ECON4150 - Introductory Econometrics

Seminar 5

Stock and Watson Chapter 8 & 9

Empirical exercise E8.3: Data

- In this exercise we use the data set *CollegeDistance.dta*
- These data are taken from the High School and Beyond survey conducted by the U.S. Department of Education in 1980, with a follow-up in 1986.
- The survey included students from approximately 1100 high schools.
- We are going to investigate the effect of distance to college on years of completed education.

Empirical exercise E8.3: Data

Series in Data Set

Name	Description
ed	Years of Education Completed (See below)
female	1 = Female/0 = Male
black	1 = Black/0 = Not-Black
Hispanic	1 = Hispanic/0 = Not-Hispanic
bytest	Base Year Composite Test Score. (These are achievement tests given to high school seniors in the sample)
dadcoll	1 = Father is a College Graduate/ 0 = Father is not a College Graduate
momcoll	1 = Mother is a College Graduate/ 0 = Mother is not a College Graduate
incomehi	1 = Family Income > $25,000 per year/ 0 = Income ≤ $25,000 per year.
ownhome	1= Family Owns Home / 0 = Family Does not Own Home
urban	1 = School in Urban Area / = School not in Urban Area
cue80	County Unemployment rate in 1980
stwmfg80	State Hourly Wage in Manufacturing in 1980
dist	Distance from 4yr College in 10's of miles
tuition	Avg. State 4yr College Tuition in $1000's

Empirical exercise E8.3: Data

```
. sum ed dist female bytest tuition black hispanic incomehi ownhome dadcoll momcoll cue80 stwmfg80
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ed	3796	13.82929	1.813969	12	18
dist	3796	1.724921	2.133836	0	16
female	3796	.5453109	.4980083	0	1
bytest	3796	51.00193	8.819251	28.95	71.36
tuition	3796	.9131396	.2835778	.43418	1.40416
black	3796	.1925711	.394371	0	1
hispanic	3796	.1498946	.3570151	0	1
incomehi	3796	.2863541	.4521164	0	1
ownhome	3796	.8192835	.3848338	0	1
dadcoll	3796	.2020548	.4015858	0	1
momcoll	3796	.1393572	.3463645	0	1
cue80	3796	7.654874	2.86577	1.4	24.9
stwmfg80	3796	9.556499	1.364411	6.59	12.15

Empirical exercise E8.3: question a)

```
. regress ed dist female bytest tuition black hispanic incomehi ownhome dadcoll mom
> coll cue80 stwmfg80, robust

Linear regression                               Number of obs =      3796
                                                F( 12,  3783) =    168.48
                                                Prob > F      =    0.0000
                                                R-squared     =    0.2836
                                                Root MSE      =    1.5378
```

ed	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
dist	-.0366613	.0120749	-3.04	0.002	-.0603352	-.0129874
female	.1429742	.0502718	2.84	0.004	.0444118	.2415366
bytest	.0930377	.003014	30.87	0.000	.0871284	.0989469
tuition	-.1910519	.0985259	-1.94	0.053	-.3842209	.0021171
black	.3506095	.0674301	5.20	0.000	.2184066	.4828125
hispanic	.3617649	.0764184	4.73	0.000	.2119397	.5115902
incomehi	.3718305	.0622177	5.98	0.000	.2498471	.4938138
ownhome	.1385475	.0649795	2.13	0.033	.0111492	.2659459
dadcoll	.5709712	.0763028	7.48	0.000	.4213726	.7205698
momcoll	.3778102	.0834999	4.52	0.000	.214101	.5415193
cue80	.0286753	.0095229	3.01	0.003	.0100049	.0473458
stwmfg80	-.0425003	.0199355	-2.13	0.033	-.0815857	-.0034148
_cons	8.920823	.2434585	36.64	0.000	8.4435	9.398145

Empirical exercise E8.3: question a)

$$\widehat{\beta}_{dist} = -0.037$$

- We have estimated a linear regression model, so the effect on Y of a unit change in X is constant and equals β.
- If Dist increases from 2 to 3, education is predicted to decrease by 0.037 years.
- If Dist increases from 6 to 7, education is predicted to decrease by 0.037 years

Empirical exercise E8.3: question b)

```
1 . gen ln_ed=ln(ed)

2 . regress ln_ed dist female bytest tuition black hispanic incomehi ownhome dadcoll
  > momcoll cue80 stwmfg80, robust

Linear regression                               Number of obs =      3796
                                                F( 12,  3783) =    173.89
                                                Prob > F      =    0.0000
                                                R-squared     =    0.2853
                                                Root MSE      =    .10918
```

ln_ed	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
dist	-.0026072	.0008651	-3.01	0.003	-.0043032	-.0009111
female	.0103059	.0035664	2.89	0.004	.0033137	.0172981
bytest	.0066561	.0002133	31.21	0.000	.0062379	.0070742
tuition	-.0139382	.0070081	-1.99	0.047	-.0276783	-.0001982
black	.0261676	.0048091	5.44	0.000	.0167389	.0355963
hispanic	.0259986	.0054098	4.81	0.000	.0153922	.0366049
incomehi	.0265197	.0044	6.03	0.000	.0178931	.0351463
ownhome	.0098332	.0046395	2.12	0.034	.000737	.0189295
dadcoll	.0405374	.0053518	7.57	0.000	.0300446	.0510302
momcoll	.0266016	.0058414	4.55	0.000	.0151491	.0380541
cue80	.0020357	.0006768	3.01	0.003	.0007088	.0033626
stwmfg80	-.0028642	.0014142	-2.03	0.043	-.0056368	-.0000916
_cons	2.265819	.0172772	131.15	0.000	2.231946	2.299693

Empirical exercise E8.3: question b)

$$\widehat{\beta}_{dist} = -0.0026$$

- We have estimated a log-linear regression model.

$$ln(y) = a + b \cdot x$$

- Taking the derivative of both sides of the equation (using the chain rule) gives

$$\frac{1}{y} dy = b \cdot dx \quad \longrightarrow \quad 100 \cdot \frac{\triangle y}{y} \approx 100 \cdot b \cdot \triangle x$$

- **Interpretation of β:** A change in X by one unit is associated with a $100 \cdot \beta$ percent change in Y

- If Dist increases from 2 to 3 education is predicted to decrease by 0.26%.

- If Dist increases from 6 to 7 education is predicted to decrease by 0.26%.

- These values, in percentage terms, are the same because the regression is a linear function relating ln(ED) and Dist.

Empirical exercise E8.3: question c)

```
1 . gen dist2=dist^2

2 . regress ed dist dist2 female bytest tuition black hispanic incomehi ownhome dadco
  > ll momcoll cue80 stwmfg80, robust

Linear regression                               Number of obs =      3796
                                                F( 13,  3782) =    155.93
                                                Prob > F      =    0.0000
                                                R-squared     =    0.2844
                                                Root MSE      =    1.5372
```

ed	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
dist	-.0811732	.0251112	-3.23	0.001	-.1304061	-.0319403
dist2	.0046413	.0020542	2.26	0.024	.0006139	.0086687
female	.1433144	.0502511	2.85	0.004	.0447925	.2418363
bytest	.0926367	.0030243	30.63	0.000	.0867072	.0985661
tuition	-.1928193	.0985524	-1.96	0.050	-.3860403	.0004016
black	.3339309	.0683045	4.89	0.000	.2000136	.4678482
hispanic	.3333104	.0778789	4.28	0.000	.1806216	.4859991
incomehi	.3694975	.0623003	5.93	0.000	.2473521	.4916429
ownhome	.14327	.0648817	2.21	0.027	.0160636	.2704765
dadcoll	.5611581	.0765802	7.33	0.000	.4110157	.7113006
momcoll	.3777022	.0835025	4.52	0.000	.2139878	.5414166
cue80	.0259537	.009587	2.71	0.007	.0071574	.0447499
stwmfg80	-.0425539	.0199267	-2.14	0.033	-.081622	-.0034858
_cons	9.012167	.2498793	36.07	0.000	8.522256	9.502078

Empirical exercise E8.3: question c)

- When Dist increases from 2 to 3, the predicted change in ED is:

$$\triangle \widehat{ED} = \left(-0.081 \times 3 + 0.0046 \times 3^2\right) - \left(-0.081 \times 2 + 0.0046 \times 2^2\right) = -0.058$$

- This means that the number of years of completed education is predicted to decrease by 0.058 years.

- When Dist increases from 6 to 7, the predicted change in ED is:

$$\triangle \widehat{ED} = \left(-0.081 \times 7 + 0.0046 \times 7^2\right) - \left(-0.081 \times 6 + 0.0046 \times 6^2\right) = -0.021$$

- This means that the number of years of completed education is predicted to decrease by 0.021 years.

Empirical exercise E8.3: question d)

- The regression in (c) adds the variable Dist2 to the regression (a).
- If the coefficient on Dist2 is statistically significant from zero this suggests that the addition of Dist2 is important.

$$H_0 : \beta_{dist^2} = 0 \qquad H_1 : \beta_{dist^2} \neq 0$$

- The t-statistic is shown in the Stata output

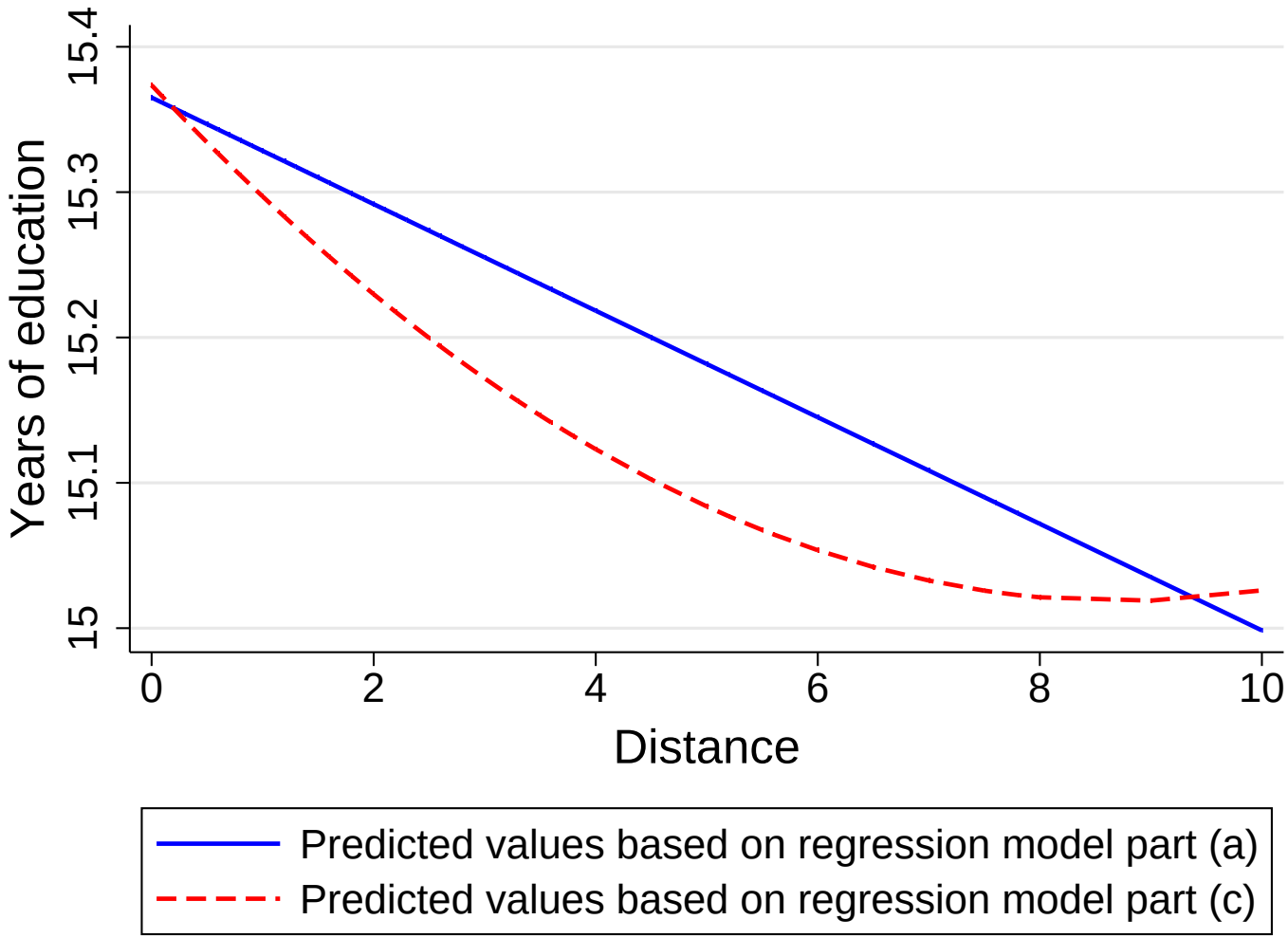
$$t = \frac{\beta_{dist^2} - 0}{SE(\beta_{dist^2})} = 2.26$$

- $t = 2.26 > 1.96$, so the coefficient on Dist2 is significantly different from zero at a 5% significance level.
- This indicates that the regression model in (c) is better than the regression model in (a).

Empirical exercise E8.3: question e)

```
 1 . qui regress ed dist female bytest tuition black hispanic incomehi ownhome dadcoll ///
   >  momcoll cue80 stwmfg80, robust

 2 . matrix b=e(b)

 3 . matrix list b

   b[1,13]
             dist      female      bytest     tuition       black    hispanic    incomehi
   y1  -.03666128   .14297422   .09303769   -.1910519   .35060952   .36176494   .37183046

            ownhome     dadcoll     momcoll       cue80    stwmfg80       _cons
   y1    .13854754   .57097117   .37781017   .02867534  -.04250025   8.9208225

 4 . gen y_predict_a=b[1,1]*dist+b[1,2]*1+b[1,3]*58+b[1,4]*0.95+b[1,5]*0+b[1,6]*1+ ///
   >  b[1,7]*0+b[1,8]*0+b[1,9]*1+b[1,10]*1+b[1,11]*7.1+b[1,12]*10.06+b[1,13]

 5 .
 6 .
 7 . qui regress ed dist dist2 female bytest tuition black hispanic incomehi ownhome ///
   >  dadcoll momcoll cue80 stwmfg80, robust

 8 . matrix b=e(b)

 9 . matrix list b

   b[1,14]
             dist       dist2      female      bytest     tuition       black    hispanic
   y1  -.08117324    .0046413   .14331438   .09263667  -.19281935   .33393089   .33331038

           incomehi     ownhome     dadcoll     momcoll       cue80    stwmfg80       _cons
   y1    .36949749   .14327003   .56115815   .37770219   .02595365  -.04255391    9.012167

10 . gen y_predict_c=b[1,1]*dist+b[1,2]*dist2+b[1,3]*1+b[1,4]*58+b[1,5]*0.95+b[1,6]*0+ ///
   >  b[1,7]*1+b[1,8]*0+b[1,9]*0+b[1,10]*1+b[1,11]*1+b[1,12]*7.1+b[1,13]*10.06+b[1,14]
```

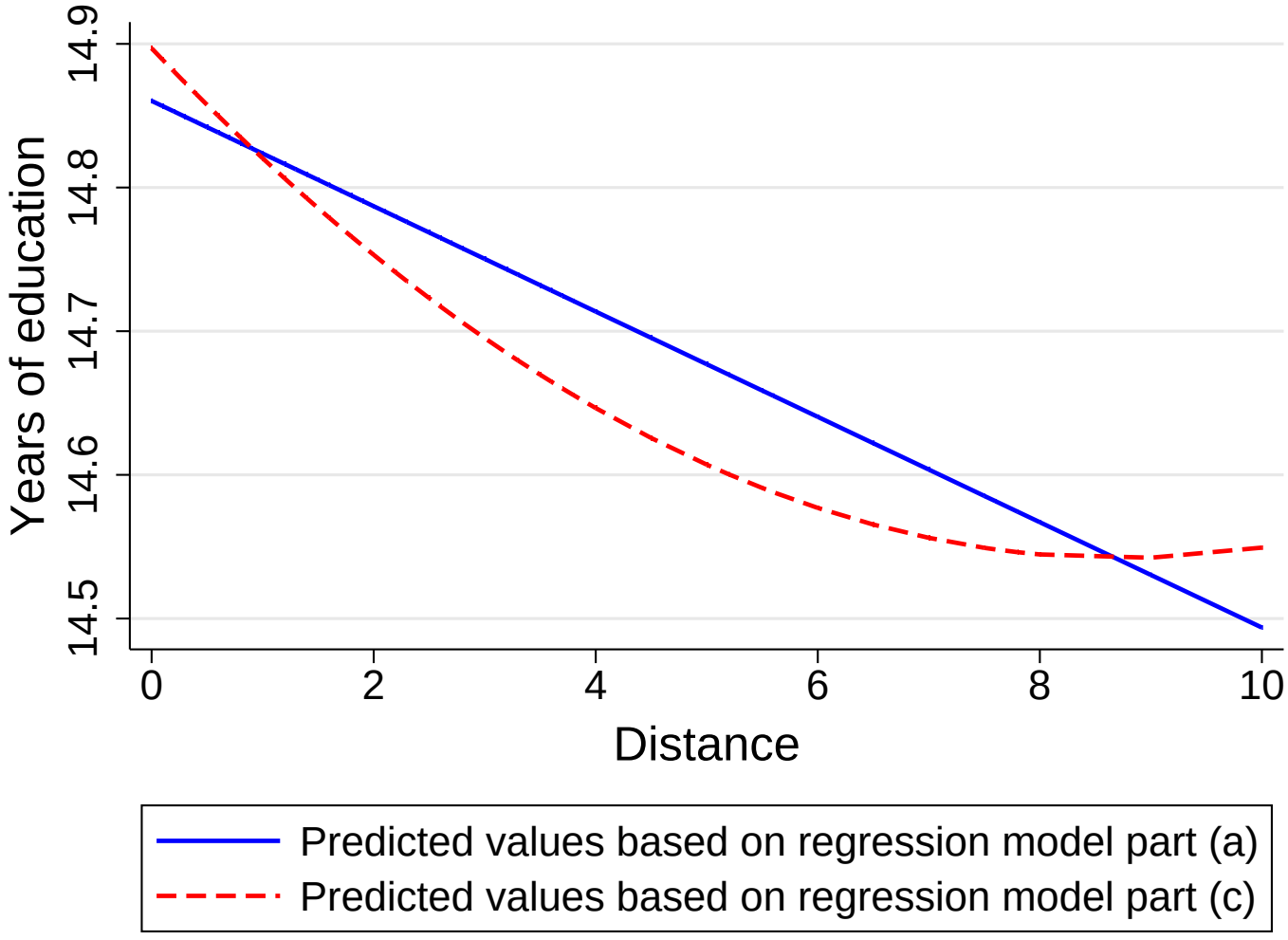
Empirical exercise E8.3: question e)



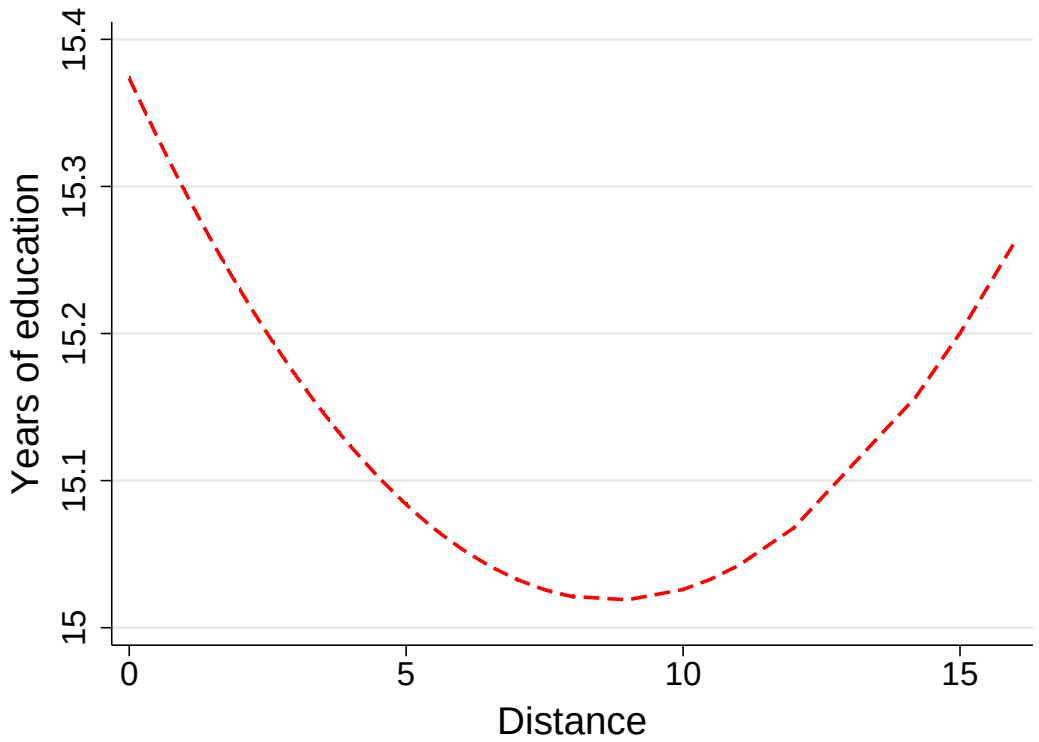
- The quadratic regression in (c) is steeper for small values of Dist than for larger values.
- The quadratic function is essentially flat when Dist=10.

Empirical exercise E8.3: question e)

- The only change in the regression functions for a white male is that the intercept would shift.
- The functions have the same slopes.



Empirical exercise E8.3: question e)



- The regression function becomes positively sloped for Dist > 10.
- There are only 44 of the 3796 observations with Dist > 10. This is approximately 1% of the sample.
- Thus, this part of the regression function is very imprecisely estimated.

Empirical exercise E8.3: question f)

```
1 . gen interaction=dadcoll*momcoll

2 . regress ed dist dist2 female bytest tuition black hispanic incomehi ownhome dadcoll
>  momcoll cue80 stwmfg80 interaction, robust

Linear regression                               Number of obs =      3796
                                                F( 14,  3781) =    145.73
                                                Prob > F      =    0.0000
                                                R-squared     =    0.2854
                                                Root MSE      =    1.5363
```

ed	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
dist	-.0810001	.025094	-3.23	0.001	-.1301992	-.0318011
dist2	.0046773	.0020564	2.27	0.023	.0006455	.0087091
female	.1406184	.0502133	2.80	0.005	.0421707	.2390661
bytest	.0925664	.0030234	30.62	0.000	.0866388	.0984939
tuition	-.1939714	.0985584	-1.97	0.049	-.3872042	-.0007387
black	.3305619	.0683148	4.84	0.000	.1966244	.4644994
hispanic	.3297465	.0779131	4.23	0.000	.1769907	.4825024
incomehi	.3623156	.0622537	5.82	0.000	.2402615	.4843697
ownhome	.1412131	.0649487	2.17	0.030	.0138752	.2685511
dadcoll	.6538031	.087084	7.51	0.000	.483067	.8245392
momcoll	.5693549	.1218052	4.67	0.000	.3305445	.8081652
cue80	.0257697	.00959	2.69	0.007	.0069677	.0445716
stwmfg80	-.0415432	.0199035	-2.09	0.037	-.0805658	-.0025206
interaction	-.3664802	.1639813	-2.23	0.025	-.6879805	-.0449799
_cons	9.00197	.2500197	36.01	0.000	8.511783	9.492157

Empirical exercise E8.3: question f)

- The estimated coefficient is $\widehat{\beta}_{interaction} = -0.366$.
- There are different ways to interpret the interaction term

1. The effect of having a mother with a college degree on years of education is smaller when the father has a college degree.

$$\widehat{ED}(momcoll = 1) - \widehat{ED}(momcoll = 0) = 0.569 - 0.366 \cdot dadcoll$$

2. The effect of having a father with a college degree on years of education is smaller when the mother has a college degree.

$$\widehat{ED}(dadcoll = 1) - \widehat{ED}(dadcoll = 0) = 0.654 - 0.366 \cdot momcoll$$

Empirical exercise E8.3: question g)

i) predicted difference between Jane and Mary's years of education

$$\widehat{ED}(Jane) - \widehat{ED}(Mary) =$$
$$(0.654 \times 1 + 0.569 \times 0 - 0.366 \times 0) - (0.654 \times 0 + 0.569 \times 0 - 0.366 \times 0) = 0.654$$

ii) predicted difference between Alexis and Mary's years of education

$$\widehat{ED}(Alexis) - \widehat{ED}(Mary) =$$
$$(0.654 \times 0 + 0.569 \times 1 - 0.366 \times 0) - (0.654 \times 0 + 0.569 \times 0 - 0.366 \times 0) = 0.569$$

iii) predicted difference between Bonnie and Mary's years of education

$$\widehat{ED}(Bonnie) - \widehat{ED}(Mary) =$$
$$(0.654 \times 1 + 0.569 \times 1 - 0.366 \times 1) - (0.654 \times 0 + 0.569 \times 0 - 0.366 \times 0) = 0.857$$

Empirical exercise E8.3: question h)

We add two interaction terms to the regression:

1. Dist x Incomehi
2. Dist^{2}x Incomehi

With Incomehi:

- equal to 1 if Family Income > \$25,000 per year
- equal to 0 if Income $\leq$ \$25,000 per year.

Empirical exercise E8.3: question h)

```
1 . regress ed dist dist2 female bytest tuition black hispanic incomehi ownhome dadcoll
>  momcoll cue80 stwmfg80 dist_incomehi dist2_incomehi, robust

Linear regression                               Number of obs =       3796
                                                F( 15,  3780) =     136.54
                                                Prob > F      =     0.0000
                                                R-squared     =     0.2853
                                                Root MSE      =     1.5365
```

ed	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
dist	-.1106924	.0281294	-3.94	0.000	-.1658427	-.0555422
dist2	.0065296	.0022158	2.95	0.003	.0021853	.010874
female	.1440742	.050229	2.87	0.004	.0455957	.2425527
bytest	.0928315	.0030209	30.73	0.000	.0869087	.0987542
tuition	-.2093653	.099146	-2.11	0.035	-.4037502	-.0149803
black	.3365495	.0684223	4.92	0.000	.2024013	.4706978
hispanic	.3262323	.0777223	4.20	0.000	.1738505	.478614
incomehi	.2188902	.0898749	2.44	0.015	.0426822	.3950982
ownhome	.1458133	.06492	2.25	0.025	.0185316	.2730949
dadcoll	.5731094	.0766517	7.48	0.000	.4228267	.723392
momcoll	.381609	.0835843	4.57	0.000	.2177344	.5454837
cue80	.0262236	.0095864	2.74	0.006	.0074285	.0450187
stwmfg80	-.0429091	.0199047	-2.16	0.031	-.081934	-.0038842
dist_incomehi	.1305136	.0620471	2.10	0.035	.0088646	.2521626
dist2_incomehi	-.0093797	.0062451	-1.50	0.133	-.0216238	.0028643
_cons	9.053221	.2506604	36.12	0.000	8.561778	9.544663

Empirical exercise E8.3: question h)

- To answer the question whether the effect of dist on ED differs significantly between individuals from high or low income families we can perform an F-test

```
 test dist_incomehi=dist2_incomehi=0

( 1)  dist_incomehi - dist2_incomehi = 0
( 2)  dist_incomehi = 0

       F(  2,  3780) =      2.49
            Prob > F =      0.0834
```

- In this regression model we cannot reject the null hypothesis that the effect of dist on ED is the same for individuals from high or low income families.
  - at a 1% or 5% significance level
  - but we reject the null hypothesis at the 10% level

Empirical exercise E8.3: question i)

Linear regression

Number of obs = 3796
F(13, 3782) = 155.93
Prob > F = 0.0000
R-squared = 0.2844
Root MSE = 1.5372

ed	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
dist	-.0811732	.0251112	-3.23	0.001	-.1304061	-.0319403
dist2	.0046413	.0020542	2.26	0.024	.0006139	.0086687
female	.1433144	.0502511	2.85	0.004	.0447925	.2418363
bytest	.0926367	.0030243	30.63	0.000	.0867072	.0985661
tuition	-.1928193	.0985524	-1.96	0.050	-.3860403	.0004016
black	.3339309	.0683045	4.89	0.000	.2000136	.4678482
hispanic	.3333104	.0778789	4.28	0.000	.1806216	.4859991
incomehi	.3694975	.0623003	5.93	0.000	.2473521	.4916429
ownhome	.14327	.0648817	2.21	0.027	.0160636	.2704765
dadcoll	.5611581	.0765802	7.33	0.000	.4110157	.7113006
momcoll	.3777022	.0835025	4.52	0.000	.2139878	.5414166
cue80	.0259537	.009587	2.71	0.007	.0071574	.0447499
stwmfg80	-.0425539	.0199267	-2.14	0.033	-.081622	-.0034858
_cons	9.012167	.2498793	36.07	0.000	8.522256	9.502078

- The effect of distance to college on years of education is negative but the effect becomes less negative for larger initial distance.
- Can we interpret the results as giving a consistent estimate of the causal effect of distance on education?

Empirical exercise E9.3: question a)

- We take the regression of E 8.3 part c)
- Lets consider the following threats to internal validity
  - Omitted variables
  - Functional form misspecification
  - Measurement error
  - Sample selection
  - Simultaneous causality
  - Heteroskedasticity and/or correlated error terms

Empirical exercise E9.3: question a)

Omitted variables: This is potentially important.

- For example, family background characteristics of students living close to college might differ from those who live far from college. These background characteristics might affect years of education.

Misspecification of the function form: We investigated this in E8.3. Difficult to say what is the correct functional form, but quadratic seems better than linear model.

Errors-in-variables: This is potentially important.Since the data are from a survey there might be measurement error both in the dependent variable as in the independent variables.

Empirical exercise E9.3: question a)

Sample Selection: This is a random sample of high school seniors, so sample selection within this population is unlikely to be a problem.

Simultaneous causality: The argument here would be that parents who want to send their children to college may locate closer to a college (maybe more an omitted variable bias problem).

Inconsistency of standard errors: Heteroskedasticity-robust standard errors were used.

- The data represent a random sample so that correlation across the error terms is unlikely to be a problem.
- If there are multiple individuals from the same high school it would be good to cluster the se's at the level of the high school.

Empirical exercise E9.3: question b)

Linear regression

Number of obs = 943
F(13, 929) = 27.02
Prob > F = 0.0000
R-squared = 0.2310
Root MSE = 1.4866

ed	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
dist	-.0906026	.0396222	-2.29	0.022	-.168362	-.0128432
dist2	.0040898	.0026249	1.56	0.120	-.0010615	.0092411
female	.0509253	.0990533	0.51	0.607	-.1434689	.2453195
bytest	.0733229	.006504	11.27	0.000	.0605588	.0860871
tuition	-.5234936	.2425345	-2.16	0.031	-.9994725	-.0475147
black	.058945	.1802014	0.33	0.744	-.2947041	.412594
hispanic	.1980937	.115382	1.72	0.086	-.0283459	.4245332
incomehi	.4132809	.1213661	3.41	0.001	.1750974	.6514644
ownhome	.1993312	.1265398	1.58	0.116	-.0490058	.4476682
dadcoll	.4690392	.1337894	3.51	0.000	.2064746	.7316037
momcoll	.3619455	.1633909	2.22	0.027	.0412875	.6826035
cue80	.0450695	.0226237	1.99	0.047	.00067	.089469
stwmfg80	.0315789	.0443372	0.71	0.476	-.0554338	.1185915
_cons	9.21347	.5201512	17.71	0.000	8.192662	10.23428

- Coefficients on Dist and Dist2 using only data from western states are very similar to the estimated coefficients from the data with the other states.
- Se's are larger in the CollegeDistanceWest data set because the number of observations is smaller.