Simple Linear Regression II: Predictions and Model Analysis

Inferences about a Single Predicted Value

We are interested in finding an "estimate" (guess) or a ***prediction*** for the value of the random variable $Y \mid x$. Note the essential difference:

- An ***estimate*** is a statistical statement on the value of an unknown, but fixed, population parameter.
- A ***prediction*** is a statistical statement on the value of an essentially random quantity.

We define a $100(1-\alpha)\%$ prediction interval $[L_1, L_2]$ for a random variable X by

$$P[L_1 \leq X \leq L_2] = 1 - \alpha.$$

As a ***predictor*** $\widehat{Y \mid x}$ for the value of $Y \mid x$ we use the estimator for the mean, i.e., we set

$$\widehat{Y \mid x} = \widehat{\mu}_{Y|x} = B_0 + B_1 x.$$

In order to find a prediction interval, we need to analyze the distribution of $\widehat{Y \mid x}$.

Inferences about a Single Predicted Value

Recall that $\widehat{\mu}_{Y|x}$ follows a normal distribution with mean $\mu_{Y|x}$ and variance $\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right)\sigma^2$. Furthermore, $Y \mid x$ is normally distributed with mean $\mu_{Y|x}$ and variance σ^2.

Hence $\widehat{Y \mid x} - Y \mid x$ is normally distributed and, furthermore,

$$\mathrm{E}[\widehat{Y \mid x} - Y \mid x] = \mu_{Y|x} - \mu_{Y|x} = 0,$$

$$\mathrm{Var}[\widehat{Y \mid x} - Y \mid x] = \left(\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right)\sigma^2 + \sigma^2 = \left(1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right)\sigma^2.$$

Thus, after standardizing and dividing by S/σ we obtain the T_{n-2} random variable

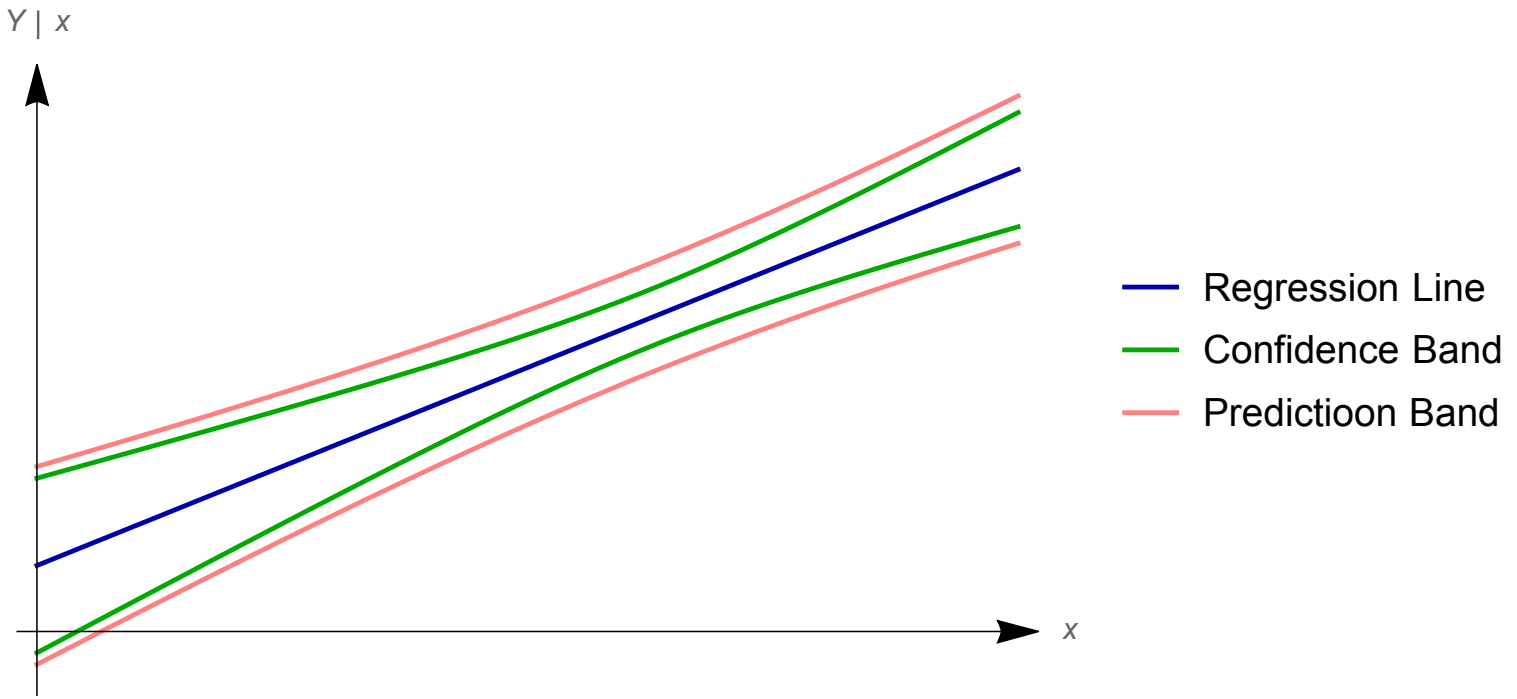
$$T_{n-2} = \frac{\widehat{Y \mid x} - Y \mid x}{S\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}}$$

Inferences about a Single Predicted Value

We thus obtain the following $100(1-\alpha)\%$ prediction interval for $Y \mid x$:

$$\widehat{Y \mid x} \pm t_{\alpha/2,n-2} S \sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}} \tag{25.1}$$

The limits of the confidence interval (24.6) and the prediction interval (25.1), plotted as functions of x, are commonly called ***confidence bands*** and ***prediction bands*** for the regression.



Confidence and Prediction Intervals

25.1. Example. Continuing with the data from Example 24.1, Mathematica gives confidence bands (24.6) for the estimated mean as

```
conf = model["MeanPredictionBands", ConfidenceLevel → 0.95]
```

$$\left\{13.6013 - 0.0794677\,x - 2.06866\sqrt{0.331656 - 0.0114296\,x + 0.00010882\,x^2}\,,\right.$$
$$\left.13.6013 - 0.0794677\,x + 2.06866\sqrt{0.331656 - 0.0114296\,x + 0.00010882\,x^2}\right\}$$

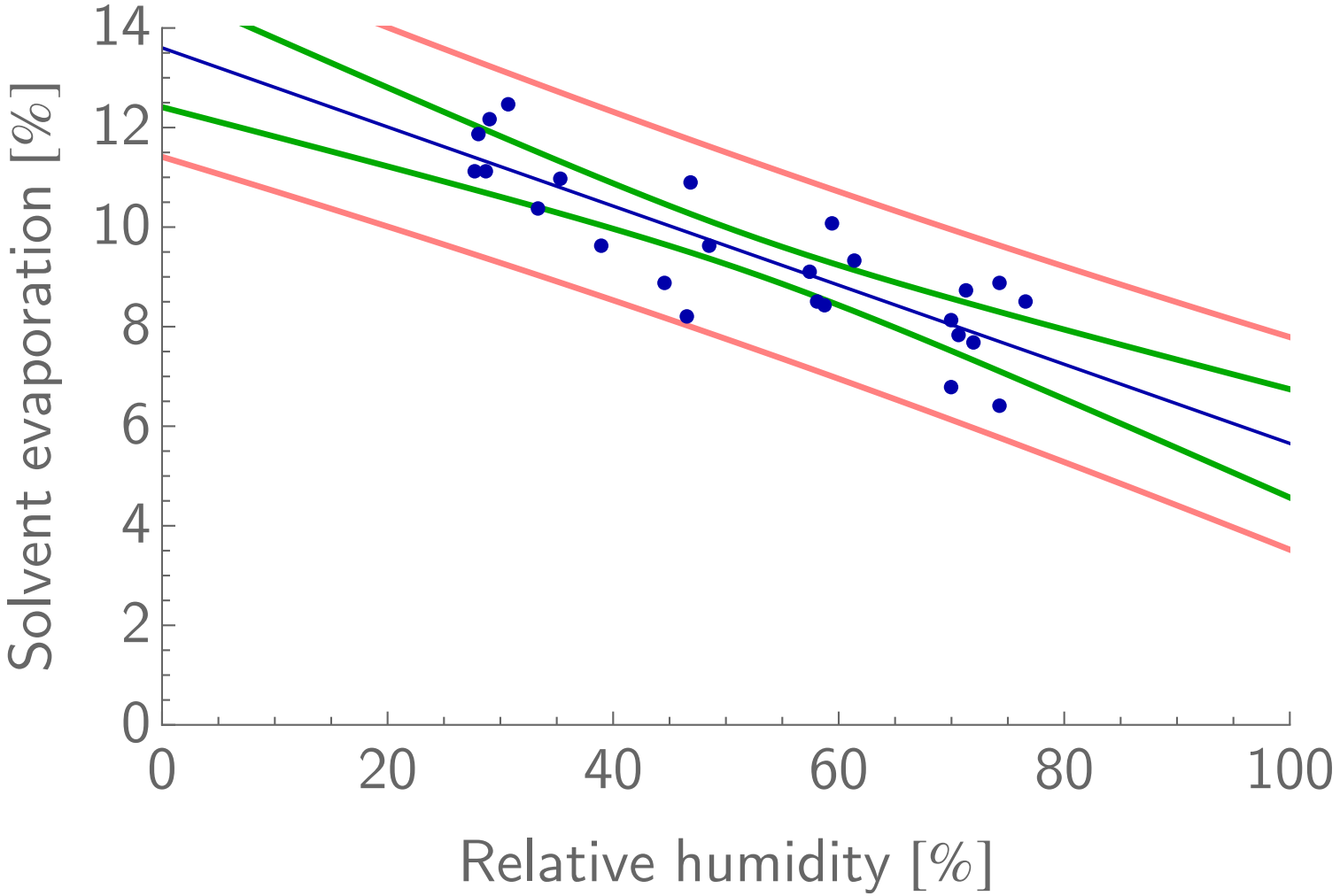
Prediction bands (25.1) are given by

```
pred = model["SinglePredictionBands", ConfidenceLevel → 0.95]
```

$$\left\{13.6013 - 0.0794677\,x - 2.06866\sqrt{1.12011 - 0.0114296\,x + 0.00010882\,x^2}\,,\right.$$
$$\left.13.6013 - 0.0794677\,x + 2.06866\sqrt{1.12011 - 0.0114296\,x + 0.00010882\,x^2}\right\}$$

Confidence and Prediction Intervals

Below, the prediction bands are shown in red, while the confidence bands for the estimated mean are green:



Analysis of the Model

Achievements so far:

- Inferences on model parameters β_0, β_1.
- Inferences on estimated mean $\widehat{\mu}_{Y|x}$.
- Prediction for $Y \mid x$.

But is our linear model actually appropriate?

Crucial Quantities:

- The total variation of the response variable,

$$\mathrm{SS_T} = S_{yy} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

We will also call this the ***Total Sum of Squares***. It represents the variation of Y regardless of any model.

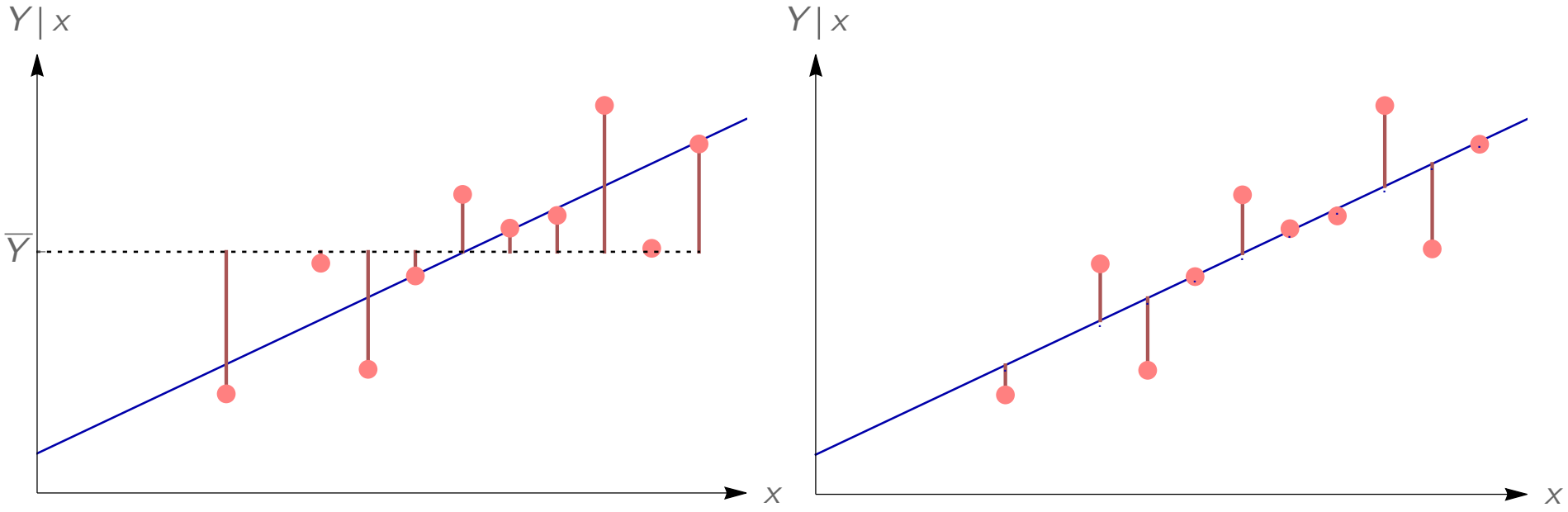
Analysis of the Model

Crucial Quantities:

- The ***Error Sum of Squares***

$$SS_E = \sum_{i=1}^{n} \big(Y_i - (b_0 + b_1 x)\big)^2.$$

It represents the variation of Y that remains after we have applied the model.



Coefficient of Determination

Of course,

$$\mathrm{SS_E} \leq \mathrm{SS_T}$$

and we define the the ***coefficient of determination***

$$R^2 := \frac{\mathrm{SS_T} - \mathrm{SS_E}}{\mathrm{SS_T}}.$$

Sometimes, one reports $R^2 \cdot 100\%$.

The coefficient R^2 expresses the ***proportion of the total variation in Y that is explained by the linear model***.

Connection to Correlation

Recall from (24.5) that

$$\mathrm{SS_E} = S_{yy} - B_1 S_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}},$$

so that

$$R^2 = \frac{\mathrm{SS_T} - \mathrm{SS_E}}{\mathrm{SS_T}} = \frac{S_{xy}^2}{S_{xx} S_{yy}}.$$

The right-hand side is exactly the square of the estimator (22.1) for the the correlation coefficient ρ_{XY}.

Since the correlation ϱ_{XY} measures the linearity of the relationship between X and Y, this is not surprising.

Connection to Significance of Regression

The statistic that we have used in the Test for Significance of regression 24.6 is

$$\frac{B_1}{\sqrt{S^2/S_{xx}}} = \frac{S_{xy}/S_{xx}}{\sqrt{\mathrm{SS_E}/[(n-2)S_{xx}]}} \tag{25.2}$$

$$= \frac{R}{\sqrt{1-R^2}}\sqrt{n-2}, \tag{25.3}$$

and we can see that is expressible entirely using the coefficient R^2.

Hence, R^2 alone includes enough information to conduct the test for significance of regression.

Test for Correlation

Conversely, we can adapt the above discussion to perform a two-sided Fisher test for a vanishing correlation in a bivariate normal distribution:

25.2. Test for Correlation. Let (X, Y) follow a bivariate normal distribution with correlation coefficient $\varrho \in (-1, 1)$. Let R be the estimator (22.1) for ϱ. Then

$$H_0 \colon \varrho = 0$$

is rejected at significance level α if

$$\left| \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \right| > t_{\alpha/2,n-2}.$$

Lack-of-Fit and Pure Error

Problem:

- R^2 measures how much of the total variation is explained by the linear model.
- If R^2 is not large, then the model does not explain a significant amount of the fluctuation of the measured values y_i.
- In short, SS_E is large.

Why could SS_E be large?

- Either σ^2 is very large ***(pure error)***
- or the model is wrong. ***(lack-of-fit error)***

To tell which of the two predominates, we need to be able to take ***repeated measurements*** of $Y \mid x_i$ for the same value of x_i.

Repeated Measurements

We can directly measure pure error (due to σ^2) if we have ***repeated measurements*** available. That is, at one or more points x_i, $i = 1, \ldots, k$, we have at least two observations on Y.

Let Y_{ij} denote the jth observation of $Y \mid x_i$, where $j = 1, \ldots, n_i$.

The total number of observations is

$$n = n_1 + n_2 + \cdots + n_k = \sum_{i=1}^{k} n_i.$$

Recall: Repeated measurements are treated just like any other measurements in regression analysis.

Internal Sum of Squares

For each $i = 1, \ldots, k$ we can view $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ as a random sample of size n_i of the random variable $Y_i = Y \mid x_i$.

An unbiased estimator for $\mu_{Y|x_i}$ is the sample mean

$$\overline{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}.$$

The statistic

$$\sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2$$

measures the natural variability of $Y \mid x_i$ and is called an ***internal sum of squares***.

Error Sum of Squares (Pure Error)

By summing over all internal sums of squares we obtain the ***error sum of squares due to pure error***,

$$\begin{aligned} \mathrm{SS}_{\mathrm{E,pe}} &:= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_i)^2 \qquad (25.4)\\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}Y_{ij}^2 - \sum_{i=1}^{k}\frac{1}{n_i}\Big(\sum_{j=1}^{n_i}Y_{ij}\Big)^2 \end{aligned}$$

It is not difficult to see that

$$\frac{1}{\sigma^2}\,\mathrm{SS}_{\mathrm{E,pe}} = \sum_{i=1}^{k}\frac{1}{\sigma^2}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_i)^2$$

follows a chi-squared distribution with $n-k$ degrees of freedom.

Error Sum of Squares

Note that $SS_{E,lf} \leq SS_E$ since

$$SS_E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - (b_0 + b_1 x_i)\right)^2$$

and in general

$$\sum_{j=1}^{n_i} \left(Y_{ij} - z\right)^2$$

is minimized if $z = \overline{Y}_i$.

Error Sum of Squares (Lack of Fit)

We therefore define the ***error sum of squares due to lack of fit*** by

$$\mathrm{SS}_{\mathrm{E,lf}} := \mathrm{SS}_{\mathrm{E}} - \mathrm{SS}_{\mathrm{E,pe}} .$$

Since $\mathrm{SS}_{\mathrm{E}} = \mathrm{SS}_{\mathrm{E,pe}} + \mathrm{SS}_{\mathrm{E,lf}}$, it seems reasonable that

$$\frac{1}{\sigma^2} \mathrm{SS}_{\mathrm{E,lf}}$$

might follow a chi-squared distribution with

$$(n-2) - (n-k) = k - 2$$

degrees of freedom.

In fact, it can be shown that this is true and that $\mathrm{SS}_{\mathrm{E,lf}}$ is actually a sum of squares.

Testing for Lack of Fit

25.3. Test for Lack of Fit. Let $x_1, \ldots, x_k$ be regressors and $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$, $i = 1, \ldots, k$, the measured responses at each of the regressors. Let $\mathrm{SS}_{\mathrm{E,pe}}$ and $\mathrm{SS}_{\mathrm{E,lf}}$ be the pure error and lack-of-fit sums of squares for a linear regression model. Then

$$H_0 : \text{the linear regression model is appropriate}$$

is rejected at significance level α if the test statistic

$$F_{k-2,n-k} = \frac{\mathrm{SS}_{\mathrm{E,lf}}/(k-2)}{\mathrm{SS}_{\mathrm{E,pe}}/(n-k)}$$

satisfies $F_{k-2,n-k} > f_{\alpha,k-2,n-k}$.

Testing for Lack of Fit

25.4. Example. Consider these data on X, the temperature, in degrees centigrade, at which a chemical reaction is conducted, and Y, the percentage yield obtained:

x_i	30	40	50	60	70
Y_{i1}	13.7	15.5	18.5	17.7	15.0
Y_{i2}	14.0	16.0	20.0	18.1	15.6
Y_{i3}	14.6	17.0	21.1	18.5	16.5

Here $k = 5$, $n_1, \ldots, n_5 = 3$ and $n = 15$. For $x_1 = 30$ we have $\overline{y}_1 = 14.1$ and the internal sum of squares

$$(13.7 - 14.1)^2 + (14.0 - 14.1)^2 + (14.6 - 14.1)^2 = 0.42$$

In the same way we calculate the other two internal sums of squares and obtain the pure error sum of squares

$$\mathrm{SS}_{\mathrm{E,pe}} = 6.453.$$

Testing for Lack of Fit

For our data we can calculate $S_{yy} = 66.6437$, $S_{xy} = 154$ and $b_1 = 0.051$. The total error sum of squares is given by

$$\mathrm{SS_E} = S_{yy} - b_1 S_{xy} = 58.583.$$

The lack-of-fit sum of squares is

$$\mathrm{SS_{E,lf}} = \mathrm{SS_E} - \mathrm{SS_{E,pe}} = 52.13.$$

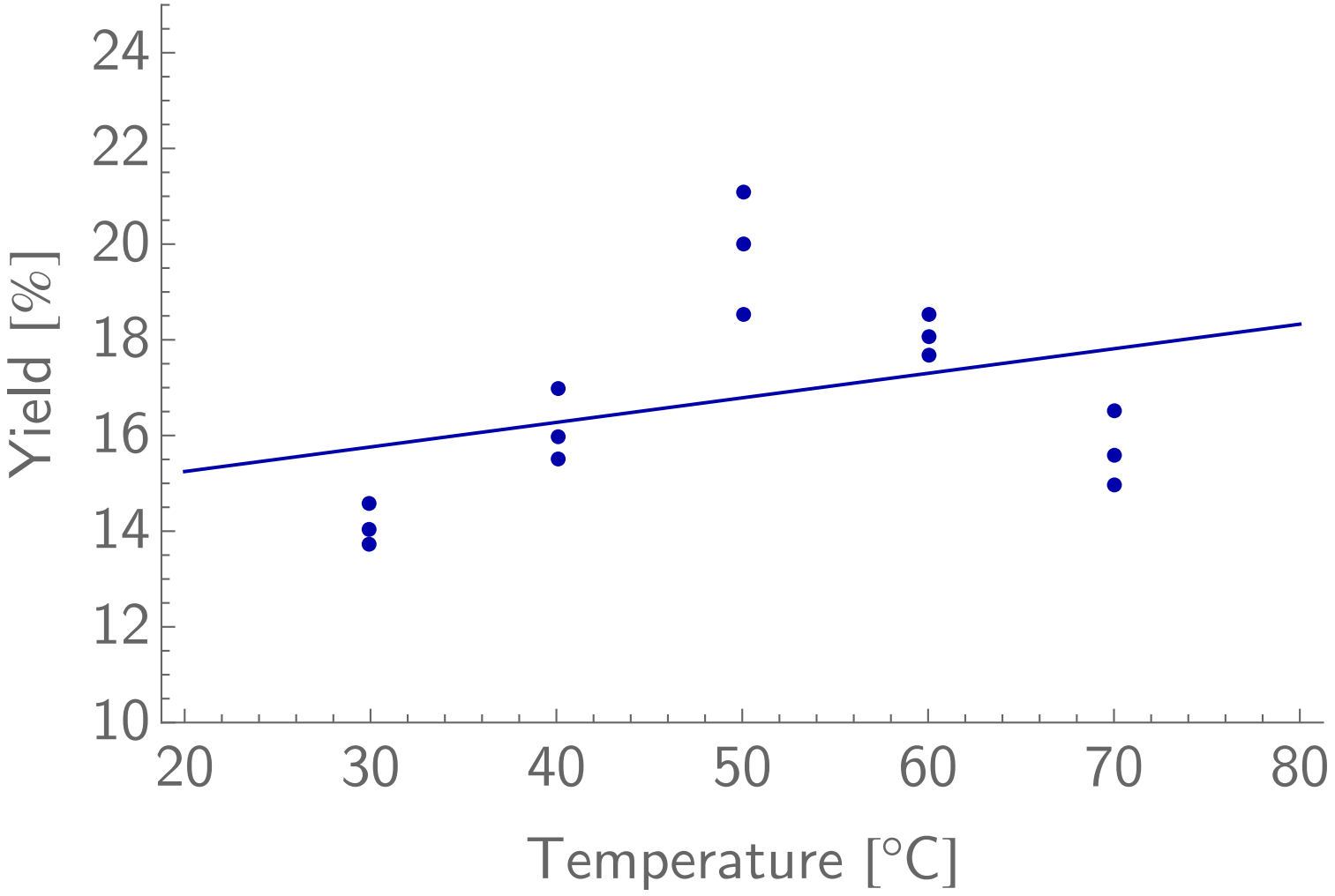
The observed value of the statistic

$$F_{k-2,n-k} = F_{3,10} = \frac{\mathrm{SS_{E,lf}}/(k-2)}{\mathrm{SS_{E,pe}}/(n-k)} = \frac{52.13/3}{6.453/10} = 26.928.$$

Based on the $F_{3,10}$ distribution, we can reject H_0 with $P < 0.05$ ($f_{0.05,3,10} = 3.708$). There is evidence that a linear regression model is not appropriate.

Testing for Lack of Fit

It is clear from the graph below that the linear model is indeed not suitable:



Residual Analysis

The residuals e_i, $i = 1, \ldots, n$ give important information on the model:

- Are they consistent with the assumption of equal variance σ^2?
- Are they consistent with the assumption of a normal distribution?
- Does the linear model seem appropriate?

Plotting the residuals vs. the values of x_i also shows potential gaps in the data.

Never extrapolate the regression model beyond the range of the regressors. Avoid leaving wide gaps in the range of the x_i.

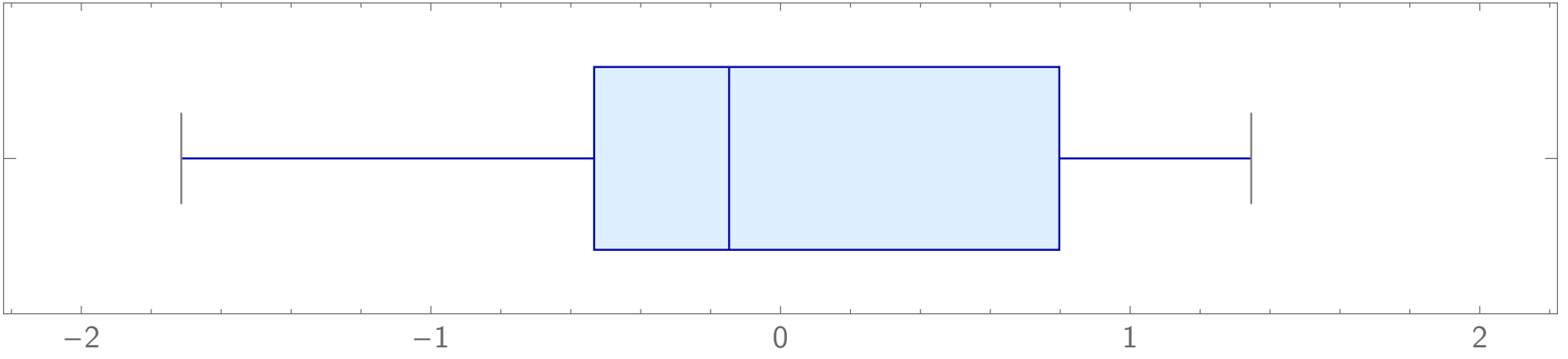
Residual Analysis

25.5. Example. Continuing with the data from Example 24.1, Mathematica gives the residuals as follows

```
residuals = model["FitResiduals"]
```

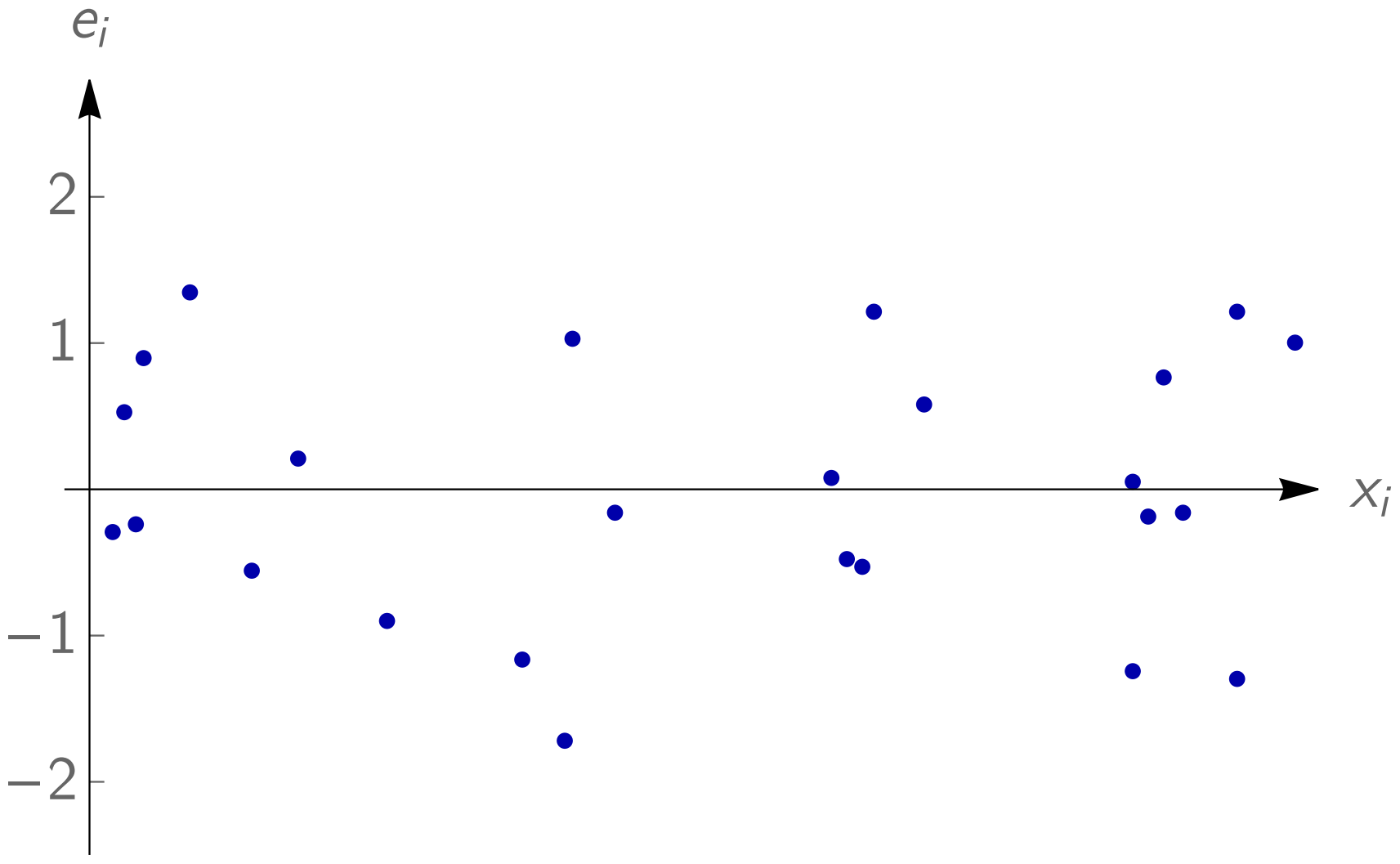
```
{0.203884, -0.300071, 1.34628, -0.528625, 0.577991, 0.764721,
 -1.28893, 0.993847, -0.182959, 0.0680671, -1.71402, 0.895291,
 0.531716, -0.894139, 1.01776, -0.147142, 1.21111, 0.0614135, -1.23859,
 1.21107, -0.171704, -0.484252, -1.15707, -0.547105, -0.22855}
```

A boxplot does not yield strong evidence against the normality assumption:



Residual Analysis

The residual plot does not show any obvious issues:



The Anscombe Quartet

Francis J. Anscombe (19182001). Boilly, Julien-Leopold. (1820). Yale Bulletin and Calendar. November 2, 2001. Vol. 30, No. 9

- $\overline{x} = 9$,
- $\frac{1}{n-1} S_{xx} = 11$,
- $\overline{y} = 7.50$,
- $\frac{1}{n-1} S_{yy} = 4.122$ or 4.127,
- $R^2 = 0.816$,
- $\widehat{\mu}_{Y|x} = 3.00 + 0.500x$

Finally, this example by Francis Anscombe illustrates why it is always important to actually look at the data instead of relying on numerical quantities. In the following four graphs, the data all have these same statistics up to the precision given.

Literature: Anscombe, F. J. ***Graphs in Statistical Analysis***. American Statistician. 27 (1): 17–21. (1973).

Plot the data!