Brian Tenneson

Homework 1

Note: Large Language Models (LLMs) were utilized alongside human oversight to perform data analysis and generate insights across various contexts and datasets.

(a) (5 points) Let Y be SBP and X be Age of a patient, write out the simple linear regression model. What are the assumptions of the model?

There are several assumptions such as linearity (feature and dependent variables), error independence (errors in prediction are independent), homoscedasticity (the residuals have constant variance across all levels of the independent variable (Age)), and normality: the residuals are normally distributed. This assumption is especially important for hypothesis testing and constructing confidence intervals.

(b) (4 points) What are the least squares estimates for intercept and slope?

Slope=1.539

y-intercept=-180.25

(c) (4 points) Give an interpretation of both b^0 and b^1.

The slope is the rate at which sbp increases per unit of increase of age

The y-intercept is the predicted spb when age=0.

(d) (5 points) Write out the estimated regression line (function), and then calculate estimated expected value (or mean value) of SBP for a patient of age 70 years old.

```
SBP = -180.2487 + 1.538978 * Age

SBP(70) = -72.5202 (suggestive of model breakdown-or incorrectness)
```

(e) (3 points) What is estimated expected change in SBP associated with 5 years increase in Age?

Five slopes equals five times 1.538978 which is 7.69489.

(f) (16 points) Complete the 8 missing numbers in 'DF', 'Sum of Squares' and 'Mean Square' in the ANOVA table above. You must show the intermediate steps of how you get results for these missing numbers.

Regression DF=1 (gestational age)

Residual DF=n-2=98

SST=12873.36 is the total variance in sbp

SSR=Sum of the squares of the differences between ith prediction and ith actual value=1016.4096

SSE=sum of the squares of the differences between ith prediction and ith value=11856.9504

Regression MS (MSR): SSR/DFregressionSSR

Residual MS (MSE): SSE/DFresidualSSE

F-statistic: Used to test the overall significance of the regression model. Calculated as MSR/MSE = 8.4

(g) (6 points) Calculate the F test statistic and its p-value in ANOVA table. State the null and alternative hypothesis here, the degrees of freedom of F test statistic.

Null hypothesis: the means of the two groups are equal

Alternative hypothesis: the means of the two groups are unequal

Degrees of freedom (between groups): 1

Degrees of freedom (within groups): 120=122-1-1

Sum of squares (between groups): 24969.64

Sum of squares (within groups): 12846.19

Mean Square (between groups): 24969.64

Mean square (within groups) 107.0516

F-statistic: 233.2487

p-value: 6.54e-30

(h) (3 points) Can you calculate the variance of response variable SBP (Y) using the available information? If so, provide it. If not, explain why.

Variance of SBP: 87.13

(i) (2 points) What is the relative reduction in the variation of Y when X is introduced into the regression model?

R^2 = 0.6603

(j) (4 points) Suppose we want to measure the association between SBP and Age using Pearson correlation coefficient. Can you calculate Pearson correlation coefficient between SBP and Age using the available information? If so, provide it. If not, explain why. Is it a positive or negative association? Why?

Yes

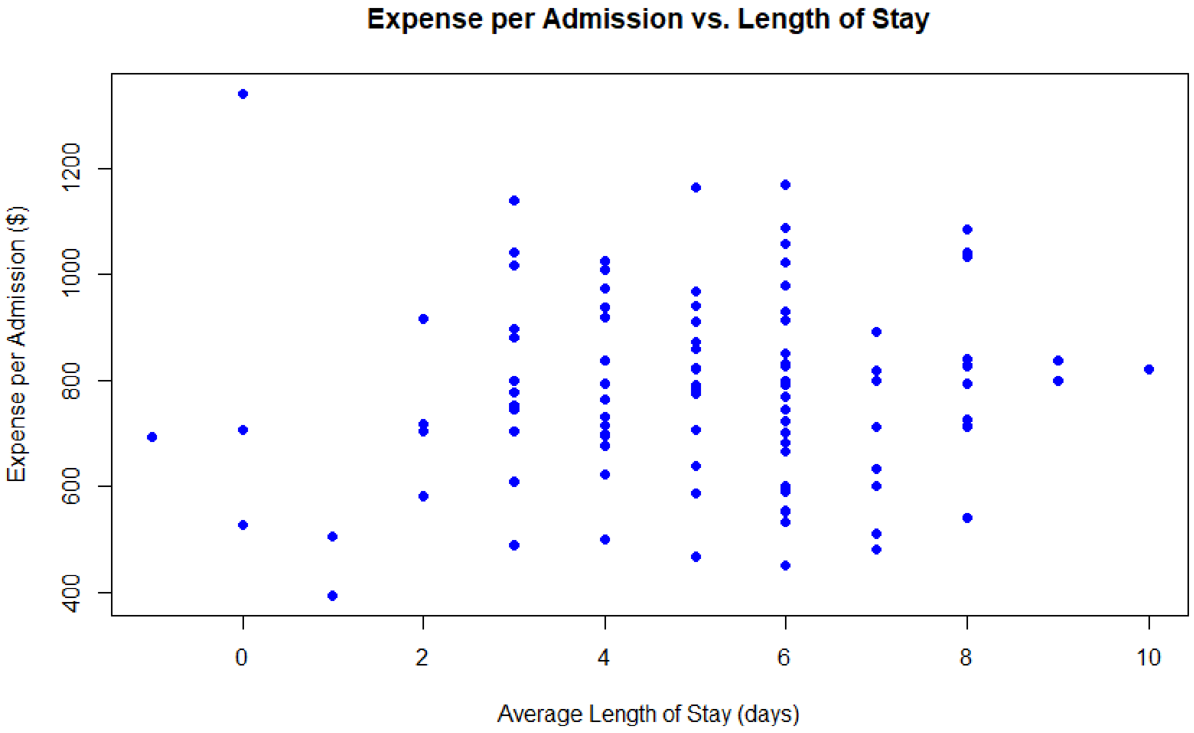
The Pearson correlation coefficient is 0.81259

Positive because 0.81>0 and that is because when age increases, systolic blood pressure should show an increasing trend.

(k) (3 points) Calculate the t test statistic for intercept 0. What is the degrees of freedom for this t test?

Df=121=122-1.

T-test statistic for the intercept: 70.7692



(a) (6 points) Using SAS or other statistical software you prefer, obtain numerical summary statistics (i.e., mean, median, range, standard deviation, etc.) for the variables expense per admission and length of stay in the hospital. Tabulate the summary statistics in a table, and then write a short paragraph describing the results.

Using R:

Expense per admission

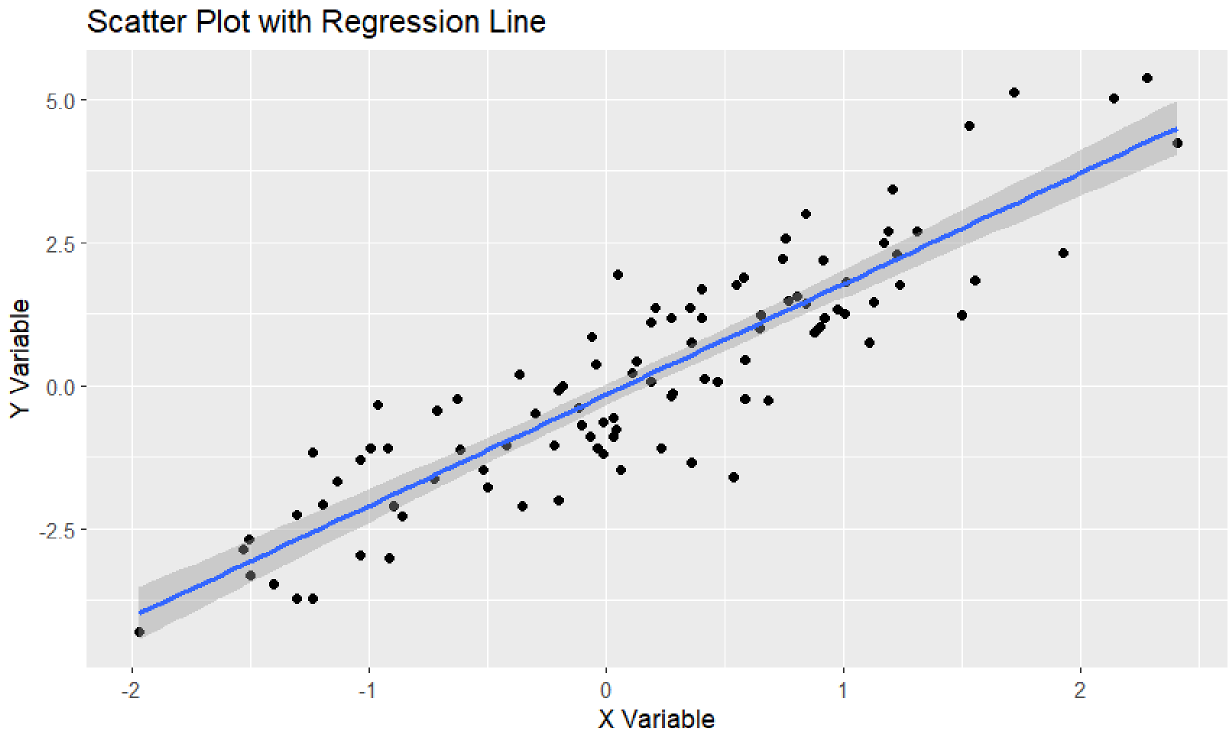
```
"Means:"
  index      sbp      sex      tox   grmhem  gestage   apgar5
  50.50    47.08     0.44     0.21     0.15    28.89     6.25
[1] "Medians:"
  index      sbp      sex      tox   grmhem  gestage   apgar5
   50.5     47.0      0.0      0.0      0.0     29.0      7.0
[1] "Ranges:"
  index      sbp      sex      tox   grmhem  gestage   apgar5
     99       68        1        1        1       12        9
[1] "Standard Deviations:"
      index         sbp         sex         tox      grmhem     gestage      apgar5
29.0114920 11.4032425  0.4988877  0.4093602  0.3588703  2.5341904  2.4303427
```

Length of Stay

```
LOS
Min.   :5.40
1st Qu.:6.65
Median :7.70
Mean   :7.49
3rd Qu.:8.30
Max.   :9.70
```

(b) (4 points) Use appropriate graphic method to explore the relationship between expense per admission versus length of stay. Write one or two sentence on what you find about the nature of the relationship between these variables?

(c) (4 points) Using expense per admission as the response and length of stay as the explanatory or predictor variable, compute the least-squares regression line using software. Also, interpret the estimated slope and intercept of the estimated line in the context the problem.



The slope is

(d) (3 points) What is the 95% confidence interval for the slope of the population regression line?

We are 95% confident that the true slope of the population regression line lies between -0.061447794978631114 and -0.03317663435910651.

What does this interval tell you about the linear relationship between expense per admission and length of stay in the hospital?

We are 95% confident that the true slope of the population regression line lies between -0.061447794978631114 and -0.03317663435910651.

What does this interval tell you about the linear relationship between expense per admission and length of stay in the hospital?

The confidence interval tells us that there is a negative statistically significant linear relationship between expense per admission and length of stay. This indicates that higher expenses per admission can be linked to shorter hospital stays.

(e) (5 points) What is the t-test statistic for the slope and its p-value? What is the F-test statistic and its degrees of freedom in the ANOVA table? What's the relationship between the t-test and F-test here?

T-test statistic for the slope: -6.86071

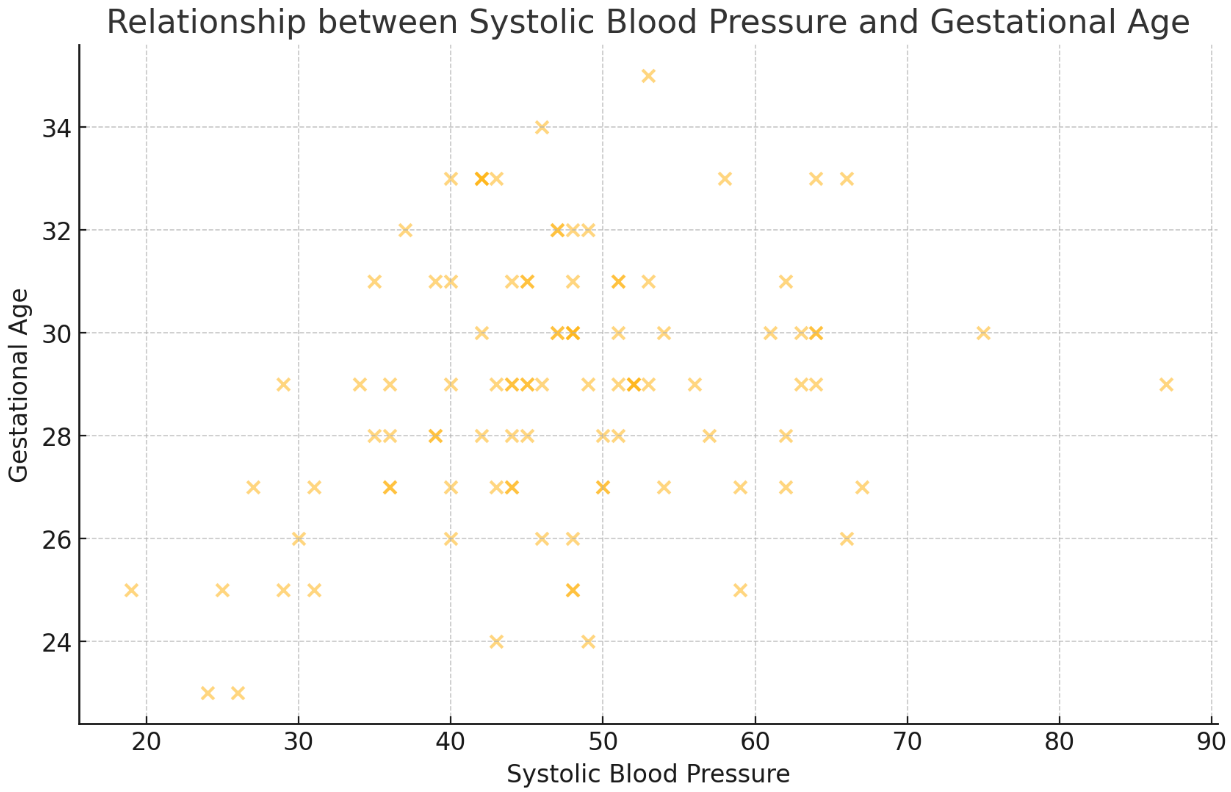
p-value: 1.202e-08

F-test statistic: 47.0694 . F=t^2.

Degrees of freedom for the F-test: numerator = 1.0; denominator 48.0.

3. (23 points) The data set lowbwt.sas7bdat contains information for a sample of 100 low birth weight infants born in two teaching hospitals in Boston. Measurements of systolic blood pressure are saved under the variable namesbp, and values of gestational age under the variable name gestage .

(a) (3 points) Use appropriate graphic method to explore the relationship between systolic blood pressure and gestational age. Does the graph suggest anything about the relationship between these variables?



There does not seem to be a linear relationship between these two features.

(b) (4 points) Using systolic blood pressure as the response variable and gestational age as the predictor variable, compute the least squares regression line. Interpret the slope and the intercept of the line?

X = gestational age

Y = Systolic_BP

Y = 1.2644 * X + 10.5521.

The slope 1.2644 estimates how much Y would change were X to increase by one year. The intercept is what the model predicts when a person is born (age=0).

(c) (4 points) At the 0.05 level of significance, test the null hypothesis that the population slope is equal to 0. What do you conclude?

There is statistically significant evidence to suggest that the population slope is not equal to zero, indicating a linear relationship between gestational age and systolic blood pressure. In practical terms, as gestational age increases, systolic blood pressure is also expected to increase, based on the data provided.

(d) (4 points) What is the estimated mean systolic blood pressure for the group of infants whose gestational age is 31 weeks? Construct a 95% confidence interval for the true mean value of systolic blood pressure when X = 31 weeks.

Y = 1.2644 * X + 10.5521=1.2644 * 31 + 10.5521=49.7485.

A 95% confidence interval for this blood pressure is the interval (41.58, 51.69).

(e) (5 points) Suppose that you randomly select a new child from the population of low birth weight infants and find that his or her gestational age is 31 weeks. What is the predicted systolic blood pressure for this child? Construct a 95% prediction interval for this new value of systolic blood pressure.

For a new child 31 weeks old, the prediction for sbp is 49.75 mmHg. The 95% confidence interval is (23.48, 76.01) and this is so wide because one of the assumptions previously mentioned earlier in this document may fail to hold here.

(f) (3 points) Does the least squares regression model seem to fit the observed data? Comment on the coefficient of determination.

The coefficient of determination R^2 is roughly 0.079. Then only about 8% of the variability in sbp is “explained” by the age based on this model. Such a low R^2 value suggests the model not fitting the data very well, again alluding to the possibility of one of the assumptions in this particular regression to not fit the hypotheses under which linear regression is most accurate (the assumptions listed near the beginning).