W203-1 - Fall 2014 - Final Exam

Juan Jose Carin

Sunday, December 14, 2014

Contents

Part 1: Multiple Choice

I've included comments in some sections of this Part, (not—only—to justify my choice but) to document my own reasoning when selecting the right answer, for future reference.

1. d. ANOVA

Comment: *ANOVA is the omnibus test for the overall effect. It tests whether at least one of the coefficients (b_1 or b_2 in this case) is not equal to zero.*

2. b. All coefficients for each independent variable equal zero.

Comment: *See the comment to the previous question.*

3. e. None of the above

Comment: *Rejecting the null hypothesis mentioned in the previous answer involves that at least one of the independent variables—but maybe not all, which discards (a), and the ANOVA test does not inform which one of them, and that discards (c)—has a significant contribution to the model. The null hypothesis does not involve b_0, so (b) is also discarded. And maybe a linear relationship may not be the best model for these data—the opposite of (d)—but the ANOVA test could be still significant, making us reject the null hypothesis.*

4. b. Maximum number of pushups in 3 minutes

Comment: *Because this variable is very likely to be highly correlated with X2 (triceps strength), which would involve high collinearity.*

5. d. You examine a random sample of internet users in one Californian county, and compare them to a random sample of internet users in an adjacent California county where the local Internet Service Provider (ISP) has chosen to block access to online dating sites.

Comment: *In a natural experiment, the treatment (the independent variable of interest, which is the use of online dating sites in this case) varies through some naturally occurring or unplanned event that happens to be exogenous to the outcome (the dependent variable of interest, which is the relationship satisfaction in this case).*

6. e. Ecological validity

Comment: *Because it is the extent to which research results can be applied to real life situations outside of research settings. The way the question is phrased, it is very unlikely that any police officer will admit to be biased against a minority—he would probably detect what the question is trying to measure, especially if he has beeen assigned to treatment 1, answering "no" to it... although he would act differently in a real life situation.*

7. b. The power of the test

Comment: *Unlike the other choices, it does not depend on the sample we're working with, but on other factors such as—mainly—the statistical significance criterion used in the test, the magnitude of the effect of interest in the population—i.e., the effect size—, and the sample size used to detect the effect.*

8. a. Random variation in each sample drawn from a population will lead to larger or smaller *p*-values.

Comment: *The statement is true (p-values will vary), but on average the type-1 error should be 5%—supposed this was the selected statistical significance criterion—, not higher. The other statemnts* ***do*** *help explain why more than 5% of published results appear to be type-1 errors.*

Let's see it with an example. Suppose our hypothesis is that men's height is significantly different that women's height. Each variable is normally distributed—it wouldn't have to be the case, we just need to know that the sampling mean will be normally distributed—with a standard deviation equal to 5. Let's say that men's average height is 180 cm, and women's average height is 170. The variance of the difference of both variables will be the sum of their variances ($5^2 + 5^2 = 50$)*. If we take 100 random samples of 50 men and 50 women, the average difference in height will be normally distributed, with a mean of 10* ($180 - 170$) *and a standard deviation of* $\sqrt{50}/\sqrt{50} = 1$*. If then we calculate the probability of finding that mean difference, for the 100 samples, on average we would obtain that only in 5 out of 100 samples we would obtain extreme values with only a 5% probability. To prove my point, I run 100 simulations and calculate the number of them in which there were more than 5 False Positives (which should not exceed 5 in most of the cases).*

```
   # A simulation to prove my answer
false_positives <- rep(0, 100)
for(i in 1:100){
   x <- data.frame(matrix(NA, nrow = 50, ncol = 100))
   y <- x
   x <- apply(x, 2, function(x){rnorm(50, 180, 5)})
   y <- apply(x, 2, function(x){rnorm(50, 170, 5)})
   xydif <- x - y
   xydif_avg <- apply(xydif, 2, mean)
   xydif_prob <- pnorm(xydif_avg, 10, 1)
   false_positives[i] <- length(xydif_prob[abs(xydif_prob) > .975])
}
false_positives
```

```
##   [1] 1 5 0 3 2 3 3 3 4 4 3 2 3 2 6 3 4 4 3 1 0 2 3 6 3 5 1 4 3 2 2 1 4 3 3
##  [36] 3 1 2 1 2 1 3 3 1 6 5 2 0 4 2 2 1 1 1 1 2 2 6 2 1 3 3 2 2 3 4 3 0 3 3
##  [71] 1 1 3 2 3 3 2 1 1 3 0 2 0 3 2 0 4 3 2 1 2 3 4 2 4 1 3 1 3 2
```

```
length(false_positives[false_positives > 5])
```

```
## [1] 4
```

Part 2: Test Selection

I've also included comments in some sections of this Part, (not—only—to justify my choice but) to document my own reasoning when selecting the right answer, for future reference. I've also run some chunks of R script to plot a graph or perform the statistical procedure which I think it's the most appropriate, to illustrate my choice—the cleaning just deals with missing values (without worrying about whether non-missing values make sense or not, as I do in **Part 3**), and the tests may be not give good or significant results (but, from my point of view, they are the most appropriate ones for this dataset and the statement quoted in each point).

1. a. chi-square

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## |           Row Percent |
## |          Column Percent |
## |           Total Percent |
## |-------------------------|
##
## Total Observations in Table:  1796
##
##                               | Dating_sample1$use_reddit
## Dating_sample1$marital_status |       No  |      Yes  | Row Total |
## ------------------------------|-----------|-----------|-----------|
##                      Divorced |      182  |       12  |      194  |
##                               |  183.414  |   10.586  |           |
##                               |   93.814% |    6.186% |   10.802% |
##                               |   10.718% |   12.245% |           |
##                               |   10.134% |    0.668% |           |
## ------------------------------|-----------|-----------|-----------|
##           Living with partner |       96  |        5  |      101  |
##                               |   95.489  |    5.511  |           |
##                               |   95.050% |    4.950% |    5.624% |
##                               |    5.654% |    5.102% |           |
##                               |    5.345% |    0.278% |           |
## ------------------------------|-----------|-----------|-----------|
##                       Married |      894  |       35  |      929  |
##                               |  878.308  |   50.692  |           |
##                               |   96.233% |    3.767% |   51.726% |
##                               |   52.650% |   35.714% |           |
##                               |   49.777% |    1.949% |           |
## ------------------------------|-----------|-----------|-----------|
##            Never been married |      366  |       43  |      409  |
##                               |  386.683  |   22.317  |           |
##                               |   89.487% |   10.513% |   22.773% |
##                               |   21.555% |   43.878% |           |
##                               |   20.379% |    2.394% |           |
## ------------------------------|-----------|-----------|-----------|
##                     Separated |       42  |        2  |       44  |
##                               |   41.599  |    2.401  |           |
##                               |   95.455% |    4.545% |    2.450% |
##                               |    2.473% |    2.041% |           |
##                               |    2.339% |    0.111% |           |
## ------------------------------|-----------|-----------|-----------|
##                       Widowed |      118  |        1  |      119  |
##                               |  112.507  |    6.493  |           |
##                               |   99.160% |    0.840% |    6.626% |
##                               |    6.949% |    1.020% |           |
##                               |    6.570% |    0.056% |           |
## ------------------------------|-----------|-----------|-----------|
##                  Column Total |     1698  |       98  |     1796  |
##                               |   94.543% |    5.457% |           |
## ------------------------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  30.64787     d.f. =  5     p =  1.099259e-05
##
```

```
##
##
##        Minimum expected frequency: 2.400891
## Cells with Expected Frequency < 5: 1 of 12 (8.333333%)
```

2. c. Logistic regression

```
##
## Call:
## mlogit(formula = region ~ 1 | life_quality, data = Dating_sample2,
##     method = "nr", print.level = 0)
##
## Frequencies of alternatives:
##   Midwest Northeast     South      West
##   0.23029   0.16577   0.38665   0.21729
##
## nr method
## 4 iterations, 0h:0m:0s
## g'(-H)^-1g = 0.0014
## successive function values within tolerance limits
##
## Coefficients :
##                          Estimate Std. Error t-value Pr(>|t|)
## Northeast:(intercept)  -0.207886   0.176830 -1.1756 0.239743
## South:(intercept)       0.476757   0.145882  3.2681 0.001083 **
## West:(intercept)        0.348655   0.162110  2.1507 0.031498 *
## Northeast:life_quality -0.046082   0.062329 -0.7393 0.459700
## South:life_quality      0.015581   0.050746  0.3070 0.758807
## West:life_quality      -0.159179   0.058460 -2.7229 0.006472 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -2974
## McFadden R^2:  0.0020444
## Likelihood ratio test : chisq = 12.185 (p.value = 0.0067757)
```

```
##                        exp.model1.coefficients.
## Northeast:(intercept)                 0.8122996
## South:(intercept)                     1.6108414
## West:(intercept)                      1.4171595
## Northeast:life_quality                0.9549637
## South:life_quality                    1.0157035
## West:life_quality                     0.8528434
```

```
##                            2.5 %    97.5 %
## Northeast:(intercept)  0.5743791 1.1487720
## South:(intercept)      1.2102569 2.1440158
## West:(intercept)       1.0314075 1.9471849
## Northeast:life_quality 0.8451479 1.0790485
## South:life_quality     0.9195424 1.1219206
## West:life_quality      0.7605154 0.9563801
```

3. d. Multiple Regression

Comment: *Though the proper answer would have been* ***Simple Regression*** *(but we can also say the latter is just an specific case of the multiple regression, when there is only one independent variable).*

(a) is discarded because Fisher's exact test is a version of the chi-square test designed for small samples and 2x2 contingency tables—our sample is quite big and, above all, the variable years_in_relationship is not categorical but continuous. (b) is discarded because the assumptions of homogeneity of variance and normality are not met. (c) is discarded because the scores of years_in_relationship come from different respondents, not the same at different moments.

```
## Dating_sample3[, "flirted_online"]: No
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.8573, p-value < 2.2e-16
```

```
##
## ------------------------------------------------------------
## Dating_sample3[, "flirted_online"]: Yes
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.5818, p-value < 2.2e-16


## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    1  321.15 < 2.2e-16 ***
##       1839
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


##
## Call:
## lm(formula = years_in_relationship ~ flirted_online, data = Dating_sample3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.959 -14.959  -3.216   8.041  81.041
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        15.9595     0.4049   39.42   <2e-16 ***
## flirted_onlineYes -12.7439     0.8853  -14.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.45 on 1839 degrees of freedom
## Multiple R-squared:  0.1013, Adjusted R-squared:  0.1008
## F-statistic: 207.2 on 1 and 1839 DF,  p-value: < 2.2e-16
```

Figure 1: Bar chart of the mean number of years in relationship depending on whether the respondent has flirted online or not

4. b. ANOVA

Comment: *Since the outcome variable (adults_in_household) is continuous, and the predictor variable (lgbt) is categorical, with more than 2 categories.*

The most appropriate test would be a ***robust ad-hoc*** *version of ANOVA, because the assumption of homogeneity of variance is not met, and there is not an specific hypothesis to make a planned contrast.*

```
## Dating_sample4$lgbt: Bisexual
## [1] 2.309524
## ------------------------------------------------------------
## Dating_sample4$lgbt: Gay
## [1] 2.111111
```

```
## ------------------------------------------------------------
## Dating_sample4$lgbt: Straight
## [1] 2.054321
## ------------------------------------------------------------
## Dating_sample4$lgbt: Other
## [1] 1.6
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lgbt           3    3.8  1.2771   1.527  0.206
## Residuals   2104 1759.8  0.8364
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Dating_sample4$adults_in_household and Dating_sample4$lgbt
##
##          Bisexual Gay  Straight
## Gay      1.00     -    -
## Straight 0.44     1.00 -
## Other    0.61     1.00 1.00
##
## P value adjustment method: bonferroni
```

```
## Dating_sample4[, "lgbt"]: Bisexual
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.8645, p-value = 0.0001464
##
## ------------------------------------------------------------
## Dating_sample4[, "lgbt"]: Gay
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.8102, p-value = 2.618e-05
##
## ------------------------------------------------------------
## Dating_sample4[, "lgbt"]: Straight
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.7939, p-value < 2.2e-16
##
## ------------------------------------------------------------
## Dating_sample4[, "lgbt"]: Other
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.7709, p-value = 0.04595
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value  Pr(>F)
## group    3  2.0965 0.09868 .
##       2104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "Taking bootstrap samples. Please wait."
```

```
## $psihat
##      con.num    psihat        se   ci.lower  ci.upper p-value
## [1,]       1 0.1083916 0.2299607 -0.4720280 0.7517483  0.5675
## [2,]       2 0.2287433 0.2000956 -0.2770180 0.7696739  0.2530
## [3,]       3 0.8205128 0.4892994 -0.7179487 1.6153846  0.2105
## [4,]       4 0.1203517 0.1187112 -0.2085297 0.4303030  0.3220
## [5,]       5 0.7121212 0.4620260 -0.8030303 1.2727273  0.2600
## [6,]       6 0.5917695 0.4479111 -0.7818930 0.9827160  0.4060
##
## $crit.p.value
## [1] 0.009
##
```

```
## $con
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    1    1    0    0    0
## [2,]   -1    0    0    1    1    0
## [3,]    0   -1    0   -1    0    1
## [4,]    0    0   -1    0   -1   -1
```

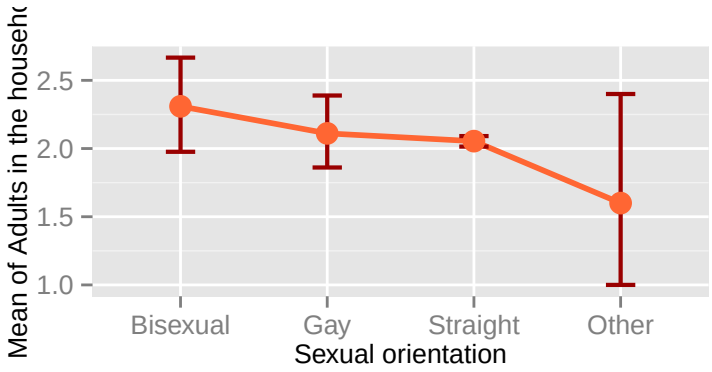

Figure 2: Line chart with error bars of the mean number of adults in the respondent's household depending on sexual orientation

5. a. Pearson correlation

Comment: *Because both variables are continuous.*

```
##
##  Pearson's product-moment correlation
##
## data:  Dating_sample6$age and Dating_sample6$children0_17
## t = -1.8815, df = 542, p-value = 0.06044
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.163514462  0.003535262
## sample estimates:
##         cor
## -0.08055523
```

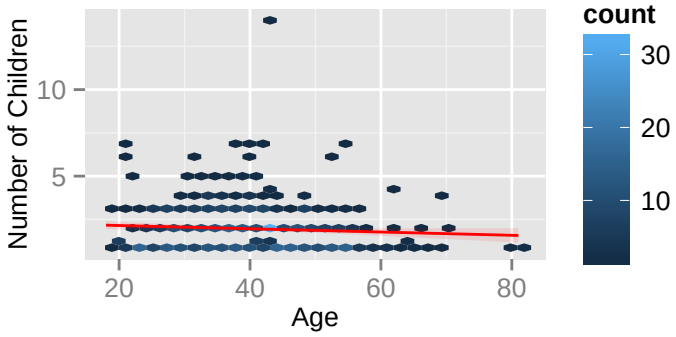

Figure 3: Scatterplot of age against number of children with a regression line (and 95% confidence interval) added)

6. b. Wilcoxon Rank-Sum Test

Comment: *Since the outcome variable (number of children) is continuous, the predictor variable (sex) is categorical (with only 2 categories), and the assumption of normality is not met in the Femlae group (besides, there are very few observations in the sample).*

```
## Dating_sample7[, "sex"]: Female
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.684, p-value = 0.00647
##
## ------------------------------------------------------------
## Dating_sample7[, "sex"]: Male
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.9643, p-value = 0.6369
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  1.2656 0.3036
##        6
```

```
## Dating_sample7[, "sex"]: Female
## [1] 2
## ------------------------------------------------------------
## Dating_sample7[, "sex"]: Male
## [1] 3
```

```
##
##  Wilcoxon rank sum test
##
## data:  children0_17 by sex
## W = 1.5, p-value = 0.055
## alternative hypothesis: true location shift is not equal to 0
```

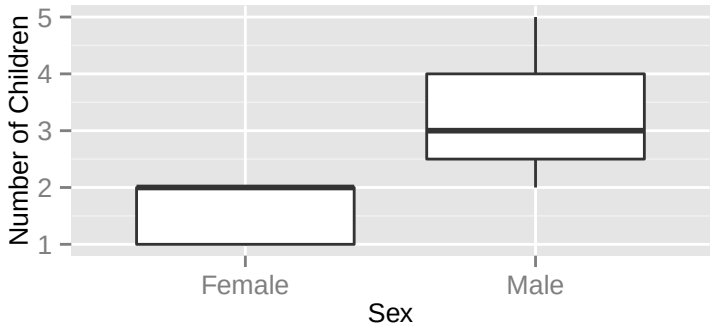

Figure 4: Boxplot of number of children at 31, split by gender

Part 3: Data Analysis

1. OLS Regression

a. The *life_quality* variable measures quality of life on a 5-point scale, where $1 = excellent$ and $5 = poor$. Reverse the scale so that $5 = excellent$ and $1 = poor$. What is the mean quality of life in the sample?

First, we reverse the scale.

```
   # The followind code also works but I prefer the one that I've finally used
# life_quality <- as.character(Dating$life_quality)
# life_quality <- ifelse(life_quality == "5", "now1", life_quality)
# life_quality <- ifelse(life_quality == "4", "now2", life_quality)
# life_quality <- ifelse(life_quality == "2", "now4", life_quality)
# life_quality <- ifelse(life_quality == "1", "now5", life_quality)
# life_quality <- ifelse(life_quality == "now1", "1", life_quality)
# life_quality <- ifelse(life_quality == "now2", "2", life_quality)
# life_quality <- ifelse(life_quality == "now4", "4", life_quality)
# life_quality <- ifelse(life_quality == "now5", "5", life_quality)
# life_quality <- factor(life_quality)
# life_quality[life_quality == "Refused" | life_quality == "Don't know"] <- NA
# life_quality <- factor(life_quality)

life_quality <- Dating$life_quality
life_quality[life_quality == "Refused" | life_quality == "Don't know"] <- NA
life_quality <- factor(life_quality)
life_quality <- factor(life_quality, levels=rev(levels(life_quality)))
levels(life_quality) <-
   as.character((4*as.numeric(levels(life_quality))) %% 5+1)
```

Then we make a small test to check we've done that correctly (i.e., 5 is recoded as 1, 1 as 5, 4 as 2, and 2 as 4... and the sum of the numeric values is also correct).

```
head(Dating$life_quality)
```

```
## [1] 2 2 3 5 3 4
## Levels: 1 2 3 4 5 Don't know Refused
```

```
as.numeric(Dating$life_quality[1]) + as.numeric(Dating$life_quality[2])
```

```
## [1] 4
```

```
head(life_quality)
```

```
## [1] 4 4 3 1 3 2
## Levels: 1 2 3 4 5
```

```
as.numeric(life_quality[1]) + as.numeric(life_quality[2])
```

```
## [1] 8
```

Finally, we calculate the mean value of the recoded variable.

```r
mean(as.numeric(life_quality), na.rm = TRUE)
```

```
## [1] 3.392921
```

As shown above, the **mean quality of life in the sample** now is **3.39**.

We'll use these recoded values in the rest of the assignment.

b. What is the mean of *years_in_relationship* in the sample?

```r
years_in_relationship <- Dating$years_in_relationship
years_in_relationship <- as.numeric(as.character(years_in_relationship))

mean(years_in_relationship, na.rm = TRUE)
```

```
## [1] 13.47697
```

As shown above, the **mean of Years in Relationship in the sample** (after converting the variable to a character string before converting it to a numeric vector), is now **13.48**.

c. To run a nested regression in R, your first step will be to select just the rows in your datase that have no missing values in your final OLS model. In this case, you will want just the rows that have non-missing values for *life_quality*, *years_in_relationship*, and *use_internet*. How many cases does this leave you with?

First we check how many cases we originally had, and how many values of the variables *life_quality* and *years_in_relationship* are non-missing.

```r
dim(Dating)[1]
```

```
## [1] 2252
```

```r
length(life_quality[!is.na(life_quality)])
```

```
## [1] 2232
```

```r
length(years_in_relationship[!is.na(years_in_relationship)])
```

```
## [1] 2193
```

Then—as we already did with the two previous variables—we recode *use_internet* and check how many of its cases are non-missing (for every variable we assume that not only missing values, but also "Don't know" or "Refused" are equivalent to NAs because they do not provide any information).

```r
use_internet <- Dating$use_internet
use_internet[use_internet == "Refused" | use_internet == "Don't know" |
               use_internet == " "] <- NA
use_internet <- factor(use_internet)
length(use_internet[!is.na(use_internet)])
```

```
## [1] 1126
```

The final step is to put the three variables in the same (new) dataframe and discard the missing values.

```
Dating_P3_preliminar <- data.frame(life_quality, years_in_relationship,
                                   use_internet)
Dating_P3 <- Dating_P3_preliminar[complete.cases(Dating_P3_preliminar), ]
dim(Dating_P3)[1]
```

```
## [1] 1090
```

```
Dating_P3$life_quality <- as.numeric(Dating_P3$life_quality)
```

As shown above, we have **1090 cases left**.

But... if we make a more detailed analysis, we find out that some of the left values do not make sense: some of these cases correspond to people who have had a relationship longer than their whole life, or that started it when they were children.

Assuming that nobody would start a relationship when he or she is younger than 10-years old, we find out 3 possible cases that may contain wrong data (people that started their current relationship when they were 5, 0... and –16 years old!).

```
Dating_P3_preliminar$age <- Dating$age
subset(Dating_P3_preliminar,
       age -10 < years_in_relationship)
```

```
##     life_quality years_in_relationship use_internet age
## 321            5                    20           No  25
## 365            3                    86           No  86
## 366            3                    97          Yes  81
```

```
Dating_P3_preliminar <- Dating_P3_preliminar[Dating_P3_preliminar$age >
                                               10 + years_in_relationship, ]
Dating_P3_rev <- Dating_P3_preliminar[complete.cases(Dating_P3_preliminar), ]
dim(Dating_P3_rev)[1]
```

```
## [1] 1087
```

Anyway, we'll use the 1090 cases in the rest of the assignment (assuming it was the age which was not properly entered in some cases).

d. Fit an OLS model to the data from the previous step that predicts *life_quality* as a linear function of *years_in_relationship*. What is the slope coefficient you get? Is it statistically significant? What about practically significant?

First of all, let's see the main statistics of both variables, and a couple of graphs reprenting them.

```
(round(stat.desc(Dating_P3$life_quality, basic = FALSE), 3))
```

```
##       median         mean      SE.mean CI.mean.0.95          var
##        3.000        3.394        0.033        0.065        1.200
##      std.dev     coef.var
##        1.095        0.323
```

```
round(stat.desc(Dating_P3$years_in_relationship, basic = FALSE), 3)
```

```
##       median         mean      SE.mean CI.mean.0.95          var
##        4.000       12.883        0.509        0.999      282.405
##      std.dev     coef.var
##       16.805        1.304
```

Note that the mean (and the rest of statistics) of both *life_quality* and *years_in_relationship* are different to those reported in ***(a)*** and ***(b)***, because we are now considering less cases (discarding those with missing values in the other variable as well as in *use_internet*).

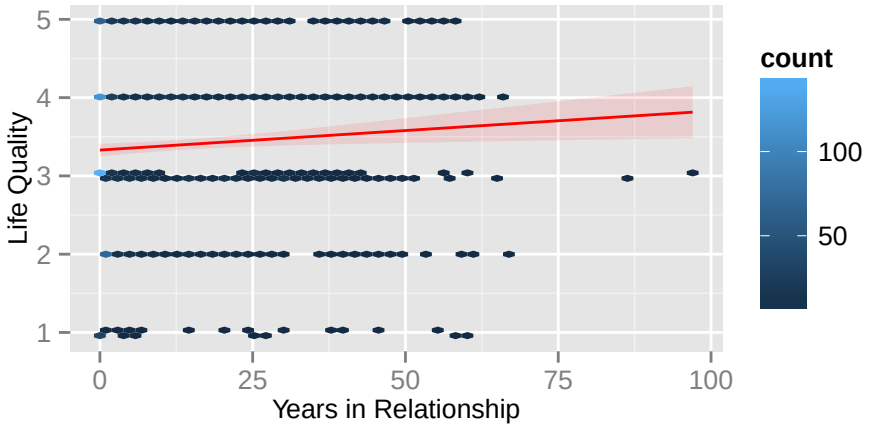

Figure 5: Scatterplot (using hexagon binning) of Years in Relationship against Life Quality with a regression line (and 95% confidence interval) added)

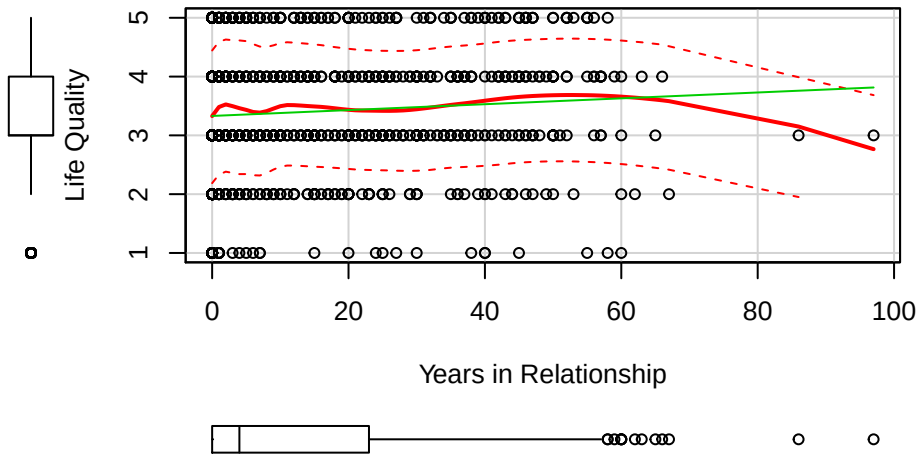

Figure 6: Scatterplot of Years in Relationship against Life Quality with linear and non-linear regression curves added)

It seems that most of the people in the sample (more than 62% of them) have a Life Quality of 3 or 4. There is also a lot of cases (more than 38%) whose Years in Relationship is 0. Overall, there seems to be a linear trend, so as years in relationsip increase, life quality grows (from 3 to 4).

The next step is to build the model and quantify its (statistical and practical) significance.

```
model1 <- lm(life_quality ~ years_in_relationship, Dating_P3)
(summary_model1 <- summary(model1))
```

```
##
## Call:
## lm(formula = life_quality ~ years_in_relationship, data = Dating_P3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6296 -0.4799 -0.3302  0.6698  1.6698
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.33022    0.04170  79.853   <2e-16 ***
## years_in_relationship  0.00499    0.00197   2.533   0.0115 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.093 on 1088 degrees of freedom
## Multiple R-squared:  0.005861,   Adjusted R-squared:  0.004947
## F-statistic: 6.414 on 1 and 1088 DF,  p-value: 0.01146
```

```
confint(model1)
```

```
##                              2.5 %      97.5 %
## (Intercept)           3.248386371 3.412046810
## years_in_relationship 0.001123941 0.008855252
```

The slope coefficient is 0.0049896 (positive, as the previous graphs suggested—so the higher the number of years in relationship, the higher the life quality): as years in relationship increase by one unit, life quality increases by 0.005 units. It is **statistically significant** ($p = 0.011 < 0.05$, and hence—as shown above—its confidence interval does not cross 0—though the lower bound is quite close to it).

Another way to see the relationship between both variables is using the standardized beta estimates:

```
lm.beta(model1)
```

```
## years_in_relationship
##            0.07655665
```

As years in relationship increase by one standard deviation (16.8), life quality increases by 0.0766 standard deviations ($0.0766 * 1.095 = 0.084$). I.e., for every 16.8 years that a respondent spends in a relationship, his or her life quality increases by less than 0.1 units. So **the practical significance is small**.

Aditionally, let's check the goodness of fit. The value of R-squared is about 0.0059: the years in relationship accounts for less than 0.6% of the variation in life quality. (The adjusted R-squared is 0.0049, quite similar to the previous one, so the cross-validity of the model is quite good.)

e. Now fit a second OLS model to the data. Keep *life_quality* as your dependent variable, but now use both *years_in_relationship* and *use_internet* as your explanatory variables. What is the slope coefficient for *use_internet*? Is it statistically significant? What about practically significant?

First of all, let's see the values of *use_internet*, compare the mean of *life_quality* depending on whether the respondent uses Internet or not, and plot the new scatterplot.

```
##  No Yes
## 180 910
```

```
##
##  Welch Two Sample t-test
##
## data:  life_quality by use_internet
## t = -4.3547, df = 243.459, p-value = 1.964e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5896189 -0.2223469
## sample estimates:
##  mean in group No mean in group Yes
##          3.055556          3.461538
```

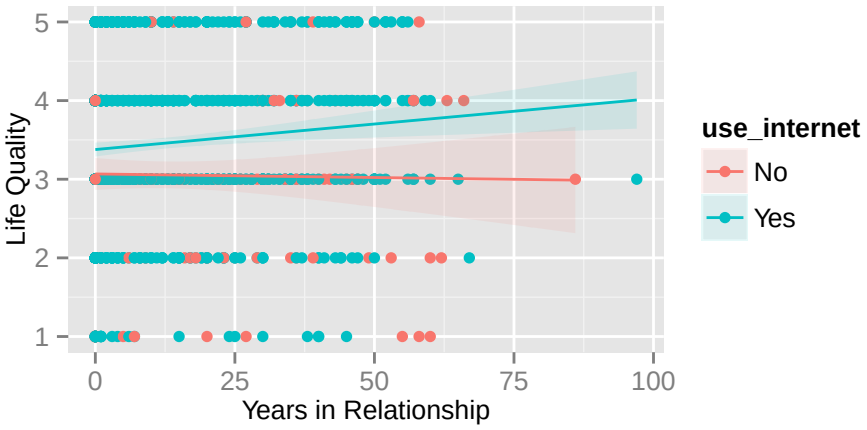

Figure 7: Scatterplot of Years in Relationship against Life Quality, depending on the Use of Internet, with a regression line (and 95% confidence interval) added)

Now we build the new model, adding that variable as a predictor of the Life Quality.

```
##
## Call:
## lm(formula = life_quality ~ years_in_relationship + use_internet,
##     data = Dating_P3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.61852 -0.53523 -0.01881  0.60195  2.00568
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.994316   0.084309  35.516  < 2e-16 ***
## years_in_relationship 0.004899   0.001952   2.509   0.0122 *
## use_internetYes       0.403738   0.088325   4.571 5.41e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.083 on 1087 degrees of freedom
## Multiple R-squared:  0.02461,    Adjusted R-squared:  0.02282
## F-statistic: 13.71 on 2 and 1087 DF,  p-value: 1.314e-06
```

```
##                             2.5 %      97.5 %
## (Intercept)           2.828888461 3.159742801
## years_in_relationship 0.001068203 0.008730185
## use_internetYes       0.230430325 0.577045456
```

The slope coefficient for *use_internet* is 0.4037379. So (assuming no change in *years_in_relationship*) as the use of Internet increases by one unit, life quality increases by 0.404 units. Since *use_internet* is a dichotomous variable, that means that using Internet implies an increase of 0.404 units in life quality, on average. This coefficient is **highly statistically significant** ($p = 5.41e-06$). And that moderate change in **life_quality** involves a **medium practical significance**.

The value of R-squared is now about 0.0246: the two predictors account for less than 2.5% of the variation in life quality. (The adjusted R-squared is 0.0228, quite similar to the previous one, so the cross-validity of the model is quite good.)

f. Compute the *F*-ratio and associated *p*-value between your two regression models. Assess the improvement from your first model to your second.

```
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: life_quality ~ years_in_relationship
## Model 2: life_quality ~ years_in_relationship + use_internet
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1   1088 1298.7
## 2   1087 1274.2  1    24.493 20.894 5.41e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(model1)
```

```
## [1] 3290.248
```

```
AIC(model2)
```

```
## [1] 3271.494
```

The *F* statistic is 20.89 and its associated *p*-value is 5.4e-06. That means that the second model (statistically) significantly improves the fit of the model to the data compared to the first model.

2. Logistic Regression

a. What are the odds that a respondent in the sample has flirted online at some point (*flirted_online*)?

These odds are the probability of having flirted online divided by the probability of not having done so. All we need to calculate them is the number of respondents who have flirted online at some point, and the number of respondents who have never flirted online.

$$odds = \frac{P(flirted\ \ online)}{1-P(flirted\ \ online)} = \frac{respondents\ \ who\ \ have\ \ flirted\ \ online\ \ /\ \ TOTAL}{respondents\ \ who\ \ have\ \ not\ \ flirted\ \ online\ \ /\ \ TOTAL}$$

$$odds = \frac{respondents\ \ who\ \ have\ \ flirted\ \ online}{respondents\ \ who\ \ have\ \ not\ \ flirted\ \ online}$$

```
flirted_online <- Dating$flirted_online
flirted_online[flirted_online == "Refused" | flirted_online == "Don't know" |
                 flirted_online == " "] <- NA
flirted_online <- factor(flirted_online)

(odds <-
    length(flirted_online[!is.na(flirted_online) & flirted_online == "Yes"]) /
    length(flirted_online[!is.na(flirted_online) & flirted_online == "No"]))
```

```
## [1] 0.2613636
```

As shown above, **the odds are 0.261** (considering the 1887 respondents in the sample that answered affirmatively or negatively). I.e., for every 100 respondents who have not flirted online, there are only about 26 who have flirted online (or that, for every respondent who have flirted online, there are almost 4 who have not).

b. Conduct a logistic regression to predict *flirted_online* as a function of where a respondent lives (*usr*). What Akaike Information Criterion (AIC) does your model have?

Before we conduct the regression, let's also plot the number of respondents who have flirted onine or not depending on the area they live.

```
##
##       Rural Suburban Urban
##   No    302      708   485
##   Yes    48      180   162
```

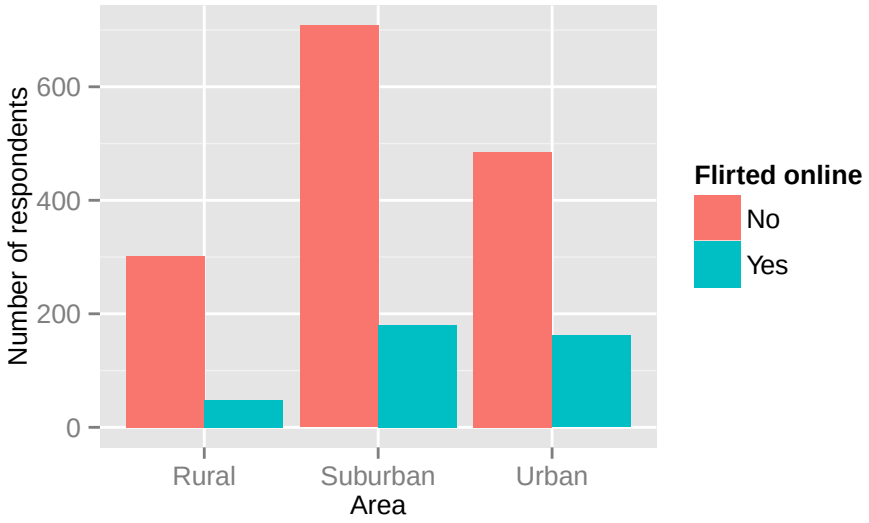

Figure 8: Count of respondents depending on the Area they live, and whether they have flirted online or not

Note: Now the sample is reduced to 1885 respondents since 2 of them did not report the area where they live.

The previous plot already gives us an idea of how the area where a respondent lives is related to having flirted online: the odds of not having flirted online seem to be about 6—this is just a very rough estimation based on the plot and the table above—in rural areas, 4 in suburban areas, and 3 in urban areas. Put it the other way, the odds of having flirted online seem to increase when moving from rural to urban areas.

```
##
## Call:
## glm(formula = flirted_online ~ usr, family = binomial(), data = Dating_P4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7592  -0.7592  -0.6731  -0.5432   1.9934
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.8392     0.1554 -11.837  < 2e-16 ***
## usrSuburban   0.4697     0.1764   2.663  0.00774 **
## usrUrban      0.7427     0.1799   4.127 3.67e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1922.0  on 1884  degrees of freedom
## Residual deviance: 1903.4  on 1882  degrees of freedom
## AIC: 1909.4
##
## Number of Fisher Scoring iterations: 4
```

The coefficients for Suburban and Urban (being Rural the baseline) are positive (and both statistically significant, even when the Wald statistics are sometimes underestimated), which corroborates our previous estimation based on the plot.

As shown above, **the AIC value is 1909.4**. It is 6 units higher than the residual deviance—i.e., the deviance of the model we've built—because the predictor variable has 3 categories (and, as we know, $AIC = deviance + 2k = -2LL + 2k$).

The deviance of the model is lower than the null deviance, so this model we've built is better at predicting whether a respondent has flirted online than the null model. And this overall improvement—compared to the null model—is (statistically) significantly better:

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: flirted_online
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                  1884     1922.0
## usr   2   18.646      1882     1903.4 8.934e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. According to your model, how much bigger are the odds that an urban respondent has flirted online than the odds that a rural respondent has flirted online? Is this effect practically significant?

We need to calculate the **Odds Ratio**:

```
## (Intercept) usrSuburban    usrUrban
##   0.1589404   1.5995763   2.1015464
```

```
##                2.5 %    97.5 %
## (Intercept) 0.115820 0.2132617
## usrSuburban 1.140536 2.2802470
## usrUrban    1.487442 3.0154854
```

The odds than an urban respondent has flirted online are about 2.1 times bigger than the odds that a rural respondent has flirted one. I.e., a 110% increase in the odds from living in a Urban area rather than in a Rural one.

The value is much bigger than 1 (and the confidence interval does not cross 1), so **the effect** of living in a Urban area (compared to a Rural one) **is practically significant**.