Homework 03

Spencer Pease

1/29/2020

(Q1) Trends in Dichotomized Age

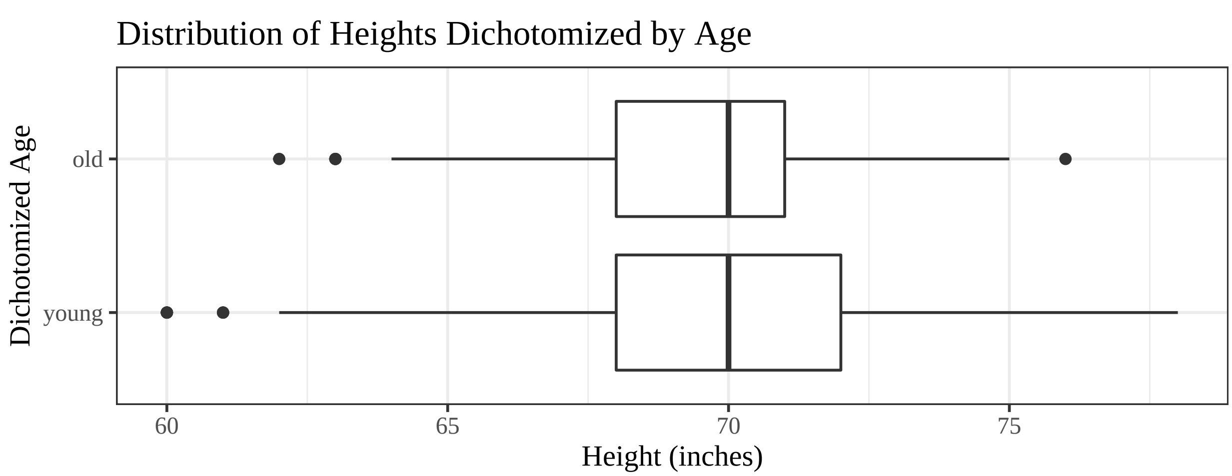
In this question, we classify observations with ages below 55 years as *young* and ages 55 years or greater as *old*. In regression models we will fit the binary variable *age* with $young = 0$ and $old = 1$.

(Q1.a)

We first look at the distribution of heights between age groups:

Table 1: Summary statistics for height by dichotomized age

age	variable	n	valid	missing	mean	sd	min	q25	median	q75	max
young	height	2832	2832	0	69.827	2.535	60	68	70	72	78
old	height	322	322	0	69.345	2.431	62	68	70	71	76



From this we can predict that a simple linear regression will show a change in age group from young to old is associated a slight decrease in height. Given that there are more observations in the *young* age group, which has a larger variance, the model-based standard error estimates will be conservative (larger). The difference between the variances is small enough that the conservative nature of the model will be slight.

Table 2: Inference table for height by dichotomized age

Estimate	Naive SE	Robust SE
-0.482	0.148	0.143

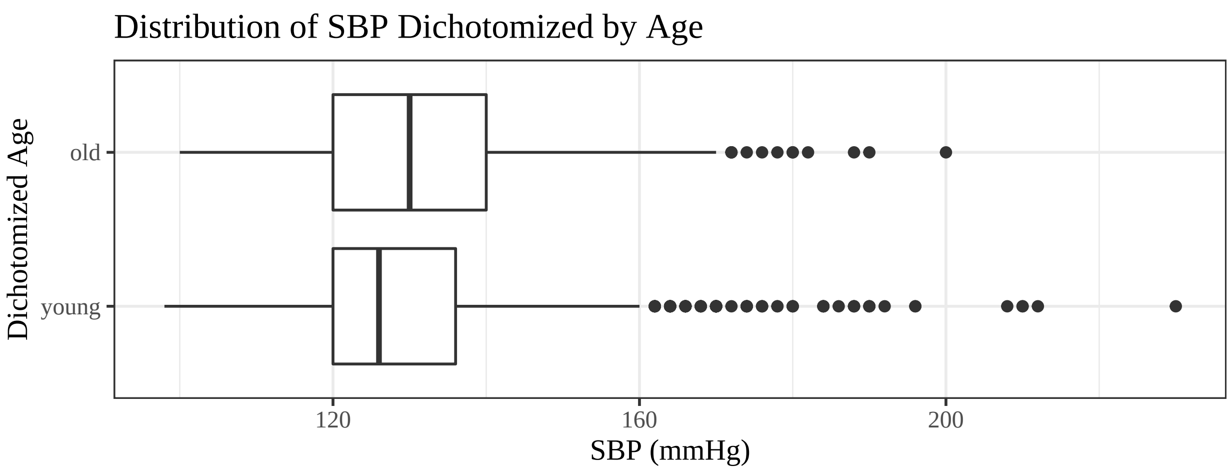
Fitting the model $E(height|age) = \beta_0 + \beta_1 \cdot age$ and comparing the model-based standard error estimates to the robust standard error estimates show that this is the case: The model-based SE is slightly conservative.

(Q1.b)

We can also look at the distribution of systolic blood pressure between age groups:

Table 3: Summary statistics for SBP by dichotomized age

age	variable	n	valid	missing	mean	sd	min	q25	median	q75	max
young	sbp	2832	2832	0	128.167	14.806	98	120	126	136	230
old	sbp	322	322	0	132.727	17.112	100	120	130	140	200



This distribution indicates that a simple linear regression will show a positive difference in SBP in the direction of increased age. Since the young age group has more observations, the variance of SBP in this group will influence the model-based standard error estimates. In this case the young group has smaller variance, so the model-based SE estimates will be anti-conservative (smaller).

Table 4: Inference table for SBP by dichotomized age

Estimate	Naive SE	Robust SE
4.559	0.886	0.992

Fitting the model $E(sbp|age) = \beta_0 + \beta_1 \cdot sbp$ and comparing the model-based standard error estimates to the robust standard error estimates confirms this: the model-based SE is anti-conservative.

(Q2) Education and Economic Benefit

(Q2.a)

Table 5: Inference table for the association between log(wage) and years of education

Estimate	Robust SE	95%L	95%H	t value	Pr(>|t|)
0.035	0.012	0.012	0.058	2.955	0.003

For this question, we fit a simple linear regression using robust standard error estimates with years of education as the predictor of interest and and log-transformed hourly wage as the response. Robust standard error estimates was chosen over classical model-based estimates because it relaxes the assumption that all levels of the predictor have the same variance in response. From the model we estimate an first-order trend with an intercept of 1.903 log-dollar hourly wage for zero years of education, and a 0.035 log-dollar difference in log-wage for every additional year of education.

While the 95% confidence interval and P-value (see above table) suggest that this association between $log(wage)$ and years of education is statistically significant, a real-world spot check of the model makes us think it unlikely that this trend is determined completely by years of education, especially at lower values. These results are likely confounded with age, and at some point there have to be diminishing returns on the value and additional year of education provides, meaning the first-order trend probably doesn't describe the entire relationship.

(Q2.b)

Our linear model does not have any power to suggest a causal relationship between years of education and $log(wage)$, so it is not an appropriate question to ask how much completing additional years of education will increase her wage.

(Q2.c)

Table 6: Model estimated mean log(wage) for 12 years of education

Estimate	2.5%	97.5%
2.324	2.241	2.408

(Q2.d)

Another way to construct an estimate confidence interval is with the formula:

$$\bar{X} \pm crit val(t_{n-1}) \times \frac{\hat{SD}(X)}{\sqrt{n}}$$

Using this formula, we estimate with a95% confidence interval that the mean $log(wage)$ for 12 years of education
is: 2.242 +/- 0.11

(Q2.e)

Table 7: Model predicted mean log(wage) for 12 years of education

Prediction	2.5%	97.5%
2.324	1.069	3.58

(Q2.f)

Another way to construct a 95% prediction interval is with the formula:

$$\bar{X} \pm crit val(t_{n-1}) \times \hat{SD}(X)\sqrt{1 + \frac{1}{n}}$$

Using this formula, we predict with 95% confidence that the $log(wage)$ for an individual with 12 years of education is: 2.242 +/- 1.099

(Q3)

Ignored