Week ON2

Elmer V Villanueva

09 March 2020

SET YOUR WORKING DIRECTORY!

```
setwd("C:/Users/1/Dropbox/00 - Working Folder/Teaching/DPH112/2019-2020/Week ON3")
```

Annoucements and clarifications

- During the online phase of this module, your attendance tied to your submission of the assignment. The assignment has a strict deadline and no extensions available. This is because the assignment solutions are released shortly after the deadline. In the event that you are unable to submit the assignment but wish to have tour attendance recorded, then you need to let me know by 6:00 PM on the Monday.
- Many of you have asked about the number of decimal points we you should follow when reporting estimates of correlation and regression. The rules from the previous semester are followed here.

Rule 1 - for variances, report estimates to +2 decimal places as the data; for all other statistics, report estimates to +1 decimal places as the data. This is important when reporting intercept and slope parameters

Rule 2 - for proportions, report to four decimal places; equivalently, for percentages, report to two decimal places. Since correlations are proportions, you need to report these accordingly.

Rule 3 - if $p \geq 0.001$, then report to three decimal places; if $p < 0.001$, then convert to scientific notation and report the coefficient to 2 decimal places.

Let's use data from last weeks Example 1 A as an example.

Table 1. Average peak plasma digoxin concentration with water and percentage change in concentration following intake of grapefruit juice.

Cmax (ngl/ml)	GFJ % Change
2.34	29.5
2.46	40.7
1.87	5.3
3.09	23.3
5.59	-45.1
4.05	-35.3
6.21	-44.6

Enter the data and always *visualise* as a first step.

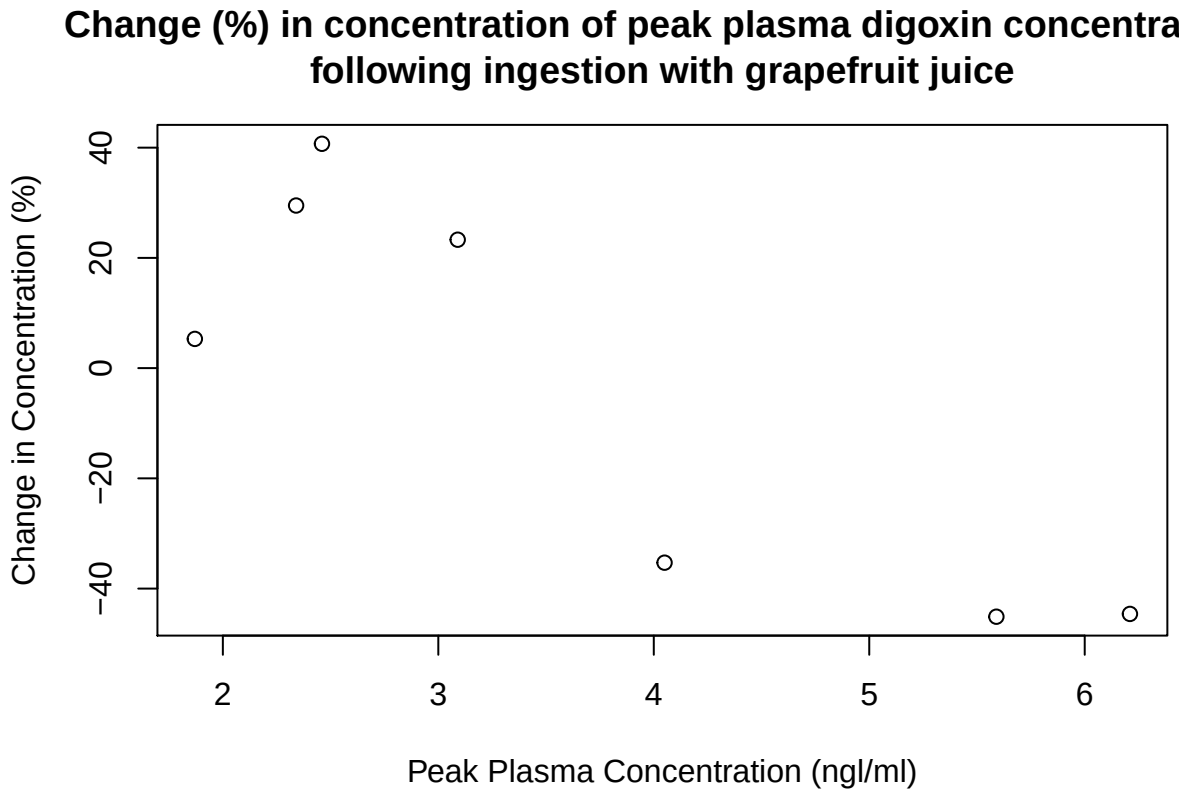
```
CMAX <- c(2.34, 2.46, 1.87, 3.09, 5.59, 4.05, 6.21)
GFJ <- c(29.5, 40.7, 5.3, 23.3, -45.1, -35.3, -44.6)
PARKER <- data.frame(CMAX, GFJ)
str(PARKER)
```

```
## 'data.frame':    7 obs. of  2 variables:
##  $ CMAX: num  2.34 2.46 1.87 3.09 5.59 4.05 6.21
##  $ GFJ : num  29.5 40.7 5.3 23.3 -45.1 -35.3 -44.6
```

```
head(PARKER)
```

```
##   CMAX   GFJ
## 1 2.34  29.5
## 2 2.46  40.7
## 3 1.87   5.3
## 4 3.09  23.3
## 5 5.59 -45.1
## 6 4.05 -35.3
```

```
plot(GFJ ~ CMAX, data = PARKER,
     main = "Change (%) in concentration of peak plasma digoxin concentration \nfollowing ingestion wit
     ylab = "Change in Concentration (%)",
     xlab = "Peak Plasma Concentration (ngl/ml)")
```



Estimate the correlation

```
cor(GFJ,CMAX)
```

```
## [1] -0.8648571
```

The answer is -0.8649 or -86.49%.

Estimate the intercept and slope parameters, and include the p-values.

```
PARKER.LM <- lm(GFJ ~ CMAX, data = PARKER)
summary(PARKER.LM)
```

```
##
## Call:
## lm(formula = GFJ ~ CMAX, data = PARKER)
##
## Residuals:
##       1       2       3       4       5       6       7
##   8.159  21.642 -24.982  16.227  -4.615 -24.111   7.680
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   65.855     19.646   3.352   0.0203 *
## CMAX         -19.023      4.938  -3.852   0.0120 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.42 on 5 degrees of freedom
## Multiple R-squared:  0.748,  Adjusted R-squared:  0.6976
## F-statistic: 14.84 on 1 and 5 DF,  p-value: 0.01197
```

The intercept is 65.86 (p=0.020) and the slope is -19.02 (p=0.012).

The application of these rules was suspended during the last two weeks. They come into effect from Monday. Deductions will be applied to answers that do not follow these rules.

- NEVER PROVIDE RAW CODE AS THE ANSWER TO YOUR ASSIGNMENTS. Raw R code is inappropriate in scientific writing. You need to take the results and state your answer to the question. If you submit raw code in an assignmetn or a report, you will be marked zero for that question. In reports, raw R code is typically included in a separate appendix.

To clarify, this is wrong:

> My answer to the correlation is

```
cor(GFJ,CMAX)
```

```
## [1] -0.8648571
```

This is correct:

> My answer to the correlation is -0.8649 or -86.49%.

```
cor(GFJ,CMAX)
```

```
## [1] -0.8648571
```

Reading

Read and understand Vittinghoff et al., Chapter 3.3.

In this chapter, we will continue learning about the simple linear regression model $Y = \beta_0 + \beta_1 X$ and its estimator $y_i = b_0 + b_1 x_i + \epsilon_i$. Many of our discussions hee are relevant to much more complex models, so we develop our understanding in this simple situation. We will spend this week and next week in discussions about simple linear regression. Then, we will move on to the more complex case of multiple linear regression models.

The relationship between simple linear regression and analysis of variance

Let us assume that we had *no information* about the independent variable. Our regression model becomes $Y = \beta_0$. What does this look like? Let us use Example 2A from last week to visualise this model.

Example 1A

Recall that the head of a lab wanted to know how long it took for 25 lab technicians to process blood samples for the presence of a particular genetic marker. Her data appear below.

Table 2. Lot size and hours to completion of DNA analysis

Lot Size	Hours
80	399
30	121
50	221
90	376
70	361
60	224
120	546
80	352
100	353
50	157
40	160
70	252
90	389
20	113
110	435
100	420
30	212
50	268
90	377
110	421
30	273
90	468
40	244
80	342
70	323

Let's enter the data and produce a scatterplot.

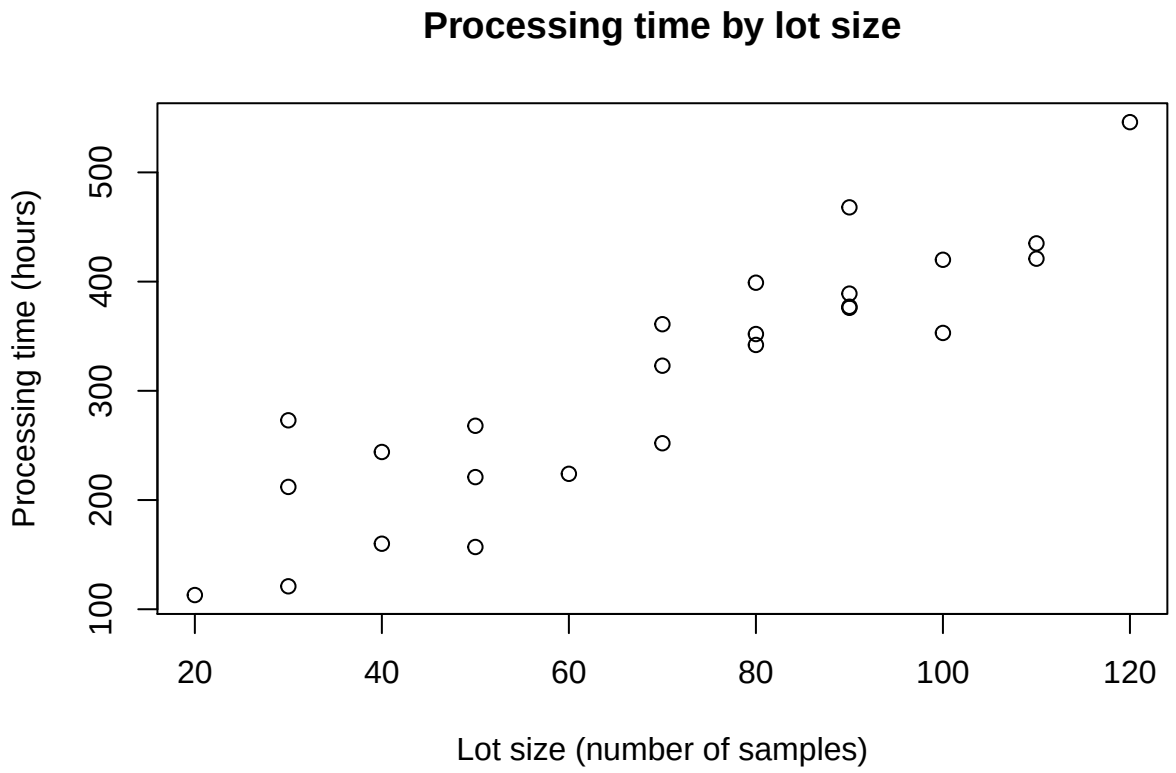
```
LOT <- c(80, 30, 50, 90, 70, 60, 120, 80, 100, 50, 40, 70, 90,
         20, 110, 100, 30, 50, 90, 110, 30, 90, 40, 80, 70)
HOURS <- c(399, 121, 221, 376, 361, 224, 546, 352, 353, 157, 160, 252, 389,
           113, 435, 420, 212, 268, 377, 421, 273, 468, 244, 342, 323)
LAB <- data.frame(LOT, HOURS)
str(LAB)
```

```
## 'data.frame':    25 obs. of  2 variables:
##  $ LOT  : num  80 30 50 90 70 60 120 80 100 50 ...
##  $ HOURS: num  399 121 221 376 361 224 546 352 353 157 ...
```

```r
head(LAB)
```

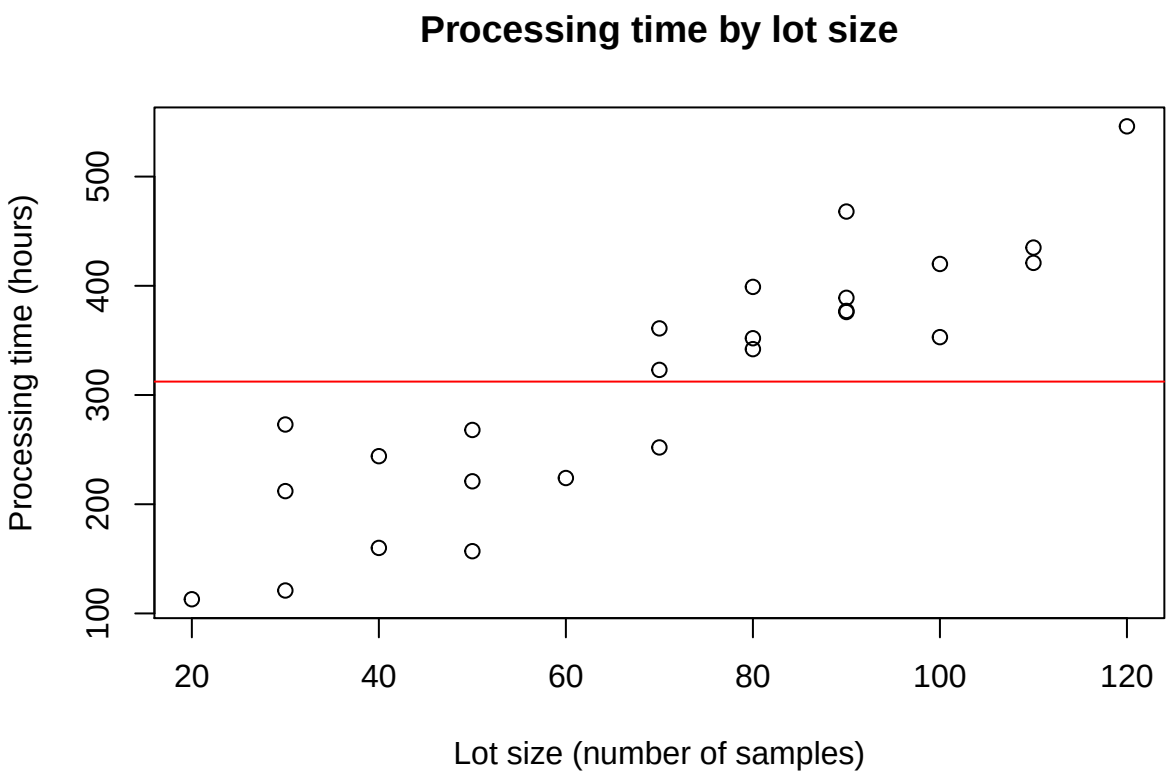
```
##   LOT HOURS
## 1  80   399
## 2  30   121
## 3  50   221
## 4  90   376
## 5  70   361
## 6  60   224
```

```r
plot(HOURS ~ LOT, data = LAB,
     main = "Processing time by lot size",
     ylab = "Processing time (hours)",
     xlab = "Lot size (number of samples)")
```



The model $y = b_0$ appears as

```r
plot(HOURS ~ LOT, data = LAB,
     main = "Processing time by lot size",
     ylab = "Processing time (hours)",
     xlab = "Lot size (number of samples)")
abline(h = mean(HOURS), col = "red")
```



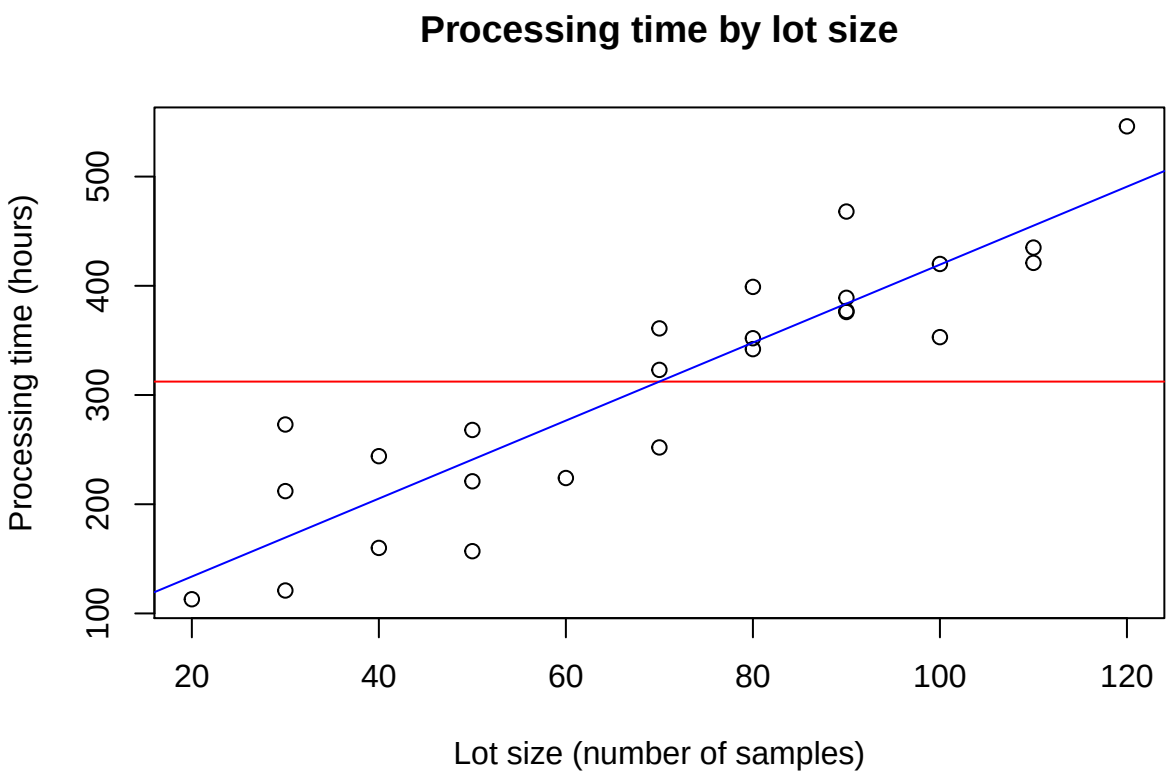
You can see that the model is quite uninformative. Nevertheless, we can calculate the *vertical distance* of each point y_i from the $\bar{y}$ line. If we square this value and summate it, we arrive at the idea of the *total deviation*. We learned about this deviation last semester, only it was called the *total sum of squares* or *SST*.

$SST = \sum (y_i - \bar{y})^2$

Now, let's say that we gained new information about the independent variable and can now include it in our model. Our new model is $Y = \beta_0 + \beta_1 X$ estimated as $\hat{y}_i = b_0 + b_1 x_i$ (note the use of tha "hat" symbol to emphasise that this is a prediction).

What does this model look like?

```
LAB.LM <- lm(HOURS ~ LOT, data = LAB)
plot(HOURS ~ LOT, data = LAB,
     main = "Processing time by lot size",
     ylab = "Processing time (hours)",
     xlab = "Lot size (number of samples)")
abline(h = mean(HOURS), col = "red")
abline(LAB.LM, col = "blue")
```



You can see that the blue line is a better fit than the red line. That is to say, the extra information we included about the independent variable improves the performance of our model.

Similar to the calculation of SST, we can calculate the *vertical distance* of each point $\hat{y}_i$ from the $\bar{y}$ line. If we square this value and summate it, we arrive at the idea of the *model dediation* or *explained deviation*. Again, we learned about this deviation last semester, only it was called the *model sum of squares*. In regression, we call this deviation *regression sum of squares* or *SSR*.

$SSR = \sum (\hat{y}_i - \bar{y})^2$

Finally, we note that the model $\hat{y}_i = b_0 + b_1 x_i$ is not perfect. We can calulate the *vertical distance* of each point y_i from the $\hat{y}_i$ line. If we square this value and summate it, we arrive at the idea of the *model dediation* or *unexplained deviation*. This final type of deviation is called *error sum of squares*, *residual sum of squares* or *SSE*.

$SSE = \sum (y_i - \hat{y}_i)^2$

The three deviations are related as follows:

$SST = SSR + SSE$, which expands to $\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$.

The ANOVA table from the regression model is given by `R` quite easily.

```
anova(LAB.LM)
```

```
## Analysis of Variance Table
##
## Response: HOURS
##           Df Sum Sq Mean Sq F value    Pr(>F)
## LOT        1 252378  252378  105.88 4.449e-10 ***
```

```
## Residuals 23  54825    2384
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results show that $SSR = 252378$ on $df = 1$ and $SSE = 54825$ on $df = 23$. SST is not given, but is the simple sum as given above.

Consistent with the analysis of variance presented last semester, we can divide SSR and SSE with their respective degrees of freedom to arrive at the means squares. The mean square of the regression in the above example is $MSR = SSR/(df_{SSR}) = 252378/1 = 252378$. The mean square error is $MSE = SSE/df_{SSE} = 54825/23 = 2384$. The MSE provides us with an estimate of the variance σ^2.

The final step is the calculation of the F statistic as $F = MSR/MSE = 252378/2384 = 105.88$. The F test compares this value against a threshold value with a $df = (1, n-2)$ to arrive at a p-value. What hypothesis does this p-value test? This p-value tests the null hypothesis $H_0 : y = b_0$ with the alternative hypothesis $H_A : y = b_0 + b_1 x$. That is, the p-value gives a global test of the importance of the extra information provided by knowing about the independent variable compared to not knowing about the independent variable.

Let us interpret the p-value from the F statistic in the example above. If we assume that lot size is unrelated to processing time and we repeat the observation with samples of 25 many, many times, the probability of observing the model $y = 62.4 + 3.6x$ or more extreme models is 4.45×10^{-10}.

Note that the p-value for the F test is given in the last line of the regression table in R.

```
summary(LAB.LM)
```

```
##
## Call:
## lm(formula = HOURS ~ LOT, data = LAB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.876 -34.088  -5.982  38.826 103.528
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.366     26.177   2.382   0.0259 *
## LOT            3.570      0.347  10.290 4.45e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.82 on 23 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8138
## F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

The relationship between simple linear regression and correlation

In the first lecture, we learned that the correlation coefficient is given by

$$r = \frac{Cov(X,Y)}{s_X s_Y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})\sum(y_i - \bar{y})}$$

In addition, in the previous lecture, we gave the equation for the estimate of the slope parameter as

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

What is the relationship between these two values? It turns out that another formula for the slope parameter uses the correlation coefficient as a term:

$$b_1 = r\frac{s_Y}{s_X}$$

This demonstrates the relationship between regression and correlation.

The coefficient of determination

What is a natural way to evaluate the strength of the model $\hat{y}_i = b_0 + b_1 x_i$? We already know that $SST = SSR + SSE$. Thus, it is intuitively appealing to speculate that if a regression equation does a good job of describing the relationship between two variables, the explained or regression sum of squares should constitute a large proportion of the total sum of squares. It would be of interest, then, to determine the magnitude of this proportion by computing the ratio of the explained sum of squares to the total sum of squares. The result is called the *coefficient of determination* and we give it the symbol r^2.

$$r^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

The largest value that r^2 can assume is 1, a result that occurs when all the variation in the y_i is explained by the regression. When $r^2 = 1$ all the observations fall on the regression line. The lower limit of r^2 is 0. This result is obtained when the regression line and the line drawn through $\hat{y}$ coincide. In this situation none of the variation in the y_i is explained by the regression.

When r^2 is large, then, the regression has accounted for a large proportion of the total variability in the observed values of Y, and we look with favour on the regression equation. On the other hand, a small r^2 which indicates a failure of the regression to account for a large proportion of the total variation in the observed values of Y, tends to cast doubt on the usefulness of the regression equation for predicting and estimating purposes.

The coefficent of determination is shown in the second-last line of the regression output of R.

```
summary(LAB.LM)
```

```
##
## Call:
## lm(formula = HOURS ~ LOT, data = LAB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.876 -34.088  -5.982  38.826 103.528
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.366     26.177   2.382   0.0259 *
## LOT            3.570      0.347  10.290 4.45e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.82 on 23 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8138
## F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

In this output, r^2 is listed as `Multiple R-squared` and is equal to 0.8215. The interpretation is straightforward. The model accounted for 82.15% of the variability in the data.

Another way to caluclate the coefficient of determination is to square the coefficient of correlation.

```
cor(HOURS, LOT); cor(HOURS, LOT)^2
```

```
## [1] 0.9063848
```

```
## [1] 0.8215335
```

This further demonstrates the relationship between regression and correlation.

Using the regrssion equation

In the previous lecture, we demonstrated the use of the predicted model $\hat{y}_i = b_0 + b_1 x_i$ to estimate new values. We will expand on this concept further.

There are two ways in which the equation can be used. It can be used to predict what value Y is likely to assume given a particular value of X. When the normality assumption of the model is met, we can construct a *prediction interval* for this predicted value of Y. We may also use the regression equation to estimate the mean of the subpopulation of Y values assumed to exist at any particular value of X. Again, if the assumption of normally distributed populations holds, a confidence interval for this parameter may be constructed. The predicted value of Y and the point estimate of the mean of the subpopulation of Y will be numerically equivalent for any particular value of X but, as we will see, the prediction interval will be wider than the confidence interval.

Predicting Y for a given X

The $100(1-\alpha)$ percent prediction interval for Y is given by

$$\hat{y} \pm t_{1-\alpha/2} s_y \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x}^2)}}$$

where x_p is the value of X at which we wish to obtain a prediction.

Example 1B

Recall the Example 2B in last week's lecture. The lab hires a new lab technician and, after a period of training, has been assigned to perform DNA analysis. The lab head is interested in the length of time it will take this new person to process 57 samples.

```
predict(LAB.LM, data.frame(LOT=57))
```

```
##        1
## 265.8674
```

Now we want to construct a 95% prediction interval around this estimate.

```
predict(LAB.LM, data.frame(LOT=57),
        interval = "prediction")
```

```
##        fit      lwr      upr
## 1 265.8674 162.4467 369.2881
```

Thus, we predict that the processing time of 57 samples is 265.9 hours with a 95% *prediction* interval of 162.4 to 369.3.

Predicting the *mean* of Y for a given X

The $100(1-\alpha)$ percent prediction interval for $\mu_{Y|X}$ is given by

$$\hat{y} \pm t_{1-\alpha/2} s_y \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x}^2)}}$$

where x_p is the value of X at which we wish to obtain a prediction.

Example 1B

The lab head is interested in the **mean** length of time it will take to process 57 samples.

```
predict(LAB.LM, data.frame(LOT=57),
        interval = "confidence")
```

```
##        fit      lwr      upr
## 1 265.8674 243.6166 288.1181
```

We predict that the processing time of 57 samples is still 265.9 hours with a 95% *confidence* interval of 243.6 to 288.1.

Note that the confidence interval is much thinner than the prediction interval. This can be shown in the following graph. Here, we need to use the added functionality of `ggplot2`.

```
if(!require(ggplot2)){install.packages("ggplot2")}
```

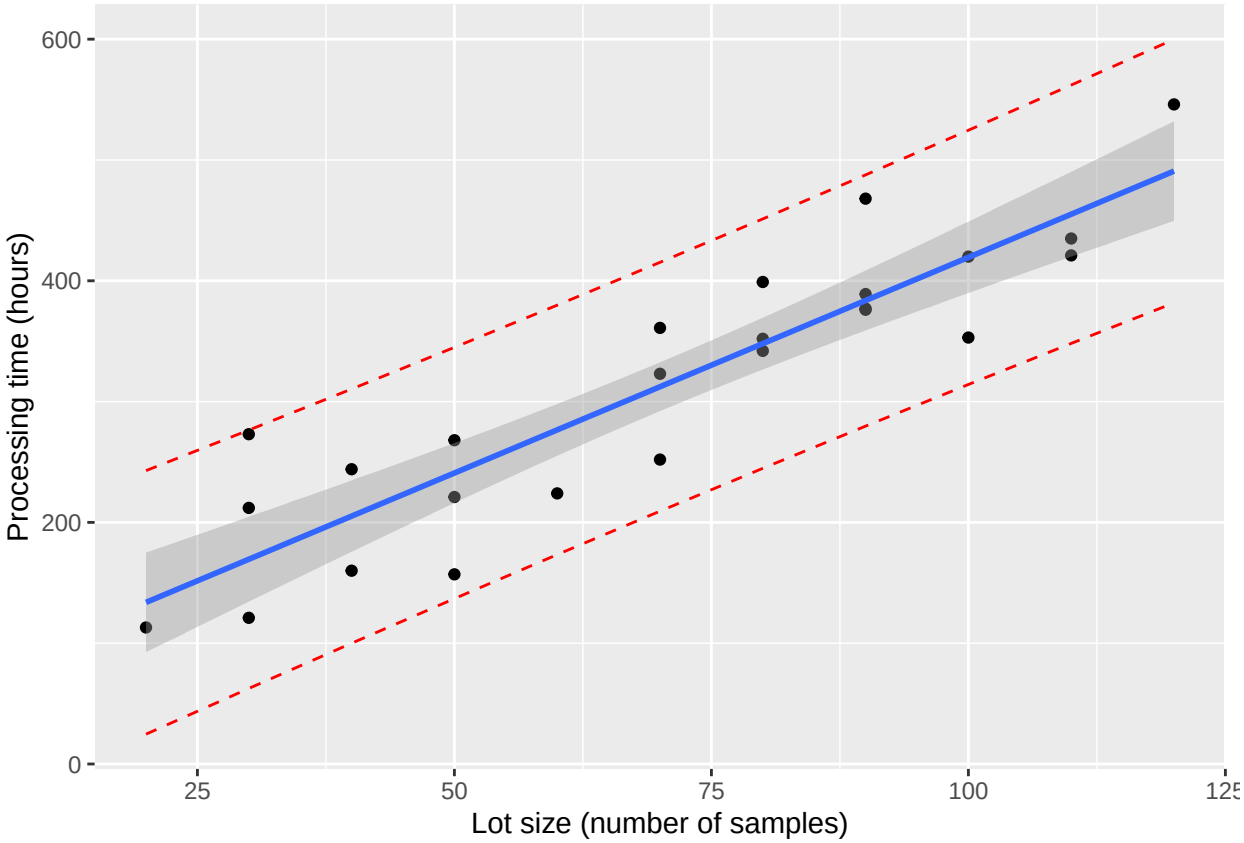
```
## Loading required package: ggplot2
```

```
library(ggplot2)

LAB.PRED <- predict(LAB.LM, interval = "prediction")
```

```
## Warning in predict.lm(LAB.LM, interval = "prediction"): predictions on current data refer to _future_
```

```
LAB2 <- cbind(LAB, LAB.PRED)

G <- ggplot(data = LAB2, aes(y=HOURS, x=LOT)) +
  geom_point() +
  stat_smooth(method = lm) +
  geom_line(aes(y = lwr), color = "red", linetype = "dashed") +
  geom_line(aes(y = upr), color = "red", linetype = "dashed") +
  ylab("Processing time (hours)") +
  xlab("Lot size (number of samples)")
G
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Here, the blue line is the estimated regression line $\hat{y}$. The grey band represents the *confidence* interval. The red band is the *prediction* interval. Note that both intervals are thinnest at $\bar{x}$ and widen as one moves away from the mean.

THE END