MPO1: Quantitative Research Methods

Session 6: F-tests for goodness of fit, Non-linearity and Model Transformations, Dummy variables, Interactions

Thilo Klein

University of Cambridge
Judge Business School

χ^2 and F Distributions

Chi-squared Distribution $\chi_K{}^2$

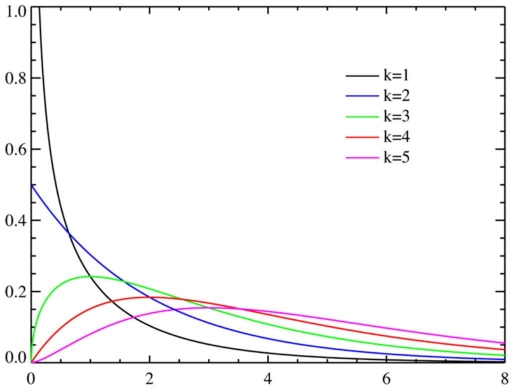
- If $Y_i \sim N(0,1)$, then
- $\sum_{i=1}^{K} Y_i{}^2 \sim \chi_K{}^2$ distribution, with K degrees of freedom

$$pdf: \; f(y,K) = \begin{cases} \frac{1}{2^{K/2}\Gamma(K/2)} y^{(K/2)-1} e^{-y/2} & for\ y > 0 \\ 0 & for\ y \leq 0 \end{cases}$$

  - $\Gamma(\cdot)$ is the Gamma function
- $E(\sum_{i=1}^{K} Y_i{}^2) = K$

χ^2 and F Distributions

Chi-squared Distribution ${\chi_K}^2$



χ^2 and F Distributions

F Distribution

- If $U_1 \sim {\chi_{df_1}}^2$, $U_2 \sim {\chi_{df_2}}^2$ and U_1, U_2 are independent, then

$$X = \frac{U_1/df_1}{U_2/df_2} \sim F_{df_1,df_2}$$

- pdf of an F distributed random variable, X with df_1 and df_2 degrees of freedom is:

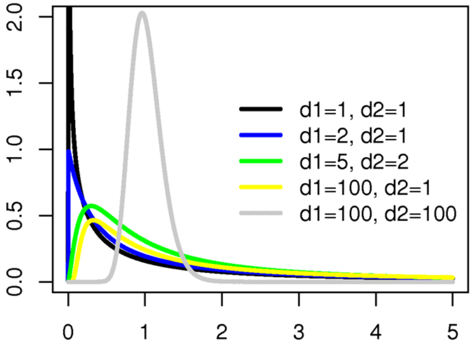
$$f(x) = \frac{\sqrt{\frac{(df_1\,x)^{df_1}\;df_2^{df_2}}{(df_1\,x+df_2)^{df_1+df_2}}}}{x\,\mathrm{B}\left(\frac{df_1}{2},\frac{df_2}{2}\right)}$$

  - $B(\cdot,\cdot)$ is the Beta function
- $E(X) = \frac{df_2}{df_2-2}$ for $df_2 > 0$

CAMBRIDGE Judge Business School

χ^2 and F Distributions

F-distribution



F Tests of fit

F-test of R^2

$$Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u_i$$

$$H_0 : \beta_1 = \cdots = \beta_K = 0 \quad H_a : \text{at least one } \beta \neq 0$$

$$\begin{aligned}
\frac{ESS/(K-1)}{RSS/(n-K)} &= \frac{\frac{ESS}{TSS}/(K-1)}{\frac{RSS}{TSS}/(n-K)} \\
&= \frac{R^2/(K-1)}{(1-R^2)/(n-K)} \sim F(K-1, n-K)
\end{aligned}$$

Application

CAMBRIDGE Judge Business School

F Tests of fit

Another application: incremental contribution of a set of variables

- $Y = \beta_1 + \beta_2 X_2 + u : \quad RSS_1$
- $Y = \beta_1 + \beta_2 X_2 + +\beta_3 X_3 + \beta_4 X_4 + u : \quad RSS_2$
- $H_0 : \beta_3 = \beta_4 = 0; \quad H_a : \beta_3 \neq 0$ or $\beta_4 \neq 0$ or both β_3 and $\beta_4 \neq 0$

$$\frac{\text{Increase in ESS}}{\text{cost in d.f.}} \Big/ \frac{\text{remaining RSS}}{\text{d.f. remaining}} \sim F(\text{cost, d.f. remaining})$$

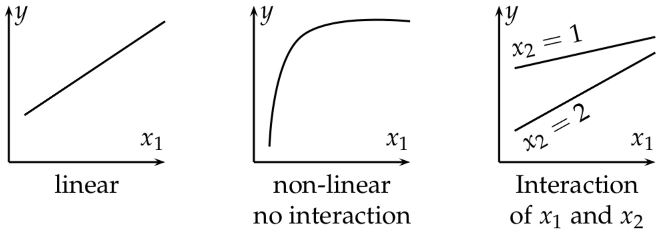
$$\frac{(RSS_1 - RSS_2)/(df_1 - df_2)}{RSS_2/df_2} \sim F(df_1, df_2)$$

- Note: $F_{1,n}$ is the squared Student t_n distribution
- A series of independent t tests is not the same as an F test: why?

CAMBRIDGE Judge Business School

Plan for today

Non-linear regression functions



If the dependence between Y and X is non-linear, the marginal effect of X is not constant.

Approach:

- non-linear functions of a single independent variable
  - Polynomials in X; Logarithmic transformation
- Interactions

Model Building 1: Variable transformations

Why variable transformations?

- Transformations: suitable mathematical functions applied to variables
- Sometimes sensible to transform the dependent and/or explanatory variables through one-to-one functions, and estimate the model with these transformed variables. Why?
  - May make more sense from a theoretical or data generating point of view.
  - Mulitple linear regression more reliable when predictors have reasonably symmetric distributions and are not too highly skewed in distribution
  - Many variables of interest are positively skewed: a log transformation works well to transform such variables

CAMBRIDGE Judge Business School

Model Building 1: Variable transformations

Linearity and Nonlinearity

- Linear in variables and parameters:
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$
- Linear in parameters, nonlinear in variables:
  - $Y = \beta_0 + \beta_1 {X_1}^2 + \beta_2 \sqrt{X_2} + \beta_3 log X_3 + u$
  - $Z_1 = {X_1}^2,\ Z_2 = \sqrt{X_2},\ Z_3 = log X_3$
  - $Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + u$
  - *Cosmetic* transformations sufficient to make the model linear in variables
- Nonlinear in parameters: Cannot estimate with OLS - but other methods exist
  - $Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + (\beta_1(1 - \beta_2)) Z_3 + u$

CAMBRIDGE Judge Business School

Model Building 1: Variable transformations

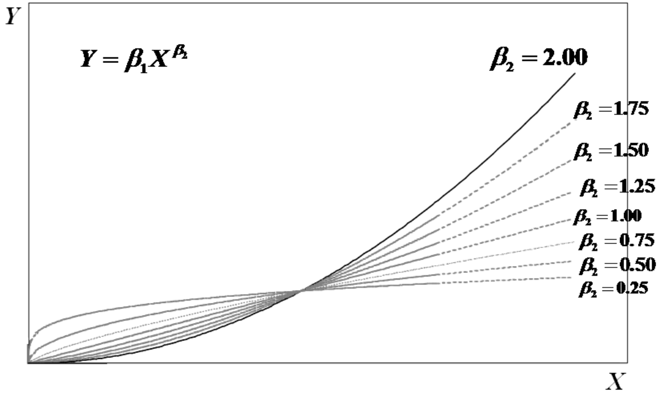
Double-logarithmic models and Elasticity

- Sometimes a stronger linear relationship between $logY$ and $logX$, than between Y and X (Why?)
- Examples: demand functions: 1% change in price leads to (constant) $x\%$ change in quantity demanded
- Proportionate change in Y linearly related to proportionate change in X
- Double-logarithmic model: constant elasticity of Y with respect to X
- Elasticity $= \frac{dY/Y}{dX/X} = \frac{dY/dX}{Y/X}$

CAMBRIDGE Judge Business School

Model Building 1: Variable transformations

Double-logarithmic models and Elasticity: figure



Model Building 1: Variable transformations

Double-logarithmic models and Elasticity (2)

- $Y = \beta_0 X^{\beta_1}$
  - $\frac{dY}{dX} = \beta_0 \beta_1 X^{\beta_1 - 1}$
  - $\frac{Y}{X} = \frac{\beta_0 X^{\beta_1}}{X} = \beta_0 X^{\beta_1 - 1}$
  - Elasticity $= \frac{dY/dX}{Y/X} = \frac{\beta_0 \beta_1 X^{\beta_1 - 1}}{\beta_0 X^{\beta_1 - 1}} = \beta_1$
- Simple to fit a constant elasticity model to data: linearize the model by taking the logarithms of both sides

$$\begin{aligned} logY &= log\left(\beta_0 X^{\beta_1}\right) = log\beta_0 + log\left(X^{\beta_1}\right) \\ &= log\beta_0 + \beta_1 logX = b_0 + b_1 logX \end{aligned}$$

- The constant b_0 is the estimate of $log\beta_0$
- To obtain estimate of β_1, exponentiate the estimated regression coefficient b_1

CAMBRIDGE Judge Business School

Model Building 1: Variable transformations

Semi-logarithmic models

- Another kind of a multiplicative relationship:
  - e.g., between additional years of experience (or education) and earnings
- The semi-logarithmic specification allows the increment to increase with level of education
  - $Y = \beta_0 e^{\beta_1 X}$
  - $\frac{dY}{dX} = \beta_0 \beta_1 e^{\beta_1 X} = \beta_1 Y$
  - $\frac{dY/Y}{dX} = \beta_1$

CAMBRIDGE Judge Business School

Model Building 1: Variable transformations

Polynomial models

- $Y = \beta_1 + \beta_2 X + \beta_3 X^2 + \beta_4 X^3 + u$
- Difficult to justify powers greater than 3, unless strong theoretical reasons to fit higher power
- Center X: deviations of X from its mean (or median) can reduce collinearity between X and higher powers
- A polynomial function may be used when
  - the true response function is polynomial
  - the true response function is unknown but a polynomial is a good approximation of its shape
- General principle: hierarchy
  - Keep X in the model, if X^2 is significant
  - Keep X and X^2 in the model, if X^3 is significant

CAMBRIDGE Judge Business School

Model Building 1: Variable transformations

Polynomial regression model: why is this example interesting?

Sample: 75 "services" firms from the North of England, observed in 2002-3

Dependent variable: Annual growth rate of the firm

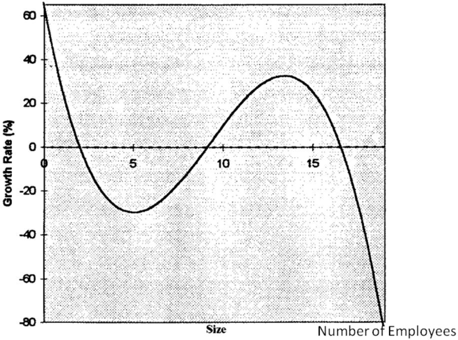
Expl Vars	**Estimates**	
CONSTANT	0.66***	
EDUC	–0.28***	: no A levels = 0; A levels = 1
TIMTR	0.46E-4***	: period respondent in business (years).
SIZE	–0.43***	: Opening employment full time equivalents
SIZESQ	0.06***	: SIZE squared
SIZECUB	–0.002***	: SIZE cubed
PPROF	–0.37***	: % of empl. accounted for by professionals
TURB	0.00***	: sum of birth and death rates in the industry
EDUCxPPROF	0.33***	: interaction term

$R^2 = 0.22$ *** Significant at 1 per cent.

CAMBRIDGE Judge Business School

Model Building 1: Variable transformations

Polynomial regression model: example, graph of mean growth conditioned on size



$$\text{Growth rate} = \beta_0 + \beta_1 \text{Size} + \beta_2 \text{Size}^2 + \beta_3 \text{Size}^3 + \text{other effects} + u$$

CAMBRIDGE Judge Business School

Model Building 1: Variable transformations

Interactions between explanatory variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2) + u$$

Transformation in practice

- $log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_2)^2 + \beta_4 log(X_3) + \beta_5 X_4 + \beta_6 X_1 X_4 + \beta_7 (\frac{1}{X_5}) + u$
- Danger: overfitting the model, Mining the sample

Dummy variables

Case: Energy costs and refrigerator pricing

- Refrigerators manufactured by a large appliance manufacturer
- The engineering division claim to have designed a new more efficient machine
  - Will cost 80 GBP more to manufacture
  - Users will save 20 GBP per year in energy costs
- Should you recommend building this?
- Q: What would customers pay to save on energy costs?

CAMBRIDGE Judge Business School

Dummy variables

Case: Energy costs and refrigerator pricing - explore

- Summary stats: Price, Ecost
- Simple regression of Price on Ecost
- Do the estimates make sense?

Dummy variables

Energy costs refrigerator price: simple regression

Call:
lm(formula = price ~ ecost)

Residuals:

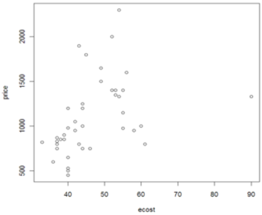
Min	1Q	Median	3Q	Max
-546.28	-304.74	-68.99	190.92	1073.77

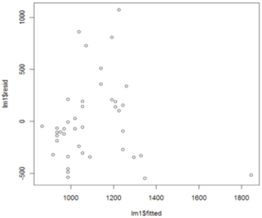
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	300.157	290.463	1.033	0.30779	
Ecost	17.150	6.075	2.823	0.00746	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 392.6 on 39 degrees of freedom
Multiple R-squared: 0.1696, Adjusted R-squared: 0.1484
F-statistic: 7.968 on 1 and 39 DF, p-value: 0.007458





CAMBRIDGE Judge Business School

Dummy variables

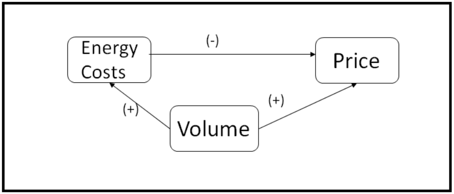
Case: Energy costs and refrigerator pricing (3)

- Other things affect price besides just energy costs
  - Size
  - Features
  - Brand
  - Design
  - Orientation(freezer on top, side by side..)
  - Others?
- Some of these other variables that impact price are also related to energy costs; notably, size
- A bigger refrigerator costs more to buy and it uses more energy

CAMBRIDGE
Judge Business School

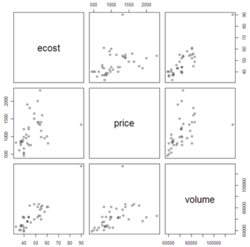
Dummy variables

Energy costs refrigerator price: Correlations



cor(data.frame(ecost, price, volume))

	ecost	price	volume
ecost	1.00	0.41	0.89
price	0.41	1.00	0.48
volume	0.89	0.48	1.00



Dummy variables

Case: Energy costs and fridge pricing - mult. regression

- How does changing energy costs impact price when volume (and other variables) are held fixed
- Multiple regression: Price on Volume and Ecost

Call: lm(formula = price ~ **volume** + ecost)

Residuals:

Min	1Q	Median	3Q	Max
-646.44	-253.73	-79.95	120.97	1194.09

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-342.89642	474.80105	-0.722	0.4746	
volume	0.02177	0.01289	1.689	0.0993	.
ecost	-2.42797	13.02064	-0.186	0.8531	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 383.6 on 38 degrees of freedom

Multiple R-squared: 0.2277, Adjusted R-squared: 0.187

F-statistic: 5.6 on 2 and 38 DF, p-value: 0.007387

CAMBRIDGE Judge Business School

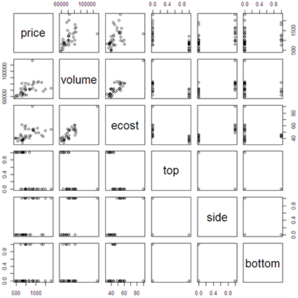
Dummy variables

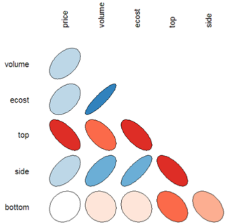
Case: Energy costs and refrigerator pricing - types of fridges

- Three types of fridges:
  - Freezer at the top
  - Freezer at the side
  - Freezer at the bottom
- Question: Will the location of the Freezer make a difference to the price at which you can sell the fridge?

Dummy variables

Energy costs refrigerator price: Freezer positions





	price	volume	ecost	top	side	bottom
price	1.00	0.48	0.41	-0.66	0.54	0.16
volume	0.48	1.00	0.89	-0.56	0.65	-0.10
ecost	0.41	0.89	1.00	-0.66	0.78	-0.15
top	-0.66	-0.56	-0.66	1.00	-0.67	-0.41
side	0.54	0.65	0.78	-0.67	1.00	-0.39
bottom	0.16	-0.10	-0.15	-0.41	-0.39	1.00

CAMBRIDGE Judge Business School

Dummy variables

Case: Energy costs and refrigerator pricing - dummy variables in data

- Data with dummy variables:

Dummy variables

Case: Energy costs and refrigerator pricing - regression with dummies

- Run a multiple regression with dummy variables to separate out the top, bottom and side types
- Run separate regressions for top, bottom and side types
  - What is the intuition?
- Interpret the coefficients

Dummy variables

Case: Energy costs refrigerator price - regression with dummies

```
Call: lm(formula = price ~ volume + ecost + top + side)

Residuals:
   Min      1Q     Median    3Q     Max
-438.52  -146.51  -69.94   86.04  1024.22

Coefficients:
               Estimate    Std. Error   t value   Pr(>|t|)
(Intercept)  918.19515   435.27647    2.109    0.041925 *
volume         0.02886     0.01007    2.865    0.006915 **
ecost        -38.57106    12.72378   -3.031    0.004491 **
top         -517.39793   131.99344   -3.920    0.000381 ***
side         345.84275   163.74242    2.112    0.041681 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
---
Residual standard error: 294.9 on 36 degrees of freedom
Multiple R-squared: 0.5675,    Adjusted R-squared: 0.5195
F-statistic: 11.81 on 4 and 36 DF,  p-value: 3.143e-06
```

CAMBRIDGE Judge Business School

Dummy variables

Omitted Variables cause bias

- In the first equation specified, the regression coefficient is CORRECT.
  - On average, a refrigerator that uses a lot of energy does cost more.
  - It also tends to be larger than average, and large refrigerators cost more
  - This indirect relationship dominates the direct, negative relationship between energy costs and price
- The effects of the missing volume and orientation variables were being picked up by the coefficient on energy cost
  - (biasing it, if what you really wanted was the effect of ecost keeping volume and orientation constant)

Dummy variables

Omitted Variables cause bias (2)

- The estimate without dummy variables measures:
  - How much price changes on average when energy costs change by 1
  - Letting other variables float (allowing them to change as they have tended to change within our data set)
- The coefficient on energy cost with dummy variable controls measures:
  - How much price changes when energy cost changes by 1, while holding both volume and orientation FIXED
  - Variables included in the regression are considered fixed
  - Omitted variables are not
- The company should go ahead and launch the new fridge.
- The expected price premium will be: $(-38.57)(-20) = 771$

CAMBRIDGE Judge Business School

Dummy variables

Regression with dummies - Changing the base category

```
lm(formula = price ~ volume + ecost + top + bottom)
Coefficients:
              Estimate     Std. Error   t value   Pr(>|t|)
(Intercept)  1264.03790   497.02682     2.543     0.01542 *
volume          0.02886     0.01007     2.865     0.00692 **
ecost         -38.57106    12.72378    -3.031     0.00449 **
top          -863.24068   173.40619    -4.978     1.61e-05 ***
bottom       -345.84275   163.74242    -2.112     0.04168 *
Residual standard error: 294.9 on 36 degrees of freedom
Multiple R-squared: 0.5675,    Adjusted R-squared: 0.5195
F-statistic: 11.81 on 4 and 36 DF,  p-value: 3.143e-06
```

```
lm(formula = price ~ volume + ecost + side + bottom)
Coefficients:
              Estimate    Std. Error   t value   Pr(>|t|)
(Intercept)  400.79722   400.89226     1.000     0.324098
volume         0.02886     0.01007     2.865     0.006915 **
ecost        -38.57106    12.72378    -3.031     0.004491 **
side         863.24068   173.40619     4.978     1.61e-05 ***
bottom       517.39793   131.99344     3.920     0.000381 ***
Residual standard error: 294.9 on 36 degrees of freedom
Multiple R-squared: 0.5675,    Adjusted R-squared: 0.5195
F-statistic: 11.81 on 4 and 36 DF,  p-value: 3.143e-06
```

Slope Dummy variables

Slope Dummy variables

- Examine data
- Run separate regressions for each type of fridge
- Compare with a single equation with intercept dummy variables and slope dummy variables.
- What do you expect to see?

Dummy variables

Separate regressions

lm(formula = price ~ volume + ecost, data = fridge[top == 1,])

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-3.780e+02	5.379e+02	-0.703	0.493743	
volume	3.746e-02	7.904e-03	4.740	0.000317	***
ecost	-3.286e+01	1.323e+01	-2.484	0.026289	*

Residual standard error: 123 on 14 degrees of freedom
Multiple R-squared: 0.6176, Adjusted R-squared: 0.563
F-statistic: 11.31 on 2 and 14 DF, p-value: 0.001195

lm(formula = price ~ volume + ecost, data = fridge[bottom == 1,])
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4796.35187	4129.06503	1.162	0.2978	
volume	0.06043	0.02508	2.409	0.0609	.
ecost	-177.39028	113.46895	-1.563	0.1787	

Residual standard error: 328.7 on 5 degrees of freedom
Multiple R-squared: 0.5377, Adjusted R-squared: 0.3528
F-statistic: 2.907 on 2 and 5 DF, p-value: 0.1453

lm(formula = price ~ volume + ecost, data = fridge[side == 1,])
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1632.27207	766.80331	2.129	0.053	.
volume	0.01377	0.02000	0.689	0.503	
ecost	-23.80089	22.89475	-1.040	0.317	

Residual standard error: 397 on 13 degrees of freedom
Multiple R-squared: 0.0919, Adjusted R-squared: -0.04781
F-statistic: 0.6578 on 2 and 13 DF, p-value: 0.5344

CAMBRIDGE Judge Business School

Dummy variables

Regression with slope dummy variables

```
lm(formula = price ~ volume + ecost + top + side + top_vol + top_ecost + side_vol +side_ecost)

Residuals:
     Min       1Q    Median      3Q     Max
 -490.68  -115.78    -61.38   72.70  953.74

Coefficients:
              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  4.796e+03   3.716e+03    1.291    0.2060
volume       6.043e-02   2.257e-02    2.677    0.0116 *
ecost       -1.774e+02   1.021e+02   -1.737    0.0920 .
top         -5.174e+03   3.935e+03   -1.315    0.1979
side        -3.164e+03   3.760e+03   -0.842    0.4063
top_vol     -2.297e-02   2.952e-02   -0.778    0.4422
top_ecost    1.445e+02   1.070e+02    1.351    0.1861
side_vol    -4.666e-02   2.705e-02   -1.725    0.0942 .
side_ecost   1.536e+02   1.035e+02    1.484    0.1477
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 295.8 on 32 degrees of freedom
Multiple R-squared: 0.6132,     Adjusted R-squared: 0.5164
F-statistic:  6.34 on 8 and 32 DF,  p-value: 6.29e-05
```

CAMBRIDGE Judge Business School

Slope Dummy variables

Interpreting slope Dummy variables coefficients

- Nothing much is significant!
- Problem: rampant multi-collinearity
- But useful exercise to interpret:
  - No difference in base price between Top, Bottom and Side fridges
  - With each cc increase in volume of Bottom fridges, price goes up by 6 pence (significant at 5% level)
  - No significant difference from this for top fridges
  - Side fridge prices go up by 1.3 pence per cc. The difference between bottom and side (4.7 pence per cc) is significant at 10% level.)
  - As energy cost goes up, price for bottom fridges goes down by 177
  - No different for Top or Side fridges

Slope Dummy variables

Comparing Regressions

- These are the same equations that we saw in the three simple regressions that we started with.
- The multiple regression is able to duplicate the performance of the two simple ones.
- It can also test the significance of the difference between the two slopes