Homework 01

Spencer Pease

1/15/2020

(Q1)

(Q1.a)

This study is an observational study, as it is looking at the results a clinical study, not performing the clinical study itself.

(Q1.b)

This article presents the findings without any mention of the confidence or level of significance of the findings. Without these qualifications, the article implies that the results should be taken as the truth, which they then use to assert a casual relationship that could be adapted to future interventions.

(Q2)

(Q2.a)

α represents the length of the hanging string when no external weight is attached to it (the "unstretched" length).

(Q2.b)

One label capturing what β measures is the "elasticity", or "stretchiness coefficient". β signals how much the length of the string will change for a given amount of weight suspended from the string.

(Q2.c)

We should expect beta to be larger for a rubber band than a rope. From experience, a rubber band will stretch much further than a rope will under the same force.

(Q2.d)

It is reasonable to estimate the length of the string under a 30-ounce weight because we are estimating within the range of data used to fit the model, meaning the estimate is likely to follow the truth observed in the data.

(Q2.e)

It is somewhat reasonable to estimate the length of the string under an 80-ounce weight, because even though the estimate is outside the range of the weight in the data used to fit the model, it is not too far from that range, and 80 ounces is not an unrealistic weight to estimate in this domain.

(Q2.f)

It is not reasonable to estimate the length of the string under a 120-ounce weight, since we know the string will break under weights greater than 100 ounces, making it an unrealistic weight to estimate for our model.

(Q2.g)

It is completely reasonable to use the fitted model to estimate $E(Y|x = -10)$, since the expected value E has only to do with the model, not what the model represents. The model can determine the expected value of -10, even though a -10-ounce weight is impossible in the real world.

(Q2.h)

I would prefer the estimate, as that accounts for errors in the measurement of our data points.

(Q2.i)

Table 1: Inference table for regression slope.

Term	Point est.	P-val	2.5%	95.7%
weight	0.751	1.12e-08	0.586	0.915

When fit using the provided data, we estimate the length of the unweighted string ($\hat{\alpha}$) to be **21.07 inches**, and the difference in length between two attached weights one ounce apart ($\hat{\beta}$) is estimated to be **0.751 inches**. With 95% confidence, we estimate the true difference in length between groups per ounce to be as reported in the above table (*table 1*). The reported *P-value* suggests that we reject the null hypothesis that attached weight does not affect the length of the hanging string.

(Q2.j)

TBD

(Q2.k)

With r defined as the correlation between length and weight, we can find β_1 using the equation:

$$\beta_1 = r \cdot \frac{sd(length)}{sd(weight)}$$

Comparing this value to the slope produced by the linear regression, we find that the difference in values is **2.22e-16**, which is as close to zero as the limits of machine precision will allow. Therefore, these two methods produce the same result.

(Q3)

(Q3.a)

In order to test the association between height and weight for every unique integer height in the data, Chris will need to perform **19** t-tests. If he wants an overall significance level of $\alpha = 0.05$, then each individual test will need a significance threshold of $\alpha^c = \alpha \div n = 0.00263$.

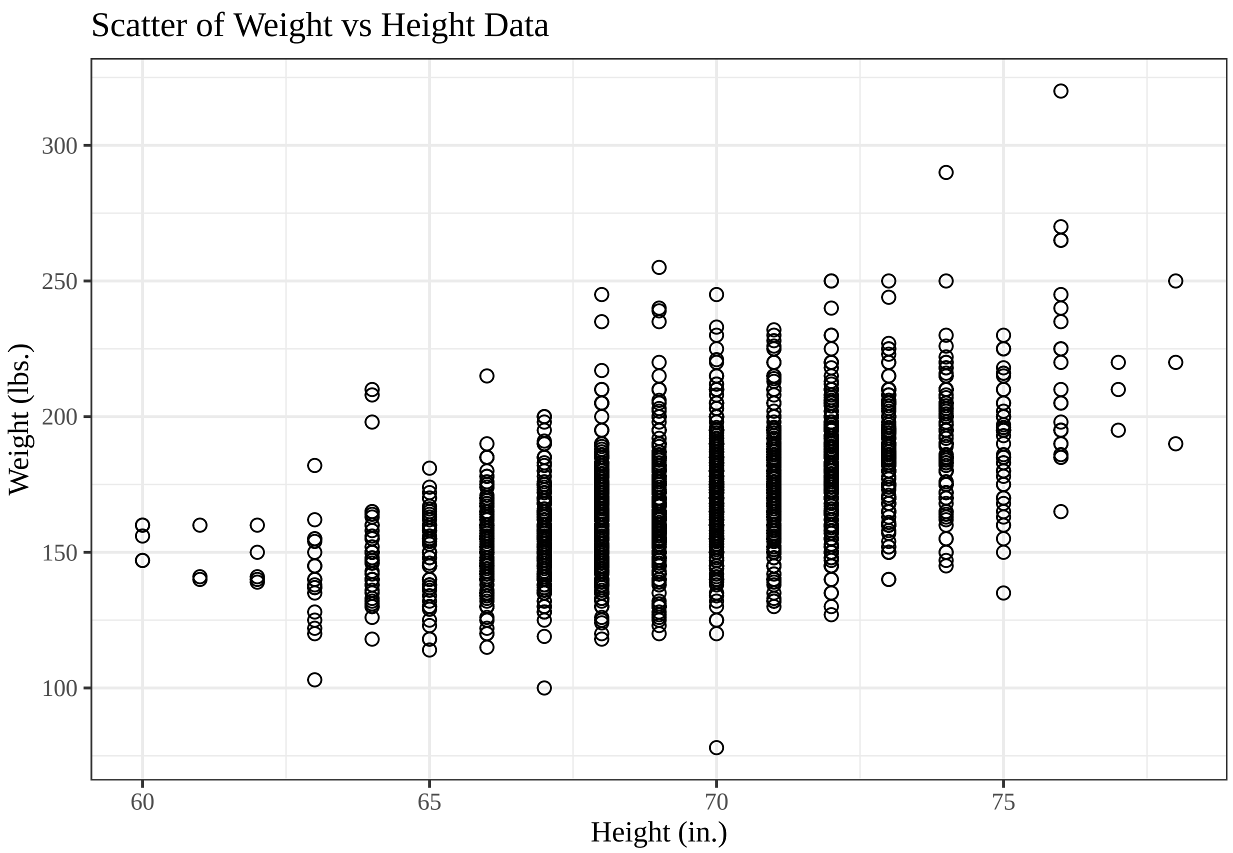
(Q3.b)

The merit of Angela's approach is dichotomizing the heights into two groups means there are fewer test to perform with larger samples. The issue with arbitrary binning is that it ignores the relationship between data points on the edge of the threshold. Also, she is basing her model on what will give the most significant results, instead of using setting up the model to model the actual phenomena.

(Q3.c)

Term	Point est.	P-val	2.5%	95.7%
height	4.446	5.61e-231	4.2	4.693

From the fitted model, we estimate that the difference in weight between two groups differing by one inch in height is **4.446 inches**. We are 95% confident that the true difference is between **4.2** and **4.693** inches, and that this result would be highly unlikely to observe if there was no true association.



(Q3.d)

From this limited data, I think there is at least a first-order linear trend. Using a linear model, we can only look at linear trends, so any additional relationship is unobservable in this data. This means that in our summary we can't say that the relationship is linear, only that there are at least linear trends.

(Q3.e)

In the lower and upper ends of height, the distribution of weights is more narrow, suggesting that the assumption of homoscedasticity is not valid.

(Q3.f)

Since data with a height of 70 inches was in the range of data used to fit the model, it is reasonable to use the model to estimate this height.

(Q3.g)

It is less reasonable

(Q3.h)

It is unreasonable, as this is a different population than was fit with the model.

(Q3.i)

Table 3: wcgs model in metric

Term	Point est.	P-val	2.5%	95.7%
height_cm	0.794	5.61e-231	0.75	0.838

We see that changing the units doesn't affect the p-values of the model.