Methods Homework 8

Tim Vigers

04 November 2018

1. Show that SSTotal = SSModel + SSError

$$SS_{\text{Total}} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}((Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}))^2 =$$

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

From the lecture notes we know that:

$$SS_{\text{Error}} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \text{ and } SS_{\text{Model}} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

So:

$$SS_{\text{Total}} = SS_{\text{Error}} + SS_{\text{Model}} + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

$$2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) =$$

$$2\sum_{i=1}^{n} Y_i\hat{Y}_i - \bar{Y}Y_i - \hat{Y}_i^2 + \bar{Y}\hat{Y}_i$$

This can be rearranged to:

$$2(\sum_{i=1}^{n}\hat{Y}_i(Y_i - \hat{Y}_i) - \bar{Y}(Y_i - \hat{Y}_i)) = 2(\sum_{i=1}^{n}\hat{Y}_i(Y_i - \hat{Y}_i) - \sum_{i=1}^{n}\bar{Y}(Y_i - \hat{Y}_i))$$

$\bar{Y}$ is a constant, so it can be pulled out of the sum, meaning:

$$\sum_{i=1}^{n}\bar{Y}(Y_i - \hat{Y}_i) = \bar{Y}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \bar{Y} * 0 = 0$$

Because $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$:

$$\sum_{i=1}^{n}\hat{Y}_i(Y_i - \hat{Y}_i) = \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 X_i)(Y_i - \hat{Y}_i) = \hat{\beta}_0\sum_{i=1}^{n}(Y_i - \hat{Y}_i) + \hat{\beta}_1\sum_{i=1}^{n}X_i(Y_i - \hat{Y}_i)$$

As before, we've pulled out the constants and know both of the above sums are equal to 0. Therefore:

$$2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 2(0 + 0 - 0) = 0$$

Which leaves us with:

$$SS_{\text{Total}} = SS_{\text{Error}} + SS_{\text{Model}} + 0$$

2. Regressions

A. SAS output:

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: chol Cholesterol (mg/100mL)

Number of Observations Read	7
Number of Observations Used	7

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27382	27382	7.74	0.0388
Error	5	17699	3539.79126		
Corrected Total	6	45081			

Root MSE	59.49614	R-Square	0.6074
Dependent Mean	339.14286	Adj R-Sq	0.5289
Coeff Var	17.54309		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	97.39533	89.78167	1.08	0.3275	-133.39580	328.18646
wtkg	Weight (kg)	1	3.72738	1.34017	2.78	0.0388	0.28236	7.17241

B. Least squares regression equation

$$\text{Intercept estimate: } 97.39533 = \hat{\beta}_0$$

$$\text{Variable estimate: } 3.72738 = \hat{\beta}_1$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

$$\hat{\text{Cholesterol}} = 97.39533 + 3.72738 * \text{Weight (kg)}$$

C. Inference about the intercept

i. The estimated intercept is 97.39533. This means that on average, if someone weighed 0 kilograms, their estimated cholesterol would be 97.39533 mg/100mL. This probably isn't biologically meaningful.
ii. SAS provides a 95% CI for the intercept, which is between -133.39580 and 328.18646. This means that we're 95% certain that someone who weighs 0 kilograms would have cholesterol between -133.39580 and 328.18646. Again, a lot of that range is biologically irrelevant (negative cholesterol doesn't really mean anything), but it's what the regression tells us.
iii. SAS also tests this hypothesis for us and gives a p-value of 0.3275. This means that at a significance level of 0.05, we can't reject the null hypothesis that the intercept is 0.
iv. It doesn't make sense to look at this intercept particularly closely, since we probably don't care what the estimated cholesterol is for someone who weighs 0 kilograms, as it's physically impossible.

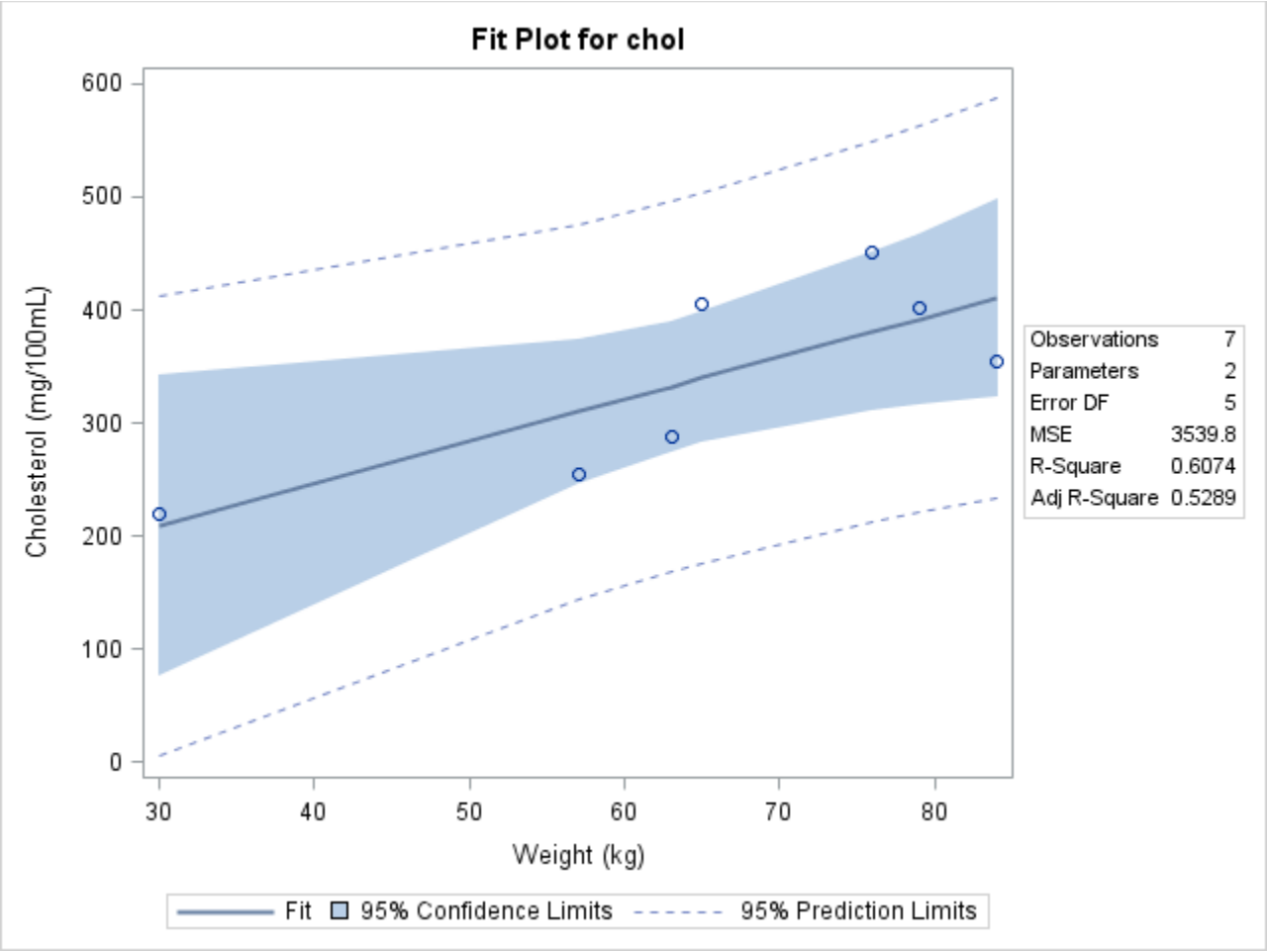
D. Inference about the slope

i. The estimated slope is 3.72738, which means that on average we would expect cholesterol to increase by 3.72738 mg/100mL for every 1 kg increase in weight.
ii. SAS also provides a 95% CI for the slope, which is between 0.28236 and 7.17241. So, for every 1 kg increase in weight, we are 95% certain that cholesterol would increase between 0.28236 and 7.17241 mg/100mL.
iii. SAS also tests the hypothesis that the true slope is equal to 0 for us. The p-value is 0.0388, so at the 0.05 significance level, we can reject the null hypothesis that the true slope is 0. This indicates a real relationship between weight and cholesterol.

E. Effect summary

There is a significant increase in cholesterol as weight increases ($p < 0.05$). On average, cholesterol increases by 3.72738 mg/100mL (95% CI: 0.28 to 7.17 mg/100mL) for every 1 kg increase in weight.

F. Scatterplot (included in PROC REG)


Fit Plot for chol
Cholesterol (mg/100mL)
600
500
400
300
200
100
0
30
40
50
60
70
80
Weight (kg)
Observations 7
Parameters 2
Error DF 5
MSE 3539.8
R-Square 0.6074
Adj R-Square 0.5289
Fit
95% Confidence Limits
95% Prediction Limits

3. Moser & Stevens paper

Moser and Stevens compare three different approaches to comparing the means of two independent, normally distributed populations: The Smith/Welch/Satterthwaite (SWS) test, the t test, and what they refer to as the "sometimes t test" (ST), which is a preliminary variance test follow by either an SWS test or a t test. In order to compare the tests, they examined the power and size of each test under different conditions of sample sizes and variance ratios. They found that with equal sample sizes, all three tests have the same power and size, which makes a variance test unnecessary, and means that either the SWS or t test is appropriate. When sample sizes are unequal, but the variance ratio is close to 1, they recommend the t test. And finally, when sample size is unequal, and the variance ratio is known to be different from 1, they recommend the SWS test. So, in the future, I will eliminate the preliminary variance testing step, and base my choice between an SWS test and a t test mostly on sample size (assuming there's nothing really bizarre happening with data, both populations are normally distributed, etc.).