Statistics for Engineering, 4C3/6C3

Assignment 5

Kevin Dunn, kevin.dunn@mcmaster.ca **Due date: 08 March 2013**

Note: Assignment objectives

- Build least squares models in R.
- Extract useful information about the model outputs.
- Investigate and understand multiple linear regression (MLR) models.

Question 1 [12]

No need to use software. Question from the final exam, 2011.

Some data were collected from tests where the compressive strength, x, used to form concrete was measured, as well as the intrinsic permeability of the product, y. There were 16 data points collected. The mean x-value was $\overline{x} = 3.1$ and the variance of the x-values was 1.52. The average y-value was 40.9. The estimated covariance between x and y was -5.5.

The least squares estimate of the slope and intercept was: $y = 52.1 - 3.6x$.

1. What is the expected permeability when the compressive strength is at 5.8 units?
2. Calculate the 95% confidence interval for the slope if the standard error from the model was 4.5 units. Is the slope coefficient statistically significant?
3. Provide a rough estimate of the 95% prediction interval when the compressive strength is at 5.8 units (same level as for part 1). What assumptions did you make to provide this estimate?
4. Now provide a more accurate, calculated 95% prediction confidence interval for the previous part.

Solution

1. It is $\hat{y} = 52.1 - 3.6(5.8) = 31.22$
2. From the definition:

$$\begin{aligned} S_E^2(b_i) &= \frac{S_E^2}{\sum_j \left(x_j - \overline{\mathrm{x}}\right)^2} \\ &= \frac{4.5^2}{\sum_j \left(x_j - \overline{\mathrm{x}}\right)^2} \end{aligned}$$

We need the denominator term, which can be found by back-calculation:

$$\begin{aligned} \mathcal{V}(x) = 1.52 &= \frac{\sum_j \left(x_j - \overline{\mathrm{x}}\right)^2}{n-1} \\ \sum_j \left(x_j - \overline{\mathrm{x}}\right)^2 &= 1.52 \times (16-1) = 22.8 \end{aligned}$$

So the 95% confidence interval for the slope, b_i:

$$\begin{aligned} b_i &\pm c_t S_E(b_i) \\ -3.6 &\pm 2.14\sqrt{\frac{4.5^2}{22.8}} \\ -3.6 &\pm 2.02 \end{aligned}$$

where $c_t = 2.14$ from the t-distribution with $n - k = 16 - 2$ degrees of freedom.

Since this confidence interval *does not* span zero, we conclude the slope coefficient is statistically significant.

3. A rough estimate would be at $\hat{y} \pm 2S_E$, in other words, 31.2 ± 9.0, which is $[22.2, 40.2]$
4. A more accurate prediction interval is given by $\hat{y}_i \pm c_t\sqrt{V\{\hat{y}_i\}}$, where:

$$\begin{aligned} V\{\hat{y}_i\} &= S_E^2\left(1 + \frac{1}{n} + \frac{(x_i - \overline{\mathrm{x}})^2}{\sum_j (x_j - \overline{\mathrm{x}})^2}\right) \\ &= 4.5^2\left(1 + \frac{1}{16} + \frac{(5.8 - 3.1)^2}{22.8}\right) \\ &= 27.99 \end{aligned}$$

and represents the variance of the predicted $\hat{y}_i$ at the given value of $x_i = 5.8$.

The confidence interval, or prediction interval for this $\hat{y}_i$ is $\pm c_t\sqrt{V\{\hat{y}_i\}} = \pm 2.14\sqrt{27.99} = \pm 11.3$, a bit larger than the rough estimate above.

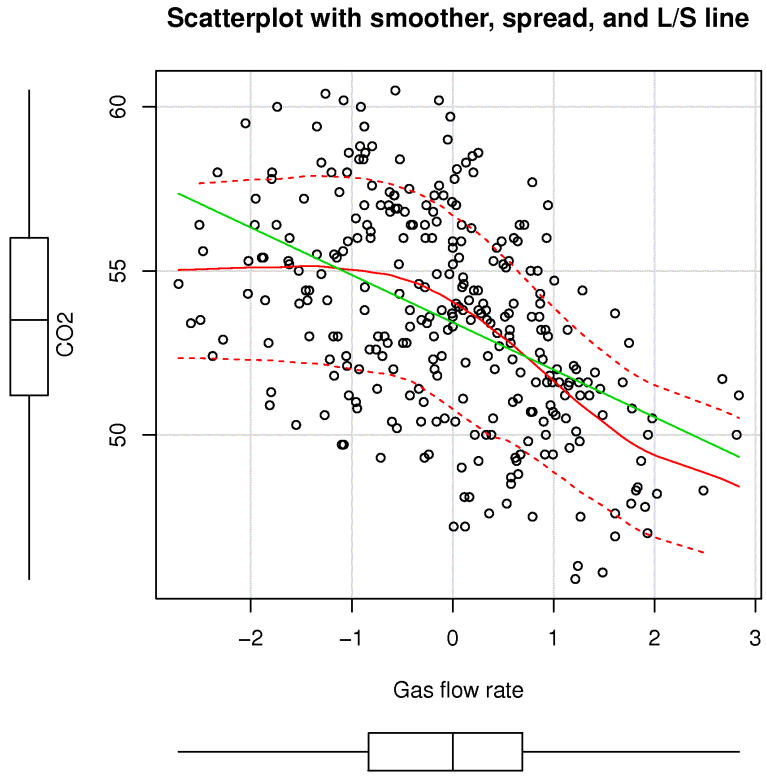
Question 2 [10]

Use the gas furnace data from the website to answer these questions. The data represent the gas flow rate (centered) from a process and the corresponding CO_2 measurement.

1. Make a scatter plot of the data to visualize the relationship between the variables. How would you characterize the relationship?
2. Calculate the variance for both variables, the covariance between the two variables, and the correlation between them, $r(x, y)$. Interpret the correlation value; i.e. do you consider this a strong correlation?
3. Now calculate a least squares model relating the gas flow rate as the x variable to the CO_2 measurement as the y-variable. Report the intercept and slope from this model.
4. Report the R^2 from the regression model. Compare the squared value of $r(x, y)$ to R^2. What do you notice? Now reinterpret what the correlation value means (i.e. compare this interpretation to your answer in part 2).
5. Switch x and y around and rebuild your least squares model. Compare the new R^2 to the previous model's R^2. Is this result surprising? How do interpret this?

Solution

1. Relationship: the data are negatively correlated.



I've chosen to use the `sp` or `scatterplot` function from the `car` library. It shows the scatterplot smoother (a.k.a. loess line) as solid red, the spread around the smoother (dashed red), the least squares regression line (black) and boxplots for each axis.

This is a great example of an information-rich visualization: packing the maximum amount of information into a small space. This plot answers so many questions we might have about the data.

2. The `cov(...)` command supplies the variance and covariance, and the `cor(...)` command gives the correlation.

   - Variance of input gas flow rate = 1.15 [gas flow units] ^{2}
   - Variance of CO_2 = 10.3 [CO_2 units] ^{2}
   - Covariance between input gas flow and CO_2 = -1.66 [gas flow units][CO_2 units]
   - Correlation = -0.48, i.e. around -0.5.

   From my experience with data, I personally would interpret this as a reasonably strong correlation. There is reasonably strong linear behaviour in the data cloud shown above, enough of a relationship to confidently say that "the CO_2 output does decrease at higher gas flow rates".

3. From the R model output:

   - intercept is -1.44 units of CO_2
   - slope is $53.4 \frac{\text{[units of CO}_2\text{]}}{\text{[units of gas flow]}}$

4.
   - From the R model output: $R^2 = 0.2347$
   - From earlier, the squared correlation is $(-0.484)^2 = 0.2347$, the same value.
   - Correlation can be interpreted as the square root of the R^2 value when regressing y on x (i.e. fitting a linear model to y using x as the input).

- Most novices would be misled and consider an R^2 value of 0.23 quite low. But notice that there is a repeatable and consistent negative linear relationship between x and y in this data.

5. This shows the interesting result that when regressing x on y (instead of the usual regression of y on x), that we get the same R^2 value. Note however that the *intercept* and *slope* are different between the two regressions.

   This also calls into question the interpretation of the R^2 value in regression. R^2 is just the square of the correlation coefficient. Recall from class the slide on the Wikipedia examples of correlation: there were examples where $r(x, y) = \sqrt{R^2}$ was zero, but still a strong *relationship* existing in the data. So we should interpret R^2 as a measure only of the *linear relationship* between two variables. And bear its quadratic nature in mind - interpreting the correlation is actually easier, and more "linear", in that a 0.2 improvement in correlation means the same thing when going from $r = 0.2$ to 0.4, as it does when going from $r = 0.7$ to 0.9 (not so for R^2).

```
gas <- read.csv('http://datasets.connectmv.com/file/gas-furnace.csv')
summary(gas)

library(car)
bitmap('CO2-gas-furnace-raw-data.png', type="png256",
        width=6, height=6, res=300, pointsize=14)

# Use the "sp" (scatterplot) function from the "car" library
sp(gas$InputGasRate, gas$CO2, xlab="Gas flow rate", ylab="CO2",
   main="Scatterplot with smoother, spread, and L/S line")
dev.off()

# (Co)variance and correlation
cov(gas)
cor(gas)

# Linear model:
model <- lm(gas$CO2 ~ gas$InputGasRate)
summary(model)

# ANOVA values
y.mean <- mean(gas$CO2)
RegSS <- sum((predict(model) - y.mean)^2)
RSS <- sum(residuals(model)^2)
TSS <- sum((gas$CO2 - y.mean)^2)
mean.square.residual <- RSS / model$df.residual

# Test normality of residuals
bitmap('CO2-gas-furnace-residuals.png', type="png256",
        width=6, height=6, res=300, pointsize=14)
par(mar=c(4.2, 4.2, 0.5, 0.5))
qqPlot(model)  # the qqPlot "knows" what to do with a model object
dev.off()
```

Question 3 [15]

In this question we consider the bioreactor yield data set and fit a linear model using all x-variables simultaneously to predict the yield.

1. Provide the interpretation for each coefficient in the model, and also comment on each one's confidence interval when interpreting it.
2. Compare the 3 slope coefficient values the case when you regress yield onto each x-variable on its own.:
   - $\hat{y} = 102.5 - 0.69T$, where T is tank temperature
   - $\hat{y} = -20.3 + 0.016S$, where S is impeller speed

- $\hat{y} = 54.9 - 16.7B$, where B is 1 if baffles are present and $B = 0$ with no baffles

Explain why your coefficients do not match.

3. Are the residuals from the multiple linear regression model normally distributed?
4. In this part we are investigating the variance-covariance matrices used to calculate the linear model.
   (a) First center the x-variables and the y-variable that you used in the model.

   Note: feel free to use MATLAB, or any other tool to answer this question. If you are using R, then you will benefit from the R tutorial on the course website. Also, read the help for the `model.matrix(...)` function to get the $\mathbf{X}$-matrix. Then read the help for the `sweep(...)` function, or more simply use the `scale(...)` function to do the mean-centering.

   (b) Show your calculated $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{y}$ variance-covariance matrices from the centered data.

   (c) Explain why the interpretation of covariances in $\mathbf{X}^T\mathbf{y}$ match the results from the full MLR model you calculated in part 1 of this question.

   (d) Calculate $\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ and show that it agrees with the estimates that R calculated (even though R fits an intercept term, while your $\mathbf{b}$ does not).
5. What would be the predicted yield for an experiment run without baffles, at 4000 rpm impeller speed, run at a reactor temperature of 90 °C?

Solution

1. The full linear model that relates bioreactor yield to 3 factors is:

$$y = 52.5 - 0.47x_T + 0.0087x_S - 9.1x_B$$

   where x_T is the temperature value in °C, x_S is the speed in RPM and x_B is a coded variable, 0=no baffles and 1=with baffles.

   - *Temperature effect*: $-0.74 < \beta_T < -0.21$, with $b_T = -0.47$ indicates that increasing the temperature by 1 °C will decrease the yield on average by 0.47 units, holding the speed and baffle effects constant. The confidence interval does not span zero, indicating this coefficient is significant. An ad-hoc way I sometimes use to gather the effect of a variables is to ask what is the effect over the entire range of temperature, ~ 40°C:
     - So a $\Delta y = -0.74 \times 40 = -29.6$ % decrease in yield
     - So a $\Delta y = -0.21 \times 40 = -8.4$ % decrease in yield

     A tighter confidence interval will have these two values even closer, but given the range of the y's in the data cover about 35% units, this temperature effect is important, and will have a noticeable effect at either end of the confidence interval.
   - *Speed effect*: $0.34 < \beta_S < 17.0822$ with $b_S = 8.7$ per 1000 RPM: indicates that increase the RPM by 1000 units will increase the yield by about 8.7 units, holding the other factors constant. While the confidence interval does not span zero, it is quite wide.
   - *Baffles effect*: $-15.9 < \beta_B < -2.29$ with $b_B = -9.1$ indicates the presence of baffles decreases yield on average by 9.1 units, holding the temperature and speed effects constant. The confidence interval does not span zero, indicating this coefficient is significant. It is an important effect to consider when wanting to change yield.
2. The separate effects are:
   - $\hat{y} = 102.5 - 0.69T$, where T is tank temperature
   - $\hat{y} = -20.3 + 0.016S$, where S is impeller speed
   - $\hat{y} = 54.9 - 16.7B$, where B is 1 if baffles are present and $B = 0$ with no baffles

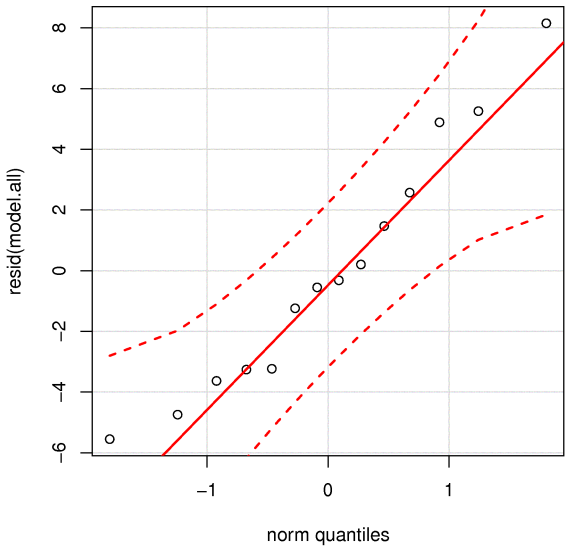
The signs of the coefficients between MLR and OLS (ordinary least squares) are in agreement, but not the magnitudes. The problem is that when building the single-variable regression model we place all the other effects into the residuals. For example, a model considering only temperature, but ignoring speed and baffles is essentially saying:

$$y = b_0 + b_T x_T + e$$
$$y = b_0 + b_T x_T + (e' + b'_S x_S + b'_B x_B)$$

i.e. we are lumping the effect of speed and baffles which we have omitted from the model, into the residuals, and we should see structure in our residuals due to these omitted effects.

Since the objective function for least squares is to minimize the sum of squares of the residuals, the effect of speed and baffles can be "smeared" into the coefficient we are estimating, the b_T coefficient, and this is even more so when any of the x-variables are correlated with each other.

3. The residuals from the multiple linear regression model are normally distributed. This can be verified in the q-q plot below:



4. The $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{y}$ variance-covariance matrices from the centered data, where the order of the variables is: temperature, speed and then baffles:

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1911 & -9079 & 36.43 \\ -9079 & 1844000 & -1029 \\ 36.43 & -1029 & 3.43 \end{bmatrix}$$
$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} -1310 \\ 29690 \\ -57.3 \end{bmatrix}$$

The covariances show a negative relationship between temperature and yield (-1310), a positive relationship between speed and yield (29690) and a negative relationship between baffles and yield (-57.3). Unfortunately, covariances are unit-dependent, so we cannot interpret the relative magnitude of these values: i.e. it would be wrong to say that speed has a greater effect than temperature because its covariance magnitude is larger. If we had two x-variables with the same units, then we could compare them fairly, but not in this case where all 3 units are different.

We can calculate

$$\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \begin{bmatrix} -0.471 \\ 0.0087 \\ -9.1 \end{bmatrix}$$

which agrees with the estimates that R calculated (even though R fits an intercept term, while we do not estimate an intercept).

5. The predicted yield yield for an experiment run without baffles, at 4000 rpm impeller speed, run at a reactor temperature of 90 °C would be 45%:

$$\hat{y} = 52.5 - 0.47x_T + 0.0087x_S - 9.1x_B$$
$$\hat{y} = 52.5 - 0.47(90) + 0.0087(4000) - 9.1(0) = \mathbf{45.0}$$

All the code for this question is given below:

```
bio <- read.csv('http://datasets.connectmv.com/file/bioreactor-yields.csv')
summary(bio)

# Temperature-Yield model
model.temp <- lm(bio$yield ~ bio$temperature)
summary(model.temp)

# Impeller speed-Yield model
model.speed <- lm(bio$yield ~ bio$speed)
summary(model.speed)

# Baffles-Yield model
model.baffles <- lm(bio$yield ~ bio$baffles)
summary(model.baffles)

# Model of everything
model.all <- lm(bio$yield ~ bio$temperature + bio$speed + bio$baffles)
summary(model.all)
confint(model.all)

# Residuals normally distributed? Yes
library(car)
bitmap('bioreactor-residuals-qq-plot.png', type="png256",
         width=6, height=6, res=300, pointsize=14)
par(mar=c(4.2, 4.2, 1.5, 0.5))
qqPlot(resid(model.all))
dev.off()

# Calculate X matrix and y vector
data <- model.matrix(model.all)
X <- data[,2:4]
y <- matrix(bio$yield)

# Center the data first
X <- scale(X, scale=FALSE)
y <- scale(y, scale=FALSE)

# Now calculate variance-covariance matrices
XTy <- t(X) %*% y
XTX <- t(X) %*% X
b <- solve(XTX) %*% XTy
# b agrees with R's calculation from ``model.all''
```

Question 4 [8]

The grades from a recent midterm exam are available, as well as the time taken by the student to write the exam. It was an "infinite" time midterm, so there was no time pressure to finish within the allocated period.

The data are available on the data set website.

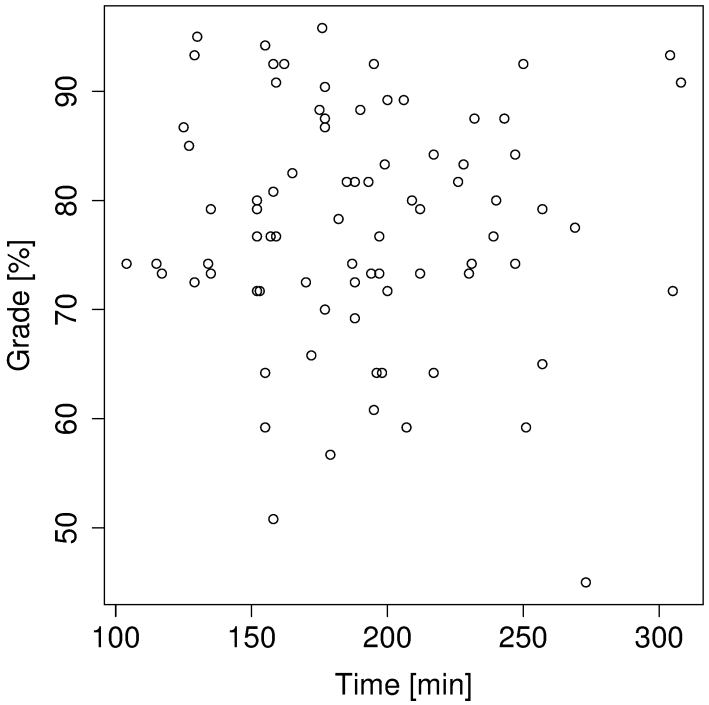
1. Use the data to confirm whether or not the amount of time used to write the test has an influence on the grade value obtained.
2. Do the regular least squares assumptions apply in this instance? Assess the assumptions individually.

Solution

This answer contains more content than expected for the solution. It contains material that 600-level students should read and be comfortable. Thanks to Sean, Jervis and Xin for helping me out with the solution.

1. In order to determine if the amount of time spent writing the test impacts a student's grade on the test, we start by plotting the data and note that there is no predominant trend in the data. There seems to be no relationship between time and grade, because of the seemingly random scatter of data on the following plot:

```
grades <- read.csv("http://datasets.connectmv.com/file/unlimited-time-test.csv")
y <- grades$Grade
x <- grades$Time
plot(x, y, xlab="Time [min]", ylab="Grade [%]")
summary(grades)
```



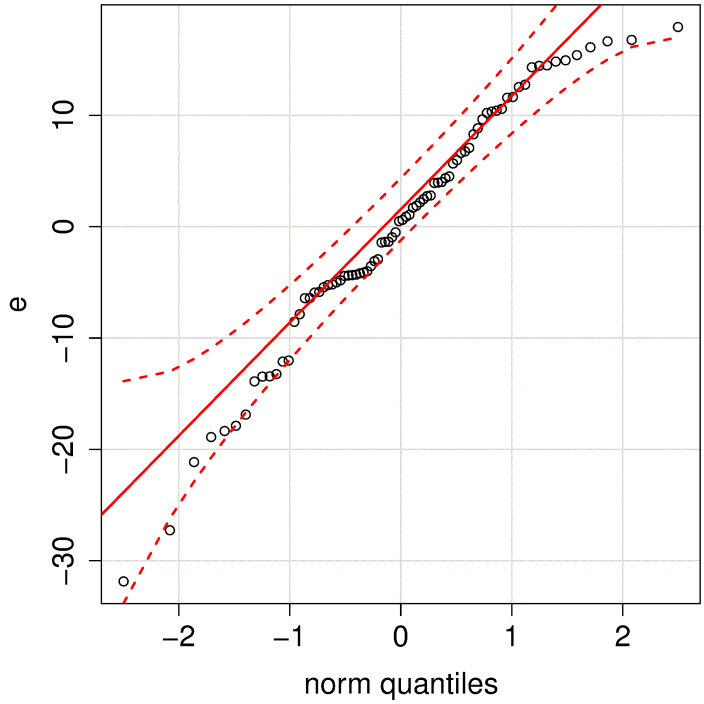
To verifys we construct a linear model and find that $b_0 = 79.7$ and $b_1 = -0.01$.

The standard error of the model residuals, as calculated by R, is 10.82. The y-values of grades have an IQR of 14.2 with median 77.9. Therefore, the S_E of the residuals is approximately 0.75 of the IQR of grades. This tells us that the model is not a very good predictor of y.

The residuals are normally distributed in the q-q plot, and residuals plotted with respect to time are randomly distributed. We might therefore think that this model is representative of the data, just with much uncertainty and very weakly correlated. This is easily dismissed by analyzing the confidence interval of the slope coefficient, b_0.

```
model <- lm(y ~ x)
b0 <- model$coefficients[1]
b1 <- model$coefficients[2]
e <- resid(model)
y.hat <- b0 + b1*x
```

```
library(car)   # Plots to determine validity of model
qqPlot(e)      # Are the residuals normally distributed?
plot(x, e)     # Is there a correlation between to the residuals?
abline(h=0, lty=2)
```



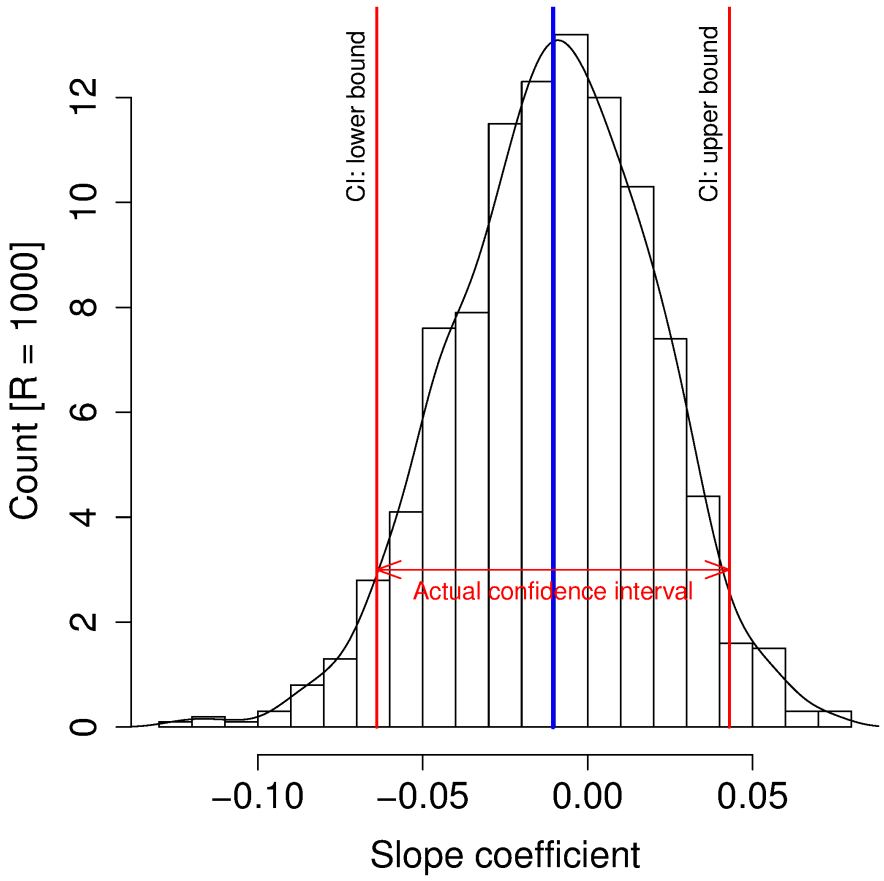
We have R calculate the 95% confidence interval for the slope coefficient and find that it lies within a range from -0.06 to +0.04. It spans 0 fairly evenly. It is therefore likely that time has no effect on grade received on this test.

```
# Calculate standard errors with subsets of n=14
n = length(x)
den <- sum((x - mean(x)) * (x - mean(x)))
SE <- sqrt(sum(e^2) / (n-2))
SE.b1 <- sqrt(SE^2 / den)

# Confidence intervals for the least squares parameters
alpha = 0.95
b1 = coef(model)[2]
t.critical = qt(1-(1-alpha)/2, df=(n-2))
b1.LB <- b1 - t.critical*SE.b1
b1.UB <- b1 + t.critical*SE.b1
```

[*This part is not required at all for full grade, but is interesting*]: We can confirm our answer with the bootstrapping technique. The data set consists of 80 rows. So randomly select 80 rows from the dataset, allowing for duplicates. Fit a least squares model and calculate the slope coefficient. Repeat this 1000 times, each time taking a different random sample. In this way, any outliers in the data set will sometimes be excluded and sometimes included.

Then, this distribution of slope coefficients is plotted and analyzed. If the highest density of slopes calculated is close to the original model's slope coefficient and the bootstrapped slope values lie above and below it, we can be sure that our original slope coefficient computed is fair and accurate. If we don't see any outlier bars in the histogram, we can be sure that the source data set in representative and any conclusions drawn (i.e. that time has no effect on the test score) are reliable *for the entire data set*. This is indeed the case here: *there are no outlier students!* Here's the bootstrapping code. It is generic and can be used for any future linear model you might build. Bootstrapping is a really neat statistical technique that is widely applied in science and engineering. I encourage you to read up about it.

```
ls.fun <- function(data, i){
  # `i` contains the indices of the rows to use in the least squares model
  print (i)
  d.subset <- data[i,]
  d.ols <- lm(d.subset$Grade ~ d.subset$Time)

  # Return just the slope coefficient
  return(coef(d.ols)[2])
}

set.seed(42)  # pick any random seed to initial the RNG
library(boot)
grades.boot <- boot(grades, ls.fun, R=1000)

library(MASS)
truehist(grades.boot$t, col=0, xlab="Slope coefficient", ylab="Count [R = 1000]")
lines(density(grades.boot$t))

ymax=15
segments(x0=coef(model)[2], y0=0, y1 = ymax, col="blue", lwd=3)
segments(x0=b1.LB, y0=0, y1 = ymax, col="red", lwd=2)
segments(x0=b1.UB, y0=0, y1 = ymax, col="red", lwd=2)
arrows(x0=b1.LB, x1=b1.UB, y0=3, y1=3, angle=20, length=0.2, col="red", code=3)
text(x=(b1.LB+b1.UB)/2, y=3, "Actual confidence interval", pos=1, col="red")

delta=0.01
text(x=b1.LB-delta, y=(ymax-4), "CI: lower bound", srt=90, pos=4)
text(x=b1.UB-delta, y=(ymax-4), "CI: upper bound", srt=90, pos=4)
```



2. Least square model assumptions are stated and addressed individually.

   Assumption 1: Model is linear

   Linearity of model can be checked by plotting input data (time) against model residuals (on y-axis), as shown below. The residuals are evenly and randomly distributed on both sides of zero, with no observable structure. This indicates that a linear model is suitable to represent this particular data set, and the only residuals left are errors. However, since the linear model's slope coefficient is not significant, as shown previously, this assumption becomes trivial.

   Assumption 2: Variance of y is constant at all values of x

   This assumption can also be checked by plotting x values against residuals. If the variance of y (grade) changes at different x values (time), the residuals will be closer together at one end of the x-axis and more scattered at the other end, which was not observed. Therefore, the variance of y can be assumed to be constant at all x values. This observation is also consistent with the previous conclusion that time has no effect on grades.

   Also, a visual plot of the data shows the grades have roughly the same spread at all levels of time.

   Assumption 3: Errors are normally distributed

   This assumption can be easily tested by a q-q plot of the linear model residuals. As shown in the graph earlier, all model residuals fall within the 95% confidence interval for a normal distribution. Therefore, this assumption still applies in this case.

   Assumption 4: Each error is independent of the other

   Independence of error can be assessed by using the autocorrelation function estimation. Using the `acf(residuals)` autocorrelation command in R, only the first error observation shows significant autocorrelation (outside blue boundary), which by definition will always be true. The remaining errors show no significant autocorrelation, which confirms that they are independent of each other.

   Assumption 5: Assume x are fixed and independent of the error

   The x values in this case are the time each student used to complete the midterm. Time to complete each midterm can be easily and relatively accurately measured. Also, measuring the time likely will not affect a student's performance and grades. Therefore, the x values (time) can be considered fixed, which means they will not contribute to the model's error. In addition, the independence of x values can be assessed by the autocorrelation function estimation as well. Only the first observation show significant autocorrelation, which confirms independence in x values.

   Assumption 6: All y_i values are independent of each others.

   The independence of y values directly depend on the independence of errors, which was previously confirmed in assessment of assumption 4. In fact, if the autocorrelation function estimation was applied on the observed y-values, a graph identical to that shown in assumption 4 will be generated, indicating that the y values are also independent of each other. This observation also makes physical sense as the students supposedly completed the midterm on their own. Therefore their grades should be independent of each other.

END