Problem Set 3

Spencer Papay - ssp2170

2/14/2017

Question 1

a) U_i represents all other variables that influence a student's score beyond the time they had, such as intelligence, preparedness, and test taking ability.

b) The experiment designed U_i and X_i as independent (since time doesn't affect other factors), so it follows that $E(U_i|X_i)$ are independent.

c) $\hat{Y}_i = 49 + 0.24X_i$

$X_i = 90$

$\hat{Y}_i = 49 + 0.24 \times 90$

$\hat{Y}_i = 70.6$

d) Estimated gain $= 0.24 \times 10 = 2.4$ points

e) The predicted score is $49 + 0.24 \times 240 = 106.6$. This prediction should not be taken seriously because it is extrapolating the data beyond the scope of the experiment's length and the covariate's bound. Just because an exam is longer does not mean all students score more than is possible.

Question 2

a)

```
m1 <- lm(ed~dist, data=college)
coef(m1)
```

```
## (Intercept)        dist
## 13.95585611 -0.07337271
```

The estimated intercept is about 13.96, which translates to 139.6 miles away, with an estimated slope of -0.073, meaning each additional 10 miles distance between high school and the nearest college reduces years of education by .073.

b) With a 20 mile distance, since distances are measured in 10s of miles, $X = 2$.

Years of completed education 20 miles away:

```
b0.hat + 2*b1.hat
```

```
## (Intercept)
##    13.80911
```

With a 20 mile distance, since distances are measured in 10s of miles, $X = 1$.

Years of completed education 10 miles away:

```
b0.hat + 1*b1.hat
```

```
## (Intercept)
##    13.88248
```

Bob's years of college completed increases by .073 if his high school was 10 miles closer.

c)

```r
summary(m1)
```

```
##
## Call:
## lm(formula = ed ~ dist, data = college)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9559 -1.8091 -0.6624  2.0515  4.4844
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.95586    0.03772 369.945   <2e-16 ***
## dist        -0.07337    0.01375  -5.336    1e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.807 on 3794 degrees of freedom
## Multiple R-squared:  0.00745,    Adjusted R-squared:  0.007188
## F-statistic: 28.48 on 1 and 3794 DF,  p-value: 1.004e-07
```

The reported value of $R^2 = 0.00745$, so only a miniscule fraction of variance in education completed is explained by distance from one's high school to the nearest college. This makes sense as many other factors are at play.

d)

```r
coeftest(m1, vcov=vcovHC(m1, type = "HC1"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 13.955856   0.037811 369.0934 < 2.2e-16 ***
## dist        -0.073373   0.013433  -5.4619 5.012e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The standard error of the slope is about 0.013, which yields a t-value of -5.46, giving an associated p-value of 5.012×10^{-8}. We can therefore reject the null hypothesis at all above-stated levels.

e)

```r
se.b1.hat <- coeftest(m1, vcov=vcovHC(m1, type = "HC1"))[2,2]
```

f)

```r
lb = b1.hat - se.b1.hat*1.96
ub = b1.hat + se.b1.hat*1.96
```

```r
lb
```

```
##        dist
## -0.09970223
```

```r
ub
```

```
##        dist
## -0.04704319
```

The confidence interval is (-0.0997, -0.0470)

g)

```
m2 = lm(ed~dist, data=college, subset=female==1)
b0.hat.women <- coef(m2)[1]
b1.hat.women <- coef(m2)[2]
```

```
b0.hat.women
```

```
## (Intercept)
##    13.93587
```

```
b1.hat.women
```

```
##        dist
## -0.06416757
```

```
coeftest(m2, vcov=vcovHC(m2, type = "HC1"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 13.935867   0.051324 271.5293 < 2.2e-16 ***
## dist        -0.064168   0.018444  -3.4791 0.0005135 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
se.b1.hat.women <- coeftest(m2, vcov=vcovHC(m2, type = "HC1"))[2,2]
lb_women = b1.hat.women - se.b1.hat.women*1.96
ub_women = b1.hat.women + se.b1.hat.women*1.96
```

```
lb_women
```

```
##      dist
## -0.100317
```

```
ub_women
```

```
##        dist
## -0.02801814
```

The confidence interval is (-0.1003, -0.0280)

h)

```
m3 = lm(ed~dist, data=college, subset=female==0)
b0.hat.men <- coef(m3)[1]
b1.hat.men <- coef(m3)[2]
```

```
b0.hat.men
```

```
## (Intercept)
##    13.97899
```

```
b1.hat.men
```

```
##        dist
## -0.08383705
```

```
coeftest(m3, vcov=vcovHC(m3, type = "HC1"))
```

```
##
```

```
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 13.978992   0.055920 249.9813 < 2.2e-16 ***
## dist        -0.083837   0.019573  -4.2833 1.943e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
se.b1.hat.men <- coeftest(m3, vcov=vcovHC(m3, type = "HC1"))[2,2]
lb_men = b1.hat.men - se.b1.hat.men*1.96
ub_men = b1.hat.men + se.b1.hat.men*1.96
```

```
lb_men
```

```
##       dist
## -0.1222004
```

```
ub_men
```

```
##        dist
## -0.04547365
```

The confidence interval is (-0.1222, -0.0455)

i)

We can compute the t-value for this independent data using the formula below:

$t = \frac{\hat{B}_{1,women} - \hat{B}_{1,men}}{\sqrt{SE\left(\hat{B}_{1,women}\right)^2 + SE\left(\hat{B}_{1,men}\right)^2}}$

$t = \frac{0.01967}{0.02689} = 0.7314$

This t-value indicates that the null hypothesis cannot be rejected at any significance level above .6384, which is the p-value. Therefore, the distance to college does not affect men and women differently, and is therefore due to other or unexplained factors.