BS2280 - Econometrics 1

Lecture 5 - Part 2: Multiple Regression Analysis I

Dr. Yichen Zhu

Module Evaluation for BS2280 Econometrics I

SCAN ME

Use QR code above or click the following link:

`https://cloud.evasys.co.uk/aston/online.php?pswd=GYH5R`

Structure of today's lecture

Intended Learning Outcomes

- Understanding hypothesis test and predictions under multiple regression model
- Applying Hedonic Pricing model
- Interpreting and testing the goodness of fit

Multiple Regression Model

- Simple regression model:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{1}$$

- Example

$$EARNINGS_i = \beta_1 + \beta_2 S_i + u_i$$

- That is often too **simplistic**!!!
- What factors other than years of schooling can affect wages of graduates?

Multiple Regression Model

- We now extend the simple regression model by adding another variable to it, i.e. out-of-school years of experience (*EXP*)

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

- The multiple regression model allows two or more X variables in the model
- Hence, Y will depend on several X variables
- How do we symbolise these variables in our multiple regression model?

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$$

Example: Determinants of Earnings

- We used a simple regression model to analyse the impact of years of schooling on hourly wags.

$$EARNINGS_i = \beta_1 + \beta_2 S_i + u_i$$

```
> lm(EARNINGS~S, data=EAWE21)

Call:
lm(formula = EARNINGS ~ S, data = EAWE21)

Coefficients:
(Intercept)            S
     0.7647       1.2657
```

$$\widehat{EARNINGS}_i = 0.765 + 1.266 S_i$$

Example: Determinants of Earnings

- We now extend the simple regression model by adding another variable to it, i.e. out-of-school years of experience (*EXP*)

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

```
> lm(EARNINGS~S+EXP, data=EAWE21)

Call:
lm(formula = EARNINGS ~ S + EXP, data = EAWE21)

Coefficients:
(Intercept)            S          EXP
   -14.6683       1.8776       0.9833
```

$$\widehat{EARNINGS}_i = -14.668 + 1.877 S_i + 0.983 EXP_i$$

Interpretation of Coefficients

Simple Regression Model	Multiple Regression Model
$Y_i = \beta_1 + \beta_2 X_i + u_i$ $\hat{\beta}_1$ and $\hat{\beta}_2$	$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$ $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$, ..., $\hat{\beta}_k$
$\hat{\beta}_1$: Intercept	$\hat{\beta}_1$: Intercept
$\hat{\beta}_2$: A one unit change in X leads to a $\hat{\beta}_2$ unit change in Y	$\hat{\beta}_2$: On average, a one unit change in X_2 leads to a $\hat{\beta}_2$ unit change in Y, **controlling for the effects of other** X **variables**
	$\hat{\beta}_3$: On average, a one unit change in X_3 leads to a $\hat{\beta}_3$ unit change in Y, **controlling for the effects of other** X **variables** ...
	$\hat{\beta}_k$: On average, a one unit change in X_k leads to a $\hat{\beta}_k$ unit change in Y, **controlling for the effects of other** X **variables**

Interpretation of Coefficients

$$\widehat{EARNINGS}_i = -14.668 + 1.877 S_i + 0.983 EXP_i$$

- We need to attach units of measurement to X and Y as per the data set being used!!!!
- Determining whether each coefficient is statistically significant uses the same concept as with the simple regression model
- In our example:
  On average, every additional schooling year increases hourly earnings by \$1.88, controlling for the effects of other X variables
  On average, every additional year of out-of-school experience completed raises hourly earnings by \$0.98, ceteris paribus
- Controlling for the effects of other X variables or Ceteris paribus means: if two individuals, e.g. Yichen and Chiara, have the same years of out of school experience (EXP), then if Chiara completes an additional grade of schooling (S) compared to Yichen, we predict that Chiara will earn a \$1.88 higher hourly rate.

Background

	Simple Regression Model	Multiple Regression Model
Model	$Y_i = \beta_1 + \beta_2 X_i + u_i$	$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$
Estimator	$\hat{\beta}_1$ and $\hat{\beta}_2$	$\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, ..., \hat{\beta}_k$ general form: $\hat{\beta}_i$, where $i = 1, 2, 3, ..., k$
Null Hypothesis	$H_0 : \beta_1 = \beta_1^0$ $H_0 : \beta_2 = \beta_2^0$	$H_0 : \beta_1 = \beta_1^0$ $H_0 : \beta_2 = \beta_2^0, ..., H_0 : \beta_k = \beta_k^0$ general form: $H_0 : \beta_i = \beta_i^0$, where $i = 1, 2, 3, ..., k$
Alternative Hypothesis	$H_1 : \beta_1 \neq \beta_1^0$ $H_1 : \beta_2 \neq \beta_2^0$	general form: $H_1 : \beta_i \neq \beta_i^0$, where $i = 1, 2, 3, ..., k$
Test statistic	$t = \frac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)}$ $t = \frac{\hat{\beta}_2 - \beta_2^0}{s.e.(\hat{\beta}_2)}$	$t = \frac{\hat{\beta}_i - \beta_i^0}{s.e.(\hat{\beta}_i)}$, where $i = 1, 2, 3, ..., k$
Reject H_0 if	$\|t\| > t_{crit}$	$\|t\| > t_{crit}$
Degrees of Freedom	$n - k = n - 2$	$n - k$
	k is the number of regression coefficients in the model,including the intercept n: number of observations in the model, sample size	

Hypothesis Testing: Statistical Significance

	Simple Regression Model	**Multiple Regression Model**
Model	$Y_i = \beta_1 + \beta_2 X_i + u_i$	$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$
Estimator	$\hat{\beta}_1$ and $\hat{\beta}_2$	$\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, ..., \hat{\beta}_k$ general form: $\hat{\beta}_i$, where $i = 1, 2, 3, ..., k$
Null Hypothesis	$H_0 : \beta_1 = 0 \quad H_0 : \beta_2 = 0$	$H_0 : \beta_1 = 0 \quad H_0 : \beta_2 = 0, ..., H_0 : \beta_k = 0$ general form: $H_0 : \beta_i = 0$, where $i = 1, 2, 3, ..., k$
Alternative Hypothesis	$H_1 : \beta_1 \neq 0 \quad H_1 : \beta_2 \neq 0$	general form: $H_1 : \beta_i \neq 0$, where $i = 1, 2, 3, ..., k$
Test statistic	$t = \dfrac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)} = \dfrac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}$ $t = \dfrac{\hat{\beta}_2 - \beta_2^0}{s.e.(\hat{\beta}_2)} = \dfrac{\hat{\beta}_2}{s.e.(\hat{\beta}_2)}$	$t = \dfrac{\hat{\beta}_i - \beta_i^0}{s.e.(\hat{\beta}_i)} = \dfrac{\hat{\beta}_i}{s.e(\hat{\beta}_i)}$ where $i = 1, 2, 3, ..., k$
Reject H_0 if	$\lvert t \rvert > t_{crit}$	$\lvert t \rvert > t_{crit}$
Degrees of Freedom	$n - k = n - 2$	$n - k$

Hypothesis Testing: Statistical Significance

Example
Test whether coefficient of S and EXP is statistically significant.

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

	Hypothesis testing for S	**Hypothesis testing for EXP**
Null Hypothesis	$H_0 : \beta_2 = 0$	$H_0 : \beta_3 = 0$
Alternative Hypothesis	$H_1 : \beta_2 \neq 0$	$H_1 : \beta_3 \neq 0$

Hypothesis Testing: Statistical Significance

```
> summary(earnfit2)

Call:
lm(formula = EARNINGS ~ S + EXP, data = EAWE21)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.098  -6.440  -2.113   3.782  76.907 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -14.6683     4.2884  -3.420 0.000677 ***
S             1.8776     0.2237   8.392 5.01e-16 ***
EXP           0.9833     0.2098   4.686 3.60e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.13 on 497 degrees of freedom
Multiple R-squared:  0.1242,	Adjusted R-squared:  0.1207 
F-statistic: 35.24 on 2 and 497 DF,  p-value: 4.86e-15
```

$$\widehat{EARNINGS}_i = \underset{(4.2884)}{-14.6683} + \underset{(0.2237)}{1.8776}\ S_i + \underset{(0.2098)}{0.9833}\ EXP_i \tag{2}$$

Note: Standard Errors (s.e.) in brackets

Statistical Significance of β_3, use t test

$$\widehat{EARNINGS}_i = \underset{(4.2884)}{-14.6683} + \underset{(0.2237)}{1.8776}\ S_i + \underset{(0.2098)}{0.9833}\ EXP_i \tag{3}$$

Note: Standard Errors (s.e.) in brackets

1. State the null and alternative hypotheses

Null Hypothesis	$H_0 : \beta_3 = 0$
Alternative Hypothesis	$H_1 : \beta_3 \neq 0$

2. Select the significance level. Significance level $\alpha = 5\%$
3. Select and calculate the test statistics

   Do not know the population variance σ^2, so use t statistic: $t = \frac{\hat{\beta}_3 - \beta_3^0}{s.e.(\hat{\beta}_3)} = \frac{\hat{\beta}_3}{s.e.(\hat{\beta}_3)} = \frac{0.9833}{0.2098} = 4.686$
4. Set the decision rule. $n = 500$, degree of freedom = $n - k = 500 - 3 - 497$, $t_{crit,5\%} = 1.96$
5. Make statistical decisions. Make statistical decisions. $|t| = 4.686 > t_{crit,5\%} = 1.96$, reject the null $H_0 : \beta_3 = 0$. *EXP* is statistical significant, *EXP* will affect *EARNINGS*.

Statistical Significance, use *p*-values

Alternatively, we can use also the *p*-values in a regression to decide whether a coefficient is statistically significant

- ***p*-values**: probability of obtaining the corresponding t statistic as a matter of chance, if the null hypothesis $H_0 : \beta = 0$ is true.
- *p*-values < $\alpha(usually 1\%, 5\%, 10\%)$, reject the null hypothesis $H_0 : \beta = 0$

- The rules to make this decision are:
- If $p < 1\%$, variable is very significant (i.e. at the 1% level)
- If $1\% < p < 5\%$, variable is significant (i.e. at the 5% level)
- If $5\% < p < 10\%$, variable is fairly significant (i.e. at the 10% level)
- If $p > 10\%$, variable is not significant (i.e. stat. insignificant)

- If the coefficient is significant we reject the null hypothesis

Student Task

```
> summary(earnfit2)

Call:
lm(formula = EARNINGS ~ S + EXP, data = EAWE21)

Residuals:
    Min      1Q  Median      3Q     Max
-21.098  -6.440  -2.113   3.782  76.907

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.6683     4.2884  -3.420 0.000677 ***
S             1.8776     0.2237   8.392 5.01e-16 ***
EXP           0.9833     0.2098   4.686 3.60e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.13 on 497 degrees of freedom
Multiple R-squared:  0.1242,	Adjusted R-squared:  0.1207
F-statistic: 35.24 on 2 and 497 DF,  p-value: 4.86e-15
```

- The t statistic critical is $t_{crit,5\%} = 1.965$
- Are the coefficient S statistically significant?
- Interpret both the t and p-value

Predictions

- Once we have estimated a regression we can calculate predictions of Y given values of the X variables
- We can substitute values into X to generate a predicted value of Y, $\hat{Y}$
- Our model predicts the average hourly wages for different levels of education and different years of work experience.
- For example, what are the predicted earnings for an individual who had schooling years of 12 and had 10 years of out-of-school work experience
- Calculate this by plugging in these values for S and EXP:

$$\widehat{EARNINGS}_i = -14.668 + 1.877S_i + 0.983EXP_i$$

Person	**Education Years**	**Work Experience**	**Predicted Earnings**
Person 1	$S_1 = 12$	$EXP_1 = 10$	$\widehat{EARNINGS}_1 = -14.668 + 1.877 \times 12 + 0.983 \times 10$ = \$17.686

Predictions

- Once we have estimated a regression we can calculate predictions of Y given values of the X variables
- We can substitute values into X to generate a predicted value of Y, $\hat{Y}$
- Our model predicts the average hourly wages for different levels of education and different years of work experience.
- For example, what are the predicted earnings for an individual who had schooling years of 12 and had 10 years of out-of-school work experience
- Calculate this by plugging in these values for S and EXP:

$$\widehat{EARNINGS}_i = -14.668 + 1.877 S_i + 0.983 EXP_i$$

Person	Education Years	Work Experience	Predicted Earnings
Person 1	$S_1 = 12$	$EXP_1 = 10$	$\widehat{EARNINGS}_1 = -14.668 + 1.877 \times 12 + 0.983 \times 10$ = \$17.686

Predictions

- Once we have estimated a regression we can calculate predictions of Y given values of the X variables
- We can substitute values into X to generate a predicted value of Y, $\hat{Y}$
- Our model predicts the average hourly wages for different levels of education and different years of work experience.
- For example, what are the predicted earnings for an individual who had schooling years of 12 and had 10 years of out-of-school work experience
- Calculate this by plugging in these values for S and EXP:

$$\widehat{EARNINGS}_i = -14.668 + 1.877 S_i + 0.983 EXP_i$$

Person	Education Years	Work Experience	Predicted Earnings
Person 1	$S_1 = 12$	$EXP_1 = 10$	$\widehat{EARNINGS}_1 = -14.668 + 1.877 \times 12 + 0.983 \times 10 = \$17.686$

Predictions

- Once we have estimated a regression we can calculate predictions of Y given values of the X variables
- We can substitute values into X to generate a predicted value of Y, $\hat{Y}$
- Our model predicts the average hourly wages for different levels of education and different years of work experience.
- For example, what are the predicted earnings for an individual who had schooling years of 12 and had 10 years of out-of-school work experience
- Calculate this by plugging in these values for S and EXP:

$$\widehat{EARNINGS}_i = -14.668 + 1.877 S_i + 0.983 EXP_i$$

Person	Education Years	Work Experience	Predicted Earnings
Person 1	$S_1 = 12$	$EXP_1 = 10$	$\widehat{EARNINGS}_1 = -14.668 + 1.877 \times 12 + 0.983 \times 10 = \$17.686$

Predictions

- Once we have estimated a regression we can calculate predictions of Y given values of the X variables
- We can substitute values into X to generate a predicted value of Y, $\hat{Y}$
- Our model predicts the average hourly wages for different levels of education and different years of work experience.
- For example, what are the predicted earnings for an individual who had schooling years of 12 and had 10 years of out-of-school work experience
- Calculate this by plugging in these values for S and EXP:

$$\widehat{EARNINGS}_i = -14.668 + 1.877 S_i + 0.983 EXP_i$$

Person	Education Years	Work Experience	Predicted Earnings
Person 1	$S_1 = 12$	$EXP_1 = 10$	$\widehat{EARNINGS}_1 = -14.668 + 1.877 \times 12 + 0.983 \times 10$ = \$17.686

Predictions

- When carrying out predictions from a multiple regression model, we need to distinguish between an in-sample prediction vs an out-of-sample prediction
- **in-sample prediction**: if in our dataset we have an individual who has exactly these values of $S = 12$ and $EXP = 10$
- **out-of-sample prediction**: if in our dataset there is no individual who has exactly these values of $S = 12$ and $EXP = 10$
- **Out-of-sample predictions are usually less reliable**

Predictions

- When carrying out predictions from a multiple regression model, we need to distinguish between an in-sample prediction vs an out-of-sample prediction
- **in-sample prediction**: if in our dataset we have an individual who has exactly these values of $S = 12$ and $EXP = 10$
- **out-of-sample prediction**: if in our dataset there is no individual who has exactly these values of $S = 12$ and $EXP = 10$
- **Out-of-sample predictions are usually less reliable**

Predictions

- When carrying out predictions from a multiple regression model, we need to distinguish between an in-sample prediction vs an out-of-sample prediction
- **in-sample prediction**: if in our dataset we have an individual who has exactly these values of $S = 12$ and $EXP = 10$
- **out-of-sample prediction**: if in our dataset there is no individual who has exactly these values of $S = 12$ and $EXP = 10$
- **Out-of-sample predictions are usually less reliable**

Predictions

- When carrying out predictions from a multiple regression model, we need to distinguish between an in-sample prediction vs an out-of-sample prediction
- **in-sample prediction**: if in our dataset we have an individual who has exactly these values of $S = 12$ and $EXP = 10$
- **out-of-sample prediction**: if in our dataset there is no individual who has exactly these values of $S = 12$ and $EXP = 10$
- **Out-of-sample predictions are usually less reliable**

Student Task

What is the predicted *birthweight* if a mother smokes 20 *cigarettes* a day, the child is the *second* born and the family *income* is USD 30,000?

Note:
bwghtg: birthweight, grams
cigs: cigarettes smoked per day while pregnant
faminc: 1988 family income, in USD 1,000
parity: birth order of child

```
> lm(bwghtg~cigs+faminc+parity, data=bwght)

Call:
lm(formula = bwghtg ~ cigs + faminc + parity, data = bwght)

Coefficients:
(Intercept)         cigs       faminc       parity
   3237.920      -13.527        2.776       45.823
```

Hedonic Pricing Model

- Hedonic pricing supposes that a good or service has a number of characteristics that individually give it value to the purchaser.
- The market price of the good is a function, typically a linear combination, of the prices of the characteristics.

$$P_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$$

Hedonic Pricing Model: Determinants of House Prices

- Firstly, select variables (characteristics) that create value to the house

  Note:
  P_i: the price (in £000s) of the i^{th} house
  S_i: the size (in square feet) of the i^{th} house
  N_i: the neighbourhood quality of the i^{th} house (1= worst, 4= best)
  Y_i: the size of the plot of land (garden etc) around the i^{th} house (in square feet)
  A_i: the age of the i^{th} house

- We estimate the following regression:

$$P_i = \beta_1 + \beta_2 S_i + \beta_3 N_i + \beta_4 Y_i + \beta_5 A_i + u_i$$

Hedonic Pricing Model: Determinants of House Prices

$$P_i = \beta_1 + \beta_2 S_i + \beta_3 N_i + \beta_4 Y_i + \beta_5 A_i + u_i$$

```
Call:
lm(formula = P ~ S + N + Y + A, data = housing)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.788348  21.582829  -0.315   0.7548
S            0.098980   0.009774  10.127 2.40e-12 ***
N           27.017088   4.435585   6.091 4.27e-07 ***
Y            0.004510   0.001458   3.093   0.0037 **
A           -0.186016   0.245033  -0.759   0.4524
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.15 on 38 degrees of freedom
Multiple R-squared:  0.9159,    Adjusted R-squared:  0.9071
F-statistic: 103.5 on 4 and 38 DF,  p-value: < 2.2e-16
```

$$P_i = -6.788 + 0.098 S_i + 27.017 N_i + 0.004 Y_i - 0.186 A_i + u_i$$

Hedonic Pricing Model: Determinants of House Prices

$$P_i = -6.788 + 0.098S_i + 27.017N_i + 0.004Y_i - 0.186A_i + u_i$$

- **Bigger houses command higher selling prices**: On average, every additional square foot in size raises prices by £98.98 (unit of P_i is £000s), controlling for the effects of other X variables.
  This coefficient is statistically significant at the 1% level (because p-value < 1%).
- **A good neighbourhood increases the prices significantly!**: A one category better neighbourhood increases price by £27017 ((unit of P_i is £000s)), ceteris paribus.
  This coefficient is statistically significant at the 1% level (because p-value < 1%).
- **Houses with bigger gardens have a higher selling price**: On average, every additional square foot in garden size raises prices by £4.50 ((unit of P_i is £000s)), ceteris paribus.
  This coefficient is statistically significant at the 1% level (because p-value < 1%).
- **Age will not affect the selling prices.** Age has a negative impact, but it is not statistically significant (because p-value > 1%)! We cannot reject the null hypothesis of $H_0 : \beta_5 = 0$.
- **Intercept**: If the house has no size, bad neighbourhood, no garden and no age, the selling price is -£6788. but it is not statistically significant (because p-value > 1%)! We cannot reject the null hypothesis of $H_0 : \beta_1 = 0$.

Hedonic Pricing Model: Determinants of House Prices

$$P_i = -6.788 + 0.098S_i + 27.017N_i + 0.004Y_i - 0.186A_i + u_i$$

- **Bigger houses command higher selling prices**: On average, every additional square foot in size raises prices by £98.98 (unit of P_i is £000s), controlling for the effects of other X variables.
  This coefficient is statistically significant at the 1% level (because p-value < 1%).
- **A good neighbourhood increases the prices significantly!**: A one category better neighbourhood increases price by £27017 ((unit of P_i is £000s)), ceteris paribus.
  This coefficient is statistically significant at the 1% level (because p-value < 1%).
- **Houses with bigger gardens have a higher selling price**: On average, every additional square foot in garden size raises prices by £4.50 ((unit of P_i is £000s)), ceteris paribus.
  This coefficient is statistically significant at the 1% level (because p-value < 1%).
- **Age will not affect the selling prices.** Age has a negative impact, but it is not statistically significant (because p-value > 1%)! We cannot reject the null hypothesis of $H_0 : \beta_5 = 0$.
- **Intercept**: If the house has no size, bad neighbourhood, no garden and no age, the selling price is -£6788. but it is not statistically significant (because p-value > 1%)! We cannot reject the null hypothesis of $H_0 : \beta_1 = 0$.

Hedonic Pricing Model: Determinants of House Prices

$$P_i = -6.788 + 0.098 S_i + 27.017 N_i + 0.004 Y_i - 0.186 A_i + u_i$$

- **Bigger houses command higher selling prices**: On average, every additional square foot in size raises prices by £98.98 (unit of P_i is £000s), controlling for the effects of other X variables.
  This coefficient is statistically significant at the 1% level (because p-value < 1%).
- **A good neighbourhood increases the prices significantly!**: A one category better neighbourhood increases price by £27017 ((unit of P_i is £000s)), ceteris paribus.
  This coefficient is statistically significant at the 1% level (because p-value < 1%).
- **Houses with bigger gardens have a higher selling price**: On average, every additional square foot in garden size raises prices by £4.50 ((unit of P_i is £000s)), ceteris paribus.
  This coefficient is statistically significant at the 1% level (because p-value < 1%).
- **Age will not affect the selling prices.** Age has a negative impact, but it is not statistically significant (because p-value > 1%)! We cannot reject the null hypothesis of $H_0 : \beta_5 = 0$.
- **Intercept**: If the house has no size, bad neighbourhood, no garden and no age, the selling price is -£6788. but it is not statistically significant (because p-value > 1%)! We cannot reject the null hypothesis of $H_0 : \beta_1 = 0$.

Hedonic Pricing Model: Determinants of House Prices

$$P_i = -6.788 + 0.098S_i + 27.017N_i + 0.004Y_i - 0.186A_i + u_i$$

- **Bigger houses command higher selling prices**: On average, every additional square foot in size raises prices by £98.98 (unit of P_i is £000s), controlling for the effects of other X variables.
  This coefficient is statistically significant at the 1% level (because p-value < 1%).
- **A good neighbourhood increases the prices significantly!**: A one category better neighbourhood increases price by £27017 ((unit of P_i is £000s)), ceteris paribus.
  This coefficient is statistically significant at the 1% level (because p-value < 1%).
- **Houses with bigger gardens have a higher selling price**: On average, every additional square foot in garden size raises prices by £4.50 ((unit of P_i is £000s)), ceteris paribus.
  This coefficient is statistically significant at the 1% level (because p-value < 1%).
- **Age will not affect the selling prices.** Age has a negative impact, but it is not statistically significant (because p-value > 1%)! We cannot reject the null hypothesis of $H_0 : \beta_5 = 0$.
- **Intercept**: If the house has no size, bad neighbourhood, no garden and no age, the selling price is -£6788. but it is not statistically significant (because p-value > 1%)! We cannot reject the null hypothesis of $H_0 : \beta_1 = 0$.

Hedonic Pricing Model: Determinants of House Prices

$$P_i = -6.788 + 0.098S_i + 27.017N_i + 0.004Y_i - 0.186A_i + u_i$$

- **Bigger houses command higher selling prices**: On average, every additional square foot in size raises prices by £98.98 (unit of P_i is £000s), controlling for the effects of other X variables.
  This coefficient is statistically significant at the 1% level (because p-value < 1%).
- **A good neighbourhood increases the prices significantly!**: A one category better neighbourhood increases price by £27017 ((unit of P_i is £000s)), ceteris paribus.
  This coefficient is statistically significant at the 1% level (because p-value < 1%).
- **Houses with bigger gardens have a higher selling price**: On average, every additional square foot in garden size raises prices by £4.50 ((unit of P_i is £000s)), ceteris paribus.
  This coefficient is statistically significant at the 1% level (because p-value < 1%).
- **Age will not affect the selling prices.** Age has a negative impact, but it is not statistically significant (because p-value > 1%)! We cannot reject the null hypothesis of $H_0 : \beta_5 = 0$.
- **Intercept**: If the house has no size, bad neighbourhood, no garden and no age, the selling price is -£6788. but it is not statistically significant (because p-value > 1%)! We cannot reject the null hypothesis of $H_0 : \beta_1 = 0$.

Goodness of Fit R^2

- To evaluate the explanatory power of the estimated model, we can use R^2 again
- In the multiple regression model, it tells us how much of the variation in Y can be explained with the variation in all X variables

$$R^2 = \frac{\textit{Explained Varations in all } X}{\textit{Total Varations in } Y} = \frac{ESS}{TSS}$$

Goodness of Fit R^2

Simple regression model:

$$\widehat{EARNINGS}_i = 0.765 + 1.266 S_i$$

Multiple regression model:

$$\widehat{EARNINGS}_i = -14.668 + 1.877 S_i + 0.983 EXP_i$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.7647     2.8038   0.273    0.785
S             1.2657     0.1855   6.824 2.58e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.36 on 498 degrees of freedom
Multiple R-squared:  0.08551       Adjusted R-squared:  0.08368
F-statistic: 46.57 on 1 and 498 DF,  p-value: 2.579e-11

-------------------------------------------------------------------

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.6683     4.2884  -3.420 0.000677 ***
S             1.8776     0.2237   8.392 5.01e-16 ***
EXP           0.9833     0.2098   4.686 3.60e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.13 on 497 degrees of freedom
Multiple R-squared:  0.1242        Adjusted R-squared:  0.1207
```

Goodness of Fit R^2

- The R^2 from the multiple regression model is higher!
- This is a property of R^2: More X variables on the right-hand side of a regression model, R^2 has a tendency to increase (or at least not decrease)

Problem: Given that the R^2 will be higher in a model with more X variables, is it fair to say that the model with S and EXP is better than the model with only S?

Adjusted R^2: $\bar{R}^2$

- The Adjusted R^2, denoted as $\bar{R}^2$ makes an attempt to compare different numbers of X variables regression models, when it comes to model fit:

$$\bar{R}^2 = R^2 - \frac{k-1}{n-k}(1 - R^2)$$

  n: number of observations in the model
  k: the number of regression coefficients in the model, including the intercept term

- As k increases, so does the negative adjustment
- The negative adjustment is like a penalty imposed on the R^2 for increasing the number of X variables in the model

Limitations of $\bar{R}^2$ and R^2

- $\bar{R}^2$ is useful when comparing across two models with potentially different number of X variables and different number of observations
- Measure has limitations and is therefore not widely used as a diagnostic statistics
- If we want to understand how good a model is in terms of its fit, we still look at its R^2 value
- Do not use R^2 as your only tool to evaluate the strength of your model!!!
- Frequently, misspecification can cause a misleading high R^2.

Student Task

We estimate two regressions.

1. Regression of number of police officers on crime
2. Regression of number of police officers on crime, per capita income and population

Calculate the Adjusted R^2 for both regressions Using Adjusted R^2, explain if adding more variables in regression two improved the goodness of fit of the model.

$$\bar{R}^2 = R^2 - \frac{k-1}{n-k}(1 - R^2)$$

Student Task

1. Regression of number of police officers on crime
2. Regression of number of police officers on crime, per capita income and population

```
Call:
lm(formula = officers ~ crimes, data = crime)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.418291  75.587257  -0.072    0.943
crimes       0.023804   0.001611  14.777   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 298.9 on 44 degrees of freedom
Multiple R-squared:  0.8323,      Adjusted R-squared:
F-statistic: 218.4 on 1 and 44 DF,  p-value: < 2.2e-16
-----------------------------------------------------------------
Call:
lm(formula = officers ~ crimes + pcinc + pop, data = crime)

Coefficients:
               Estimate   Std. Error t value Pr(>|t|)
(Intercept) 586.5479917 270.9496169   2.165   0.0361 *
crimes        0.0137920   0.0040023   3.446   0.0013 **
pcinc        -0.0934626   0.0366929  -2.547   0.0146 *
pop           0.0011855   0.0004426   2.678   0.0105 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 270.3 on 42 degrees of freedom
Multiple R-squared:  0.8691,      Adjusted R-squared:
F-statistic: 92.94 on 3 and 42 DF,  p-value: < 2.2e-16
```

```
> nobs(crimefit)
[1] 46
```

What to do next:

- Attempt homework 5
- Read chapter 3.1 - 3.3, 3.5 of Dougherty