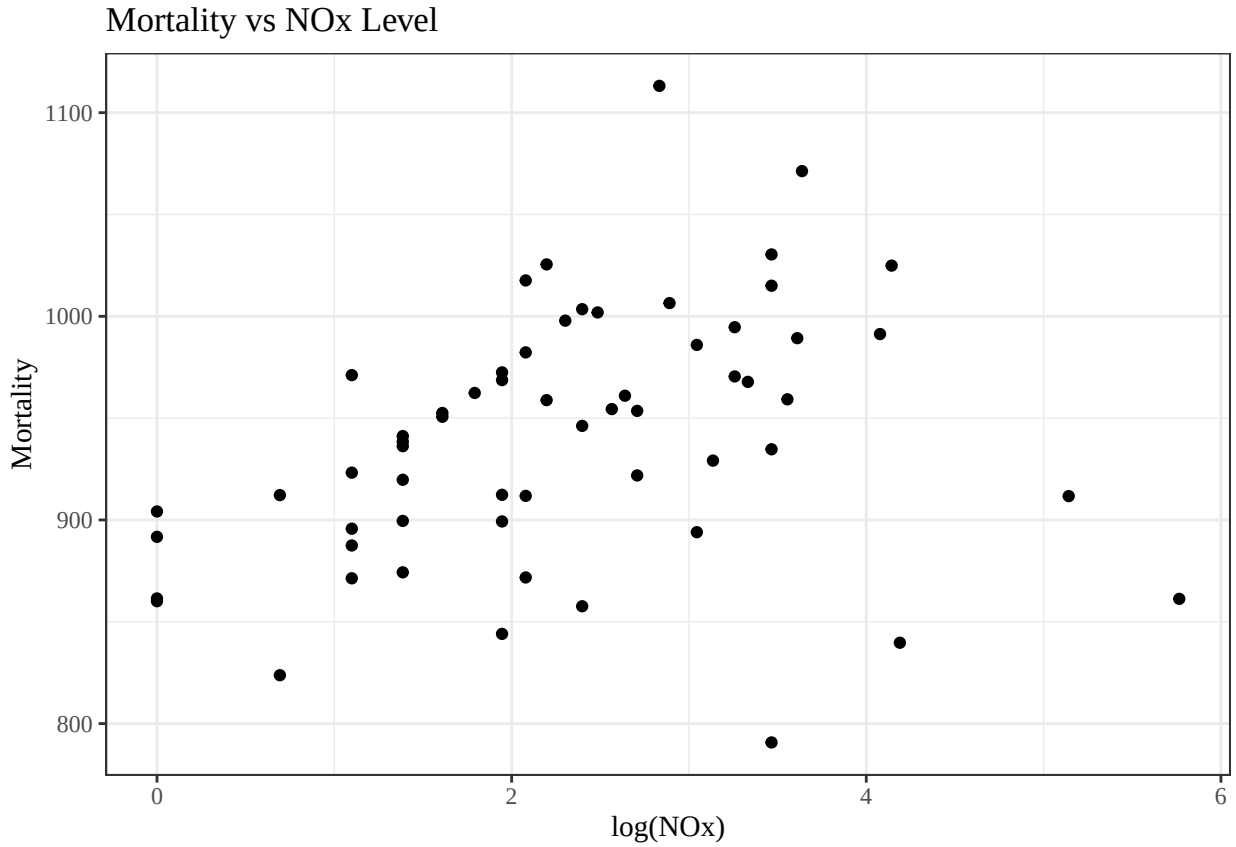
Homework 04

Spencer Pease

2/5/2020

(Q1)



The scatter of *mortality* vs $log(NOx)$ shows at least a small positive first-order trend. There are four locations however that shun this trend and maintain lower mortality for comparatively high $log(NOx)$ levels.

(Q2)

Regression model:

$$E[mortality|log(NOx)] = \beta_0 + \beta_1 \times log(NOx)$$

Table 1: Inferential statistics of Mortality vs log(NOx)

Parameter	Estimate	Robust SE	95%L	95%H	t value	Pr(>|t|)
log(NOx)	15.099	7.983	-0.882	31.08	1.891	0.064

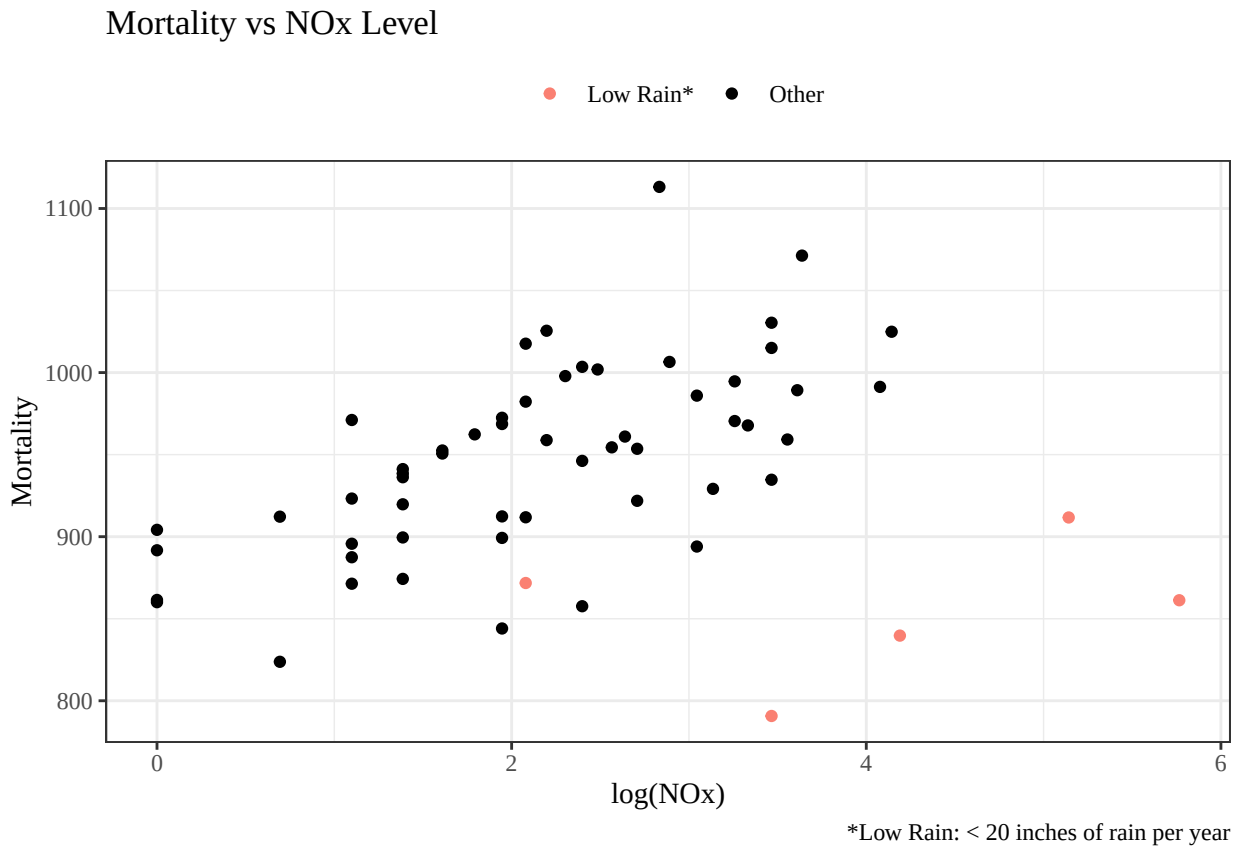
Using the above linear regression model with robust standard error estimates to assess the first-order relationship between mortality and $log(NOx)$, we estimate that an increase in one unit $log(NOx)$ is associated with a 15.1 unit increase in mortality (95% CI: -0.88, 31.08). A P-value greater than .05 suggests this relationship is not statistically significant.

(Q3)

Using robust standard error estimates avoids the assumption of equal variance within each group of our data (homoscedasticity), meaning the model makes fewer assumptions about the data in general.

This model is still based on the assumptions that the data are independent and the sample size is sufficiently large for the central limit theorem to apply when testing associations. The point estimate and confidence interval also rely on the assumption that the model is a good fit to the true distribution.

(Q4)



Rainfall should be an effect modifier, since it appears to have an association with both NOx pollution level and mortality. It doesn't make sense for pollution to determine rainfall, so rain is not in the casual pathway of interest.

(Q5)

Regression model:

$$E[mortality|log(NOx), rain] = \beta_0 + \beta_1 \times log(NOx) + \beta_2 \times rain + \beta_3 \times log(NOx) \cdot rain$$

Table 2: Inferential statistics of Mortality vs log(NOx), adjusted for Rain

Parameter	Estimate	Robust SE	95%L	95%H	t value	Pr(>|t|)
log(NOx)	-22.699	7.953	-38.631	-6.768	-2.854	6.04e-03
Rain	-0.479	0.844	-2.169	1.212	-0.567	5.73e-01
Log(NOx):Rain	1.463	0.223	1.016	1.909	6.561	1.83e-08

(Q6)

Fitting the model displayed in *Q5* using robust standard error estimates, we produce the estimated coefficients in the above table. From these values we estimate that for a location with no rainfall, one unit difference in $log(NOx)$ will change mortality by -22.7 units (95% CI: -38.63, -6.77). Between two groups with a one inch per year difference in rainfall and equal $log(NOx)$, mortality will in change by -0.48 units (95% CI: -2.17, 1.21). However, when comparing two groups that differ by one unit in both rainfall per year and $log(NOx)$, the resulting change in difference in mortality is -21.72 units, Which is less negative than just the sum of the $log(NOx)$ and rain coefficients. Of all of these estimates, rainfall on its own is not statistically significant. Overall, the relationship between mortality and $log(NOx)$ tends to be negative unless there is a significant difference in amount of rainfall (approximately 16 inches per year) between two groups.

(Q7)

In Aridia, less rain means that the negative impact of NOx level on mortality is diminished. This means reducing pollution might not be as effective at lowering mortality as other locations with more rain.

In Seattle, large amounts of rain increase the impact of NOx on mortality, so lowering pollution will have a larger impact on mortality.

(Q8)

(Q8.a)

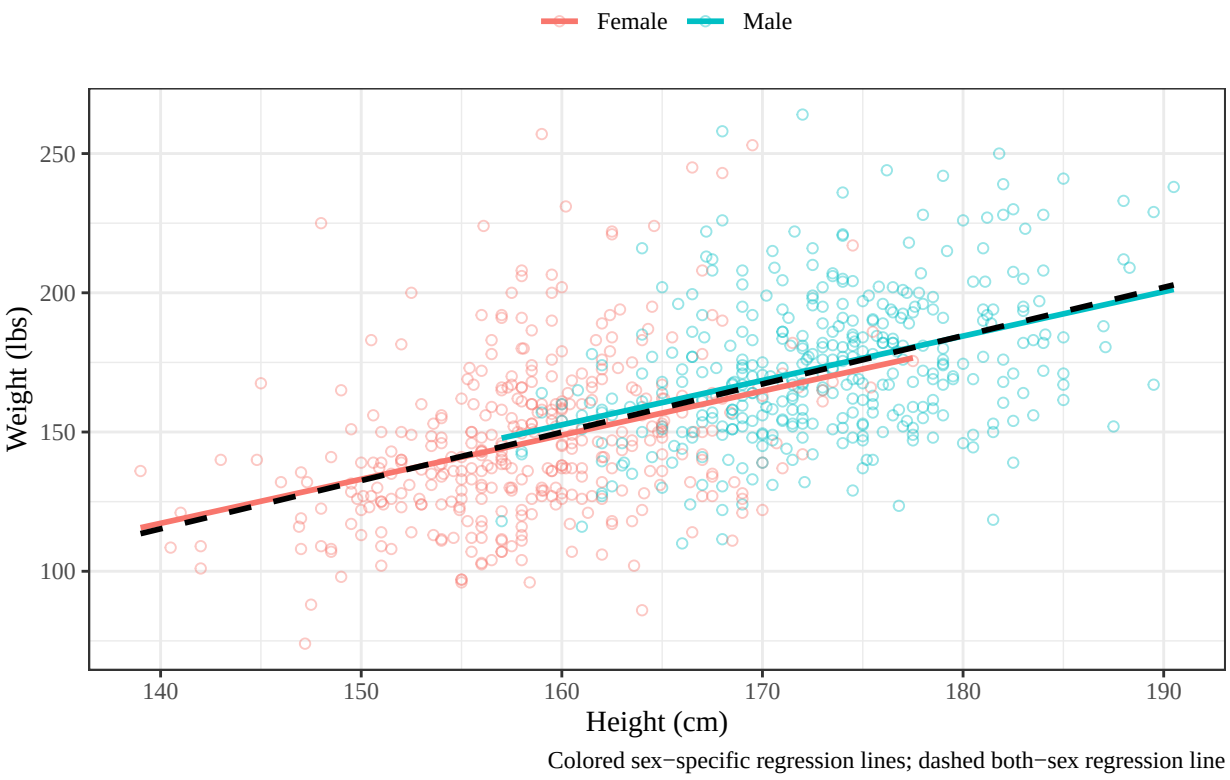


Table 3: Inferential statistics of Weight vs Height for separate models containing both sexes and individual sex data subsets

Data	Parameter	Estimate	Robust SE	95%L	95%H	t value	Pr(>|t|)
both	Height	1.734	0.096	1.545	1.923	18.014	2.43e-60
male	Height	1.594	0.202	1.196	1.991	7.878	3.89e-14
female	Height	1.584	0.216	1.158	2.009	7.321	1.57e-12

(Q8.b)

Scientifically, sex is an important variable to consider when predicting weight from height. It is well documented that a person's sex plays a role in their expected height and weight. Regardless if sex has a statistically significant role in our sample, our model should reflect what we expect in the true population. To exclude sex from our model would be to ignore a documented phenomena in our population.

(Q8.c)

The estimated slopes from the sex-specific models are nearly identical, suggesting that the relationship between weight and height is similar for both males and females.

Table 4: Summary of distribution of heights by sex

male	variable	n	valid	missing	mean	sd	min	q25	median	q75	max
0	height	369	369	0	158.515	6.354	139	155.0	158.1	162.4	177.5
1	height	366	366	0	173.099	6.467	157	168.7	173.0	177.3	190.5

From the data in our sample, we see that males have a greater height than females on average. This means there is likely an interaction between height and sex.

(Q8.d)

Altogether, sex plays the role of a confounder.

(Q8.e)

The estimated slope of the both-sex regression is slightly greater than both of the estimated sex-specific regression slopes. Since the both-sex regression includes both the larger heights and weights of the males and the lower heights and weights of the females, the model β_1 will represent this more extreme range of data as a greater estimated slope.

(Q8.f)

Table 5: Inferential statistics of Weight vs Height, adjusted for Sex

Parameter	Estimate	Robust SE	95%L	95%H	t value	Pr(>|t|)
height	1.589	0.148	1.298	1.879	10.744	4.22e-25
male	3.757	2.911	-1.957	9.471	1.291	1.97e-01

The estimated slope of the multiple regression model nearly matches the estimated slopes of the sex specific models. This is what we would expect, since adjusting for sex allows the model to capture separately the range of height and weight values associated with each sex while maintaining the same relationship between the two variables across sexes.

(Q9)

The denominator of (4.4) includes the term $(1 - r^2_{X,W})$, which will increase in magnitude (and thereby decrease $var(\hat{\gamma}_1)$) as the correlation between X and W decreases. In this scenario X is the treatment and W is diabetes, which are uncorrelated with each other. This minimizes $r^2_{X,W}$, which decreases the variance of our estimated coefficient overall, creating a more efficient model for estimation.

Checking for an interaction between the treatment and diabetes in this analysis is a good idea, because if there is an interaction between the two then it would change the scientific question we should be asking since we would know more about the behavior of our population.