Regression with One Regressor: Hypothesis Tests and Confidence Intervals

Ming-Ching Luoh

2022.2.21.

Testing Hypotheses

Confidence intervals

Regression When X is Binary

Heteroskedasticity and Homoskedasticity

Weighted Least Squares

Summary and Assessment

Testing Hypotheses About One of the Regression Coefficients

Suppose a skeptic suggests that reducing the number of students in a class has no effect on learning or, specifically, test scores. The skeptic thus asserts the hypothesis,

$$H_0 : \beta_1 = 0$$

We wish to test this hypothesis using data— reach a tentative conclusion whether it is correct or not.

Null hypothesis and **two-sided** alternative:

$$H_0\colon \beta_1 = 0 \;\; vs. \;\; H_1\colon \beta_1 \neq 0$$

or, more generally,

$$H_0\colon \beta_1 = \beta_{1,0} \;\; vs. \;\; H_1\colon \beta_1 \neq \beta_{1,0}$$

Null hypothesis and **one-sided** alternative:

$$H_0\colon \beta_1 = \beta_{1,0} \;\; vs. \;\; H_1\colon \beta_1 < \beta_{1,0}$$

In economics, it is almost always possible to come up with stories in which an effect could “go either way," so it is standard to focus on two-sided alternatives.

In general,

$$t = \frac{\text{estimator} - \text{ hypothesized value}}{\text{S.E. of the estimator}}$$

where the $S.E.$ of the estimator is the square root of an estimator of the variance of the estimator.
Applied to a hypothesis about β_1:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

where $\beta_{1,0}$ is the hypothesized value of β_1.

Formula for $SE(\hat{\beta}_1)$

Recall the expression for the variance of $\hat{\beta}_1$ (large n):

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\mathrm{Var}\left((X_i - \bar{X})u_i\right)}{n(\sigma_X^2)^2} = \frac{\sigma_v^2}{n\sigma_X^4}$$

where $v_i = (X_i - \bar{X})u_i$.

Estimator of the variance of $\hat{\beta}_1$ is

$$\begin{aligned}\hat{\sigma}^2_{\hat{\beta}_1} &= \frac{1}{n} \times \frac{\text{estimator of } \sigma_v^2}{(\text{estimator of } \sigma_X^2)^2} \\ &= \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2\hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}\end{aligned}$$

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$

where $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ is the residual.

- There is no reason to memorize this.
- It is computed automatically by regression software.
- $SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}$ is reported by regression software.
- It is less complicated than it seems. The numerator estimates $\mathrm{Var}(v)$, the denominator estimates $\mathrm{Var}(X)$.

The calculation of the t-statistic:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}}$$

- Reject at 5% significance level if $|t| > 1.96$.
- p-value is $p = \Pr\left(|t| > |t^{act}|\right)$=probability in tails outside $|t^{act}|$.
- Both the previous statements are based on large-n approximation.

Example: Test Scores and STR, California data

Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$

Regression software reports the standard errors:

$$SE(\hat{\beta}_0) = 10.4, SE(\hat{\beta}_1) = 0.52$$

t-statistic testing $\beta_{1,0} = 0$ is $\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \frac{-2.28-0}{0.52} = -4.38$

- The 1% two-sided significance level is 2.58, so we reject the null at 1% significance level.
- Alternatively, we can compute the p-value.

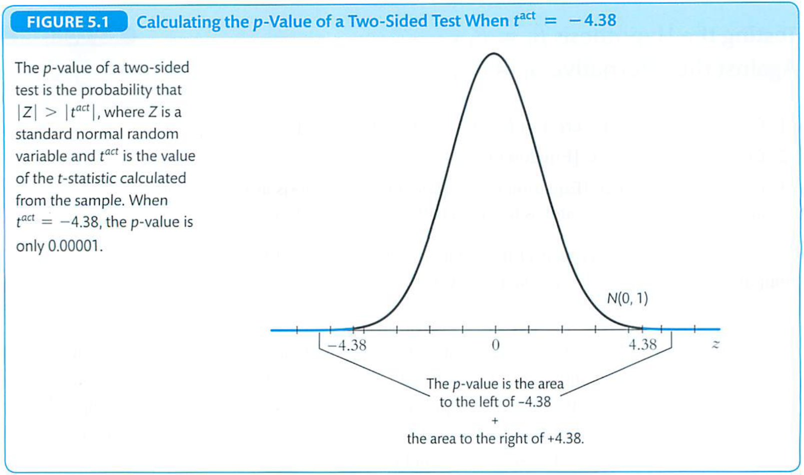


FIGURE 5.1 Calculating the p-Value of a Two-Sided Test When $t^{act} = -4.38$

The p-value of a two-sided test is the probability that $|Z| > |t^{act}|$, where Z is a standard normal random variable and t^{act} is the value of the t-statistic calculated from the sample. When $t^{act} = -4.38$, the p-value is only 0.00001.

The p-value is the area to the left of −4.38
+
the area to the right of +4.38.

The p-value based on the large-n standard normal approximation to the t-statistic is 0.00001.

Confidence intervals for a Regression Coefficient

In general, if the sample distribution of an estimator is nomal for large n, then a 95% confidence interval can be constructed as estimator ±1.96× standard error, that is

$$\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)$$

Example: Test Scores and STR, California data

Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$

Regression software reports the standard errors.

$$SE(\hat{\beta}_0) = 10.4, SE(\hat{\beta}_1) = 0.52$$

95% confidence interval for $\hat{\beta}_1$:

$$\begin{aligned} \hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1) &= \{-2.28 \pm 1.96 \times 0.52\} \\ &= (-3.30, -1.26) \end{aligned}$$

Equivalent statements:

- The 95% confidence interval does not include zero.
- The hypothesis $\beta_1 = 0$ is rejected at the 5% level.

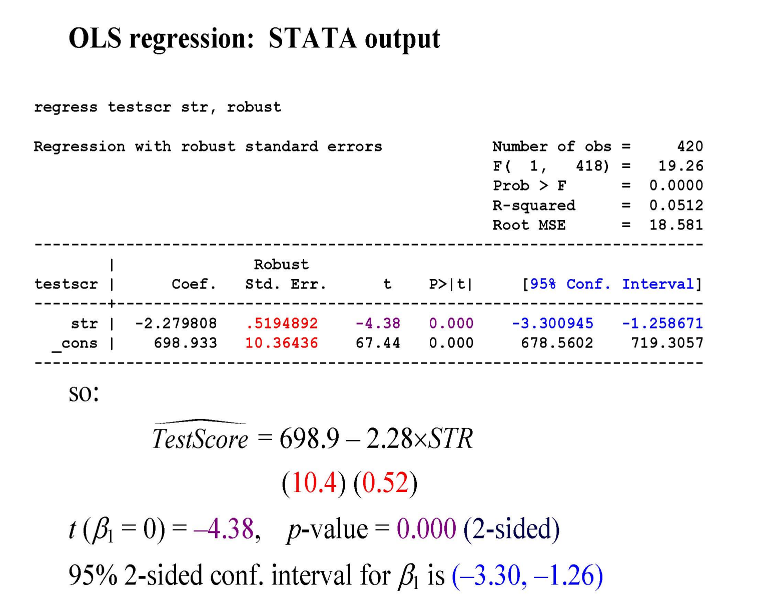
A concise (and conventional) way to report regressions:

Put standard errors in parentheses below the estimated coefficients to which they apply.

$$\begin{aligned}\widehat{TestScore} &= \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR, R^2 = .05, SER = 18.6\end{aligned}$$

This expression gives a lot of information.

- The estimated regression line is $\widehat{Test\ Score} = 698.9 - 2.28 \times STR$.
- The standard error of $\hat{\beta}_0$ is 10.4.
- The standard error of $\hat{\beta}_1$ is 0.52
- The R^2 is .05; the standard error of the regression is 18.6.



OLS regression: STATA output

```
regress testscr str, robust

Regression with robust standard errors                 Number of obs =     420
                                                       F(  1,   418) =   19.26
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0512
                                                       Root MSE      =  18.581
-------------------------------------------------------------------------
        |               Robust
testscr |      Coef.   Std. Err.      t    P>|t|      [95% Conf. Interval]
--------+----------------------------------------------------------------
    str |  -2.279808   .5194892    -4.38   0.000     -3.300945   -1.258671
  _cons |    698.933   10.36436    67.44   0.000      678.5602    719.3057
-------------------------------------------------------------------------
```

so:

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

$$(10.4)\ (0.52)$$

$t\,(\beta_1 = 0) = -4.38$, p-value = 0.000 (2-sided)

95% 2-sided conf. interval for β_1 is $(-3.30, -1.26)$

Regression When X is a Binary Variable

- Sometimes a regressor is binary:
  - $X = 1$ if female, $= 0$ if male
  - $X = 1$ if treated (experimental drug), $= 0$ if not
  - $X = 1$ if small class size, $= 0$ if not
- A binary variable is sometimes called a dummy variable or an indicator variable.
- So far, β_1 has been called a “slope," but that doesn’t make much sense if X is binary. How do we interpret regression with a binary regressor?

Interpreting regressions with a binary regressor

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

where X is binary ($X_i = 0$ or 1).

$$\begin{aligned} \text{When } X_i &= 0, Y_i = \beta_0 + u_i \\ \text{When } X_i &= 1, Y_i = \beta_0 + \beta_1 + u_i \end{aligned}$$

that is:

$$\begin{aligned} \mathrm{E}(Y_i|X_i = 0) &= \beta_0 \\ \mathrm{E}(Y_i|X_i = 1) &= \beta_0 + \beta_1 \end{aligned}$$

so:

$$\beta_1 = \mathrm{E}(Y_i|X_i = 1) - \mathrm{E}(Y_i|X_i = 0)$$

which is the population difference in group means.

Example: TestScore and STR, California data

Let

$$\begin{aligned} D_i &= 1 \text{ if } STR < 20 \\ &= 0 \text{ if } STR \geq 20 \end{aligned}$$

The OLS estimate of the regression line relating *Test Score* to D (with standard errors in parentheses) is:

$$\begin{aligned} \widehat{Test\ Score} = \quad & 650.0 \quad +7.4 \times D \\ & (1.3) \qquad (1.8) \end{aligned}$$

Difference in means between groups = 7.4;

$$SE = 1.8, t = \frac{7.4}{1.8} = 4.0$$

Compare the regression results with the group means, computed directly:

Class Size	Average Score($\bar{Y}$)	Std. Dev.	N
Small($STR < 20$)	657.4	19.4	238
Large($STR \geq 20$)	650.0	17.9	182

- Estimation: $\bar{Y}_{small} - \bar{Y}_{large} = 657.4 - 650.0 = 7.4$.
- Test: $t = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)} = \frac{7.4}{1.83} = 4.05$.

This is the same as in the regression.

Heteroskedasticity and Homoskedasticity

Heteroskedasticity, Homoskedasticity, and the Formula for the Standard Errors of $\hat{\beta}_0$ and $\hat{\beta}_1$

- What do these two terms mean?
- Consequences of homoskedasticity.
- Implication for computing standard errors.

What do these two terms mean?

- If $\mathrm{Var}(u|X=x)$ is constant— that is, the variance of the conditional distribution of u given X does not depend on X, then u is said to be homoskedasticity (變異數齊一).
- Otherwise, u is said to be **heteroskedastic** (變異數不齊一).

Homoskedasticity in a picture

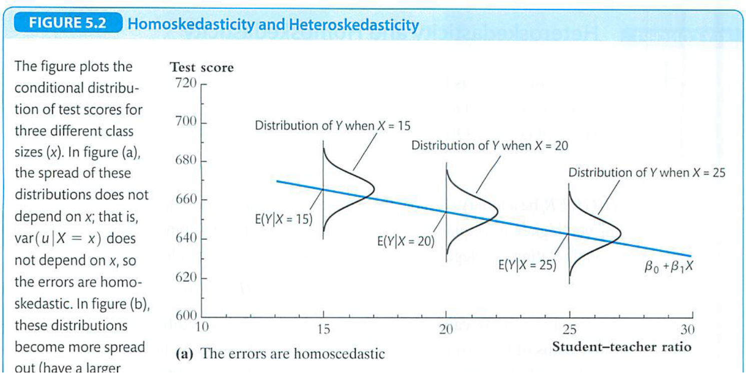


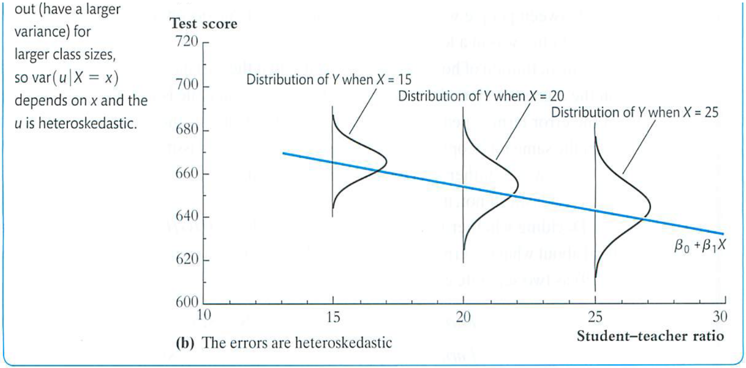
FIGURE 5.2 Homoskedasticity and Heteroskedasticity

The figure plots the conditional distribution of test scores for three different class sizes (x). In figure (a), the spread of these distributions does not depend on x; that is, $\text{var}(u|X = x)$ does not depend on x, so the errors are homoskedastic. In figure (b), these distributions become more spread out (have a larger

(a) The errors are homoscedastic

- $E(u|X = x) = 0$, u satisfies Least Squares Assumption #1.
- The variance of u does **not** depend on x.

Heteroskedasticity in a picture



out (have a larger variance) for larger class sizes, so $\text{var}(u|X = x)$ depends on x and the u is heteroskedastic.

(b) The errors are heteroskedastic

- $E(u|X = x) = 0$, u satisfies Least Squares Assumption #1.
- The variance of u depends on x.

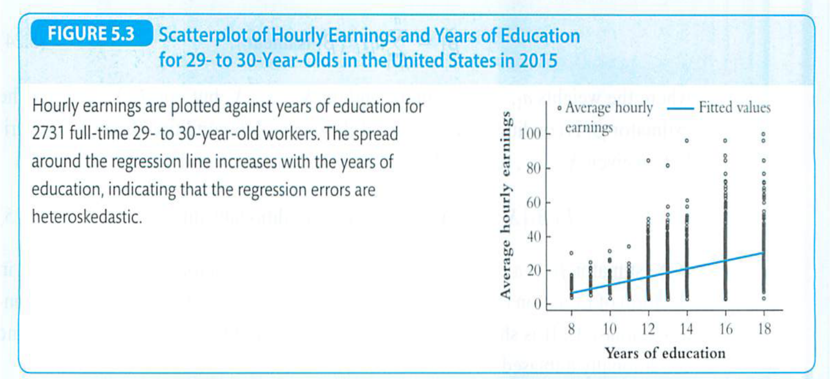


FIGURE 5.3 Scatterplot of Hourly Earnings and Years of Education for 29- to 30-Year-Olds in the United States in 2015

Hourly earnings are plotted against years of education for 2731 full-time 29- to 30-year-old workers. The spread around the regression line increases with the years of education, indicating that the regression errors are heteroskedastic.

Heteroskedastic or homoskedastic?

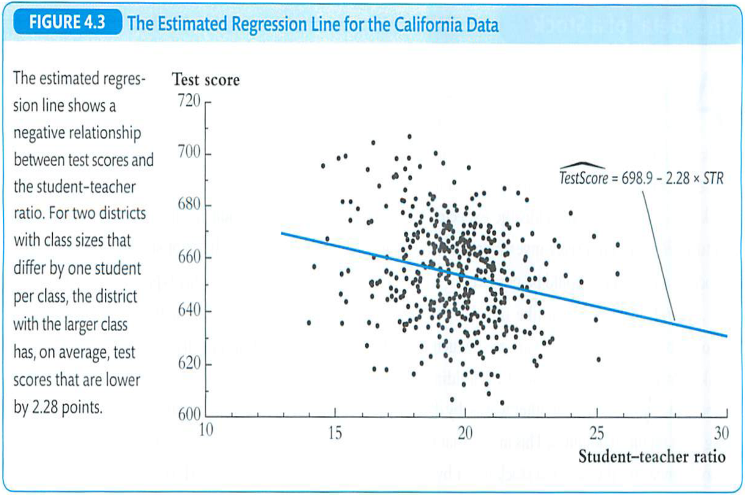


FIGURE 4.3 The Estimated Regression Line for the California Data

The estimated regression line shows a negative relationship between test scores and the student-teacher ratio. For two districts with class sizes that differ by one student per class, the district with the larger class has, on average, test scores that are lower by 2.28 points.

Is heteroskedasticity present in the class size data?

So far we have assumed that u is heteroskedastic. Recall the three least squares assumptions:

1. The conditional distribution of u given X has mean zero, that is, $\mathrm{E}(u|X = x) = 0$.
2. $(X_i, Y_i), i = 1, \cdots, n$, are $i.i.d.$
3. Large outliers are rare.

- Heteroskedasticity and homoskedasticity concern $\mathrm{Var}(u|X = x)$. Because we have not explicitly assumed homoskedastic errors, we have implicitly allowed for heteroskedasticity.

What if the errors are in fact homoskedastic?

- You can prove that OLS has the lowest variance among estimators that are linear in Y, a result called the Gauss-Markov theorem.

- The formula for the variance of $\hat{\beta}_1$ and the OLS standard error simplifies.
  If $\text{Var}(u_i|X_i = x) = \sigma_u^2$, then

$$\text{Var}(\hat{\beta}_1) = \frac{\text{Var}\left[(X_i - \mu_X)u_i\right]}{n(\sigma_X^2)^2} = \cdots = \frac{\sigma_u^2}{n\sigma_X^2}$$

Note: $\text{Var}(\hat{\beta}_1)$ is inversely proportional to $\text{Var}(X)$. More spread in X means more information about $\hat{\beta}_1$.

Because the nominator of $\mathrm{Var}(\hat{\beta}_1)$ is

$$\begin{aligned}
& \mathrm{Var}\left[(X_i-\mu_X)u_i\right] \\
= \ & \mathrm{E}\left(\left[(X_i-\mu_X)u_i-\mathrm{E}((X_i-\mu_X)u_i)\right]^2\right) \\
= \ & \mathrm{E}\left(\left[(X_i-\mu_X)u_i\right]^2\right), \text{ since } \mathrm{E}((X_i-\mu_X)u_i)=0 \\
= \ & \mathrm{E}\left((X_i-\mu_X)^2u_i^2\right) \\
= \ & \mathrm{E}\left((X_i-\mu_X)^2\mathrm{Var}(u_i|X_i)\right), \\
& \text{by law of iterated expectations} \\
= \ & \sigma_X^2\sigma_u^2
\end{aligned}$$

Gauss-Markov conditions:

i. $E(u_i|X_1,\cdots,X_n) = 0.$

ii. $\mathrm{Var}(u_i|X_1,\cdots,X_n) = \sigma_u^2, 0 < \sigma_u^2 < \infty$ for $i = 1,\cdots,n$

iii. $E(u_iu_j|X_1,\cdots,X_n) = 0, i = 1,\cdots,n, i \neq j.$

Gauss-Markov Theorem:

- Under the Gauss-Markov conditions, the OLS estimator $\hat{\beta}_1$ is BLUE (Best Linear Unbiased Estimator). That is,

$$\mathrm{Var}(\hat{\beta}_1|X_1,\cdots,X_n) \leq \mathrm{Var}(\tilde{\beta}_1|X_1,\cdots,X_n)$$

  for all linear conditionally unbiased estimators $\tilde{\beta}_1$.

  See Appendix 5.2 for proof.

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i - \bar{Y}\sum_{i=1}^{n}(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= \sum_{i=1}^{n}\hat{a}_i Y_i
\end{aligned}$$

where $\hat{a}_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

$\hat{\beta}_1$ is a linear unbiased estimator.

Proof of the Gauss-Markov Theorem

- $$\begin{aligned} \text{Let } \tilde{\beta}_1 &= \sum_{i=1}^{n} a_i Y_i = \sum_{i=1}^{n} a_i(\beta_0 + \beta_1 X_i + u_i) \\ &= \beta_0 \left(\sum_{i=1}^{n} a_i \right) + \beta_1 \left(\sum_{i=1}^{n} a_i X_i \right) + \sum_{i=1}^{n} a_i u_i \end{aligned}$$

  is a linear unbiased estimator of β_1.

- For $\tilde{\beta}_1$ to be conditionally unbiased,

  $$\begin{aligned} & \mathrm{E}(\tilde{\beta}_1 | X_1, \cdots, X_n) \\ = \; & \beta_0 \left(\sum_{i=1}^{n} a_i \right) + \beta_1 \left(\sum_{i=1}^{n} a_i X_i \right) + \sum_{i=1}^{n} a_i \mathrm{E}(u_i | X_1, \cdots, X_n) \\ = \; & \beta_0 \left(\sum_{i=1}^{n} a_i \right) + \beta_1 \left(\sum_{i=1}^{n} a_i X_i \right) = \beta_1, \end{aligned}$$

  we need $\sum_{i=1}^{n} a_i = 0$ and $\sum_{i=1}^{n} a_i X_i = 1$.

With the above two conditions for a_i, $\tilde{\beta}_1 - \beta_1 = \sum_{i=1}^n a_i u_i$.

$$\begin{aligned}
\mathrm{Var}(\tilde{\beta}_1 | X_1, \cdots, X_n) &= \mathrm{E}\left((\sum_{i=1}^n a_i u_i)^2 | X_1, \cdots, X_n \right) \\
&= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \mathrm{E}\left(u_i u_j | X_1, \cdots, X_n \right) = \sigma_u^2 \sum_{i=1}^n a_i^2
\end{aligned}$$

Let $a_i = \hat{a}_i + d_i$, then $\sum_{i=1}^n a_i^2 = \sum_{i=1}^n \hat{a}_i^2 + \sum_{i=1}^n d_i^2 + 2\sum_{i=1}^n \hat{a}_i d_i$, and

$$\begin{aligned}
\sum_{i=1}^n \hat{a}_i d_i &= \frac{\sum_{i=1}^n (X_i - \bar{X}) d_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
\sum_{i=1}^n (X_i - \bar{X}) d_i &= \sum_{i=1}^n (a_i - \hat{a}_i) X_i - \bar{X} \sum_{i=1}^n (a_i - \hat{a}_i) \\
&= \left(\sum_{i=1}^n a_i X_i - \sum_{i=1}^n \hat{a}_i X_i \right) - \bar{X} \left(\sum_{i=1}^n a_i - \sum_{i=1}^n \hat{a}_i \right) = 0
\end{aligned}$$

Therefore,

$$\begin{aligned}\sigma_u^2 \sum_{i=1}^{n} a_i^2 - \sigma_u^2 \sum_{i=1}^{n} \hat{a}_i^2 &= \sigma_u^2 \sum_{i=1}^{n} d_i^2 \\ \mathrm{Var}(\tilde{\beta}_1 | X_1, \cdots, X_n) - \mathrm{Var}(\hat{\beta}_1 | X_1, \cdots, X_n) &= \sigma_u^2 \sum_{i=1}^{n} d_i^2\end{aligned}$$

$\tilde{\beta}_1$ has a greater conditional variance than $\hat{\beta}_1$ if $d_i \neq 0$ for any $i = 1, \cdots, n$. If $d_i = 0$ for all i, then $\tilde{\beta}_1 = \hat{\beta}_1$, which proves that OLS is BLUE.

General formula for the standard error of $\hat{\beta}_1$ is the square root of

$$\hat{\sigma}^2_{\hat{\beta}_1} \quad = \quad \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$

Special case under homoskedasticity is

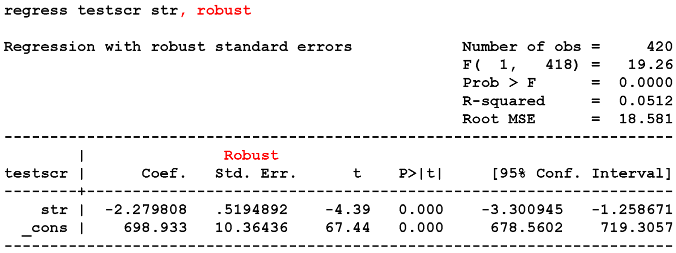
$$\hat{\sigma}^2_{\hat{\beta}_1} \quad = \quad \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

- The homoskedasticity-only formula for the standard error of $\hat{\beta}_1$ and the “heteroskedasticity-robust" formula (the formula that is valid under heteroskedasticity) differ. In general, you get *different standard errors using the different formulas*.

- Homoskedasticity-only standard errors are the default setting in regression software - sometimes the only setting (e.g. Excel). To get the general “heteroskedasticity-robust" standard errors you must override the default.

- If you don’t override the default and there is in fact heteroskedasticity, will get the wrong standard errors (and wrong t-statistics and confidence intervals).

Heteroskedasticity-robust standard errors in STATA



```
regress testscr str, robust

Regression with robust standard errors                Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581
-------------------------------------------------------------------------
        |               Robust
testscr |      Coef.   Std. Err.      t    P>|t|      [95% Conf. Interval]
--------+----------------------------------------------------------------
    str |  -2.279808   .5194892    -4.39   0.000     -3.300945   -1.258671
  _cons |    698.933   10.36436    67.44   0.000      678.5602    719.3057
-------------------------------------------------------------------------
```

Use the “, robust” option!!!

- If you use the “, robust" option, the STATA computes heteroskedasticity-robust standard errors.
- Otherwise, STATA computes homoskedasticity-only standard errors.

The critical points:

- If the errors are homoskedastic and you use the heteroskedastic formula for standard errors (the one we derived), you are OK.
- If the errors are heteroskedastic and you use the homoskedasticity-only formula for standard errors, the standard errors are wrong.
- The two formulas coincide (when n is large) in the special case of homoskedasticity.
- The bottom line: you should always use the heteroskedasticity-based formulas- these are conventionally called the heteroskedasticity-robust standard errors.

Weighted Least Squares

- Since OLS under homoskedasticity is efficient, traditional approach is trying to transform a heteroskedastic model into a homoskedastic model.
- Suppose the conditional variance of u_i is known as a function of X_i

$$\mathrm{Var}(u_i|X_i) = \lambda h(X_i)$$

- Then we can divide both sides of the single-variable regression model by $\sqrt{h(X_i)}$ to obtain

  $$\tilde{Y}_i = \beta_0 \tilde{X}_{0i} + \beta_1 \tilde{X}_{1i} + \tilde{u}_i$$

  where

  - $\tilde{Y}_i = Y_i/\sqrt{h(X_i)}$, $\tilde{X}_{0i} = 1/\sqrt{h(X_i)}$,
  - $\tilde{X}_{1i} = X_i/\sqrt{h(X_i)}$, $\tilde{u}_i = u_i/\sqrt{h(X_i)}$,
  - $\mathrm{Var}(\tilde{u}|X_i) = \frac{\mathrm{Var}(u_i)}{h(X_i)} = \lambda$.

- The **WLS** estimator is the OLS estimator obtained by regressing $\tilde{Y}_i$ on $\tilde{X}_{0i}$ and $\tilde{X}_{1i}$.

- However, $h(X_i)$ is usually unknown, then we have to estimate $h(X_i)$ first to obtain $\widehat{h(X_i)}$, then replace $h(X_i)$ with $\widehat{h(X_i)}$. This is called **feasible** WLS.
- More importantly, the function form of $h(\cdot)$ is usually unknown, then there is no way to systematically estimate $h(X_i)$. This is why, in practice, we usually only run OLS with **robust** standard error.

Summary and Assessment

- The initial policy question:

  Suppose new teachers are hired so the student-teacher ratio falls by one student per class. What is the effect of this policy intervention (this “treatment") on test scores?

- Does our regression analysis give a convincing answer? Not really - districts with low STR tend to be ones with lots of other resources and higher income families, which provide kids with more learning opportunities outside school. This suggests that $corr(u_i, STR_i) < 0$, so $\mathrm{E}(u_i|X_i) \neq 0$.