Problem Set 2

Cody Frisby

January 23, 2016

#1 Intent of the Researches?

I think the intent of the researches is to create a reliable model that allows them to not only predict blood pressure for people between 23 and 67 but also to have an understanding of perhaps a healthy range for a given age. If there is a linear relationship between the variables of interest, a linear regression model would provide a simple and reliable way to predict and infer relationships between age and blood pressure.

#2 Find SSY, SSX, SXY, $\bar{x}$, $\bar{y}$.

- ssx = $\sum(x^2)$ - $\frac{\sum(x)^2}{n} = 5.2119 \times 10^4 - \frac{1.164241 \times 10^6}{24} = 3608.9583333$
- ssy = $\sum(y^2)$ - $\frac{\sum(y)^2}{n} = 4.74053 \times 10^5 - \frac{1.1148921 \times 10^7}{24} = 9514.625$
- sxy = $\sum(xy) - \frac{\sum(x)\sum(y)}{n} = 155921 - \frac{3602781}{24} = 5805.125$
- $\bar{x} = \frac{\sum(x)}{n} = \frac{1079}{24} = 44.9583333$
- $\bar{y} = \frac{\sum(y)}{n} = \frac{3339}{24} = 139.125$

#3 Determine the correlation coefficient between x and y

$$Cor(X,Y) = \frac{SXY}{\sqrt{SSX * SSY}}$$

$$0.9906604 = \frac{5805.125}{\sqrt{3608.9583333 * 9514.625}}$$

#4 Is SLR appropriate?

From the correlation coefficient and the scatter plot it appears that simple linear regression would be appropriate.

#5 Least Squares regression Line

Equation of least squares regression line:

$$BloodPressure = 66.8080817 + 1.608532 * Age$$

The intercept, 66.8081, is computed $\bar{y} - \bar{x} * \hat{\beta}_1$ and is interpreted as the mean blood pressure when someone's age is 0. $\hat{\beta}_1$, the coefficient to age, is equal to 1.6085 and is computed $\frac{sxy^2}{ssx}$. This is to be interpreted that for every unit increase in age there will be an approximate 1.609 increase in blood pressure, *on average*.

#6 Estimated common standard deviation for sub populations:

This can be calculated by the formula

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-2}}$$

Where $SSE = SSY - \frac{SXY^2}{SSX}$

So for our data $\hat{\sigma} = \sqrt{\frac{9514.625 - \frac{33699476}{3608.9583}}{22}} = 2.835615$

#7 Coefficient of determination:

This can be calculated from the sample data a couple of ways. Its value is always between 0 and 1. Its interrpretation is the porportion of the variation in the response variable that can be explained by the predictor variable. Above, where we calculated to coefficient of correlation, we can use the square of it's value to obtain the coefficient of determination. So, $0.9906604^2 = 0.981408$ is the value for the coefficient of determination, also represented by r^2. So, approximately 98% of the variation in blood pressure can be explained by age.

#8 Mean blood pressure for age 52.

To find the mean blood pressure for someone who is 52 we would simply plug 52 in for age into our least squares equation

$$BloodPressure = 66.8080817 + 1.608532 * Age$$

$$150.4517462 = 66.8080817 + 1.608532 * 52$$

#9 Discuss Assumptions

- Linearity:
  The assumption of linearity is valid. There appears to be a strait line relationship between the two variables. There are not any outliers that we need to be concered about.

- Equal Variances:
  This assumption involves assuming that the variance, or standard deviations, for the sub populations are approximately equal. Graphically this entails that we have random scatter of our models standardized residuals about a horizontal line at zero. This can be a subjective exercise, but for our assumption to be valid, we do not want to see any obvious patterns or funneling. It appears by this plot that the homogeneity of variances is valid.

- Normality:
  This assumption involves that each subpopulation of blood pressure determined by age is a Gaussian population. One way to graphically investigate this is to create a qqplot. If this plot falls on an approximate strait line with few points straying off this strait line then our assumption of normality is valid. From the qqplot on the printout it appears this assumption is valid. Another plot is the histogram of the residuals. This one looks ok too. The data looks approx. Gaussian.

#10 For people 52 years old find a 95% confidence interval for blood pressure:

From number 8 above we showed that the predicted value of blood pressure when age is 52 = 150.4517462. Now to calculate a 95% confidence interval. The general form for our least square regression paramater

estimators is given by:

$$\hat{\theta} - (tablevalue) * SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + (tablevalue) * SE(\hat{\theta})$$

We calculate $SE(\hat{\mu}(x))$ by the following:

$$SE(\hat{\mu}(x)) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{SSX}}$$

$$0.6674612 = 2.8356146\sqrt{\frac{1}{24} + \frac{(52 - 44.95833)^2}{3608.9583333}}$$

And our t statistic, with 22 degrees of freedom and $\alpha = 0.05$, = 2.0739. Now plugging in $\hat{\mu}(x)$, t statistic, and $SE(\hat{\mu}(x))$ to our general form confidence interval we get:

$$150.4517462 - 2.0738731 * 0.6674612 \leq \mu(52) \leq 150.4517462 + 2.0738731 * 0.6674612$$

$$149.0675164 \leq \mu(52) \geq 151.835976$$

$$[149.0675164, 151.835976]$$

The 95% confidence interval for x = 52

#11 Find a 95% prediction interval:

From number 8 above we showed that the predicted value of blood pressure when age is 52 = 150.4517462. Now to calculate a 95% prediction interval:

$$SE(\hat{Y}(x)) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SSX}}$$

$$2.9131108 = 2.8356146\sqrt{1 + \frac{1}{24} + \frac{(52 - 44.95833)^2}{3608.9583333}}$$

$$150.4517462 - 2.0738731 * 2.9131108 \leq Y(52) \leq 150.4517462 + 2.0738731 * 2.9131108$$

$$144.4103243 \leq Y(52) \leq 156.4931682$$

95% prediction interval for x = 52

$$[144.4103243, 156.4931682]$$

#12 Find 95% confidence interval for β_1

Our estimate for β_1 is 1.608532. The standard error for $\hat{\beta}_1$ is:

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SSX}}$$

So for our data $SE(\hat{\beta}_1) = \frac{2.8356}{\sqrt{3608.9583}} = 0.0472$. And our t statistic, with 22 degrees of freedom and $\alpha = 0.05$, $= 2.0739$

$$\hat{\beta}_1 - 2.0739 * SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + 2.0739 * SE(\hat{\beta}_1)$$

So, the confidence interval is calculated by plugging into the above equation:

$$1.608532 - 2.0738731 * 0.0472016 \leq \beta_1 \leq 1.608532 + 2.0738731 * 0.0472016$$

95% confidence interval for β_1:

$$[1.510642, 1.706422]$$

#13 Find simultaneous confidence intervals for x=42 and x=52

The interval formulation is the same as above:

$$\hat{\theta} - (tablevalue) * SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + (tablevalue) * SE(\hat{\theta})$$

The table value will be different since we want simultaneous confidence for m = 2. Looking up the value using the table on pg. 684 in the text book with 0.95 confidence and df=22 we find 2.405.
First, for x = 52:

$$150.4517462 - 2.405 * 0.6674612 \leq \mu(52) \leq 150.4517462 + 2.405 * 0.6674612$$

$$148.8465021 \leq \mu(52) \leq 152.0569904$$

And for x = 42:

$$134.3664261 - 2.405 * 0.6674612 \leq \mu(42) \leq 134.3664261 + 2.405 * 0.6674612$$

$$132.761182 \leq \mu(42) \leq 135.9716703$$

So, the 95% simultaneous confidence interval is:

$$[132.761182, 135.9716703 \ \ and \ \ 148.8465021, 152.0569904]$$

#14 How useful is this model?

The Residual standard error from our model is 2.8356146. If $\mu_y(x) = x$ then 99% of individuals that are within $\pm$ 8 (the max distance from the mean for the Dr) of the average will be between $\mu_y(x) - z_{0.995}\sigma \ \ and \ \ \mu_y(x) + z_{0.995}\sigma$. So, $2.5758293 * 2.8356146 = 7.304059$. And a confidence interval:

$$[\hat{\mu}(x) - 7.304059, \ \hat{\mu}(x) + 7.304059]$$

This is within the 8 value that the Dr is interested in, and, assuming robustness of our asumptions, we conclude that the model will be useful to him/her.

Note: I got this one wrong so I am going to edit it here
First we need to calculate a confidence interval for σ. The formula for this is as follows

$$C\left[\sqrt{\frac{(df)\hat{\sigma}^2}{\chi^2_{1-\alpha:df}}} \leq \sigma \leq \sqrt{\frac{(df)\hat{\sigma}^2}{\chi^2_{\alpha/2:df}}}\right] = 1-\alpha$$

$$C\left[\sqrt{\frac{22 \times 8.0407099}{33.9244385}} \leq \sigma \leq \sqrt{\frac{22 \times 8.0407099}{12.3380146}}\right] = 0.9$$

90% confidence interval for σ

$$[2.2835062, 3.786482]$$

Now using the z score from above and that we are 95% confident that $\sigma \leq 3.786482$ we can say that $2.5758293 \times 3.786482 = 9.7533313$. We cannot conclude that the model will be useful to the Dr.

#15 Conclusions

This experiment appears to have genterated data that built a very good linear model. By good I mean one with small errors. It seems like this could be of benefit to the medical community in that it should predict, with minimal error, where one should be given their age and if they are not perhaps identify potential health risks with the individual. Additionaly, this study should lead to other studies where we confirm or modify our findings.