8. Multiple Regression: Relationships Between Several Variables

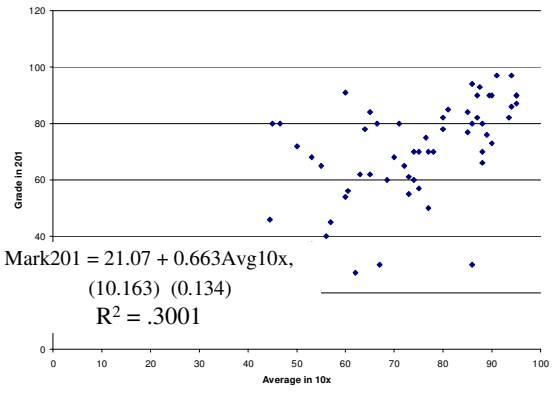
ASW (Chapter 13)

1

A) Introduction

- Multiple regression allows:
  - Full set of explanatory variables.
  - Ceteris paribus effects:
    - Impact of Δx_1 on y, holding constant impacts of x_2, x_3, etc.

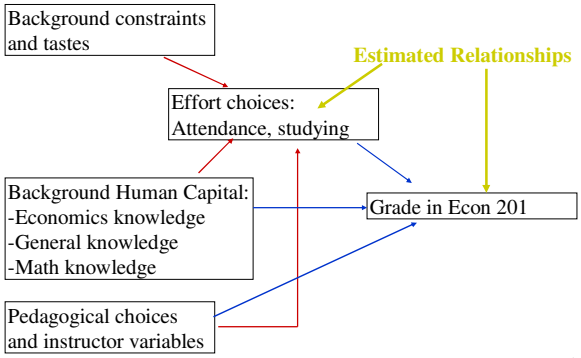
2



Mark201 = 21.07 + 0.663Avg10x,
(10.163) (0.134)
R^2 = .3001

Source: H. King's data.

Education Production Function Model



4

Mark201 as a Function of:

CONSTANT	-12.882 (11.92)
Effort Variables	
ATTEND (1 unit = 10% of classes)	5.434 (1.420)*
HRCOURS (studying at home)	very insignificant
Initial Economics Ability	
AVG10X	very insignificant
BOTH1012	9.605 (3.3645)*
TIMEGAP (since Econ 100)	1.6577 (0.644)**
Initial General Ability	
TIMEHS	-0.691 (.410)***
ENGL100 (number of ENGL courses)	-4.626 (2.023)**
UNIGPA	0.771 (.157)*
FEMALE	-7.505 (2.627)*

* (**, ***) indicates significance at the 1% (5%, 10%) level.

5

Mark201 as a Function of:

Initial Math Ability	
CALCHIGH	-5.638 (5.059)
MATH103	-8.351 (2.994)*
MATH105	-31.236 (3.995)*
MATH110	-10.08 (3.475)*
CALCAVG	0.119 (.061)***
Pedagogical Variables	
EVENSTYL (Study methodology)	8.488 (3.464)**
ASSTMISS	-11.894 (3.133)*
R^2 (Adjusted R^2)	0.7657 (0.6912)
F-Statistic	10.274*

* (**, ***) indicates significance at the 1% (5%, 10%) level.

6

Multiple Regression Steps

1. Economic theory → equation to estimate.
2. Put ALL possible, theoretically relevant variables in regression.
3. Find a good sample, do regression, and run test statistics on significance of coefficients.
4. If coefficients are strongly insignificant (|t-stat| < 1), throw out those variables.
5. Redo regression and you have a final "tested-down" version.

7

B) Multiple Regression Model & Least Squares

(1) $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_p x_p + \varepsilon$

- $E(\varepsilon) = 0$ on average, so:
  (2) $E(y) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_p x_p$
- Sample data –> estimated regression equation:
  (3) $\hat{y} = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_p x_p$
- Pick $b_0, b_1, \ldots b_p$ to minimize $\Sigma(e_i)^2 = \Sigma(y_i - \hat{y}_i)^2$.
  - Using matrix algebra!

8

Consumption and Multiple Regression

- We saw that consumption was correlated with real GDP, unemployment and the interest rate.
- We could run separate regressions.

9

Consumption (1985-2004)

Consumption = 26659 + 0.541RealGDP, R^2 = .9933
(4374) (0.005)

Consumption = 757647 – 30015URate, R^2 = .3556
(41123) (4574)

Consumption = 733944 – 21042IntRate, R^2 = .3920
(34821) (2967)

Consumption = 357996 + 3305TTrend, R^2 = .9610
(3513) (75.35)

10

Multiple R	0.997966
R Square	0.995936
Adj. R Square	0.99572
Standard Error	5126.059
Observations	80

	df	*SS*	*MS*	*F*	*Sig F*
Regression	4	4.83E+11	1.21E+11	4595.408	8.28E-89
Residual	75	1.97E+09	26276478		
Total	79	4.85E+11			

CONSUMPTION as a function of:

	Coeff's	*St. Error*	*t Stat*	*P-value*
Intercept	71671.94	21746.78	3.29575	0.001501
GDP	0.439409	0.025284	17.37873	5.14E-28
U Rate	739.6108	664.9002	1.112364	0.269535
Int Rate	592.8069	406.4753	1.458408	0.148903
Time trend	719.9698	148.0774	4.862117	6.23E-06

Consumption cont'd…

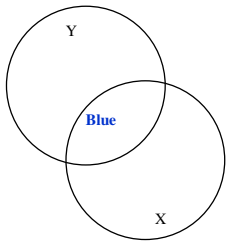
Consumption = 71672 + 0.439RealGDP + 739URate
(21747)* (0.025)* (665)

+ 593IntRate + 720TTrend, R^2 = 0.9959
(406) (148)*

with standard errors in brackets.
* – significant at the 1% level

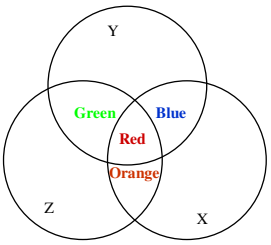
12

Multiple Regression and Venn Diagrams



13

Multiple Regression and Venn Diagrams



14

Ceteris Paribus

- Bottom line: regression analysis shows the impact of one explanatory variable, holding constant the other explanatory variables.

$$b_1 = \frac{\Delta Y}{\Delta X_1}\bigg|_{\Delta X_2 = \Delta X_3 = \ldots = \Delta X_n = 0} = \frac{\partial Y}{\partial X_1}$$

15

R^2 and R^2(adjusted)

- Recall: Total sum of squares
  = sum of squares due to regression
  + sum of squares due to error
- SST = SSR + SSE
- $R^2 = \frac{SSR}{SST}$ = proportion of variation in y explained by the regression equation.
- R^2 always increases when we add extra variables, even if they are useless.
- C = f(Y only) → R^2 = .9933.
- C = f(Y, URate, IntRate, TTrend) → R^2 = .9959

16

Adjusted R^2

- Why?
- More variables → SSE goes down → SSR up → R^2 up.
- The adjusted R^2 (R^2_a or $\bar{R}^2$) tries to avoid overestimation:

$$R^2_a = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where p = number of slope variables.

- R^2 must rise fast enough to compensate for p rising as you add variables.

17

Consumption Function

Multiple R	0.997966
R Square	0.995936
Adj. R Square	**0.99572**
Standard Error	5126.059
Observations	80

	df	*SS*	*MS*	*F*	*Sig F*
Regression	4	4.83E+11	1.21E+11	4595.408	8.28E-89
Residual	75	1.97E+09	26276478		
Total	79	4.85E+11			

	Coeff's	*St. Error*	*t Stat*	*P-value*
Intercept	71671.94	21746.78	3.29575	0.001501
GDP	0.439409	0.025284	17.37873	5.14E-28
U Rate	739.6108	664.9002	1.112364	0.269535
Int Rate	592.8069	406.4753	1.458408	0.148903
Time trend	719.9698	148.0774	4.862117	6.23E-06

C) Model Assumptions

- Reminder: –> OLS estimator works best only if certain conditions about the model are met.
  - Error terms are random, centered around 0, no pattern.

19

D) Testing for Significance

- Is there a relationship between x and y?
- If just one x-variable → t-test and F-test identical.
- If 2 x-variables → different purposes.
- F-test: *overall significance*.
- If F-test holds → use t-test to look at significance of *individual* x_i-variable.

20

The F-test

(1) $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_p x_p + \varepsilon$

- F-test tests significance of all slope variables.
  H_o: $\beta_o = \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_p = 0$.
  H_A: one of the β's is not equal to 0.
- F = Mean square due to regression / Mean square due to error = MSR / MSE

21

MSR and MSE

- MSR = sum of squares due to regression / regression degrees of freedom
  = sum of squares due to regression / no. of slope variables
  = SSR / p
- MSE = Sum of squares due to error / error degrees of freedom
  = SSE / n − p − 1

22

F-test

- Determine hypothesis (β's = 0), determine level of significance (α = 0.05), calculate F-statistic.
- F-statistic = MSR ← p degrees of freedom
  MSE ← n−p−1 degrees of freedom
- If H_o is true, MSR ≈ MSE → F-statistic ≈ 1.
- If H_o is false
  → F-statistic is higher than critical value from the table/Excel.
  → p-value is < level of significance.
  → at least one β ≠ 0.

23

Consumption Function

Multiple R	0.997966
R Square	0.995936
Adj. R Square	**0.99572**
Standard Error	5126.059
Observations	80

	df	*SS*	*MS*	*F*	*Sig F*
Regression	4	4.83E+11	1.21E+11	4595.408	8.28E-89
Residual	75	1.97E+09	26276478		
Total	79	4.85E+11			

Reject null hypothesis if probability value < α.
Probability value = .0000000000…0008
→ equation is highly significant.

Excel's ANOVA Table

ANOVA

	df	*SS*	*MS*	*F*	*Sig F*
Regression	p	SSR	MSR = SSR/p	F=MSR/MSE	p-value
Residual	n – p – 1	SSE	MSE = SSE/(n-p-1)		
Total	n – 1	SST			

25

t-tests

- Significance of one variable → use the t-test.

Consumption Function

	Coeff's	*St. Error*	*t Stat*	*P-value*
Intercept	71671.94	21746.78	3.29575	0.001501
GDP	0.439409	0.025284	17.37873	5.14E-28
U Rate	739.6108	664.9002	1.112364	0.269535
Int Rate	592.8069	406.4753	1.458408	0.148903
Time trend	719.9698	148.0774	4.862117	6.23E-06

Significant

Insignificant

26

Joint Coefficient Tests

- You can use special F-tests to **jointly** test if two or more variables are significant → see Econ 324.

27

Multicollinearity

- Our explanatory variables may be correlated with
  - Each other.
  - The residuals.
  - Other outside variables –> that is, they are really dependent variables too.
- Example: real GDP and unemployment in our consumption model.
- Multicollinearity = strong correlation among explanatory variables.

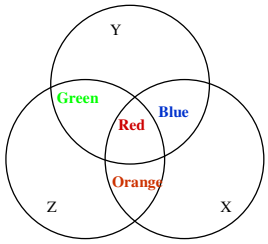
28

Multicollinearity cont'd

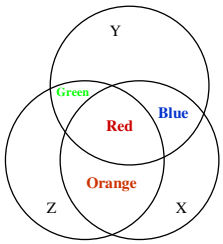
- Impact of multicollinearity:
  - Strong F-test, but negligible t-tests.
  - Strange values for b's and s_b's.

29

Multicollinearity and Venn Diagrams



Low Multicollinearity



Strong Multicollinearity

30

E) Qualitative Independent Variables

- Quantitative variables → on a scale.
- Dummy or qualitative variables → on or off.
  - Gender.
  - **Level** of schooling (did you get a degree?)
  - Recession year.
  - Unemployed or employed or out of labour force.

31

Simple Dummy Variables

- Only 2 choices.
- Create x-variables, coded = 1 if on, 0 if off.
- Gender example:

Wages	Female	Gender
22,000	No	0
25,000	Yes	1
37,000	No	0

- Example:
  Wages = $b_0 + b_1$POTEXP + b_2FEMALE
- If $b_2 \neq 0$ → female and male wages are different.

32

Interpreting Dummy Variables

- If male (so FEMALE = 0):
  $E(Wages|Male) = b_0 + b_1POTEXP + b_2(0)$
- If female (so FEMALE = 1)
  $E(Wages|Female) = b_0 + b_1POTEXP + b_2(1)$
- $E(Wages|Female) - E(Wages|Male) =$
  $b_0 + b_1POTEXP + b_2 - (b_0 + b_1POTEXP)$
  $= b_2$
- b_2 gives the wage differential due to being female.

33

Wages, Gender and Potential Experience

Wages as Dependent Variable

R Square	0.123983
Ad. R Square	0.122199
St. Error	26632.74
Observations	985

	df	*SS*	*MS*	*F*	*Sig. F*
Regression	2	9.86E+10	4.93E+10	69.492	5.94E-29
Residual	982	6.97E+11	7.09E+08		
Total	984	7.95E+11			

	Coefficients	*St. Error*	*t Stat*	*P-value*
Intercept	42827.65	1970.29	21.737	7.68E-86
Female	-14915	1707.774	-8.734	1.05E-17
POTEXP	592.1337	85.63828	6.914	8.45E-12

Data Source: Statistics Canada, *2001 Census PUMF*

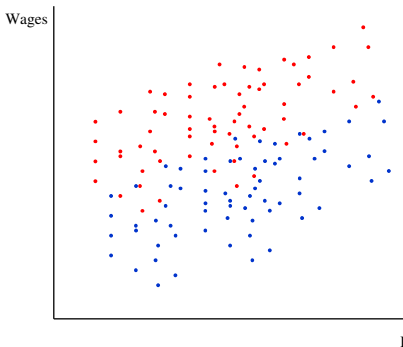
34

Wages, Gender, Potential Experience cont'd

- $E(Wages|Males) = 42827.65 + 592.14POTEXP$
  $= 42827.65 + 592.14*(17.02)$
  $= 52905.87$ on average.
- $E(Wages|Females) = 42827.65 + 592.14POTEXP$
  $- 14915(1)$
  $= 42827.65 - 14915 + 592.14*(17.02)$
  $= 27912.65 + 592.14*(17.02)$
  $= 37990.87$ on average.

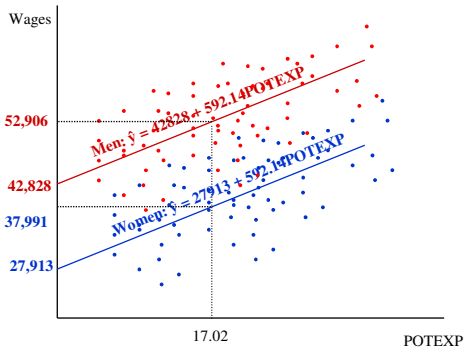
35

Female and Male Regression Lines



36

Female and Male Regression Lines



37

The Dummy Trap

- Warning: a dummy variable for women + a dummy for men AND an intercept term → regression will fail due to perfect multicollinearity.

38

Dummies with More than 2 Choices

- Education choices:
  - less than High School ← **Leave out as reference group**
  - only high school
  - technical school or community college
  - some university
  - Bachelor's degree
  - graduate degree
- Set up dummies for FIVE out of the six choices:

$Wages = b_0 + b_1POTEXP + b_2FEMALE + b_3HS + b_4TECH + b_5SOMEUNI + b_6BACH + b_7GRAD$

- b_5 compares wages of SOMEUNI to less than HS.

39

Interactive Dummy Variables

- We could also allow the slope variables to differ between men and women:

$Wages = b_0 + b_1POTEXP + b_2FEMALE + b_3FEMALE*POTEXP$

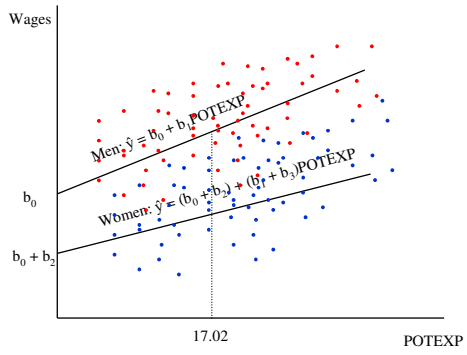
$E(Wages|Male) = b_0 + b_1POTEXP$

$E(Wages|Female) = (b_0 + b_2) + (b_1 + b_3)POTEXP$

- Allows men and women to have different impacts from POTEXP.

40

Female and Male Regression Lines Again



41

Logs and Dummies

- Suppose:
  Ln(Wages) = $b_0 + b_1$POTEXP + b_2FEMALE
- b_1 = %ΔWages due to 1 more unit of POTEXP.
- b_2 ≈ %ΔWages from being female.
  - Exact %ΔWages = $e^{b2} - 1$.
    = $2.71828^{b2} - 1$.

42

Ln Wages and Potential Experience

Multiple R	0.206157
R Square	0.042501
Adj R Square	0.04055
St. Error	0.858435
Observations	985

Women get paid about 24.2% less in wages

	df	*SS*	*MS*	*F*	*Sig F*
Regression	2	32.120	16.06	21.794	5.48E-10
Residual	982	723.65	0.737		
Total	984	755.77			

Each year of potential experience yields 1% more wages

	Coeff.	*St Err*	*t Stat*	*P-value*
Intercept	10.43658	0.063507	164.34	0
Female	-0.27747	0.055045	-5.04	5.52E-07
POTEXP	0.010174	0.00276	3.686	0.00024

43

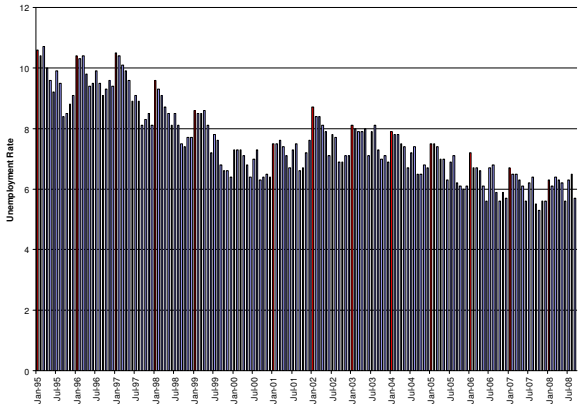
Seasonally Adjusted Variables

Example 1: Home Sales

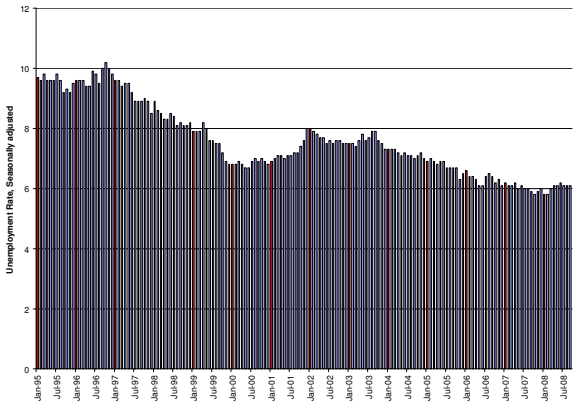
- The home sales handout shows seasonal impacts.
- To remove seasonal impacts, we use dummy variables for each season → leaves us with results that are not influenced by the seasons.

44

Example 2: Unemployment Rates



Example 2: Unemployment Rates, Seasonally Adjusted



F) Example: Explaining Wages

- A simplified version of A. Ferrer and W.C. Riddell, "The role of credentials in the Canadian labour market," *Can. Journal of Economics*, Nov. 2002.
- What is the impact of schooling years, credentials, and experience on Saskatchewan wages?
- General, theoretical hypotheses:
  - Human capital model: only schooling years matter.
  - Screening/signalling model: only credentials matter.

47

Mincerian Wage-generating Function

Ln(Weekly wages) = β_0 + β_1*(FEMALE)*
+ β_2*(Schooling Years)*
+ β_3*(Potential Experience)* + β_4*(Potential Experience)*^{2}
+ β_5*(HS Certificate)* + β_6*(Tech/Coll Cert)*
+ β_7*(Univ. Undergrad Certificate)*
+ β_8*(Bachelors)* + β_9*(PostGrad Cert)*
+ β_{10}*(Masters)* + β_{10}*(Medical/Dental Doctor)*
+ β_{11}*(PHD)* + ε

48

Specific Hypotheses

- HK-Only Model: $b_5 = b_6 = \ldots b_{11} = 0$ if schooling years in regression.
- Screening-Only Model: $b_2 = 0$ if credentials in regression.
- BOTH models:
  - Experience matters.
  - Males and females should have different equations.

49

Data

- 2001 Census, PUMF, Saskatchewan weekly wages, for *all* levels of schooling.
  - Only those between 21 and 64, working full-time, not in school.
  - Removed: immigrants, self-employed, those with weekly wages < $250.

50

Simple Regression

Ln (Weekly Wages) = 5.95 + 0.043SchoolingYears,
(0.035) (0.002)

$R^2 = 0.043$, F-stat = 272.56

- The prob. value = 0.000000…. for both estimates and for the F-stat.

51

Dependent Variable: Ln (Weekly Wages)

Variable	*Both*	*Females*	*Males*
Intercept	5.762 (0.047)*	5.299 (0.072)*	5.772 (0.064)*
Female	-0.326 (0.012)*	–	–
Syears	0.033 (0.004)*	0.048 (0.005)*	0.027 (0.005)*
POTEXP	0.034 (0.002)*	0.027 (0.003)*	0.040 (0.003)*
POTEXPSQ	-0.001 (0.000)*	0.000 (0.000)*	-0.001 (0.000)*

Cont'd

St. Errors in brackets:

* (**, ***)– significant at the 1% (5%, 10%) level

52

Dependent Variable: Ln (Weekly Wages)

Variable	*Both*	*Females*	*Males*
Hschool	-0.001 (0.018)	0.034 (0.026)	-0.018 (0.025)
TECHCOLL	0.060 (0.018)*	0.047 (0.025)***	0.064 (0.025)**
UNICERT	0.056 (0.041)	0.025 (0.048)	0.065 (0.069)
BACH	0.299 (0.028)*	0.302 (0.037)*	0.252 (0.040)*
POSTCERT	0.356 (0.062)*	0.336 (0.083)*	0.333 (0.089)*
MEDICAL	0.904 (0.139)*	0.740 (0.182)*	0.963 (0.202)*

Cont'd

* (**, ***)– significant at the 1% (5%, 10%) level

53

Dependent Variable: Ln (Weekly Wages)

Variable	*Both*	*Females*	*Males*
MASTERS	0.282 (0.053)*	0.358 (0.076)*	0.206 (0.072)*
PHD	0.485 (0.098)*	0.483 (0.146)*	0.465 (0.131)*
No. of obs.	6003	2587	3416
R^2 (adj)	0.2344	0.2114	0.1671
F-Test	154.138*	64.037*	63.281*

St. Errors in brackets:

* (**, ***)– significant at the 1% (5%, 10%) level

54

Interpretation: Schooling Years & Experience

Variable	*Both*	*Females*	*Males*
Female	-27.8%	–	–
Syears	3.3%	5.0%	2.8%
POTEXP	3.4%	2.7%	4.0%
POTEXPSQ	-0.1%	0.0%	-0.1%

Schooling years do matter → human capital model is NOT rejected

Females have a higher return to schooling than males, but a lower return to potential experience.

Females earn about 28% less, controlling for schooling and potential experience.

55

Interpretation: Certificate Effects

Variable	*Both*	*Females*	*Males*
Hschool	Not significantly different than less than HS		
TECHCOLL	6.2%	4.8%	6.7%
UNICERT	Not significantly different than less than HS		
BACH	34.9%	35.2%	28.7%
POSTCERT	42.7%	40.0%	39.5%
MEDICAL	146.9%	109.6%	162.0%
MASTERS	32.6%	43.0%	22.9%
PHD	62.5%	62.2%	59.2%

Certificate effects are generally positive for post-secondary education → the Signalling model is NOT rejected.
Mixed HK-signalling Model is true!?

56

Interpretation: Certificate Effects

Variable	*Both*	*Females*	*Males*
Hschool	Not significantly different than less than HS		
TECHCOLL	6.2%	4.8%	6.7%
UNICERT	Not significantly different than less than HS		
BACH	34.9%	35.2%	28.7%
POSTCERT	42.7%	40.0%	39.5%
MEDICAL	146.9%	109.6%	162.0%
MASTERS	32.6%	43.0%	22.9%
PHD	62.5%	62.2%	59.2%

Less than HS, High School graduate, and university certificate all pay about the same. Technical college certificate (SIAST) pays a bit more, especially for males.

57

Interpretation: Certificate Effects

Variable	*Both*	*Females*	*Males*
Hschool	Not significantly different than less than HS		
TECHCOLL	6.2%	4.8%	6.7%
UNICERT	Not significantly different than less than HS		
BACH	34.9%	35.2%	28.7%
POSTCERT	42.7%	40.0%	39.5%
MEDICAL	146.9%	109.6%	162.0%
MASTERS	32.6%	43.0%	22.9%
PHD	62.5%	62.2%	59.2%

Returns to university education are strong, and generally higher for women.

58

Interpretation: Certificate Effects

Variable	*Both*	*Females*	*Males*
Hschool	Not significantly different than less than HS		
TECHCOLL	6.2%	4.8%	6.7%
UNICERT	Not significantly different than less than HS		
BACH	34.9%	35.2%	28.7%
POSTCERT	42.7%	40.0%	39.5%
MEDICAL	146.9%	109.6%	162.0%
MASTERS	32.6%	43.0%	22.9%
PHD	62.5%	62.2%	59.2%

A university degree is worth MUCH more than a SIAST degree (in 2001).

59

Certificate and Human Capital Effects

- Pure Human Capital model → rejected.
- Pure Certificate model → rejected.
- Mixed human capital and certificate model → NOT rejected.

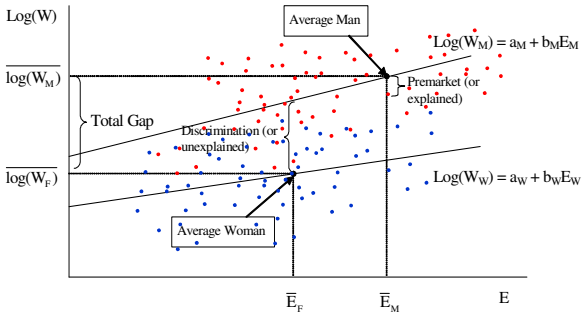
60

Male-Female Wage Differences Revisited

- Role of actual vs. potential experience?
  - Return males = 4% > 2.7% = return females.
  - M. Drolet, "New Evidence on Gender Pay Differentials: Does Measurement Matter?", *Canadian Public Policy*, March 2002:
    - Average man (age 39) has 18.3 years of *full-time* labour market experiences.
    - Average woman (also age 39): only 14.4 years of *full-time* experience.

61

Decomposing Male-Female Wage Differences



62

10 Most Frequent Jobs for Men, 2006 Census

Retail salespersons and sales clerks	285,800
Truck drivers	276,200
Retail trade managers	192,200
Janitors, caretakers and building superintendents	154,100
Farmers and farm managers	147,800
Material handlers	147,000
Automotive service technicians, truck and bus mechanics and mechanical repairers	143,000
Carpenters	142,400
Construction trades helpers and labourers	133,600
Sales, marketing and advertising managers	102,600

10 Most Frequent Jobs for Women, 2006 Census

Retail salespersons and sales clerks	400,000
Cashiers	255,500
Registered nurses	249,400
General office clerks	244,200
Secretaries (except legal and medical)	237,300
Elementary school and kindergarten teachers	214,600
Food counter attendants, kitchen helpers and related occupations	194,800
Early childhood educators and assistants	157,700
Food and beverage servers	152,000
Light duty cleaners	147,400

Total University Degrees (thousands)

Year	Men	Women	Female/Male Ratio
1991	1347.8	1058	0.78
1993	1537	1244.8	0.81
1995	1614.8	1392	0.86
1997	1735.4	1511.6	0.87
1999	1846.5	1675.3	0.91
2001	2024.5	1880.7	0.93
2003	2203.1	2103.1	0.95
2005	2345.2	2345.2	1.00

Source: Labour Force Historical Review, 2005

65