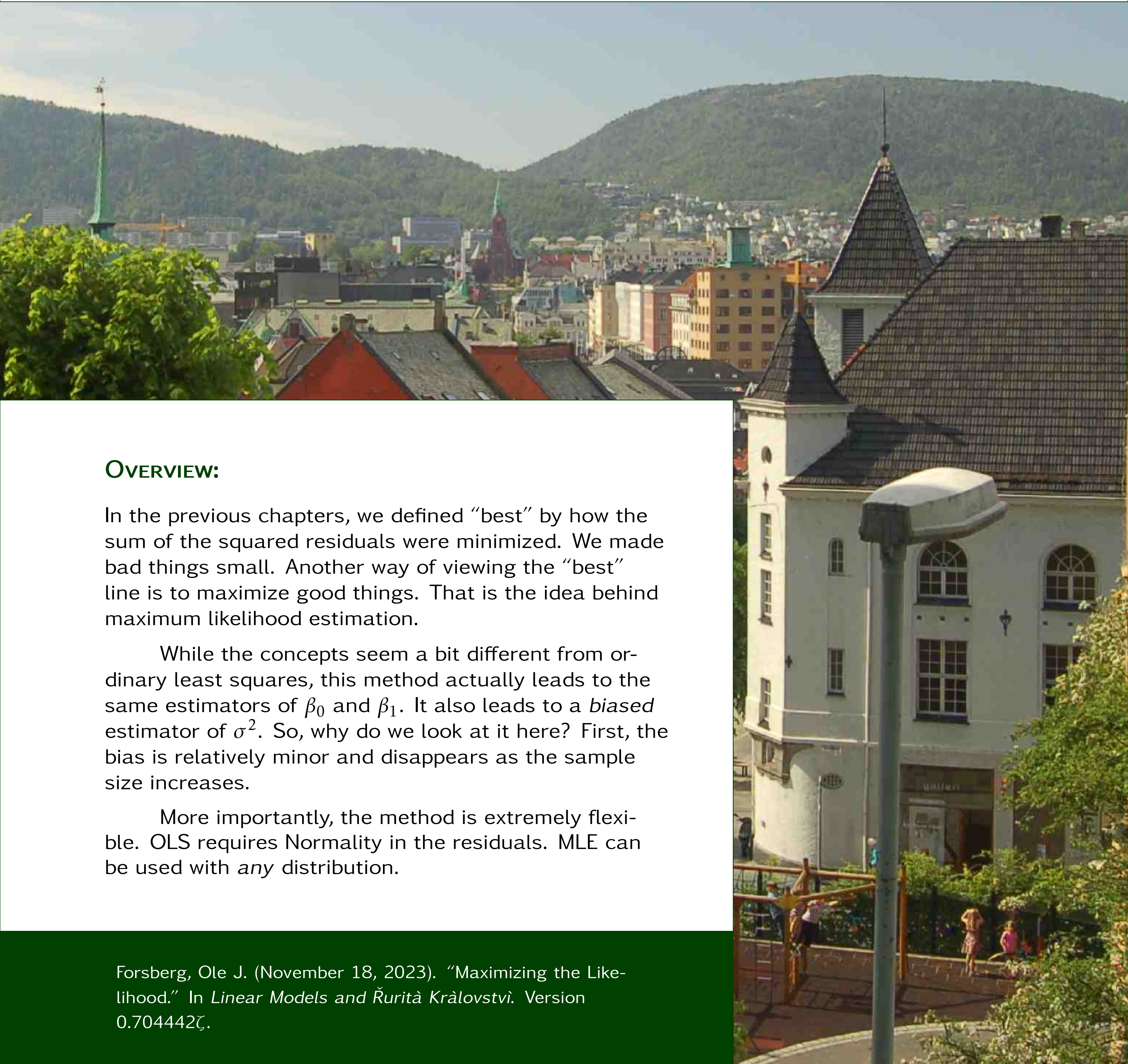
Chapter 9:

Maximizing the Likelihood



Overview:

In the previous chapters, we defined "best" by how the sum of the squared residuals were minimized. We made bad things small. Another way of viewing the "best" line is to maximize good things. That is the idea behind maximum likelihood estimation.

While the concepts seem a bit different from ordinary least squares, this method actually leads to the same estimators of β_0 and β_1. It also leads to a *biased* estimator of σ^2. So, why do we look at it here? First, the bias is relatively minor and disappears as the sample size increases.

More importantly, the method is extremely flexible. OLS requires Normality in the residuals. MLE can be used with *any* distribution.

Forsberg, Ole J. (November 18, 2023). "Maximizing the Likelihood." In *Linear Models and Řurità Kràlovstvì*. Version 0.704442ζ.

Chapter Contents

❧❧❧

In the previous chapters, we have progressed from our desire to minimize some function of the residuals. This led to several related techniques:

- ordinary least squares
- weighted least squares
- generalized least squares
- ordinary least absolutes

All of these techniques sought to make the ‘bad’ things as small as possible, to produce a model that minimizes these residuals. However, our first definition of “best” from Page 16 was based on making good things as large as possible, where that “good thing” is the likelihood of observing this particular data.[1]

The theory is that the estimate most likely to have produced the observed data is the “best” estimate. Note that this differs from previous estimation methods in both the objective function and the size we desire. Bigger is better. . . bigger in terms of the “likelihood.”

[1] It is called the “likelihood of the data given our parameter estimates.”

9.1: The Likelihood

From a theoretical standpoint, the likelihood is just a generalization of probability. Where probability is bounded by both 0 and 1, the likelihood is only bounded below by 0. Values with higher probabilities are more likely to be observed. The same is true of likelihood: Values with higher likelihoods are more likely to be observed.

In the discrete case, the likelihood and the probability (mass) are the same. In the continuous case, the likelihood is the probability density. In other words, you really *have* come across the likelihood before. In your previous statistics course, the likelihood was called the "probability density" for continuous random variables and the "probability mass" for discrete random variables.

The difference between the likelihood and the probability mass or density is only one of emphasis. The probability mass (or density) is a function of observable values given the parameters of the distribution.

The likelihood is a function of the parameters, given the observed values (data). That difference is illustrated in the next two examples.

Example 9.1: Given that the success probability of a binomial random variable is $\pi = 0.25$, what is the probability of observing exactly one success out of two trials? ●

Solution: The probability mass function of the binomial distribution is

$$f(x;\ \pi, n) = \binom{n}{x} \pi^x\ (1-\pi)^{n-x} \tag{9.1}$$

In this particular instance, the probability mass function is

$$f(x;\ \pi = 0.25, n = 2) = \binom{2}{x} 0.25^x\ (1-0.25)^{2-x} \tag{9.2}$$

And now calculating the probability gives

$$\begin{aligned} f(1;\ \pi = 0.25, n = 2) &= \binom{2}{1}\, 0.25^1\ (1 - 0.25)^{2-1} && (9.3) \\ &= 2\ 0.25^1\ (0.75)^1 && (9.4) \\ &= 2\ (0.25)\ (0.75) && (9.5) \\ &= 0.375 && (9.6) \end{aligned}$$

Thus, the probability I observe exactly one success in two trials, given the success probability is 0.25 is just 0.375, which is a probability of 3 in 8. ♦

Example 9.2: Given that I observed exactly on success in two trials, what is the likelihood that the success probability is $\pi = 0.25$? ●

Solution: The likelihood for a binomial random variable is

$$f(\pi;\ x, n) = \binom{n}{x}\, \pi^x\ (1 - \pi)^{n-x} \tag{9.7}$$

In this particular instance, the likelihood is

$$f(\pi;\ x = 1, n = 2) = \binom{2}{1}\, \pi^1\ (1 - \pi)^{2-1} \tag{9.8}$$

Thus, the value of the likelihood for $\pi = 0.25$ is

$$\begin{aligned} f(0.25;\ x = 1, n = 2) &= \binom{2}{1}\, 0.25^1\ (1 - 0.25)^{2-1} && (9.9) \\ &= 2\ (0.25)^1\ (0.75)^1 && (9.10) \\ &= 2\ (0.25)\ (0.75) && (9.11) \\ &= 0.375 && (9.12) \end{aligned}$$

Thus, the likelihood that $\pi = 0.25$ is 0.375. Is this a lot? It depends on the number of data points. In general, the larger your sample size, the smaller the likelihood. Thus, the likelihood can only *meaningfully* be interpreted when in relation to other likelihoods based on the same data. ♦

same data

The probability and the likelihood are *numerically* the same. The use (interpretation), however, is different. With probability, we are looking at a function of possible outcomes. With likelihood, we are looking at a function of possible values of the parameters. Thus, in the first case, we could ask questions about which value of x is most likely. In the second case, we would ask questions about which value of π is most likely.

Likelihood cries out to be maximized. Is $\pi = 0.25$ the maximum likelihood in the previous example? No. Calculate the likelihood of $\pi = 0.40$ to see that 0.25 is not the maximum (the value of π that produces the largest likelihood value). If you calculated $f(0.40;\ x = 1, n = 2) = 0.48$, then you did the calculations correctly.

Note that $f(0.40;\ x = 1, n = 2) > f(0.25;\ x = 1, n = 2)$. Thus, $\pi = 0.25$ is not the maximum likelihood estimate of π in this case. What is? Such optimization requires using calculus. From the above, you should be able to see that the objective function is

$$Q(\pi) = \binom{2}{1} \pi^1\ (1-\pi)^{2-1} = 2\ \pi\ (1-\pi) \tag{9.13}$$

This is a function of the parameter, since we are trying to determine the value of π that is *most* likely, given the data. The optimization proceeds as expected:

$$\frac{d}{d\pi}\ Q(\pi) = 2(1-2\pi) \tag{9.14}$$

$$0 \overset{\text{set}}{=} 2(1-2\hat{\pi}) \tag{9.15}$$

$$0 = 1 - 2\hat{\pi} \tag{9.16}$$

$$1 = 2\hat{\pi} \tag{9.17}$$

$$\frac{1}{2} = \hat{\pi} \tag{9.18}$$

Thus, given that we observed 1 success in 2 trials, the maximum likelihood estimator of π is $\hat{\pi} = 0.500$. For some reason, I am not surprised at this outcome. Are you?

In general, one can show that the maximum likelihood estimator of π is $\hat{\pi} = x/n$, where x is the number of successes and n is the number of trials. I will leave that as an exercise. **exercise**

A second distribution that you probably saw in your previous statistics class is the Poisson distribution. It has just one parameter, λ, the average rate. In this example, we will determine the maximum likelihood estimator of λ.

Example 9.3: Let Y be the number of Ruritanians walking through the door of the Valné Shromáždění, the general assembly building of Ruritania. The King would like use to estimate the average entering between noon and 1pm. To do this, we simply measure the number of people who entered the building during that hour on Monday.

That number is $y = 17$. With this information, let us calculate the estimate of λ using maximum likelihood estimation. ●

Solution: The likelihood for a discrete distribution, like the Poisson, is just the probability mass function:

$$\mathcal{L}(\lambda;\, y) = \frac{e^{-\lambda}\, \lambda^{y}}{y!} \tag{9.19}$$

That is the likelihood of each observation. Here, we only took one measurement (on Monday). Thus this is also the entire likelihood.

The next step is to maximize the likelihood with respect to the parameter, λ:

$$\frac{d}{d\lambda}\mathcal{L}(\lambda;\, y) = \frac{d}{d\lambda}\left(\frac{e^{-\lambda}\, \lambda^{y}}{y!}\right) \tag{9.20}$$

$$= \frac{-e^{-\lambda}\, y(\lambda^{y-1}) + e^{-\lambda}\lambda^{y}}{y!} \tag{9.21}$$

Now, set this equal to zero and solve for the parameter.

$$0 \overset{\text{set}}{=} -e^{-\hat{\lambda}}\, y(\hat{\lambda}^{y-1}) + e^{-\hat{\lambda}}\hat{\lambda}^{y} \tag{9.22}$$

Since λ is constrained to be positive, we have the following simplification

$$0 = -y(\hat{\lambda}^{y-1}) + \hat{\lambda}^{y} \tag{9.23}$$

$$0 = -y + \hat{\lambda} \tag{9.24}$$

Thus, the maximum likelihood estimate of λ is $\hat{\lambda} = y = 17$.

And so, we report to His Majesty that our estimate of the average number of people passing through the doors of the Valné Shromáždění is 17 per hour. ♦

surprised?

Example 9.4: His Majesty liked our report, especially our font. However, he asked an excellent question: “Bylo by lepší měřit více než jednou?”

To address his point, I decided to take multiple measurements over several days. So, for the next week, I measured the number of people entering the Valné Shromáždění an hour at a time, randomly selecting the time of day. Here is that data: $15, 20, 23, 34, 23$.

With that new data what is the maximum likelihood estimator of λ, given these $n = 5$ measurements? ●

Solution: From the previous example, we know that the likelihood of a single observation is

$$\mathcal{L}(\lambda;\, y) = \frac{e^{-\lambda}\ \lambda^{y}}{y!} \tag{9.25}$$

Thus, the likelihood of n independent observations is

$$\mathcal{L}(\lambda;\, y, n) = \prod_{i=1}^{n} \frac{e^{-\lambda}\ \lambda^{y_i}}{y_i!} \tag{9.26}$$

Since there is a product involved, it will be easier to maximize the logarithm of the likelihood,

$$l(\lambda;\, y, n) = \sum_{i=1}^{n} (-\lambda + y_i\ \log\lambda - \log y_i!) \tag{9.27}$$

And so, we maximize this function with respect to λ to obtain our estimator:

$$\frac{d}{d\lambda} l(\lambda;\, y, n) = \frac{d}{d\lambda} \sum_{i=1}^{n} (-\lambda + y_i\ \log\lambda - \log y_i!) \tag{9.28}$$

$$= \sum_{i=1}^{n} -1 + \sum_{i=1}^{n} \frac{y_i}{\lambda} \tag{9.29}$$

$$= -n + \frac{n\overline{y}}{\lambda} \tag{9.30}$$

Now, setting this equal to zero and solving for the estimator gives us

$$0 \overset{\text{set}}{=} -n + \frac{n\overline{y}}{\hat{\lambda}} \tag{9.31}$$

$$n = \frac{n\overline{y}}{\hat{\lambda}} \tag{9.32}$$

$$\tag{9.33}$$

Thus, with multiple measurement, the maximum likelihood estimator of λ is

$$\hat{\lambda} = \overline{y} \tag{9.34}$$

Before moving on, think about the result to ensure that it makes sense. This is always an important step! ♦

Warning: *At the end of every result, you should think about its consequences. Make sure the results make sense. If they do not, then double-check your work* or *see the world in a more subtle light.*

Another important distribution is the exponential distribution. It is used to model the time until some event occurs. Actuaries may use it to model (estimate) the time until a person dies or gets into an automobile accident or gets sued or some other wonderful event.

It has a single parameter, λ, which is the rate.[2] This means that the average will be $1/\lambda$. Double-check that this actually makes sense.

The following examples deals with this distribution.

[2] If you are having *déjà vu* again, do not worry. There is an intimate connection between the Poisson and exponential distributions. If the time between arrivals follows an exponential distribution, then the number of arrivals follows a Poisson distribution.

Example 9.5: His Majesty has some additional work for us. He would like to estimate the average lifetime of Ruritanians.

Let us use maximum likelihood estimation to provide an estimator for λ, the average rate of a person dying (NOT the average time until death). ●

Solution: The probability density function for the exponential distribution, when parameterized on its rate, is

$$f(x;\ \lambda) = \lambda\ e^{-\lambda x} \tag{9.35}$$

Thus, the likelihood function for a single observation is

$$\mathcal{L}(\lambda;\ x) = \lambda\ e^{-\lambda x} \tag{9.36}$$

And, the likelihood function for n independent observations is

$$\mathcal{L}(\lambda;\ x, n) = \prod_{i=1}^{n}\ \lambda\ e^{-\lambda x_i} \tag{9.37}$$

As this is a product, the log-likelihood will be easier to differentiate. It is

$$l(\lambda;\ x, n) = \sum_{i=1}^{n}\ (\log\lambda - \lambda x_i) \tag{9.38}$$

Now, we maximize it.

$$\begin{aligned}
\frac{d}{d\lambda} l(\lambda;\ x, n) &= \frac{d}{d\lambda} \sum_{i=1}^{n}\ (\log\lambda - \lambda x_i) && (9.39)\\
&= \sum_{i=1}^{n}\ \frac{1}{\lambda} - \sum_{i=1}^{n}\ x_i && (9.40)\\
&= \frac{n}{\lambda} - n\overline{x} && (9.41)\\
0 &\overset{\text{set}}{=} \frac{n}{\hat{\lambda}} - n\overline{x} && (9.42)\\
0 &= \frac{1}{\hat{\lambda}} - \overline{x} && (9.43)\\
\hat{\lambda} &= \frac{1}{\overline{x}} && (9.44)
\end{aligned}$$

♦

> ***Note***: From this, it can be shown that the maximum likelihood estimator of the mean of an exponential distribution is
>
> $$\hat{\mu} = \overline{x} \tag{9.45}$$
>
> All it takes is knowing that the expected value of an exponential distribution is $\frac{1}{\lambda}$.

Since the original question dealt with the average age, we would want to calculate $\hat{\mu}$, not $\hat{\lambda}$. I leave it as an exercise to show that the maximum likelihood estimator of μ for the following parameterization of the exponential distribution

$$f(x;\mu) = \frac{1}{\mu}\, e^{-x/\mu} \tag{9.46}$$

is $\hat{\mu} = \overline{x}$.

> ***Note***: It should be noted that the maximum likelihood estimator is awesome in that functions "pass through." In other words, it can be shown that
>
> $$\widehat{f(x)}_{\text{MLE}} = f(\widehat{x}_{\text{MLE}}) \tag{9.47}$$
>
> In words, the maximum likelihood estimator of a function of a parameter is that function of the maximum likelihood estimator of the parameter.

This is as good a time as any. There are two "drawbacks" to using maximum likelihood to estimate parameters. The first is that there is no guarantee that the estimator is unique. The second is that there is no guarantee that the estimator is unbiased.

While these seem bad, there is a nifty theorem that states the MLE is asymptotically unbiased; that is, as the sample size increases, its bias goes to zero.

9.2: The MLE and the CLM

Recall that the classical linear model assumes

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}, \tag{9.48}$$

with $\mathbf{E} \sim \mathcal{N}\left(\mathbf{0};\ \sigma^2\mathbf{I}\right)$. When we fit this model using ordinary least squares (OLS), we obtained the following estimators:

$$b_0 = \overline{y} - b_1\overline{x} \tag{9.49}$$

$$b_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \tag{9.50}$$

Let us see what we get when we fit this model using maximum likelihood methods.

Theorem 9.1. *The maximum likelihood estimator of β_0 is*

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \tag{9.51}$$

This is equivalent to the OLS estimator of the y-intercept.

Proof. The first step is to determine the likelihood function. The second step is to maximize that likelihood with respect to the parameter. As is usual, one maximizes the logarithm of the likelihood instead of the likelihood itself. It is generally easier.

Remember the conditional distribution of y. With that in mind, here is the likelihood for *one* observation:

$$\mathcal{L}(\mu, \sigma^2;\ x, y) = \frac{1}{\sqrt{2\pi\sigma^2}}\ \exp\left[-\frac{1}{2}\ \frac{(y-\mu)^2}{\sigma^2}\right] \tag{9.52}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}}\ \exp\left[-\frac{1}{2}\ \frac{(y-\hat{y})^2}{\sigma^2}\right] \tag{9.53}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}}\ \exp\left[-\frac{1}{2}\ \frac{\left(y-(\beta_0+\beta_1 x)\right)^2}{\sigma^2}\right] \tag{9.54}$$

Jeeee-willikers! That is just the probability density function for the normal distribution, where μ (as always) represents an expected value.

That was for a single observation. However, we rarely deal with just one data point. We deal with n of them. We remember from our introductory statistics course that **if** the data are independent, then $\mathbb{P}\left[A \cap B\right] = \mathbb{P}\left[A\right]\mathbb{P}\left[B\right]$. That means that the likelihood of observing all of our data is just the product of the individual likelihoods.

With that, we have

$$\mathcal{L}\left(\beta_0, \beta_1, \sigma^2;\ \mathbf{x}, \mathbf{Y}\right) = \prod_{i=1}^{n}\ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2}{\sigma^2}\right] \tag{9.55}$$

Since you may not have seen the notation before, $\prod$ is the product symbol just like $\sum$ is the summation symbol.

The next step is to maximize this likelihood. From calculus, we recall the product formula for derivatives. Just try applying the product formula here. You will shortly go bald from pulling out your hair. There is no easy way to maximize this likelihood function. *¡Què làstima!*

one-to-one and onto

However, if we apply a bijection to this likelihood, then maximizing *that* function is equivalent to maximizing the original likelihood... equivalent in terms of the value that produces the maximum.

Because the likelihood has a lot of products, and because it is easier to maximize a sum, we use the logarithm function. The log-likelihood function of the above function is just

$$l\left(\beta_0, \beta_1, \sigma^2;\ \mathbf{x}, \mathbf{Y}\right) = \sum_{i=1}^{n} \left(-\frac{1}{2}\log\left(-2\pi\sigma^2\right) - \frac{1}{2}\frac{\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2}{\sigma^2}\right) \tag{9.56}$$

Taking the derivative of a summation is so much easier than taking the derivative of a product... so much easier!

And, now that we have a practically differentiable function, we use calculus to maximize it with respect to β_0:

$$\begin{aligned}
\frac{\partial}{\partial\beta_0} l(\beta_0, \beta_1, \sigma^2;\ \mathbf{x}, \mathbf{Y}) &= \frac{\partial}{\partial\beta_0}\sum_{i=1}^{n} \left(-\frac{1}{2}\log\left(-2\pi\sigma^2\right) - \frac{1}{2}\frac{\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2}{\sigma^2}\right) \\
&= \sum_{i=1}^{n} -\frac{1}{2}\ \frac{2\left(y_i - (\beta_0 + \beta_1 x_i)\right)(-1)}{\sigma^2} && (9.57) \\
&= \sum_{i=1}^{n} \frac{y_i - \beta_0 - \beta_1 x_i}{\sigma^2} && (9.58)
\end{aligned}$$

Now, set this to zero and solve for $\hat{\beta}_0$:

$$0 \overset{\text{set}}{=} \sum_{i=1}^{n} \frac{y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i}{\sigma^2} \tag{9.59}$$

$$0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i \tag{9.60}$$

$$\sum_{i=1}^{n} \hat{\beta}_0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i \tag{9.61}$$

$$n\hat{\beta}_0 = n\overline{y} - n\hat{\beta}_1 \overline{x} \tag{9.62}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{9.63}$$

Thus, we have shown that $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$, as we desired. Note that this is also the OLS estimator of the y-intercept. Very interesting! □

Theorem 9.2. *The maximum likelihood estimator of β_1 is*

$$\hat{\beta}_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} \tag{9.64}$$

Proof. From our proof of the estimator of β_0, we have the following as our log-likelihood function:

$$l\Big(\beta_0, \beta_1, \sigma^2;\, \mathbf{x}, \mathbf{Y}\Big) = \sum_{i=1}^{n} \left(-\frac{1}{2}\log\Big(-2\pi\sigma^2\Big) - \frac{1}{2}\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{\sigma^2} \right) \tag{9.65}$$

And so, the proof proceeds by taking the derivative with respect to β_1 and solving for $\hat{\beta}_1$.

$$\begin{aligned} \frac{\partial}{\partial \beta_1} l(\beta_0, \beta_1, \sigma^2;\, \mathbf{x}, \mathbf{Y}) &= \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} \left(-\frac{1}{2}\log\Big(-2\pi\sigma^2\Big) - \frac{1}{2}\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{\sigma^2} \right) \\ &= -\frac{1}{2} \sum_{i=1}^{n} \frac{2\,(y_i - (\beta_0 + \beta_1 x_i))\,(-x_i)}{\sigma^2} && (9.66) \\ &= \sum_{i=1}^{n} \frac{x_i y_i - \beta_0 x_i - \beta_1 x_i^2}{\sigma^2} && (9.67) \end{aligned}$$

Setting this to zero and solving for the estimator, $\hat{\beta}_1$ gives

$$\begin{aligned}
0 &\overset{\text{set}}{=} \sum_{i=1}^{n} \frac{x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2}{\sigma^2} && (9.68)\\
&= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \hat{\beta}_0 x_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 && (9.69)\\
&= \sum_{i=1}^{n} x_i y_i - n\overline{x}\hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 && (9.70)\\
&= \sum_{i=1}^{n} x_i y_i - n\overline{x}\left(\overline{y} - \hat{\beta}_1 \overline{x}\right) - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 && (9.71)\\
&= \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} + n\hat{\beta}_1 \overline{x}^2 - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 && (9.72)
\end{aligned}$$

Moving the $\hat{\beta}_1$ terms to the left side gives

$$\begin{aligned}
\sum_{i=1}^{n} \hat{\beta}_1 x_i^2 - n\hat{\beta}_1 \overline{x}^2 &= \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} && (9.73)\\
\hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \right) &= \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} && (9.74)
\end{aligned}$$

And finally,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \qquad (9.75)$$

surprise!

We have seen this before. It is the OLS estimator of the slope parameter. No surprise.

To finish the proof, use algebra to show that the final equation above is equivalent to $\hat{\beta}_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$. □

Theorem 9.3. *The maximum likelihood estimator of σ^2 is*

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \tag{9.76}$$

Proof. It has been a while, so I will leave this as an exercise for you to prove this. I have already shown the log-likelihood function. All you have to do is differentiate with respect to σ^2, solve for $\hat{\sigma}^2$, and use algebra to move things into the right form. □

exercise

Note that the above formula (Eqn 9.76) is equivalent to

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2 \tag{9.77}$$

This should raise a red (maybe only yellow or a nice chartreuse) flag, as this is a biased estimator of σ^2.

Why?

9.2.1 Consequences

I leave it as an exercise to prove the following consequences:

1. $\hat{\beta}_0$ is unbiased for β_0.
2. $\hat{\beta}_1$ is unbiased for β_1.
3. $\hat{\sigma}^2$ is biased for σ^2.

Because the maximum likelihood estimators are identical to the ordinary least square estimators, and because we have not altered the Normality assumption of the classical linear model, everything from Chapters 2 and 3 hold.

Well, that is not entirely true. Remember that the MLE estimator of σ^2 is not the same as the OLS estimator. Thus, the test statistic and confidence interval will need to be altered a bit. However, the differences are minor for large samples.[3]

[3] And this is the problem that William Sealy Gosset had to deal with (see Section S.4.6). Things easily work for large samples. He had to deal with small samples in his work.

9.2.2 Multivariate Distributions*

There is one prerequisite to this textbook that would make things a little easier: an introduction to multivariate distribution. Thus far, I have "hand-waved" over the topic. Here, I will *briefly* discuss the topic.

The following is a univariate distribution:

$$f(x;\ \mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\ \exp\left[-\frac{1}{2}\ \frac{(x-\mu)^2}{\sigma^2}\right] \tag{9.78}$$

This is the (in)famous probability density function for the normal distribution. It is a function of just one variable (value), x. This is what makes it univariate. The prefix *uni* stands for neither the University of Northern Iowa nor edible sea urchin gonads. It is a Latin combining form for "one." Thus, "univariate" indicates "one variable."

Ewww!

The following is one example of a *bi*variate distribution:

$$f(x,y) = \frac{\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \tag{9.79}$$

This is a distribution where X and Y are distributed jointly normal, and where they have correlation ρ between them. There are a lot of symbols there because it is written in scalar form. Were we to write it in matrix form, we could generalize all of these "—variate" distributions into one form.

MVN

If the random vector $\mathbf{Y}$ follows a multivariate normal distribution such that $\mathbf{Y} \sim \mathcal{N}_n(\mu;\ \mathbf{\Sigma})$, then

$$f\,(\mathbf{Y};\ \mu,\mathbf{\Sigma}) = (2\pi)^{-n/2}\ \left|\ \mathbf{\Sigma}\ \right|^{-1/2}\ \exp\left[-\frac{1}{2}\ (\mathbf{x}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right] \tag{9.80}$$

Here, n is the number of *variables* that are jointly normal. This means that each random variable follows a normal distribution, given the values of the others. The vector μ is a column vector of expected values for each X_i. Finally, $\mathbf{\Sigma}$ is the correlation matrix between the n random variables. If the x values are independent, then $\mathbf{\Sigma} \in \mathcal{D}_n$ (diagonal). If the x values are independent and identically distributed, then $\mathbf{\Sigma} = \sigma^2\mathbf{I}_n$.

If $n = 1$, then the multivariate normal reduces to the univariate normal. If $n = 2$, then it reduces to the bivariate normal, where the off-diagonal entries in $\mathbf{\Sigma}$ are equal to $\rho\sigma_1\sigma_2$ and the diagonal entries are σ_1^2 and σ_2^2.

That is, if $n = 2$, then

$$\not\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{9.81}$$

Still, not much in this subsection is important — if the observations are independent. If the observations are independent, then the n-variate normal is just the product of the n univariate normals. The same is true for any distribution.

If the observations are *not* independent, then the joint distribution — the actual distribution we care about — is not so simple. In many cases, it has not been entirely formulated. For instance, what is the multivariate Binomial distribution? That is, what is the distribution of $\left\{y_1, y_2, y_3, \ldots, y_n\right\}$, given correlation amongst those n measurements? Even better: How could we measure such correlation?[4]

[4] In such cases, Dai, Ding, and Wahba (2013) may give you some insight into the difficulty of these questions — and their answers! This really makes you appreciate random sampling, where both independence and identical distribution hold.

9.3: Conclusion

From this chapter, we have discovered how to perform maximum likelihood estimation (MLE). The steps are as usual: formulate the objective function, use calculus to maximize it.

The maximum likelihood estimators for the two main parameters of the classical linear model are the same as the ordinary least squares estimators. Thus, they are both unbiased. The maximum likelihood estimator of the error variance, however, is biased. We know this because it does not equal the MSE, which is unbiased.

Thus, it appears as though maximum likelihood gives us nothing helpful. However, this is not true. First, there is a theorem (beyond the scope of this course) that proves *all* maximum likelihood estimators are asymptotically unbiased. In other words, if your estimator is a maximum likelihood estimator, you have nothing to prove with respect to asymptotic bias (Panchenko 2006, Thode et al. 2002). All other estimators (like OLS) require separate proofs. So, we gain there.

Second, ordinary least squares requires that the conditional distribution of the dependent variable is normal. Maximum likelihood does not have that as a requirement. This allows us to go beyond the classical linear model and the requirement of Normality. In fact, the next part of this class examines this feature of maximum likelihood estimation.

9.4: End-of-Chapter Materials

9.4.1 R Functions

This chapter had no R functions. It was all mathematics and concepts. Yay!!

9.4.2 Exercises

1. Prove that the maximum likelihood estimator of π is x/n in a Binomial experiment.

2. Prove Theorem 9.3.

3. Prove $\hat{\beta}_0$ is unbiased for β_0.

4. Prove $\hat{\beta}_1$ is unbiased for β_1.

5. Prove $\hat{\sigma}^2$ is biased for σ^2 and that the bias is exactly $\frac{n-1}{n}\sigma^2$.

9.4.3 Theory Readings

- Bin Dai, Shilin Ding, and Grace Wahba (2012). “Multivariate Bernoulli Distribution.” *Bernoulli* 19(**4**): 1465–1483. doi: 10.3150/12-BEJSP10

- Dmitry Panchenko (2006). “Lecture 3: Properties of MLE: consistency, asymptotic normality. Fisher information.” *Open Courseware/MIT*. URL = https://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/lecture3.pdf

- Aaron Thode, Michele Zanolin, Eran Naftali, Ian Ingram, Purnima Ratilal, and Nicholas C. Makris (2002). “Necessary conditions for a maximum likelihood estimate to become asymptotically unbiased and attain the Cramer–Rao lower bound. II. Range and depth localization of a sound source in an ocean waveguide.” *The Journal of the Acoustical Society of America*. 112(**1890**). doi: 10.1121/1.1496765