DATA557 - HW07

Will Wright

February 24, 2019

DATA 557

Homework Assignment 7

Question 1

Data set for Question 1: 'cells.csv'

Summary: a randomized clinical trial of immune cell stimulation in 40 patients

Variables:

id: subject id # dose: drug dose (0, 10, or 100mg) sex: sex (0=female, 1=male) age: age (yrs) count0: pre-treatment cell count count1: post-treatment cell count (the response variable)

Use linear regression to estimate the effect of dose on post-treatment cell count, with adjustment for sex, age, and pre-treatment cell count. State the interpretation of the estimated dose effect. Using your model, conduct a test of the null hypothesis of no effect of dose after adjustment for sex, age, and pre-treatment cell count. Do you think this test is valid? Provide justification by assessing the model assumptions.

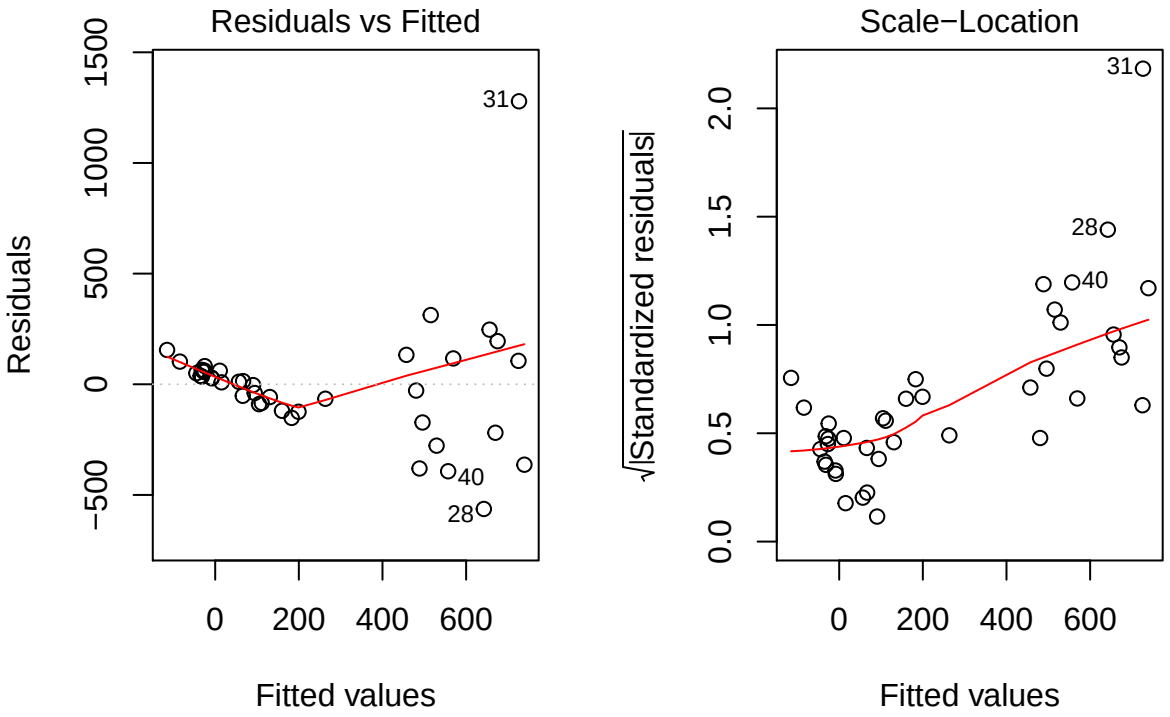
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	274.317	174.480	1.572	0.125
dose	5.810	1.021	5.689	0.000
sex	26.230	96.015	0.273	0.786
age	-5.933	3.116	-1.904	0.065
count0	0.815	0.753	1.082	0.287

Interpretation: When sex, age, and pre-treatment cell counts are held constant, this model shows that for every 1mg increase in dose, post-treatment cell count increases 5.81, on average.

The p-value for dose is essentially 0 so we can reject the null hypothesis that dose has no effect on post-treatment cell count with sex, age, and pre-treatment cell count held constant.

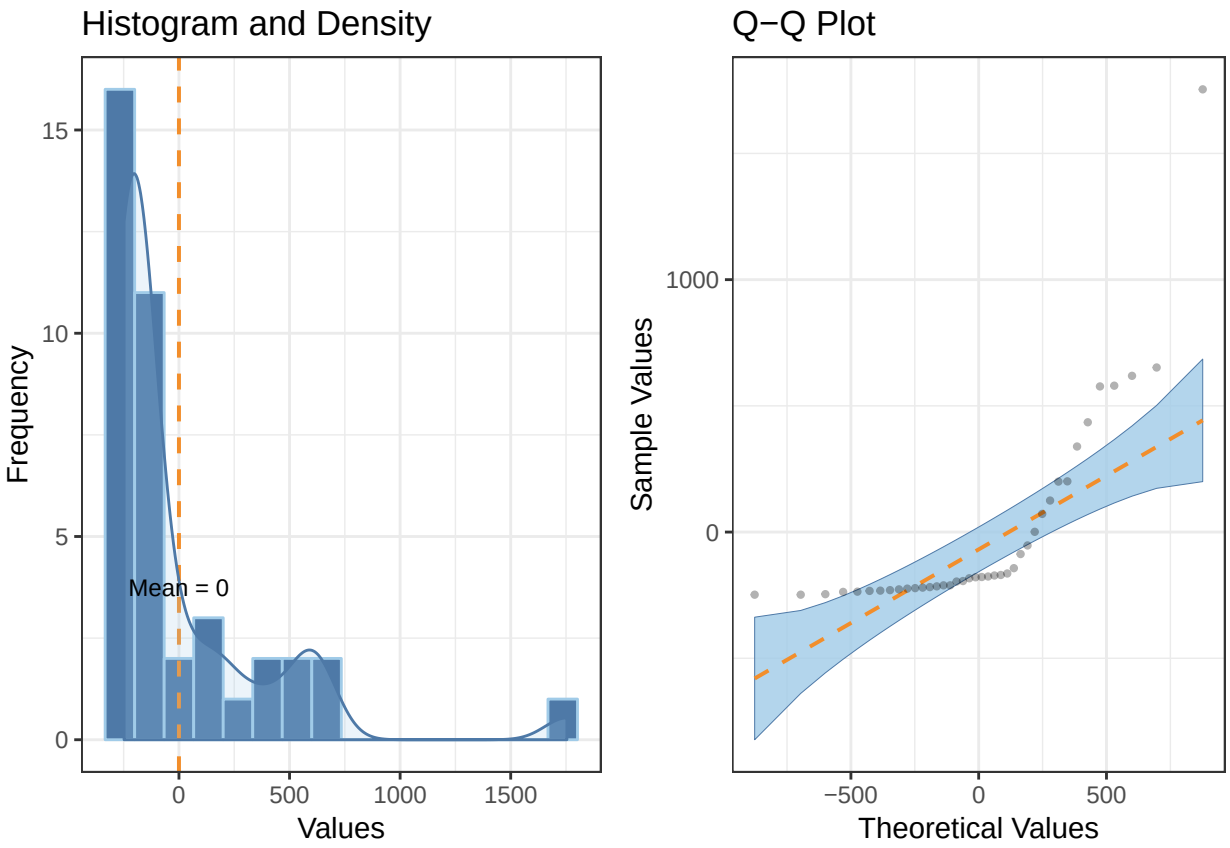
Assessing Model Assumptions

Independence: Given that the study is a randomized clinical trial, independence likely holds. We do not, however, have all the relevant information to truly assess. It would be useful, for instance, to know how patients were selected, if it was a double-blind experiment, and if dropout occurred.



Linearity: The Residuals vs Fitted plot shows that the residuals have two linear trends–first decreasing, then increasing.

Constant Variance: The Scale-Location plot shows non-constant variance with more variance at greater values.



Normality or Large Sample Size: The residuals are pretty-clearly non-normal, which a major skew to the left. With only 40 data points, reliance on the CLT is questionable.

Overall Assessment: I do not believe this test is valid given how poorly the model satisfies the assumptions (especially non-constant variance).

Question 2

Data set for Question 2: 'Sales.csv'

Variables:

LAST_SALE_PRICE: the sales price of the home
SQFT: area of the house (sq. ft.)
LOT_SIZE: area of the lot (sq. ft.)
BEDS: number of bedrooms
BATHS: number of bathrooms

You are a real estate agent in Seattle who is often asked by potential clients the question "If I added a bathroom to my house, how much would it increase the house's value." House price depends on many factors in addition to the number of bathrooms, so use linear regression to adjust for some of these other confounding variables (use only the variables listed above for your analyses). Your answer should be one that you can use for any client (i.e., for the "average" house). However, in addition you would like to have answers specifically for owners of small houses and owners of large houses (as reflected by the area of the house). Provide two different estimates of the value of an additional bathroom: one for houses of below-average size, and one for houses of above-average size. (Note: Ignore the fact that the data set is a few years old and treat it as if it is current.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	93641.61	17889.274	5.24	0.000
SQFT	355.37	7.842	45.32	0.000
LOT_SIZE	-2.82	0.974	-2.90	0.004
BEDS	-93104.80	6665.232	-13.97	0.000
BATHS	86936.19	8611.747	10.10	0.000

Interpretation: All else held constant, the average home increases in value by $86,936 for every bathroom added.

Trifurcating into Small, Medium, and Large Houses: In order to split the dataset into a small/medium/large bucketing, the data will be divided such that an equal number of data points falls into each bucket (1565 houses per bucket).

Small Houses:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	313085.7	17002.18	18.41	0.000
SQFT	173.5	13.82	12.56	0.000
LOT_SIZE	-12.7	1.03	-12.29	0.000
BEDS	-16157.0	5750.12	-2.81	0.005
BATHS	37456.7	7033.27	5.33	0.000

For small houses (max size of 1720sqft), every additional bath increases the house price by $37,457.

Large Houses

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-118565.40	61073.04	-1.94	0.052
SQFT	411.27	17.34	23.71	0.000
LOT_SIZE	-2.18	1.82	-1.19	0.233
BEDS	-120218.09	15448.73	-7.78	0.000

	Estimate	Std. Error	t value	Pr(>|t|)
BATHS	131568.51	19107.56	6.89	0.000

For large houses (min size of 2420sqft), every additional bathroom adds $131,569 to the value of the house.