Code

36-309 / 36-749 Homework 9: Review, Chi-Squared Tests, and Logistic Regression

Due Wednesday, November 23, 11:59pm

Sanaz Saaadatifar

This homework focuses on data I've analyzed in my research. The dataset contains information about reddit.com (reddit.com).

For those not familiar with reddit.com (hereafter called Reddit), I will give a brief description here. Reddit is a website where users anonymously submit posts – questions, jokes, recipes, memes, gifs, and many other things – and other users reply with "upvotes", "downvotes", and comments. Within Reddit, there are many "subreddits," which are basically subforums dedicated to particular topics.

Some of my collaborators and I are interested in assessing to what extent there are gender biases on Reddit. For example, are Reddit posts made by women treated differently than posts made by men? This is difficult to assess because Reddit is anonymous, and thus we usually do not know users' gender. However, on the /r/relationships subreddit – where users ask for advice about relationships (which could be romantic, platonic, professional, etc.) – it is very common for users to "declare" their gender. In this way, the /r/relationships subreddit is uniquely different from most other parts of Reddit. For example, someone may make a post with the title, "My [25F] roommate [24M] refuses to do the dishes." In this case, the poster has "declared" that they are a 25-year-old female with a 24-year-old male roommate. A while back, my collaborators and I downloaded every Reddit post ever made, focused on the /r/relationships subreddit, and algorithmically labeled each post as being made by a "male" or "female" author (or at least the ones that had a [F] or [M] label, like in the aforementioned example). We ignored non-binary posters, because there did not seem to be a consistent, systematic way (like the [M]/[F] format) that these posters declared their gender.

In this homework, you will focus on a random set of 10,000 posts from this subreddit. Here's the dataset:

Hide

```
reddit = read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall2022/main/redditData36309.csv")
#ensure categorical variables are factors
reddit$postID = factor(reddit$postID)
reddit$male = factor(reddit$male)
reddit$success = factor(reddit$success)
```

Here are the columns/variables in this dataset:

- `postID` : a unique number that (uncreatively) ranges from 1 to 10000.
- `num_comments` : The number of comments that post received.
- `word.count` : The length (in words) of the post.
- `male` : An indicator variable equal to one if the author has declared themselves as male and equal to zero if they've declared themselves as female.
- `success` : An indicator variable equal to one if the author received 18 or more comments and equal to 0 otherwise.

In Questions 1 and 2, we will use `num_comments` as the outcome variable. In Question 3, we will use `success` as the outcome variable. The first half of this homework is review of old material to help prepare you as we move towards the end of the semester; the second half of the homework is about new material.

Question 1: Internal, External, and Construct Validity (18 points)

a. (8pts) For this part, answer the following two questions.

- (4pts) In this homework, we will use the number of comments as the "construct" for the reception of a post. In other words, more comments will be considered "better reception" for a post. Do you think number of comments has good construct validity? State Yes or No, and discuss in 1-2 sentences.

[No. number of comments cannot necessarily be considered as a better reception. Maybe it is the opposite. Maybe people are commenting more to show their high disagreement or dis reception for a post.]

- (4pts) In this homework, we are focusing on the /r/relationships subreddit. Thus, our findings are restricted to this subreddit, but the hope is that these findings can be generalized to other parts of Reddit (i.e., other subreddits about different topics). In this sense, do you think that this dataset has good external validity? Discuss in 1-2 sentences. (**Hint**: Everything you need to know to answer this question is in the description of the homework, so you do not need to be familiar with Reddit to answer this question.)

[no, I do not think it has high external validity. Because the subjects are different, the target audience of the relational reddit might be different from audience of another reddit topic, so we cannot generalize the relational reddit's results to other reddit posts.]

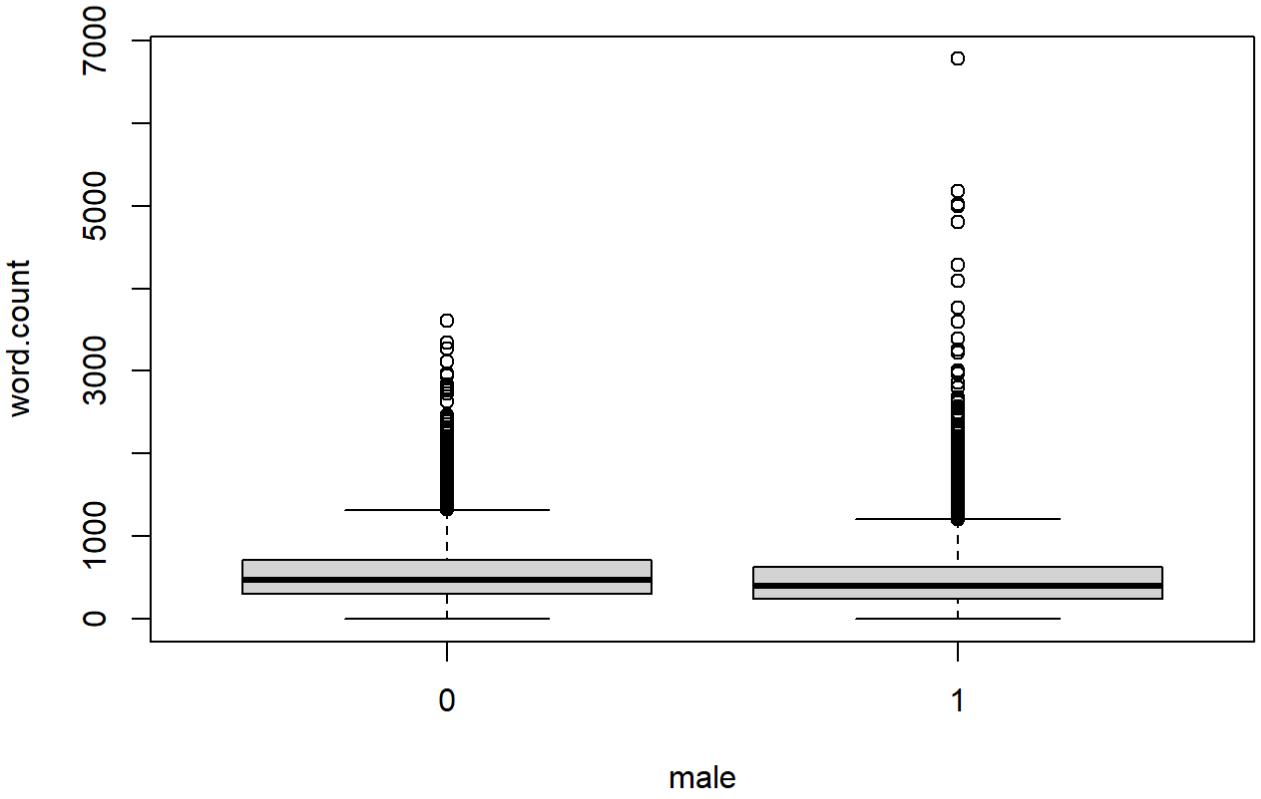
b. (10pts) Note that `word.count` is an explanatory variable, and `male` will be considered our "treatment" variable. In this part, we will assess the internal validity of this dataset in terms of `word.count` variable. For this part, complete the following **two** tasks:

- (5pts) First, create some form of graphical EDA to assess internal validity in terms of `word.count` . After creating your EDA, in 1-3 sentences, explain how your EDA relates to internal validity, and

discuss whether you think there is high internal validity based on your EDA.

Hide

```
boxplot(word.count ~ male, data = reddit)
```



[The boxplot shows word.count across gender groups. it shows that posts related to male group have a lot of outliers (mor than female group), and also median of female group's word.count is slightly higher than the male gorup.it is not clear how much the outliers would affect the results but medians are not that different. overall i think word.count is not exactly same across gender groups. and NO. internal validity does not hold much considering the outliers and medians, but to be sure, t.test is require]

- (5pts) Now, run a statistical test that allows you to assess internal validity in terms of `word.count`. By "statistical test," I mean an analysis that provides a p-value, and thus formally tests a null hypothesis. After running your test, in 1-3 sentences, explain how your test relates to internal validity, and discuss whether you think there is high internal validity based on your test.

Hide

```
t.test(word.count ~ male, data = reddit)
```

```
##
##  Welch Two Sample t-test
##
## data:  word.count by male
## t = 7.728, df = 9997.3, p-value = 1.197e-14
## alternative hypothesis: true difference in means between group 0 and group 1 is not equ
al to 0
## 95 percent confidence interval:
##  46.80492 78.61842
## sample estimates:
## mean in group 0 mean in group 1
##        563.7883        501.0767
```

[I used t.test to asses whether word.count is same across male and female groups which is required for internal validity assessment. null hypothesis is that mean word.count for male group equals to female group or in other word internal validity holds. Since the P-value is less than 0.05 (1.197e-14) we conclude that we reject the null hypothesis, and the t.test shows that mean word.count of female is higher than the male group with 95% confidence. So internal validity does not hold.]

Question 2: Linear Regression and ANCOVA (27pts)

In this question, we will consider the following statistical models:

Model 1: `num_comments` linearly regressed on `word.count`.

Model 2: `num_comments` linearly regressed on `word.count` and `male` (but NOT their interaction).

Model 3: `num_comments` linearly regressed on `word.count`, `male`, and the interaction between `word.count` and `male`.

Model 1 is *simple linear regression* because it involves a single quantitative explanatory variable and a quantitative outcome variable. Model 2 is an *additive ANCOVA model* because it involves (1) a quantitative explanatory variable, (2) a categorical explanatory variable, and (3) a quantitative outcome variable. Model 3 is an *interactive ANCOVA model* because it involves an interaction between the two explanatory variables.

If you have trouble with this question, it may be helpful to refer back to Homework4.

a. (9pts) For this part, answer the following two questions.

- (5pts) First, write code running **Model 1**, **Model 2**, and **Model 3**; in your code, define your models as `linReg1`, `linReg2`, and `linReg3`, and print out the `summary()` output for each model.

Hide

```
linReg1 = lm(num_comments ~ word.count, data = reddit)
summary(linReg1)
```

```
##
## Call:
## lm(formula = num_comments ~ word.count, data = reddit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -66.73  -16.27  -11.89   -2.46 1050.08
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.990482   0.735817  21.732   <2e-16 ***
## word.count   0.009260   0.001099   8.429   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.78 on 9998 degrees of freedom
## Multiple R-squared:  0.007056,   Adjusted R-squared:  0.006957
## F-statistic: 71.05 on 1 and 9998 DF,  p-value: < 2.2e-16
```

Hide

```
linReg2 = lm(num_comments ~ word.count + male, data = reddit)
summary(linReg2)
```

```
##
## Call:
## lm(formula = num_comments ~ word.count + male, data = reddit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -59.87  -16.38  -11.15   -2.25 1047.12
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.600600   0.891981  21.974  < 2e-16 ***
## word.count   0.008658   0.001099   7.877 3.70e-15 ***
## male1       -6.384936   0.896402  -7.123 1.13e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.67 on 9997 degrees of freedom
## Multiple R-squared:  0.01207,    Adjusted R-squared:  0.01187
## F-statistic: 61.07 on 2 and 9997 DF,  p-value: < 2.2e-16
```

Hide

```
linReg3 = lm(num_comments ~ word.count + male + word.count * male, data = reddit)
summary(linReg3)
```

```
##
## Call:
## lm(formula = num_comments ~ word.count + male + word.count *
##     male, data = reddit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -52.57  -16.19  -11.21   -2.38 1046.40
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      18.799179   1.124531  16.717  < 2e-16 ***
## word.count        0.010080   0.001638   6.153 7.89e-10 ***
## male1            -5.000291   1.484392  -3.369 0.000758 ***
## word.count:male1 -0.002585   0.002209  -1.170 0.241918
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.67 on 9996 degrees of freedom
## Multiple R-squared:  0.0122, Adjusted R-squared:  0.01191
## F-statistic: 41.17 on 3 and 9996 DF,  p-value: < 2.2e-16
```

Then, using your R output above, write out the *prediction equation* for **Model 2**. Please write your answer in terms of a *single* prediction equation, rather than two prediction equations. Be sure to clearly define any notation you use to write your equation. (**Hint**: In order to do this, you will have to define some new notation involving indicator variables.)

[Mean num_comments = ß0 + ß1 * word.count + ßm * male1 = 19.6 + 0.0086 * word.count -6.38 * male1 so if the subject is female the Mean num_comments = 19.6 + 0.0086 * word.count, and if the subject is male the Mean num_comments = 19.6 + 0.0086 * word.count - 6.38 = 13.22 + 0.0086 * word.count.] [ß0 is the intercept which is the mean num_comments when wordcount is 0 or the post does not have any words. ß1 is the change in the mean num_comments when wordcount increases by 1 and the gender variable is held constant. And ßm is the addition the intercept when the subject is male.]

- (4pts) Now look at the estimated *Intercept* for Model 2. In 1-2 sentences, write your interpretation of this estimated intercept in terms of the estimated mean number of comments.

[ß0 is the intercept which is the mean num_comments when wordcount is 0 or the post does not have any words.And ßm is the addition the intercept when the subject is male.]

b. (9pts) First use the `anova()` function to determine whether you prefer `linReg1`, `linReg2`, or `linReg3`. In your answer, first display the `anova()` output, then state which model you prefer (Model 1, Model 2, or Model 3), and discuss in 1-2 sentences how you arrived at your answer.

Hide

```
anova(linReg1, linReg2, linReg3)
```

```
## Analysis of Variance Table
##
## Model 1: num_comments ~ word.count
## Model 2: num_comments ~ word.count + male
## Model 3: num_comments ~ word.count + male + word.count * male
##   Res.Df      RSS Df Sum of Sq       F    Pr(>F)
## 1   9998 20046188
## 2   9997 19944967  1    101221 50.7368 1.129e-12 ***
## 3   9996 19942234  1      2732  1.3695    0.2419
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[the p-value for the Model 3 row is 0.24; because this is greater than 0.05, this suggests that we prefer Model 2 over Model 3. Meanwhile, the p-value for the Model 2 row is 1.129e-12; because this is less than 0.05, this suggests that we prefer Model 2 over Model 1. Thus, we prefer Model 2 the most out of the three models.]

Now look at your `anova()` output. You should see two p-values: One in Row 2, one in Row 3. Using ßs in your answer, what is the null hypothesis for the p-value in Row 2? What is the null hypothesis for the p-value in Row 3? In your answer, please be sure to specify what your ßs refer to.

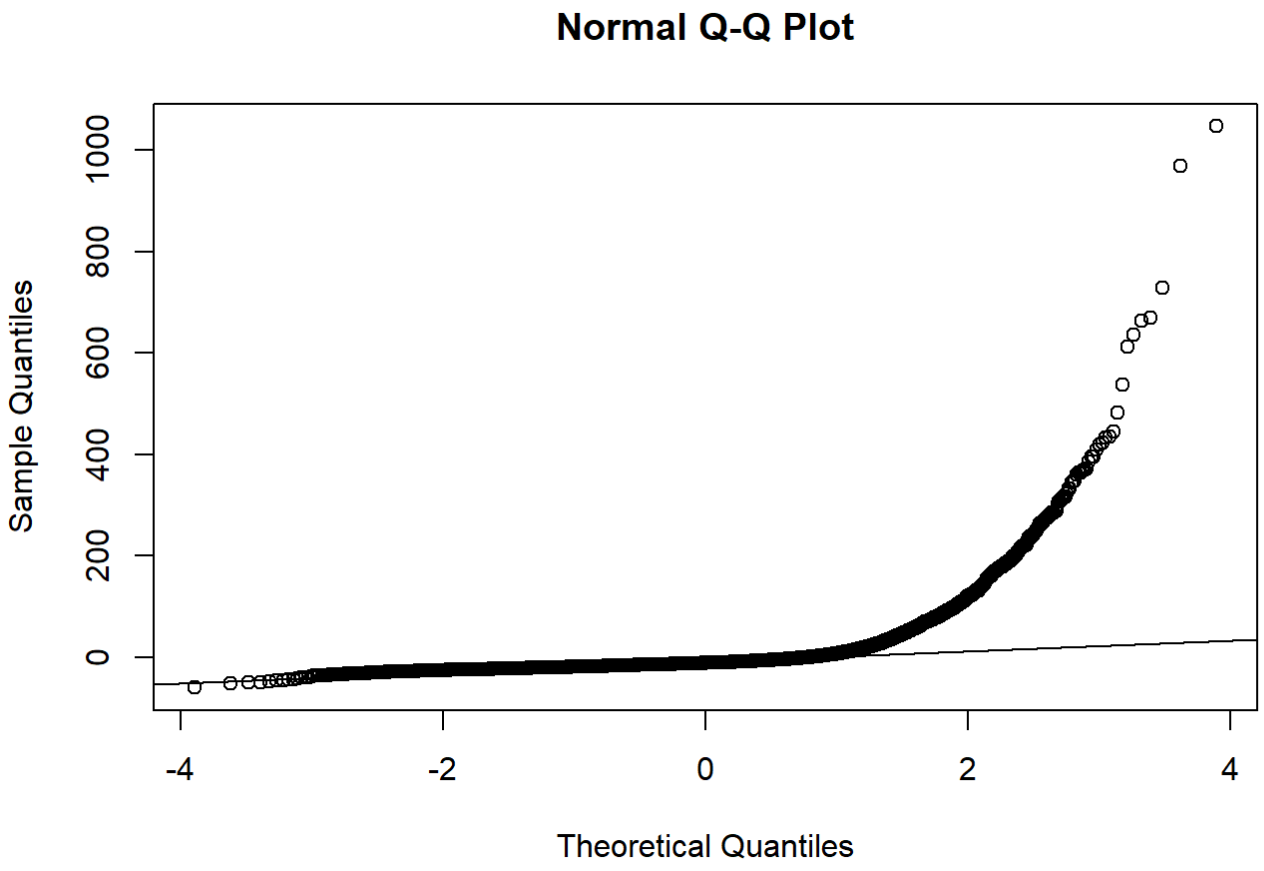
[In the third row, the p-value corresponds to the null hypothesis H0 : βw,m = 0, because βw,m is the only additional coefficient between Model 3 and Model 2. Meanwhile, in the second row, the p-value corresponds to the null hypothesis H0 : βm= 0, because βm is the only additional coefficient between Model 2 and Model 1. Thus, because we fail to reject the null hypothesis for third row, we do not find evidence that βw,m coefficient is non-zero; but for the second row, we have enough evidence that βm coefficient is non-zero]

c. (9pts) *For the model you preferred in Part B*, make a quantile-normal plot of the residuals, and interpret it (in 1-2 sentences) in terms of the assumptions made for this linear regression model. Then, make a residual-vs-fit plot, and interpret it (in 1-2 sentences) in terms of the assumptions made for this linear regression model. Be sure to discuss to what extent you believe the appropriate assumptions are violated, according to your residual plots.

Hide

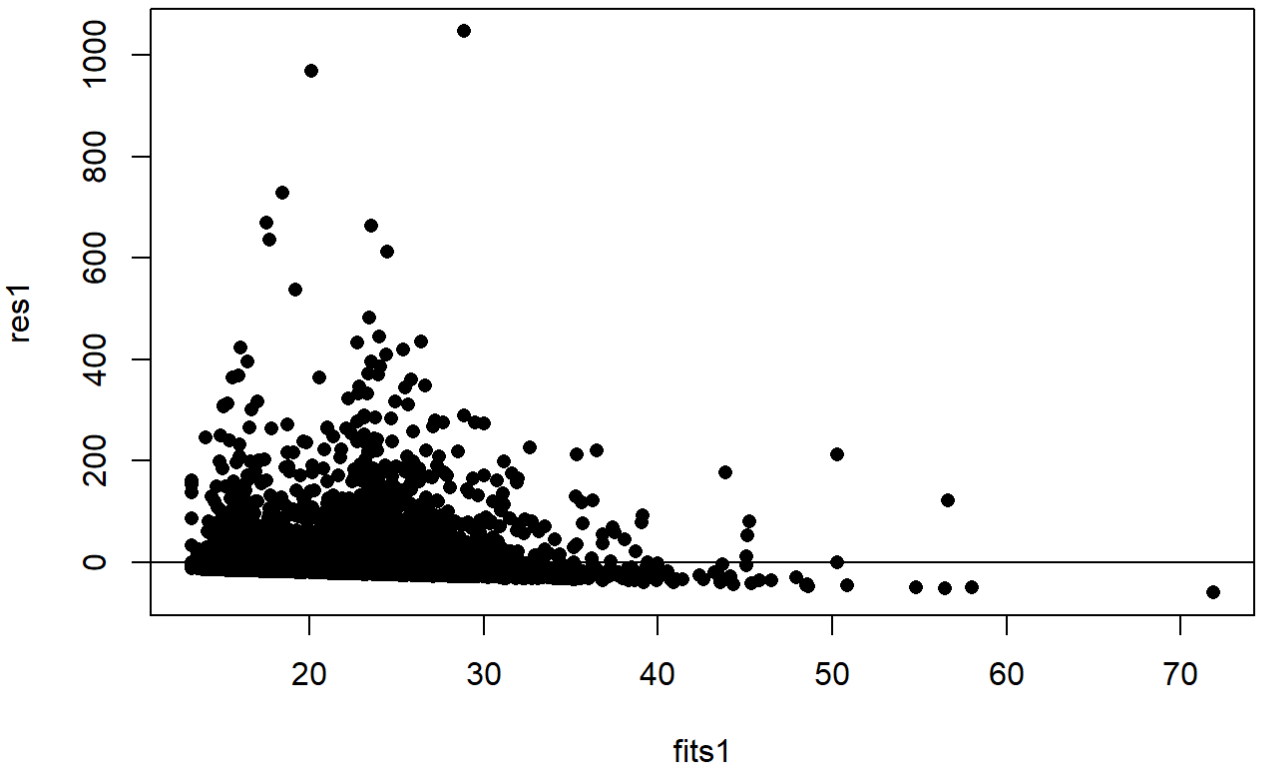
```
#residuals
res1 = residuals(linReg2)
#fits
fits1 = fitted(linReg2)


qqnorm(res1)
qqline(res1)
```



Hide

```
plot(fits1, res1,
     pch = 16)
abline(h=0)
```



[Assuming that independency and fixed X hold especially that if we consider subjects are independent of each other, we use quantile-normal plot of the residuals to check the normality assumption, and we use the residual-vs-fit plot to check for both Equal variance and Linearity. Q-Q plot shows that normality dies not hold because so many points are aligned on the Q-Q plot line. The residual-vs-fit plot shows that points are not equally distributed along the h=0 line, the are concentrated on left, so equal variance does not hold. More over linearity does not hold either because it seems like there is funneling how points are positioned.]

Question 3: Logistic Regression (40 points)

As an alternative to using a quantitative outcome variable, it is very common to dichotomize the outcome into a “success” or “failure” and use that as the outcome instead. After doing some basic EDA, we can find that the third quartile of `num_comments` is 18. In the dataset, I have defined a post as “successful” if it received 18 or more comments. This is captured in the `success` variable, which we will use as the outcome for this question. (As an aside: It is very common to use observed quartiles as a rule-of-thumb for dichotomizing a quantitative variable.)

a. (12pts) For this part, answer the following two questions.

- (6pts) In this homework, we will include `word.count` and `male` (but not their interaction) as explanatory variables. Write out the statistical model for logistic regression using `success` as the outcome variable and `word.count` and `male` (but not their interaction) as explanatory variables. Please write your model using ß symbols. (**Hint**: Your statistical model should specify and describe the *distribution* placed on the `success` variable for the logistic regression model described here.) **[Y = 1 indicates success and Y = 0 indicates failure. log (P (success = 1)/P (success = 0)) = ˆβ0 +**

ˆβ1 word.count + ˆβ2 male. P (success = 1)/P (success = 0) = exp (ˆβ0 + ˆβ1 word.count + ˆβ2 male). P (success = 1) = exp (ˆβ0 + ˆβ1 word.count + ˆβ2 male)/(1 + exp (ˆβ0 + ˆβ1 word.count + ˆβ2 male)). success iid∼ Bern(p), where p = exp (ˆβ0 + ˆβ1 word.count + ˆβ2 male)/(1 + exp (ˆβ0 + ˆβ1 word.count + ˆβ2 male))]

- (6pts) Your statistical model should include an intercept coefficient, as well as a coefficient for `male`. What is the interpretation of the intercept coefficient *in terms of the log-odds of success*? And what is the interpretation of the coefficient for `male` *in terms of the log-odds of success*?

[ˆβ0 is the intercept coefficient of this model representing the estimated log-odds of success for comments with 0 word count written by female. ˆβ2 is the coefficient for `male` which means that estimated log-odds of success for males is ˆβ2 units added to the estimated log-odds of success for females.]

b. (16pts) For this part, answer the following three questions.

- (4pts) First, run the regression model you wrote in Part A using the `glm()` function. Print out the `summary()` output from this function.

Hide

```
summary(glm(success ~ word.count + male, data = reddit, family = "binomial"))
```

```
##
## Call:
## glm(formula = success ~ word.count + male, family = "binomial",
##     data = reddit)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7570  -0.8051  -0.6831   1.3248   1.8572
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.096e+00  4.439e-02 -24.683  < 2e-16 ***
## word.count   4.180e-04  5.345e-05   7.819  5.3e-15 ***
## male1       -4.328e-01  4.655e-02  -9.298  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11334  on 9999  degrees of freedom
## Residual deviance: 11176  on 9997  degrees of freedom
## AIC: 11182
##
## Number of Fisher Scoring iterations: 4
```

- (4pts) Now look at your `summary()` output. What is the null hypothesis for the p-value in the `word.count` row? Please write your null hypothesis using mathematical notation; please use the same mathematical notation that you used in Part A.

[H0,w : β1 = 0 null hypothesis: wordcount coefficient is zero]

- (4pts) Now, using the `summary()` output, write out the *prediction equation* for this model *in terms of the probability of success*. (**Hint**: Your prediction equation should have P(`success` = 1) on the left-hand side, and the right-hand side should involve terms with exp(), i.e., the natural exponential function.)

[P (success = 1) = exp (ˆβ0 + ˆβ1 word.count + ˆβ2 male1)/(1 + exp (ˆβ0 + ˆβ1 word.count + ˆβ2 male1))= exp (-1.096e+00 + 4.180e-04 word.count - 4.328e-01male1)/(1 + exp (-1.096e+00 + 4.180e-04 word.count - 4.328e-01male1))]

- (4pts) Finally, using your prediction equation, what is the estimated probability of success for a male post that is 150 words long? In addition to your answer, please provide code that demonstrates how you arrived at your answer.

Hide

```
exp ( -1.096e+00 + 4.180e-04 * 150 - 4.328e-01)/(1 + exp (-1.096e+00 + 4.180e-04 * 150 - 4.328e-01))
```

```
## [1] 0.1875361
```

[P (success = 1) = exp (-1.096e+00 + 4.180e-04 word.count - 4.328e-01male1)/(1 + exp (-1.096e+00 + 4.180e-04 word.count - 4.328e-01male1))] [P (success = 1) = exp (-1.096e+00 + 4.180e-04 * 150 - 4.328e-01)/(1 + exp (-1.096e+00 + 4.180e-04 * 150 - 4.328e-01)) = 0.1875361 the probability is 0.1875]

c. (12pts) Using your `summary()` output from Part B, answer the following two questions.

- (6pts) Write out your interpretation of the “Estimate” in the `word.count` row in terms of the *odds of success*. Then, write out your interpretation of the “Estimate” in the `male1` row in terms of the *odds of success*.

Hide

```
exp(4.180e-04)
```

```
## [1] 1.000418
```

Hide

```
exp(-4.328e-01)
```

```
## [1] 0.6486902
```

[for every one-unit increase in word.count, (holding all other explanatory variables fixed), the odds will multiply by exp(4.180e-04)= 1.000418. thus, the odds of success are expected to increase for every one-unit increase in word.count).for male, the odds of success for males are estimated to be exp(-4.328e-01) = 0.6486902 times the odds of success for females (the odds of success for male increases compared to female).]

- (6pts) Based on your `summary()` output from the logistic regression model, what are your scientific conclusions for this dataset in terms of how gender and the word length of a post are related to the

probability of a post being “successful”? Please explain how you used the `summary()` output to arrive at your answer.

[They both are related to the probability of a post being “successful”, because co-efficent of both does not equal to 0 with the 95% confidence. The null hypothesis for word.count is that β1 equals to zero, since the p-value is less than 0.05 (5.3e-15) we reject the null. The null hypothesis for gender is that β2 equals to zero, since the p-value is less than 0.05 (< 2e-16) we reject the null.

Question 4: Chi-squared Tests (15 points)

In the previous part, we categorized `num_comments` into just two categories: a “success” or a “failure” (i.e., a high or low number of comments). Now we will consider categorizing `num_comments` into multiple categories. Note that the 25%, 50%, and 75% quantiles of `num_comments` are 4, 8, and 18, respectively:

Hide

```
quantile(reddit$num_comments, prob = c(0.25, 0.5, 0.75))
```

```
## 25% 50% 75%
##   4   8  18
```

Thus, we will create a four-category version of `num_comments`, which denotes whether someone’s post was in the first, second, third, or fourth quantiles in terms of `num_comments`:

Hide

```
#define the categorized version of num_comments
reddit$num_comments_cat = factor(ifelse(reddit$num_comments < 4, "first",
    ifelse(reddit$num_comments < 8, "second",
        ifelse(reddit$num_comments < 18, "third", "fourth"))))
#ensure that the factor is listed in an intuitive order
reddit$num_comments_cat = ordered(reddit$num_comments_cat,
                                  levels = c("first", "second", "third", "fourth"))
```

Note that, by default, `R` displays the levels of categorical variables in alphabetical order; thus, by default, `R` would display the levels of `num_comments_cat` as: “first”, “fourth”, “second”, “third”, which is not an intuitive order (hence the second line of code).

For a while I thought about making a separate question that asks you to code up this variable yourself, but then I thought, “This is the last homework, let’s calm down a bit.” Please just take a moment to see what this variable looks like (e.g., by running `reddit$num_comments_cat` in the Console), and then answer the following questions.

- (6pts) As EDA, write code to produce (and turn in) three tables:

1. A table that contains the counts of each combination of `male` and `num_comments_cat`.
2. A table that contains the percentages of each combination of `male` and `num_comments_cat`, where the percentages add up to 100% for each level of `male`.
3. A table that contains the percentages of each combination of `male` and `num_comments_cat`, where the percentages add up to 100% for each level of `num_comments_cat`.

For this part, you just have to write code that produces the three tables described above.

Hide

```
#first table
table(reddit$num_comments_cat, reddit$male)
```

```
##
##             0    1
##   first   863 1218
##   second 1149 1414
##   third  1389 1427
##   fourth 1446 1094
```

Hide

```
#second table
prop.table(table(reddit$num_comments_cat, reddit$male), margin = 2)
```

```
##
##                  0         1
##   first  0.1780483 0.2363672
##   second 0.2370538 0.2744033
##   third  0.2865690 0.2769261
##   fourth 0.2983289 0.2123035
```

Hide

```
#third table
prop.table(table(reddit$num_comments_cat, reddit$male), margin = 1)
```

```
##
##                  0         1
##   first  0.4147045 0.5852955
##   second 0.4483028 0.5516972
##   third  0.4932528 0.5067472
##   fourth 0.5692913 0.4307087
```

- (4pts) After you've made your tables, answer the following: Between Table 2 and Table 3, which one would be more useful for *assessing gender equality among different number-of-comments categories*? In other words, which one would be more useful for assessing if there are a roughly equal number of male and female posts in each level of `num_comments_cat` ? Explain your answer in 1-2 sentences.

[Third table because the sum of male and female in each num-comment category equals to 100. So it is easier to compare the male and female percentages within each level of the num-comment category.]

- (5pts) As formal analysis, conduct the appropriate chi-squared test for this dataset using `male` and `num_comments_cat` . After you do this, state your scientific conclusion from the test. In your answer, be sure to state the null hypothesis that is being tested, and whether you reject or fail to reject this hypothesis.

Hide

```
chisq.test(table(reddit$num_comments_cat, reddit$male))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(reddit$num_comments_cat, reddit$male)
## X-squared = 128.01, df = 3, p-value < 2.2e-16
```

[Null hypothesis is that num_comments_catcategories are independent from the gender. Since the p-value is less than 0.05(< 2.2e-16), we can reject the null and conclude that num_comments_cat depends of gender.]