Homework 8

Anahi Rodriguez

Due 11/2/2022

Homework Instructions

Make sure to add your name to the header of the document. When submitting the assignment on Gradescope, be sure to assign the appropriate pages of your submission to each Exercise.

The point value for each exercise is noted in the exercise title.

For questions that require code, please create or use the code chunk directly below the question and type your code there. Your knitted pdf will then show both the code and the output, so that we can assess your understanding and award any partial credit.

For written questions, please provide your answer after the indicated *Answer* prompt.

You are encouraged to knit your file as you work, to check that your coding and formatting are done so appropriately. This will also help you identify and locate any errors more easily.

Homework Setup

We'll use the following packages for this homework assignment. We'll also read in data from a csv file. To access the data, you'll want to download the dataset from Canvas, and place it in the same folder as this R Markdown document. You'll then be able to use the following code to load in the data.

```
library(ggplot2)
library(leaps)
library(faraway)
```

Exercise 1: Chick-fil-A Order Type [25 points]

For this exercise, we'll consider an extended dataset with nutritional information about menu items from Chick-fil-A. Be sure to use the updated cfa version of the dataset for Homework 8 as posted to Canvas, which is different from the Homework 7 version.

part a

Read in the cfa dataset from Canvas. When you read in the cfa file, include the argument `stringsAsFactors = T`.

```
# Use this code chunk for your answer.
setwd("~/Desktop/data")
cfa = read.csv("cfa.csv", stringsAsFactors = T)
```

part b

What proportion of menu items at Chick-fil-A include chicken?

```
# Use this code chunk for your answer.
mean(cfa$has_chicken)
```

```
## [1] 0.3017241
```

Answer: 0.3017241

part c

Fit a model predicting the calories of a menu item from the has_chicken variable. What is the estimate of the difference in mean calories between all menu items that do have chicken and all menu items that do not have chicken?

```
# Use this code chunk for your answer.
lm1 = lm(Calories ~ has_chicken, data = cfa)
summary(lm1)
```

```
##
## Call:
## lm(formula = Calories ~ has_chicken, data = cfa)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -327.0 -199.8 -103.4  103.0 4660.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   259.81      53.32   4.873 3.58e-06 ***
## has_chicken    97.19      97.07   1.001    0.319
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 479.9 on 114 degrees of freedom
## Multiple R-squared:  0.008716,   Adjusted R-squared:  2.043e-05
## F-statistic: 1.002 on 1 and 114 DF,  p-value: 0.3189
```

Answer: the difference in mean calories between all menu items that do have chicken and all menu items that do not have chicken is estimated as 97.19 calories on average

part d

Is there a statistically significant difference in mean calories between all menu items that do have chicken and all menu items that do not have chicken? Explain.

Answer: There is not a statitically significant difference in mean calories between all menu items that do have chicken and all menu items that do not have chicken as shown by the large p-value of 0.319. This p-value is too high to reject the null of no significant difference even at the 10% level.

part e

Now, let's look at the category variable. Create a table that contains a count of how many menu items fall into each possible category. *Hint: this can be done with one line of code.*

```
# Use this code chunk for your answer.
table(cfa$category)
```

```
##
##   breakfast      drinks   entr\x8ee      entree        kids       salad
##          14          19           2          10           3           3
##      sauces        side single_item       trays      treats
##          15           8          19          13          10
```

part f

What type of variable does R consider or classify the category variable as? If the category variable is included as a first-order term in a linear model, what will its contribution to the p for the model be?

Answer: R considers the category variable as a factor variable. If the category variable is included as a first-order term in a linear model, it will contribute 10 to p for the model as it is a categorical variable with 11 factors and one will not be included to work as the baseline category.

part g

Fit a model predicting the calories of a menu item from the category of that menu item and the serving size. Print a summary of this model.

```
# Use this code chunk for your answer.
lm2 = lm(Calories ~ category + Serving.size,
         data = cfa)
summary(lm2)
```

```
##
## Call:
## lm(formula = Calories ~ category + Serving.size, data = cfa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1941.95   -73.07     1.27    94.90  2518.77
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          237.85533   94.65898   2.513   0.0135 *
## categorydrinks      -561.15767  127.53438  -4.400 2.63e-05 ***
## categoryentr\x8ee      5.73825  264.19282   0.022   0.9827
## categoryentree        24.44243  144.68634   0.169   0.8662
## categorykids        -152.49420  222.41744  -0.686   0.4945
## categorysalad        134.24811  223.44177   0.601   0.5493
## categorysauces      -143.08792  130.34197  -1.098   0.2748
## categoryside         -87.16378  154.83768  -0.563   0.5747
## categorysingle_item -178.69167  123.63256  -1.445   0.1514
## categorytrays         98.68646  137.84051   0.716   0.4756
## categorytreats       -70.28658  145.18787  -0.484   0.6293
## Serving.size           0.83220    0.09966   8.351 3.13e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 349.4 on 104 degrees of freedom
## Multiple R-squared:  0.5207, Adjusted R-squared:   0.47
```

```
## F-statistic: 10.27 on 11 and 104 DF,  p-value: 1.793e-12
```

part h

What is the baseline level for this model?

Answer: The baseline level for this model is the breakfast category

part i

From the summary in part g, I notice that one of the estimates is provided as -70.3. What does this value mean?

Answer: We estimate the calories for a menu item of the type treats to have, on average, 70.3 less calories than menu items of the type breakfast, holding constant serving size

---

Exercise 2: High School Scores [30 points]

If you haven't already, you may need to download the faraway package using `install.packages(faraway)`.

For our second exercise of Homework 8, we'll use the `hsb` dataset included in the `faraway` package. You can read more about the `hsb` dataset by using `help(hsb)`

```
library(faraway)
data(hsb)
hsb = hsb
```

part a

There are 10 variables contained in the High School and Beyond dataset in addition to the `id` variable, which serves as a record of the observational unit – the student. For each of the 10 variables, record its **type**, including both the general and specific type.

Answer: Gender and schtyp are dichotomous categorical variables

race and prog are nominal categorical variables

ses is an ordinal categorical variable

read, write, math, science, and socst are discrete quantitative variables

part b

Fit a model that predicts the math score from the reading score, writing score, high school program, school type, and socioeconomic status. Print the summary, including the coefficients table, of the results. What is the value of p for this model?

```
# Use this code chunk for your answer.
lm3 = lm(math ~ read + write + prog + schtyp + ses,
         data = hsb)
summary(lm3)
```

```
##
## Call:
## lm(formula = math ~ read + write + prog + schtyp + ses, data = hsb)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -19.6770  -4.3258  -0.4242   4.4346  17.3644
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.72059    3.73370   5.282 3.45e-07 ***
## read          0.35790    0.05811   6.159 4.21e-09 ***
## write         0.29710    0.06179   4.808 3.07e-06 ***
## proggeneral  -2.74668    1.21004  -2.270  0.02432 *
## progvocation -3.94757    1.28262  -3.078  0.00239 **
## schtyppublic  0.64310    1.29208   0.498  0.61925
## seslow       -1.53970    1.34073  -1.148  0.25223
## sesmiddle    -0.04555    1.11421  -0.041  0.96743
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.427 on 192 degrees of freedom
## Multiple R-squared:  0.5459, Adjusted R-squared:  0.5294
## F-statistic: 32.98 on 7 and 192 DF,  p-value: < 2.2e-16
```

Answer: the value of p for this model is 8

part c

What is the baseline level for each of the categorical predictors in this model?

Answer: the basline level for program type is 'academic', for socioeconomic status is 'high', and for school type is 'private'

part d

Interpret the fitted intercept estimate.

Answer: We expect a student of high socioeconomic status who was enrolled in the academic program at a private school with a reading and writing score of 0, to score an average estimated math score of 19.72059

part e

From the output in part b, we'd like to determine if there's a significant difference in the mean math scores between being from a high socioeconomic class compared to being in a middle socioeconomic class, holding reading scores, writing scores, high school program, and school type constant. What about between students from a high socioeconomic class compared to a low socioeconomic class, holding reading scores, writing scores, high school program, and school type constant? Report your answer to these two tests, including numeric support in your written answer.

```
# Use this code chunk for your answer, if needed.
summary(lm3)
```

```
##
## Call:
## lm(formula = math ~ read + write + prog + schtyp + ses, data = hsb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.6770  -4.3258  -0.4242   4.4346  17.3644
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.72059    3.73370   5.282 3.45e-07 ***
## read          0.35790    0.05811   6.159 4.21e-09 ***
## write         0.29710    0.06179   4.808 3.07e-06 ***
## proggeneral  -2.74668    1.21004  -2.270  0.02432 *
## progvocation -3.94757    1.28262  -3.078  0.00239 **
## schtyppublic  0.64310    1.29208   0.498  0.61925
## seslow       -1.53970    1.34073  -1.148  0.25223
## sesmiddle    -0.04555    1.11421  -0.041  0.96743
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.427 on 192 degrees of freedom
## Multiple R-squared:  0.5459, Adjusted R-squared:  0.5294
## F-statistic: 32.98 on 7 and 192 DF,  p-value: < 2.2e-16
```

Answer: These coefficients for 'seslow' and 'sesmiddle' tell us what the estimated difference is between being low socioeconomic status compared to high socioeconomic status and similarly tell us what the estimated difference is between being middle socioeconomic status compared to high socioeconomic status. There is not a significant difference in mean math scores between being from a high socioeconomic class compared to either of low or middle socioeconomic classes, holding reading scores, writing scores, high school program, and school type constant as shown by the coefficients high p-values of 0.25223 and 0.96743 which would both be rejected even at the 25% significance level.

part f

We'd like to determine if there's a statistically significant difference of the mean math scores depending on the high school program, holding reading scores, writing scores, school type, and socioeconomic class constant. We'd like to be able to compare each set of two programs (academic vs. general, academic vs. vocation, & general vs. vocation).

Perform any necessary calculations to determine if there's a statistically significant difference between each of these sets of two programs. Report your answer for these three tests, including numeric support.

```
# Use this code chunk for your answer.
lm4 = lm(math ~ prog + read + write + schtyp + ses,
         data = hsb)
summary(lm4)
```

```
##
## Call:
## lm(formula = math ~ prog + read + write + schtyp + ses, data = hsb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.6770  -4.3258  -0.4242   4.4346  17.3644
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.72059    3.73370   5.282 3.45e-07 ***
## proggeneral  -2.74668    1.21004  -2.270  0.02432 *
## progvocation -3.94757    1.28262  -3.078  0.00239 **
## read          0.35790    0.05811   6.159 4.21e-09 ***
## write         0.29710    0.06179   4.808 3.07e-06 ***
## schtyppublic  0.64310    1.29208   0.498  0.61925
## seslow       -1.53970    1.34073  -1.148  0.25223
```

```
## sesmiddle    -0.04555    1.11421  -0.041  0.96743
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.427 on 192 degrees of freedom
## Multiple R-squared:  0.5459, Adjusted R-squared:  0.5294
## F-statistic: 32.98 on 7 and 192 DF,  p-value: < 2.2e-16
```

```
hsb$prog = relevel(hsb$prog, ref = 2)
lm7 = lm(math ~ prog + read + write + schtyp + ses,
         data = hsb)
summary(lm7)
```

```
##
## Call:
## lm(formula = math ~ prog + read + write + schtyp + ses, data = hsb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.6770  -4.3258  -0.4242   4.4346  17.3644
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.97392    3.61889   4.690 5.17e-06 ***
## progacademic  2.74668    1.21004   2.270   0.0243 *
## progvocation -1.20089    1.36388  -0.880   0.3797
## read          0.35790    0.05811   6.159 4.21e-09 ***
## write         0.29710    0.06179   4.808 3.07e-06 ***
## schtyppublic  0.64310    1.29208   0.498   0.6192
## seslow       -1.53970    1.34073  -1.148   0.2522
## sesmiddle    -0.04555    1.11421  -0.041   0.9674
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.427 on 192 degrees of freedom
## Multiple R-squared:  0.5459, Adjusted R-squared:  0.5294
## F-statistic: 32.98 on 7 and 192 DF,  p-value: < 2.2e-16
```

Answer: There is a statistically significant difference of the mean math scores between both general and vocation program types with academic program types. We see this from the p-values for the coefficients 'proggeneral' and 'progvocation'. These p-values, from model 4, tell us that at the 5% and 1% level respectively, we can reject the null that there is no difference between estimated mean math score between academic and general program types as well as between academic and vocation program types respectively. From the model 7 output, we can say that there does not exist a significant difference between estimated mean math scores between the vocational and general program type. We see this from the p-value for the coefficient 'progvocation' and its high p-value of 0.3797.

part g

Alicia isn't sure about including the school type variable and the high school program variable in the model to predict math scores. Alicia would like to perform a single statistical test to decide whether to include these two variables in the model from part b. Help Alicia perform this test. Generate the R output, report the p-value, the decision of the test, and the model that should be used going forward.

```
# Use this code chunk for your answer.
testing_model = lm(math ~ read + write + ses,
```

```
         data = hsb)

anova(lm4, testing_model)
```

```
## Analysis of Variance Table
##
## Model 1: math ~ prog + read + write + schtyp + ses
## Model 2: math ~ read + write + ses
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    192 7930.5
## 2    195 8374.7 -3   -444.16 3.5844 0.01483 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: the p-value for this f-test is 0.01483, indicating that we should reject the null that at least one of high school program type and school type has no effect on estimated mean math score for a student on average at the 5% level. Moving forward, we should use the full model that includes high school program type and school type as well as reading score, writing score, and socioeconomic status.

part h

Suppose that an additional type of school, a charter school, recently opened in the years since the hsb data were collected. Based on the model from part b, could we calculate a fitted value for a student who attended the charter school? Explain.

Answer: We could not calculate a fitted value for a student who attended the charter school because the only school type coefficient included is for public school. This value would be input as 0 for this student, and would then be incorrectly estimating a math score of a student who attended private school.

Exercise 3: US Wage Model Interpretations [15 points]

For this exercise, we'll analyze weekly wages of US male workers in 1988. This data is contained in the uswages dataframe from the faraway package. Before beginning our analyses, the starter code chunk creates a new version of the dataset that is more appropriate for regression purposes.

```
data(uswages)
usawages = uswages
usawages$geo = factor(names(uswages[,6:9])[max.col(uswages[,6:9])])
usawages = usawages[,-c(6:9)]
head(usawages)
```

```
##         wage educ exper race smsa pt geo
## 6085  771.60   18    18    0    1  0  ne
## 23701 617.28   15    20    0    1  0  we
## 16208 957.83   16     9    0    1  0  so
## 2720  617.28   12    24    0    1  0  ne
## 9723  902.18   14    12    0    1  0  mw
## 22239 299.15   12    33    0    1  0  we
```

For this exercise, we will work with the corrected `usawages` data (Note the additional "a" in "usa" at the beginning of the data frame).

part a

Fit a model to the usawages data, predicting wage from education, experience, living in a Standard Metropolitan Statistical Area (city + surrounding suburbs), and part time status.

```
# Use this code chunk for your answer.
lm5 = lm(wage ~ educ + exper + smsa + pt,
         data = usawages)
summary(lm5)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + smsa + pt, data = usawages)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -878.6 -213.8  -53.0  126.1 7524.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -264.788     50.686  -5.224 1.93e-07 ***
## educ          49.786      3.243  15.354  < 2e-16 ***
## exper          9.075      0.728  12.465  < 2e-16 ***
## smsa         111.825     21.617   5.173 2.54e-07 ***
## pt          -340.017     32.027 -10.617  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 413.6 on 1995 degrees of freedom
## Multiple R-squared:  0.1924, Adjusted R-squared:  0.1908
## F-statistic: 118.9 on 4 and 1995 DF,  p-value: < 2.2e-16
```

part b

There's a specific vocabulary term that applies to the variable for part time status. What is that vocabulary term?

Answer: factor/dummy/dichotomous variable

part c

Interpret the coefficients for education and living in a Standard Metropolitan Statistical Area in part a.

Answer: education interpretation: for each additional year of schooling an individual receives, I expect their estimated average weekly wage to increase by 49.786 dollars holding experience, for an individual working full time and who is not living in a standard metropolitan statistical area

Standard Metropolitan Statistical Area interpretation: I estimate that individuals living in a Standard Metropolitan Statistical Area will earn an average weekly wage 111.825 dollars greater than for individuals not living in a standard metropolitan statistical area - holding education and experience constant for an individual who works part time

part d

The model from part a could be written out equivalently as 4 distinct models after partitioning the data based on values recorded in 2 variables. Write out each of these 4 models, and define to what part of the data these models apply.

Answer: For those living in a standard metropolitan statistical area who are part time: estimated weekly wage in dollars = -492.98 + 49.786 * educ + 9.075 * exper

For those living in a standard metropolitan statistical area who are not part time: estimated weekly wage in dollars = -152.963 + 49.786 * educ + 9.075 * exper

For those not living in a standard metropolitan statistical area who are part time: estimated weekly wage in dollars = -604.805 + 49.786 * educ + 9.075 * exper

For those not living in a standard metropolitan statistical area who are not part time: estimated weekly wage in dollars = -264.788 + 49.786 * educ + 9.075 * exper

---

Exercise 4: Summarizing Interaction in US Wages [25 points]

For this problem, we'll continue working with the usawages dataset, but this time we'll focus on a model that includes an interaction term.

part a

Fit a model predicting wage from the geographic area that a male worker lives (`geo`), the experience level of that worker, and the interaction of the two variables. Print the summary of that model.

```
# Use this code chunk for your answer.
lm6 = lm(wage ~ geo * exper, data = usawages)
summary(lm6)
```

```
##
## Call:
## lm(formula = wage ~ geo * exper, data = usawages)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -770.6 -274.3  -82.1  165.7 6887.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 448.8918    34.2842  13.093  < 2e-16 ***
## geone        93.5695    49.3678   1.895   0.0582 .
## geoso        14.9600    46.0752   0.325   0.7455
## geowe        72.5725    51.5719   1.407   0.1595
## exper         7.6816     1.5569   4.934 8.73e-07 ***
## geone:exper  -3.0508     2.1506  -1.419   0.1562
## geoso:exper  -1.4928     2.0403  -0.732   0.4645
## geowe:exper  -0.3436     2.3835  -0.144   0.8854
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 451.5 on 1992 degrees of freedom
## Multiple R-squared:  0.03909,    Adjusted R-squared:  0.03572
## F-statistic: 11.58 on 7 and 1992 DF,  p-value: 1.714e-14
```

part b

Using the geographic area variable to separate the data into four different partitions, write out the model for each partition.

Answer: estimated wage for those living in geographic area ne = 542.4613 + 4.6308 * exper

estimated wage for those living in geographic area so = 463.8518 + 6.1888 * exper

estimated wage for those living in geographic area we = 521.4643 + 7.338 * exper

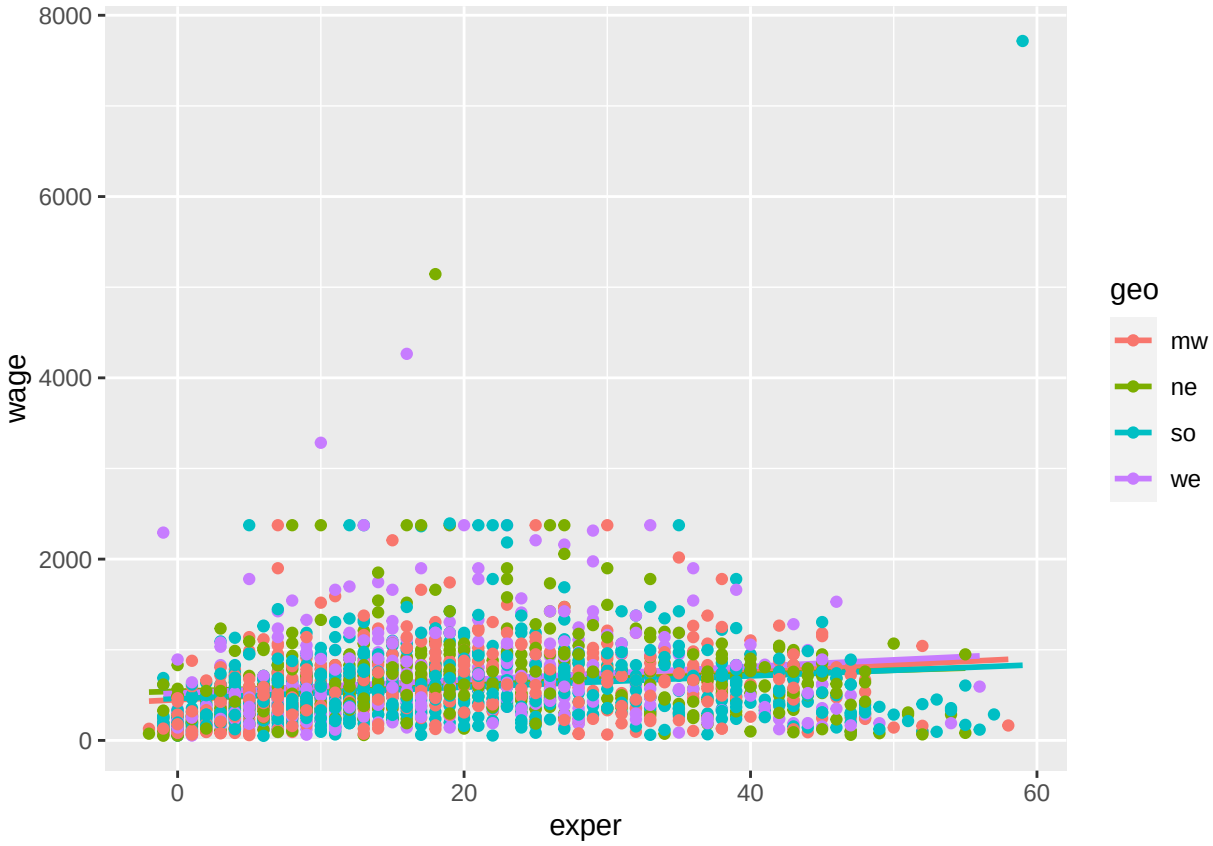
estimated wage for those living in geographic area mw = 448.8918 + 7.6816 * exper

part c

Visualize the relationship between the experience level of the worker, the geographic area, and the wage. Make sure to include appropriate summary lines in your plot representing the model fitted in part a.

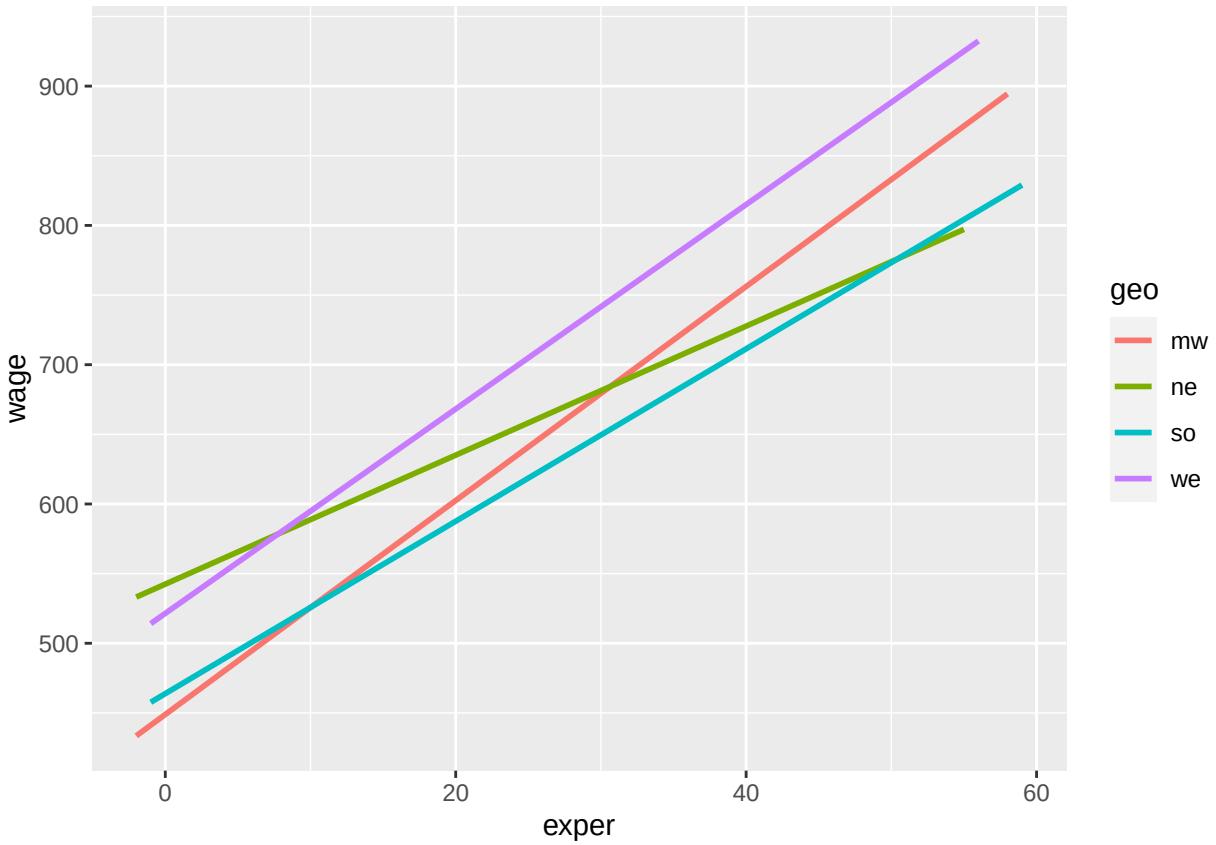
```
# Use this code chunk for your answer.
ggplot(data = usawages, aes(x = exper, y = wage, color = geo)) +
  geom_smooth(method = 'lm', se = F) +
  geom_point()
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(data = usawages, aes(x = exper, y = wage, color = geo)) +
  geom_smooth(method = 'lm', se = F) # plot w out outliers and points to better see lines
```

```
## `geom_smooth()` using formula 'y ~ x'
```



part d

Perform a single statistical test to test if at least one of the geographic regions has a different slope from the other regions. Report the p-value and a conclusion to the problem, indicating if we have evidence that at least one of the regions has a different slopes. *Hint: we are testing for the different geographic regions simultaneously with one test.*

```
# Use this code chunk for your answer.
testing_model2 = lm(wage ~ geo,
                    data = usawages)

anova(testing_model2)
```

```
## Analysis of Variance Table
##
## Response: wage
##             Df    Sum Sq Mean Sq F value  Pr(>F)
## geo          3   1711772  570591  2.7054 0.04395 *
## Residuals 1996 420968877  210906
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: the p-value for this test is 0.04395, indicating that we have sufficient evidence to suggest at the 5% level that at least one of the regions has a different slope

part e

Now, perform a single statistical test to test if at least one of the geographic regions has a different intercept from the other regions, assuming a single, constant slope for experience across all of the geographic regions. Report the p-value and a conclusion to the problem, indicating if we have evidence that at least one of the regions has a different intercept. *Hint: we are testing for the different geographic regions simultaneously with one test.*

```
# Use this code chunk for your answer.
null_model3 = lm(wage ~ geo + exper,
                 data = usawages)

testing_model3 = lm(wage ~ exper,
                    data = usawages)

anova(null_model3, testing_model3)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ geo + exper
## Model 2: wage ~ exper
##   Res.Df       RSS Df Sum of Sq     F  Pr(>F)
## 1   1995 406643983
## 2   1998 408494360 -3  -1850377 3.026 0.02851 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: the p-value for this test is 0.02851, indicating that there is sufficient evidence to suggest, at the 5% level, that at least one of the regions has a different intercept, holding experience fixed

---

Exercise 5: Formatting [5 points]

The last five points of the assignment will be earned for properly formatting your final document. Check that you have:

- included your name on the document
- properly assigned pages to exercises on Gradescope
- selected **page 1 (with your name)** and this page for this exercise (Exercise 5)
- all code is printed and readable for each question
- all output is printed
- generated a pdf file