Lab Exercises Lecture 6-8

1. Suppose a random sample of size n = 100 has been selected and the sample mean is found to be x̄ = 67. The population standard deviation is assumed to be σ = 12. Please answer the following questions.

(a) What is the standard error of the mean σx̄?

Answer: 1.2

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{100}} = 1.2$$

(b) What is the margin of error if the confidence level is (1 – α) = 0.95?

Answer: 2.352

If (1 –α) = 0.95, then zα/2 =2 = z0.025 = 1.96

```
qnorm(0.025, lower.tail = FALSE) * 12 / sqrt(100)
## [1] 2.351957
```

Therefore, zα/2σx̄ = (1.96)(1.2) = 2.352

(c) What is the 95% confidence interval estimate of *μ*?

Answer: 67 ± 2.352 or [64.648, 69.352]

x̄ ± zα/2σx̄

67 ± 2.352

[64.648, 69.352]

```
#Comment1. To find the upper bound of the confidence interval
#estimate, add the margin of error to the sample mean.

67 + qnorm(0.025, lower.tail = FALSE) * 12 / sqrt(100)

## [1] 69.35196

#Comment2. To find the lower bound of the confidence interval
#estimate, subtract the margin of error to the sample mean.

67 - qnorm(0.025, lower.tail = FALSE) * 12 / sqrt(100)

## [1] 64.64804
```

2. Suppose the Dean of a large medical college wishes to estimate the mean student age of the most recent entering class of aspiring physicians pursuing an M.D. degree. A quick pilot study reveals that 3 years might be used as an estimate of the value of σ. If a 95% confidence interval estimate with a margin of error 1 is desired, what sample size should we recommend?

Answer: n = 35

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{(PME)^2}$$

$$n = \frac{(1.96)^2(3)^2}{(1)^2} = 34.57 \approx 35$$

```
qnorm(0.025, lower.tail = FALSE)
## [1] 1.959964
(qnorm(0.025, lower.tail = FALSE) * 3 / 1) ^ 2
## [1] 34.57313
```

check:

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = (1.96)\frac{3}{\sqrt{35}} = (1.96)(0.51) \approx 1$$

```
qnorm(0.025, lower.tail = FALSE) * 3 / sqrt(35)
## [1] 0.9938831
```

……

3. Family physicians in Tampa, Florida reportedly earn an average annual salary of \$141. 300. Suppose we conduct a survey on a sample of n = 64 family physicians from New Orleans, Louisiana to test whether their mean annual salary is different from the reported mean of \$141, 300 in Tampa, and find that the sample mean is \$138, 000. Assume σ = \$18, 000. At the level of α = 0.01, test H0 : *μ* = 141, 300; against Ha : *μ*0 ≠ 141, 300. What is the p-value?

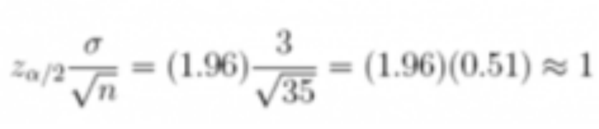
Answer:

(a) H0 : *μ* = 141, 300
(b) Ha : *μ*0 ≠ 141, 300
(c) n = 64 and α = 0.01
(d) Reject H0 if z>za/2 = z0.005 = 2.576 or z< – z a/2 = –z0.005 = –2.576. That is
RR : z>2.576 or z< – 2.576 where

$$z = \frac{\overline{x} - \mu_0}{\sigma_{\overline{x}}}$$

```
qnorm(0.005)
## [1] -2.575829
qnorm(0.995)
## [1] 2.575829
```

(e) Since x̄ = 138, 000



$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = (1.96)\frac{3}{\sqrt{35}} = (1.96)(0.51) \approx 1$$

(f) Since z = 1.47<2.576 (and thus does not fall in the RR), we do not reject H0. We cannot conclude from the evidence that New Orleans physician salary is different from Tampa physician salary.

Answer:
The p-value = (2)(p(z>1.47)) = 0.1416.

```
2 * pnorm(1.47, lower.tail = FALSE)
## [1] 0.1415618
```

Since p-value = 0.1416> α= 0:01, we do not reject H0.

………

4. Use the Cars93 data from the R package named MASS.

```
library("MASS")
> names(Cars93)
```

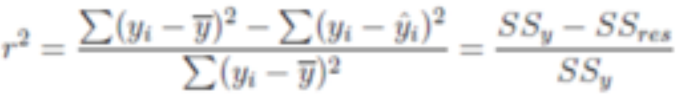
a) Find if MPG.city and EngineSize appear related in any systematic way? Comment.

Answer:

```
plot(Cars93$EngineSize, Cars93$MPG.city,
pch = 19,
xlab = 'Engine Size (liters)',
ylab = 'City Miles per Gallon ',
main = 'Relationship Between City MPG and Engine Size (liters)’)
```

Answer: The pattern of points revealed by the scatterplot suggests that the relationship is negatively related. The important question is whether the relationship is linear; the scatterplot suggests that it is probably more curvilinear than linear. To sort out that issue, we next move on to the residual plot.

b) What is the strength of association between the two variables, MPG.city and Engine Size? Find the coefficient of determination r2 using the following expression for r2. This exercise provides another opportunity to hone your coding skills.



$$r^2 = \frac{\sum(y_i - \overline{y})^2 - \sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y})^2} = \frac{SS_y - SS_{res}}{SS_y}$$

```
#Comment1. Find the total sum of squares, ss_y.

ss_y <- sum((Cars93$MPG.city - mean(Cars93$MPG.city)) ^ 2)

#Comment2. Find the residual sum of squares, ss_res.

ss_res <- sum((resid(slr2)) ^ 2)
```

#Comment3. Find the coefficient of determination.

```
(ss_y - ss_res) / ss_y

##    [1]    0.505143
```

What does the coefficient of determination r2 reveal about the regression model?

Answer: The r2 indicates the proportion of variation in the dependent variable MPG.city that is explained (or accounted for) by variation in the independent vari- able EngineSize. In this case, that proportion is 0.505143, or roughly 51%. More- over, because r2 = 0.505143, we also know that almost 49% of the variation in MPG.city remains unaccounted for, even once the association with EngineSize has been considered.

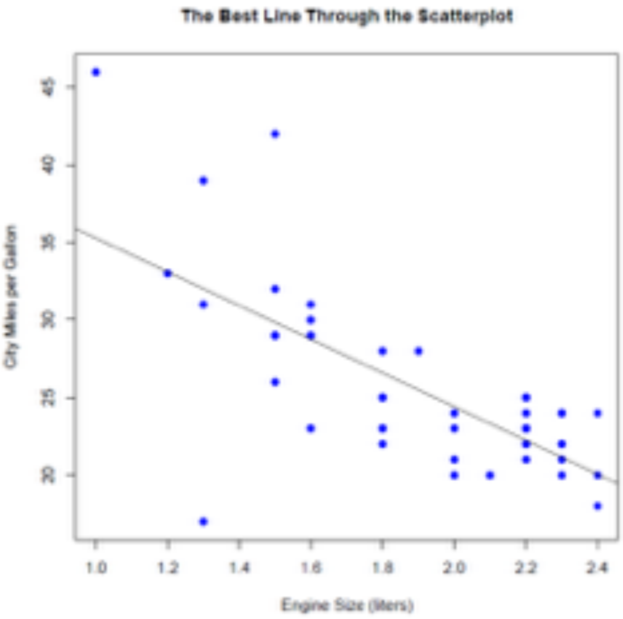
c) Create a scatterplot with MPG.city on the vertical axis, Engine Size on the horizontal axis. Add labels to both axes as well as a main title. Finally, using the abline() function, add a regression line to the scatterplot.

Answer:

```
plot(Cars93$EngineSize, Cars93$MPG.city,

xlab = 'Engine 'Size (liters)',
ylab = 'City Miles per Gallon',
main = 'The Best Line Through the Scatterplot',
pch = 19,
col = 'blue')

abline(slr2)
```



d) There is a possibility of multi-collinearity between two of the variables, Passengers and Weight. Can you think of any other way to explore whether this might be a problem?

```
#Comment. Use the cor() function to find the correlation.
cor(Cars93$Passengers, Cars93$Weight)
## [1] 0.5732935
```

Answer: While a correlation of r = 0.5732935 is a clear and unambiguous indicator of the presence of multicollinearity between these two independent variables, it is not so severe that we cannot conduct the analysis at all. In fact, some authorities report the rule-of-thumb they use as this: if Irl > 0.70—that is, if r > 0.70 or r < –0.70—we would probably not introduce both variables. Since r = 0.5732935 does not fall in that range, we include both independent variables in this analysis.

e) From the regression equation, fill the ANOVA table.

```
#Comment1. Calculations for the first row of missing values.
ss_reg <- sum((fitted(mr1) - mean(Cars93$MPG.city)) ^ 2)
ss_reg
## [1] 1778.247
ms_reg <- ss_reg / 2
ms_reg
## [1] 889.1236
#Comment2. Calculations for the second row of missing values.
ss_res <- sum((resid(mr1)) ^ 2)
ss_res
## [1] 607.6957
ms_res <- ss_res/ (70 - 2 - 1)
ms_res

## [1] 9.070084
#Comment3. Calculation for the F statistic.
f <- ms_reg / ms_res
f
## [1] 98.02815
```

………

5. Assignment

Simple linear regression, prediction: Heart and body weights

1. In the R package MASS there is a dataset called cats. Run the following commands:
library(MASS)
data(cats)
Have a look at the dataset. The variables Bwt and Hwt give the weight of the body (kg) and the heart (g), respectively. There are both male and female cats. Make a dataset with the data from males only.

2. Make a scatterplot of the data for the male cats (Bwt on x-axis, Hwt on y-axis). Does it look reasonable to use a linear regression model for the data?

3. Fit a linear regresison model for the male cats, that allows for prediction of the heart weight given the body weight. Add the fitted regression line to the scatterplot from the previous question.

4. Find the coefficients of the fitted line. How large is the expected difference in heart weight for two cats with a difference of 1 kg in bodyweight? Find a confidence interval for this difference? How large is the expected difference in heart weight for two cats with a difference of 100 g in bodyweight?

5. Use model validation plot to examine if the model is appropriate for the data.

6. Use the estimates to find the expected heart weight for a male cat that weighs 3 kg. Then try the commands (where you replace the name regModel with whatever name you gave the the model fit in question 2).

```
newObs <- data.frame(Bwt=3)
newObs
predict(regModel, newObs)
predict(regModel, newObs, interval=“predict")
```

………