MAT2377

Ali Karimnezhad

Version December 15, 2015

Comments

- These slides cover material from Chapter 7.
- In class, I may use a blackboard. I recommend reading these slides before you come to the class.
- I am planning to spend 2 lectures on this chapter.
- I am not re-writing the textbook. The reference book contains many interesting and practical examples.
- There may be some typos. The final version of the slides will be posted *after* the chapter is finished.

Bivariate data and scatterplot

Data: `Hydrocarbon level` (x) and `Oxygen level` (y):

x: 0.99, 1.02, 1.15, 1.29, 1.46, 1.36, 0.87,

1.23, 1.55, 1.40, 1.19, 1.15, 0.98, 1.01,

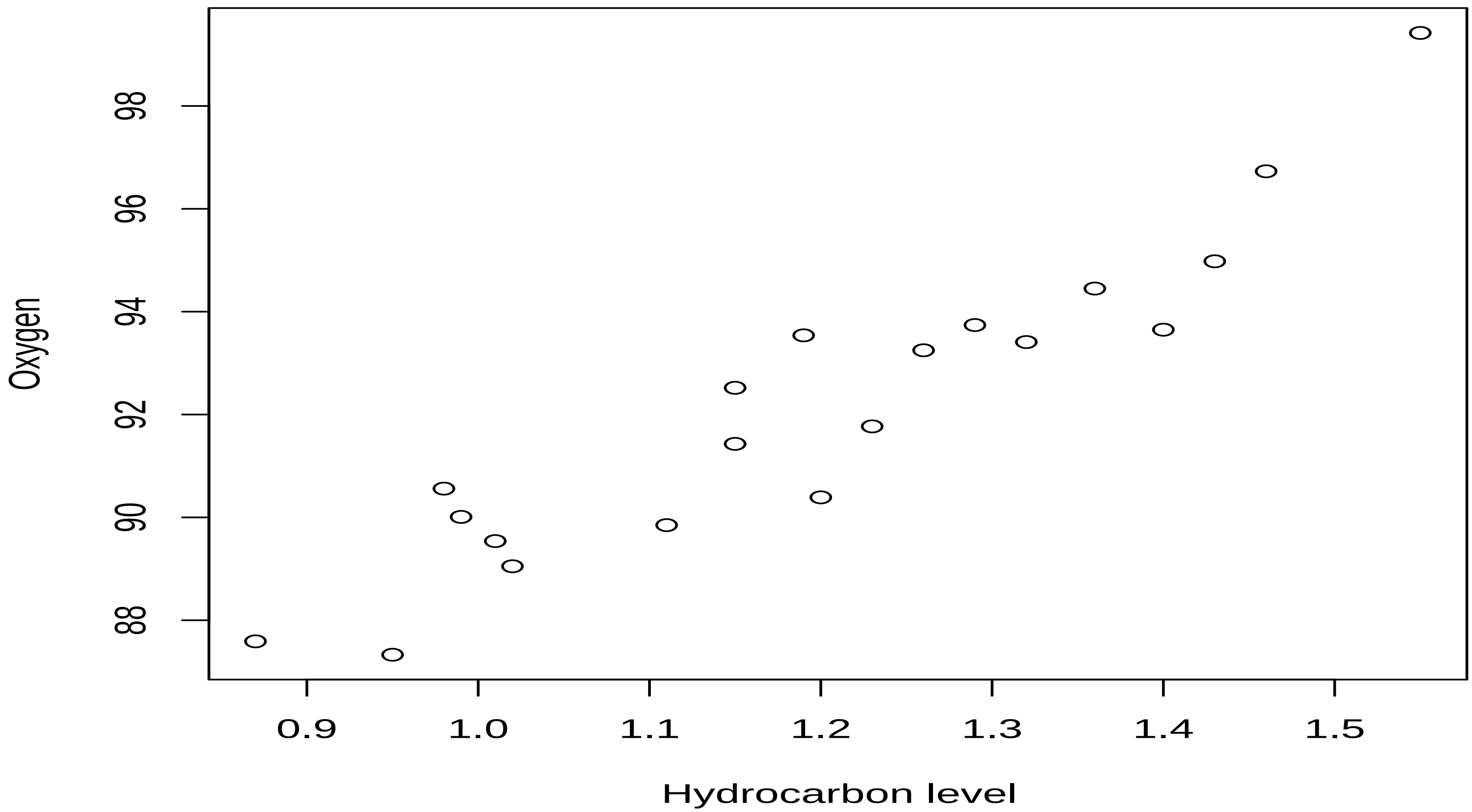
1.11, 1.20, 1.26, 1.32, 1.43, 0.95

----

y: 90.01, 89.05, 91.43, 93.74, 96.73, 94.45,

87.59, 91.77, 99.42, 93.65, 93.54, 92.52, 90.56, 89.54,

89.85, 90.39, 93.25, 93.41, 94.98, 87.33


Oxygen
98
96
94
92
90
88
0.9
1.0
1.1
1.2
1.3
1.4
1.5
Hydrocarbon level

We want to describe the relationship between these two variables. We will use regression analysis. We will assume that our model is given by

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where ϵ is a random error and β_0, β_1 are regression coefficients. The variable x is called regressor (predictor) variable and Y is called a dependent or response variable.

It is assumed that $\mathrm{E}(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = \sigma^2$. In particular,

$$\mathrm{E}(Y|x) = \beta_0 + \beta_1 x.$$

Suppose now that we have observations (x_i, y_i) from our model, so

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n.$$

Our aim is to find $\hat{\beta}_0, \hat{\beta}_1$, estimator of the unknown parameters β_0, β_1. consequently, we will find the estimated (fitted) regression line or the line of the best fit:

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

This line is obtained using the method of least squares. Having the observations y_i, $i = 1, \ldots, n$, their deviation e_i from the line $\beta_0 + \beta_1 x$ are

$$e_i = (y_i - \beta_0 - \beta_1 x_i), \quad i = 1, \ldots, n.$$

So,

$$L(\beta_0, \beta_1) := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Consequently,

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

and

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

Solving $\frac{dL}{d\beta_0} = 0$ and $\frac{dL}{d\beta_1} = 0$ we obtain the least squares estimators of β_0 and β_1:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

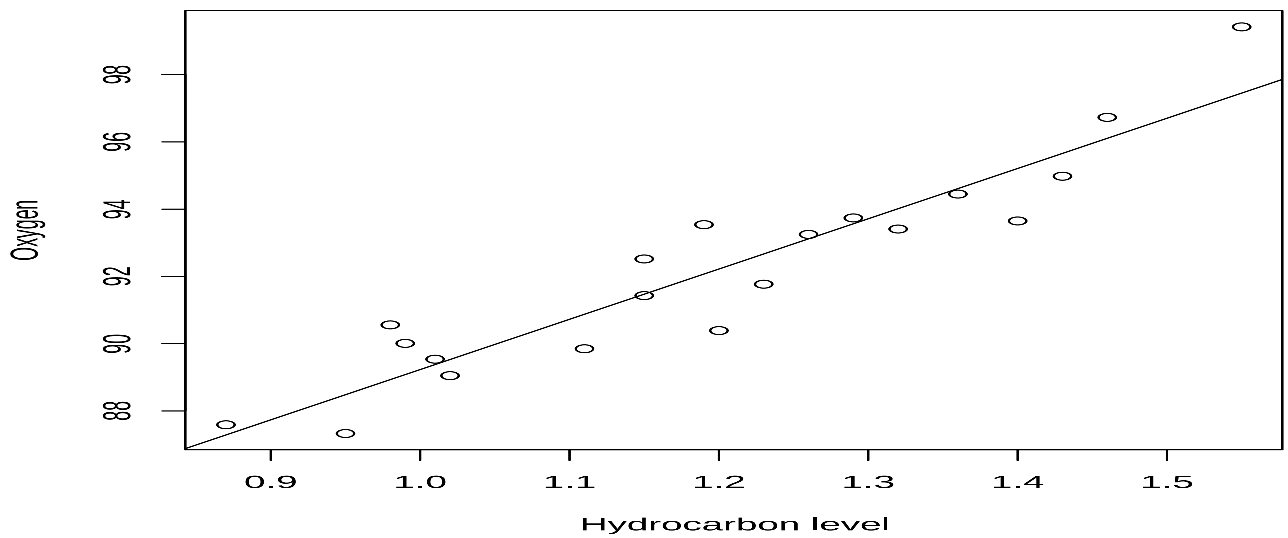
$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Equivalently,

$$\boxed{\begin{array}{rcl} S_{xy} & = & \sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i) \\ S_{xx} & = & \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2 \\ S_{yy} & = & \sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2 \end{array}}$$

Example

For hydrocarbon data we find $\sum x_i = 23.92$, $\sum y_i = 1843.21$, $\sum x_i^2 = 29.2892$, $\sum y_i^2 = 170044.5$, $\sum x_i y_i = 2214.657$, so that $S_{xy} = 10.17744$, $S_{xx} = 0.68088$, $S_{yy} = 173.3769$. Therefore, $\hat{\beta}_0 = 74.28$, $\hat{\beta}_1 = 14.95$. Consequently, the fitted regression line is $\hat{y} = 74.28 + 14.95x$.



Sample Correlation Coefficient.

For data (x_i, y_i) we define the sample correlation coefficient as

$$R = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}} \tag{1}$$

(R is defined only if S_{xx} and S_{yy} are positive).

Example: For the hydrocarbon data the sample correlation coefficient is $R = 0.93$.

Properties of R

- R is unaffected by change of scale or origin. Adding constants to x does not change $x - \bar{x}$ and multiplying x and y by constants changes numerator and denominator equally.

- R is symmetric in x and y.

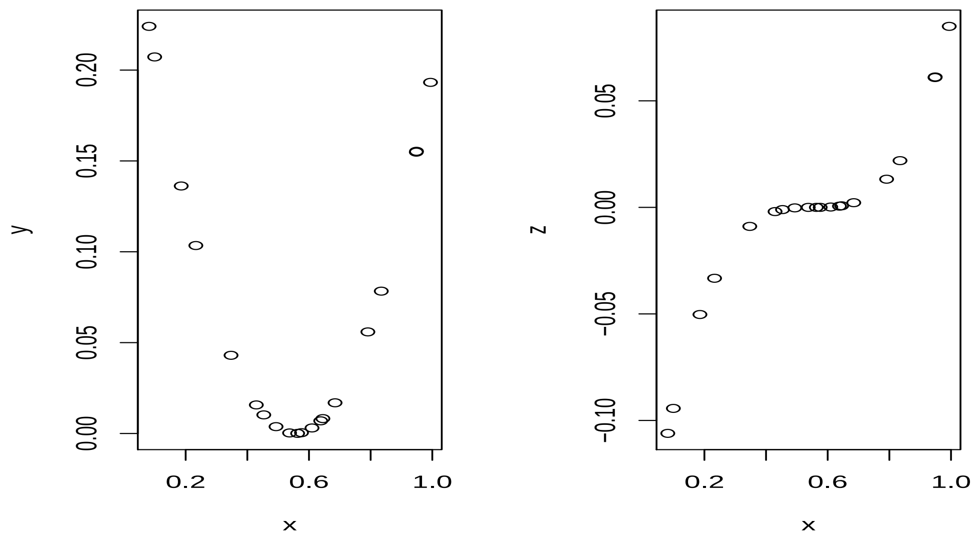
- $-1 \leq R \leq 1$.

- If $R = 1$ $(R = -1)$ then the observations (x_i, y_i) all lie on a straight line with a positive (negative) slope.

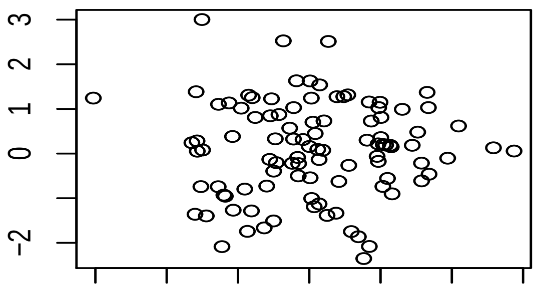
- The sign of r reflects the trend of the points.

- Note x and y can have a very strong *non-linear* relationship and $R \approx 0$.

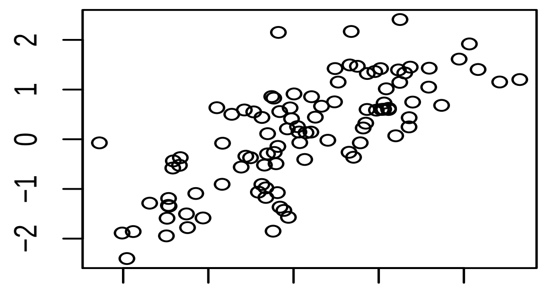
  The graph shows plots of a vector x against $y = (x - \bar{x})^2$ and $z = (x - \bar{x})^3$.



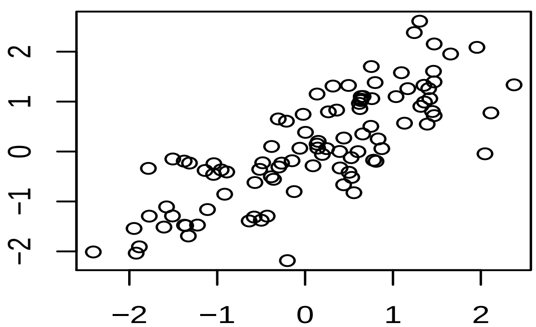
  Correlation for the above data is -0.12 and 0.93, respectively.



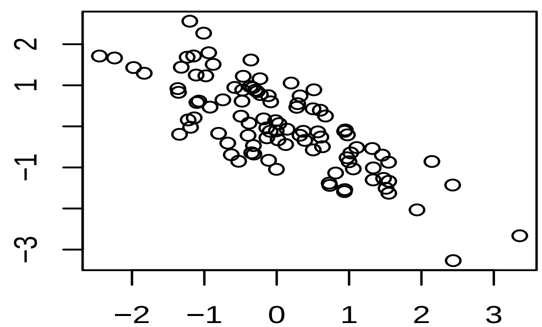
R=0



R=0.74



R=0.78



R=−0.78

Estimating σ^2

Recall that $\mathrm{Var}(\epsilon) = \sigma^2$. To estimate it we use sum of squares of residuals:

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

The estimator $\hat{\sigma}^2$ of σ^2 is given by

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} = \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}. \tag{2}$$

For the oxygen data we have $\hat{\sigma}^2 = 1.18$.

Properties of the LSE

Recall that we consider the model $Y = \beta_0 + \beta_1 x + \epsilon$, where $\mathrm{E}(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = \sigma^2$. Thus, given x, Y is a random variable with mean $\beta_0 + \beta_1 x$ and variance σ^2. Note that $\hat{\beta}_0$ and $\hat{\beta}_1$ depend on the observed y's, which are realizations of the random variable Y. Consequently, the estimators are random variables. We have

$$\mathrm{E}(\hat{\beta}_1) = \beta_1, \ \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}, \ \ \mathrm{E}(\hat{\beta}_0) = \beta_0, \ \mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$$

($\hat{\beta}_0$ and $\hat{\beta}_1$ are the unbiased estimator of β_0 and β_1, respectively). Consequently, the estimated standard errors are

$$\mathrm{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}, \ \mathrm{se}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

Hypothesis tests in linear regression

We want to test that the slope equals $\beta_{1,0}$, i.e.

$$H_0 : \beta_1 = \beta_{1,0},\ H_1 : \beta_1 \neq \beta_{1,0}.$$

Since $\hat{\beta}_1$ is approximately normal $\mathcal{N}(\beta_1, \frac{\sigma^2}{S_{xx}})$, we may use statistics:

$$\frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\sigma^2/S_{xx}}} \sim \mathcal{N}(0,1).$$

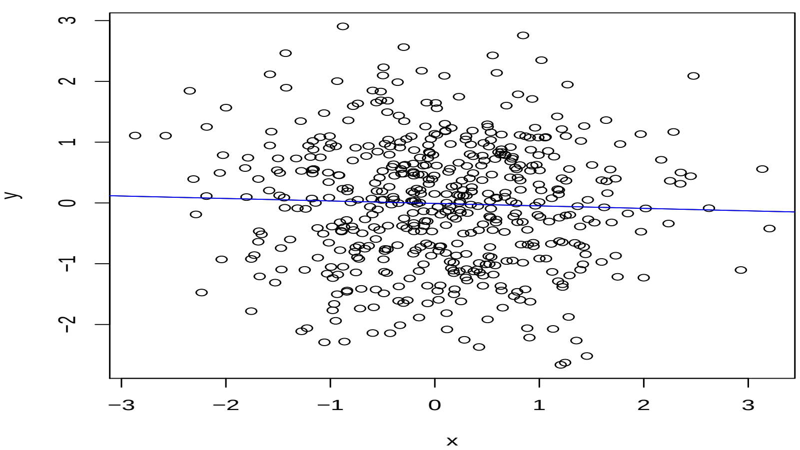
However, σ^2 is not known, thus we plug-in its estimator $\hat{\sigma}^2$ given in (2):

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t_{n-2}.$$

If now t_0 is the observed value, we reject H_0 if $|t_0| > t_{\alpha/2,n-2}$.

Significance of regression

For each bivariate data set we may fit a regression line, which aims to describe a linear relationship between x and Y. However, does it always make sense?



Of course, we may fit the regression line, $\hat{y} = -0.01 - 0.04x$, but this line does not describe at all the bivariate data set.

Having computed the regression line, we want to test whether this line is *significant*. The test for significance of regression is given by

$$H_0 : \beta_1 = 0,\ H_1 : \beta_1 \neq 0.$$

If we reject H_0, then there is a linear relationship between x and Y.

Example: Hydrocarbon data - $\hat{\beta}_1 = 14.95$, $n = 20$, $S_{xx} = 0.68$, $\hat{\sigma}^2 = 1.18$. Consequently,

$$t_0 = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = 11.35 > 2.88 = t_{0.005,18}.$$

We reject H_0 - there is a linear relationship between x and Y.

Confidence intervals - slope and intercept

The $100(1-\alpha)\%$ confidence intervals for β_1 and β_0 are:

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}$$

Example: Hydrocarbon data - $12.181 \leq \beta_1 \leq 17.713$.

Confidence intervals - mean response

We want to estimate $\mu_{Y|x_0} = \mathrm{E}(Y|x_0)$ - the mean response at x_0 (typically at the observed x_0). Of course, it can be read exactly from the regression line

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

The distance (at x_0) between the estimated and the true regression line is

$$\hat{\mu}_{Y|x_0} - \mu_{Y|x_0} = \left(\hat{\beta}_0 - \beta_0\right) + \left(\hat{\beta}_1 - \beta_1\right) x_0.$$

Now, $\mathrm{E}(\hat{\mu}_{Y|x_0}) = \mu_{Y|x_0}$ and

$$\mathrm{Var}(\hat{\mu}_{Y|x_0}) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right].$$

Note that

$$\mathrm{Var}(\hat{\mu}_{Y|x_0}) = \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) \neq \mathrm{Var}(\hat{\beta}_0) + \mathrm{Var}(\hat{\beta}_1 x_0)$$

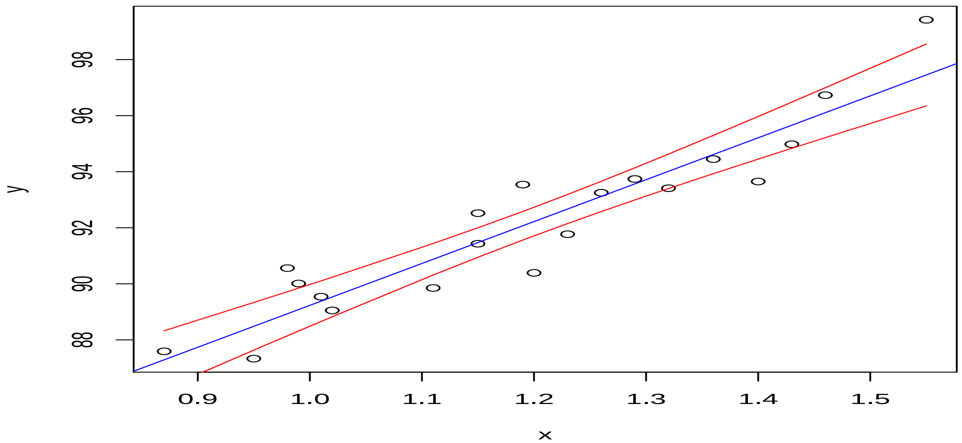
since $\hat{\beta}_0$ and $\hat{\beta}_1$ are dependent.

The confidence interval for $\mu_{Y|x_0}$ (the mean response (regression line)) is

$$\hat{\mu}_{Y|x_0} \pm t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}.$$

Example: Hydrocarbon data -

$$\hat{\mu}_{Y|x_0} \pm 2.101\sqrt{1.18\left[\frac{1}{20} + \frac{(x_0-1.196)^2}{0.68}\right]}$$



A lot of observations are outside the confidence interval (small sample (???)).

Prediction of new observations

If x_0 is the value of the regressor variable of interest, then the estimated value of the response variable Y is

$$\hat{y} = \hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

If Y_0 is the true future observation at $x = x_0$ (so, $Y_0 = \beta_0 + \beta_1 x_0 + \epsilon$) and $\hat{Y}_0$ is the predicted value, given by the above equation, then the prediction error

$$e_{\hat{p}} = Y_0 - \hat{Y}_0 = \beta_0 + \beta_1 x_0 + \epsilon - (\hat{\beta}_0 + \hat{\beta}_1 x_0) = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) x_0 + \epsilon$$

is normally distributed

$$\mathcal{N}\left(0, \sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]\right).$$

Now, we plug-in the estimator of σ to get the following confidence interval for Y_0:

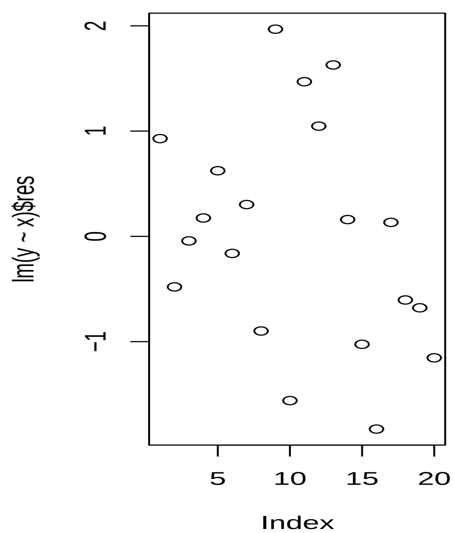
$$\hat{y}_0 \pm t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]},$$

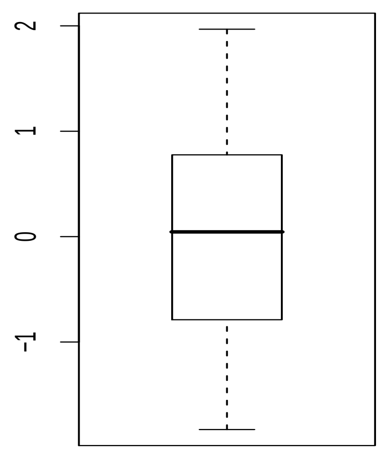
where $\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1 x_0$

Residuals

$$e_i = y_i - \hat{y}_i,$$

where y_i, $i = 1, \ldots, n$ are the observed value and $\hat{y}_i$, $i = 1, \ldots, n$ are the values obtained from the regression line, i.e. $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.





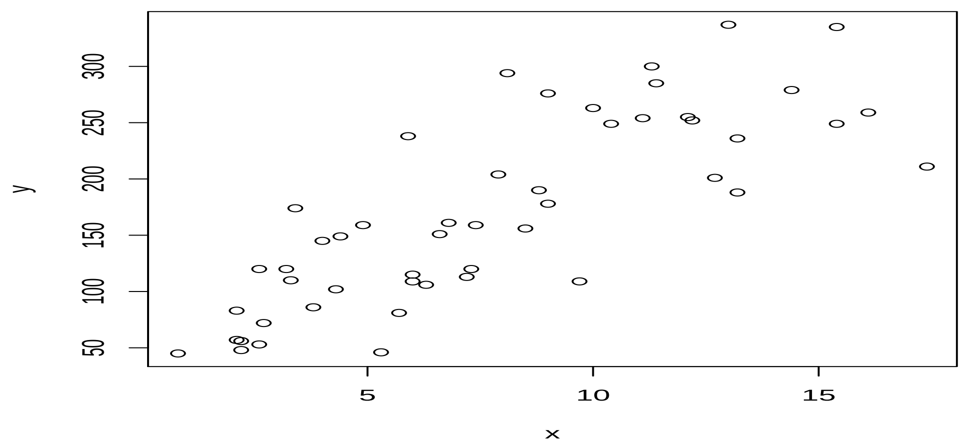
Regression Analysis - summary

1. Draw scatterplot
2. Find the regression line
3. Check the appropriateness of a linear fit (correlation coefficient, significance of regression test)
4. Check *goodness-of-fit* (confidence interval for the regression line)
5. Check model assumptions (residuals)
6. Do prediction, if appropriate

Set #1

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973.

1. x-number of murders, y - number of assaults

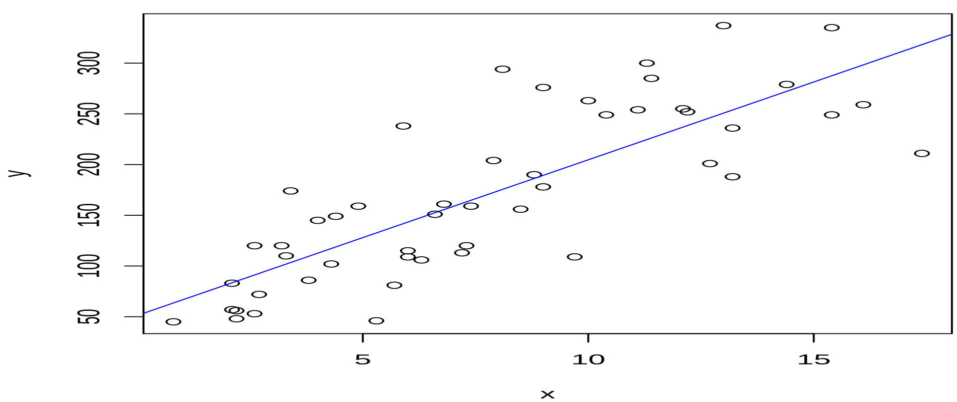


2.

$$\sum_{i=1}^{n} x_i = 389.4, \; \sum_{i=1}^{n} y_i = 8538$$

$$\sum_{i=1}^{n} x_i^2 = 3962.2, \; \sum_{i=1}^{n} y_i^2 = 1798262, \; \sum_{i=1}^{n} x_i y_i = 80756$$

Thus, $\hat{y} = 51.27 + 15.34x$

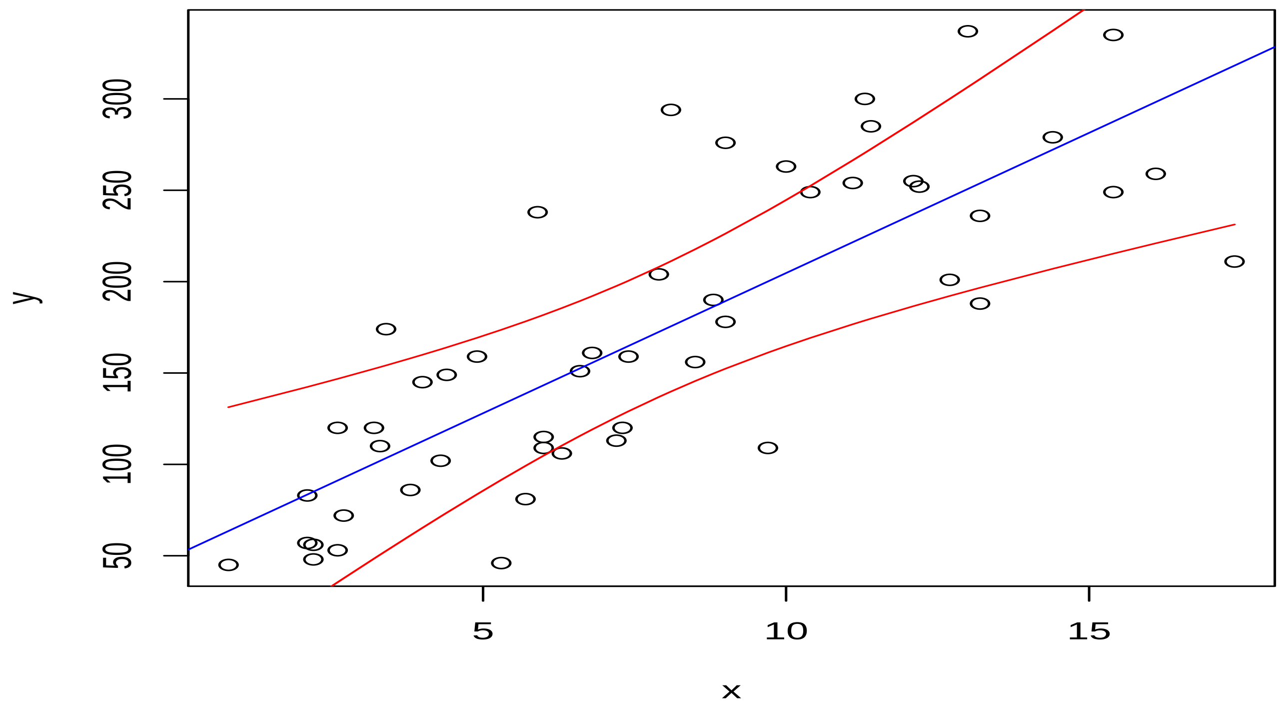


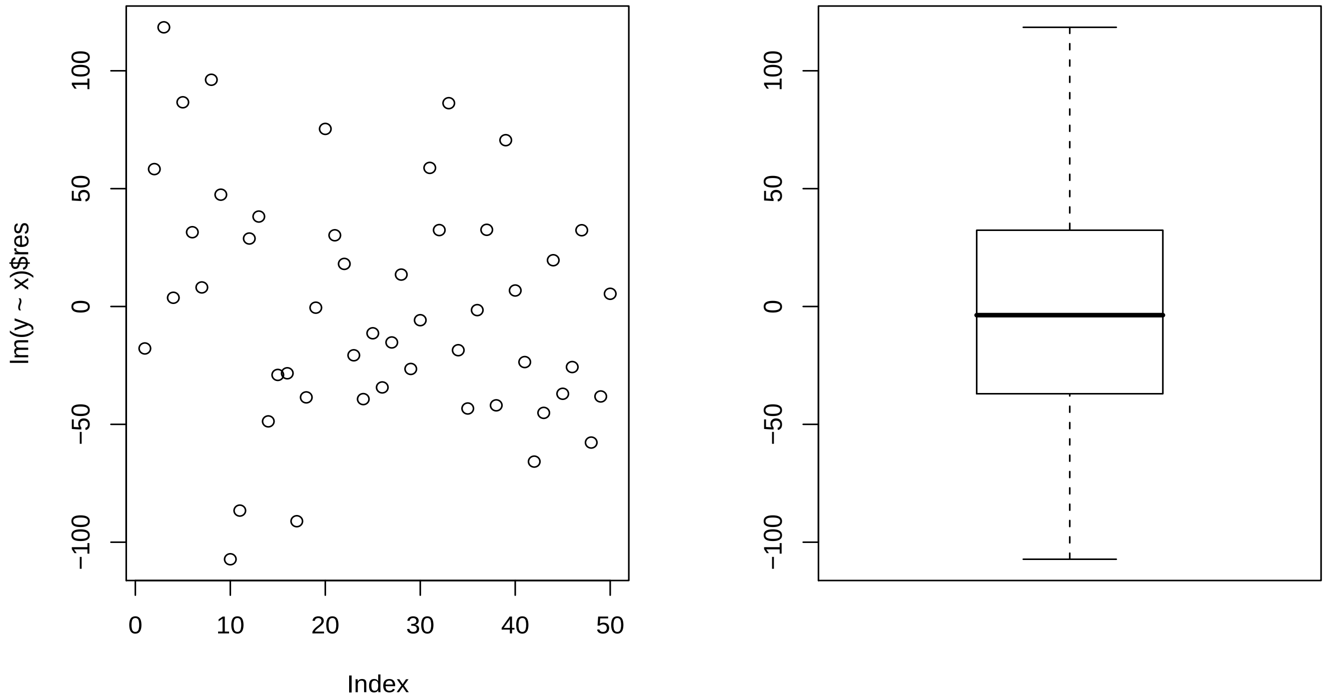
3. Correlation: $R = 0.802$.

Significance of regression: $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$. Test statistics $T_0 = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{n-2}$.. We have $\hat{\sigma}^2 = 2531.73$, $S_{xx} = 929.55$.

The observed value: $t_0 = 9.30$, $t_{0.05/2,48} \approx 2.01$ - reject H_0 - there is a linear relationship between x and y.

4. Confidence interval for the regression line





5.

6. Predict number of assaults if number of murders is $x_0 = 20$:

$$\hat{y}_0 = 51.27 + 15.34 \times 20 = 358.07,$$

Equivalent statement:

- Give a point estimate of number of assaults if number of murders is ...

Compute prediction interval

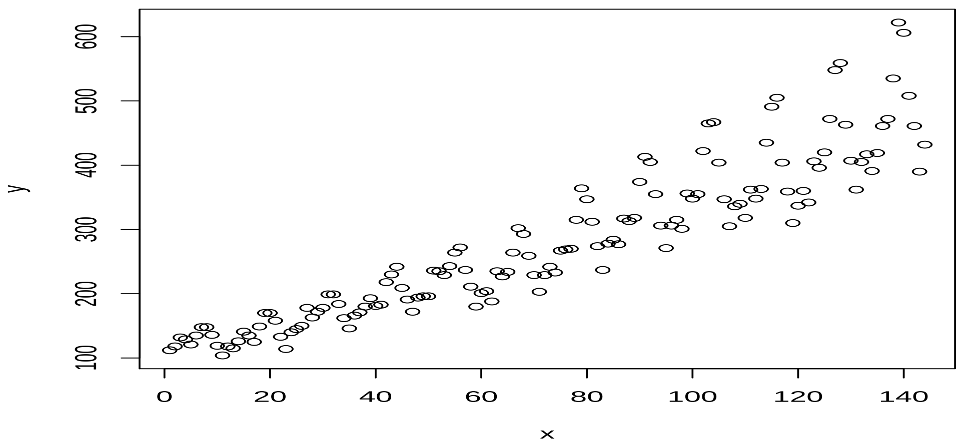
$$
\begin{aligned}
\hat{y}_0 &\pm t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \\
358.07 &\pm 2.01\sqrt{2531.73\left[1+\frac{1}{48}+\frac{(20-7.78)^2}{929.55}\right]} \\
358.07 &\pm 40.64.
\end{aligned}
$$

What change in mean number of assaults would be expected for a change of 1 in the number of assaults? - compute slope

Set #2

The classic airline data. Monthly totals of international airline passengers, 1949 to 1960.

1. x-time, $x = (1, 2, \ldots, 144)$.

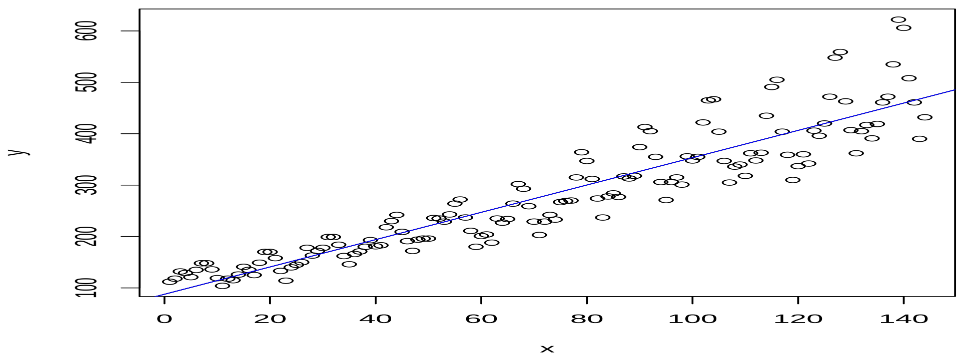


2.

$$\sum_{i=1}^{n} x_i = 10440, \sum_{i=1}^{n} y_i = 40363$$

$$\sum_{i=1}^{n} x_i^2 = 1005720, \sum_{i=1}^{n} y_i^2 = 13371737, \sum_{i=1}^{n} x_i y_i = 3587478$$

Thus, $\hat{y} = 87.653 + 2.657x$

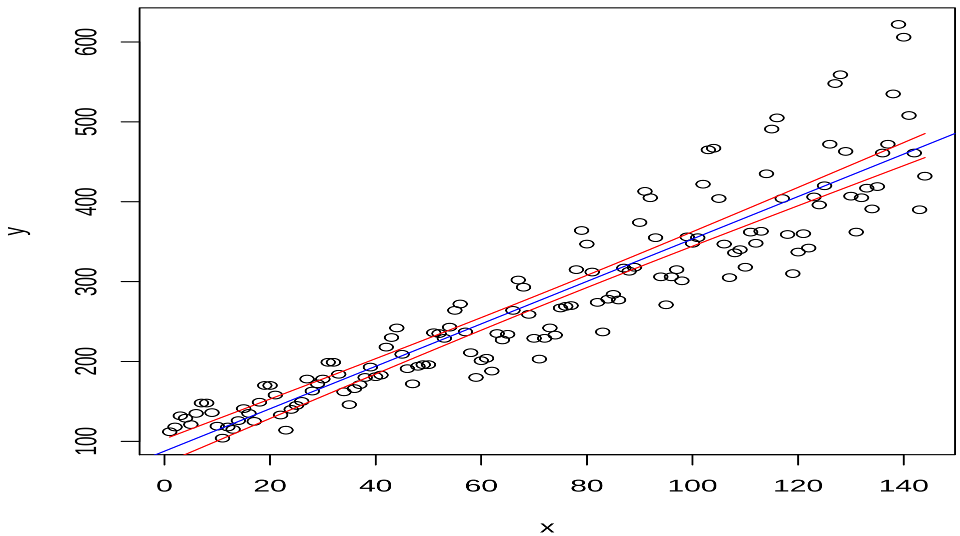


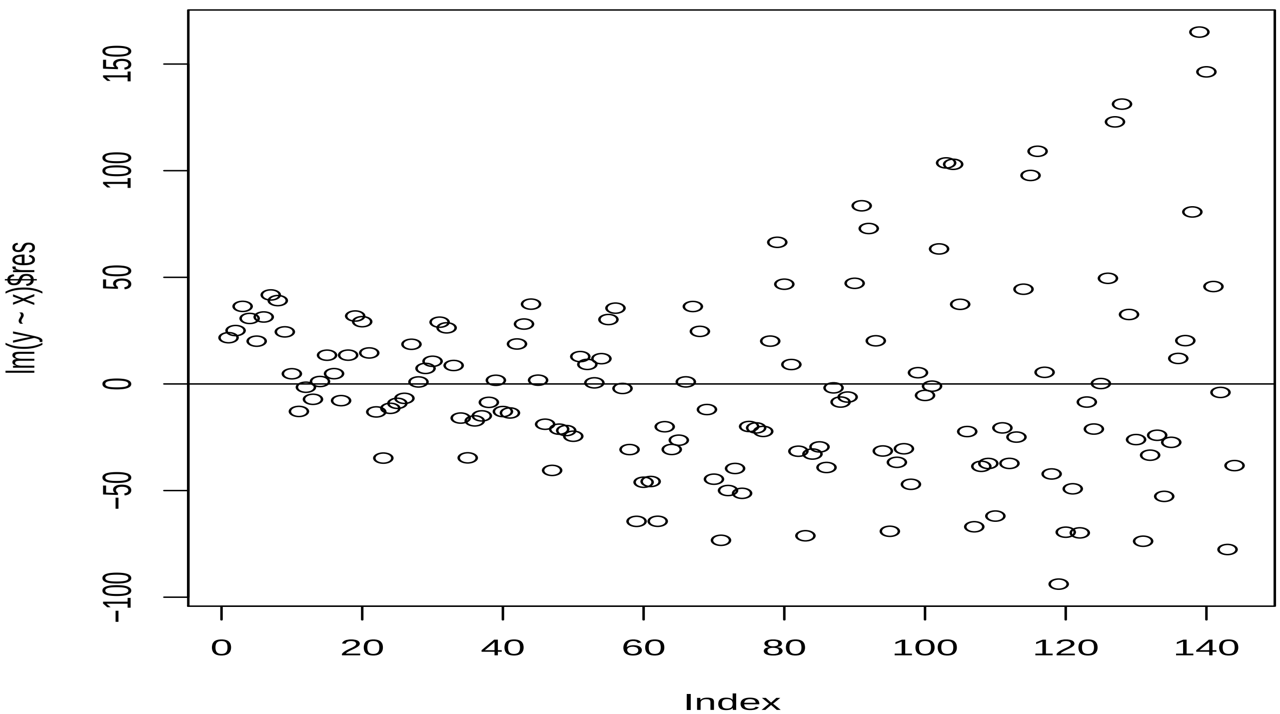
3. Correlation: $R = 0.924$.

Significance of regression: $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$. Test statistics $T_0 = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{n-2}$.. We have $\hat{\sigma}^2 = 2121.261$, $S_{xx} = 248820$.

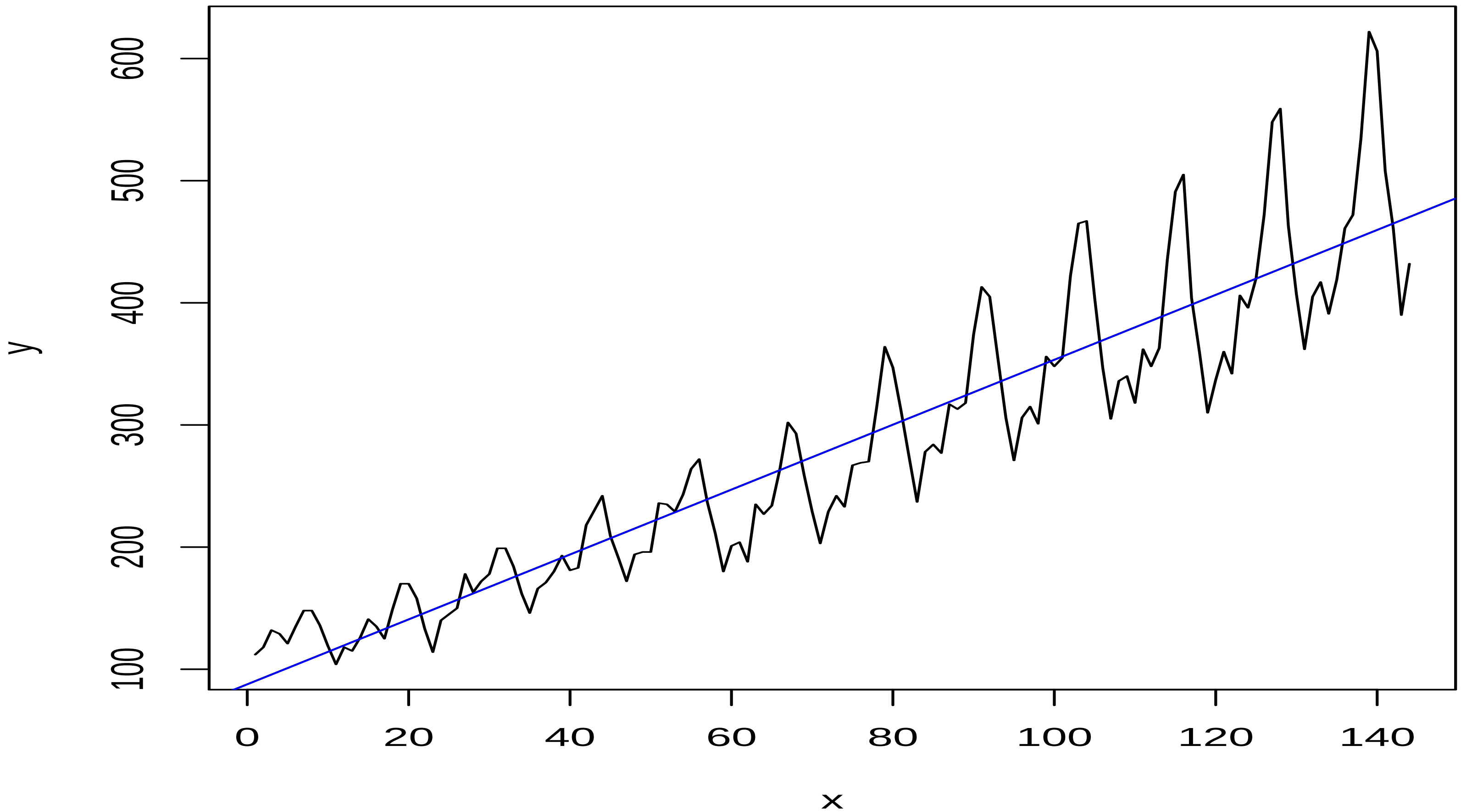
The observed value: $t_0 = 28.77644$, $t_{0.05/2,142} \approx 1.97$ - reject H_0 - there is a linear relationship between x and y.

4. Confidence interval for the reg. line $\hat{\mu}_{Y|x_0} \pm t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$.





5.



y
100
200
300
400
500
600
0
20
40
60
80
100
120
140
x