Topic 5: Simple Linear Regression-Inference

- We have estimated $\hat{\beta}_1$, which is the OLS estimator of the slope coefficient.
- Now we are going to see how we can use the estimated results from a sample to draw inferences about the population parameters β_1.
- I.e., how do we do hypothesis testing and construct confidence interval for the true parameter β_1?

1) Hypothesis testing

- Recall our example

$$TestScore = \beta_0 + \beta_1 \cdot STR + U$$

- We estimated

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

- When we run this in Stata, we get more than just the coefficients. We also get R^2, SER, and standard error of the coefficient etc.

1) Hypothesis testing

- In our example, suppose we want to test if

$$H_0 : \beta_1 = 0$$

- I.e., does class size affect student performance? If $\beta_1 = 0$ then no. If $\beta_1 \neq 0$, then yes.

- In our Topic 3 on Statistics, we could test if $\mu_Y = 0$ by setting up the hypothesis test:

$$H_0 : \mu_Y = 0 \qquad t = \frac{\bar{Y} - \mu^*}{s/\sqrt{n}} = \frac{\bar{Y}}{s/\sqrt{n}}$$

$$H_1 : \mu_Y \neq 0 \qquad \geq 1.96 \text{ or } \leq -1.96$$

1) Hypothesis testing

- We do the same thing for regression estimators $\hat{\beta}_0$ and $\hat{\beta}_1$
- Let's focus on $\hat{\beta}_1$

(1) set up Hypothesis $H_0 : \beta_1 = \beta_1^*$ v.s. $H_1 : \beta_1 \neq \beta_1^*$

(2) Calculate the t-stat $t = \dfrac{\hat{\beta}_1 - \beta_1^*}{SE\left(\hat{\beta}_1\right)}$

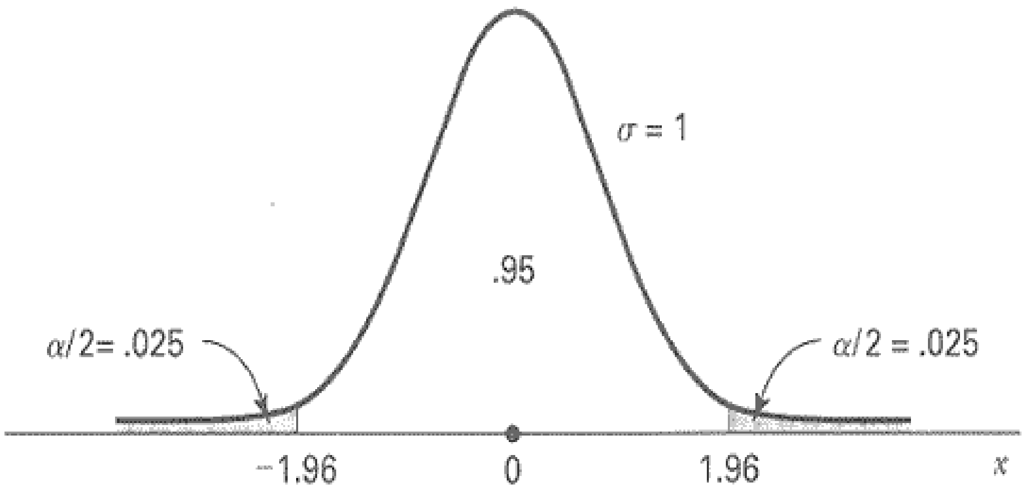
(3) Choose a pre-specified significance level α=0. 05, 0. 01, 0. 1

(4) Again, there are at least 3 ways to test the hypothesis in (1):

(i)critical value (ii)p-values (iii)confidence interval

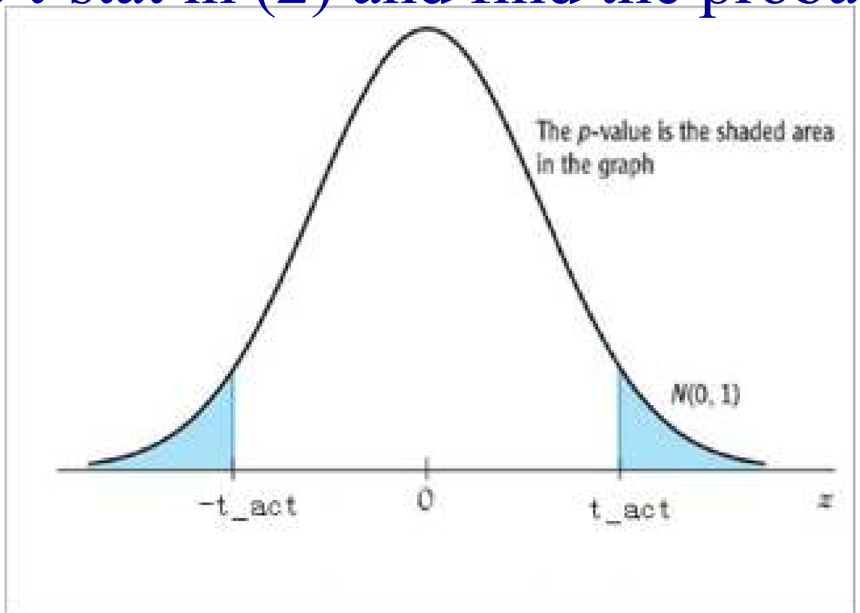
1) Hypothesis testing

- (i) critical value way:
- compare the t-stat in (2) from the standard normal distribution $Z_{\frac{\alpha}{2}}\left(= \pm 1.96 \text{ for } \alpha = 0.05\right)$
- Reject H_0 if $t \geq 1.96$ or $t \leq -1.96$



1) Hypothesis testing

- (ii) p-value way:
- Find the p-value, which the probability of observing a $\hat{\beta}_1$ different from β_1 due to sampling variation given the actual estimate $\hat{\beta}_1^{act}$
- Calculate t-stat in (2) and find the probability:



- If p-value is “large”, we don’t reject H0.

1) Hypothesis testing

- (iii) confidence interval way
- Calculate the (1-α) confidence interval for β_1

$$\left[\hat{\beta}_1 - Z_{\frac{\alpha}{2}} \cdot SE\left(\hat{\beta}_1\right), \quad \hat{\beta}_1 + Z_{\frac{\alpha}{2}} \cdot SE\left(\hat{\beta}_1\right)\right]$$

- If β_1^* falls within the interval, then we cannot reject H0.
- If β_1^* is outside the interval, then we reject H0.

1) Hypothesis testing

- All this hinges on $\hat{\beta}_1$ and $SE\left(\hat{\beta}_1\right)$.
- $SE\left(\hat{\beta}_1\right)$ is an estimator of the standard deviation of the sampling distribution of $\hat{\beta}_1$
- Recall

$$\hat{\beta}_1 \sim N\left(\beta_1, var\left(\hat{\beta}_1\right)\right),$$

where $$var\left(\hat{\beta}_1\right) = \frac{1}{n}\frac{var[(X_i - \mu_X) \cdot u_i]}{[var(X_i)]^2}$$

- An estimator of $var\left(\hat{\beta}_1\right)$ is

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

1) Hypothesis testing

- Thus,

$$SE\left(\hat{\beta}_1\right) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}} = \sqrt{\frac{1}{n}\,\frac{\frac{1}{n-2}\sum\left[(X_i - \bar{X})^2 \cdot \hat{u}_i^2\right]}{\left[\frac{1}{n}\sum(X_i - \bar{X})^2\right]^2}}$$

- This is called heteroskedasticity-robust standard error

1) Hypothesis testing

- This is also called White standard error named after Halbert White from a landmark econometrics paper:

  H. White: "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838 (1980)

  (most cited paper in the economics literature over the last 35 years)

1) Hypothesis testing

- Recall $SE(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$. Same idea for $SE\left(\hat{\beta}_1\right)$, but more complicated.
- $SE\left(\hat{\beta}_1\right)$ estimates how accurate or tight the sampling distribution of $\hat{\beta}_1$.
- Example:

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \cdot STR$$

- Let's test to see if $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{SE\left(\hat{\beta}_1\right)} = \frac{-2.28 - 0}{0.52} = -4.38 < -1.96$$

- Reject at 5% significance level.

1) Hypothesis testing

- $$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \cdot STR$$
- Let's test to see if $H_0 : \beta_1 = -2$ v.s. $H_1 : \beta_1 \neq -2$

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{SE\left(\hat{\beta}_1\right)} = \frac{-2.28 - (-2)}{0.52} = -0.53 > -1.96$$

- Fail to reject at 5% significance level.
- P-value=2*Pr(Z<-0.53)=2*0.298=0.596>0.05: fail to reject
- Confidence interval: fail to reject

$$\left[\hat{\beta}_1 - Z_{\frac{\alpha}{2}} \cdot SE\left(\hat{\beta}_1\right), \quad \hat{\beta}_1 + Z_{\frac{\alpha}{2}} \cdot SE\left(\hat{\beta}_1\right)\right]$$

$$= [-2.28 - 1.96 \cdot 0.52, \quad -2.28 + 1.96 \cdot 0.52]$$

$$= [-3.30, \quad -1.26]$$

1) Hypothesis testing

- What about beta0?

- Usually we don't care about the intercept.

- But we could do everything as we just did for beta1.

- The formula of the standard error of beta0 is different. Stata will do it for us.

2) Confidence interval

- We just saw that a confidence interval can help us with hypothesis testing, but can also use it to get an idea of what beta1 looks like.
- We want to know about beta1 (the effect of X on Y), but we only have beta1_hat (based on a sample).
- We can use information about beta1_hat to create an interval that will contain beta1 in $(1-\alpha)$*100% of repeated sample.
- For the student-teacher-ratio example, the 95% confidence interval is $[-3.30, \quad -1.26]$.

3) Binary variable

- Sometimes our X variable takes a discrete value that is not a number.
- Year of education is a number and is “continuous”: 1,2,3,…etc
- But what about male or female? This is discrete. Suppose we are interested in the impact of gender on wages.

$$Wage_i = \beta_0 + \beta_1 \cdot Gender_i + u_i$$

- Let define

$$D_i = \begin{cases} 1 \text{ if a condition holds} \\ 0 \text{ if not} \end{cases} \quad \text{e.g. } D_i = \begin{cases} 1 \text{ if } i^{th} \text{ person is male} \\ 0 \text{ if } i^{th} \text{ person is female} \end{cases}$$

3) Binary variable

- $Wage_i = \beta_0 + \beta_1 \cdot Gender_i + u_i$
- OLS results:

$$\widehat{Wage}_i = \underset{(2904)}{39565} + \underset{(3746)}{29936} \cdot D_i$$

- If a male,

  expected or predicted wage=39565+29936*1=69501

- If female,

  expected or predicted wage=39565+29936*0=39565

- The coefficient $\hat{\beta}_1$ =29936=Wage for Male-Wage for Female
- So when we do hypothesis test on beta1=0, we are testing whether mean male wage=mean female wage
- Here, e.g. $t = \dfrac{29936 - 0}{3746} = 8 > 1.96$: there is statistically significant difference between men&women income.

4) Heteroskedasticity

- So far, our OLS regression assumptions from last chapters are (1) $E(u_i|X_i)=0$; (2) iid; (3) outlier unlikely.
- So far, we allow that the conditional variance given X_i can depend on X_i
- Homoskedasticity: the conditional variance $var(u_i|X_i)$ does not depend on on X_i. I.e.

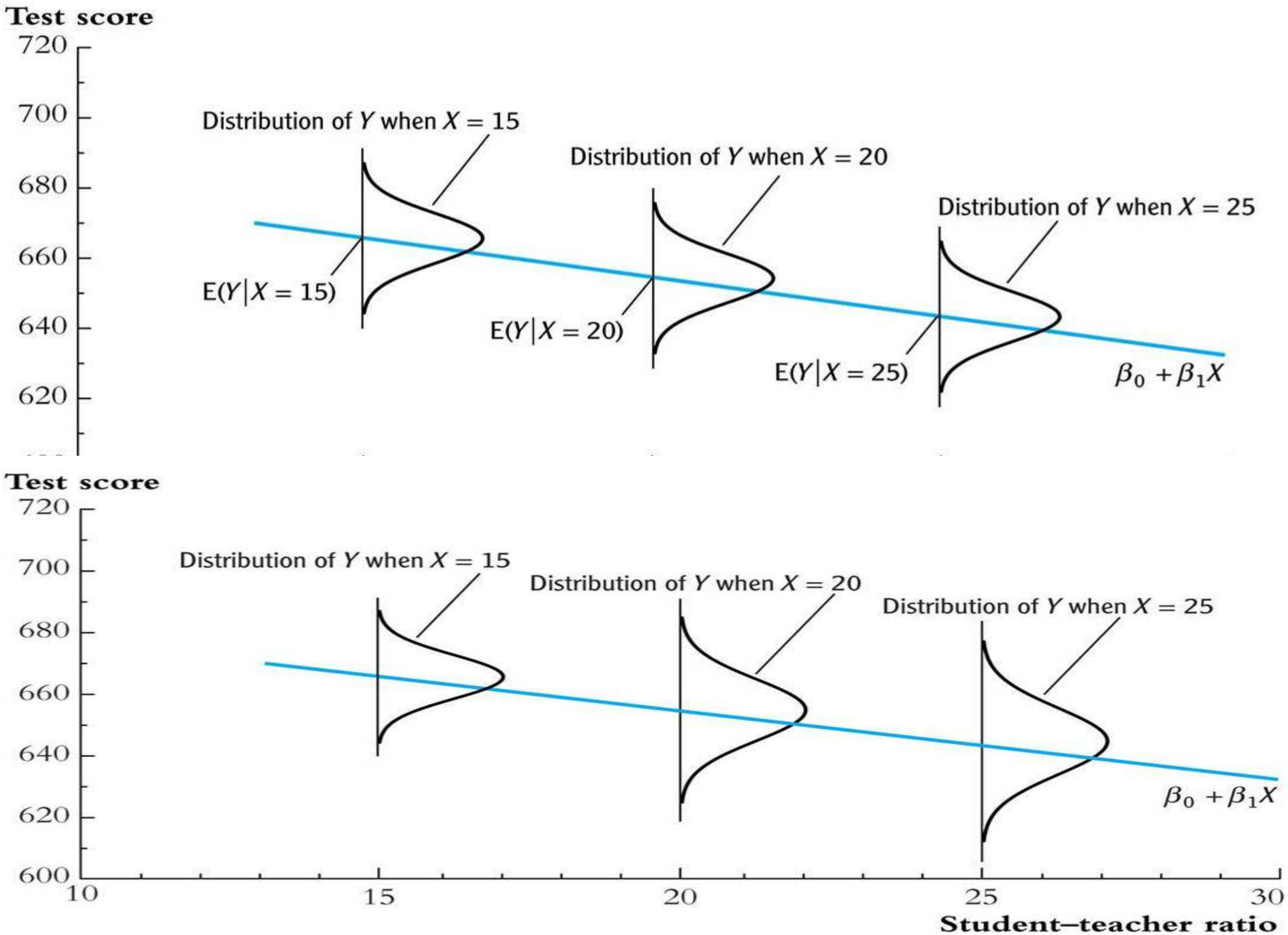
$$var(u_i|X_i) = \text{constant} = \sigma^2$$

- Heteroskedasticity: the conditional variance $var(u_i|X_i)$ can change as X_i changes. E.g.,

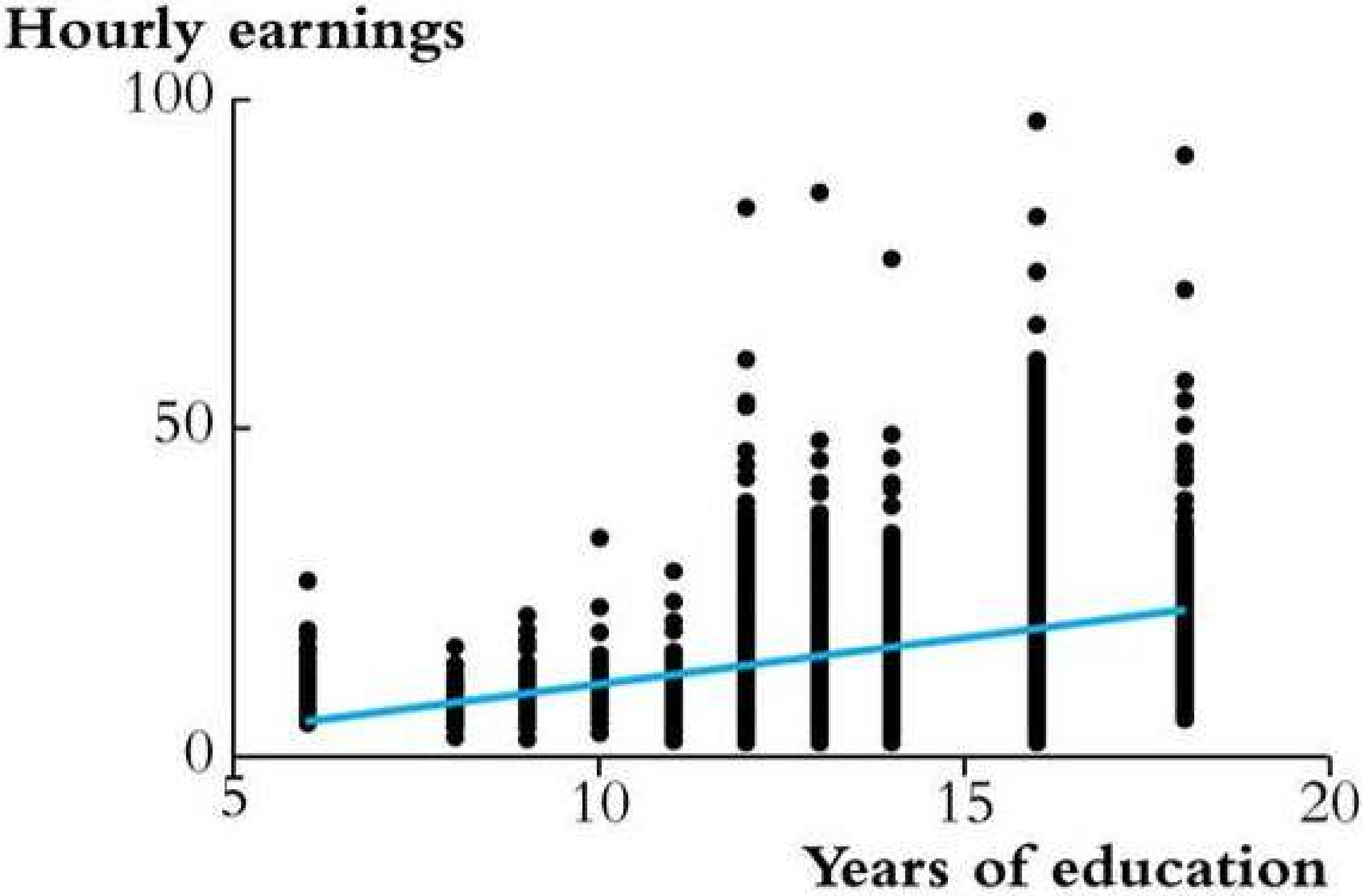
$$var(u_i|X_i) = X_i^2$$

- We can show that $var(Y_i|X_i) = var(u_i|X_i)$

4) Heteroskedasticity



4) Heteroskedasticity

➢ 

Heteroskedastic or homoskedastic?

4) Heteroskedasticity

- Why do we need to care about if it is heteroskedasticity or homoskedasticity?
- As long as the three OLS assumptions are satisfied, then OLS estimators are unbiased, consistent and asymptotically normally distributed. We can do hypothesis testing and construct confidence interval without any problem.
- There are two main reasons:
  - First, when we have homoskedasticity, the standard error of OLS estimators can be simplified.
  - Second, to show OLS estimators are efficient, we need homoskedasticity.
    - This implies that if heteroskedasticity, we can design more efficient estimators.

4) Heteroskedasticity

- Under homoskedasiticty, we can simplify the standard error.
- Recall that under heteroskedasticity:

$$\hat{\beta}_1 \sim N\left(\beta_1, var\left(\hat{\beta}_1\right)\right), \text{where } var\left(\hat{\beta}_1\right) = \frac{1}{n}\frac{var[(X_i - \mu_X) \cdot u_i]}{[var(X_i)]^2}$$

White standard error: $SE\left(\hat{\beta}_1\right) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}} = \sqrt{\frac{1}{n}\frac{\frac{1}{n-2}\sum\left[(X_i - \bar{X})^2 \cdot \hat{u}_i^2\right]}{\left[\frac{1}{n}\sum(X_i - \bar{X})^2\right]^2}}$

- Under homoskedasticity:

$$\hat{\beta}_1 \sim N\left(\beta_1, var\left(\hat{\beta}_1\right)\right), \text{where } var\left(\hat{\beta}_1\right) = \frac{1}{n}\frac{var(u_i)}{var(X_i)}$$

Homoskedasticity-only standard error:

$$SE\left(\hat{\beta}_1\right) = \sqrt{\frac{\frac{1}{n-2}\sum \hat{u}_i^2}{\sum(X_i - \bar{X})^2}}$$

4) Heteroskedasticity

- How do we know if we have homoskedasticity?
- In many case, you don't know, so just use White (heteroskedasticity-robust) standard error
- You can visually inspect the graph by plotting the residuals against Xi and look at the spread
- You can do a formal statistical test, for example conducting a White test.
- It is always safe to use White standard error. If you use White standard error, everything is fine whether homoskedasticity or heteroskedasticity
- If we use homoskedasticity-only standard error, everything is fine, except that your statistical test will be wrong if errors are heteroskedasticity.

4) Heteroskedasticity

- You need to tell Stata which standard errors you are using.
- reg testscr str

```
. reg testscr str

      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(  1,   418) =   22.58
       Model |  7794.11004     1  7794.11004           Prob > F      =  0.0000
    Residual |  144315.484   418  345.252353           R-squared     =  0.0512
-------------+------------------------------           Adj R-squared =  0.0490
       Total |  152109.594   419  363.030056           Root MSE      =  18.581

------------------------------------------------------------------------------
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .4798256    -4.75   0.000    -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000    680.3231    717.5428
------------------------------------------------------------------------------
```

4) Heteroskedasticity

- You need to tell Stata which standard errors you are using.
- reg testscr str, r

```
. reg testscr str,r

Linear regression                               Number of obs =     420
                                                F(  1,   418) =   19.26
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0512
                                                Root MSE      =  18.581
```

testscr	Coef.	Robust 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

5) Gauss-Markov Theorem

- Given our assumption:
  - (i) $E(u_i|X_i)=0$
  - (ii) (X_i, Y_i) are iid
  - (iii) outliers are unlikely

  If we further assume (iv) $var\ (u_i|X_i)= \sigma^2$ (homoskedasticity)

  Then OLS estimaors has the least variance (most efficient, best) among all other possible linear estimators
- OLS is BLUE under (i), (ii), (iii) and (iv).
- If however, we don't have (iv), i.e., the errors are heteroskedasticity, then OLS is not most efficient. And we can use more efficient estimators (GLS: generalized least square)