Chapter 7
Regression Models and Hypothesis Testing

Abstract This chapter covers the basics of adapting regression models, also known as linear fits in physics, to find the parameters that best explain data in a model and then estimate the error bars of the parameters. The analysis of the model’s reliability stimulates a discussion of $\chi^2$ and $t$-distributions and their role in testing hypotheses regarding the parameters; for example, whether a parameter can be omitted from the fit. A more elaborate method, based on the $F$-test, follows. The chapter closes with a discussion of parsimony as a guiding principle when building models.

Both in physical and social sciences as well as in economics we often try to describe the behavior of observable parameters—the measurements—in terms of model parameters. Frequently the dependence of the measurements on the parameters of a model—the fit parameters—is linear; the archetypical problem of this type is the fit of data points to a straight line. Let us therefore assume that we are given a list of $n$ data points $(s_i, y_i)$ for $i = 1, \ldots, n$ that come from varying a variable $s$ and observing another quantity $y$. Figure 7.1 shows two examples. Visual inspection indicates that the data points cluster around a straight line that we can describe by two parameters $a$ and $b$, given by $y_i = as_i + b$. Our task is thus to find the slope $a$ and the intercept $b$. This is most easily accomplished by writing one copy of the equation for each of the $n$ measurements, which defines one row in the following matrix-valued equation

$$
\begin{pmatrix}
\vdots \\
y_i \\
\vdots 
\end{pmatrix} = \begin{pmatrix}
\vdots \\
s_i \\
\vdots 
\end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}.
$$

(7.1)

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-63643-2_7) contains supplementary material, which is available to authorized users.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2021
V. Ziemann, Physics and Finance, Undergraduate Lecture Notes in Physics,
https://doi.org/10.1007/978-3-030-63643-2_7
Fig. 7.1 Two examples with data points scattered around a linear trend (asterisks, crosses) and dashed regression lines

Here the $n$ measurements $y_i$ are assembled into a column vector and the values of the independent variable $s_i$ are assembled, together with $n$ times a one, in a $n \times 2$-matrix. The unknown parameters $a$ and $b$ are assembled in a second column vector. Note that the equation has the form of $y = Ax$, where $y$ is the column vector with the $y_i$, the matrix $A$ with the $s_i$, and a column vector $x$ with $a$ and $b$. Our task is now to find values for $a$ and $b$ such that the line approximates the data points as good as possible (in some sense, defined below). Moreover, we want to assess the error bars on the fit parameters $a$ and $b$. The larger scatter of the data points on the left-hand plot in Fig. 7.1 will likely reduce our confidence in the fitted value of $a$ and $b$, compared to the values derived from the data points on the right-hand plot. The scatter of the data points $y_i$ is commonly described by adding a vector $\Delta y$ to $y = Ax + \Delta y$ and the assumption that the components $\Delta y_i$ of $\Delta y$ are sampled from Gaussian distributions with standard deviation $\sigma_i$, the error bars of the measurements $y_i$. Furthermore, increasing the number of fit parameters, for example, by fitting a third-order polynomial through the data points, will likely make the fit better. One might ask, however, whether a small improvement is really worth having to deal with an additional fit parameter? This fit parameter may even have large error bars associated. We will address questions, such as this one, throughout this chapter.

Before considering a number of examples we need to elaborate how to actually find the model parameters such as the slope or intercept in a straight-line fit.

### 7.1 Regression and Linear Fitting

In (7.1) we only try to determine two fit parameters, but we can easily generalize the method to include many, say $m$, fit parameters $x_j$ with $j = 1, \ldots, m$. Since we assume the dependence of the measurements $y_i$ on the model parameters $x_j$ to be linear, the corresponding equation is matrix valued and in general will have the
following form

\[
\begin{pmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{pmatrix}
= \begin{pmatrix}
A_{11} & A_{12} & \ldots & A_{1m} \\
A_{21} & A_{22} & \ldots & A_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
A_{n1} & A_{n2} & \ldots & A_{nn}
\end{pmatrix}\begin{pmatrix}
x_1 \\
x_2 \\
\vdots \\
x_m
\end{pmatrix} + \begin{pmatrix}
\Delta y_1 \\
\Delta y_2 \\
\vdots \\
\Delta y_n
\end{pmatrix},
\]

(7.2)

which can be also written as \( y = Ax + \Delta y \). The information about the dynamics of the system, which relates the \( n \) measurements \( y_i \) to the \( m \) fit parameters \( x_j \), is encoded in the known coefficients of the matrix \( A_{ij} \) with \( i = 1, \ldots, n \) and \( j = 1, \ldots, m \). The measurement errors are described by \( \Delta y_i \) with rms values \( \sigma_i \).

Systems of equations, defined by (7.2), are over-determined linear systems if \( n \), the number of observations or measurements, is larger than the number of parameters \( m \). Such systems can be easily solved by standard linear-algebra methods in the least squares sense. For this we minimize the so-called \( \chi^2 \), which is the squared difference between the observations and the model

\[
\chi^2 = \sum_{i=1}^{n} \left( \frac{y_i - \sum_{j=1}^{m} A_{ij} x_j}{\sigma_i} \right)^2 .
\]

(7.3)

Here we introduce the error bar or standard deviation \( \sigma_i \) of the observation \( y_i \) to assign a weight to each measurement. If an error bar of a measurement is large, that measurement contributes less to the \( \chi^2 \) and will play a less important role in the determination of the fit parameters \( x_i \). If all error bars are equal \( \sigma_i = \sigma \) for all \( i \) the system is called homoscedastic and if they differ, it is called heteroskedastic.\(^1\)

If we consider a homoscedastic system and ignore the \( \sigma_i \) for the moment, we see that (7.3) can be written as a matrix equation \( \chi^2 = (y - x' A')(y - A x) \), where \( x \) and \( y \) are column vectors, albeit of different dimensionalities \( n \) and \( m \). Minimizing the \( \chi^2 \) with respect to the fit parameters \( x \), we find \( 0 = \frac{\partial \chi^2}{\partial x} = -2 A'(y - A x) \), where \( \partial \)/\( \partial x \) denotes the gradient with respect to the components of \( x \). Solving for \( x \) we obtain

\[
x = (A' A)^{-1} A' y ,
\]

(7.4)

where \( (A' A)^{-1} A' \) is the well-known pseudo-inverse of the matrix \( A \).

For heteroscedastic systems, we introduce the diagonal matrix \( \Lambda \) that contains the inverse of the error bars on its diagonal \( \Lambda_{ii} = 1/\sigma_i \), which transforms (7.3) into \( \chi^2 = (\Lambda y' - x' A' \Lambda') (\Lambda y - A A x) \) such that both \( y \) and \( A \) are prepended with \( \Lambda \). Performing these substitutions in (7.4), we obtain

\[
x = (A' \Lambda^2 A)^{-1} A' \Lambda^2 y = J y .
\]

(7.5)

\(^1\)Greek: homo = equal, hetero = unequal, skedasis = dispersion or spread.
Here we introduce the matrix $J$ that maps $y$ onto $x$ to explicitly show that the fit parameters $x$ depend linearly on the measurements $y$. We denote the matrix by the symbol $J$ to remind us that it has the same function as a Jacobian in a coordinate transformation. In our case the transformation is linear, and consequently the Jacobian has constant matrix elements.

The error bars and the covariance matrix of the fit parameters $x_j$ can be calculated by standard error propagation techniques from the covariance matrix of the measurements $y_i$, which we is given by $C_{ij}(y) = \langle \Delta y_i \Delta y_j \rangle$. The analysis is based on the realization that the covariance matrix $C_{ij}(y)$ is defined through the second moments of the deviations $\Delta y_i$. On its diagonal $C_{ij}(y)$ contains the squared error bars of the individual measurements $\sigma_i^2$ and the off-diagonal elements carry information about correlations among the measurements. If, on the other hand, the measurements are uncorrelated, as we assume them to be, all off-diagonal elements are zero. In our example $C_{ij}(y)$ is simply the square of the inverse of the matrix $\Lambda$, namely $C(y) = (\Lambda^2)^{-1}$, which has $\sigma_1^2, \ldots, \sigma_n^2$ on its diagonal. From (7.5) follows that small deviations of the measurements $\Delta y_i$ lead to small deviations of the model parameters $\Delta x = J \Delta y$. This observation allows us to calculate the covariance matrix $C_{kl}(x) = \langle \Delta x_k \Delta x_l \rangle$. A little algebra then leads to

$$C(x) = J C(y) J^T,$$  

(7.6)

which describes how covariance matrices transfrom under a change of variables given by (7.5). Furthermore, inserting the definition of $J$ from (7.5) in the previous equation, we find after some additional algebra

$$C(x) = (A^T \Lambda^2 A)^{-1},$$  

(7.7)

which allows us to find the covariance matrix $C(x)$ of the fit parameters $x_j$ by simple matrix operations from the error bars of the initial measurements that are buried in $\Lambda$ and the system matrix $A$ that initially defined the problem in (7.2).

In order to illustrate the methodology, let us consider a few examples of regression models taken from several disciplines.

### 7.2 Examples

Here we illustrate how we can use regression methods to extract information from data provided in tabular form, either to obtain quantitative numbers or to test some model.

Our first example addresses the question whether education actually pays off. Do extra years of education increase the earnings and if so, at which rate. Moreover, can we find a quantitative description of how well education protects from unemployment? We base our analysis on information for 2018 from the US Bureau of
7.2 Examples

Fig. 7.2 The median weekly earnings (left) and the unemployment rate (right) as a function educational attainment, here specified as the years of education past the age of 15. The data are based on [1] from the US bureau of labor statistics

Labor Statistics [1], which provides data on the median weekly earnings \( w \) and the unemployment rate \( u \) based on the highest level of educational attainment. Here we characterize the attainment by the additional years \( t \) of education past the age of fifteen; we assume a high-school diploma completed after two years, a bachelor after six, a master after eight, a PhD after twelve, and a professional degree after fourteen years. The left plot in Fig. 7.2 shows the weekly wage \( w \) as a function of \( t \) as black asterisks and a straight-line fit of \( w = at + b \) as a dashed red line. The slope \( a \) of the line is \( a = 106 \) $/year of education, such that each additional year, on average, increases the weekly earnings by 106 $. So, financially, education pays off!

The right plot in Fig. 7.2 shows the unemployment rate \( u \) as a function of \( t \). The data points do not follow a straight line, but plotting \( \log(u) \) versus \( \log(t) \) reveals a linear dependence. Fitting a straight line to \( \log(u) = a \log(t) + b \) provides fit parameters \( a \approx -0.5 \) and \( b \approx 1.76 \). The dashed red line on the right plot is based on these parameters. We thus find that the unemployment rate \( u \) scales as \( 1/\sqrt{t} \). Education even protects against unemployment, at least to some extent.

The second example uses data about the number of infected persons and fatalities from the corona epidemic, provided by the Johns-Hopkins University [2]. We ask ourselves whether the reported number of infected people \( I \) and reported number of deaths \( D \) are consistent. In the basic SIR model [3] of epidemics the rate of recovered and subsequently immune patients \( dR/dt \) is proportional to \( I \). Analogously, we assume here that the rate of change of fatalities \( dD/dt \) is also proportional to the number infected \( I \), or \( dD/dt = \alpha I \) with a proportionality constant \( \alpha \) that may differ from country to country, depending on the health system and the accounting of the infected and the corona-related fatalities. Note that the reported numbers of infected persons in the data from [2] do not reflect those having recovered from the infection. We therefore expect the model to work only during the first few weeks of the epidemic, while we can neglect the recovery rate. With these considerations in mind we process the data by integrating the model equation once to obtain \( D = \alpha \int I dt \). The data files from [2] contain day-by-day values of \( D \) and \( I \). Therefore we store the values
in arrays and access the number of fatalities on day $i$ by $D_i$ and likewise the number of infected by $I_i$. The integral $K_i$ of $I_i$ is then given by $K_i = \sum_{j \leq i} I_j$ and we then try to determine $\alpha$ from $D_i = K_i \alpha$, which is of the type specified by (7.2) with $K_i$ corresponding to a single-column matrix $A$ and $I_i$ to $y_i$. We then find $\alpha$, the model parameter, from (7.4).

Figure 7.3 shows the number of reported deaths from March 1 until April 16, 2020 in three countries. The solid red line shows the integral of the infected persons $K_i$, scaled with the country-specific factor $\alpha$, which was determined by fitting the range up to 14 days before the end. We observe that the regression line for Germany underestimates the fatalities in the last 14 days, which could be explained by many unaccounted infected persons. The fit and the data for the US show good agreement between model and data, indicating that the assumptions in the model are approximately valid. The data for Italy show less actual fatalities than the model predicts. One could hypothesise that this might be attributed to a significant number of recovered persons such that the underlying assumptions of our simple model are not valid towards the end of the range. As a disclaimer, we need to stress that this example is intended to illustrate regression methods, not to develop policies of how to deal with an epidemic.

Our third example is motivated by the fact that systems in equilibrium, if slightly perturbed, perform harmonic oscillations. Thus, the dynamics is governed by $\ddot{x} + \omega^2 x = 0$ with amplitude $x$ and frequency $\omega$. We now want to determine the frequency $\omega$ and the initial conditions $x_0$ and $\dot{x}_0$ at time $t = 0$ from a sequence of measurements of the amplitude $x_n$ at times $t = n \Delta t$. This can be visualized as a stroboscopic recording of the oscillation at discrete points in time. To do so, we first realize that $x = x_0 \cos(\omega t) + (\dot{x}_0 / \omega) \sin(\omega t)$ solves the differential equation and satisfies the initial conditions. Likewise, we find that the position $x_{n+1}$ after $(n + 1) \Delta t$ is related to the position $x_n$ after $n \Delta t$ by $x_{n+1} = x_n \cos(\omega \Delta t) + (\dot{x}_n / \omega) \sin(\omega \Delta t)$ and we find the velocity $\dot{x}_{n+1} = -\omega x_n \sin(\omega \Delta t) + \dot{x}_n \cos(\omega \Delta t)$ by differentiation. Both equations can we written as a matrix-valued equation.
\[
\begin{pmatrix}
  x_{n+1} \\
  \dot{x}_{n+1}
\end{pmatrix} =
\begin{pmatrix}
  \cos(\omega \Delta t) & (1/\omega) \sin(\omega \Delta t) \\
  -\omega \sin(\omega \Delta t) & \cos(\omega \Delta t)
\end{pmatrix}
\begin{pmatrix}
  x_n \\
  \dot{x}_n
\end{pmatrix}.
\] (7.8)

Inverting this equation, we find
\[
(1/\omega) \sin(\omega \Delta t) \dot{x}_{n+1} = x_{n+1} \cos(\omega \Delta t) - x_n.
\]
Moreover, using (7.8) shifted by one time step, results in
\[
x_{n+2} = x_{n+1} \cos(\omega \Delta t) + (\dot{x}_{n+1}/\omega) \sin(\omega \Delta t).
\]
Inserting the former into the latter equation yields
\[
x_{n+2} + x_n = 2 \cos(\omega \Delta t)x_{n+1}.
\] (7.9)

This is a linear equation in the positions \(x_n\), which allows us to determine \(\cos(\omega \Delta t)\) and thereby \(\omega\) by casting the equation in the form of (7.2) with \(x_{n+2} + x_n\) populating the vector \(y\) on the left-hand side and filling the single-column matrix \(A\) with \(2x_{n+1}\). Here \(\cos(\omega \Delta t)\) corresponds to a single fit parameter, called \(x\) in (7.2). As opposed to Fourier methods, which require many consecutive samples to achieve a high frequency resolution, here we only need a minimum of three positions \(x_n\) from consecutive periods in order to determine \(\cos(\omega \Delta t)\). Using more data points makes the problem over-determined and improves the accuracy.

Now that we have determined the frequency \(\omega\), we proceed to determine the initial conditions \(x_0\) and \(\dot{x}_0\) from subsequent readings of \(x_n = x_0 \cos(n \omega \Delta t) + (\dot{x}_0/\omega) \sin(n \omega \Delta t)\). Furthermore, we assume that each measurement \(x_n\) is known to be within its error bars \(\pm \sigma\). This allows us to write the problem as a matrix-valued equation
\[
\begin{pmatrix}
  \vdots \\
  x_n \\
  \vdots
\end{pmatrix} =
\begin{pmatrix}
  \vdots \\
  \cos(n \omega \Delta t)/\sigma & (1/\omega) \sin(n \omega \Delta t)/\sigma \\
  \vdots & \vdots
\end{pmatrix}
\begin{pmatrix}
  x_0 \\
  \dot{x}_0
\end{pmatrix},
\] (7.10)

where we explicitly absorbed the matrix \(\Lambda\) with the error bars into (7.10), which is of the type of (7.2) and can be solved by the methods discussed in Sect. 7.1 yielding the initial conditions \(x_0\) and \(\dot{x}_0\) of the harmonic oscillation. The error bars of the fit parameters can be determined with (7.7).

Going beyond fitting for one or two parameters we consider a fit to a polynomial of order \(p\) to \(n\) data points \((t_i, y_i)\) for \(i = 1, \ldots, n\)
\[
y_i = a_pt_i^p + \cdots + a_1 t_i + a_0,
\] (7.11)

which is straightforward to convert to an equation similar to (7.2). Here the matrix \(A\) is of size \(n \times (p + 1)\) and the vector \(x\) contains the \(p + 1\) coefficients \(a_p\).

In Chap. 8, where we will analyze time series, we will use regression methods, similar to those discussed in this chapter to determine models for the temporal behavior of dynamics systems. In Sect. 11.1 we will encounter macroeconomic models that are described through difference equations, relating economical quantities, such as company output, profit, and investment rate, from one time period to the next. The validity of these models can be corroborated or rejected by analyzing the consistency of the models when comparing them with published data for companies.
The above manipulations allow us to calculate the fit parameters and their error bars, but we still do not know how reliable the calculations are, because we could have a small $\chi^2$ and just had misjudged the initial measurement errors $\sigma_i$. Therefore we will first introduce a parameter $R^2$ that characterizes how well the fitted model actually explains the observed data $y_i$ and then calculate the distribution of $\chi^2$ we can expect and can judge how likely the obtained $\chi^2$ actually is.

### 7.3 Goodness-of-Fit $R^2$

The goodness-of-fit parameter $R^2 = \text{SSE}/\text{SST}$ compares the spread of the measurement values $y_i$, the total sum of squares (SST), around their mean $\bar{y}$

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{with} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \quad (7.12)$$

to the explained sum of squares (SSE), defined as the spread of the values $\hat{y}_i$ predicted by the fitted model, to the mean $\bar{y}$

$$\text{SSE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \text{with} \quad \hat{y}_i = \sum_{j} A_{ij} x_j \quad (7.13)$$

Here $\hat{y}_i$ are the values estimated with the model $A_{ij}$ and the fit parameters $x_j$. $R^2$ thus characterizes how well the fitted model can explain the variation around the mean $\bar{y}$.

Let us now inspect how SST and SSE are related to the $\chi^2$ that we introduced in (7.3). We therefore insert and subtract the $\hat{y}_i$ in the definition of SST

$$\begin{align*}
\text{SST} &= \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\
&= \text{SSR} + \text{SSE} + 0 \quad (7.14)
\end{align*}$$

where the first term is the sum of squared differences of the measured values and the model predictions $y_i - \hat{y}_i = y_i - \sum_{j} A_{ij} x_j$ and is called the sum of squared residuals (SSR). We recognize it as $\sigma^2 \chi^2$ from (7.3) if all the error bars are equal and have magnitude $\sigma$. The second term we recognize as the SSE and the last term is zero, which can be shown by inserting $\hat{y}_i = \sum_{j} A_{ij} x_j$ and (7.4).
Using (7.14) to write \( \text{SSE} = \text{SST} - \text{SSR} \), the Goodness-of-fit \( R^2 \) can be written using either SSE or SSR

\[
R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}. \tag{7.15}
\]

We see that the smaller the SSR, or equivalently the \( \chi^2 \) are, the closer \( R^2 \) approaches unity, where \( R^2 = 1 \) describes a perfect fit of the model to the data.

Calculating \( R^2 \) is rather straightforward; first calculate the sample variance SST from the average and variance of the measurement or sample values. Then perform the fit procedure and determine the sum of squared residuals between the measurement samples and the corresponding fitted values SSR. Finally calculate \( R^2 \) from (7.15).

We already noted that the SSR is closely related to the \( \chi^2 \) of the fit. In the following section we derive the probability distribution function of the \( \chi^2 \) values that we can expect. It will help us to assess our confidence in the fitted parameters.

### 7.4 \( \chi^2 \)-Distribution

The quantity we minimize to find a regression model is the \( \chi^2 \), which is given by

\[
\chi^2 = \sum_{i=1}^{n} \left( \frac{y_i - f_i(x)}{\sigma_i} \right)^2 = \sum_{i=1}^{n} z_i^2, \tag{7.16}
\]

where \( y_i \) are the measurement values with error bars \( \sigma_i \) and \( f_i(x) \) is a model function with fit parameters \( x_j \). In Sect. 7.1 we used a linear dependence \( f_i(x) = \sum_j A_{ij} x_j \). If we have estimated the error bars \( \sigma_i \) correctly, all individual factors \( z_i = (y_i - f_i(x))/\sigma_i \) in the sum should be of order unity and normally (Gaussian) distributed. Thus it is natural to ask what the distribution function of \( n \) squares of normally distributed random variables is, namely the probability distribution function \( \psi_n(q) \) such that we have the probability \( \psi_n(q = \chi^2) dq \) of finding a value of \( q = \chi^2 \) within the interval \([q - dq/2, q + dq/2]\).

We can calculate this distribution by assuming that the individual constituents \( z_i \) of the sum above are normally distributed random variables and we need to find the distribution of the \( q = \sum_{i=1}^{n} z_i^2 \). We start by writing the product of \( n \) independent and normalized Gaussian distribution functions for the \( z_i \)

\[
1 = \frac{1}{(2\pi)^{n/2}} \int_{-\infty}^{\infty} dz_1 \int_{-\infty}^{\infty} dz_2 \cdots \int_{-\infty}^{\infty} dz_n e^{-(z_1^2 + z_2^2 + \cdots + z_n^2)/2}
= \frac{1}{(2\pi)^{n/2}} \int_{0}^{\infty} dr r^{n-1} \int d\Omega \ e^{-r^2/2}, \tag{7.17}
\]
where we introduce spherical coordinates with \( r^2 = z_1^2 + z_2^2 + \cdots + z_n^2 \) in the second equation. Since the integrand does not depend on angular variables we know that \( \int d\Omega = S_n \) is the surface area of an \( n \)-dimensional sphere. After substituting \( q = r^2 \) we arrive at
\[
1 = \frac{S_n}{2(2\pi)^{n/2}} \int_0^\infty q^{n/2-1} e^{-q/2} dq = \int_0^\infty \psi_n(q) dq ,
\]
which implicitly defines the probability distribution \( \psi_n(q) \) for the \( q = \chi^2 \).

The previous equation still depends on the unknown surface area \( S_n \). We can, however, determine \( S_n \) from the requirement that the distribution function is normalized. Substituting \( p = q/2 \) we find
\[
1 = \frac{S_n}{2(2\pi)^{n/2}} 2^{n/2-1} \int_0^\infty dp p^{n/2-1} e^{-p} = \frac{S_n}{2(2\pi)^{n/2}} 2^{n/2-1} \Gamma(n/2)
\]
where \( \Gamma(z) \) is the Gamma function \([4]\), defined as
\[
\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt .
\]
Solving for \( S_n \), we find
\[
S_n = \frac{2\pi^{n/2}}{\Gamma(n/2)}
\]
which we insert into (7.18) to obtain the following expression for the probability distribution function
\[
\psi_n(q) dq = \frac{1}{2^{n/2} \Gamma(n/2)} q^{n/2-1} e^{-q/2} dq .
\]

Figure 7.4 displays \( \psi_n(q) \) for \( n = 2, 5, \) and \( 10 \).

The probability of finding a \( \chi^2 \) smaller than a given limit \( \xi \) is given by the integral from zero to \( \xi \) of the probability distribution function
\[
Q_n(\xi) = \int_0^\xi \psi_n(q) dq = \int_0^\xi \frac{1}{2^{n/2} \Gamma(n/2)} q^{n/2-1} e^{-q/2} dq = P(n/2, \xi/2)
\]
where \( P(a, x) \) is the incomplete gamma function \([4]\). Now we are in a position to answer the question whether a \( \chi^2 \) that arises from a fitting procedure or regression analysis is actually probable. In particular, the probability of finding a value of \( \chi^2 \)
smaller than  \( \hat{\chi}^2 \) is given by  \( P(n/2, \hat{\chi}^2/2) \). Conversely, finding a value that is even larger than  \( \bar{\chi}^2 \), is given by  \( 1 - P(n/2, \bar{\chi}^2/2) \).

The  \( \chi^2 \)-distribution plays an important role in testing hypothesis, but also in assessing the reliability of estimates from a small number of samples. The latter is the topic of the next section.

### 7.5 Student’s  \( t \)-Distribution

Assume that you are responsible for the quality control of the base materials for your favorite beer. For example, you need to work out how many bags of barley you have to test in order to assess the average quality of the entire delivery, which could be quite large, say, a hundred bags. This was the task that William Gosset—“Student” was his pen name—faced during the early years of the 20th century. He worked for the Guinness brewery in Dublin and had to ensure the quality of the delivered barley. We follow his lead and work out how well we can estimate the true mean  \( \mu \) and variance  \( \sigma^2 \) of a Gaussian distribution from a small number of samples  \( n \).

Since all we have are the  \( n \) samples  \( x_i \), we start by calculating their average  \( \bar{X}_n \) and the variance  \( S_n^2 \). These quantities are given by

\[
\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X}_n)^2 ,
\]

(7.24)

where the  \( x_i \) are the samples. In the following analysis we assume the samples  \( x_i \) to be drawn from a Gaussian distribution  \( N(x; \mu, \sigma) \), defined by
\[
N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi \sigma}} \exp\left[ -\frac{(x - \mu)^2}{2\sigma^2} \right].
\]

(7.25)

Note that \( \bar{X}_n \) and \( S_n^2 \) are commonly called the **sample average** and the **unbiased sample variance**, respectively.\(^2\) We first need to understand how fast the estimates \( \bar{X}_n \) and \( S_n^2 \) converge towards the true values \( \mu \) and \( \sigma^2 \) as we increase \( n \). We therefore calculate the expectation value of the squared difference between the estimate and the true value

\[
\langle (\bar{X}_n - \mu)^2 \rangle = \left\langle \left( \frac{1}{n} \sum_i x_i \right)^2 \right\rangle = \left\langle \left( \frac{1}{n} \sum_i (x_i - \mu) \right)^2 \right\rangle
\]

\[
= \frac{1}{n^2} \sum_i \sum_j (x_i - \mu)(x_j - \mu) \tag{7.26}
\]

\[
= \frac{1}{n^2} \sum_i ((x_i - \mu)^2) = \frac{1}{n^2} \sum_i \sigma^2 = \frac{\sigma^2}{n},
\]

where we exploited the fact that the different \( x_i \) are statistically independent and therefore \( \langle (x_i - \mu)(x_j - \mu) \rangle = \langle (x_i - \mu)^2 \rangle \delta_{ij} = \sigma^2 \delta_{ij} \). Finally we see that the estimate \( \bar{X}_n \) on average approaches the true mean \( \mu \) with \( \sigma/\sqrt{n} \) as we increase the number of samples \( n \).

Next we need to address the reliability of an estimate based on a small number \( n \) of samples. To achieve this goal we introduce a test-statistic \( t \), which is given by the deviation of the estimate \( X_n \) from the real mean \( \mu \), but divided by the estimated standard deviation \( S_n/\sqrt{n} \).

\[
t = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} = \frac{(\bar{X}_n - \mu)/(\sigma/\sqrt{n})}{\sqrt{S_n^2/\sigma^2}}. \tag{7.27}
\]

In the second equality in (7.27) we divide numerator and denominator by \( \sigma \) to visualize that the numerator \( x = (\bar{X}_n - \mu)/(\sigma/\sqrt{n}) \) indeed stems from a normal distribution with unit variance, given by

\[
\psi_x(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{7.28}
\]

Moreover, the denominator \( y = \sqrt{S_n^2/\sigma^2} \) stems from the square root of a \( \chi^2 \)-distribution, the latter given by (7.22). Thus \( t = x/y \) is a random variable, defined by the ratio of a Gaussian random variable \( x \) and a random variable \( y \), which is derived

\(^2\)Defining the sample variance with \( n \) in the denominator yields a value that is too small and is called **biassed**, because using \( \bar{X}_n \), which is derived from the same samples as \( S_n^2 \), is closer to the samples \( x_i \) than the “true” mean \( \mu \). This is compensated by dividing by \( n - 1 \) instead, which leads to the **unbiased sample variance**, defined in (7.24).
from the root of numbers drawn from a $\chi^2$-distribution. Note that originally we start from $n$ samples or independent measurements and one degree of freedom was used to determine the average $\bar{X}_n$. Therefore the estimate of the variance $S_n$ only contains information of $\nu = n - 1$ degrees of freedom. The $\chi^2$—distribution that we need to consider is therefore one for $\nu = n - 1$ degrees of freedom.

On our way to calculate the distribution $\Phi_\nu(t)$ of the test-statistic $t$ we first determine the distribution function of the square root of the $\chi^2$ variable divided by $\sqrt{\nu}$. We introduce $y = \sqrt{q/\nu}$ and change variables from $q$ to $y$ in (7.22)

$$
\phi_y(y) = \int_0^\infty \frac{1}{2^{\nu/2} \Gamma(\nu/2)} q^{\nu/2-1} e^{-q/2} \delta\left(y - \sqrt{q/\nu}\right) dq
$$

$$
= \frac{1}{2^{\nu/2} \Gamma(\nu/2)} \int_0^\infty q^{\nu/2-1} e^{-q/2} \delta(q - vy^2) \frac{1}{(2\sqrt{q})} dq
$$

$$
= \frac{\nu^{\nu/2}}{2^{\nu/2-1} \Gamma(\nu/2)} y^{\nu-1} e^{-vy^2/2} .
$$

(7.29)

In appendix A we follow [5] and show that the sum of random variables and the sum of squares of random variables are statistically independent and therefore we can calculate the probability distribution function $\Phi_\nu(t)$ of $t = x/y$ from the product of the two distribution functions in the following way

$$
\Phi_\nu(t) = \int_0^\infty dy \phi_y(y) \int_{-\infty}^\infty dx \psi_x(x) \delta(t - x/y)
$$

$$
= \int_0^\infty \phi_y(y) \psi_x(yt) |y| dy .
$$

(7.30)

Inserting the distribution functions $\psi_x(x)$ from (7.28) and $\phi_y(y)$ from (7.29) and rearranging terms we find

$$
\Phi_\nu(t) = \frac{1}{\sqrt{2\pi}} \frac{\nu^{\nu/2}}{2^{\nu/2-1} \Gamma(\nu/2)} \int_0^\infty y^{\nu} e^{-(\nu+t^2)y^2/2} dy .
$$

(7.31)

The substitution $z = (\nu + t^2)y^2/2$ and some rearrangements lead to

$$
\Phi_\nu(t) = \frac{1}{\sqrt{\pi}} \frac{\nu^{\nu/2}}{\Gamma(\nu/2)} \frac{1}{(\nu + t^2)^{(\nu+1)/2}} \int_0^\infty z^{(\nu-1)/2} e^{-z} dz .
$$

(7.32)
The integral is given by the gamma function, defined in (7.20), as $\Gamma((\nu + 1)/2)$, such that we finally arrive at

$$
\Phi_{\nu}(t) = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma((\nu + 1)/2)}{\Gamma(\nu/2)} \left( \frac{1}{1 + t^2/\nu} \right)^{\nu/2},
$$

(7.33)

which is the commonly found form of Student’s $t$-distribution. Note, that it is expressed in terms of $\nu$, the number of degrees of freedom, rather than the number of samples $n = \nu + 1$.

It is instructive to consider the limiting cases of the distribution. In particular, for $n = \nu + 1 = 2$ we recover the Lorentz, Breit-Wigner, or Cauchy-distribution and for infinitely many degrees of freedom it can be shown that we recover the Gaussian distribution. The distributions for other values of $\nu$ lie between these extremes. Figure 7.5 shows $\Phi_{\nu}(t)$ for $\nu = 1, 2, 5, \text{and } 100$.

In order to assess the reliability of testing only a few samples we need to calculate the probability that our test statistic $t = (\bar{X}_n - \mu)/(S_n/\sqrt{n})$ lies within the central range of the distribution. If we specify this central range to lie within $\pm \hat{t}$ this probability is given by the integral

$$
A(\hat{t}, \nu) = \int_{-\hat{t}}^{\hat{t}} \Phi_{\nu}(t) dt = 1 - I_x\left(\frac{\nu}{2}, \frac{1}{2}\right)
$$

(7.34)

where $I_x(a, b)$ is the (regularized) incomplete beta function \cite{4} and $x = \nu/(\nu + \hat{t}^2)$. The second equality follows from the substitution $s = t^2$ and the definition of the incomplete beta function as an integral with the same integrand \cite{4}. On the left-hand side in Fig. 7.6 we show $A(\hat{t}, \nu)$ for $\nu = 1, 2, 5, \text{and } 100$ as a function of $\hat{t}$. The solid black line corresponds to the case $\nu = 100$, where $\Phi_{\nu}(t)$ approaches a Gaussian, which contains 95% of the distribution within two standard deviations and corresponds to $\hat{t} \approx 2$. The range containing a percentage, say 95%, of the distribution, defines the 95% confidence level. When taking fewer samples $n = \nu + 1$ the probability to find a value of $t$ within $\pm \hat{t}$ is reduced compared to the Gaussian. For example, if we only test three samples ($\nu = 2$) the 95% confidence level is only reached at $|\hat{t}| \approx 4.3$. For convenience, we show the values $|\hat{t}|$ where $A(\hat{t}, \nu)$ as a function of $\nu$ reaches 95, 90, and 80% on the right-hand plot in Fig. 7.6. The solid black curve corresponds to the 95% confidence level and indeed we find the point mentioned above at $\nu = 2$ and $|\hat{t}| \approx 4.3$ on it. Furthermore, for large values of $\nu$ it approaches $\hat{t} \approx 2$ as expected for Gaussian distributions. Below the curve for the 95% level, we find the curves for 90 and 80%. They correspond to a smaller area around the center of the distribution function and consequently a smaller probability of finding a value of $t$ closer to the center. We point out that for $\nu > 10$ the curves rapidly approach their asymptotic values, which agree with those of a Gaussian. This is the reason for the common method to use two standard deviations to specify the 95% confidence level.
Let us come back to William Gosset at Guinness and ensure the quality of the delivery of a hundred bags of barley. Since we are lazy, we only test $n = 3$ randomly selected bags that have quality indicators $x = 16, 12, \text{ and } 17$. So, what is the range of values $x$ that encloses the true mean $\mu$ with 90% confidence? To find out, we use (7.24) and calculate $\bar{X}_3 = 15$ and $S_3 \approx 2.6$ from the three samples. From the right-hand plot in Fig. 7.6 we find $|\hat{t}| \approx 2.92$ for $\nu = n - 1 = 2$ degrees of freedom at the desired 90% confidence level. From (7.27) we now determine the corresponding range of $\mu$ to be $\mu = X_3 \pm |\hat{t}|S_3/\sqrt{3} \approx 15 \pm 4.5$, which is a rather large range. We therefore decide to take a fourth sample, which turns out to be $x = 18$. Repeating the above calculation, we now find the 90% range to be $\mu \approx 15.8 \pm 3.1$. Maybe we should take a fifth sample, but that is left as an exercise.
In this section, we determined the range that contains the “true” mean $\mu$ of a distribution with a certain level of confidence. In the next section we will address a related problem: the validation or rejection of a hypothesis about the value a parameter is expected to have.

### 7.6 Hypothesis Testing and $p$-Values

In a regression analysis we might wonder whether we really need to include a certain fit parameter $\hat{X}$ or whether the model works equally well when omitting it. We can address this problem by testing the hypothesis $\hat{X} = 0$. In particular, we reject the hypothesis $\hat{X} = 0$ if the test-statistics $t = X/\sigma(X)$ is very large and lies in the tails of the distribution. Here $X$ is the value of a fitted parameter from (7.5) and $\sigma(X)$ is its error bar, extracted from (7.7). In particular, if we find a value of $t$ that lies in the tails containing 10% probability, as shown by the red areas in Fig. 7.7, we say that “the hypothesis $\hat{X} = 0$ is rejected at the 10% level.”

Once we have determined the test-statistics $t$ from the regression analysis we can calculate the probability—the $p$-value—of finding an even more extreme value as

$$p = \int_{t}^{\infty} \Phi_\nu(t') dt',$$  

where we assumed that $t$ was positive. If it is negative we have to integrate from $-\infty$ to $t$ instead. The parameter $\nu$ is the number of degrees of freedom $\nu = n - m$ of the regression, where $n$ and $m$ are defined in the context of (7.2).

---

**Fig. 7.7** Student’s $t$-distribution for one degree of freedom $\nu = n - m = 1$ with the tails, containing 5% each, indicated by the red area. The tails begin at $|t| = 6.31$. If the actually observed $t$ lies in the tail region, the hypothesis $\hat{X} = 0$ is rejected at the 10% level.
For an example let us return to the question posed in the introduction to this chapter whether all fit parameters are really necessary. Therefore we consider fitting data to a straight line as shown in Fig. 7.1. It appears reasonable to fit the linear dependence $y_i = a + b s_i$ to the dataset but we might even consider a second-order polynomial $y_i = a + b s_i + c s_i^2$, which will follow the data points even closer and result in a smaller $\chi^2$, because we have the additional parameter $c$ to approximate the dataset. But since we expect the data to lie on the straight line we state the hypothesis that $c = 0$. We test it by fitting the second-order polynomial to the dataset, then determine the error bars $\sigma(c)$ of the fit parameter $c$ using (7.7), and finally calculate the test statistics $t = c/\sigma(c)$. If we have many degrees of freedom $\nu = n - m \gg 1$, we can approximate $\Phi_\nu(t)$ by a Gaussian and check whether $t$ is larger than 2, which would indicate that our hypothesis $c = 0$ is rejected at the 10% level. Conversely, if $t$ is smaller, we corroborate the hypothesis that $c$ is consistent with zero and we might as well omit it from the fit.

The discussed method works well to test hypotheses about individual fit parameters, but occasionally we have to figure out whether we can omit a larger number of fit parameters at the same time. This is the topic of the following section.

### 7.7 F-Test

In the previous section we used the $t$-statistic to determine whether one coefficient in a regression model is compatible with zero and can be omitted. This works well if there is only a single obsolete coefficient, but might fail if several of the coefficients are significantly correlated. In that case the error bars for each of the coefficients is large and leads to the conclusion that the coefficient can be omitted, but the real origin of the problem is that the fitting procedure fails to work out whether to assign the uncertainty to one or the other(s) of the correlated coefficients. This leads to serious mis-interpretations of fit results.

One way to solve the dilemma with several potentially correlated coefficients and to determine which ones to omit is the F-test. In this test the contribution of a group of $p$ fit parameters to the $\chi^2_p$ is determined. We define this by the squared difference of the $N$ “measurement” values $y_i$ and the regression model, normalized to the measurement error $\sigma_i$, thus $s_i = (y_i - \sum_{j=1}^{p} A_{ij} x_j)/\sigma_i$. For $\chi^2_p$ we then obtain

$$\chi^2_p = \sum_{i=1}^{N} s_i^2 = \sum_{i=1}^{N} \left(\frac{y_i - \sum_{j=1}^{p} A_{ij} x_j}{\sigma_i}\right)^2 .$$  \hspace{1cm} (7.36)

In the next step we increase the number of fit parameters to $q > p$ and test whether $\chi^2_q$ is significantly smaller than $\chi^2_p$. In order to quantify this improvement, we introduce the $F$-statistic
\[ f = \frac{(\chi^2_p - \chi^2_q)/(q - p)}{\chi^2_q/(N - q)}. \] (7.37)

Here the denominator is the \( \chi^2 \) per degree of freedom of the “larger” fit, whereas the numerator is the difference \( \chi^2_p - \chi^2_q \) per additional degree of freedom, which come from the additional \( q - p \) fit parameters. The \( F \)-value thus measures the relative reduction of the \( \chi^2 \) when increasing the number of fit parameters. We point out that the denominator depends on \( N - q \) squared random number \( s_i \) and the numerator on the \( q - p \) additional \( s_i \). In Appendix A we motivate that the random numbers in the numerator and those in the denominator are independent, which allows us to derive the distribution of the \( F \)-statistic as the ratio of two independent \( \chi^2 \)-distributions; one with \( n = q - p \) degrees of freedom, the other with \( m = N - q \) degrees of freedom.

For easy reference, we again display the \( \chi^2 \)-distribution function, already shown in (7.22)

\[ \psi_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1/2} e^{-x/2} . \]

We will use it once to describe the \( \chi^2 \)-variable \( x \) with \( n = q - p \) degrees of freedom that appears in the numerator of (7.37) and once for the \( \chi^2 \)-variable \( y \) for the \( m = N - q \) degrees of freedom in the denominator. Using these variables the \( F \)-statistic \( f \) is given by

\[ f = \frac{x/n}{y/m} = \frac{m \cdot x}{n \cdot y} \quad \text{with} \quad m = N - q \quad \text{and} \quad n = q - p . \] (7.38)

Our next task is to calculate the distribution function \( \Psi(f) \) for the \( f \) from

\[ \Psi_{n,m}(f) = \int_0^\infty dy \psi_m(y) \int_0^\infty dx \psi_n(x) \delta(f - mx/ny) \]

\[ = \frac{n}{m} \int_0^\infty dy \psi_m(y) \psi_n(nyf/m)y , \] (7.39)

where we used that the \( \chi^2 \)-variable \( y \) is positive and that \( \partial f/\partial x = m/ny \). Again, the delta function collects all values of \( x \) and \( y \) that result in \( f = mx/ny \). Inserting the definition of \( \psi_m \) and \( \psi_n \) we obtain

\[ \Psi_{n,m}(f) = \frac{n/m}{2^{m/2+n/2}\Gamma(m/2)\Gamma(n/2)} \left( \frac{nf}{m} \right)^{n/2-1} \int_0^\infty y^{(m+n)/2-1} e^{-(nyf/m)/2} dy \]

\[ = \frac{\Gamma((m + n)/2)\Gamma(n/2)}{\Gamma(m/2)\Gamma(n/2)} \left( \frac{n}{m} \right)^{n/2} \left( 1 + nf/m \right)^{(n+m)/2} f^{n/2-1} . \] (7.40)
where we substitute \( z = (1 + nf/m)y/2 \) in the second step and then recognize the remaining integral as a representation of a Gamma-function with argument \((n + m)/2\). Note that the specific combination of Gamma-functions can be expressed as a Beta-function \( B(p, q) \) with [4]

\[
B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p + q)},
\]

which allows us to write the \( F - \)distribution function in the form

\[
\Psi_{n,m}(f) = \frac{1}{B(n/2, m/2)} \left( \frac{n}{m} \right)^{n/2} \frac{f^{n/2-1}}{(1 + nf/m)^{(n+m)/2}}
\]  

(7.42)

that clearly shows that it depends only on the number of degrees of freedom \( m = N - q \) and \( n = q - p \). We can now use it to assess whether we need to include \( n \) additional fit parameters in a model and whether this added complexity is worth the effort.

Figure 7.8 shows the \( F - \)distribution function for a few values of \( n \) and \( m \). In the first two cases we have a small number of degrees of freedom \( N - q = m = 5 \) where we fit \( q \) parameters to \( N \) data points. We then compare what distributions of our test-characteristic \( f \) we can expect, if we add \( n \) additional fit parameters The dashed blue line corresponds to a situation, where we add \( n = 2 \) additional fit parameters. We see that the distribution function is peaked near zero, which indicates that small values of \( f \) are very likely. Adding a third fit parameter \( (n = 3) \) causes \( \Psi_{3,5}(f) \) to assume the shape indicated by the dot-dashed red line. Now very small values near zero are less likely and the distribution shows a peak. The solid black line in Fig. 7.8 illustrates a case where we add \( n = 10 \) additional fit parameters to a fit that originally had \( N - q = m = 30 \) degrees of freedom. We observe that the peak of the distribution moves towards \( f = 1 \) but is rather broad and shows significant tails.
In order to intuitively assess whether a found $f$-statistics is likely or not, we calculate the probability that its value is even smaller or larger, depending whether we find a particularly small or large value of $f$. The probability can be expressed in terms of the cumulative distribution function of $\Psi_{n,m}(f)$ by

$$\int_{0}^{f} \Psi_{n,m}(f')df' = \frac{1}{B\left(\frac{n}{2}, \frac{m}{2}\right)} \int_{0}^{\frac{f}{m}} t^{\frac{n}{2}-1}dt \quad (7.43)$$

$$= \frac{1}{B\left(\frac{n}{2}, \frac{m}{2}\right)} \int_{0}^{\left(\frac{m}{n+f}\right)} u^{\frac{n}{2}-1}(1-u)^{\frac{m}{2}-1}du = I_\hat{x}\left(\frac{n}{2}, \frac{m}{2}\right)$$

with $\hat{x} = nf/(m + nf)$ and where we used the substitution $t = u/(1-u)$ in the second equality. $I_\hat{x}(a,b)$ is the (regularized) incomplete beta function [4], we already encountered in (7.34), and $B(a,b)$ is the beta function [4]. Now, to assess the probability, the $p$-value $p(f)$, of finding an even smaller value $f$ is given by $p(f) = I_\hat{x}(n/2, m/2)$. Note that $\hat{x}$ depends on $f$. Conversely, the probability of finding a larger value, which is relevant on the on the right-hand side of the maximum, is given by $p(f) = 1 - I_\hat{x}(n/2, m/2)$.

Rejecting hypotheses works in the same way as discussed above in Sect. 7.6. If an $F-$value $f$, computed from data, lies in the tails of the distribution and exceeds the value for a 10% tail-fraction, we say that the hypothesis is rejected at the 10% level.

### 7.8 Parsimony

In philosophy, a razor is a criterion to remove explanations that are rather improbable. The image of shaving off something unwanted, for example, one’s beard, comes to mind. One well-known example is Occam’s razor, which, translated from Latin, reads “plurality should not be posited without necessity.” Today, it is often rephrased as “the simpler solution is probably the correct one.” Applied to the our regression analysis and fitting parameters to models it guides us to seek models with as few fit parameters as possible, which is also called the principle of parsimony. It helps us to avoid adding unnecessary parameters to a model that may lead to over-fitting which causes the model parameters to be overly affected, or over-constrained, by the noise in the data. Such a model then works very well with the existing data set with its particular noise spectrum, but its predictive power to explain new data is limited.

A classical example of over-fitting is the fit of a polynomial of degree $n - 1$ to $n$ data points. With many data points the polynomial is of very high order. It perfectly fits the data set and makes the $\chi^2$ of the fit to zero, but outside the range of the original
data set large excursions typically occur. Any additional data point is unlikely to be well-described by this highly over-constrained polynomial.

A further reason to avoid too many fit parameters is that groups of parameters are highly correlated and, again, this degeneracy is heavily affected by noise. It is therefore advisable to construct a model with the least number of fit parameters—to be parsimonious, in other words. Using fewer parameters therefore makes models more robust against the spurious influence of noise.

Now that we have the methods to fit parameters to models and assess their validity and robustness, we can take a closer look at time series and extract useful information from the raw data.

**Exercises**

1. In an experiment you recorded the parameter $y$ as you changed another parameter $s$ in Table 7.1. (a) You expect a linear dependence between $s$ and $y$ and therefore set up (7.1), use the methods from Sect. 7.1, and determine the slope $a$ and intercept $b$ of a straight-line fit to the data. If you assume that all error bars are $\sigma = 1$, what values do you obtain for $a$ and $b$ and what are the error bars of these fit parameters. (b) You realize that you were less careful when recording the data points in the range $-1 < s < 4$. You therefore double the error bars for the corresponding data points, as indicated by the last row in the table, and redo the fit. What values do you obtain for $a$ and $b$ and what are their error bars?

2. Calculate the $R$-value of the data and fit from Exercise 1a.

3. Use (7.5) and (7.6) to prove that (7.7) is correct.

4. If $x$ is a Gaussian random variable, calculate the probability distribution functions of (a) $y = x - a$, (b) $y = bx$, and (c) $y = cx^2$.

5. You know that the data in the file *ex7_5.dat*, available from the book’s web page, comes from a process that can be fitted by a polynomial of third order. (a) Plot the data. (b) Find the coefficients of a third-order polynomial in a regression analysis. (c) Estimate the error bars $\sigma_y$ of the $y$–values (the “measurements”) from the rms deviation of your fit-polynomial to the data points. Note that is a very heuristic approach to estimate error bars and can be criticised! (d) Calculate the covariance matrix, based on your estimate of the error bars $\sigma_y$, and deduce the error bars of the polynomial coefficients. Is there a coefficient that is so small

| $s$ | $y$  | $\sigma$ |
|-----|------|---------|
| $-2$ | $-7.0$ | 1 |
| $-1$ | $-3.5$ | 2 |
| 0   | $-3.3$ | 2 |
| 1   | 0.1  | 2 |
| 2   | 1.6  | 2 |
| 3   | 0.3  | 2 |
| 4   | 1.5  | 2 |
| 5   | 5.5  | 1 |
as to be consistent with zero. (e) Determine its F-value \( \hat{f} \) and the probability to find an F-value that is even larger than \( \hat{f} \).

6. Let's pursue the quality control at Guinness from the end of Sect. 7.5 and assume that we test a fifth bag of barley with the test result \( x = 13 \). What is the 90% confidence range now? If we were to be satisfied with an 80% confidence level, what would that range be?

References

1. US Bureau of Labor Statistics data on educational attainment (2018). https://www.bls.gov/emp/tables/unemployment-earnings-education.htm
2. Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by Johns Hopkins CSSE (2019). https://github.com/CSSEGISandData/COVID-19
3. H. Hethcote, The mathematics of infectious diseases. SIAM Rev. 42, 599 (2000)
4. M. Abramowitz, I. Stegun, Handbook of Mathematical Functions (Dover, New York, 1972)
5. G. Casella, R.L. Berger, Statistical Inference, 2nd edn. (Brooks/Cole, 2008)