Level M Regression Models Lecture 12

Now we will focus on the construction interval estimates for various parameters of our models.

R code

Please find some R code used to fit and plot some models. You can find most data sets on Moodle. Please download data and try some of the R code as you read through these notes.

Interval estimate for $\mathbf{b}^T\boldsymbol{\beta}$

In order to construct an interval estimate, we again apply the ideas already discussed in the previous lecture. In particular, an interval estimate for $\mathbf{b}^T\boldsymbol{\beta}$ with confidence c is

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} \pm t\left(n-p; \frac{1+c}{2}\right)\sqrt{\frac{RSS}{n-p}(\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b})}.$$

Prediction interval (PI) for Y given x

The quantity of interest here is a future observation of Y, Y_f say, when x takes the value x_f, which denotes the value of the explanatory variable at the position where a prediction of Y is required. Note that the observation (Y_f, x_f) was not used to construct the regression model or estimate regression parameters. The expected value $E(Y|x_f)$ can be written in the form $\mathbf{b}_f^T\boldsymbol{\beta}$. For example, with a simple linear regression, $E(Y) = \alpha + \beta x$, we can write $E(Y|x_f) = \alpha + \beta x_f = \mathbf{b}^T\boldsymbol{\beta}$, where $\mathbf{b}^T = (1, x_f)$ and $\boldsymbol{\beta}^T = (\alpha, \beta)$.

$$\frac{(\mathbf{b}^T\hat{\boldsymbol{\beta}} - \mathbf{b}^T\boldsymbol{\beta})}{\sqrt{\frac{RSS}{n-p}(1+\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b})}}$$

is a prediction interval for y_f since

$$\frac{(\mathbf{b}^T\hat{\boldsymbol{\beta}} - \mathbf{b}^T\boldsymbol{\beta})}{\sqrt{\frac{RSS}{n-p}(1+\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b})}} \sim t(n-p)$$

Applying the same results again, a prediction interval for y_f with confidence c is

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} \pm t\left(n-p; \frac{1+c}{2}\right)\sqrt{\frac{RSS}{n-p}(1+\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b})}.$$

Simple linear regression

Construct a 95% C.I. (confidence interval) *beta* for the model $y_i = \alpha + \beta x_{1i} + \gamma x_{2i} + \epsilon_i$. We can write this C.I. as

$$\hat{\beta} \pm t(n-p; 0.975)\text{s.e.}(\hat{\beta})$$

where

$$\text{s.e.}(\hat{\beta}) = \sqrt{\frac{RSS}{n-p}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}$$

and

$$\mathbf{b} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

R automatically prints the standard error (s.e.) of each individual parameter when it fits a regression model. This makes the construction of C.I.s for each parameter very easy given these estimates.

A parameter estimate is a random variable, since it can take several values, and in probability terms the estimated standard error of a parameter estimate is an estimate of its standard deviation.

Examples

Protein in pregnancy

Continuing the example from last lecture with data
Data: $(y_i, x_i) \quad i = 1, \ldots, 19$
Model: $E(Y_i) = \alpha + \beta x_i$
We now are asked to:

2. Construct a confidence interval for β to test it's significant.
3. Predict a future observation of protein level in a health women at $x = 27$ weeks.

To answer the questions we will need the following:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ . & . \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix}$$

$$(\mathbf{X}^T\mathbf{X}) = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} = \begin{pmatrix} 19 & 456 \\ 456 & 12164 \end{pmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.524763 & -0.019672 \\ -0.019672 & 0.000820 \end{pmatrix} \qquad (2)$$

and the R output

```
pregnancy<-read.csv("PROTEIN.CSV",header=T)
fit1<-lm(formula = Protein ~ Gestation,data=pregnancy)
summary(fit1)
```

```
##
## Call:
## lm(formula = Protein ~ Gestation, data = pregnancy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16853 -0.08720 -0.01009  0.08578  0.20422
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.201738   0.083363   2.420    0.027 *
## Gestation   0.022844   0.003295   6.934 2.42e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1151 on 17 degrees of freedom
## Multiple R-squared:  0.7388, Adjusted R-squared:  0.7234
## F-statistic: 48.08 on 1 and 17 DF,  p-value: 2.416e-06
```

Confidence Interval

So, a 95% C.I. for β is

$$0.0228 \pm t(17; 0.975)\sqrt{\frac{0.2251}{17}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}$$

and since $\mathbf{b} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ ie

$$\begin{aligned} 0.0228 \pm & \quad 2.11\sqrt{\frac{0.2251}{17}0.000820} \\ 0.0228 \pm & \quad 2.11(0.003295) \end{aligned}$$

i.e. 0.0228 ± 0.0070,
i.e. (0.016, 0.030)

The fact that this interval contains only positive values tells us that there is clear evidence that the average level of protein increases with gestation. The coefficient for β is highly likely to lie somewhere between 0.02 and 0.03.

Note: A confidence interval that includes zero indicates that there insufficient evidence of a relationship between the response and the explanatory variable. In this situation we would expect the p-value for testing $H_0 : \beta = 0$ to be > 0.05.

Prediction Interval

In order to use this model in a clinical setting we need a means of telling what values of protein level are expected for a future healthy mother who attends this clinic. For example, if a woman who is 27 weeks pregnant has a protein level of 1.06, should this be regarded as unusual? A prediction interval helps us to answer this question.

A 95% P.I. for a future observation y at $x = 27$ is done in the following way:

Here $\mathbf{b} = \begin{pmatrix} 1 \\ 27 \end{pmatrix}$.

Thus

$$\begin{aligned} \boldsymbol{b}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{b} &= \begin{pmatrix} 1 & 27 \end{pmatrix} \begin{pmatrix} 0.5247 & -0.0196 \\ -0.0196 & 0.0008 \end{pmatrix} \begin{pmatrix} 1 \\ 27 \end{pmatrix} \\ &= \begin{pmatrix} -0.0064 & 0.0025 \end{pmatrix} \begin{pmatrix} 1 \\ 27 \end{pmatrix} \\ &= 0.060255 \end{aligned}$$

and the prediction interval is

$$(0.2017 + 27 \times 0.0228) \pm 2.11\sqrt{\frac{0.2251}{17}(1 + 0.060255)}$$

$$(0.57, 1.07)$$

Since it lies within this interval (just) we have strong grounds for regarding a protein level of 1.06 as unusual. (Even if it did lie outside the interval, this only says that the result is unusual).

Note that **Prediction intervals will always be wider than confidence intervals**.

Trees

The dataset refers to the volume (cubic feet) and diameter (inches) (at 54 inches above the ground) and height (feet) for a sample of 31 black cherry trees in the Allegheny National Forest Pennsylvania. The data were collected in order to and an estimate for the volume of a tree (and therefore for the timber yield), given its height and diameter. A starting point for estimating volume using these data is the geometric formula for a cylinder:

Our full model for the trees data is

$$E(Y) = \alpha + \beta x_1 + \gamma x_2$$

where Y denotes log(volume), x_1 denotes log(diameter) and x_2 denotes log(height) of 31 trees.

The fitted model produced:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} -6.632 \\ 1.983 \\ 1.117 \end{pmatrix}$$

$$RSS = 0.1855$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 96.5721 & 3.1393 & -24.1651 \\ 3.1393 & 0.8495 & -1.2275 \\ -24.1651 & -1.2275 & 6.3099 \end{pmatrix}$$

The matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ was obtained from `R`.

Construct 95% confidence intervals for β and γ to test their significance.

In order to check whether or not there is clear evidence of a relationship between each of the explanatory variables and the response, we can construct interval estimates for β and γ.

A 95% C.I. for β is given by

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} \pm t(n-3;0.975)\sqrt{\frac{RSS}{n-3}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}$$

where $\mathbf{b} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$

$$\hat{\beta} \pm t(28;0.975)\sqrt{\frac{0.1855}{28}0.8495}$$

$$1.983 \pm 0.15$$

$$(1.83, 2.13)$$

This interval does not contain zero. There is therefore clear evidence of a relationship between log (diameter) and log (volume), i.e. log diameter is a significant predictor in addition to log height, and it is highly likely that the coefficient for log diameter lies between 1.83 and 2.13.

Note also that 2 is a plausible value for the coefficient of log (diameter). This is therefore consistent with the cylindrical model discussed in chapter 2, where (V $= \pi(\frac{\mathrm{d}}{2})^2$h; log(V) = ($\pi$/4)+2logd+logh).

A 95% C.I. for γ is given by

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} \pm t(n-3;0.975)\sqrt{\frac{RSS}{n-3}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}$$

where $\mathbf{b} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$

$$\hat{\gamma} \pm t(28;0.975)\sqrt{\frac{0.1855}{28}6.3099}$$

$$1.12 \pm 0.42$$

$$(0.70, 1.54)$$

Again there is clear evidence of a relationship between log (height) and log (volume), since 0 does not lie in the interval estimate. Log height is a significant predictor in addition to log diameter. The results are again consistent with the cylindrical model since the value 1 lies in the interval. It is highly likely that the coefficient for log height lies between 0.70 and 1.54.

Confidence interval for the difference in two population means

Data: (y_{ij}); $\quad i = 1, 2; j = 1, \ldots, n_i$

Model: $E(Y_{ij}) = \mu_i$, $Y_{ij} \sim N(\mu_i, \sigma^2)$

Construct a confidence interval for $(\mu_1 - \mu_2)$

$$\mathbf{Y} = \begin{pmatrix} y_{11} \\ y_{12} \\ . \\ y_{1,n_1} \\ y_{21} \\ y_{22} \\ . \\ y_{2,n_2} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ . & . \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ . & . \\ 0 & 1 \end{pmatrix}$$

Interest is in $(\mu_1 - \mu_2)$, i.e. $\mathbf{b} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$

$$(\mathbf{X}^T\mathbf{X}) = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}, \quad (\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{j=1}^{n_2} y_{2j} \end{pmatrix}, \quad (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix} = \hat{\boldsymbol{\beta}}$$

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} = (\bar{y}_1 - \bar{y}_2), \quad \mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = (\frac{1}{n_1} + \frac{1}{n_2})$$

What about the RSS (residual sum-of-squares)?

$$\begin{aligned} RSS &= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} y_{ij}^2 - (n_1\bar{y_1}^2 + n_2\bar{y_2}^2) \\ &= RSS_1 + RSS_2 \end{aligned}$$

where $RSS_i = \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$
Interval for $\mu_1 - \mu_2$:

$$(\bar{y_1} - \bar{y_2}) \pm t\left(n_1 + n_2 - 2; \frac{1+c}{2}\right)\sqrt{\frac{RSS_1 + RSS_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Autoanalyser data

Blood plasma concentrations are usually measured using a lengthy laboratory process. A simpler, cheaper method using an autoanalyser is often used. The autoanalyser is regularly tested to see if it is performing properly. On this occasion, 12 measurements have been made on samples of known concentration (3 replicates at each of 4 concentrations).

The following model has been fitted in `R` to estimate the autoanalyser concentration from the true concentration:

$$\text{autoanalyser}_i = \beta_0 + \beta_1 \text{true}_i + \epsilon_i, \quad i = 1, \ldots, 12$$

Construct a 95% C.I. for the population mean autoanalyser concentration when the true concentration is 6 units and interpret the interval.

```
auto<-read.csv("autoanalyser.csv")
# names(auto)
auto.lm<-lm(autoanalyser~true,data=auto)
summary(auto.lm)
```

```
##
## Call:
## lm(formula = autoanalyser ~ true, data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23333 -0.09583 -0.03333  0.10417  0.21667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.68333    0.18993   3.598  0.00487 **
## true         0.85000    0.04096  20.752  1.5e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1586 on 10 degrees of freedom
## Multiple R-squared:  0.9773, Adjusted R-squared:  0.975
## F-statistic: 430.6 on 1 and 10 DF,  p-value: 1.496e-09
```

We also have

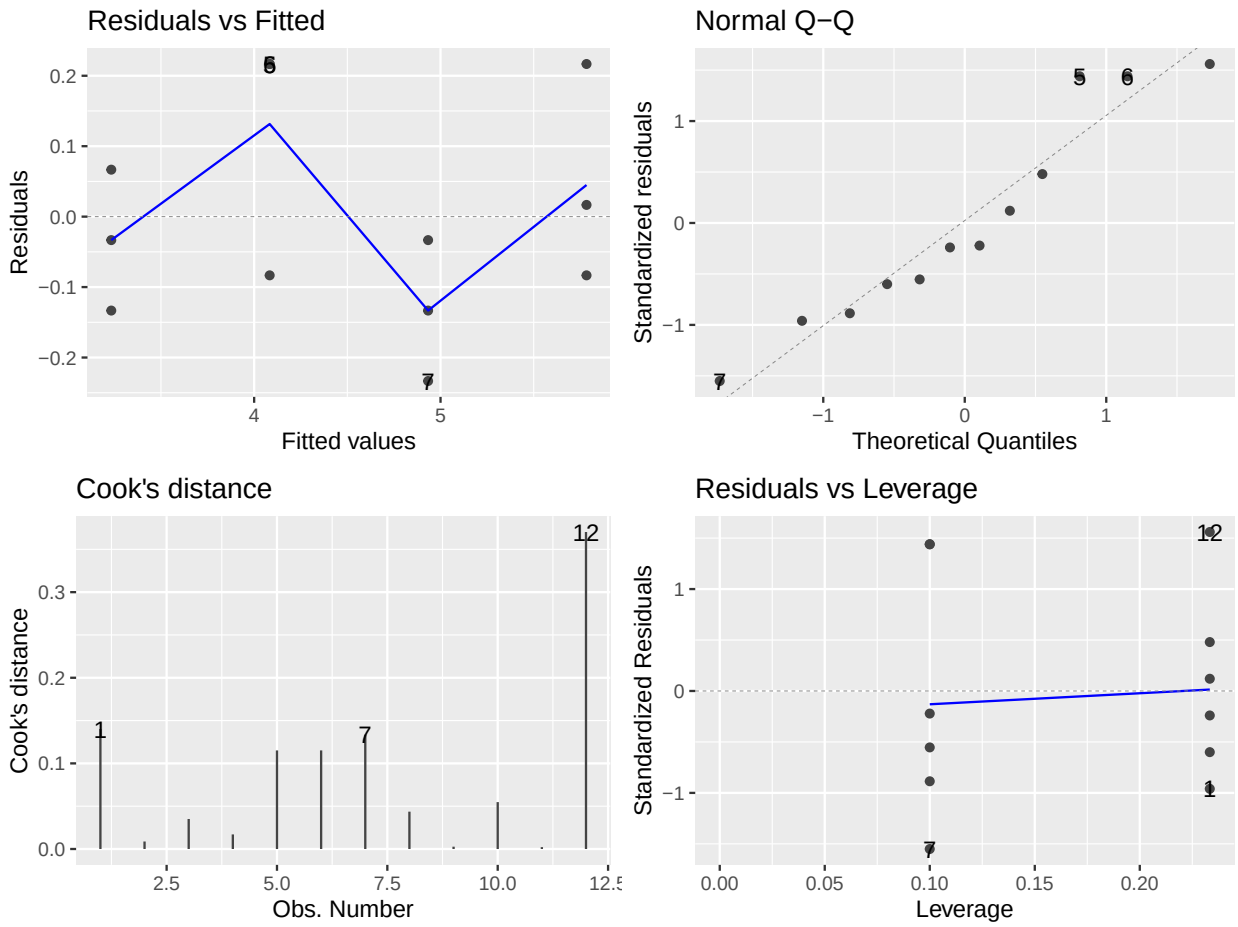
$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{180}\begin{pmatrix} 258 & -54 \\ -54 & 12 \end{pmatrix}$$

Model diagnostics

Normally at this point, you would check model assumptions and plot data to make sure your fitted model is adequate

```
#Load needed libraries
library(ggfortify)
library(gridExtra)

#Plot residuals against fitted values and Normal QQ plot
autoplot(auto.lm, which=c(1,2,4,5))
```

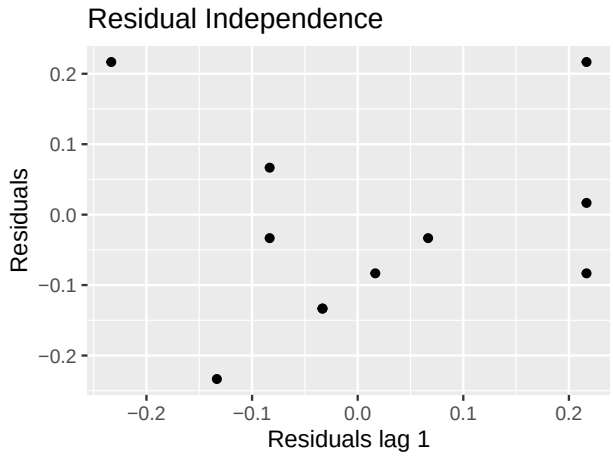


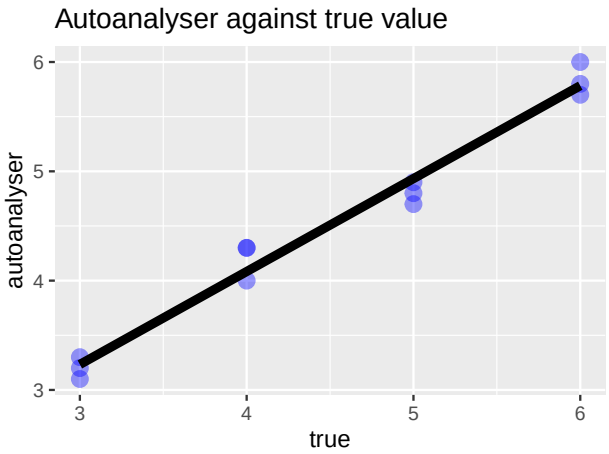

```
#Create data frame with model residuals
auto.fit<-data.frame(res=auto.lm$residuals,
                     res1=c(auto.lm$residuals[2:nrow(auto)],NA))

#Plot residuals against the previous residuals (lag 1)
plot1<-ggplot(auto.fit, aes(x=res1,y=res)) +
  geom_point() +
  labs(y="Residuals", x="Residuals lag 1", title="Residual Independence")

#Plot data
plot2<-ggplot(auto, aes(x = true, y = autoanalyser)) +
          geom_point(size=3.2, alpha = 0.4, col="blue") +
          ggtitle("Autoanalyser against true value") +
  geom_smooth(method = "lm",fullrange=TRUE, color="black",size=2,se=FALSE)

grid.arrange(plot1,plot2,ncol=2)
```





From residuals against fitted values, residuals seem randomly scattered around the zero line and so we may assume residuals have constant variance and mean zero. Form normal QQ plot, it seems reasonable to assume residuals are normally distributed. From Cook's distance and leverage there does not appear to be any potential influential observations or outliers. Also, there seems to reason to think residuals are not independent. Lastly from the scatterplot of data, the fitted line appears to provide a reasonable fit to the data and describe the relationship between autoanalyser value and true value.

Confidence Interval

We require an interval estimate for

$$\alpha + 6\beta = \mathbf{b}^T\boldsymbol{\beta}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} 1 \\ 6 \end{pmatrix}$$

$$\begin{aligned} \mathbf{b}^T\hat{\boldsymbol{\beta}} &= \hat{\alpha} + 6 \times \hat{\beta} \\ &= 0.683 + 0.85 \times 6 \\ &= 5.783 \\ RSS &= 0.252 \\ n &= 12 \end{aligned}$$

t(10;0.975) = 2.228

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{180}\begin{pmatrix} 258 & -54 \\ -54 & 12 \end{pmatrix}$$

Interval Estimate for $(\alpha + 6\beta)$: (5.61, 5.95)

The population mean autoanalyser concentration (when the true concentration is 6 units) is very likely to lie between 5.61 and 5.95 units

Prediction interval

Construct a 95% prediction interval for the autoanalyser concentration y when the true value x is 6 and interpret the interval

$$5.783 \pm 2.228\sqrt{\frac{0.252}{10}(1+0.233)}$$

$$5.783 \pm 0.393$$

(5.390, 6.176)

A future observation for the autoanalyser concentration is highly likely to lie between 5.39 and 6.18 when the true value is 6.

Additional Reading

Please see

- Sections 3.5 and 4.1 in **Linear Models with R**.
- Sections 2.7 and 2.8 in **Regression Analysis By Example**.
- Section 3.1.2 in **An Introduction to Statistical Learning**.