Homework 7

Zachary Lazerick

11/22/2021

Problem 1

Detailed interviews were conducted with over 1,000 street vendors in the city of Puebla, Mexico, in order to study the factors influencing vendors' incomes (World Development, February 1998). Vendors were defined as individuals working in the street, and included vendors with carts and stands on wheels and excluded beggars, drug dealers, and prostitutes. The researchers collected data on gender, age, hours worked per day, annual earnings, and education level. These data can be found in STREETVN.Rdata.

```
load("STREETVN.Rdata")
STREETVN = na.omit(STREETVN)
attach(STREETVN)
```

(a) Use R to fit a first-order model for mean annual earnings, $E(y)$, as a function of age (x_1) and hours worked (x_2). Summarize the results of the model fit.

```
##
## Call:
## lm(formula = EARNINGS ~ AGE + HOURS, data = STREETVN)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1105.1  -322.1   -61.0   331.9   721.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -20.352    652.745  -0.031  0.97564
## AGE           13.350      7.672   1.740  0.10738
## HOURS        243.714     63.512   3.837  0.00236 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 547.7 on 12 degrees of freedom
## Multiple R-squared:  0.5823, Adjusted R-squared:  0.5126
## F-statistic: 8.363 on 2 and 12 DF,  p-value: 0.005314
```

(b) Interpret the estimated β_i coefficients in your model separately.

β_i	**Observed Value**	**P-Value**	**Significance Level**
β_0 (Intercept)	-20.325	.97564	None
β_1 (Age)	13.350	.10738	None
β_2 (Hours)	243.714	.00236	**

From our summary of our multiple linear model, we obtain these estimated values for our β coefficients, with associated p-values and significance levels. In our model, β_0 has no practical interpretation because a Street Vendor with Age = 0 and Hours Worked = 0 is not a reasonable data point. Also, an annual income less than 0 also does not make sense. Our value for β_1 means that with every 1 unit increase in the Age of a Street Vendor, we expect that the Vendor's Income increases by 13.35, assuming that the number of Hours Worked is held constant. Our value for β_2 means that with every 1 unit increase in the number of Hours Worked of a Street Vendor, we expect that the Vendor's Income increases by 243.714, assuming that the Age of the Vendor is held constant. However, only one of these values has statistical significance.

(c) Use the output to comment on the global utility of the model (at $\alpha = .01$). Interpret the result.

From our output, we obtained an F-Statistic of 8.363 on 2 numerator and 12 denominator degrees of freedom. The associated p-value of this F-Statistic is .005314. This means that at the $\alpha = .01$ significance level, we reject H_0 and conclude that at least one of our predictors is statistically significant. However, as stated before, only one of our predictors had some statistical significance in the first place.

(d) Use the output to report and interpret the value of R_a^2.

From our output, we obtained an R_a^2 of 0.5126. This means that 51.26% of the variability in our data frame is accounted for by the model. This means that our model is not does not have a good fit for our data. Therefore, our model is not great for predictions and conclusions because only about half of the variance of our data is represented.

(e) Is age (x_1) a statistically useful predictor of annual earnings? Test using $\alpha = .01$.

Age is not a statistically useful predictor of annual earnings. As stated before in the above table, the P-Value associated with β_1 (Age) is 0.10738, which is greater than our stated significance level.

(f) Find a 95% confidence interval for β_2. Interpret the interval in the words of the problem.

```
##                     2.5 %     97.5 %
## (Intercept) -1442.561891 1401.85787
## AGE            -3.364701   30.06559
## HOURS         105.334278  382.09465
```

From our report, the 95% confidence interval for β_2 (Hours Worked) is (105.334278, 382.09465). This means that we are 95% confident that the true value for β_2 (Hours Worked) is within this calculated interval. Thus we are 95% confident that the true value for which a 1 unit increase in the Number of Hours Worked for a Street Vendor has between a 105.33 as a lower bound and 382.09 as a upper bound increase in the annual income of said Street Vendor, assuming the age of the Street Vendor is held constant.

Problem 2

Refer to the previous problem of street vendors' earnings (y).

(a) Use R to compute a 95% confidence interval for $E(y)$ for a 45-year-old vendor who works 10 hours a day (i.e., for $x_1 = 45$ and $x_2 = 10$). Interpret the confidence interval in the words of the problem.

```
##        fit      lwr      upr
## 1 3017.563 2620.252 3414.873
```

From the output, we obtain the 95% confidence interval (2620.252, 3414.873) with a calculated $E(y) = 3017.563$ for $x_1 = 45$ and $x_2 = 10$. This means the expected annual income for a 45 year old Street Vendor that works 10 hours a day, is 3017.563. However, we are 95% confident that the average value for this Street Vendor is between 2620.252 as a lower bound and 3414.873 as an upper bound.

(b) Use R to compute a 95% prediction interval for annual earnings for a 45 year old vendor who works 10 hours a day (i.e., for $x_1 = 45$ and $x_2 = 10$). Interpret the prediction interval in the words of the problem.

```
##        fit      lwr      upr
## 1 3017.563 1759.747 4275.379
```

From the output, we obtain the 95% prediction interval (1759.747, 4275.379) with a calculated $E(y) = 3017.563$ for $x_1 = 45$ and $x_2 = 10$. This means that we predict that the annual income for a 45 year-old Street Vendor that works 10 hours a day, is 3017.563. Therefore, we are confident that 95% of Street Vendors 45 years-old who work 10 hours a day annual income is between 1759.747 as a lower bound and 4275.379 as an upper bound.

Problem 3

Environmental Science and Technology (January 2005) reported on a study of the reliability of a commercial kit to test for arsenic in groundwater. The field kit was used to test a sample of 328 groundwater wells in Bangladesh. In addition to the arsenic level (micro-grams per liter), the latitude (degrees), longitude (degrees), and depth (feet) of each well was measured. These data can be found in ASWELLS.Rdata.

```
load("ASWELLS.Rdata")
ASWELLS = na.omit(ASWELLS)
attach(ASWELLS)
```

(a) Use R to fit a first-order model for arsenic level, $E(y)$, as a function of latitude, longitude, and depth. Summarize the results of the model fit.

```
##
## Call:
## lm(formula = ARSENIC ~ LATITUDE + LONGITUDE + DEPTHFT, data = ASWELLS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -135.74  -65.48  -26.38   26.97  468.55
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.351e+04  3.137e+04  -2.662  0.00816 **
## LATITUDE    -2.277e+03  5.312e+02  -4.287 2.40e-05 ***
## LONGITUDE    1.520e+03  3.741e+02   4.064 6.09e-05 ***
## DEPTHFT     -3.501e-01  1.570e-01  -2.229  0.02649 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103.4 on 320 degrees of freedom
## Multiple R-squared:  0.129,  Adjusted R-squared:  0.1209
## F-statistic:  15.8 on 3 and 320 DF,  p-value: 1.319e-09
```

(b) Interpret the estimated β_i coefficients in your model separately.

β_i	**Observed Value**	**P-Value**	**Significance Level**
β_0	$-8.351x10^4$	.00816	**
β_1 (Latitude)	$-2.277x10^3$	$2.40x10^-5$	***
β_2 (Longitude)	$1.520x10^3$	$6.09x10^-5$	***
β_3 (Depth)	$-3.501x10^-1$	.02649	*

From our summary of our multiple linear model, we obtain these estimated values for our β coefficients, with associated p-values and significance levels. In our model, β_0 has no practical interpretation because a Groundwater Well with a Latitude, Longitude, and Depth of 0 makes no sense, and would also be outside the study's scope because the well would not be located in Bangladesh. β_1 means that with every 1 unit increase in the Latitude (degrees) of a well, we expect the level of arsenic (micrograms per liter) to decrease by $2.277x10^3$, assuming that the Longitude (degrees) and Depth (in feet) of the well are held constant. β_2 means that with every 1 unit increase in the Longitude (degrees) of a well, we expect the level of arsenic (micrograms per liter) to increase by $1.520x10^3$, assuming that the Latitude (degrees) and Depth of the well

(in feet) are held constant. β_3 means that with every 1 unit increase in the Depth of the well (in feet), we expect that the level of arsenic (micrograms per liter) to decrease by $3.501x10^-1$, assuming that the Latitude (degrees) and Longitude (degrees) of the well are held constant.

(c) Use the output to comment on the global utility of the model (at $\alpha = .05$). Interpret the result.

From our output, we obtained an F-Statistic of 15.8 on 3 numerator and 320 denominator degrees of freedom. The associated p-value of this F-Statistic is $1.309x10^-9$. This means that at the $\alpha = .01$ significance level, we reject H_0 and conclude that at least one of our predictors is statistically significant. From the significance codes in our output we can see this as well, with every β coefficient having some level of statistical significance.

(d) Use the output to report and interpret the value of R_a^2.

From our output, we obtained an R_a^2 of 0.1209. This means that 12.09% of the variability in our data is accounted for by the model. Therefore, our model is not a good fit for the data.

(e) Based on the results of the previous parts, would you recommend using the model to predict arsenic level (y)? Explain.

From part b, we concluded that all of our β coefficients had some statistical significance, with our values for β_1 and $beta_2$ being greatly significant. However, our model is not a good fit of the data. This is because the reported R_a^2 value only accounts for 12.09% of the data. Overall, I would recomend this model because of the significant predictors, but feel as tho the model could also stand to be improved considerably.

Problem 4

Refer to the previous problem of arsenic level (y) in groundwater. Using the data in the ASWELLS.Rdata file, you fit a first-order model for arsenic level as a function of latitude, longitude, and depth. Based on the model statistics, the researchers concluded that the arsenic level is highest at a low latitude, high longitude, and low depth. Do you agree?

For the next two questions, we need to know the lowest latitude, highest longitude, and lowest depth that are within the scope of this study. To calculate this, we are going to use the min and max functions which return the minimum and maximum values of a data frame. For the lowest latitude, we will use min(LATITUDE), for the highest longitude, we will use max(LONGITUDE), and for the lowest depth, we will use max(DEPTHFT), because the lowest depth would be the greatest size of a well, thus the deeper the depth of the well, the bigger the well.

```
min(LATITUDE)
```

```
## [1] 23.75467
```

```
max(LONGITUDE)
```

```
## [1] 90.66169
```

```
max(DEPTHFT)
```

```
## [1] 225
```

The calls of the these functions return these respective values, which we will use for the following.

(a) Use R to compute a 95% confidence interval for $E(y)$ for the lowest latitude, highest longitude, and lowest depth that are within the range of the sample data. Interpret the confidence interval in the words of the problem.

```
##        fit      lwr     upr
## 1 163.3297 81.55829 245.101
```

From our function call, we calculated the 95% confidence interval for $E(y)$ for the lowest latitude, highest longitude, and lowest depth, within the range of the sample data to be (81.55829, 245.101), with a calculated value for $E(y) = 163.3297$. This means that we are 95% confident that the average value for the arsenic level at the lowest latitude, highest longitude, and lowest depth within the range of the sample data is within our interval. Therefore, we are 95% confident that the true highest level of arsenic in groundwater wells in Bangladesh is between 81.55829 (micrograms per liter) as a lower bound and 245.101 (micrograms per liter) as an upper bound.

(b) Use R to compute a 95% prediction interval for arsenic level for the lowest latitude, highest longitude, and lowest depth that are within the range of the sample data. Interpret the prediction interval in the words of the problem.

```
##        fit       lwr      upr
## 1 163.3297 -55.91748 382.5768
```

From our function call, we calculated the 95% confidence interval for $E(y)$ for the lowest latitude, highest longitude, and lowest depth, within the range of the sample data to be (-55.91748, 382.5768), with a calculated value for $E(y) = 163.3297$. This means that we predict that 95% of groundwater wells with the above stated values, will have an arsenic level (in micrograms per liter) of -55.91748 (micrograms per liter) as a lower bound and 382.5768 (micrograms per liter) as an upper bound. However, this entire interval is not consistent with expected arsenic levels because a negative concentration of arsenic in a groundwater well is not a possible value. So, our interval should be (0, 382.5768).

Problem 5

Refer to Problem 1 regarding street vendors' earnings (y). Recall that the vendors' mean annual earnings, $E(y)$, was modeled as a first-order function of age (x_1) and hours worked (x_2). Now, consider the interaction model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$.

(a) Use R to fit a second-order interaction model for mean annual earnings, $E(y)$. Summarize the results of the model fit.

```
##
## Call:
## lm(formula = EARNINGS ~ AGE + HOURS + AGE:HOURS, data = STREETVN)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -936.5 -281.3 -117.6  255.6  787.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1041.894   1303.593   0.799    0.441
## AGE          -13.238     29.234  -0.453    0.659
## HOURS        103.306    162.014   0.638    0.537
## AGE:HOURS      3.621      3.840   0.943    0.366
##
## Residual standard error: 550.3 on 11 degrees of freedom
## Multiple R-squared:  0.6135, Adjusted R-squared:  0.5081
## F-statistic:  5.82 on 3 and 11 DF,  p-value: 0.0124
```

(b) What is the estimated slope relating annual earnings (y) to age (x_1) when number of hours worked (x_2) is 10? Interpret the result.

Substituting $x_2 = 10$ into the interaction model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ and combining like terms yields the equation $E(y) = 2074.954 + (22.972)x_1$. This means that we expect the average annual income of a street vendor to increase by 22.972 for every 1 unit increase in the age of a street vendor, assuming that the number of hours worked by the street vendor remains fixed at 10.

(c) What is the estimated slope relating annual earnings (y) to hours worked (x_2) when age (x_1) is 40? Interpret the result.

Substituting $x_2 = 10$ into the interaction model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ and combining like terms yields the equation $E(y) = 512.374 + (248.146)x_2$. Tis means that we expect the average annual income of a street vendor to increase by 248.146 for every 1 unit increase in the number of hours worked by a street vendor, assuming that the age of the street vendor remains fixed at 40.

(d) Give the null hypothesis for testing whether age (x_1) and hours worked (x_2) interact.

H_0: $\beta_3 = 0$ H_a: $\beta_3 \neq 0$.

(e) Report the p-value of the interaction test. Give the appropriate conclusion in the words of the problem.

The p-value of the interaction test for H_0: $\beta_3 = 0$ is 0.366, with corresponding t-statistic of 0.943 on 11 degrees of freedom. Therefore, we fail to reject H_0. Thus, there is not enough statistical evidence to suggest that $\beta_3 \neq 0$. This means that there is not enough statistical evidence to suggest that there is interaction between the Age of a street vendor and Number of Hours Worked by a street vendor and how they relate to the Annual Income of a street vendor.

Problem 6

Refer to Problem 3 regarding arsenic level. Write a model for arsenic level (y) that includes first-order terms for latitude, longitude, and depth, as well as terms for interaction between latitude and depth and interaction between longitude and depth.

(a) Use R to fit the interaction model for arsenic level, $E(y)$. Summarize the results of the model fit.

```
##
## Call:
## lm(formula = ARSENIC ~ LATITUDE + LONGITUDE + DEPTHFT + LATITUDE:DEPTHFT +
##     LONGITUDE:DEPTHFT, data = ASWELLS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -176.09  -64.98  -22.58   29.23  479.00
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       16802.46   67933.02   0.247   0.8048
## LATITUDE          -1387.74    1061.31  -1.308   0.1920
## LONGITUDE           180.00     816.41   0.220   0.8256
## DEPTHFT           -1594.57     987.00  -1.616   0.1072
## LATITUDE:DEPTHFT    -10.28      11.90  -0.864   0.3882
## LONGITUDE:DEPTHFT    20.29      11.22   1.809   0.0715 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103.2 on 318 degrees of freedom
## Multiple R-squared:  0.1384, Adjusted R-squared:  0.1249
## F-statistic: 10.22 on 5 and 318 DF,  p-value: 4.294e-09
```

(b) Conduct a test (at $\alpha = .05$) to determine whether latitude and depth interact to effect arsenic level.

The p-value of the test to determine whether latitude and depth interact is .3882, with a corresponding t-statistic of -0.864 on 318 degrees of freedom. Therefore, we fail to reject H_0. Thus, there is not enough statistical evidence to suggest that $\beta_4 \neq 0$. This means there is not enough statistical evidence to suggest that latitude and depth interact to effect arsenic level.

(c) Conduct a test (at $\alpha = .05$) to determine whether longitude and depth interact to effect arsenic level.

The p-value of the test to determine whether longitude and depth interact is .0715, with a corresponding t-statistic of 1.809 on 318 degrees of freedom. Therefore, we fail to reject H_0, although it is trending towards significance. But, there is not enough statistical evidence to suggest that $\beta_5 \neq 0$. This means there is not enough statistical evidence to suggest that longitude and depth interact to effect arsenic level.

(d) Practically interpret the results of the tests, parts b and c.

As stated previously, there is not enough statistical evidence to suggest that either latitude or longitude interact with the depth of a well in Bangladesh to affect its arsenic level.