BS2280 - Econometrics 1

Lecture 3 - Part 2: Bias, Efficiency and Interpreting Coefficients

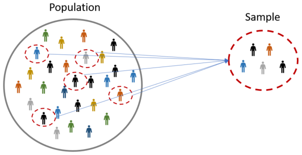
by Dr Yichen Zhu

Outline

Intended Learning Outcomes

- Interpreting coefficients of regressions
- Deriving and understanding R^2

Review: Simple Linear Regression Model - Population vs. Sample



Population	**Sample**
$Y_i = \beta_1 + \beta_2 X_i + u_i$	$Y_i = \hat{Y}_i + \hat{u}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$
parameters β_1 and β_2	coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$
u_i disturbance term	$\hat{u}_i$ residual

OLS estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ have certain desirable properties, but these properties rely on a set of assumptions we need to make!!!

Review: OLS Assumption 1

Model is linear in parameters and correctly specified

Which one describes Assumption 1?

A) $Y_i = \beta_1 + \beta_2 X_i + u_i$
B) $Y_i = \beta_1 X_i^{\beta_2} + u_i$

Review: OLS Assumption 2

There is some variation in the X variable

Which one describes Assumption 2?

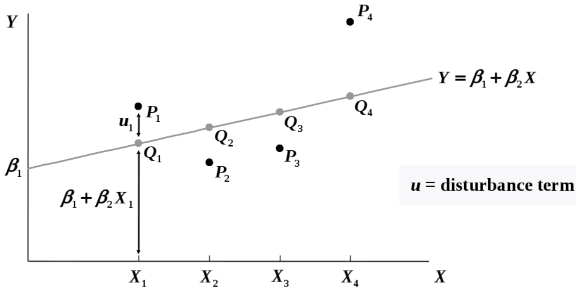
A) X can be constant
B) X cannot be constant

Review: OLS Assumption 3

Disturbance term has zero expectation

Which one describes Assumption 3?

A) $E(u_i) = 0$ for all i or $E(u_i|X_1, X_2, ..., X_n) = 0$ for all i
B) $E(u_i) \neq 0$ for all i or $E(u_i|X_1, X_2, ..., X_n) \neq 0$ for all i



Review: OLS Assumption 4

The disturbance term is homoscedastic

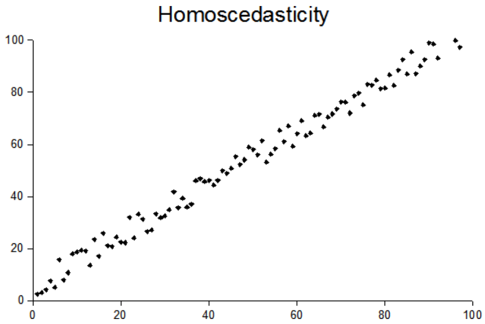
Which one describes Assumption 4?

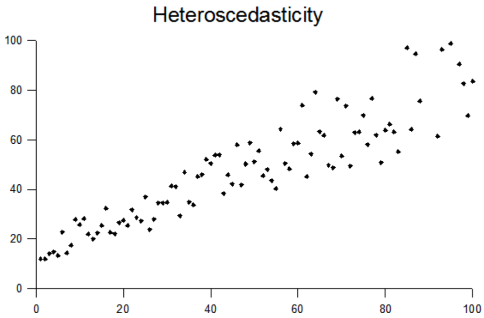
A) We assume that the error term has a constant variance, $E(u_i^2) = \sigma_u^2$ for all i

B) We assume that the error term has a non-constant variance, $E(u_i^2) \neq \sigma_u^2$ for all i

Review: OLS Assumption 4

The disturbance term is homoscedastic



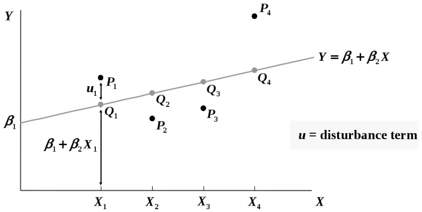


Review: OLS Assumption 5

Values of disturbance term have independent distributions

Which one describes Assumption 5?

A) We assume that the error terms are absolutely **independent** of each other. u_i is independently distributed from u_j for all $i \neq j$
B) We assume that the error terms are absolutely **dependent** of each other. u_i is dependently distributed from u_j for all $i \neq j$



Review: OLS Assumption 6

The disturbance term has a normal distribution

Which one describes Assumption 6?

A) We assume the error terms come from a **normal** distribution
B) We assume the error terms come from a **T** distribution

Check the 'quality' of OLS estimated $\hat{\beta}_1$ and $\hat{\beta}_2$

- Unbiasedness
- Consistency
- Efficiency / Precision

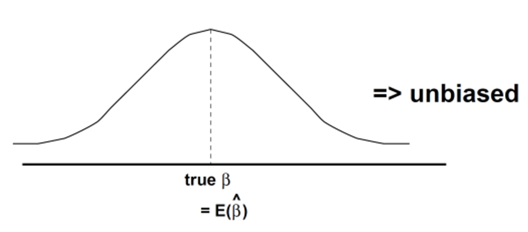
- If all our OLS assumptions hold, we get Best (most efficient) Linear Unbiased Estimators (BLUE) $\hat{\beta}_1$ and $\hat{\beta}_2$

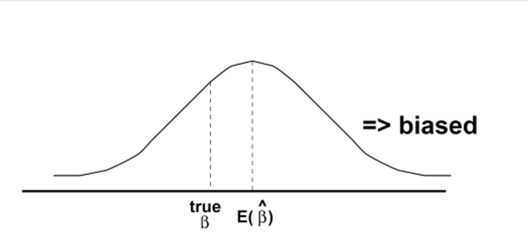
Unbiasedness

Unbiasedness

An estimator is unbiased when expected value equals population value. E.g. for $\hat{\beta}_2$: $E(\hat{\beta}_2) = \beta_2$, otherwise, the estimator is said to be biased

- This will be the case for OLS because we have Assumptions 1,2 and 3 $E(u_i) = 0$
- See Dougherty chapter 2 for mathematical prove

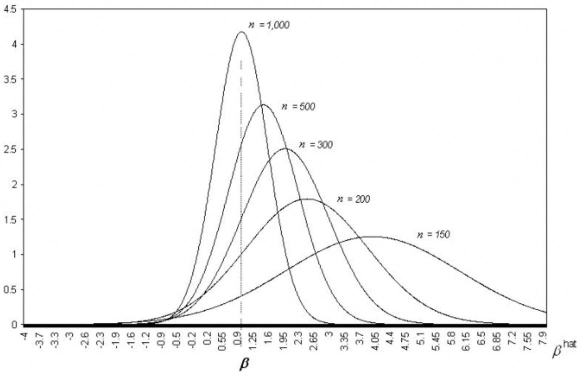




Consistency

Consistency

The larger the sample the closer our estimators ($\hat{\beta}_1$ and $\hat{\beta}_2$) should be to the population value (β_1 and β_2)

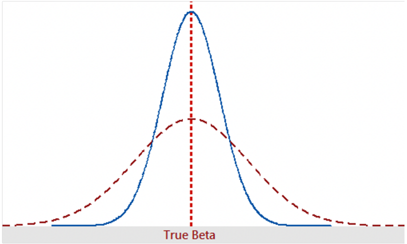


Efficiency / Precision

Efficiency / Precision

For the OLS estimator to be best (most efficient) it needs to have a lower variance than all the other estimators within the class.

- Recall that our interest lies in knowing if $\hat{\beta}_1$ and $\hat{\beta}_2$ are close estimates of true but unknown β_1 and β_2
- An estimator has higher precision if its variance is small compared to one with a larger variance



Efficiency / Precision

- Variances of $\hat{\beta}_1$ and $\hat{\beta}_2$, i.e. $\sigma^2_{\hat{\beta}_1}$ and $\sigma^2_{\hat{\beta}_2}$
- Lower variance $\sigma^2_{\hat{\beta}_1}$, our estimates are more efficient / precise

- What factors will affect the value of variance: $\sigma^2_{\hat{\beta}_1}$ and $\sigma^2_{\hat{\beta}_2}$?
- We will use $\sigma^2_{\hat{\beta}_2}$ to explain, but same intuition works for $\sigma^2_{\hat{\beta}_1}$

Efficiency / Precision

- Variance of β_2 is:

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sigma^2_{u_i}}{n\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- we define MSD(X) as:

$$MSD(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- Therefore

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X)}$$

Efficiency / Precision

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X)}$$

Efficiency will depend on three factors:

- $\sigma^2_{u_i}$: Larger residuals' variance reduce the precision of estimates (see earlier example above)
- n: Larger the number of observations (n), the smaller $\sigma^2_{\hat{\beta}_2}$
- $MSD(X)$: Larger MSD(X) leads to smaller $\sigma^2_{\hat{\beta}_2}$
  - Regression analysis - understand how variations in X_i explain variations in Y_i . Variations in u_i pick up the rest
  - Low variations in MSD(X) means low variations in X_i $\longrightarrow$ more variations from u_i $\longrightarrow$ higher variance (lower precision) in $\hat{\beta}_2$

Efficiency / Precision

```
> summary(earnfit)
Call:
lm(formula = EARNINGS ~ S, data = EAWE21.simple)

Residuals:
Min       1Q   Median       3Q      Max
-20.079  -6.726  -2.203   3.451  79.037

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
```

BLUE!!!

- If all our 6 assumptions mentioned above hold, then our OLS estimator is BLUE
- Best Linear Unbiased Estimator
- The Best property is particular of interest: the Gauss-Markov Theorem states that provided the assumptions hold, OLS estimators will have the lowest variance amongst all possible estimators
(proof of this in Dougherty textbook Appendix 2.1)
- An estimator with low variance means it comes from a sampling distribution where most of the values are concentrated around the true but unknown population value of the characteristic it is estimating

BLUE!!!

- If all our 6 assumptions mentioned above hold, then our OLS estimator is BLUE
- Best Linear Unbiased Estimator
- The Best property is particular of interest: the Gauss-Markov Theorem states that provided the assumptions hold, OLS estimators will have the lowest variance amongst all possible estimators
  (proof of this in Dougherty textbook Appendix 2.1)
- An estimator with low variance means it comes from a sampling distribution where most of the values are concentrated around the true but unknown population value of the characteristic it is estimating

BLUE!!!

- If all our 6 assumptions mentioned above hold, then our OLS estimator is BLUE
- Best Linear Unbiased Estimator
- The Best property is particular of interest: the Gauss-Markov Theorem states that provided the assumptions hold, OLS estimators will have the lowest variance amongst all possible estimators
(proof of this in Dougherty textbook Appendix 2.1)
- An estimator with low variance means it comes from a sampling distribution where most of the values are concentrated around the true but unknown population value of the characteristic it is estimating

BLUE!!!

- If all our 6 assumptions mentioned above hold, then our OLS estimator is BLUE
- Best Linear Unbiased Estimator
- The Best property is particular of interest: the Gauss-Markov Theorem states that provided the assumptions hold, OLS estimators will have the lowest variance amongst all possible estimators
  (proof of this in Dougherty textbook Appendix 2.1)
- An estimator with low variance means it comes from a sampling distribution where most of the values are concentrated around the true but unknown population value of the characteristic it is estimating

Interpretation of Coefficients I

- Last week we analysed the relationship between education and income.
- We used OLS regression to estimate β_1 and β_2

```
> earnfit <- lm(EARNINGS~S, data=EAWE21.simple)
> earnfit

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

Coefficients:
(Intercept)            S
0.7647       1.2657
```

Regression equation:

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

Interpretation of Coefficients II: β_2

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

- What do results actually mean?
- Interpretation depends on units of variables!
- To interpret β_2 we think at the margin: A one unit change in X leads to a β_2 unit change in Y.
- In our example:
  Increasing education by 1 year will increase hourly earnings by \$1.27
  Increasing education by 3 years will increase hourly earnings by $3 \times \$1.27$

Interpretation of Coefficients II: β_2

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

- What do results actually mean?
- Interpretation depends on units of variables!
- To interpret β_2 we think at the margin: A one unit change in X leads to a β_2 unit change in Y.
- In our example:
  Increasing education by 1 year will increase hourly earnings by \$1.27
  Increasing education by 3 years will increase hourly earnings by $3 \times \$1.27$

Interpretation of Coefficients II: β_2

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

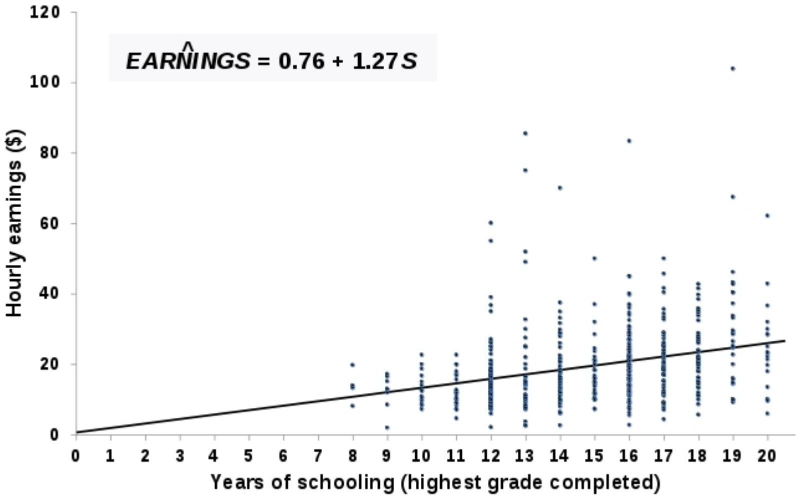
- What do results actually mean?
- Interpretation depends on units of variables!
- To interpret β_2 we think at the margin: A one unit change in X leads to a β_2 unit change in Y.
- In our example:
  Increasing education by 1 year will increase hourly earnings by \$1.27
  Increasing education by 3 years will increase hourly earnings by $3 \times \$1.27$

Interpretation of Coefficients II: β_2

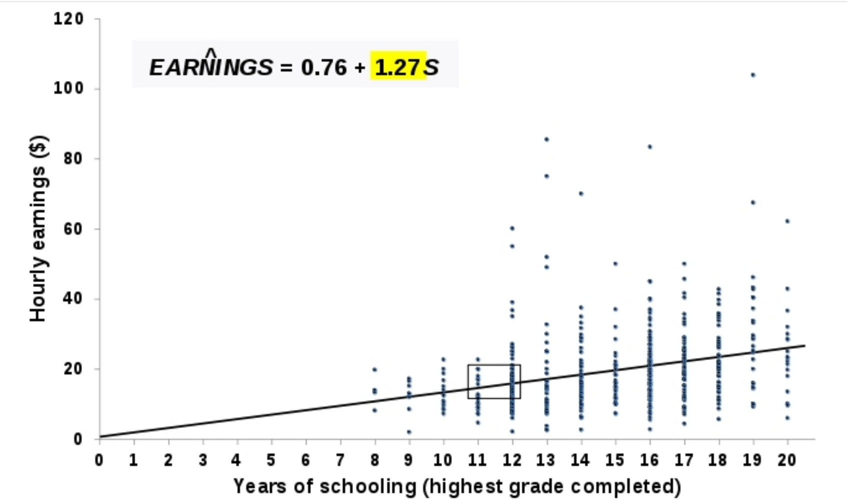
$$\widehat{EARNINGS}_i = 0.765 + 1.266 S_i$$

- What do results actually mean?
- Interpretation depends on units of variables!
- To interpret β_2 we think at the margin: A one unit change in X leads to a β_2 unit change in Y.
- In our example:
  Increasing education by 1 year will increase hourly earnings by \$1.27
  Increasing education by 3 years will increase hourly earnings by $3 \times \$1.27$

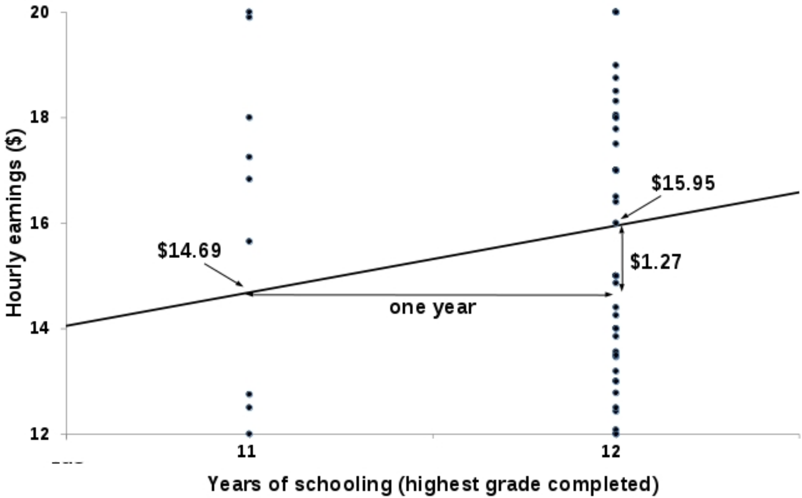
Interpretation of Coefficients II: β_2



Interpretation of Coefficients II: β_2

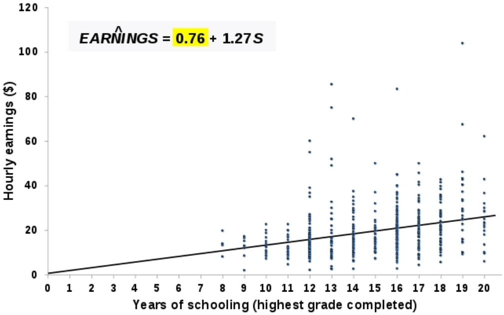


Interpretation of Coefficients II: β_2



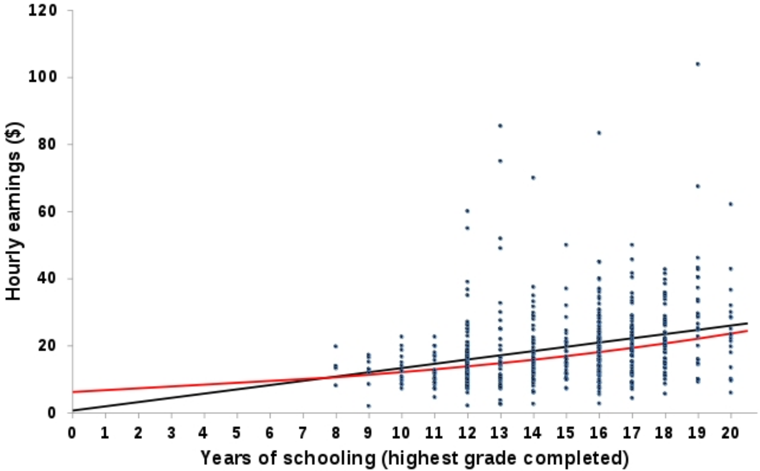
Interpretation of Coefficients III: β_1

- β_1: intercept, it shows value of Y when X = 0.
- A person with no education would have an hourly salary of $0.76.
- Is such a low wage realistic? No!
- Statistically, it makes sense. But, economically, it does not make sense.
- Solution: Limit interpretation to a specific range



Interpretation of Coefficients III: β_1

- Alternative solution: use non-linear specification. More on that later...



Student Task:

- We estimate the impact of temperature on ice cream sales over a period of 21 days.
- Temperature is measures in degree Celsius and ice cream sales in thousand GBP
- The results are: $\widehat{Sales}_t = -0.184 + 0.111\, Temp_t$
- Interpret carefully the sign and the size of both the β_1 and β_2 coefficients.

Prediction

- We can use our regression line for predictions!
- Our model predicts the average hourly wages for different levels of education.
- For example, what is the predict level of average wages?

Person	Education Years	Predicted Earnings
Person 1	$S_1 = 0$	$\widehat{EARNINGS}_1 = 0.76 + 1.27 \times 0 = \$0.76$
Person 3	$S_3 = 5$	$\widehat{EARNINGS}_3 = 0.76 + 1.27 \times 5 = \$7.11$
Person 7	$S_7 = 10$	$\widehat{EARNINGS}_7 = 0.76 + 1.27 \times 10 = \$13.46$

- If we are only interested in the accuracy of the prediction, rather than the unbiasedness and efficiency of the estimates, we can relax our OLS assumptions.

Prediction

- We can use our regression line for predictions!
- Our model predicts the average hourly wages for different levels of education.
- For example, what is the predict level of average wages?

Person	Education Years	Predicted Earnings
Person 1	$S_1 = 0$	$\widehat{EARNINGS}_1 = 0.76 + 1.27 \times 0 = \$0.76$
Person 3	$S_3 = 5$	$\widehat{EARNINGS}_3 = 0.76 + 1.27 \times 5 = \$7.11$
Person 7	$S_7 = 10$	$\widehat{EARNINGS}_7 = 0.76 + 1.27 \times 10 = \$13.46$

- If we are only interested in the accuracy of the prediction, rather than the unbiasedness and efficiency of the estimates, we can relax our OLS assumptions.

Prediction

- We can use our regression line for predictions!
- Our model predicts the average hourly wages for different levels of education.
- For example, what is the predict level of average wages?

Person	Education Years	Predicted Earnings
Person 1	$S_1 = 0$	$\widehat{EARNINGS}_1 = 0.76 + 1.27 \times 0 = \$0.76$
Person 3	$S_3 = 5$	$\widehat{EARNINGS}_3 = 0.76 + 1.27 \times 5 = \$7.11$
Person 7	$S_7 = 10$	$\widehat{EARNINGS}_7 = 0.76 + 1.27 \times 10 = \$13.46$

- If we are only interested in the accuracy of the prediction, rather than the unbiasedness and efficiency of the estimates, we can relax our OLS assumptions.

Prediction

- We can use our regression line for predictions!
- Our model predicts the average hourly wages for different levels of education.
- For example, what is the predict level of average wages?

Person	**Education Years**	**Predicted Earnings**
Person 1	$S_1 = 0$	$\widehat{EARNINGS}_1 = 0.76 + 1.27 \times 0 = \$0.76$
Person 3	$S_3 = 5$	$\widehat{EARNINGS}_3 = 0.76 + 1.27 \times 5 = \$7.11$
Person 7	$S_7 = 10$	$\widehat{EARNINGS}_7 = 0.76 + 1.27 \times 10 = \$13.46$

- If we are only interested in the accuracy of the prediction, rather than the unbiasedness and efficiency of the estimates, we can relax our OLS assumptions.

Student Task

- Referring back to our ice-cream sale example
- The results were: $\widehat{Sales}_t = -0.184 + 0.111\,Temp_t$
- What are the predicted ice-cream sales when the temperature is 20 ℃ or 35 ℃?

Student Task

- Referring back to our ice-cream sale example
- The results were: $\widehat{Sales_t} = -0.184 + 0.111\,Temp_t$
- What are the predicted ice-cream sales when the temperature is 20℃ or 35℃?

Student Task

- Referring back to our ice-cream sale example
- The results were: $\widehat{Sales}_t = -0.184 + 0.111\,Temp_t$
- What are the predicted ice-cream sales when the temperature is 20℃ or 35℃?

Goodness of Fit

- Regression analysis aims to understand whether observed variations in Y can be explained by observed variations in X.
- Variance of a variable is a measure of its dispersion, i.e. how far away the observations are from their mean value.
- Dispersion in the Y variable is given by the sum of squared deviation from the mean (Total Sum of Squares, TSS)

$$TSS = \sum (Y_i - \bar{Y})^2$$

- How much of this variation can be explained with our model?

Goodness of Fit

- Regression analysis aims to understand whether observed variations in Y can be explained by observed variations in X.
- Variance of a variable is a measure of its dispersion, i.e. how far away the observations are from their mean value.
- Dispersion in the Y variable is given by the sum of squared deviation from the mean (Total Sum of Squares, TSS)

$$TSS = \sum (Y_i - \bar{Y})^2$$

- How much of this variation can be explained with our model?

Goodness of Fit

- Regression analysis aims to understand whether observed variations in Y can be explained by observed variations in X.
- Variance of a variable is a measure of its dispersion, i.e. how far away the observations are from their mean value.
- Dispersion in the Y variable is given by the sum of squared deviation from the mean (Total Sum of Squares, TSS)

$$TSS = \sum (Y_i - \bar{Y})^2$$

- How much of this variation can be explained with our model?

Goodness of Fit

- Regression analysis aims to understand whether observed variations in Y can be explained by observed variations in X.
- Variance of a variable is a measure of its dispersion, i.e. how far away the observations are from their mean value.
- Dispersion in the Y variable is given by the sum of squared deviation from the mean (Total Sum of Squares, TSS)

$$TSS = \sum (Y_i - \bar{Y})^2$$

- How much of this variation can be explained with our model?

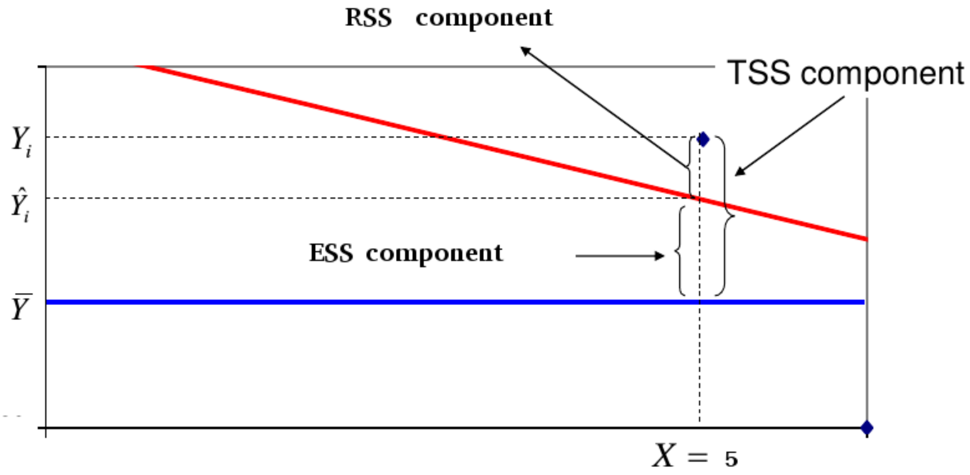
Goodness of Fit

$$\begin{array}{cccc} Y_i & = \beta_1 + \beta_2 & X_i & + & u_i \\ \text{Total variations, variations in } Y & & \text{Explained variations, variations in } X & & \text{Residual variations, variations in } u_i \\ \text{Total Sum of Squares, TSS} & & \text{Explained sum of squares, ESS} & & \text{Residual sum of squares, RSS} \\ TSS = \sum (Y_i - \bar{Y})^2 & & ESS = \sum (\hat{Y}_i - \bar{Y})^2 & & RSS = \sum (Y_i - \hat{Y}_i)^2 = \sum u_i^2 \end{array} \tag{1}$$

- Total variation in Y can be decomposed into variations arising from the X variable and variations arising from the residuals u_i

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$TSS = ESS + RSS$$


RSS component
TSS component
ESS component
Y_i
$\hat{Y}_i$
$\bar{Y}$
$X = 5$

Calculating R^2

- Create a measure of the proportion of ESS in TSS: the R^2 - the coefficient of determination

$$R^2 = \frac{Explained\ Varations\ in\ X}{Total\ Varations\ in\ Y} = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- $0 \leq R^2 \leq 1$: The R^2 is bound between 0 and 1
- $R^2 = 0$: variations in X cannot explain any variations in Y
- $R^2 = 1$: variations in X can fully explain variations in Y
- A high R^2 means that the independent variable X is good at predicting Y

- **Example**: $R^2 = 0.75$, it means that 75% of the variations in Y can be explained by variations in the X variable.
- Remaining 25% comes from unobservables through the residual term

Calculating R^2

- Create a measure of the proportion of ESS in TSS: the R^2 - the coefficient of determination

$$R^2 = \frac{\text{Explained Varations in X}}{\text{Total Varations in Y}} = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- $0 \leq R^2 \leq 1$: The R^2 is bound between 0 and 1
- $R^2 = 0$: variations in X cannot explain any variations in Y
- $R^2 = 1$: variations in X can fully explain variations in Y
- A high R^2 means that the independent variable X is good at predicting Y

- **Example**: $R^2 = 0.75$, it means that 75% of the variations in Y can be explained by variations in the X variable.
- Remaining 25% comes from unobservables through the residual term

Calculating R^2

- Create a measure of the proportion of ESS in TSS: the R^2 - the coefficient of determination

$$R^2 = \frac{Explained\ Varations\ in\ X}{Total\ Varations\ in\ Y} = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- $0 \leq R^2 \leq 1$: The R^2 is bound between 0 and 1
- $R^2 = 0$: variations in X cannot explain any variations in Y
- $R^2 = 1$: variations in X can fully explain variations in Y
- A high R^2 means that the independent variable X is good at predicting Y

- **Example**: $R^2 = 0.75$, it means that 75% of the variations in Y can be explained by variations in the X variable.
- Remaining 25% comes from unobservables through the residual term

Calculating R^2

- Create a measure of the proportion of ESS in TSS: the R^2 - the coefficient of determination

$$R^2 = \frac{Explained\ Varations\ in\ X}{Total\ Varations\ in\ Y} = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- $0 \leq R^2 \leq 1$: The R^2 is bound between 0 and 1
- $R^2 = 0$: variations in X cannot explain any variations in Y
- $R^2 = 1$: variations in X can fully explain variations in Y
- A high R^2 means that the independent variable X is good at predicting Y

- **Example**: $R^2 = 0.75$, it means that 75% of the variations in Y can be explained by variations in the X variable.
- Remaining 25% comes from unobservables through the residual term

Calculating R^2

- Create a measure of the proportion of ESS in TSS: the R^2 - the coefficient of determination

$$R^2 = \frac{\textit{Explained Varations in X}}{\textit{Total Varations in Y}} = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- $0 \leq R^2 \leq 1$: The R^2 is bound between 0 and 1
- $R^2 = 0$: variations in X cannot explain any variations in Y
- $R^2 = 1$: variations in X can fully explain variations in Y
- A high R^2 means that the independent variable X is good at predicting Y

- **Example**: $R^2 = 0.75$, it means that 75% of the variations in Y can be explained by variations in the X variable.
- Remaining 25% comes from unobservables through the residual term

Calculating R^2

- Create a measure of the proportion of ESS in TSS: the R^2 - the coefficient of determination

$$R^2 = \frac{\textit{Explained Varations in X}}{\textit{Total Varations in Y}} = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- $0 \leq R^2 \leq 1$: The R^2 is bound between 0 and 1
- $R^2 = 0$: variations in X cannot explain any variations in Y
- $R^2 = 1$: variations in X can fully explain variations in Y
- A high R^2 means that the independent variable X is good at predicting Y

- **Example**: $R^2 = 0.75$, it means that 75% of the variations in Y can be explained by variations in the X variable.
- Remaining 25% comes from unobservables through the residual term

Calculating R^2

- Create a measure of the proportion of ESS in TSS: the R^2 - the coefficient of determination

$$R^2 = \frac{\textit{Explained Varations in X}}{\textit{Total Varations in Y}} = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- $0 \leq R^2 \leq 1$: The R^2 is bound between 0 and 1
- $R^2 = 0$: variations in X cannot explain any variations in Y
- $R^2 = 1$: variations in X can fully explain variations in Y
- A high R^2 means that the independent variable X is good at predicting Y

- **Example**: $R^2 = 0.75$, it means that 75% of the variations in Y can be explained by variations in the X variable.
- Remaining 25% comes from unobservables through the residual term

```
> summary(earnfit)
Call:
lm(formula = EARNINGS ~ S, data = EAWE21.simple)

Residuals:
    Min      1Q  Median      3Q     Max
-20.079  -6.726  -2.203   3.451  79.037

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
```

$R^2 = 0.0855$, only 8.55% of the **variations in earnings** can be explained by **variations in schooling.** 91.45% of the variations in earnings is left unexplained.

Calculating R^2

- The `anova` command allows us to identify **ESS** and **RSS**:

```
> anova(earnfit) # use regression output as argument

Analysis of Variance Table

Response: EARNINGS
           Df  Sum Sq Mean Sq F value    Pr(>F)
S           1    6014  6014.0  46.568 2.579e-11 ***
Residuals 498   64315   129.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $ESS = 6014$
- $RSS = 64315$
- so $TSS = ESS + RSS = 70329$
- $R^2 = \dfrac{ESS}{TSS} = \dfrac{6014}{70329} = 0.0855$

Student Task

- Using the simplified R output below, interpret the coefficients and comment on the overall fit of the model.

```
Call:
lm(formula = lexp ~ gnppc, data = lifeexp)

Coefficients:
               Estimate  Std. Error t value Pr(>|t|)
(Intercept) 69.44836458  0.54798213 126.735  < 2e-16 ***
gnppc        0.00032342  0.00004012   8.061 3.45e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.36 on 61 degrees of freedom
(5 observations deleted due to missingness)
Multiple R-squared:  0.5158,  Adjusted R-squared:  0.5079
F-statistic: 64.98 on 1 and 61 DF,  p-value: 3.452e-11
```

- where *lexp* is the life expectancy at birth in years and *gnppc* the gross national product per capita measured in USD.

Student Task

- Using the Anova output below, calculate R^2 and explain how well the weight (wt, Weight in 1000lbs) of a car can explain the fuel consumption (mpg) of a car.

```
Response: mpg
Df Sum Sq Mean Sq F value    Pr(>F)
wt         1 847.73  847.73  91.375 1.294e-10 ***
Residuals 30 278.32    9.28
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What to do next:

- Attempt homework 3
- Revise basic R commands from R Workshop 1
- Read chapter 2.1 - 2.7 of Dougherty