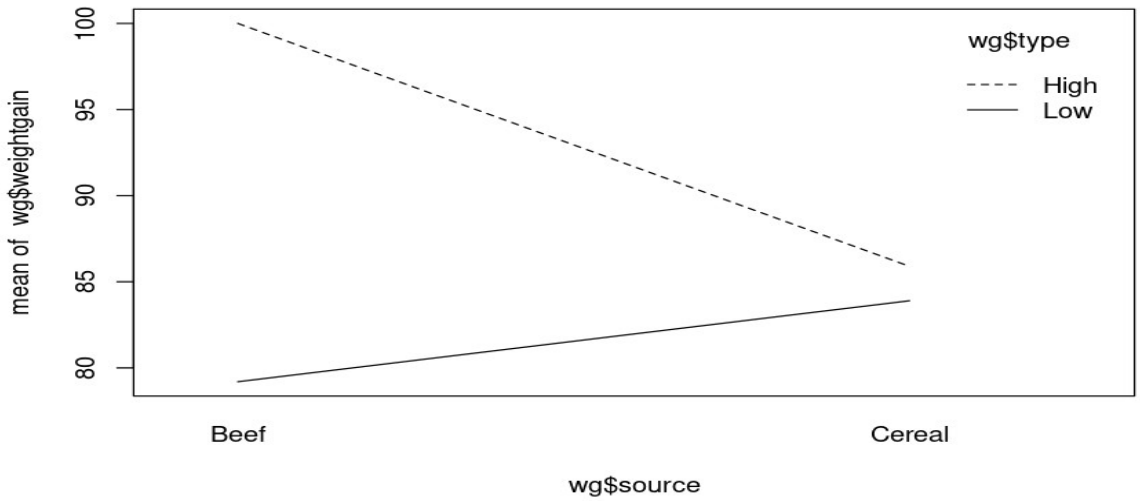
IST772 Summary Template: Chapter 9 – Interactions in ANOVA and Regression
Originality Assertion: By submitting this file you affirm that this writing is your own.

Name: Hendi Kushta
Date: 03/24/2023

****Important** Copying and/or pasting anything from the textbook will not be acceptable for your chapter notes submissions. You must write your notes in your own words and generate your own code, results, and graphs in R. This is what forces your brain to process the material that you read.**

INTRODUCTION

The general linear model includes analysis of variance and linear multiple regression. These models examine the simple, direct effects of independent variables on the dependent variable, also known as main effects. When there are multiple main effects, an interaction can occur where the dependent variable is affected differently by one independent variable depending on the status of another independent variable. An example of an interaction is the study of weight gain in rats fed different diets with varying levels of protein. The research question regarding an interaction requires simultaneous consideration of both independent variables. The same statistical test can shed light on different phrasing of the interaction research question.



The diagram in Figure represents a study of weight gain in rats fed either beef or cereal and a high or low protein diet. The Y-axis shows weight gain in grams, while the X-axis indicates the food source. The diagram includes two lines

representing high and low protein diets. The lines connect the means for the four different groups, but there are no intermediate points. The purpose of the lines is to visually connect the means belonging to the same condition of the second factor.
The non-parallel lines in the interaction diagram suggest the possibility of an interaction effect, which occurs when the effect of one independent variable on the dependent variable depends on the level of another independent variable. The ANOVA test is used to determine whether the observed differences are statistically significant and not due to chance. Factorial ANOVA is used when there are at least two independent variables, each with two or more levels. In contrast, oneway ANOVA is used when there is only one independent variable with more than two levels.

INTERACTIONS IN ANOVA

The chapter discusses the basic arrangement of a weight-gain data set, which has a dietary source factor with two levels (beef, cereal) fully crossed with a type factor that also has two levels (low protein and high protein). The phrase “fully crossed” means that for each level of the first factor, we have all levels of the second factor (and vice versa). The passage also provides two aov() calls that can be used to test the main effects of these two factors as well as the interaction between them.
There are two ways to specify the model in ANOVA test in R, one using explicit listing of effects and the other using a shortcut formula. Both ways provide the same output. The output displays the values of F, p-value, and degrees of freedom for each of the main effects and the interaction effect, as well as the overall model. The author emphasizes the importance of checking the interaction effect before making any conclusions about the main effects.

```
```{r}
aovOut = aov(weightgain ~ source + type + source:type, data=weightgain)
aovOut2 = aov(weightgain ~ source * type, data=weightgain)
summary(aovOut2)
```
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
source        1    221   220.9   0.988 0.3269
type          1   1300  1299.6   5.812 0.0211 *
source:type   1    884   883.6   3.952 0.0545 .
Residuals    36   8049   223.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

IST772 Summary Template: Chapter 9 – Interactions in ANOVA and Regression
Originality Assertion: By submitting this file you affirm that this writing is your own.

The interaction between the dietary source and type factors was not statistically significant at the conventional alpha threshold of $p < .05$, despite the promising plot. It's important to remember that any p-value greater than or equal to 0.05 is not statistically significant. Therefore, the researchers can move on to interpreting the main effects. There was no significant effect for the main effect of source, and the notation "N.S." means not significant. However, there was a significant main effect for the main effect of type, with $F(1,36) = 5.81$, $p < .05$.

The author suggests using the BayesFactor package to conduct a Bayesian analysis and see if the results support the evidence from the null hypothesis tests. The package is used to compute the Bayes factor, which quantifies the strength of evidence for one hypothesis relative to another. The author notes that this approach can provide a complementary perspective to traditional null hypothesis testing.

```{r}
aovOut3 = anovaBF(weightgain ~ source*type, data=weightgain)
aovOut3
```

```
  |============================================================
Bayes factor analysis
--------------
[1] source                       : 0.4275483 ±0.01%
[2] type                         : 2.442128  ±0.01%
[3] source + type                : 1.099992  ±0.74%
[4] source + type + source:type : 1.743962  ±0.87%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

The BayesFactor package was used to confirm the earlier findings of the analysis. The results showed that the source factor did not have a significant effect, while the type factor had weak evidence in favor of an effect. The main effects-only model had about a 1:1 odds ratio, while the interaction model had extremely weak evidence in favor of the interaction effect. It is important to compare the interaction model with the main effects-only model to determine which model is better.

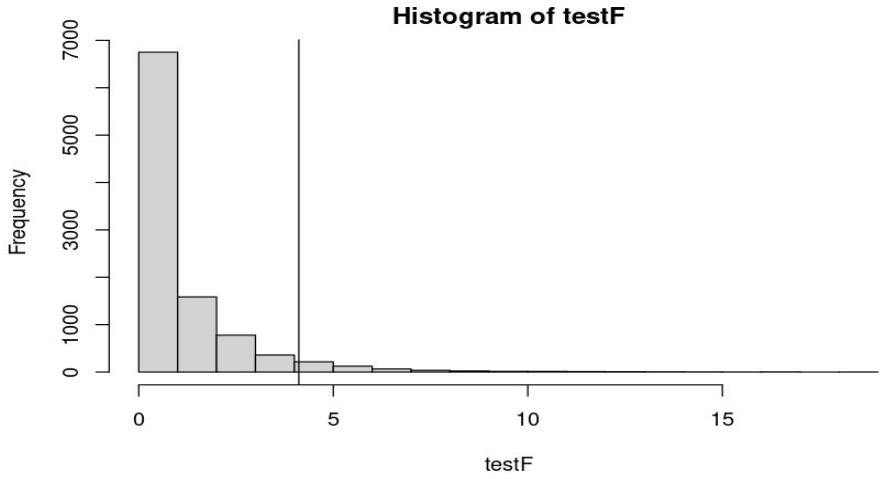
BOX ON P.187: DEGREES OF FREEDOM FOR INTERACTIONS

The ANOVA output for the test of rat weight gain showed $F(1,36) = 3.952$, $p = .0545$ for the null hypothesis test of the interaction term. The interaction has 1 degree of freedom between groups and 36 degrees of freedom within groups. The degrees of freedom for the main effects of dietary source and type are 1 less than

Originality Assertion: By submitting this file you affirm that this writing is your own.

the number of levels in each factor, leaving 37 degrees of freedom. The degrees of freedom taken by an interaction term is the product of the degrees of freedom for the main effects that go into it, in this case 1 * 1 = 1, which leaves exactly 36 degrees of freedom for the residuals that are free to vary after everything else has been calculated.

An example study with 60 participants, half male and half female, who were randomly assigned to one of three gaming conditions. The study involves two factors, gender with two levels (male and female) and game type with three levels (puzzle, adventure, and strategy). The degrees of freedom are calculated for each factor and the interaction term, with the F-test for the main effect of gender being on F(1,54) degrees of freedom, the F-test for game type being on F(2,54) degrees of freedom, and the F-test for the interaction also being on F(2,54) degrees of freedom. The rf(n, df1, df2) function is recommended to explore the shapes of the family of F-distributions.



The generated F values represent the distribution of outcomes under the null hypothesis, and we look for extreme values of the test statistic to reject the null hypothesis. The vertical line in the generated F-distribution represents the threshold of statistical significance, assuming a conventional alpha of $p < .05$.

BOX ON P.193: A WORD ABOUT STANDARD ERROR

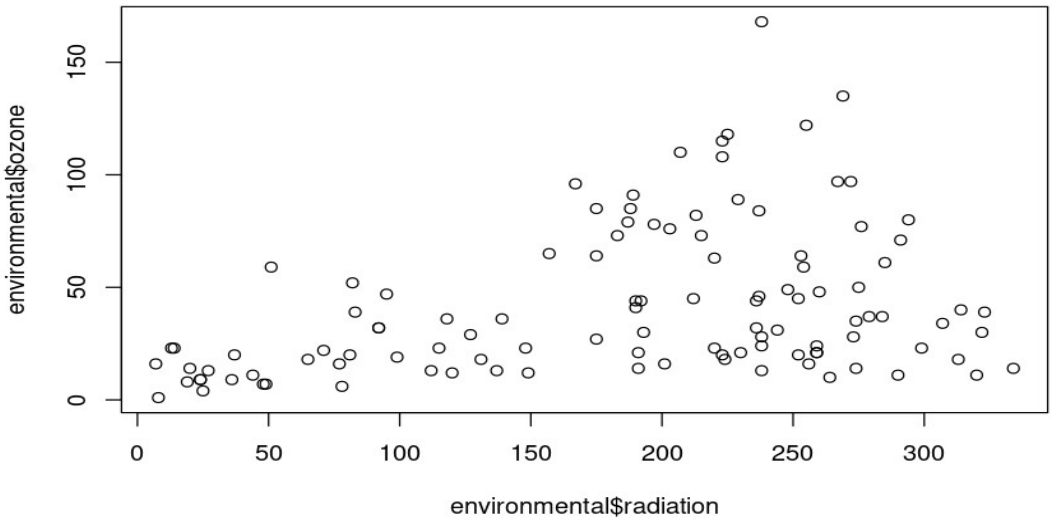
The standard error is a type of standard deviation that describes the variability of a sampling distribution of means. If we have a large random normal distribution with a mean of (near) 0 and a standard deviation of (near) 1, we can repeatedly draw samples of 100 from that population to create a sampling distribution of means. The variability of that sampling distribution can be calculated using the formula $\sigma x = \sigma / n$, where σ is the population standard deviation and n is the sample size.

IST772 Summary Template: Chapter 9 – Interactions in ANOVA and Regression
Originality Assertion: By submitting this file you affirm that this writing is your own.

The standard error is a type of standard deviation that measures the variability of a sampling distribution of means. It is calculated by dividing the population standard deviation by the square root of the sample size. The standard error becomes smaller as the sample size increases. Plus or minus two standard errors is approximately equivalent to the confidence interval, which covers the central 95% region of the sampling distribution. Confidence intervals are constructed around sample means that are normally distributed and centered on the population mean, so 95% of confidence intervals will contain the population value in the long run.
The rule of thumb suggests that if two means for two different samples of data have non-overlapping confidence intervals, a statistical test comparing them is likely to be significant. The greater the separation between the confidence intervals, the stronger the Bayes factor will tend to be. However, the rule is not perfect, as even if there is some overlap between the confidence intervals, it is still possible for the difference in means to register as significant.

INTERACTIONS IN MULTIPLE REGRESSION

The general linear model is the basis of ANOVA and regression, and there is an essential similarity to interactions in both ANOVA and multiple regression. In this section, the author uses R's built-in "environmental" dataset to explore the relationship between ozone levels and solar radiation. The author predicts that the effect of solar radiation on ozone levels will be much less pronounced when there is a strong wind, and this is an interaction prediction. The author creates a scatterplot of radiation and ozone and imposes two different regression lines on it using lm() and abline(). The code divides the data set into a high-wind portion and a low-wind portion.



IST772 Summary Template: Chapter 9 – Interactions in ANOVA and Regression
Originality Assertion: By submitting this file you affirm that this writing is your own.

The general linear model is the foundation of ANOVA and regression, and both yield similar results. In this section, the "environmental" dataset from R's "lattice" package is used to investigate the relationship between ground-level ozone and solar radiation in New York City, with wind speed as a potential moderator. A scatterplot of radiation and ozone with two regression lines (one for high wind speed and one for low wind speed) suggests that the relationship between radiation and ozone is stronger when wind speeds are low. However, since this is an observational study, it is essential to be cautious when drawing causal conclusions. The chapter then moves on to run inferential tests to determine if there is statistical evidence to support the interaction hypothesis.

```{r}
lmOut1 <- lm(ozone ~ radiation * wind, data=environmental)
summary(lmOut1)
```

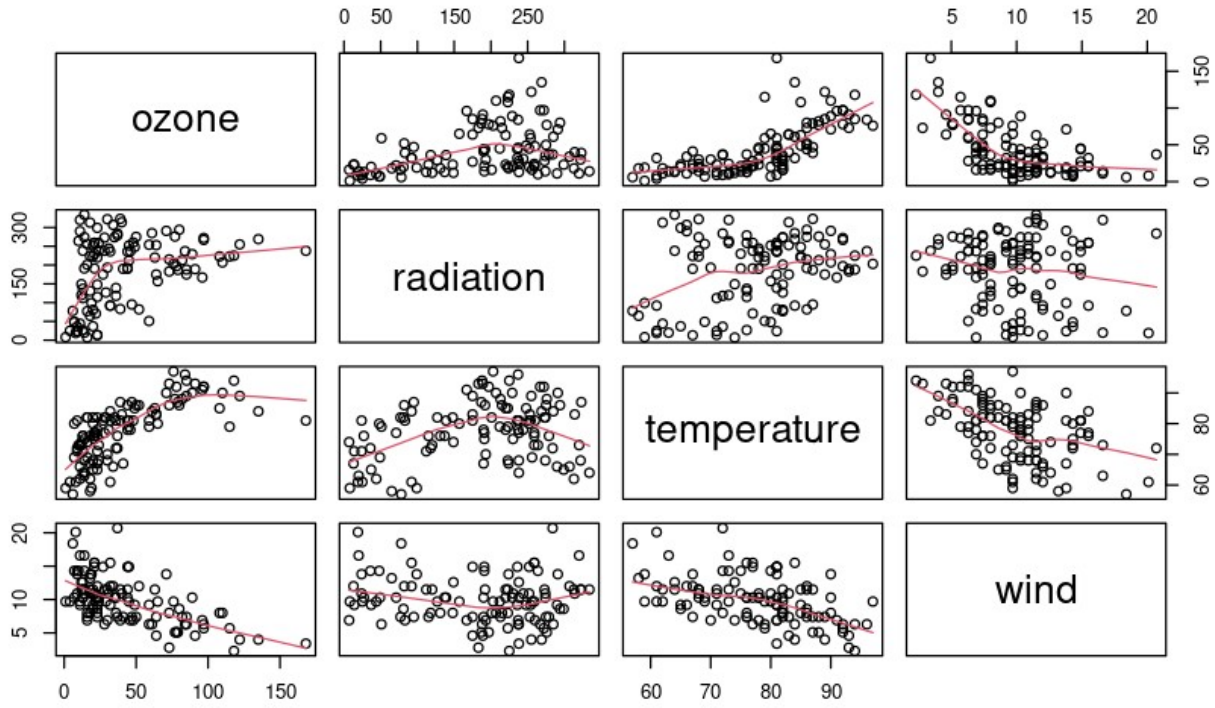
```
Call:
lm(formula = ozone ~ radiation * wind, data = environmental)

Residuals:
    Min      1Q  Median      3Q     Max
-48.680 -17.197  -4.374  12.748  78.227

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    34.48226   17.62465   1.956 0.053015 .
radiation       0.32404    0.08386   3.864 0.000191 ***
wind           -1.59535    1.50814  -1.058 0.292518
radiation:wind -0.02028    0.00724  -2.801 0.006054 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.15 on 107 degrees of freedom
Multiple R-squared:  0.4875,    Adjusted R-squared:  0.4732
F-statistic: 33.93 on 3 and 107 DF,  p-value: 1.719e-15
```

The output of a multiple regression analysis is interpreted by examining the F-test on the null hypothesis that R-squared is equal to 0, the adjusted R-squared, the coefficients, and the residuals. The observed R-squared value indicates that the independent variables accounted for about half of the variability in ozone levels. The coefficients reveal that the radiation:wind interaction term is statistically significant, radiation has a significant effect, and wind does not. The residuals suggest the possibility of a nonlinear relationship between the independent and dependent variables. Heteroscedasticity is observed in the residuals versus the independent and dependent variables, indicating that the variance of the residuals is different at different levels of the independent variable. Additional

analyses, such as including a quadratic term, may be needed to account for nonlinearity.

BOX ON P.200: DIAGNOSING RESIDUALS AND TRYING ALTERNATIVE MODELS

The chapter discusses a more in-depth analysis of residuals, particularly in the environmental data set. It identifies heteroscedasticity and a possible nonlinear connection between a predictor and the outcome variable. The chapter emphasizes the importance of scrutinizing variables beforehand to identify anomalies, such as the nonlinear relationship between radiation and ozone. Creating scatterplots of each pair of variables can help to visualize these relationships.



The pairs() command, which creates a matrix of scatterplots. The “panel=panel.smooth” argument displays a smooth curve on each scatterplot, which suggests a possible nonlinear relationship between variables. The scatterplot for radiation versus ozone shows a linear relationship at low radiation levels, but as radiation increases, the connection between radiation and ozone bends strongly to a new slope, which likely caused the anomalies observed in the

residuals. To address this anomaly, transformations such as squaring, square rooting, or taking the log of a variable can be used. For the environmental data set, a squared version of the radiation variable was added alongside the linear version to model a possible curvilinear relationship. The code to add a quadratic term for radiation is shown.

```r
env <- environmental
# Make a copy of the data set
env$radSqr <- env$radiation^2
# Add a squared version of radiation
# Include both radiation and radSqr in our new lm() model
lmOutQuad <- lm(ozone ~ radiation + wind + radSqr + radiation:wind, data=env)
summary(lmOutQuad)
# Review the model containing the new term
```

```
Call:
lm(formula = ozone ~ radiation + wind + radSqr + radiation:wind,
    data = env)

Residuals:
    Min      1Q  Median      3Q     Max
-43.985 -17.915  -2.265  14.236  78.524

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)    17.1322261 18.9310718   0.905  0.36753
radiation       0.5568418  0.1320168   4.218 5.21e-05 ***
wind           -1.1451041  1.4935471  -0.767  0.44496
radSqr         -0.0007005  0.0003106  -2.255  0.02617 *
radiation:wind -0.0205302  0.0071065  -2.889  0.00469 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.7 on 106 degrees of freedom
Multiple R-squared:  0.511,     Adjusted R-squared:  0.4925
F-statistic: 27.69 on 4 and 106 DF,  p-value: 9.6e-16
```

Adding a quadratic term to a regression model to address the nonlinear relationship between variables. The coefficient on the variable “radSqr” was found to be statistically significant, and the R-squared increased from 0.499 to 0.511. The residuals showed less heteroscedasticity and nonlinearity after adding the quadratic term. The significance of the interaction term and the linear effect of radiation were strengthened, indicating that the addition of the quadratic term was useful. To compare models with and without the quadratic term, the modelCompare() procedure is introduced later in the chapter.

IST772 Summary Template: Chapter 9 – Interactions in ANOVA and Regression
Originality Assertion: By submitting this file you affirm that this writing is your own.

BAYESIAN ANALYSIS OF REGRESSION INTERACTIONS

The chapter discusses conducting a Bayesian analysis using centered variables to enhance the analytical evidence. Two regression models are run, one with the interaction term and one without. The Bayes factor for the first model is 7.098e+11, and the Bayes factor for the second model is 4.27e+12, indicating strong support for the alternative hypothesis that radiation, wind, and their interaction predict ozone. The models are compared by creating a fraction to obtain a Bayes factor for the result.

```r
lmOutBayes1 <- lmBF(ozone ~ radiation + wind, data=stdenv)
lmOutBayes2 <- lmBF(ozone ~ radiation + wind + radiation:wind, data=stdenv)
```

The results show that the model including the interaction between radiation and wind has better odds (6:1) and is worth reporting. The interpretation of these effects requires examining the interaction plot and explaining the difference in slopes between the two regression lines. Theoretical explanations for the observed data should also be included. Generating a posterior distribution with posterior=TRUE and iterations=10000 can provide a more detailed view of the coefficients and overall R-squared.

```r
mcmcOut <- lmBF(ozone ~ radiation + wind + radiation:wind, data=stdenv,
posterior=TRUE, iterations=10000)
summary(mcmcOut)
```

How to generate a posterior distribution to obtain a more detailed view of each coefficient and the overall R-squared. The mean values of the distributions match the coefficients in the conventional regression output fairly well, although they are slightly smaller. The 95% highest density intervals (HDI) for each coefficient show the likely distribution of the population values for each coefficient. The HDI for radiation, wind, and the interaction term do not straddle 0, indicating that the coefficient of radiation is credibly positive in the population, somewhere between 0.067 and 0.170, and that the coefficients on wind and the interaction term are both credibly negative in the population.