MAT 2377
Probability and Statistics for Engineers

Chapter 7
Linear Regression and Correlation

P. Boily (uOttawa)

Winter 2021

Based on course notes by Rafał Kulik

Contents

Scenario – Motivation

Consider the following data, consisting of $n = 20$ paired measurements (x_i, y_i) of hydrocarbon levels (x) and pure oxygen levels (y) in fuels:

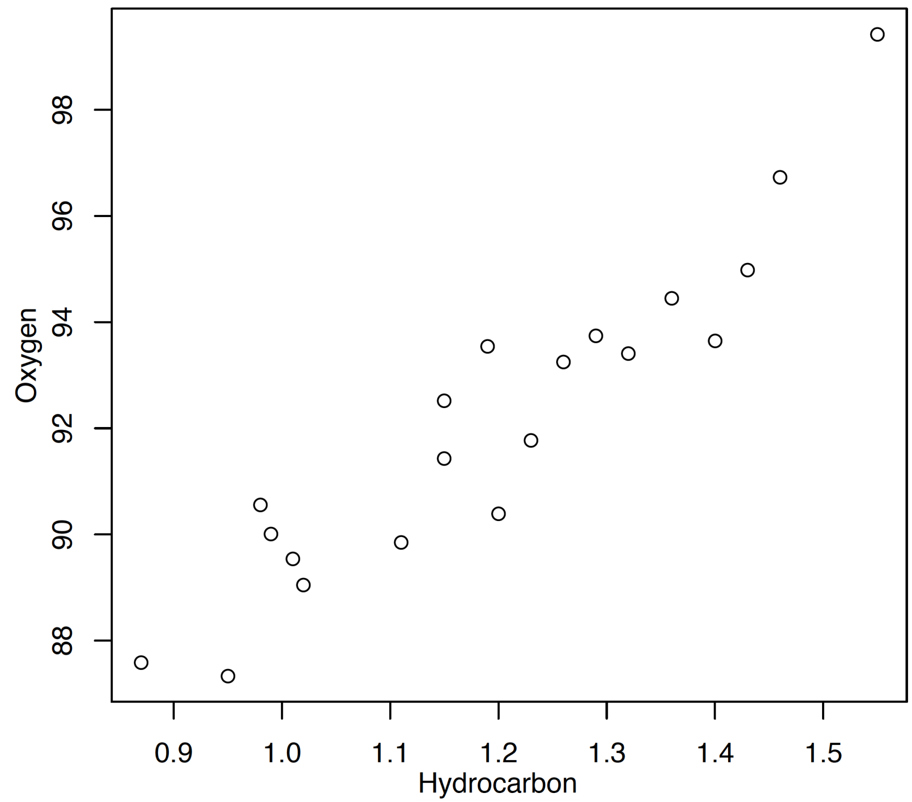
```
x:  0.99  1.02  1.15  1.29  1.46  1.36  0.87  1.23  1.55  1.40
y: 90.01 89.05 91.43 93.74 96.73 94.45 87.59 91.77 99.42 93.65

x:  1.19  1.15  0.98  1.01  1.11  1.20  1.26  1.32  1.43  0.95
y: 93.54 92.52 90.56 89.54 89.85 90.39 93.25 93.41 94.98 87.33
```

Goals:

- measure the **strength of association** between x and y
- **describe** the relationship between x and y

A graphical display provides an initial description of the relationship.



It seems that points lie around a hidden line!

7.1 – Coefficient of Correlation

For paired data (x_i, y_i), $i = 1, \ldots, n$, the **sample correlation coefficient** of x and y is

$$\rho_{XY} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}}.$$

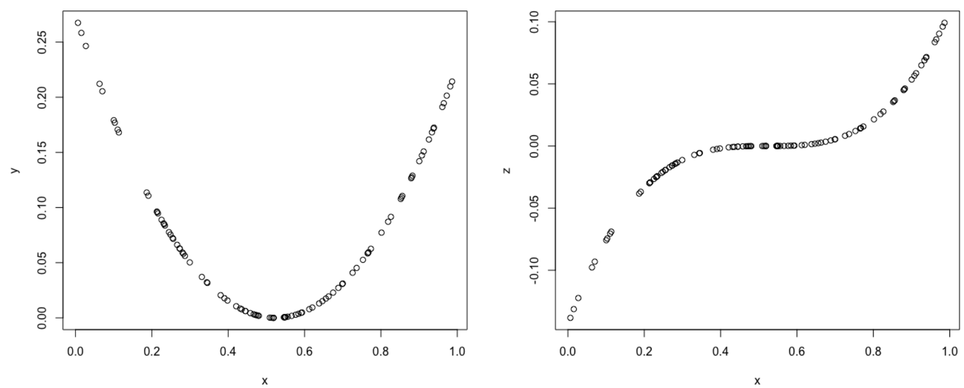
The coefficient ρ_{XY} is defined only if $S_{xx} \neq 0$ and $S_{yy} \neq 0$, i.e. neither x_i nor y_i are constant. The variables x and y are **uncorrelated** if $\rho_{XY} = 0$ (or very small, in practice), and **correlated** if $\rho_{XY} \neq 0$ (or $|\rho_{XY}|$ is "large", in practice).

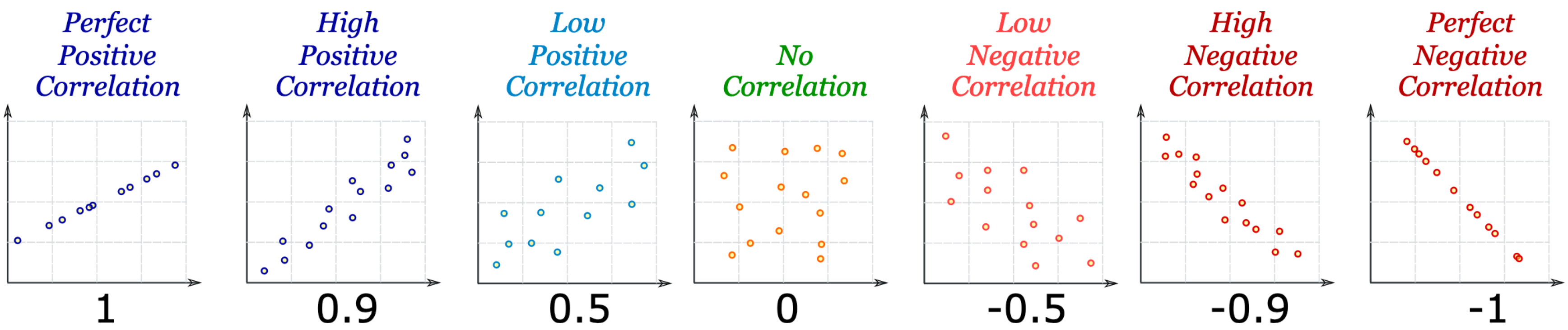
Example: for the data on the previous slide, we have $S_{xy} \approx 10.18$, $S_{xx} \approx 0.68$, $S_{yy} \approx 173.38$, and $\rho_{XY} \approx \frac{10.18}{\sqrt{0.68 \cdot 173.38}} \approx 0.94$.

Properties of ρ_{XY}

- ρ_{XY} is unaffected by changes of scale or origin. Adding constants to x does not change $x - \overline{x}$ and multiplying x and y by constants changes both the numerator and denominator equally;
- ρ_{XY} is symmetric in x and y (i.e. $\rho_{XY} = \rho_{YX}$) and $-1 \leq \rho_{XY} \leq 1$; if $\rho_{XY} = \pm 1$, then the observations (x_i, y_i) all lie on a straight line with a positive (negative) slope;
- the sign of ρ_{XY} reflects the trend of the points;
- a high correlation coefficient value $|\rho_{XY}|$ does not necessarily imply a **causal relationship** between the two variables;

- note that x and y can have a very strong **non-linear** relationship without ρ_{XY} reflecting it (-0.12 on the left, 0.93 on the right).





Computing ρ_{XY} with R

```
> x=c(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=c(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)

> plot(x,y) # will produce the scatterplot on slide 3
> cor(x,y)
    0.9367154

> Sxy=sum((x-mean(x))*(y-mean(y)))
> Sxx=sum((x-mean(x))^2)
> Syy=sum((y-mean(y))^2)
> rho=Sxy/(sqrt(Sxx*Syy))
> rho
    0.9367154
```

7.2 – Simple Linear Regression

Regression analysis can be used to describe the relationship between a **predictor variable** (or regressor) X and a **response variable** Y. Assume that they are related through the model

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where ε is a **random error** and β_0, β_1 are the **regression coefficients**.

It is assumed that $\mathrm{E}[\varepsilon] = 0$, and that the error's variance $\sigma_\varepsilon^2 = \sigma^2$ is constant. Then the model can be re-written as

$$\mathrm{E}[Y|X] = \beta_0 + \beta_1 X.$$

Suppose that we have observations $(x_i, y_i),\ i = 1, \ldots, n$ so that

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

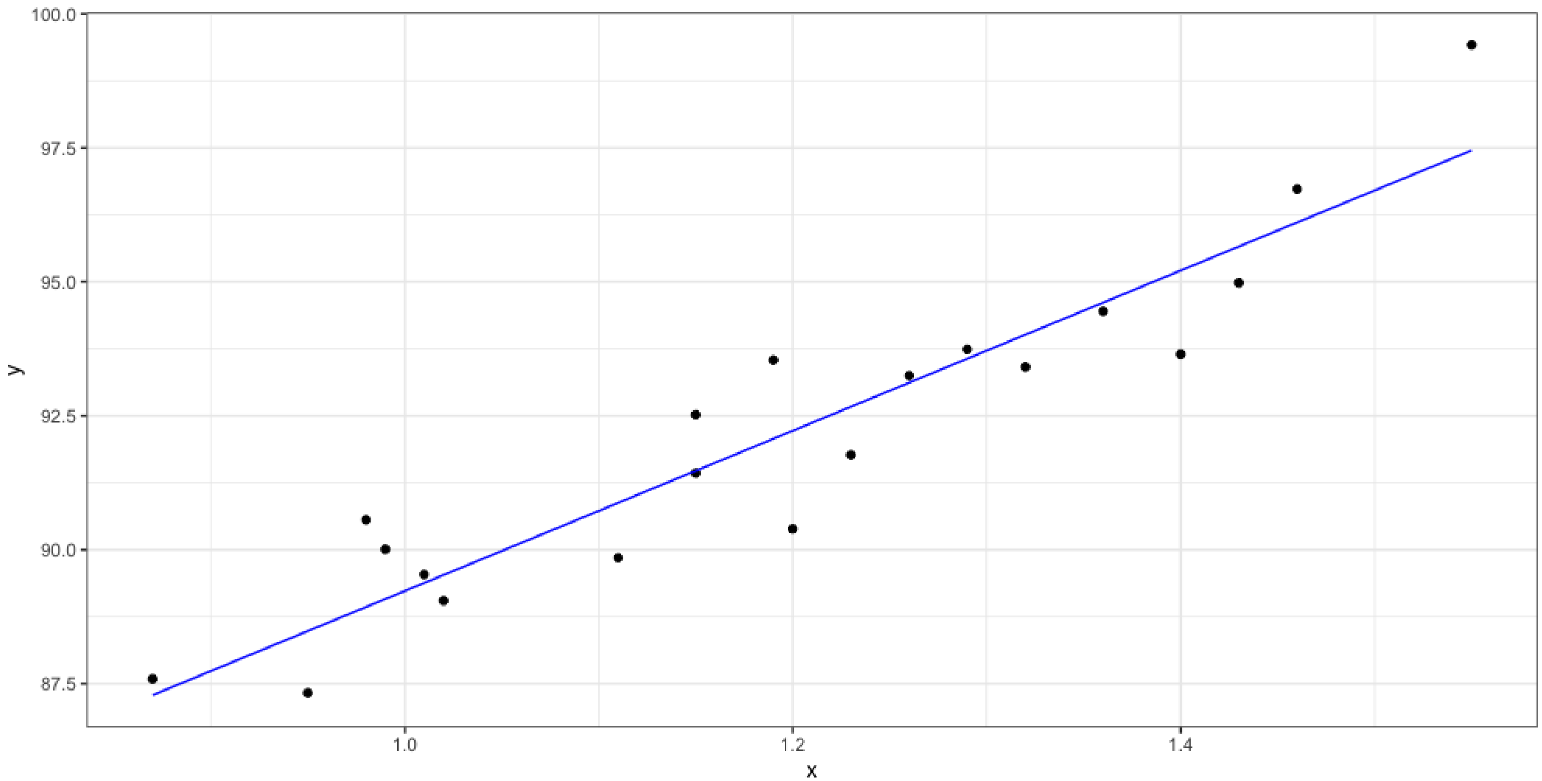
The aim is to find **estimators** b_0, b_1 of the unknown parameters β_0, β_1, in order to obtain the **estimated (fitted) regression line**

$$\hat{y}_i = b_0 + b_1 x_i$$

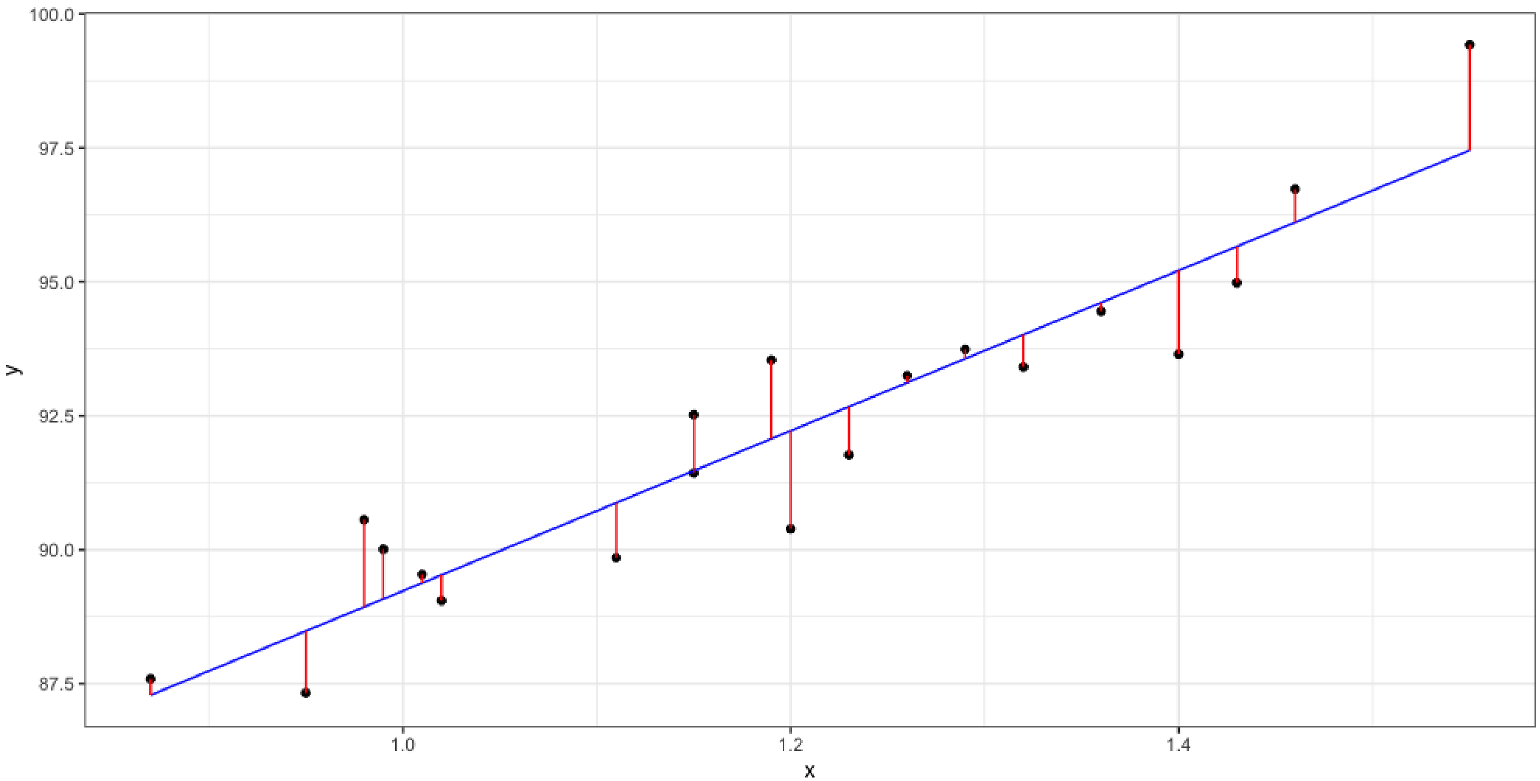
The **residual** or error in predicting y_i using $\hat{y}_i$ is thus

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i, \quad i = 1, \ldots, n.$$

How do we find the estimators? How do we determine if the fitted line is a good model for the data?



fitted line: $\hat{y} = 74.28 + 14.95x$



residuals: $e_i = y_i - \hat{y}_i$

Consider the **Sum of Squared Errors** (SSE):

$$\mathrm{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

(It can be shown that $\mathrm{SSE}/\sigma^2 \sim \chi^2(n-2)$, but that's outside the scope of this course). The optimal values of b_0 and b_1 are those that minimize the SSE. As such, solving

$$0 = \frac{\mathrm{dSSE}}{\mathrm{d}b_0} = -2\sum(y_i - b_0 - b_1 x_i) = -2n(\overline{y} - b_0 - b_1\overline{x})$$

$$0 = \frac{\mathrm{dSSE}}{\mathrm{d}b_1} = -2\sum(y_i - b_0 - b_1 x_i)x_i = -2\left(\sum x_i y_i - n b_0 \overline{x} - b_1 \sum x_i^2\right)$$

yields the **least squares estimators** b_0, b_1 or β_0, β_1, respectively.

From $\frac{\mathrm{dSSE}}{\mathrm{d}b_0} = 0$, we get

$$\overline{y} - b_0 - b_1\overline{x} = 0 \quad \Rightarrow \quad b_0 = \overline{y} - b_1\overline{x}.$$

For the second coefficient, note that

$$\begin{aligned} S_{xy} &= \sum (x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - n\overline{xy} \\ S_{xx} &= \sum (x_i - \overline{x})^2 = \sum x_i^2 - n\overline{x}^2, \end{aligned}$$

which can be re-written as

$$\begin{aligned} \sum x_i y_i &= S_{xy} + n\overline{xy} \\ \sum x_i^2 &= S_{xx} + n\overline{x}^2. \end{aligned}$$

From $\frac{\mathrm{dSSE}}{\mathrm{d}b_1} = 0$, we get

$$\begin{gathered}
\sum x_i y_i - n b_0 \overline{x} - b_1 \sum x_i^2 = 0 \\
(S_{xy} + n\overline{xy}) - n b_0 \overline{x} - b_1 (S_{xx} + n\overline{x}^2) = 0 \\
S_{xy} + n\overline{xy} - n(\overline{y} - b_1 \overline{x})\overline{x} - b_1 S_{xx} - n b_1 \overline{x}^2 = 0 \\
S_{xy} + n\overline{xy} - n\overline{xy} + n b_1 \overline{x}^2 - b_1 S_{xx} - n b_1 \overline{x}^2 = 0 \\
S_{xy} - b_1 S_{xx} = 0 \\
b_1 = \frac{S_{xy}}{S_{xx}}.
\end{gathered}$$

The estimators are also linear combinations of the observed responses y_i:

$$b_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} u_i y_i, \quad b_0 = \overline{y} - b_1 \overline{x} = \sum_{i=1}^{n} v_i y_i.$$

Example: for the fuels data, we've already found that

$$S_{xy} \approx 10.18,\ S_{xx} \approx 0.68,\ \text{and } S_{yy} = 173.38.$$

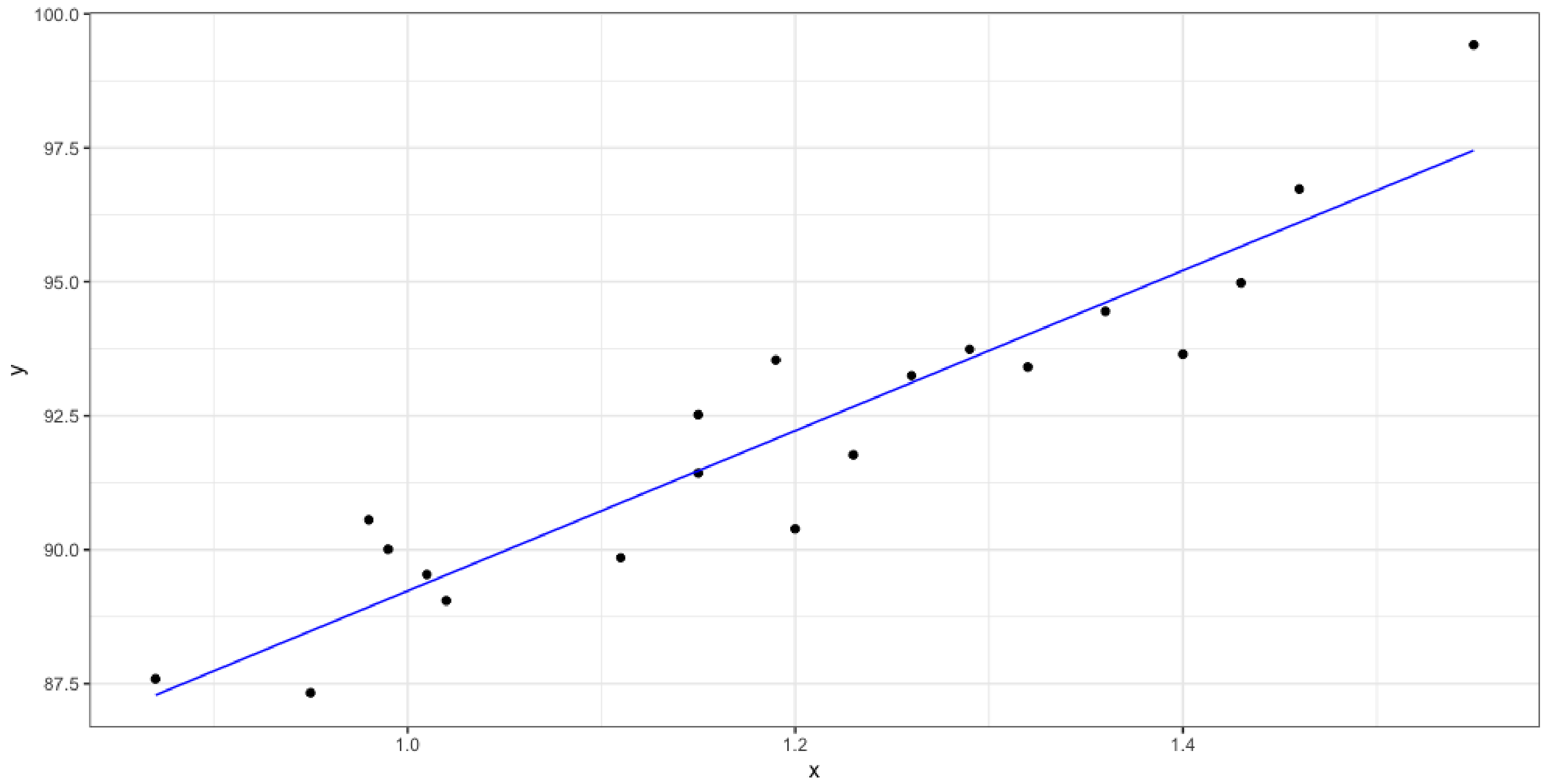
Thus, $b_1 = \frac{10.18}{0.68} = 14.95$. Since

$$n = 20,\ \overline{x} = 1.20,\ \text{and } \overline{y} = 92.16,$$

we also have $b_0 = 92.16 - 20(1.20) = 74.28$.

Consequently, the **fitted regression line** is

$$\hat{y} = 74.28 + 14.95x.$$



fitted line: $\hat{y} = 74.28 + 14.95x$

Estimating σ^2

Recall that the variance of the error term is $\sigma_\varepsilon^2 = \sigma^2$. To estimate σ^2 we use

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

The question is: which denominator should we use?

For a population, we would use n. For a sample, we would use $n-1$. For the regression error, the **unbiased estimator** of σ^2 is in fact

$$\hat{\sigma}^2 = \mathrm{MSE} = \frac{\mathrm{SSE}}{n-2} = \frac{S_{yy} - b_1 S_{xy}}{n-2},$$

where the SSE has $n-2$ **degrees of freedom**, because 2 parameters had to be estimated in order to obtain $\hat{y}_i$: b_0 and b_1.

Example: what is the estimated variance of the noise in the linear model for the fuels data?

Solution: since $S_{xy} \approx 10.18,\ S_{yy} = 173.38,\ b_1 = 14.95,$ and $n = 20$, we have

$$\hat{\sigma}^2 = \frac{173.38 - 14.95(10.18)}{20 - 2} \approx 1.18.$$

The following code shows how to plot the line of best fit, obtain the estimators of β_1, β_2, and extract the **mean squared error** (MSE) in `R`, assuming that `x`, `y`, `Sxx`, and `Sxy` have been assigned/computed in a previous step.

```
> library(ggplot2)   ### for line of best fit, residual plots
> fuels=data.frame(x,y)
> model <- lm(y ~ x, data=fuels)   ### R function for linear regression
```

```
> summary(model)   ### we will explain this output later
    Call: lm(formula = y ~ x, data = fuels)

    Residuals:
         Min       1Q   Median       3Q      Max
    -1.83029 -0.73334  0.04497  0.69969  1.96809

    Coefficients:
                Estimate Std. Error t value Pr(>|t|)
    (Intercept)   74.283      1.593   46.62  < 2e-16 ***
    x             14.947      1.317   11.35 1.23e-09 ***
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Residual standard error: 1.087 on 18 degrees of freedom
    Multiple R-squared: 0.8774, Adjusted R-squared: 0.8706
    F-statistic: 128.9 on 1 and 18 DF,  p-value: 1.227e-09
```

```
> ggplot(model) + geom_point(aes(x=x, y=y)) +   ### plotting line of best fit
    geom_line(aes(x=x, y=.fitted), color="blue" ) +
    theme_bw()

> ggplot(model) + geom_point(aes(x=x, y=y)) +   ### plotting residuals
    geom_line(aes(x=x, y=.fitted), color="blue" ) +
    geom_linerange(aes(x=x, ymin=.fitted, ymax=y), color="red") +
    theme_bw()

> n=length(x)
> sigma2 = (Syy-as.numeric(model$coefficients[2])*Sxy)/(n-2)   ### directly
> sigma2
    1.180545
> summary(model)$sigma^2   ### getting the MSE from the summary
    1.180545
```

Properties of the Least Square Estimators

Recall that the simple linear regression model is

$$Y = \beta_0 + \beta_1 X + \varepsilon, \text{ with } \mathrm{E}[\varepsilon] = 0, \sigma_\varepsilon^2 = \sigma^2.$$

Given X, Y is a random variable with mean $\beta_0 + \beta_1 X$ and variance σ^2:

$$\mathrm{E}[Y|X] = \beta_0 + \beta_1 X, \quad \mathrm{Var}[Y|X] = \sigma^2.$$

Note that b_0 and b_1 depend on the observed x's and y's, which are realizations of the random variables X and Y. As a result, the **estimators are random variables**, that is to say: different realizations (observed data) lead to different estimates b_0, b_1 for β_0, β_1.

It can be shown that

$$\mathrm{E}[b_0] = \beta_0, \qquad \sigma^2_{b_0} = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right] = \sigma^2 \frac{\sum_{i=1}^n x_i^2}{nS_{xx}},$$

$$\mathrm{E}[b_1] = \beta_1, \qquad \sigma^2_{b_1} = \sigma^2/S_{xx}.$$

We say that b_0 and b_1 are **unbiased estimators** of β_0 and β_1, respectively. The **estimated standard errors** (replacing σ^2 by $\mathrm{MSE} = \hat{\sigma}^2$ in the expressions for $\sigma^2_{b_1}$ and $\sigma^2_{b_0}$ above) are

$$\mathrm{se}(b_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]} \quad \text{and} \quad \mathrm{se}(b_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}.$$

Example: find the estimated standard error for b_0 and b_1 in the fuels data.

Solution: we have $n = 20$, $\overline{x} = 1.20$, $S_{xx} = 0.68$, and $\hat{\sigma}^2 = 1.18$, so that

$$\mathrm{se}(b_0) = \sqrt{1.18\left[\frac{1}{20} + \frac{1.20^2}{0.68}\right]} \approx 1.593 \quad \text{and} \quad \mathrm{se}(b_1) = \sqrt{\frac{1.18}{0.68}} \approx 1.317.$$

This information is also available in the `R` output:

```
> summary(model)$coefficients

            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 74.28331   1.593473 46.61723 3.171476e-20
x           14.94748   1.316758 11.35173 1.227314e-09
```

7.3 – Hypothesis Testing for Linear Regression

With standard errors, we can **test hypotheses** on the regression parameters.

We try to determine if the true parameters β_0, β_1 take on specific values, and whether the line of best fit describes a bivariate dataset well.

The steps are the same as in Chapter 6:

1. set up a null hypothesis H_0 and an alternative hypothesis H_1
2. compute a test statistic (often by some form of standardizing)
3. find a critical region/$p-$value for the test statistic under H_0
4. reject or fail to reject H_0 based on the critical region/$p-$value

Hypothesis Test for the Intercept β_0

We might be interested in testing whether the true intercept β_0 is equal to some **candidate value** $\beta_{0,0}$, i.e.

$$H_0 : \beta_0 = \beta_{0,0} \text{ against } H_1 : \beta_0 \neq \beta_{0,0}.$$

The linear regression model requires normal errors $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, which implies that $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$, $i = 1, \dots, n$.

Since b_0 is a linear function of the observed responses y_i, it has normal distribution with mean β_0 and variance $\sigma^2 \frac{\sum x_i^2}{nS_{xx}}$. Therefore, under H_0,

$$Z_0 = \frac{b_0 - \beta_{0,0}}{\sqrt{\sigma^2 \frac{\sum x_i^2}{nS_{xx}}}} \sim \mathcal{N}(0, 1).$$

But σ^2 is not known, so the test statistic with $\hat{\sigma}^2 = \mathrm{MSE}$

$$T_0 = \frac{b_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2 \frac{\sum x_i^2}{n S_{xx}}}} \sim t(n-2)$$

follows a Student $t-$distribution with $n-2$ degrees of freedom.

Alternative Hypothesis	**Critical/Rejection Region**
$H_1 : \beta_0 > \beta_{0,0}$	$t_0 > t_\alpha(n-2)$
$H_1 : \beta_0 < \beta_{0,0}$	$t_0 < -t_\alpha(n-2)$
$H_1 : \beta_0 \neq \beta_{0,0}$	$\lvert t_0 \rvert > t_{\alpha/2}(n-2)$

where t_0 is the observed value of T_0 and $t_\alpha(n-2)$ is the $t-$value satisfying $P(T > t_\alpha(n-2)) = \alpha$, and $T \sim t(n-2)$.

Reject H_0 if t_0 in the critical region.

Hypothesis Test for the Slope β_1

We might be interested in testing whether the true slope β_1 is equal to some **candidate value** $\beta_{1,0}$, i.e.

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

The linear regression model requires normal errors $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, which implies that $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$, $i = 1, \ldots, n$.

Since b_1 is a linear function of the observed responses y_i, it has normal distribution with mean β_1 and variance $\frac{\sigma^2}{S_{xx}}$. Therefore, under H_0,

$$Z_0 a = \frac{b_1 - \beta_{1,0}}{\sqrt{\sigma^2 / S_{xx}}} \sim \mathcal{N}(0, 1).$$

But σ^2 is not known, so the test statistic with $\hat{\sigma}^2 = \mathrm{MSE}$

$$T_0 = \frac{b_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t(n-2)$$

follows a Student $t-$distribution with $n-2$ degrees of freedom.

Alternative Hypothesis	**Critical/Rejection Region**
$H_1 : \beta_1 > \beta_{1,0}$	$t_0 > t_\alpha(n-2)$
$H_1 : \beta_1 < \beta_{1,0}$	$t_0 < -t_\alpha(n-2)$
$H_1 : \beta_1 \neq \beta_{1,0}$	$\lvert t_0 \rvert > t_{\alpha/2}(n-2)$

where t_0 is the observed value of T_0 and $t_\alpha(n-2)$ is the $t-$value satisfying $P(T > t_\alpha(n-2)) = \alpha$, and $T \sim t(n-2)$.

Reject H_0 if t_0 in the critical region.

Examples: use the fuels dataset and assume the quantities/models (`n`, `sigma2`, `Sxx`, `x`, `model`) have been assigned/computed in a previous step.

a) Test for $H_0 : \beta_0 = 75$ against $H_1 : \beta_0 < 75$ at $\alpha = 0.05$.

b) Test for $H_0 : \beta_1 = 10$ against $H_1 : \beta_1 > 10$ at $\alpha = 0.05$.

c) Test for $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at $\alpha = 0.05$.

Solution: the following code shows that we fail to reject H_0 for a), but that we reject H_0 in favour of H_1 for b) and c).

```
> b0 = as.numeric(model$coefficients[1])  ### LS parameters
> b1 = as.numeric(model$coefficients[2])  ### LS parmeters
> beta00 = 75  ### for a)
> beta10 = 10  ### for b)
```

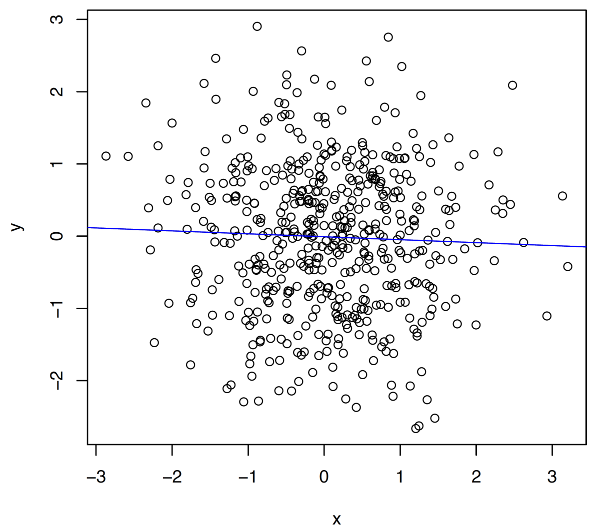
```
# a)
> t0a = (b0-beta00)/sqrt(sigma2*sum(x^2)/n/Sxx)  ### test statistic
> crit_t005_18a = qt(0.05,n-2)                   ### critical value
> t0a < crit_t005_18a                            ### test for critical region
    FALSE                                        ### fail to reject H0

# b)
> t0b = (b1-beta10)/sqrt(sigma2/Sxx)  ### test statistic
> crit_t005_18b = - qt(0.05,n-2)      ### critical value
> t0b > crit_t005_18b                 ### test for critical region
    TRUE                              ### reject H0

# c)
> t0c = b1/sqrt(sigma2/Sxx)           ### test statistic
> crit_t0025_18c = - qt(0.025,18)     ### critical value
> abs(t0c) > crit_t0025_18c           ### test for critical region
    TRUE                              ### reject H0
```

Significance of Regression

As long as $S_{xx} \neq 0$ (at least two distinct values of X in the data), we can fit a regression line to the observations using the **least squares framework**. Recall that one of the goals of linear regression is to **describe a linear relationship** between X and Y... if one exists.



The regression line for the dataset on the previous slide is

$$\hat{y} = -0.01 - 0.04x,$$

but this line does not describe the bivariate data set at all, which is more like a diffuse blob. The relationship between X and Y in that dataset is simply not linear.

Given a regression line, we may want to test whether it is **significant**. The test for **significance of the regression** is

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0.$$

If we reject H_0 in favour of H_1, then the evidence suggests that there is a linear relationship between X and Y.

Example: in the fuels dataset, we have $b_1 = 14.95$, $n = 20$, $S_{xx} = 0.68$, $\hat{\sigma}^2 = 1.18$. We test for significance of the regression at $\alpha = 0.01$:

$$H_0 : \beta_1 = 0 \ , \quad \text{against} \quad H_1 : \beta_1 \neq 0.$$

Since the observed value of the test statistic is

$$t_0 = \frac{b_1 - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = 11.35 > 2.88 = t_{0.01/2}(18) \, ,$$

where $t_{0.01/2}(18)$ is the critical value of Student's $t-$distribution with 18 degrees of freedom at $\alpha = 0.01$ for two-sided tests, we reject H_0 and conclude that there is a linear relationship between X and Y (at $\alpha = 0.01$).

(Use `-qt(0.01/2,18)` to get the critical value in `R`.)

7.4 – Confidence and Prediction Intervals for Linear Regression

We can also build **confidence intervals** (C.I.) for the regression parameters and **prediction intervals** (P.I.) for the predicted values.

The steps are the same as in Chapter 5:

1. find a point estimate W for the parameter β or the prediction Y
2. find the appropriate standard error $\mathrm{se}(W)$
3. select a confidence level α and find the appropriate critical value $k_{\alpha/2}$
4. build the $100(1-\alpha)\%$ interval $W \pm k_{\alpha/2} \cdot \mathrm{se}(W)$

C.I. for the Intercept β_0 and the Slope β_1

Since we estimate the error variance with $\hat{\sigma}^2 = \mathrm{MSE}$, we need to use Student's $t-$distribution with $n-2$ degrees of freedom (remember that we use the data to estimate 2 parameters).

The $100(1-\alpha)\%$ C.I. for β_0 and β_1 are:

$$\beta_0 : \qquad b_0 \pm t_{\alpha/2}(n-2)\mathrm{se}(b_0) = b_0 \pm t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2\frac{\sum x_i^2}{nS_{xx}}}$$

$$\beta_1 : \qquad b_1 \pm t_{\alpha/2}(n-2)\mathrm{se}(b_1) = b_1 \pm t_{\alpha/2}(n-2)\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

The caveat regarding the interpretation of confidence intervals still applies.

Example: build 95% and 99% C.I. for β_0 and β_1 in the fuels data.

Solution: from previous examples, we have $b_0 = 74.283$, $b_1 = 14.947$, $\mathrm{se}(b_0) = 1.593$, $\mathrm{se}(b_1) = 1.317$, $t_{0.025}(18) = 2.10$ and $t_{0.005}(18) = 2.88$.

Then, for $\alpha = 0.05$, we have

$$\beta_0 : \; 74.283 \pm 2.10(1.593) = (70.93, 77.63)$$
$$\beta_1 : \; 14.497 \pm 2.10(1.317) = (12.18, 17.71)$$

and for $\alpha = 0.01$, we have

$$\beta_0 : \; 74.283 \pm 2.88(1.593) = (69.70, 78.87)$$
$$\beta_1 : \; 14.497 \pm 2.88(1.317) = (11.15, 18.74).$$

Confidence Intervals for the Mean Response

We might also be interested in estimating $\mu_{Y|x_0} = \mathrm{E}[Y|x_0]$, the **mean response** at an observed x_0 (in practice, there could be more than one response at the predictor, due to replication in an experiment, say).

The predicted value can be read directly from the regression line:

$$\hat{\mu}_{Y|x_0} = b_0 + b_1 x_0.$$

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

$$\hat{\mu}_{Y|x_0} - \mu_{Y|x_0} = (b_0 - \beta_0) + (b_1 - \beta_1)\, x_0.$$

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

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

Note that

$$\mathrm{Var}[\hat{\mu}_{Y|x_0}] = \mathrm{Var}[b_0 + b_1 x_0] \neq \mathrm{Var}[b_0] + \mathrm{Var}[b_1 x_0]$$

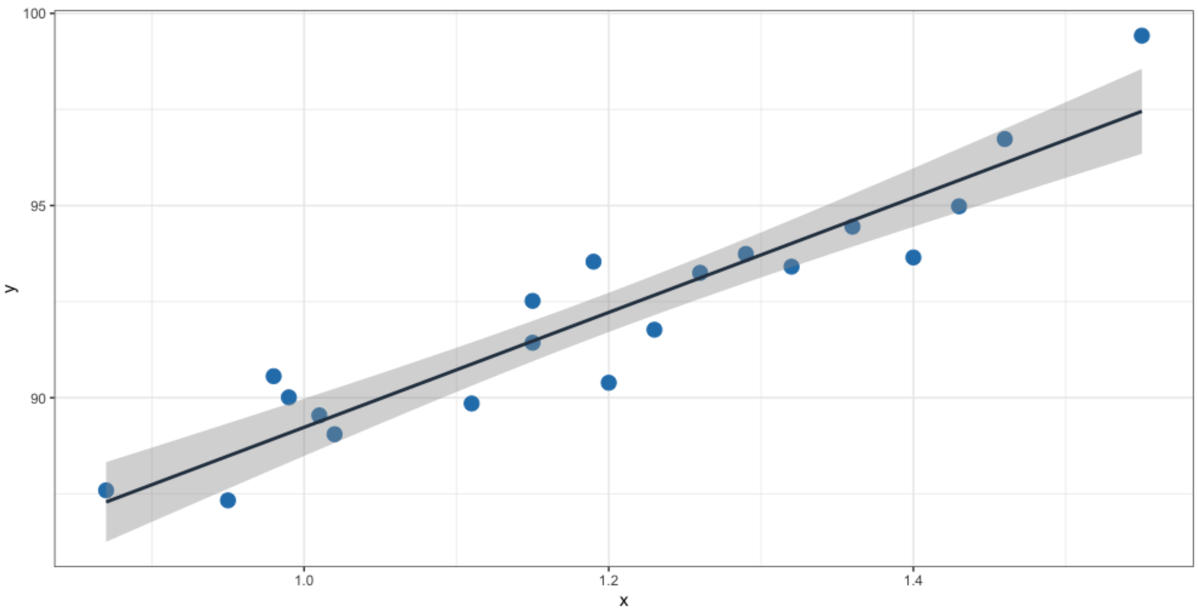
since b_0 and b_1 are dependent.

With the usual $t_{\alpha/2}(n-2)$, the $100(1-\alpha)\%$ C.I. for the **mean response** $\mu_{Y|x_0}$ (or for the line of regression) is

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

Example: for the fuels dataset, the 95% C.I. for $\mu_{Y|x_0}$ is

$$74.28 + 14.95x_0 \pm 2.10\sqrt{1.18\left[\frac{1}{20} + \frac{(x_0 - 1.12)^2}{0.68}\right]}.$$



A fair number of the observations are found outside the 95% C.I. for the mean response, potentially because of the relatively small sample size.

The R code to produce this chart is shown below:

```
> ggplot(fuels, aes(x=x, y=y)) +
    geom_point(color='#2980B9', size = 4) +
    geom_smooth(method=lm, color='#2C3E50') +
    theme_bw()
```

Predicting New Observations

If x_0 is the value of interest for the regressor (predictor), then the estimated value of the response variable Y is

$$\hat{y} = \hat{Y}_0 = b_0 + b_1 x_0.$$

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

$$e_{\hat{p}} = Y_0 - \hat{Y}_0 = \beta_0 + \beta_1 x_0 + \varepsilon - (b_0 + b_1 x_0) = (\beta_0 - b_0) + (\beta_1 - b_1)x_0 + \varepsilon$$

has normal distribution with zero mean and variance $\sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]$.

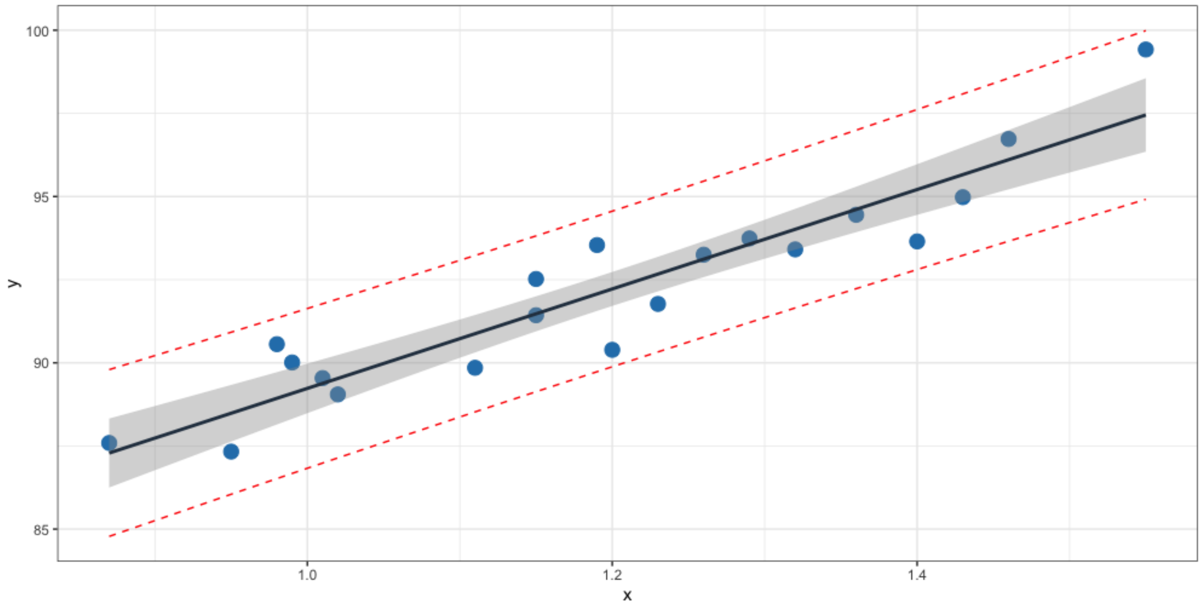
Substitute σ^2 by its estimator $\hat{\sigma}^2 = \mathrm{MSE}$ and we get a $100(1-\alpha)\%$ **prediction interval** for Y_0:

$$b_0 + b_1 x_0 \pm t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]},$$

where $t_{\alpha/2}$ is the critical value of Student's $t-$distribution with $n-2$ degrees of freedom at α.

Example: for the fuels dataset, the 95% P.I. for $\mu_{Y|x_0}$ is

$$74.28 + 14.95x_0 \pm 2.10\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(x_0 - 1.12)^2}{0.68}\right]}.$$



None of the observations are found outside the 95% P.I. for new observations. In general, for a given α, the prediction interval is wider than the confidence interval, which is not surprising: the CLT implies that the mean response has a smaller variance than the predicted responses.

The R code that produces the chart on the previous slide is

```
## build P.I. for various regressors
> preds <- predict(model, interval="prediction")

## put data in a new dataframe
> new.fuels <- cbind(fuels, preds)

> ggplot(new.fuels, aes(x=x, y=y)) +
    geom_point(color='#2980B9', size = 4) +
    geom_smooth(method=lm, color='#2C3E50') +
    geom_line(aes(y=lwr), color = "red", linetype = "dashed") +
    geom_line(aes(y=upr), color = "red", linetype = "dashed") +
    theme_bw()
```

7.5 – Analysis of Variance

The test for **significance of regression**,

$$H_0 : \beta_1 = 0 \ \text{ against } \ H_1 : \beta_1 \neq 0,$$

can be restated in term of the **analysis-of-variance** (ANOVA), given by the following table:

Source of Variation	**Sum of Squares**	**df**	**Mean Square**	F^*	$p-$**Value**
Regression	SSR	1	MSR	$\frac{\mathrm{MSR}}{\mathrm{MSE}}$	$P(F > F^*)$
Error	SSE	$n-2$	MSE		
Total	SST	$n-1$			

In this table, the $F-$statistic $F^* \sim F(1, n-2)$, and

$$\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad \mathrm{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2, \quad \mathrm{SST} = \sum_{i=1}^{n}(y_i - \overline{y})^2,$$

$$\mathrm{MSR} = \frac{\mathrm{SSR}}{1}, \quad \mathrm{MSE} = \frac{\mathrm{SSE}}{n-2}, \quad \text{and} \quad F^* = \frac{\mathrm{MSR}}{\mathrm{MSE}} = \frac{\mathrm{SSR}/1}{\mathrm{SSE}/n-2}$$

The **rejection region** for the null hypothesis $H_0 : \beta_1 = 0$ is still given by

$$\left| \frac{b_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2/S_{xx}}} \right| > t_{\alpha/2}(n-2),$$

but it can also be written as $F^* > f_\alpha(1, n-2)$, where $f_\alpha(1, n-2)$ is the critical $F-$value of the F-distribution with $\nu_1 = 1$ and $\nu_2 = n-2$ df.

Example: the $F-$statistic can be found in the output of the linear regression summary in `R`. For the fuels dataset, it is:

```
Residual standard error: 1.087 on 18 degrees of freedom
Multiple R-squared: 0.8774, Adjusted R-squared: 0.8706
F-statistic: 128.9 on 1 and 18 DF,  p-value: 1.227e-09
```

The critical value for $\alpha = 0.05$ is $f_{0.05}(1, 18) =$ `qf(0.95,1,18)` $= 4.41$.

Since

$$F^* = 128.9 > f_{0.05}(1, 18) = 4.4,$$

we reject the null hypothesis H_0 in favour of the regression being significant at $\alpha = 0.05$.

7.6 – Coefficient of Determination

For observations (x_i, y_i), $i = 1, \ldots, n$, we define the **coefficient of determination** as

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}},$$

where SSE and SST are as in the ANOVA.

The coefficient of determination is the proportion of the variability in the response that is explained by the fitted model. Note that R^2 always lies between 0 and 1; when $R^2 \approx 1$, the fit is considered to be very good.

BE CAREFUL: in practice, R^2 is not always the best way to determine the **goodness-of-fit** of the regression. There are factors (such as the number of observations) which can affect the coefficient of determination.

Example: the coefficient of determination R^2 statistic can be found in the output of the linear regression summary in `R`. For the fuels dataset, it is:

```
Residual standard error: 1.087 on 18 degrees of freedom
Multiple R-squared: 0.8774, Adjusted R-squared: 0.8706
F-statistic: 128.9 on 1 and 18 DF,  p-value: 1.227e-09
```

At $R^2 = 0.8774$, about 88% of the variability in the response Y can be explained by line of best fit.

Appendix – Summary of Regression Analysis

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, or confidence interval for the regression line
5. Check model assumptions (residuals)
6. Offer predictions, if appropriate

Example: US Arrests

This dataset US Arrests contains statistics, in arrests per $100,000$ residents about various crimes in 1973, for each of the $n = 50$ US states.

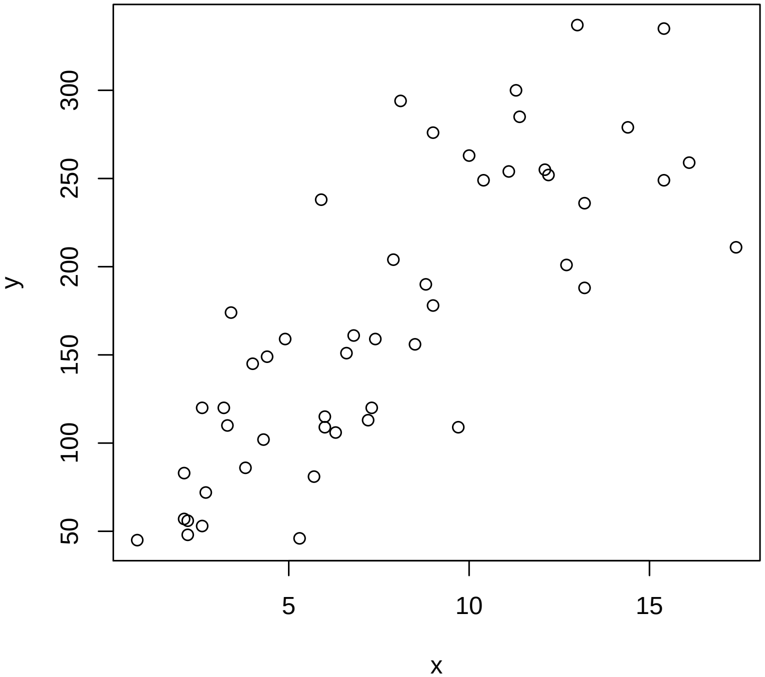
1. The response is y: number of assaults, and the regressor is x: number of murders, for each of the 50 states.

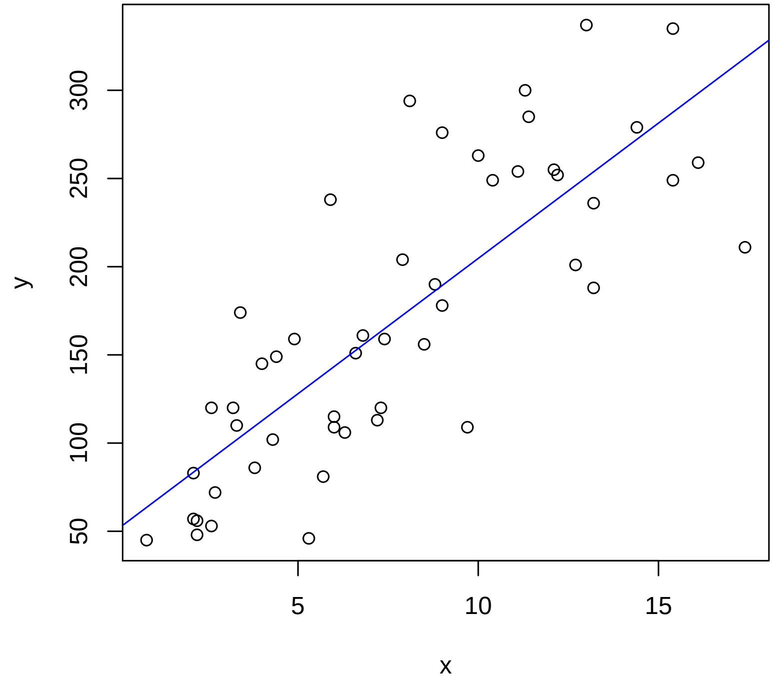
2. We have

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

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

The line of best fit is $\hat{y} = 51.27 + 15.34x$.

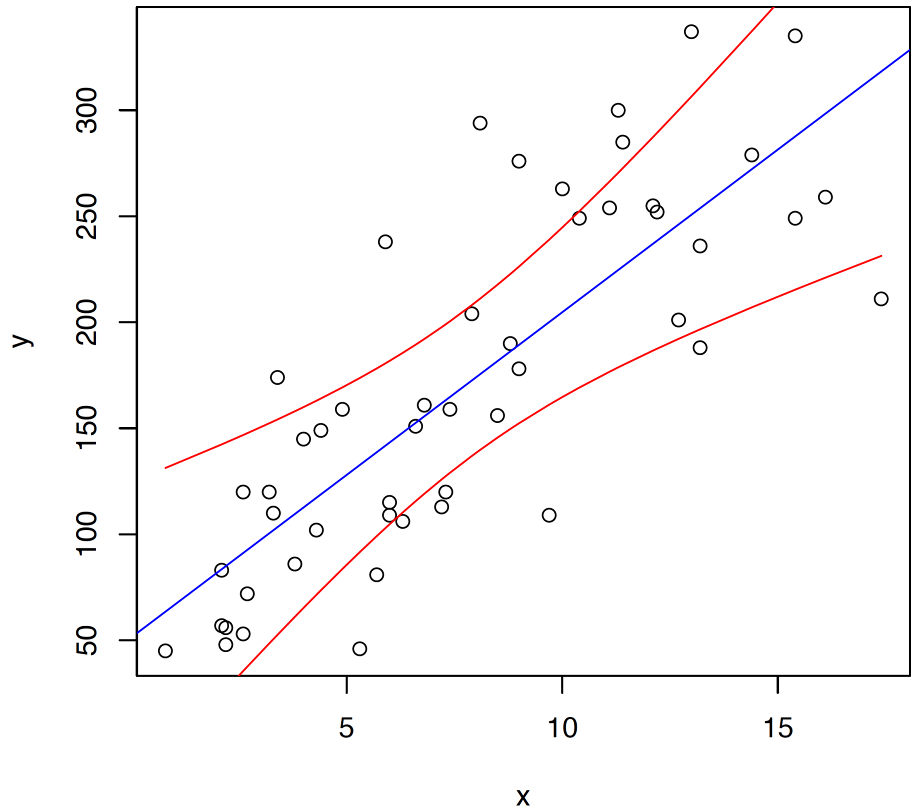




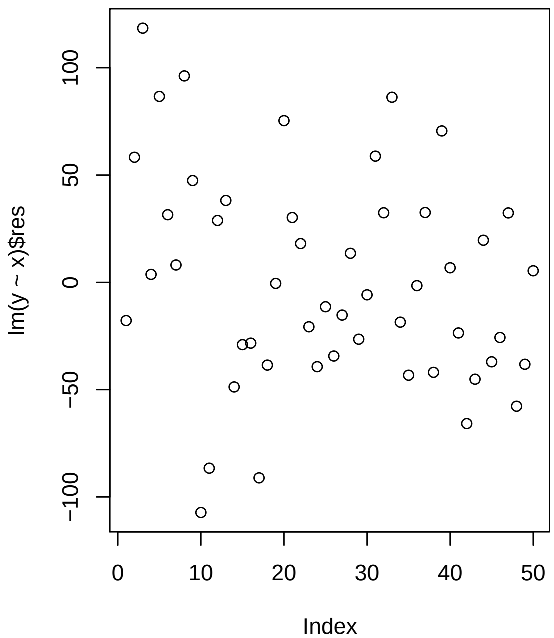
3. The correlation coefficient is $\rho = 0.802$, which suggests that there is a linear relationship between x and y. We test for the significance of the regression:
$$H_0 : \beta_1 = 0, \quad \text{against} \quad H_1 : \beta_1 \neq 0;$$
the test statistic
$$T_0 = \frac{b_1 - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t(n-2),$$
with $\hat{\sigma}^2 = 2531.73$ and $S_{xx} = 929.55$.

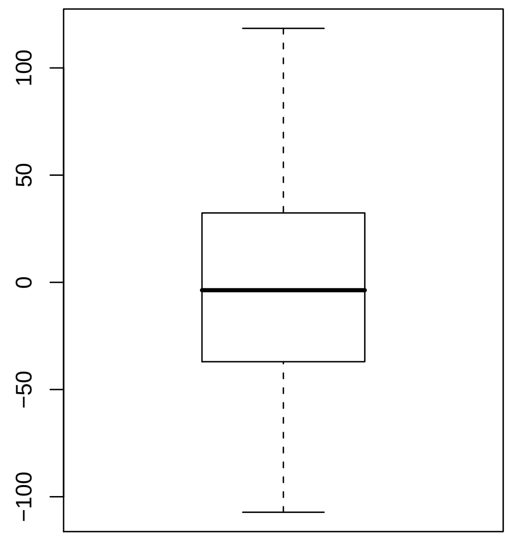
The observed value of the test statistic is $t_0 = 9.30$; since
$$t_{0.05/2}(48) \approx 2.01 < t_0 = 9.30,$$
we reject H_0 in favour of a linear relationship between x and y.

4. The 95% C.I. for the regression line is shown below:



5. The regression is a fairly good fit as the residuals show no systematic pattern: they seem uniformly distributed around 0.





6. As the regression seems to be a good model of the situation, it might have good predictive power (over its domain). We can predict the number of assaults in a state if the number of murders is $x_0 = 20$:

$$\hat{y}_0 = 51.27 + 15.34(20) = 358.07.$$

An equivalent way to ask for this answer is to look for a point estimate of the number of assaults in a state if the number of murders is 20.

The prediction interval for the number of assault in a state if $x_0 = 20$ is

$$358.07 \pm 2.01\sqrt{2531.73\left[1 + \frac{1}{48} + \frac{(20-7.78)^2}{929.55}\right]} = 358.07 \pm 40.64.$$

Example: Airline Data

This is a classic dataset, tracking the monthly totals of international airline passengers from 1949 to 1960. It is available in `R` as `AirPassengers`.

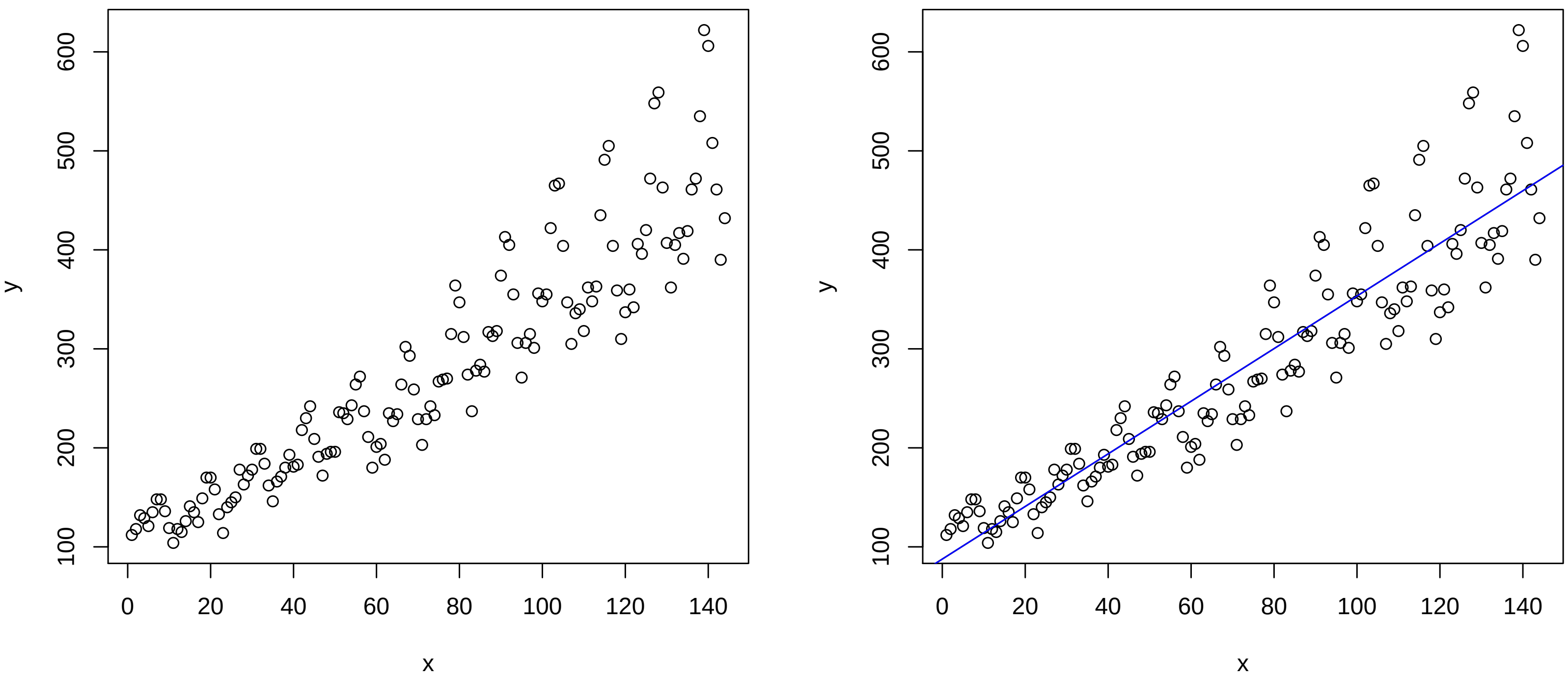
1. The response is y: number of monthly passengers, and the regressor is x: the number of month since January 1, 1949, $x = (1, 2, \ldots, 144)$.

2. We have

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

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

The line of best fit is $\hat{y} = 87.653 + 2.657x$.



3. The correlation coefficient is $\rho = 0.924$, which suggests that there is a strong linear relationship between x and y. We test for the significance:

$$H_0 : \beta_1 = 0, \text{ against } H_1 : \beta_1 \neq 0;$$

the test statistic

$$T_0 = \frac{b_1 - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t(n-2),$$

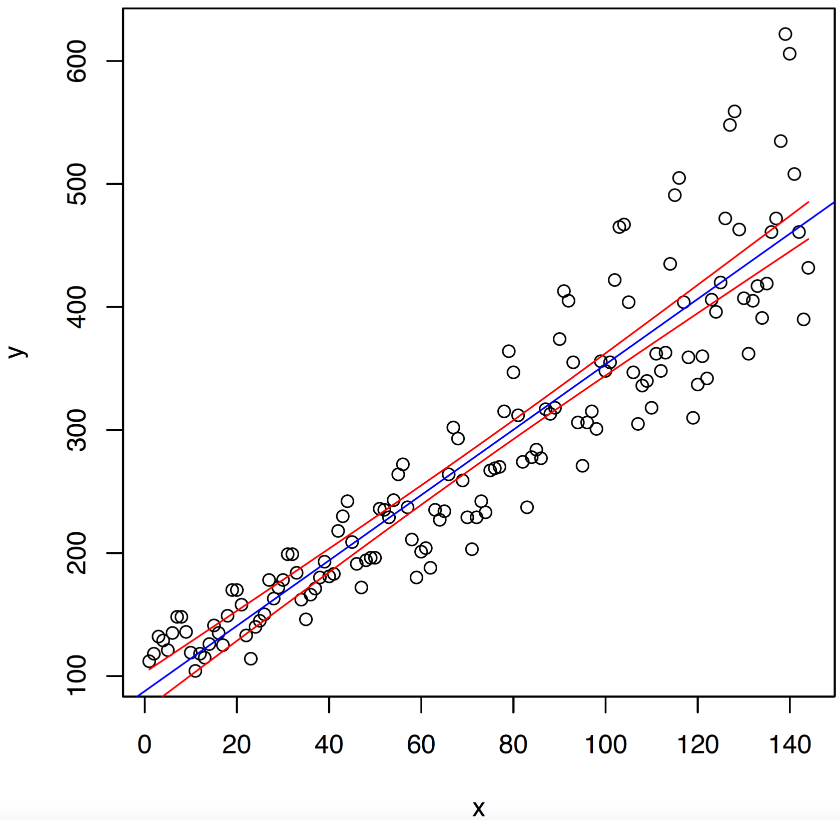
with $\hat{\sigma}^2 = 2121.261$ and $S_{xx} = 248820$.

The observed value of the test statistic is $t_0 = 28.77644$; since

$$t_{0.05/2}(142) \approx 1.97 < t_0 = 28.78,$$

we reject H_0 in favour of a linear relationship between x and y.

4. The 95% C.I. for the regression line is shown below:



5. The residuals show some structure: the variance of the error in not constant and increases with x. This suggests that data transformations need to be conducted before proceeding with linear regression.