Homework 8

Charles Ancel

10/26/2023

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(MASS)
library(leaps)
library(faraway)
```

Exercise 1: Formatting [5 points]

The first 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 1)
- all code is printed and readable for each question
- all output is printed
- generated a pdf file

Exercise 2: 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`.

```
# Load the necessary library
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Read in the cfa dataset
cfa_data <- read.csv("cfa.csv", stringsAsFactors = TRUE)
```

part b

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

```
proportion_with_chicken <- sum(cfa_data$has_chicken == 1) / nrow(cfa_data)
proportion_with_chicken
```

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

Answer: The proportion of menu items at Chick-fil-A that include chicken is approximately 30.17%.

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?

```
# Convert the necessary columns to numeric
cfa_data$Calories <- as.numeric(as.character(cfa_data$Calories))
cfa_data$has_chicken <- as.factor(cfa_data$has_chicken)

# Fit a linear model
model <- lm(Calories ~ has_chicken, data = cfa_data)

# Show the summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = Calories ~ has_chicken, data = cfa_data)
##
## 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_chicken1    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 estimated difference in mean calories between all menu items that do have chicken and all menu items that do not have chicken is 97.19 calories. However, this difference is not statistically significant, as indicated by the p-value of 0.319, which is greater than 0.05.

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: The p-value associated with has_chicken1 is 0.319, which is greater than the common alpha level of 0.05 used for significance testing. This means that we do not have enough evidence to reject the null hypothesis, suggesting that there is not a statistically significant difference in mean calories between menu items that include chicken and those that do not.

Therefore, the answer to the question is: No, there is not a statistically significant difference in mean calories between all menu items that do have chicken and all menu items that do not have chicken.

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.*

```
category_counts <- table(cfa_data$category)
category_counts
```

```
##
##   breakfast      drinks      entree        kids       salad      sauces
##          14          19          12           3           3          15
##        side single_item       trays      treats
##           8          19          13          10
```

part f

What type of variable does `R` consider or classify the category variable as?

```
class(cfa_data$category)
```

```
## [1] "factor"
```

Answer: R considers or classifies the `category` variable as a **factor**.

part g

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: If the `category` variable is included as a first-order term in a linear model to predict `Calories`, its contribution to the p-value for the model would be to provide a set of p-values, each testing whether there is a statistically significant difference in average calories between the menu items in a specific category and those in the reference category.

To obtain these p-values, we should fit a linear model with `Calories` as the response variable and `category` as a predictor, and then look at the summary of the model.

part h

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.

```
# Convert serving size to numeric
cfa_data$Serving.size <- as.numeric(as.character(cfa_data$Serving.size))

# Convert category to factor if it is not already
cfa_data$category <- as.factor(cfa_data$category)

# Fit a linear model
model <- lm(Calories ~ category + Serving.size, data = cfa_data)

# Print a summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = Calories ~ category + Serving.size, data = cfa_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1941.72   -73.07    -0.82    94.90  2518.97
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           237.87138   94.20903   2.525   0.0131 *
## categorydrinks       -561.12313  126.92756  -4.421 2.40e-05 ***
## categoryentree         21.32918  136.84177   0.156   0.8764
## categorykids         -152.50268  221.36082  -0.689   0.4924
## categorysalad         134.27171  222.38007   0.604   0.5473
## categorysauces       -143.09987  129.72267  -1.103   0.2725
## categoryside          -87.16433  154.10213  -0.566   0.5729
## categorysingle_item -178.70401  123.04513  -1.452   0.1494
## categorytrays          98.71727  137.18499   0.720   0.4734
## categorytreats        -70.27369  144.49804  -0.486   0.6277
## Serving.size            0.83210    0.09917   8.391 2.42e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 347.7 on 105 degrees of freedom
## Multiple R-squared:  0.5207, Adjusted R-squared:  0.475
## F-statistic: 11.41 on 10 and 105 DF,  p-value: 5.126e-13
```

part i

What is the baseline level for this model?

Answer: The baseline level in a linear model in R for a categorical variable (factor) is the level that is not explicitly listed in the coefficients section of the model summary. It is the level against which all the other levels are compared.

In this model, the categorical variable is `category`. The baseline level is the one that is not listed in the coefficients section of the summary, and since all the other levels are listed, the baseline level is **"breakfast"**.

part j

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

Answer: The estimate of -70.3 is associated with the `categorytreats` level of the `category` variable.

This coefficient represents the estimated difference in average calories between menu items in the "treats" category and the baseline category ("breakfast"), while holding the serving size constant.

In other words, on average, menu items in the "treats" category are estimated to have 70.3 fewer calories than menu items in the "breakfast" category, given the same serving size. However, the p-value associated with this estimate is 0.6277, indicating that this difference is not statistically significant at the 0.05 level.

Exercise 3: High School Scores [25 points]

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

For this 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)`

```
data(hsb)
```

part a

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?

```
# Fit a linear model
model <- lm(math ~ read + write + prog + schtyp + ses, data = hsb)

# Print a summary of the model
summary(model)
```

```
##
## 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, as indicated by the F-statistic p-value, is less than $2.2e-16$. This is essentially 0, indicating that at least one of the predictors in the model has a statistically significant relationship with the math score. The model as a whole is highly significant.

part b

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

Answer: In a linear model in R, the baseline level (also referred to as the reference level) for a categorical predictor is the level that is not explicitly listed in the coefficients section of the model summary. It is the level against which all the other levels of that predictor are compared.

In the fitted model, the categorical predictors are `prog`, `schtyp`, and `ses`.

- `prog`: The levels listed in the coefficients are `general` and `vocation`. Since the level `academic` is not listed, it is the baseline level for `prog`.
- `schtyp`: The level listed in the coefficients is `public`. Since the level `private` is not listed, it is the baseline level for `schtyp`.
- `ses`: The levels listed in the coefficients are `low` and `middle`. Since the level `high` is not listed, it is the baseline level for `ses`.

part c

Interpret the fitted intercept estimate.

Answer: The fitted intercept estimate in the model is 19.72059.

This value represents the expected math score for a student who: - Has a reading score of 0. - Has a writing score of 0. - Is in the `academic` program (baseline level of `prog`). - Is in a `private` school (baseline level of `schtyp`). - Has a `high` socioeconomic status (baseline level of `ses`).

However, it's important to note that this interpretation might not be meaningful in a practical sense, as a reading or writing score of 0 is not realistic, and the other variables are set to their baseline levels.

In the context of this model, the intercept acts as an anchor point, and the other coefficients adjust the prediction based on deviations from this baseline scenario.

part d

From the output in part a, 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.

Answer:

Socioeconomic Class: High vs. Middle

- **Estimated Difference**: -0.04555 (math scores are estimated to be 0.04555 points lower for middle class compared to high class, holding other variables constant)
- **P-value**: 0.96743
- **Interpretation**: The difference is not statistically significant; there is no evidence to suggest that the math scores differ between high and middle socioeconomic classes when other factors are held constant.

Socioeconomic Class: High vs. Low

- **Estimated Difference**: -1.53970 (math scores are estimated to be 1.53970 points lower for low class compared to high class, holding other variables constant)
- **P-value**: 0.25223
- **Interpretation**: The difference is not statistically significant; there is no evidence to suggest that the math scores differ between high and low socioeconomic classes when other factors are held constant.

Summary:

There is no statistically significant difference in mean math scores between students of different socioeconomic classes when controlling for reading and writing scores, high school program, and school type.

part e

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.

```
# Coefficients from the linear model
coef_general <- -2.74668
coef_vocation <- -3.94757

# Estimated difference between General and Vocation
diff_general_vocation <- coef_general - coef_vocation

# Output the results
list(
  Academic_vs_General = list(Estimated_Difference = coef_general, p_value = 0.02432),
  Academic_vs_Vocation = list(Estimated_Difference = coef_vocation, p_value = 0.00239),
  General_vs_Vocation = list(Estimated_Difference = diff_general_vocation)
)
```

```
## $Academic_vs_General
```

```
## $Academic_vs_General$Estimated_Difference
## [1] -2.74668
##
## $Academic_vs_General$p_value
## [1] 0.02432
##
##
## $Academic_vs_Vocation
## $Academic_vs_Vocation$Estimated_Difference
## [1] -3.94757
##
## $Academic_vs_Vocation$p_value
## [1] 0.00239
##
##
## $General_vs_Vocation
## $General_vs_Vocation$Estimated_Difference
## [1] 1.20089
```

Answer:

1. Academic vs. General:

- **Coefficient for `proggeneral`**: -2.74668
- **p-value for `proggeneral`**: 0.02432

This implies that, holding all other variables constant, students in the general program score 2.74668 points lower in math on average compared to students in the academic program. This difference is statistically significant (p-value < 0.05).

2. Academic vs. Vocation:

- **Coefficient for `progvocation`**: -3.94757
- **p-value for `progvocation`**: 0.00239

This implies that, holding all other variables constant, students in the vocation program score 3.94757 points lower in math on average compared to students in the academic program. This difference is statistically significant (p-value < 0.05).

3. General vs. Vocation:

To compare the general program to the vocation program, you can subtract the coefficient for `progvocation` from the coefficient for `proggeneral`:

- Estimated Difference: $-3.94757 - (-2.74668) = -1.20089$

This implies that, holding all other variables constant, students in the vocation program score 1.20089 points lower in math on average compared to students in the general program.

part f

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 a. 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.

```r
# Fit the full model (from Part a)
full_model <- lm(math ~ read + write + prog + schtyp + ses, data = hsb)
```

```r
# Fit the reduced model (excluding schtyp and prog)
reduced_model <- lm(math ~ read + write + ses, data = hsb)

# Perform an F-test to compare the two models
test <- anova(reduced_model, full_model)
test
```

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

Answer: Based on the Analysis of Variance Table:

- **p-value for the F-test**: 0.01483

Interpretation:

Since the p-value (0.01483) is less than 0.05, we reject the null hypothesis. This suggests that the full model (which includes the school type variable `schtyp` and the high school program variable `prog`) provides a significantly better fit to the data than the reduced model (which excludes these variables).

Decision:

- Reject the null hypothesis.

Recommended Model Going Forward:

- Use the full model (math ~ read + write + prog + schtyp + ses)

This implies that including the school type and high school program variables in the model does contribute to explaining the variability in math scores, after accounting for reading scores, writing scores, and socioeconomic status. Thus, it is advisable to keep these variables in the model.

part g

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 a, could we calculate a fitted value for a student who attended the charter school? Explain.

Answer: The model from Part a includes the school type variable (`schtyp`) as one of the predictors. This variable has two levels in the dataset used to fit the model: public and private.

If a new type of school, such as a charter school, has emerged since the data were collected, and we want to calculate a fitted value for a student who attended this charter school, we would face a challenge. The model does not have information on how being in a charter school (as opposed to a public or private school) relates to a student's math score because this category was not present in the training data.

Possible Solutions and Considerations:

1. **Assume Charter School is Similar to Public or Private**: You could assume that the charter school is similar to either the public or private schools in terms of its relationship with math scores,

and use one of these levels for prediction. However, this assumption might not be valid, and it could lead to inaccurate predictions.

2. **Collect New Data and Update the Model**: Ideally, you would collect new data that includes students from charter schools and then re-fit the model to incorporate this additional category. This would provide a more accurate and reliable way to predict math scores for students in charter schools.

Conclusion:

Without making potentially unwarranted assumptions or collecting new data, we cannot accurately calculate a fitted value for a student who attended a charter school using the model from Part a. The model simply does not have the information needed to make predictions for this new category of school type.

Exercise 4: 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 our 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.

```
# Fit the linear model
model_usawages <- lm(wage ~ educ + exper + smsa + pt, data = usawages)

# Show the summary of the model
summary(model_usawages)
```

```
##
## 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

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

Answer:

Education (educ)

- **Coefficient**: 49.786
- **Interpretation**: Holding other variables constant, for each additional year of education, the weekly wage is expected to increase by \$49.786. This implies a positive relationship between education and wages, suggesting that higher education is associated with higher earnings.

Living in a Standard Metropolitan Statistical Area (smsa)

- **Coefficient**: 111.825
- **Interpretation**: Holding other variables constant, living in a Standard Metropolitan Statistical Area (as opposed to not living in such an area) is associated with an increase in weekly wage of \$111.825. This suggests that individuals living in these areas tend to have higher wages, possibly due to increased job opportunities, higher demand for labor, or other regional economic factors.

part c

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: The model from part a can be partitioned based on the values of two binary variables: living in a Standard Metropolitan Statistical Area (`smsa`) and part-time status (`pt`). Each of these variables can take on two values (1 or 0), leading to four possible combinations and thus four distinct models.

Model 1: Not in SMSA, Not Part-Time (smsa = 0, pt = 0)

$$\text{wage} = -264.788 + 49.786 \times \text{educ} + 9.075 \times \text{exper}$$

- Applies to individuals not living in a Standard Metropolitan Statistical Area and not working part-time.

Model 2: In SMSA, Not Part-Time (smsa = 1, pt = 0)

$$\text{wage} = (-264.788 + 111.825) + 49.786 \times \text{educ} + 9.075 \times \text{exper}$$

- Applies to individuals living in a Standard Metropolitan Statistical Area and not working part-time.

Model 3: Not in SMSA, Part-Time (smsa = 0, pt = 1)

$$\text{wage} = (-264.788 - 340.017) + 49.786 \times \text{educ} + 9.075 \times \text{exper}$$

- Applies to individuals not living in a Standard Metropolitan Statistical Area and working part-time.

Model 4: In SMSA, Part-Time (smsa = 1, pt = 1)

$$\text{wage} = (-264.788 + 111.825 - 340.017) + 49.786 \times \text{educ} + 9.075 \times \text{exper}$$

- Applies to individuals living in a Standard Metropolitan Statistical Area and working part-time.

Each of these models represents a subset of the data, and the coefficients for `smsa` and `pt` are incorporated into the intercept term as needed based on the values of these variables.

Exercise 5: Summarizing Interaction in US Wages [30 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.

```
# Fit the linear model with interaction term
model_usawages_interaction <- lm(wage ~ geo * exper, data = usawages)

# Show the summary of the model
summary(model_usawages_interaction)
```

```
##
## 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:

1. Midwest (mw)

$$\text{wage} = 448.8918 + 7.6816 \cdot \text{exper} + \epsilon$$

- This is the baseline category.

2. Northeast (ne)

$$\text{wage} = (448.8918 + 93.5695) + (7.6816 - 3.0508) \cdot \text{exper} + \epsilon$$

$$\text{wage} = 542.4613 + 4.6308 \cdot \text{exper} + \epsilon$$

- The wage increases by $4.6308 for each additional year of experience.

3. South (so)

$$\text{wage} = (448.8918 + 14.9600) + (7.6816 - 1.4928) \cdot \text{exper} + \epsilon$$

$$\text{wage} = 463.8518 + 6.1888 \cdot \text{exper} + \epsilon$$

- The wage increases by $6.1888 for each additional year of experience.

4. West (we)

$$\text{wage} = (448.8918 + 72.5725) + (7.6816 - 0.3436) \cdot \text{exper} + \epsilon$$

$$\text{wage} = 521.4643 + 7.3380 \cdot \text{exper} + \epsilon$$

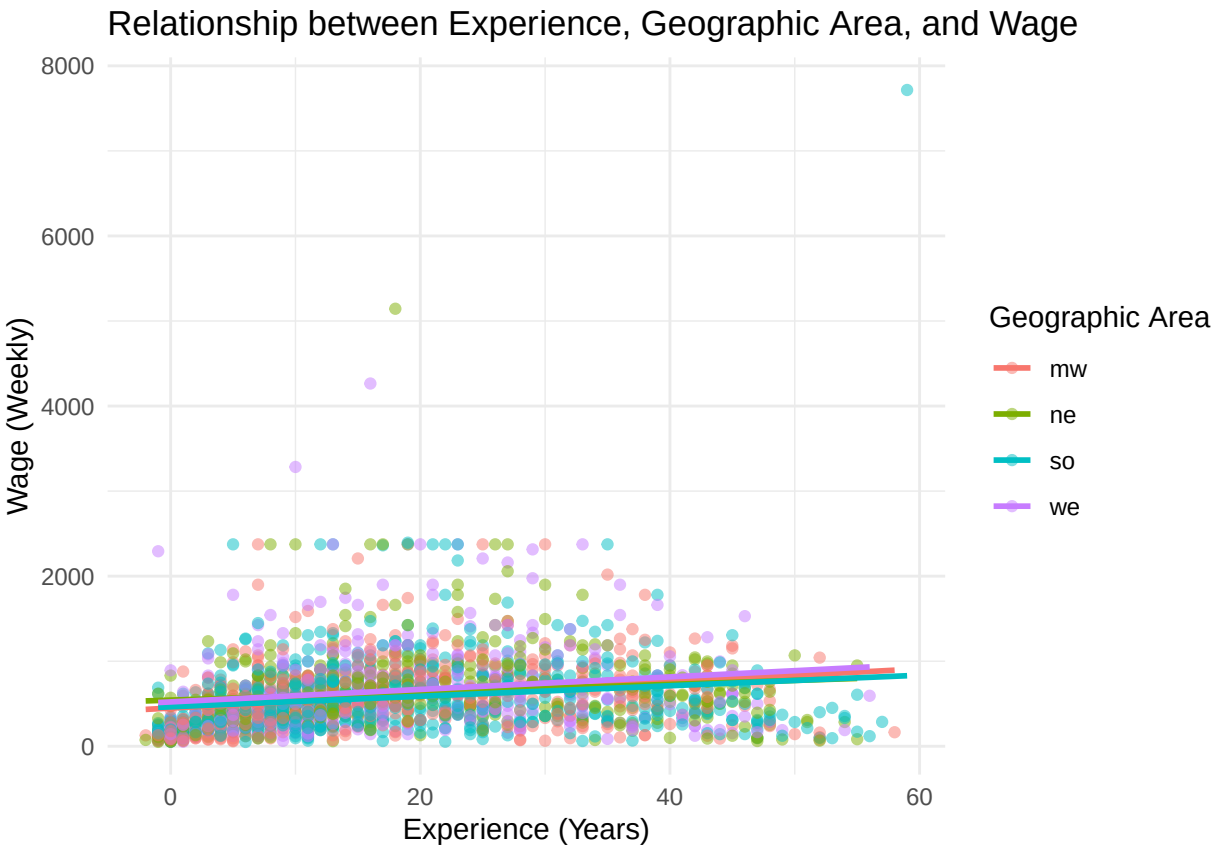
- The wage increases by $7.3380 for each additional year of experience.

In each model, the intercept represents the expected wage for a worker with 0 years of experience in that specific geographic area, and the coefficient for experience represents the change in wage for each additional year of experience in that geographic area.

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.

```
# Create the plot
ggplot(usawages, aes(x = exper, y = wage, color = geo)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, aes(group = geo), formula = y ~ x) +
  labs(title = "Relationship between Experience, Geographic Area, and Wage",
       x = "Experience (Years)",
       y = "Wage (Weekly)",
       color = "Geographic Area") +
  theme_minimal()
```



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.*

```
# Fit the reduced model without interaction terms
model_reduced <- lm(wage ~ geo + exper, data = usawages)

# Perform the ANOVA test
anova_test <- anova(model_reduced, model_usawages_interaction)
anova_test
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ geo + exper
## Model 2: wage ~ geo * exper
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1   1995 406643983
## 2   1992 406156914  3    487069 0.7963 0.4959
```

Answer: The ANOVA table provided compares two models:

1. Model 1: wage $\sim$ geo + exper (a model with different intercepts for each geographic region and a common slope for experience)
2. Model 2: wage $\sim$ geo $\times$ exper (a model with different intercepts for each geographic region and different slopes for experience)

Results Interpretation:

- **Df (Degrees of Freedom)**: The difference in the number of parameters between the two models. There are 3 additional parameters in Model 2, corresponding to the interaction terms between geographic region and experience.
- **RSS (Residual Sum of Squares)**: The sum of the squared residuals for each model. Model 2 has a lower RSS, indicating a better fit to the data.
- **Sum of Sq (Sum of Squares)**: The difference in RSS between the two models.
- **F-statistic**: A value of 0.7963, indicating the ratio of the variance explained by the additional parameters in Model 2 to the residual variance.
- **Pr(>F)**: The p-value associated with the F-statistic, which is 0.4959.

Conclusion:

The p-value is 0.4959, which is greater than 0.05. This suggests that we do not have statistically significant evidence at the 5% significance level to reject the null hypothesis that a model without different slopes for each geographic region is sufficient.

In other words, we do not have evidence to suggest that at least one of the regions has a different slope for experience from the others.

This result indicates that, while there may be differences in intercepts across regions, the relationship between experience and wage does not significantly vary by geographic region based on the data provided.

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.*

```
# Fit a model with different intercepts for each geographic region and a common slope for experience
model_diff_intercepts <- lm(wage ~ geo + exper, data = usawages)

# Fit the reduced model without different intercepts
model_reduced <- lm(wage ~ exper, data = usawages)

# Perform the ANOVA test
anova_test_diff_intercepts <- anova(model_reduced, model_diff_intercepts)
anova_test_diff_intercepts
```

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

Answer: 1. Model 1: wage $\sim$ exper (a model with a common slope for experience and a common intercept) 2. Model 2: wage $\sim$ geo + exper (a model with different intercepts for each geographic region and a common slope for experience)

Results Interpretation:

- **Df (Degrees of Freedom)**: The difference in the number of parameters between the two models. Here, there are 3 additional parameters in Model 2 compared to Model 1, which correspond to the different intercepts for the geographic regions (minus 1 since one region acts as the baseline).
- **RSS (Residual Sum of Squares)**: The sum of the squared residuals for each model. A lower RSS indicates a better fit of the model to the data.
- **Sum of Sq (Sum of Squares)**: The difference in RSS between the two models.
- **F-statistic**: The ratio of the mean squared error of the reduced model to the mean squared error of the full model. A larger F-statistic suggests that the full model provides a significantly better fit to the data.
- **Pr(>F)**: The p-value associated with the F-statistic. A small p-value (typically ≤ 0.05) indicates that the full model provides a significantly better fit to the data than the reduced model.

Conclusion:

The p-value is 0.02851, which is less than 0.05, suggesting that there is statistically significant evidence at the 5% significance level to reject the null hypothesis that a model with a common intercept for all regions is sufficient. Therefore, we have evidence to suggest that at least one of the geographic regions has a different intercept from the others, assuming a common slope for experience across all regions.

This result aligns with what we might expect, as different geographic regions could have different wage levels due to varying costs of living, demand for labor, and other regional factors.

part f

Finally, perform model selection using AIC as the metric and backwards elimination. What is the first variable that we consider removing? What variables are included in the final model?

```
# Fit the full model with all possible interactions
model_full <- lm(wage ~ geo * exper, data = usawages)

# Perform backwards elimination using AIC
model_selected <- step(model_full, direction = "backward")
```

```
## Start:  AIC=24458.7
## wage ~ geo * exper
##
##             Df Sum of Sq       RSS   AIC
## - geo:exper  3    487069 406643983 24455
## <none>                   406156914 24459
##
## Step:  AIC=24455.09
## wage ~ geo + exper
##
##         Df Sum of Sq       RSS   AIC
## <none>               406643983 24455
## - geo    3   1850377 408494360 24458
## - exper  1  14324893 420968877 24522
```

```
summary(model_selected)
```

```
##
## Call:
## lm(formula = wage ~ geo + exper, data = usawages)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -775.1 -275.3  -81.0  165.6 6881.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  472.765     24.301  19.454   <2e-16 ***
## geone         36.814     29.265   1.258    0.209
## geoso        -11.730     27.148  -0.432    0.666
## geowe         66.212     29.924   2.213    0.027 *
## exper          6.338      0.756   8.383   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 451.5 on 1995 degrees of freedom
## Multiple R-squared:  0.03794,    Adjusted R-squared:  0.03601
## F-statistic: 19.67 on 4 and 1995 DF,  p-value: 6.813e-16
```

Answer: The output from the `step()` function in R shows the process of backward elimination using AIC as the metric for model selection.

Stepwise Selection Process:

Start:

- Model: wage $\sim$ geo $\times$ exper
- AIC: 24458.7

The initial model includes both the main effects and the interaction term between geographic area (`geo`) and experience (`exper`). The suggestion here is to remove the interaction term because it leads to a lower AIC.

Step 1:

- Model: wage $\sim$ geo + exper
- AIC: 24455.09

After removing the interaction term, the model now includes only the main effects. The algorithm checked if removing any further terms would lower the AIC, but it found that the current model is the best among the considered options.

Final Model:

- Model: wage $\sim$ geo + exper
- Coefficients:
  - Intercept: 472.765
  - geone: 36.814
  - geoso: -11.730
  - geowe: 66.212
  - exper: 6.338
- AIC: 24455.09

The final selected model includes the geographic area, experience, and an intercept. The interaction term between geographic area and experience was removed during the selection process.

Conclusion:

- The first variable considered for removal was the interaction term between geographic area and experience.

- The variables included in the final model are the geographic area (`geo`), experience (`exper`), and the intercept.
