Assignment 6

Sri Harsha Sudalagunta

2023-04-02

Problem 1.

1. (3 points)

The following regression output is for predicting RFFT score of 500 randomly sampled individuals from the PREVEND data based on age (years) - 4Points.

(a) Do these data provide statistically significant evidence at the alpha = 0.01 significance level that age is associated with RFFT score? State the null and alternative hypotheses, report the relevant p-value, and state your conclusion. (3 points)

```
library(oibiostat)
data("prevend")
data("prevend.samp")
regression = lm(prevend.samp$RFFT~prevend.samp$Age)
summary(regression)
```

```
##
## Call:
## lm(formula = prevend.samp$RFFT ~ prevend.samp$Age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.879 -16.845  -1.095  15.524  58.564
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      137.54972    5.01614   27.42   <2e-16 ***
## prevend.samp$Age  -1.26136    0.08953  -14.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.19 on 498 degrees of freedom
## Multiple R-squared:  0.285,  Adjusted R-squared:  0.2836
## F-statistic: 198.5 on 1 and 498 DF,  p-value: < 2.2e-16
```

Yes, the data provided is statistically significant even when the alpha = 0.01 as the given p- value i.e., 0.000 is less then 0.01.

Null Hypothesis: There is no change in RFFT score associated with a change in 1 year of age.

H0 = 0

Alternate Hypothesis: There is a change in RFFT score associated with a change in 1 year of age.

HA != 0

But, based on the given values there is a statistical evidence that age is negatively associated with RFFT score.

(b) Compute and interpret a 99% confidence interval for the population slope. (1 point)

```
confint(regression, level = 0.99)
```

```
##                         0.5 %    99.5 %
## (Intercept)      124.579291 150.52014
## prevend.samp$Age  -1.492848  -1.02987
```

With 99% confidence, the interval (-1.49, -1.03) points contains the population average difference in RFFT score between individuals who differ in age by 1 year; the older individual is predicted to have a lower RFFT score.

Problem 2.

2. (2 pts)

The data collected from a random sample of 170 married couples in Britain, where both partners' ages are below 65 years, and fits a model predicting wife's age from husband's age. Wife's age has a mean of 40.68 years, with standard deviation 11.41 years. Husband's age has a mean of 42.92 years, with standard deviation 11.76 years. From software, the residual standard error is s = 3.95 – 4 Points.

(a) Use the summary statistics to calculate a 95% confidence interval for the average age of wives whose husbands are 55 years old. (1 point)

n =170 residual sample error(s) =3.95 sample mean(x) of the husband's age =42.92 years standard deviation = 11.76 years

Standard Error = s*sqrt(1/n + (x-mean(x))ˆ2 / ((n-1)*(sd)ˆ2)))

```
3.95*sqrt((1/170)+(55-42.92)^2/((170-1)*(11.76)^2))
```

```
## [1] 0.4349651
```

Standard Error = 0.435 The predicted wife's age is y = b0+b1(x)

```
y = 1.574+(0.9112*55)
y
```

```
## [1] 51.69
```

The wife's age is 51.69 years

95% confidence interval for the predicted wife age is CI = yś t (alpha/2, n-2)* SE

```
CI = 51.69 + qt(0.975,168)*0.435
CI
```

```
## [1] 52.54877
```

```
CI = 51.69 - qt(0.975,168)*0.435
CI
```

```
## [1] 50.83123
```

So, the 95% confidence intervals are 50.8 and 52.55

(b) You meet a married man from Britain who is 55 years old. Predict his wife’s age and give a 95% prediction interval for her age. (1 point)

Lets compute the standard error Standard Error = s*sqrt(1+1/n + (x-mean(x))ˆ2 / ((n-1)*(sd)ˆ2)))

```
SE=3.95*sqrt(1+(1/170)+(55-42.92)^2/((170-1)*(11.76)^2))
SE
```

```
## [1] 3.973877
```

The predicted wife age is 51.69

95% predicted interval is PI = yś t (alpha/2, n-2)* SE

```
PI = 51.69 - qt(0.975,168)*3.97
PI
```

```
## [1] 43.85248
```

```
PI = 51.69 + qt(0.975,168)*3.97
PI
```

```
## [1] 59.52752
```

(c) Repeat parts (a) and (b) using the approximate formulas for the appropriate standard errors (2 points)

Part (a)

The approximate 95% confidence interval is s/sqrt(n)

```
3.95/sqrt(170)
```

```
## [1] 0.3029512
```

The approximate standard error is 0.303 CI = yś t (alpha/2, n-2)* SE

```
CI = 51.69 - qt(0.975,168)*0.303
CI
```

```
## [1] 51.09182
```

```
CI = 51.69 + qt(0.975,168)*0.303
CI
```

```
## [1] 52.28818
```

The approximate confidence interval is between 51.091 and 52.28

Part (b)

The approximate 95% predicted interval is s*sqrt(1+(1/n))

```
3.95*(sqrt(1+(1/170)))
```

```
## [1] 3.961601
```

The approximate predicted error is 3.96

PI = yś t (alpha/2, n-2)* SE

```
CI = 51.69 - qt(0.975,168)*3.96
CI
```

```
## [1] 43.87223
```

```
CI = 51.69 + qt(0.975,168)*3.96
CI
```

```
## [1] 59.50777
```

The 95% confidence interval is between 43.8 and 59.5 ## Problem 3

3. The Child Health and Development Studies investigate a range of topics. One study considered all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area. The variable smoke is coded 1 if the mother is a smoker, and 0 if not. The variable parity is 1 if the child is the first born, and 0 otherwise. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, based on the smoking status of the mother and whether the child is the first born. - 4 Points.

(a) Write the equation of the regression model. (0.5 Point)

As per the given data, the equation of regression model is Regression Equation = 123.57 - 8.96(smoke)-1.98(parity)

(b) Interpret the model slopes in the context of the data. (0.5 Point)

A child born to a mother who smokes has a birth weight about 9 ounces less, on average, than one born to a mother who does not smoke, holding birth order constant. A child who is the first born has birth weight about 2 ounces less, on average, than one who is not first born, when comparing children whose mothers were either both smokers or both nonsmokers.

(c) Calculate the estimated difference in mean birth weight for two infants born to non-smoking mothers, if one is first born and the other is not. (1 Point)

The birth weight of the infant born to mother who is non-smoker and the infant is 1st born is

1st_born = 123.57 - 8.96(0)-1.98(1) 1st-born = 121.59

The birth weight of the infant born to mother who is non-smoker and the infant is not 1st born is

2nd_born = 123.57 - 8.96(0)-1.98(0) 2nd_born = 123.57

Estimated difference in mean birth weight for two infants born = 121.59-123.57 = 1.98

(d) Calculate the estimated difference in mean birth weight for two infants born to mothers who are smokers, if one is first born and the other is not. (1 Point)

The birth weight of the infant born to mother who is smoker and the infant is 1st born is

1st_born = 123.57 - 8.96(1)-1.98(1) 1st-born = 112.63

The birth weight of the infant born to mother who is smoker and the infant is not 1st born is

2nd_born = 123.57 - 8.96(1)-1.98(0) 2nd_born = 114.61

Estimated difference in mean birth weight for two infants born = 112.63-114.61 = 1.98

(e) Calculate the predicted mean birth weight for a first born baby born to a mother who is not a smoker (1 Point)

The predicted weight for a first born baby is

Regression Equation = 123.57 - 8.96(0)-1.98(parity) = 121.59 ounces

Problem 4.

4.

(a) Write the equation of the regression model that includes all of the variables. (1 Points) As the per the given information

Regression Equation = -80.41 + 0.44 (gestation) -3.33(parity) -0.01(age) + 1.15(height) + 0.05(weight) -8.40(smoke).

(b) Interpret the slopes of gestation and age in this context. (0.5 Points)

The model predicts a 0.44ounce increase in the birth weight of the baby for each additional day of pregnancy, all else held constant. Similarly, the model predicts a 0.01 ounce decrease in the birth weight of the baby for each additional year in mother's age, all else held constant.

(c) The coefficient for parity is different than in the linear model shown in the(previous question) Why might there be a difference? (1 Point)

Parity might be correlated with one of the other variables in the model, which complicates model estimation.

(d) Calculate the residual for the first observation in the data set. (0.5 Point)

baby weight = 120.58. e = 120 -120.58 = -0.58.

(e) The variance of the residuals is 249.28, and the variance of the birth weights of all babies in the data set is 332.57. Calculate the R2 and the adjusted R2. Note that there are 1,236 observations in the data set. (1 Point)

R2 = (Total Variance - Residual Variance)/ Total Variance

```
R2 = (332.57-249.28)/332.57
R2
```

```
## [1] 0.2504435
```

Adjusted R2 = 1 - [(1-R2)(n-1)/(n-k-1)]

```
AR2 = 1-(1-0.2504)*(1236-1)/(1236-6-1)
AR2
```

```
## [1] 0.2467404
```

Problem 5.

5.(a) What proportion of the variability of a child's systolic blood pressure is explained by this model? (0.5 Points)

The given R square = 0.597 It means 59.7% of the variability of a child's systolic blood pressure is explained by this model

(b) Does the least squares line indicate statistically significant associations between each of the parent's systolic blood pressures and that of the child? Explain your answer. (1 Point)

The Critical values of the parents systolic blood pressures are greater the critical t value for the respective degrees of freedom. So, we conclude that the association between the each of the parents and child are statistically significant.

(c) What is the predicted systolic blood pressure for a child whose mother's and father's systolic blood pressure is 125 mm Hg and 140 mm Hg, respectively? (1 Point)

= -15.69+0.415(125)+0.423(140)

```
predicted = -15.69+(0.415*125)+(0.423*140)
predicted
```

```
## [1] 95.405
```

(d) A colleague tells you that something must be wrong with your model because your fitted intercept is negative, but blood pressures are never negative. How do you respond? (0.5 Point)

The negative intercept doesn't mean it is incorrect.The Predicted blood pressure will be negative only when the systolic blood pressures of both mother and father are zero, which is not realistic to have blood pressure as zero. So, the negative intercept of the given equation is accurate.

(e) Briefly describe three different plots for assessing the appropriateness or fit of the above regression model.(1 Point)

Scatter plot with ab line - to check the normality of the distribution and regression line Residual Plot - to check the constant variability across the data. QQ plot: A plot of the residuals against a theoretical normal distribution.