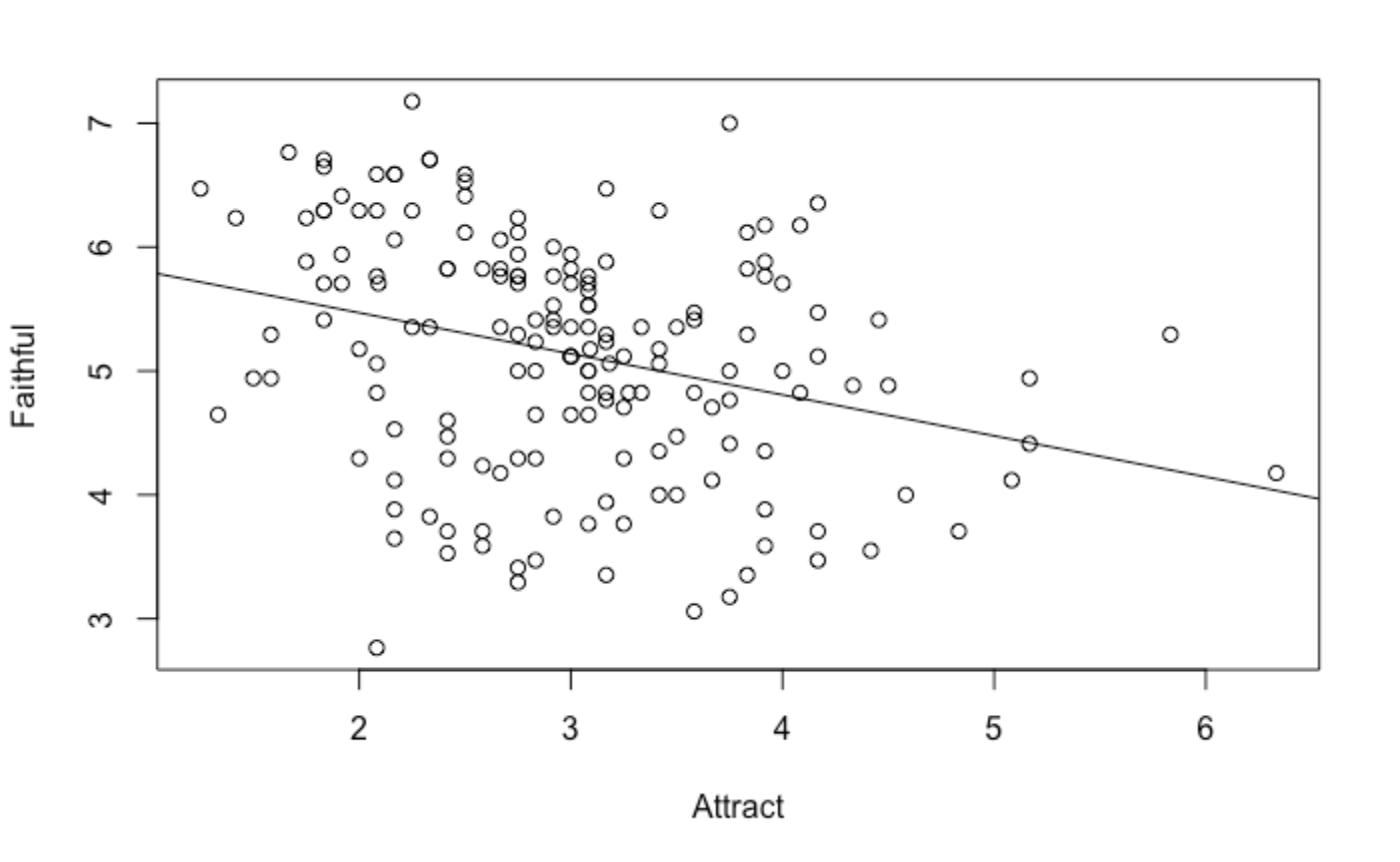
Name: Zach Yek STAT 351 Exam #2

Directions: You may either type or hand write your exam. Either way, please print it out and turn it into my office no later than 6pm on Tuesday, April 25. Complete each problem to the best of your ability. Round all final answers no fewer than two decimal places. While you may use your materials, all work needs to be completely individually. Any evidence of cheating will result in a grade of zero.

1. (25 pts.) College students were asked to look at a photograph of an opposite-sex adult face and to rate the person, on a scale from 1 (low) to 10 (high), for attractiveness. They were also asked to rate trustworthiness, faithfulness, and sexual dimorphism (i.e., how masculine a male face is and how feminine a female face is). Overall, 68 students (34 males and 34 females) rated 170 faces (88 men and 82 women). The data are in the file **FaithfulFaces**, which includes the variables Attract, Trust, Faithful, SexDimorph, FaceSex (M for males and F for females), and RaterSex (M for male raters and F for female raters).
   1. Create a simple linear model of the relationship between Faithful (response) and Attract. Is this a good model? Provide two pieces of statistical evidence that support that this is a good or poor model.



Text, application

Description automatically generated

This is an *okay* model, since while `Attract` is a strong predictor with p-value , the R-squared value of 0.09 and F-statistic of 16.67 both leave something to be desired.

* 1. We wonder if we can add predictors to this model make it better. Create a new model adding Trust and SexDimorph as predictors. Has the model improved? Discuss p-values and R2 of the new model.

Text

Description automatically generated

After adding `Trust` and `SexDimorph` as predictors to our model, we do see an improvement in the model’s (1) R-squared, and (2) F-statistic, signaling that this model does a better job at explaining the variation in the data. Additionally, all 3 predictors are statistically significant, with p-values less than the standard significance level .

* 1. How much has the model improved? Use the F test to examine the impact the two additional predictors had on the new model.

We begin by establishing the null and alternate hypotheses.

* + - ; the additional predictors do not have a significant impact on the model
    - , for ; at least one of the additional predictors has a significant impact on the model

Then, we calculate the F-statistic for the additional predictors as

where model1 refers to the 1st model, model2 the 2nd, and k the number of predictors. Using the values

shown above, we have

Moreover, with a standard significance level , the critical F-value for the F-distribution on 2

and 166 degrees of freedom is 3.074. Since 39.585 > 3.074, we have enough evidence to reject the null

in favor of the alternative. Thus, we conclude that the multiple regression model with 3 predictors is a

significantly better fit than the simple linear regression model with 1 predictor, and that the additional

predictors contribute significantly to the overall fit of the model.

* 1. Are there collinearity issues with this model? Describe using a correlation matrix.

Diagram

Description automatically generated

Text

Description automatically generated

Indeed there are. Specifically, we notice that `Attract` and `SexDimorph` are fairly strongly correlated, with

correlation coefficient 0.604. It’s also worth pointing out that `Attract` and `Trust` are also weakly

correlated, with correlation coefficient 0.364.

1. (20 pts.) An experiment was conducted to compare three different methods of repairing a meniscus (cartilage in the knee). Eighteen lightly embalmed cadaveric specimens were used, with each being randomly assigned to one of the three treatments: vertical suture, meniscus arrow, FasT-Fix. Each knee was evaluated on three different response variables: load at failure, stiffness, and displacement. The data are located in the file **Meniscus**.  For this exercise we will concentrate on the stiffness response variable (variable name stiffness) by displacement. Assume .
   1. Give the hypotheses in symbols and words that would be tested in an ANOVA procedure for this dataset.

; there is no significant relationship between Stiffness and Displacement

; there is a significant relationship between Stiffness and Displacement

* 1. Conduct an ANOVA. Report the F-statistic, degrees of freedom and the p-value.

Graphical user interface, text

Description automatically generated

F-statistic: 2.505

Degrees of freedom: 1 and 16

p-value: 0.133

* 1. Is it appropriate to report the effect size for this test? Why or why not? If so, compute the effect size between the methods with the largest and smallest means.

Note: we’re switching our predictor from Displacement to Method; the response is still Stiffness.

It is appropriate, since the F-statistic alone does not provide any information on the magnitude of the effect, nor the practical significance of the difference in means; only whether the means of the levels (methods) are significantly different from each other. Thus, reporting the effect size can help us interpret the practical significance of our results, providing more information than statistical significance alone.

Graphical user interface, application

Description automatically generated

We find that the estimated effect size between the method with the largest (method 3) and smallest (method 2) means is 1.512. The confidence interval for the effect size is also reported.

* 1. Compute a 95% confidence interval for the difference in means between methods 2 and 3. You do not need to include the interpretation (hint: there are two samples here).

(0.053, 2.971)

1. (25 pts.) Suzanne Rohrback used a novel approach in a series of experiments to examine calcium-binding proteins. The variable Calcium is the log of the free calcium concentration and ProteinProp is the proportion of protein bound to calcium. The data are found in **Fluor**.
2. Fit a regression model for predicting ProteinProp from Calcium. Write down the fitted regression equation.

Chart, scatter chart

Description automatically generated

Text

Description automatically generated

1. Are the conditions for inference reasonably satisfied for this model? Discuss the appropriate plots.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

While there do seem to be grouping/trends in the data from the residuals vs fitted plot (as well as from the y vs x scatterplot), we believe that there is little cause for concern as the data do seem to be fairly well-behaved; the residuals appear normally distributed from the Normal Q-Q plot. Hence, we find that the conditions for inference are, indeed, reasonably satisfied.

1. State the test-statistic, degrees of freedom and p-value.

Test statistic:

Degrees of freedom:

p-value:

1. State the conclusion for this problem in context.

With our hypotheses being

* + - ; there is no relationship between the log of the free calcium concentration and the proportion of protein bound to calcium
    - ; there is a relationship between the log of the free calcium concentration and the proportion of protein bound to calcium

We find that there is enough evidence to reject the null in favor of the alternative, since the stated p-

value falls below our significance level of 0.05. Thus, we conclude that there is indeed a relationship

between the log of the free calcium concentration and the proportion of protein bound to calcium.

1. Using a calcium level of -8.5, develop a 95% confidence interval for the mean proportion of protein bound to calcium. You do not need to state the interpretation.

Graphical user interface

Description automatically generated with medium confidence

(0.541, 0.614)

1. Using a calcium level of -8.5, develop a 95% prediction interval for a single proportion of protein bound to calcium. You do not need to state the interpretation.

Graphical user interface, text

Description automatically generated

(0.334, 0.821)

1. (20 pts.) Cholesterol levels are measured on a sample of 21 volunteers. Our response variable is HDL (high-density lipoprotein, or “good” cholesterol). We fit a multiple regression model with three predictors: total cholesterol (*Chol*), total triglycerides (*Triglyc*), and the presence (1) or absence (0) of a sticky substance called sinking pre-beta (*SPB*).

Call:

lm(formula = HDL ~ Chol + Triglyc + SPB, data = HDL)

Residuals:

Min 1Q Median 3Q Max

-11.5891 -2.4883 0.4158 2.4449 12.3951

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 29.49598 6.22106 4.741 0.000189 \*\*\*

Chol 0.03641 0.03469 1.049 0.308656

Triglyc 0.05893 0.02777 2.122 0.048821 \*

SPB 8.86114 2.89701 3.059 0.007105 \*\*

1. Using the output above, write down the fitted model for those individuals with SPB. (Hint: look at datatype).
2. Interpret the coefficient for individuals with SPB.

Since SPB is a binary variable with 1 indicating that an individual has SPB, the output above tells us that the model thinks that individuals with SPB have 8.861 units more HDL than those who don’t.

1. Based on the plots below and the output above, would you remove any predictors from the model? Explain.

Two scatterplots are shown.
"The first scatter-plot titled Residuals versus Fitted plots fitted values in the horizontal axis ranging from 40 to 65 with increments of 10 and Residuals in the vertical axis ranging from negative 10 to 15 with increments of 5. A curve with a negative slope originates at (40, 3), drops down gradually, and terminates at (70, negative 2). A cluster of dots lies between 40 and 55 along the horizontal axis and between negative 10 and 10 along the vertical axis. Four dots lie between 60 and 65 along the horizontal axis and between negative 5 and 0 along the vertical axis. A dot lies significantly off at (65, 12).                                 
The second scatter-plot titled Normal Q-Q plots Theoretical Quantiles in the horizontal axis ranging from negative 2 to 2 with increments of 1 and Standardized residuals in the vertical axis which ranges from negative 2 to 2 with increments of 1 unit. A dotted line with a uniform positive slope originates at (negative 2, negative 1.2) and terminates at (2, 1.2). A cluster of dots lies between negative 2 and 1 along the horizontal axis and between negative 2 and 0.8 along the vertical axis. Three dots lie between 1.2 and 2 along the horizontal axis and between 1 and 2 along the vertical axis. All data in the plot are approximate."


We notice from the above output that `Chol` is not a statistically significant predictor, with a p-value of

0.309. Additionally, the Normal Q-Q plot seems to imply that our model’s residuals are not normally

distributed. Thus, I would refit the model using only `Triglyc` and `SPB` as predictors, and investigate

further if any issues persist.

1. (10 pts.) Knit your output and upload it to OnCourse.