Unit 9 – R for Normal Theory Regression
Homework

SOLUTIONS

Before you Begin
Download from the course website
companies.Rdata
hersdata_small.Rdata

Description of companies.dta

This data set contains 30 observations. It is from a study reported in the January 1981 issue of *Forbes* magazine of the characteristics of the 30 largest chemical companies. The two variables that we will use in a simple linear regression analysis are the following: **eps5** and **salesgr5**

Codebook

Variable Name	Variable Coding	Description
salesgr5	Continuous	**Independent (X)**: Per cent annual compound growth rate of sales, computed from the most recent five years compared with the previous five years.
eps5	Continuous	**Dependent (Y)**: Per cent annual compound growth in earnings per share, computed from the most recent five years compared with the preceding five years.

NOTE!!!
Dear all – in the knitted file below, some commands have been "commented" out. These correspond to instances where I was creating a new variable and wanted to check my work along the way.

Initialize R Studio Session. Clear the Decks

```
setwd("~/Desktop")              # Set working directory
getwd()                         # Check working directory

## [1] "/Users/cbigelow/Desktop"

options(scipen=999)             # Turn off scientific notation
rm(list = ls())                 # Clear the Decks
```

Load Data

```
load(file="companies.Rdata")
load(file="hersdata_small.Rdata")
```

QUESTIONS 1-7 Simple Linear Regression using companies.Rdata

Q1 Label the variable eps5 as y: earnings per share.
Label the variable salesgr5 as x: growth in sales.

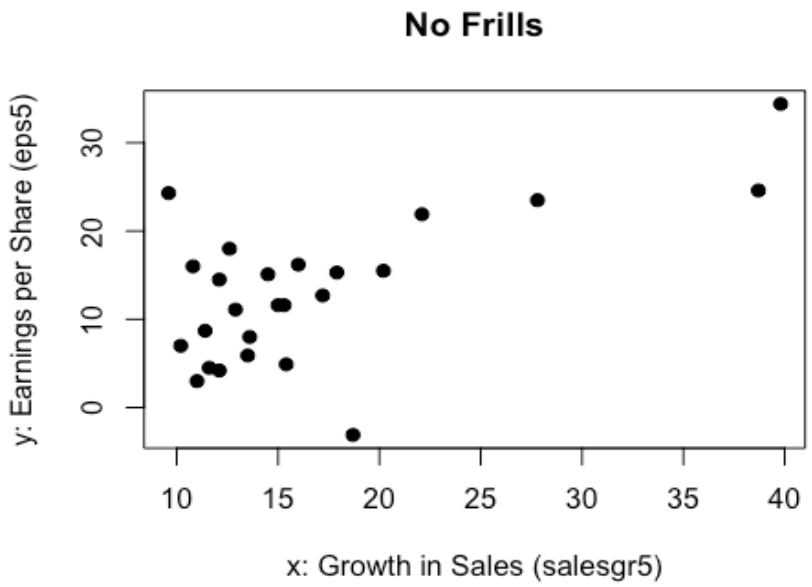
```
library(Hmisc)

Hmisc::label(companies$eps5) <- "y: earnings per share"
Hmisc::label(companies$salesgr5) <- "x: growth in sales"
```

Q2 Produce an XY Scatterplot with any aesthetics you like

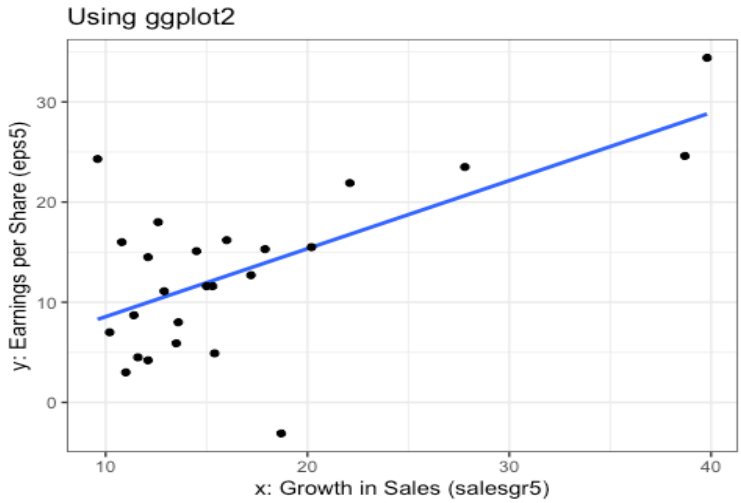
```
library(ggplot2)

# using {base}
# plot(x,y)
plot(companies$salesgr5, companies$eps5,
     main="No Frills",
     xlab="x: Growth in Sales (salesgr5)",
     ylab="y: Earnings per Share (eps5)", pch=19)
```



Intepretation – Scatter plot suggests a linear relationship. It also suggests some sparseness of data for large values of X: Growth in Sales.

```
# Using {ggplot2}
ggplot(data=companies, aes(x=salesgr5,y=eps5)) +
  geom_smooth(method=lm, se=FALSE) +
  geom_point() +
  xlab("x: Growth in Sales (salesgr5)") +
  ylab("y: Earnings per Share (eps5)") +
  ggtitle("Using ggplot2") +
  theme_bw()
```



Intepretation – SAME. Scatter plot suggests a linear relationship. It also suggests some sparseness of data for large values of X: Growth in Sales.

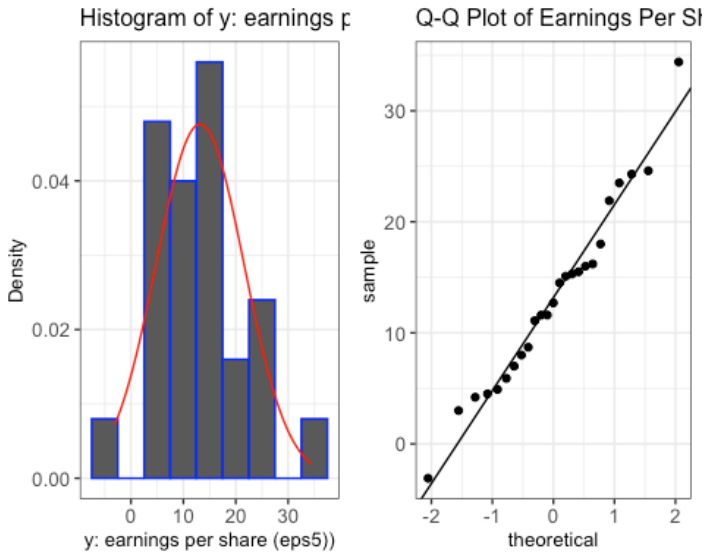
Q3 Graphically: Assess normality of y=eps5

```
library(ggplot2)
library(gridExtra)

# p1 = histogram w overlay normal
# ggplot(DATAFRAME, aes(x=VARIABLENAME)) + optins
p1 <- ggplot2::ggplot(data=companies, aes(x=eps5)) +
         geom_histogram(binwidth=5, colour="blue",
                        aes(y=..density..)) +
         stat_function(fun=dnorm,
                       color="red",
                       args=list(mean=mean(companies$eps5),
                                 sd=sd(companies$eps5))) +
         ggtitle("Histogram of y: earnings per share (eps5)") +
         xlab("y: earnings per share (eps5))") +
         ylab("Density") +
         theme_bw() +
         theme(axis.text = element_text(size = 10),
               axis.title = element_text(size = 10),
               plot.title = element_text(size = 12))
```

```
# p2 = quantile-quantile plot
p2 <- ggplot2::ggplot(data=companies, aes(sample=eps5)) +
          stat_qq() +
          geom_abline(intercept=mean(companies$eps5),
               slope = sd(companies$eps5)) +
          ggtitle("Q-Q Plot of Earnings Per Share (eps5)") +
          theme_bw() +
          theme(axis.text = element_text(size = 10),
               axis.title = element_text(size = 10),
               plot.title = element_text(size = 12))

gridExtra::grid.arrange(p1, p2, ncol=2)
```



Interpretation: The histogram with overlay normal suggests reasonableness of the assumption of normality of Y.

Q4 Numerically: Assess normality of y=eps5

```
# Shapiro-Wilk Test (Null:  normality)
shapiro.test(companies$eps5)

##
##  Shapiro-Wilk normality test
##
## data:  companies$eps5
## W = 0.97392, p-value = 0.7447
```

Interpretation – These test results are consistent with the histogram and overlay normal. The Shapiro-Wilk test of the null hypothesis of normality fails to reject (p-value = .74). We will assume that the assumption of normality of Y is reasonably satisfied for these data. Onward!

Q5 Estimate the Straight Line Regression

```
# MODELNAME <- lm(YVAR ~ XVAR, data=DATAFRAME)
model_q5 <- lm(eps5 ~ salesgr5, data=companies)
summary(model_q5)

##
## Call:
## lm(formula = eps5 ~ salesgr5, data = companies)
##
## Residuals:
## y: earnings per share
##      Min       1Q   Median       3Q      Max
## -17.5665  -3.4511  -0.3534   3.5674  16.0144
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.7650     3.1248   0.565 0.577658
## salesgr5      0.6792     0.1688   4.023 0.000531 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.554 on 23 degrees of freedom
## Multiple R-squared:  0.413,  Adjusted R-squared:  0.3875
## F-statistic: 16.18 on 1 and 23 DF,  p-value: 0.0005314
```

Intepretation: The fitted simple linear regression model is eps5 = 1.76 + 0.679 * salesgr5. For each one percentage increase in annual compound growth of sales, it is estimated that there is a 0.679 percentage increase in annual compound growth in earnings per share.

Q6 How much of the variability is explained by model in Q5? Peform the overall F test.

```
# R-Squared
r2_simple <- summary(model_q5)$r.squared
r2_simple <- 100*round(r2_simple,digits=2)
paste("Percent Variance Explained, R-squared =",r2_simple,"%")

## [1] "Percent Variance Explained, R-squared = 41 %"

# Overall F-Test
anova(model_q5)

## Analysis of Variance Table
##
## Response: eps5
##           Df Sum Sq Mean Sq F value    Pr(>F)
## salesgr5   1 695.26  695.26  16.184 0.0005314 ***
## Residuals 23 988.10   42.96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The R-squared value is 41% which says that 41% of the variability in Y is explained by the fitted model. The value of the overall F-test is 16.184 with an accompanying p-value of .0005. The null hypothesis of zero slope has led to a very unlikely outcome prompting statistical rejection of the null hypothesis. Conclude that the fitted line explains statistically significantly more of the variability in Y=eps5 than is explained by the fit of "no model" (the fit is instead the average of Y).

Q7 Test null: slope = 0 Produce 95% CI of estimated slope

```
# t-test of slope = 0
summary(model_q5)

##
## Call:
## lm(formula = eps5 ~ salesgr5, data = companies)
##
## Residuals:
## y: earnings per share
##      Min       1Q   Median       3Q      Max
## -17.5665  -3.4511  -0.3534   3.5674  16.0144
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.7650     3.1248   0.565 0.577658
## salesgr5      0.6792     0.1688   4.023 0.000531 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.554 on 23 degrees of freedom
## Multiple R-squared:  0.413,  Adjusted R-squared:  0.3875
## F-statistic: 16.18 on 1 and 23 DF,  p-value: 0.0005314

# 95% CI of slope
paste("95% CI Estimate of Slope")

## [1] "95% CI Estimate of Slope"

round(confint(model_q5, 'salesgr5', level=0.95),2)

##          2.5 % 97.5 %
## salesgr5  0.33   1.03

# 95% CI for all terms in model
paste("95% CI Estimates -Intercept and Slope")

## [1] "95% CI Estimates -Intercept and Slope"

round(confint(model_q5, level=0.95),2)

##             2.5 % 97.5 %
## (Intercept) -4.70   8.23
## salesgr5     0.33   1.03
```

t-test of slope: The value of the t-statistic is 4.023 with an accompanying p-value = .0005. This is equivalent to the overall F-test. The null hypothesis has led to a very unlikely outcome, prompting statistical rejection of the null hypothesis. Conclude that the fitted line explains statistically significantly more of the variability in Y=eps5 than is explained by the average of Y (no model).

95% CI estimate of slope = (0.33, 1.03). With 95% confidence I estimate that the slope of Y=eps5 on X=salesgr5 is between 0.33 and 1.03.

QUESTIONS 8-13 Multiple Linear Regression of Y = Glucose among Non-diabetics only

Description of hersdata_small.dta

These data are a simple random sample of 1000 observations from the HERS study called *hersdata_small.dta*. The HERS study was a randomized clinical trial of hormone therapy (estrogen plus progestin) for the reduction of cardiovascular disease risk in post-menopausal women with established coronary disease. Study participants were n=2,763 women who were: (1) post-menopausal (2) with coronary disease; and (3) with an intact uterus. The data set for this homework is a simple random sample of n=1000.

```
library(dplyr)

# some checking before creating restricted data
#class(hersdata_small$diabetes)
#table(hersdata_small$diabetes)

# create dataset for questions 8-13
nondiab <- hersdata_small %>%                                  # Begin with full dataset
    filter(diabetes=="no") %>%                                 # Keep only diabetes="no"
    select(glucose,exercise,age,drinkany,BMI,physact) %>%      # Retain only variables of interest
    na.omit()                                                  # Retain complete observations ONLY
```

Q8 Fit Y=glucose to: exercise among non-diabetics only

```
model_q8 <- lm(glucose ~ exercise, data=nondiab)
summary(model_q8)

##
## Call:
## lm(formula = glucose ~ exercise, data = nondiab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.7569  -6.7569  -0.7569   5.2431  29.1538
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  97.7569     0.4520  216.28 < 0.0000000000000002 ***
## exerciseyes  -1.9107     0.6999   -2.73              0.00648 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.438 on 746 degrees of freedom
## Multiple R-squared:  0.009893,   Adjusted R-squared:  0.008566
## F-statistic: 7.454 on 1 and 746 DF,  p-value: 0.00648
```

Dear class – YIKES! I forgot to create 0/1 indicator for the character variableexercise. R seemed to forgive me! It created the indicator exerciseyes.

(1) The fitted line is Y = 97.76 – (1.82)*exerciseyes for Y=glucose

(2) R-squared = .0099 tells us that less than 1% of the variation in glucose is explained.

(3) Even so, the p-value of .006 for the overall F test (value=7.45) tells us that this fitted model performs better than the null model of using the average Y only.

Q9 Fit Y=glucose to: exercise, age, drinkany, and BMI.

```
model_q9 <- lm(glucose ~ age + drinkany + BMI + exercise, data=nondiab)
summary(model_q9)

##
## Call:
## lm(formula = glucose ~ age + drinkany + BMI + exercise, data = nondiab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.0278  -6.1495  -0.3959   5.1964  31.1003
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 81.20638    4.11626  19.728 < 0.0000000000000002 ***
## age          0.05326    0.05087   1.047                0.295
## drinkanyyes -0.40222    0.67575  -0.595                0.552
## BMI          0.46605    0.06613   7.047     0.00000000000417 ***
## exerciseyes -1.09086    0.68816  -1.585                0.113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.147 on 743 degrees of freedom
## Multiple R-squared:  0.07376,    Adjusted R-squared:  0.06877
## F-statistic: 14.79 on 4 and 743 DF,  p-value: 0.00000000001234
```

Dear class – Ditto. I forgot to create 0/1 indicators for drinkany and exercise. R created for me the 2 indicators I needed: drinkanyyes and exerciseyes

(1) The fitted line is now:
Y = 81.21 + (0.05)*age – (0.40)*drinkanyyes + (0.47)*BMI – (1.09)*exerciseyes

(2) R-squared = .0738 says that 7.4% of the variation in glucose is explained by this model.

(3) The p-value of the overall F-test (value=14.79 is < .00001. So again, the fitted model performs better than the null model of using the average Y only.

Q10 Partial F-test of exercise controlling for: age, drinkany, BMI

```
model_q10 <- lm(glucose ~ age + drinkany + BMI, data=nondiab)
anova(model_q10, model_q9)

## Analysis of Variance Table
##
## Model 1: glucose ~ age + drinkany + BMI
## Model 2: glucose ~ age + drinkany + BMI + exercise
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    744 62375
## 2    743 62165  1    210.24 2.5128 0.1134
```

The partial F-test (null hypothesis: exercise is not significant in the adjusted model) has p-value = .11. Do NOT reject. Conclude that, controlling for age, drink and BMI, exercise is not statistically significant for the prediction of glucose.

Q11 Create design variables for 5 levels of physact

```
library(summarytools)

## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE):
## running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/
## Resources/library/tcltk/libs//tcltk.so'' had status 1

# Create factor variable physactf
mylevels <- c("much less active", "somewhat less active", "about as active",
              "somewhat more active","much more active")
nondiab$physactf <- factor(nondiab$physact,levels=mylevels)


# Create design variables (I created 5 even though I only needed to create 4)
# newvariable <- with(data=dataframe,ifelse(conditionfor1,1,0), na.rm=TRUE)
nondiab$muchless01 <- with(data=nondiab,ifelse(physact=="much less active",1,0),na.rm=TRUE)
nondiab$somewhatless01 <- with(data=nondiab,ifelse(physact=="somewhat less active",1,0),na.rm=TRUE)
nondiab$aboutas01 <- with(data=nondiab,ifelse(physact=="about as active",1,0),na.rm=TRUE)
nondiab$somewhatmore01 <- with(data=nondiab,ifelse(physact=="somewhat more active",1,0),na.rm=TRUE)
nondiab$muchmore01 <- with(data=nondiab,ifelse(physact=="much more active",1,0),na.rm=TRUE)

# CHECK
# What could go wrong:  putting a leading summarytools:: caused a fail.
#with(nondiab, ctable(physactf, muchless01,prop="n",totals=FALSE))
#with(nondiab, ctable(physactf, somewhatless01,prop="n",totals=FALSE))
#with(nondiab, ctable(physactf, aboutas01,prop="n",totals=FALSE))
#with(nondiab, ctable(physactf, somewhatmore01,prop="n",totals=FALSE))
#with(nondiab, ctable(physactf, muchmore01,prop="n",totals=FALSE))
```

Q12 Fit Y=glucose to the design variables for physact you created in Q11

```
# user defined 0/1 design variables
model_q12a <- lm(glucose ~ muchless01 + somewhatless01 + somewhatmore01 + muchmore01, data=nondiab)
paste("Model q12a:  User defined 0/1 variables (referent = about as active)")

## [1] "Model q12a:  User defined 0/1 variables (referent = about as active)"

(summary(model_q12a))$coefficients

##                  Estimate Std. Error     t value    Pr(>|t|)
## (Intercept)    97.6074380  0.6040603 161.5855971 0.000000000
## muchless01      2.9834711  1.5400575   1.9372465 0.053093917
## somewhatless01  0.6497048  1.0981214   0.5916512 0.554264150
## somewhatmore01 -2.1943106  0.8401353  -2.6118537 0.009187097
## muchmore01     -1.1788666  1.1251403  -1.0477507 0.295094173
```

Q13 Fit Y=glucuse to a factor variable physactf. Compare your answer to Q12

```
# By default, R sets REFERENT = level 1 (here physact = "much less active")
model_q13a <- lm(glucose ~ physactf, data=nondiab)
paste("Model q13a:  R default  (referent = much less active)")

## [1] "Model q13a:  R default  (referent = much less active)"

(summary(model_q13a))$coefficients

##                               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                 100.590909   1.416647 71.006339 0.0000000000
## physactfsomewhat less active  -2.333766   1.687563 -1.382921 0.1671046530
## physactfabout as active       -2.983471   1.540058 -1.937247 0.0530939169
## physactfsomewhat more active  -5.177782   1.532262 -3.379175 0.0007649714
## physactfmuch more active      -4.162338   1.705268 -2.440870 0.0148844418

# Tell R to set REFERENT = level 3 (here physact = "about as active")
# Challenge -  You need to know integer storage of your predictor
nondiab <- within(nondiab, physactf <- relevel(physactf, ref = 3))
model_q13b <- lm(glucose ~ physactf, data = nondiab)
paste("Model 13b:  User specifies referent (referent=about as active)")

## [1] "Model 13b:  User specifies referent (referent=about as active)"

(summary(model_q13b))$coefficients

##                               Estimate Std. Error     t value    Pr(>|t|)
## (Intercept)                 97.6074380  0.6040603 161.5855971 0.000000000
## physactfmuch less active     2.9834711  1.5400575   1.9372465 0.053093917
## physactfsomewhat less active  0.6497048  1.0981214   0.5916512 0.554264150
## physactfsomewhat more active -2.1943106  0.8401353  -2.6118537 0.009187097
## physactfmuch more active     -1.1788666  1.1251403  -1.0477507 0.295094173
```

The coefficients (question 12 model and question 13 model) match!

QUESTIONS 14-19 Multiple Linear Regression of Y = LDL on entire sample

```
library(dplyr)

# create dataset for questions 14-19
entire <- hersdata_small %>%                          # Begin with full dataset
          select(LDL,HT,statins) %>%                  # Retain only variables of interest
          na.omit()                                   # Retain complete observations ONLY
```

Q14 - Two way contingency table analysis of HT and statins

```
library(summarytools)

summarytools::ctable(entire$HT,entire$statins,prop="r")

## Cross-Tabulation, Row Proportions
## HT * statins
## Data Frame: entire
##
## ----------------- --------- ------------ ------------ --------------
##                     statins           no          yes          Total
##                HT
##   hormone therapy             331 (65.4%)   175 (34.6%)   506 (100.0%)
##           placebo             303 (61.7%)   188 (38.3%)   491 (100.0%)
##             Total             634 (63.6%)   363 (36.4%)   997 (100.0%)
## ----------------- --------- ------------ ------------ --------------

chisq.test(entire$HT,entire$statins,correct=FALSE)       # NO continuity correction for large n

##
##  Pearson's Chi-squared test
##
## data:  entire$HT and entire$statins
## X-squared = 1.4768, df = 1, p-value = 0.2243
```

The chi square test (null hypothesis: no association) is not statistically significant (p-value = .22). Do NOT reject. In this sample, there is no statistically significant evidence of an association of receipt of hormone treatment with statin use.

Q15 - Single predictor model Y=LDL to X=HT

```
library(summarytools)

# Preliminary - create 0/1 indicator of randomization to hormone therapy (HT="hormone therapy")
# newvariable <- with(data=dataframe,ifelse(conditionfor1,1,0), na.rm=TRUE)
entire$HT01 <- with(data=entire,ifelse(HT=="hormone therapy",1,0),na.rm=TRUE)
#summarytools::ctable(entire$HT,entire$HT01,prop="n",totals=FALSE)

model_q15 <- lm(LDL ~ HT01, data=entire)
summary(model_q15)

##
## Call:
## lm(formula = LDL ~ HT01, data = entire)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -100.538  -23.738   -3.738   21.235  249.035
##
```

```
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 144.3646     1.6635  86.783 <0.0000000000000002 ***
## HT01          0.5738     2.3351   0.246               0.806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.86 on 995 degrees of freedom
## Multiple R-squared:  6.068e-05,  Adjusted R-squared:  -0.0009443
## F-statistic: 0.06038 on 1 and 995 DF,  p-value: 0.8059
```

(1) The fitted line is Y = 144.36 + (0.57)*HT01 for Y=LDL
(2) R-squared < < .0001 tells us that less than .01% of the variation in LDL is explained.
(3) So, not surprisingly, the overall F test (value=0.81) is NOT statistically significant. Do NOT reject. Conclude that LDL is not associated with HT in this sample.

Q16 - Single predictor model Y=LDL to X=statins

```
library(summarytools)

# Preliminary - create 0/1 indicator of use of statins (statins="yes")
# newvariable <- with(data=dataframe,ifelse(conditionfor1,1,0), na.rm=TRUE)
entire$statins01 <- with(data=entire,ifelse(statins=="yes",1,0),na.rm=TRUE)
#summarytools::ctable(entire$statins,entire$statins01,prop="n",totals=FALSE)

model_q16 <- lm(LDL ~ statins01, data=entire)
summary(model_q16)

##
## Call:
## lm(formula = LDL ~ statins01, data = entire)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -106.659  -23.259   -3.872   19.941  242.341
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  151.059      1.425 106.022 < 0.0000000000000002 ***
## statins01    -17.588      2.361  -7.448    0.000000000000205 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.88 on 995 degrees of freedom
## Multiple R-squared:  0.05281,    Adjusted R-squared:  0.05186
## F-statistic: 55.48 on 1 and 995 DF,  p-value: 0.0000000000002046
```

(1) The fitted line is now Y = 151.06 - (17.59)*statins01
(2) R-squared = .0528 says that 5% of the variation in LDL is explained.
(3) Here the p-value for the overall F test is < .0001. Reject. Conclude that this sample gives statistically significant evidence that LDL is associated with use of statins (just as you thought!).
(4) From the fitted line, we see that statin use is associated with lower LDL levels, because the regression coefficient is negative.

Q17 - Create interaction HTstatins that is an interaction of HT and statins.

```
entire$HTstatins <- entire$HT01*entire$statins01
```

Q18 - Fit the 3 predictor model with predictors = HT,statins and HTstatins.

```
model_q18 <- lm(LDL ~ HT01 + statins01 + HTstatins, data=entire)
summary(model_q18)

##
## Call:
## lm(formula = LDL ~ HT01 + statins01 + HTstatins, data = entire)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -106.420  -22.953   -3.921   19.786  242.079
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 151.3208     2.0630  73.350 < 0.0000000000000002 ***
## HT01         -0.5009     2.8552  -0.175                0.861
## statins01   -18.1676     3.3340  -5.449          0.0000000638 ***
## HTstatins     1.1614     4.7308   0.245                0.806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.91 on 993 degrees of freedom
## Multiple R-squared:  0.05287,    Adjusted R-squared:  0.05001
## F-statistic: 18.48 on 3 and 993 DF,  p-value: 0.0000000000114
```

(1) We now have fitted Y = 151.32 – (0.5)*HT01 - (18.17)*statins01 + (1.16)*HTstatins

(2) Oh my goodness. R-squared = .0529 is essentially the same as the R-squared in the simpler model that did not include the interaction (there, R-squared = .0528).

Q19 - Partial F test of HTstatins controlling for HT and statins

```
model_q19 <- lm(LDL ~ HT01 + statins01, data=entire)
anova(model_q19,model_q18)

## Analysis of Variance Table
##
## Model 1: LDL ~ HT01 + statins01
## Model 2: LDL ~ HT01 + statins01 + HTstatins
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    994 1280607
## 2    993 1280529  1    77.718 0.0603 0.8061
```

Just as we expected. The partial F-test (null hypothesis: the addition of the interaction to the model co ntaining the main effects) has p-value = .81. Do NOT reject. Conclude that, in this sample, there is no statistically significant evidence of an interaction of hormone therapy and statin use for prediction of L DL, after controlling for the main effects of hormone therapy and statin use. Note – in a real world anal ysis, we would choose as our final model the single predictor model containing statins