FINANCIAL MODELLING

LINEAR REGRESSION AND ASSUMPTION VIOLATIONS

Dr Issam Malki
School of Finance and Accounting
Westminster Business School

February 2024

Aims and Objectives

- Multiple Linear Regression
  - Multicollinearity
  - Dummy variables
    - Intercept dummy
    - Interaction terms
  - Functional form
    - Transformations
    - Testing for miss-specification
  - Stability of the model
    - Testing stability using dummy variables
    - Structural breaks tests

- Applications

Assumption Violations: Multicollinearity

- Collinearity: High correlation exists among two or more independent variables
- This means the correlated variables contribute redundant information to the multiple regression model
- Types of multicollinearity
  - Perfect multicollinearity
    - Impossible to estimate the model
  - Imperfect (near perfect) multicollinearity
    - One or more variables will be dropped from the regression
    - e.g. suppose $x_3 = 2x_2$, then

Assumption Violations: Multicollinearity

- The presence of multicollinearity
  - No new information provided
  - Can lead to unstable coefficients (large standard error and low t-values)
  - Coefficient signs may not match prior expectations
- Consequences
  - Coefficients differ from the values expected by theory or experience, or have incorrect signs
  - Coefficients of variables believed to be a strong influence have small t statistics indicating that their values do not differ from 0
  - All the coefficient student t statistics are small, indicating no individual effect, but the overall F statistic indicates a strong effect for the total regression model

Assumption Violations: Multicollinearity

Detection of Multicollinearity

- Observe the signs and t statistics of the estimated coefficients
  - Wrong signs may indicate the presence of multicollinearity
  - Too many low t statistics
- Correlation matrix
  - Highly correlated variables are possibly linearly dependent and could be the cause of multicollinearity.
  - Run regressions involving these correlated variables. Use t statistic to establish dependency.
- Extension of the correlation matrix
  - VIF: Variance inflation
  - Rule of thumb: VIF>10, multicollinearity is problematic

Assumption Violations: Multicollinearity

Dealing with Multicollinearity

- Do nothing
  - Something comes with data
  - If the model is OK we can live with multicollinearity
- Remove one or more of the highly correlated independent variables.
  - Remove the most problematic one according to VIF or correlation matrix.
  - This might lead to a bias in coefficient estimation.
- Change the model specification, including possibly a new independent variable that is a function of several correlated independent variables.
- Obtain additional data that do not have the same strong correlations between the independent variables

Assumption Violations: Multicollinearity- Application

Dependent Variable: SALARY
Method: Least Squares
Date: 11/06/18 Time: 19:02
Sample: 1 50
Included observations: 50

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	531.7914	201.1622	2.643595	0.0115
BONUS	0.056737	0.334828	0.169451	0.8663
EXPER	19.56880	13.56184	1.442931	0.1565
SALES	0.067042	0.020832	3.218233	0.0025
PCTOWN	-42.69000	35.61469	-1.198663	0.2374
PROFITS	-0.039299	0.293650	-0.133830	0.8942
TENURE	-1.299592	8.873796	-0.146453	0.8843
VALUATE	0.219031	0.808064	0.271057	0.7877

R-squared	0.310833	Mean dependent var	920.1200
Adjusted R-squared	0.195972	S.D. dependent var	697.6053
S.E. of regression	625.5261	Akaike info criterion	15.86071
Sum squared resid	16433882	Schwarz criterion	16.16663
Log likelihood	-388.5177	Hannan-Quinn criter.	15.97721
F-statistic	2.706161	Durbin-Watson stat	1.916155
Prob(F-statistic)	0.020756		

Assumption Violations: Multicollinearity- Application

Covariance Analysis: Ordinary
Date: 11/06/18 Time: 19:04
Sample: 1 50
Included observations: 50

Correlation	SALARY	BONUS	EXPER	SALES	PCTOWN	PROFITS	TENURE	VALUATE
SALARY	1.000000							
BONUS	0.145450	1.000000						
EXPER	0.153776	0.488015	1.000000					
SALES	0.456654	0.116691	-0.052477	1.000000				
PCTOWN	-0.244807	0.173993	0.298364	-0.109781	1.000000			
PROFITS	0.086059	0.335293	0.187521	0.143875	0.052752	1.000000		
TENURE	0.088177	0.282987	0.467012	0.120960	0.198889	0.191123	1.000000	
VALUATE	-0.135723	0.220330	0.356798	-0.029709	0.866030	0.102561	0.062488	1.000000

Assumption Violations: Multicollinearity- Application

Dependent Variable: PCTOWN
Method: Least Squares
Date: 11/06/18 Time: 19:06
Sample: 1 50
Included observations: 50

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.656479	0.407582	1.610668	0.1138
VALUATE	0.019214	0.001601	12.00024	0.0000

R-squared	0.750007	Mean dependent var	1.853800
Adjusted R-squared	0.744799	S.D. dependent var	5.531459
S.E. of regression	2.794350	Akaike info criterion	4.932254
Sum squared resid	374.8027	Schwarz criterion	5.008735
Log likelihood	-121.3063	Hannan-Quinn criter.	4.961378
F-statistic	144.0057	Durbin-Watson stat	2.058378
Prob(F-statistic)	0.000000		

Assumption Violations: Multicollinearity- Application

Variance Inflation Factors
Date: 11/06/18 Time: 19:08
Sample: 1 50
Included observations: 50

Variable	Coefficient Variance	Uncentered VIF	Centered VIF
C	40466.23	5.170968	NA
BONUS	0.112110	2.162999	1.455689
EXPER	183.9236	4.384182	1.851904
SALES	0.000434	2.047357	1.126194
PCTOWN	1268.406	5.417087	4.860076
PROFITS	0.086231	1.314389	1.162518
TENURE	78.74425	7.146862	1.589947
VALUATE	0.652967	5.407007	5.082989

Assumption Violations: Multicollinearity- Application

Dependent Variable: SALARY
Method: Least Squares
Date: 11/06/18 Time: 19:09
Sample: 1 50
Included observations: 50

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	533.9853	198.8220	2.685746	0.0102
BONUS	0.058896	0.331107	0.177876	0.8597
EXPER	20.75696	12.69489	1.635065	0.1093
SALES	0.068165	0.020195	3.375402	0.0016
PCTOWN	-34.22133	16.91070	-2.023649	0.0492
PROFITS	-0.031712	0.289147	-0.109675	0.9132
TENURE	-2.258085	8.050636	-0.280485	0.7805

R-squared	0.309627	Mean dependent var	920.1200
Adjusted R-squared	0.213296	S.D. dependent var	697.6053
S.E. of regression	618.7502	Akaike info criterion	15.82246
Sum squared resid	16462630	Schwarz criterion	16.09014
Log likelihood	-388.5614	Hannan-Quinn criter.	15.92439
F-statistic	3.214200	Durbin-Watson stat	1.937359
Prob(F-statistic)	0.010703		

Dependent Variable: SALARY
Method: Least Squares
Date: 11/06/18 Time: 19:09
Sample: 1 50
Included observations: 50

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	541.9603	202.0014	2.682953	0.0103
BONUS	0.055867	0.336523	0.166014	0.8689
EXPER	22.82458	13.35437	1.709147	0.0946
SALES	0.072769	0.020379	3.570697	0.0009
PROFITS	-0.008365	0.293996	-0.028454	0.9774
TENURE	-5.424316	8.220867	-0.659823	0.5129
VALUATE	-0.630675	0.389855	-1.617719	0.1130

R-squared	0.287257	Mean dependent var	920.1200
Adjusted R-squared	0.187804	S.D. dependent var	697.6053
S.E. of regression	628.6951	Akaike info criterion	15.85435
Sum squared resid	16996074	Schwarz criterion	16.12203
Log likelihood	-389.3587	Hannan-Quinn criter.	15.95628
F-statistic	2.888383	Durbin-Watson stat	1.982461
Prob(F-statistic)	0.018689		

Dummy Variables

- Dummy variables can be used in situations in which the categorical variable of interest is included in a regression.
- Dummy variables can also be useful
  - Experimental design to identify possible causes of variation in the value of the dependent variable
  - Measuring differences between categories
- Dummy variables Structure
  - Binary: takes value one or zero
  - Ordinal: takes values from zero (positive integers)

Dummy Variables- Intercept Dummy Variables

Dummy variables are binary (0,1)

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + \varepsilon_i$$

Y_i = Salary

X_i = sales

D_i = 1 if postgraduate degree,

D_i = 0 otherwise.

Higher return to education if the effect is positive and significant

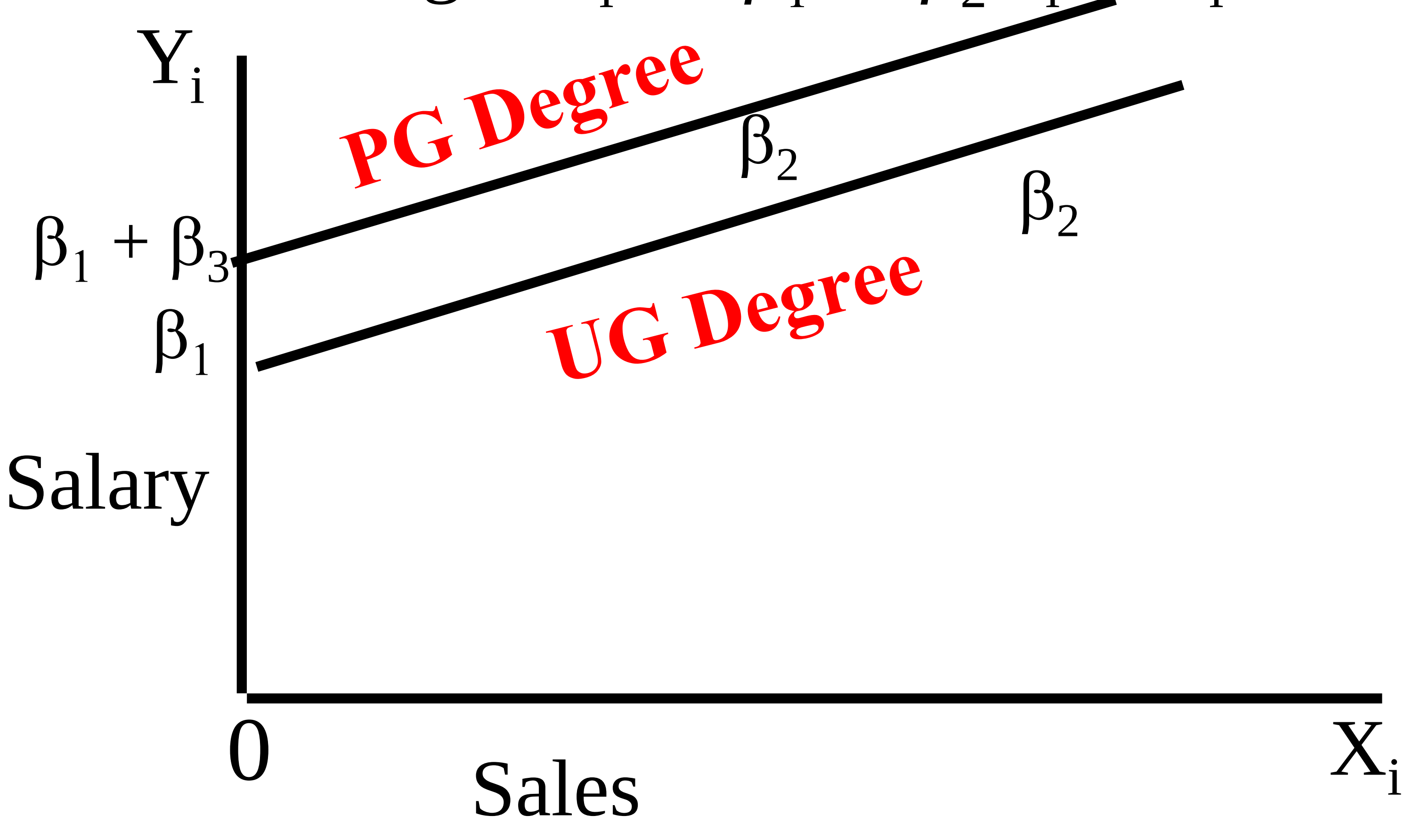
H_0: $\beta_3 = 0$

H_1: $\beta_3 > 0$

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + \varepsilon_i$$

MA/MSc degree: $Y_i = (\beta_1 + \beta_3) + \beta_2 X_i + \varepsilon_i$

No advertising: $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$



Dummy Variables- Intercept Dummy Variables

Dependent Variable: SALARY
Method: Least Squares
Date: 11/06/18 Time: 20:38
Sample: 1 50
Included observations: 50

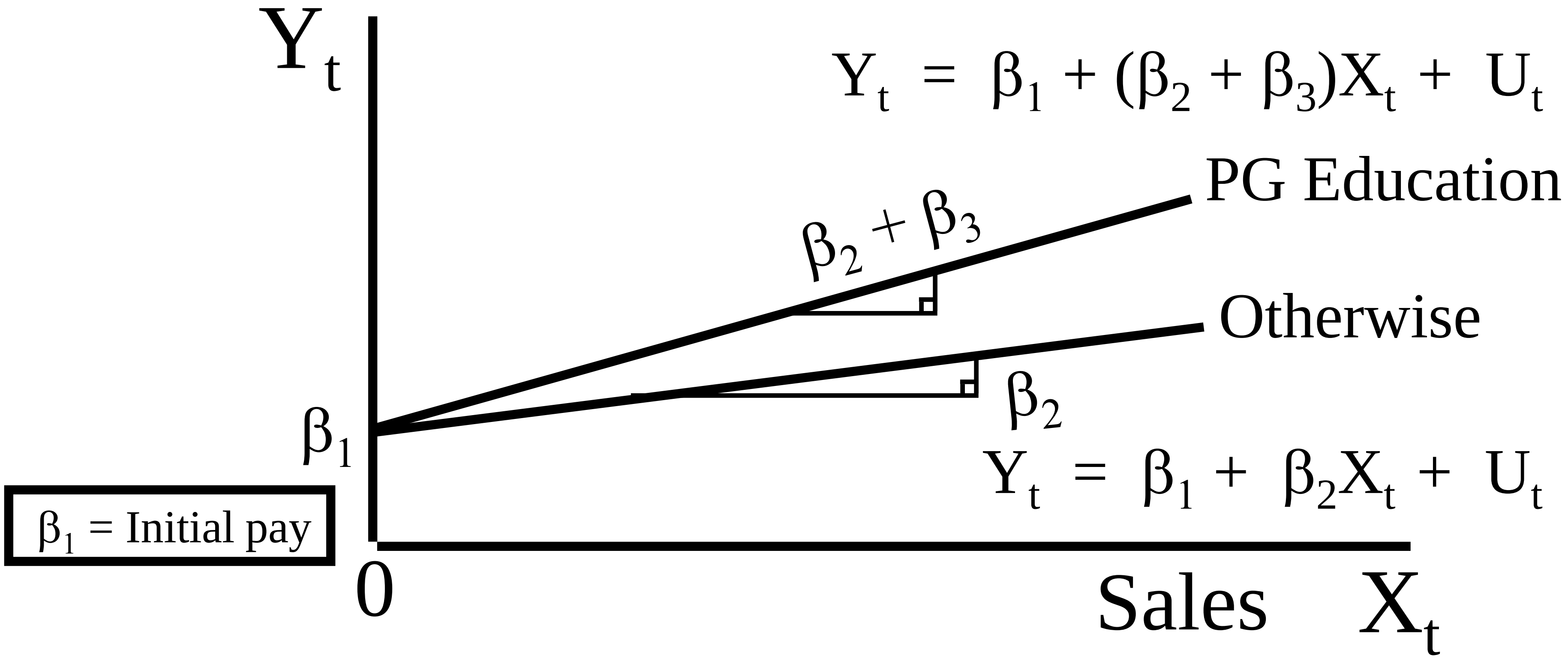
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	876.2444	155.1371	5.648194	0.0000
SALES	0.067348	0.018883	3.566644	0.0008
DEDUC	-397.6123	172.4119	-2.306176	0.0256

R-squared	0.288990	Mean dependent var	920.1200
Adjusted R-squared	0.258734	S.D. dependent var	697.6053
S.E. of regression	600.6158	Akaike info criterion	15.69191
Sum squared resid	16954748	Schwarz criterion	15.80663
Log likelihood	-389.2978	Hannan-Quinn criter.	15.73560
F-statistic	9.551573	Durbin-Watson stat	1.921815
Prob(F-statistic)	0.000330		

DUMMY VARIABLES- SLOPE DUMMY

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 D_t X_t + \varepsilon_t$$

PG Education : $D_t = 1$	Otherwise: $D_t = 0$



DUMMY VARIABLES- INTERCEPT DUMMY VARIABLES

Dependent Variable: SALARY
Method: Least Squares
Date: 11/06/18 Time: 20:47
Sample: 1 50
Included observations: 50

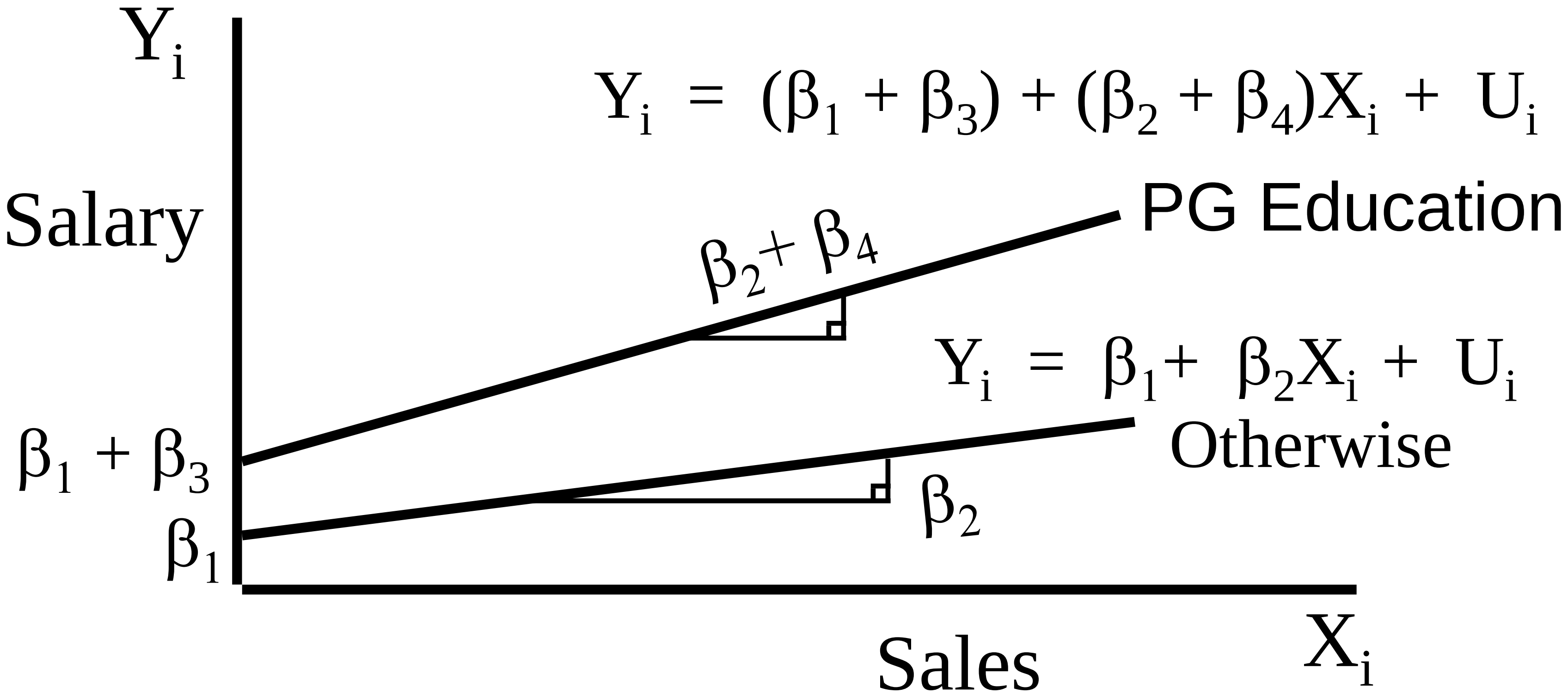
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	605.8911	100.1752	6.048311	0.0000
SALES	0.140173	0.022324	6.279019	0.0000
EDSALES	-0.115305	0.024782	-4.652679	0.0000

R-squared	0.458116	Mean dependent var	920.1200
Adjusted R-squared	0.435057	S.D. dependent var	697.6053
S.E. of regression	524.3389	Akaike info criterion	15.42028
Sum squared resid	12921769	Schwarz criterion	15.53500
Log likelihood	-382.5069	Hannan-Quinn criter.	15.46396
F-statistic	19.86721	Durbin-Watson stat	1.765972
Prob(F-statistic)	0.000001		

Dummy Variables- Both intercept and slope dummies

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + \beta_4 D_i X_i + U_i$$

PG Education: $D_i = 1$	Otherwise: $D_i = 0$



Dummy Variables- Both intercept and slope dummies

Dependent Variable: SALARY
Method: Least Squares
Date: 11/06/18 Time: 20:50
Sample: 1 50
Included observations: 50

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	501.2252	166.6186	3.008219	0.0043
DEDUC	164.6755	208.9947	0.787941	0.4348
SALES	0.152664	0.027454	5.560726	0.0000
EDSALES	-0.134033	0.034411	-3.895069	0.0003

R-squared	0.465332	Mean dependent var	920.1200
Adjusted R-squared	0.430463	S.D. dependent var	697.6053
S.E. of regression	526.4667	Akaike info criterion	15.44687
Sum squared resid	12749689	Schwarz criterion	15.59983
Log likelihood	-382.1718	Hannan-Quinn criter.	15.50512
F-statistic	13.34491	Durbin-Watson stat	1.699000
Prob(F-statistic)	0.000002		

DUMMY VARIABLES- MORE THAN TWO CATEGORIES

If a qualitative variable has m categories:

- Introduce all m categories and drop the intercept term

- Keep the intercept and introduce m-1 categories

If we do not follow the above guidelines then we fall into what is known as the dummy variable trap (perfect multicollinearity)

Dummy Variables- Testing for Qualitative Effect

1. Test for differences in intercept.

2. Test for differences in slope.

3. Test for differences in both intercept and slope.

Dummy Variables- Testing for Qualitative Effect- Salary differences due to education

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + \beta_4 D_i X_i + U_i$$

intercept

$$H_0:\ \beta_3 = 0 \text{ vs. } H_1:\ \beta_3 > 0$$

$\boxed{\text{Testing for the effect of education.}}$

$$\frac{\hat{\beta}_3 - 0}{Se(\hat{\beta}_3)} \sim t^c_{\alpha;n-4}$$

slope

$$H_0:\ \beta_4 = 0 \text{ vs. } H_1:\ \beta_4 > 0$$

$\boxed{\text{Testing for the effect of education}}$

$$\frac{\hat{\beta}_4 - 0}{Se(\hat{\beta}_4)} \sim t^c_{\alpha;n-4}$$

Dummy Variables- Testing for Qualitative Effect- Salary differences due to education

Testing: H_o: $\beta 3 = \beta 4 = 0$

H_1 : otherwise

$$\frac{(RSS_R - RSS_U)/2}{RSS_U/(n-4)} \sim F_{\alpha;\,2,\,n-4}$$

intercept and slope

Wald Test:
Equation: Untitled

Test Statistic	Value	df	Probability
F-statistic	11.04682	(2, 46)	0.0001
Chi-square	22.09364	2	0.0000

Functional Form- Miss-specification Tests

- Sources of miss-specification
  - We omit a relevant variable, or
  - We include an irrelevant variable, or
  - We use an incorrect functional form

- Detection using informal tests
  - Refer back to theory
  - Changes in signs and significance when adding new variables
  - Changes in Adjusted R-squared.
  - Changes in residuals patterns.
- Detection using formal tests
  - Ramsey Test
  - Known as: RESET

Functional Form- Miss-specification Tests

- RESET is based on the explanatory power of the fitted values
- Consider the model

$$\boxed{Y_i = b_1 + b_2 X_{2i} + V_i}$$

- Steps
  - Estimate the model and save fitted values $\hat{Y}_i$
  - Construct proxies to capture general miss-specification based on the fitted values: $\hat{Y}_i^2, \hat{Y}_i^3, \hat{Y}_i^4$
  - Estimate the model above including the proxies above

$$\boxed{Y_i = b_1 + b_2 X_{2i} + b_3 \hat{Y}_i^2 + b_4 \hat{Y}_i^3 + b_5 \hat{Y}_i^4 + U_i}$$

FUNCTIONAL FORM- MISS-SPECIFICATION TESTS

- Steps
  - Estimate the model and save fitted values $\widehat{Y}_i$
  - Construct proxies to capture general miss-specification based on the fitted values: $\hat{Y}_i^2$, $\hat{Y}_i^3$, $\hat{Y}_i^4$
  - Estimate the model above including the proxies above.
  - Compute the F statistic of the joint significance of the terms: $\hat{Y}_i^2$, $\hat{Y}_i^3$, $\hat{Y}_i^4$
  - The null hypothesis: the model has correct specification
  - Reject the null if the F-statistic is above the critical value.

- Remark
  - RESET is easy to apply but cannot tell us the reason for the mis-specification (i.e. omitted variable or functional form)

FUNCTIONAL FORM- MISS-SPECIFICATION TESTS

Ramsey RESET Test
Equation: EQ01
Specification: SALARY C BONUS EXPER SALES PCTOWN PROFITS TENURE VALUATE
Omitted Variables: Powers of fitted values from 2 to 4

	Value	df	Probability
F-statistic	2.725114	(3, 39)	0.0572
Likelihood ratio	9.515485	3	0.0232

F-test summary:

	Sum of Sq.	df	Mean Squares
Test SSR	2847941.	3	949313.8
Restricted SSR	16433882	42	391282.9
Unrestricted SSR	13585940	39	348357.4

LR test summary:

	Value	df
Restricted LogL	-388.5177	42
Unrestricted LogL	-383.7600	39

STABILITY TESTS

- Recall

Dependent Variable: SALARY
Method: Least Squares
Date: 11/06/18 Time: 20:50
Sample: 1 50
Included observations: 50

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	501.2252	166.6186	3.008219	0.0043
DEDUC	164.6755	208.9947	0.787941	0.4348
SALES	0.152664	0.027454	5.560726	0.0000
EDSALES	-0.134033	0.034411	-3.895069	0.0003

R-squared	0.465332	Mean dependent var	920.1200
Adjusted R-squared	0.430463	S.D. dependent var	697.6053
S.E. of regression	526.4667	Akaike info criterion	15.44687
Sum squared resid	12749689	Schwarz criterion	15.59983
Log likelihood	-382.1718	Hannan-Quinn criter.	15.50512
F-statistic	13.34491	Durbin-Watson stat	1.699000
Prob(F-statistic)	0.000002		

Wald Test:
Equation: Untitled

Test Statistic	Value	df	Probability
F-statistic	11.04682	(2, 46)	0.0001
Chi-square	22.09364	2	0.0000

STABILITY TESTS

- This means we can estimate two models for the two groups
  - As long as the sample allows- otherwise keep the same structure with dummies.
  - Dummy variables can be used to test the stability of the relationship.

- A more general framework is to use formal tests of structural breaks.

- There are two types
  - Tests with known structural breaks
  - Tests with unknown structural breaks

Stability Tests

- Test with known breaks
  - Popular test: Chow break point
  - Relevant when we have issues of comparing between categories and groups.
  - For time series you need to know the date of the occurrence of the change.
  - For cross section you need to sort the data by group

Chow Breakpoint Test: 22
Null Hypothesis: No breaks at specified breakpoints
Varying regressors: All equation variables
Equation Sample: 1 50

F-statistic	3.787004	Prob. F(8,34)	0.0029
Log likelihood ratio	31.85687	Prob. Chi-Square(8)	0.0001
Wald Statistic	30.29603	Prob. Chi-Square(8)	0.0002

Stability Tests

- Test with known breaks
  - Step 1: Estimate the model using the full sample. Save RSS.
  - Step 2: split the sample into two sub-samples. Estimate the model and save their RSS (RSS_1 and RSS_2).
  - The null hypothesis: the model is stable. The alternative hypothesis: the model has a structural break at the point defined.
  - Step 3: Compute the F-statistic

$$F = \frac{(RSS - (RSS_1 + RSS_2))/k}{(RSS_1 + RSS_2)/(N_1 + N_2 - 2k)}$$

  - Step 4: If the F exceeds the critical value, reject the null.

STABILITY TESTS

- Test with unknown breaks
  - Popular test: Quandt test
  - Relevant when we do not know when the changes would occur.

THANK YOU