Economics 6400: Econometrics

Lecture 8: Heteroskedasticity and various other topics

CSU, East Bay

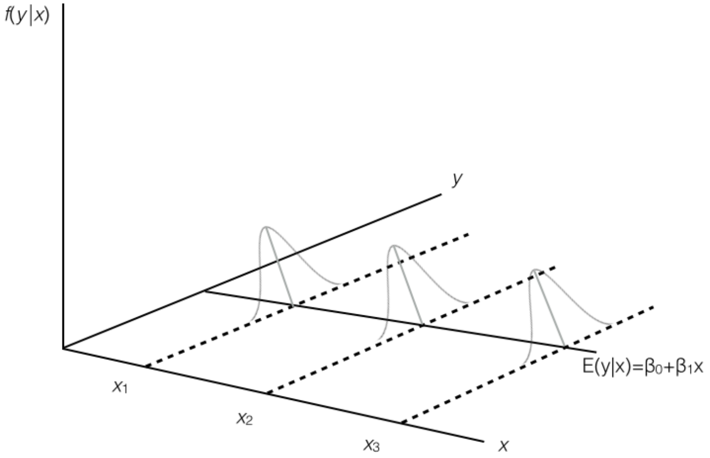
November 21, 2017

In lecture 4...

- We discussed the Gauss-Markov Assumptions.
- The four assumptions are used to establish unbiasedness and the fifth assumption is used to derive the variance formulas:
  1. Linearity: $y = \beta_0 + \beta_1 x + u$
  2. We have a random sample of size n, $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$
  3. Sample variation of explanatory variable: The sample outcomes on x are not all the same value
  4. Zero conditional mean: $E(u|x) = 0$
  5. $\text{Var}(u|x_1, x_2 \ldots, x_k) = \sigma^2$
- Under these assumptions, the OLS estimators are the **best linear unbiased estimators (BLUEs)** of the regression coefficients
  - **Best** means smallest variance

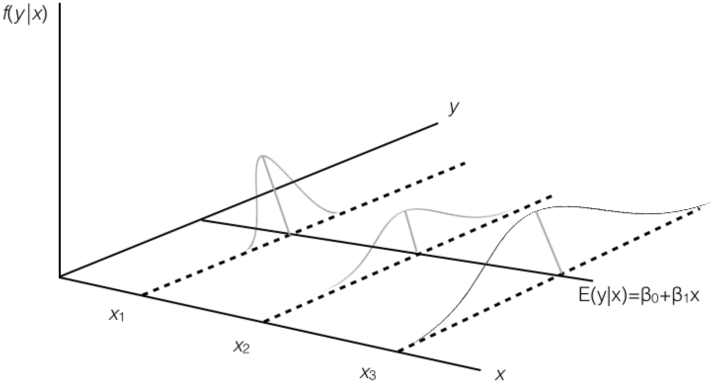
Fifth Gauss-Markov Assumption: Homoskedasticity

- Variation of the unobserved error, u, conditional on the explanatory variables, is constant: $\text{Var}(u_i | x_{i1}, x_{i2} \ldots, x_{ik}) = \sigma^2$



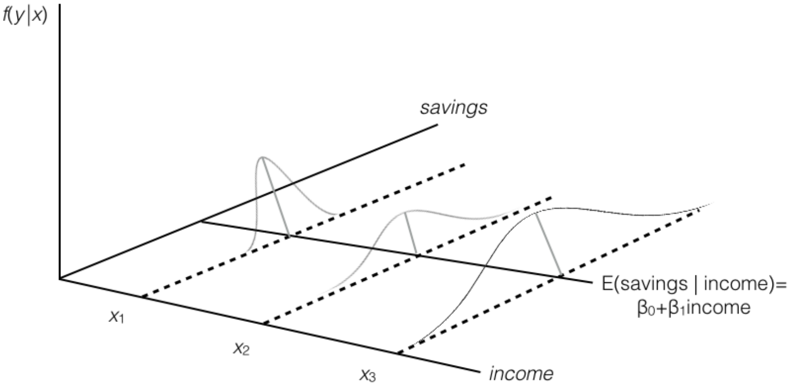
Heteroskedasticity

- If homoskedasticity assumption fails then the variance of the observed factors changes across different segments of the population: $\text{Var}(u_i|x_{i1}, x_{i2} \ldots, x_{ik}) = \sigma_i^2$



Heteroskedasticity

- As incomes rise, families have more discretionary income and more choice about whether to save or spend so σ_i^2 is likely to increase with income



Consequences of heteroskedasticity for OLS

- OLS still unbiased!
- Interpretation of R^2 unchanged
  - Note that $R^2 \approx 1 - \frac{\sigma_u^2}{\sigma_y^2}$ involves **unconditional** variances
- Heteroskedasticity invalidates variance formulas for OLS estimators =(
  - Usual F tests and t tests are not valid under heteroskedasticity
- Fortunately OLS standard errors and related statistics have been developed that are robust to heteroskedasticity of unknown form
- These formulas are only valid in large samples

Heteroskedasticity-robust OLS standard errors

- Valid estimator of $\text{Var}(\hat{\beta}_j)$ is:

$$\widehat{Var}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}$$

  where

  - $\hat{r}_{ij}$ is the ith residual from regressing x_j on others RHS variables
  - SSR_j is the sum of squared residuals from the same regression

- Square-root is **heteroskedasticity-robust standard error**
  - Often the variance estimator is multiplied by $\frac{n}{n-k-1}$ before taking the square-root so that regular OLS standard errors are obtained if $\hat{u}_i^2 = \hat{u}^2$ for all i (i.e. homoskedasticity)
  - Note: $\frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}^2}{SSR_j^2} = \frac{\hat{u}^2 \sum_{i=1}^{n} \hat{r}_{ij}^2}{SSR_j^2} = \frac{\hat{u}^2 SSR_j}{SSR_j^2} = \frac{\hat{u}^2}{SSR_j} = \frac{\hat{u}^2}{SST_j(1-R_j^2)}$
  - Multiplying by $\frac{n}{n-k-1}$ gives $\frac{\hat{\sigma}^2}{SST_j(1-R_j^2)}$

Calculating robust standard errors

- Example: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$

i	y_i	x_{i1}	x_{i2}	$\hat{u}_i$	$\hat{u}_i^2$	$\hat{r}_{i1}$	$\hat{r}_{i1}^2$	$\hat{u}_i^2 \hat{r}_{i1}^2$
1	3.1	11	2	0.05	0.0025	-0.4663	0.2175	0.0005
2	3.2	12	22	0.00	0.00	1.7805	3.1704	0.0000
3	3	11	2	-0.05	0.0025	-0.4663	0.2175	0.0005
4	6	8	44	0.00	0.00	-0.8479	0.7189	0.0000
							$\sum = 4.3242$	$\sum = 0.00109$

- SSR_1 from regression $x_1 = \delta_0 + \delta_1 x_2 + v$ is 4.32418953
- $\widehat{Var}(\hat{\beta}_1) = \frac{0.00109}{4.32418953^2} = 5.815 \times 10^{-5}$
- Robust standard error = $\sqrt{n/(n-k-1) \cdot 5.815 \times 10^{-5}} = 0.01525$

Comparing regular and robust standard errors using Stata's "robust" option

```
. reg wage educ exper
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.6442721	.0538061	11.97	0.000	.5385695	.7499747
exper	.0700954	.0109776	6.39	0.000	.0485297	.0916611
_cons	-3.390539	.7665661	-4.42	0.000	-4.896466	-1.884613

```
. reg wage educ exper, robust
```

wage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.6442721	.0651869	9.88	0.000	.5162117	.7723324
exper	.0700954	.0109943	6.38	0.000	.048497	.0916938
_cons	-3.390539	.8648747	-3.92	0.000	-5.089595	-1.691484

- Robust standard errors tend to be larger (though they can be smaller) so coefficients are typically less significant

Inference with robust standard errors

- t statistic obtained as before but now standard error in the denominator is robust

$$t = \frac{estimate - hypothesized\ value}{robust\ standard\ error}$$

- Regular standard errors are still preferable with homoskedasticity since t statistics have *exact* t distributions regardless of sample size
  - Robust standard errors and t statistics only valid when sample size becomes large!
  - Worthwhile to check for the presence of heteroskedasticity
- Regular F statistic no longer valid but robust F (Wald) statistic can be obtained

Testing for heteroskedasticity: Breusch-Pagan test

- Null hypothesis assumes homoskedasticity (no heteroskedasticity)

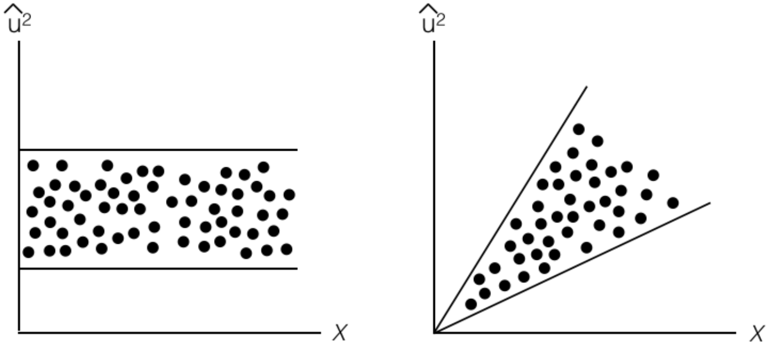
  $$H_0 : Var(u|x_1, x_2, \ldots, x_k) = Var(u|x) = \sigma^2$$

  Since we assume $E(u|x) = 0$, null hypothesis is equivalent to:

  $$H_0 : Var(u|x) = E(u^2|x) - [E(u|x)]^2 = E(u^2|x) = E(u^2) = \sigma^2$$

- Does the data indicate this is not true, i.e. is u^2 related to one or more of the explanatory variables?

Informal testing for heteroskedasticity: Graphical method



- Little evidence of heteroskedasticity on the left; definite pattern on the right

Testing for heteroskedasticity: Breusch-Pagan test

- Does the following linear regression have explanatory power?

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k + v$$

- Null hypothesis of homoskedasticity is:

$$H_0 : \delta_1 = \delta_2 = \ldots = \delta_k = 0$$

F statistic:

$$F = \frac{R^2_{\hat{u}^2}/k}{(1 - R^2_{\hat{u}^2})/(n-k-1)} \sim F_{k,n-k-1}$$

- If the p-value is sufficiently small, we can reject the null hypothesis of homoskedasticity

- Consider the simple house price equation:

$$\widehat{price} = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

```
. reg price lotsize sqrft bdrms

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  3,    84) =   57.46
       Model |  617130.701     3  205710.234           Prob > F      =  0.0000
    Residual |  300723.805    84  3580.0453            R-squared     =  0.6724
-------------+------------------------------           Adj R-squared =  0.6607
       Total |  917854.506    87  10550.0518           Root MSE      =  59.833

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lotsize |   .0020677   .0006421     3.22   0.002     .0007908    .0033446
       sqrft |   .1227782   .0132374     9.28   0.000     .0964541    .1491022
       bdrms |   13.85252   9.010145     1.54   0.128    -4.065141    31.77018
       _cons |  -21.77031   29.47504    -0.74   0.462    -80.38466    36.84405
------------------------------------------------------------------------------
```

Heteroskedasticity in house price equations

```
. predict u, r

. gen u_sq = u*u

. reg u_sq lotsize sqrft bdrms

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  3,    84) =    5.34
       Model |   701213780     3   233737927           Prob > F      =  0.0020
    Residual |  3.6775e+09    84  43780003.5           R-squared     =  0.1601
-------------+------------------------------           Adj R-squared =  0.1301
       Total |  4.3787e+09    87  50330276.7           Root MSE      =  6616.6

------------------------------------------------------------------------------
        u_sq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lotsize |   .2015209   .0710091     2.84   0.006     .0603116    .3427302
       sqrft |   1.691037    1.46385     1.16   0.251    -1.219989    4.602063
       bdrms |    1041.76    996.381     1.05   0.299    -939.6526    3023.173
       _cons |  -5522.795   3259.478    -1.69   0.094    -12004.62    959.0348
------------------------------------------------------------------------------
```

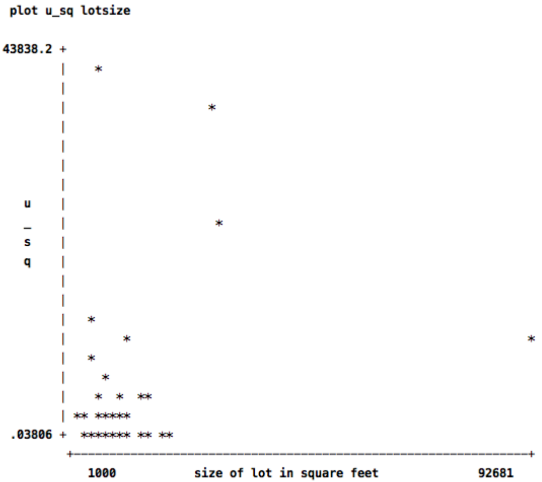
$$F = \frac{0.1601/3}{(1-0.1601)/84} = 5.34 > \text{critical value of } 2.7 \Rightarrow \text{Reject } \mathrm{H}_0$$

Testing for heteroskedasticity: Breusch-Pagan test

- If you suspect that heteroskedasticity is driven by certain variables, you can simply regress $\hat{u}^2$ on those independent variables and carry out the appropriate F test using the degrees of freedom determined by the initial regression to recover $\hat{u}$
  - If one suspects that the heteroskedasticity is caused by a single variable, an t test is appropriate
  - It may be the case that the heteroskedasticity is driven by an outlier

Heteroskedasticity driven by outliers?

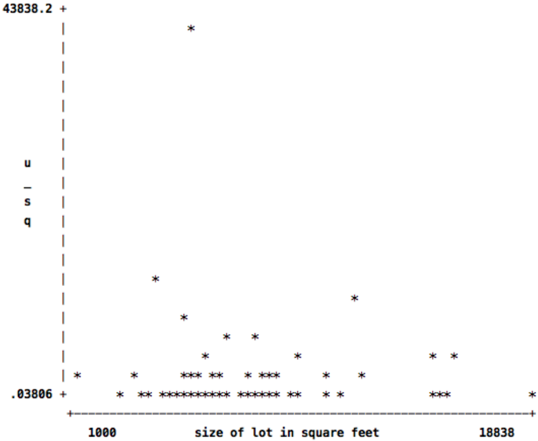
- Spread of $\hat{u}$ seems to be driven by outliers (3 very large houses)



Heteroskedasticity driven by outliers?

- Less evidence of heteroskedasticity with 4 large houses removed

```
. plot u_sq lotsize if lotsize < 20000
```



Heteroskedasticity in house price equations without outliers

```
. reg u_sq lotsize sqrft bdrms if lotsize < 20000

      Source |       SS       df       MS              Number of obs =      84
-------------+------------------------------           F(  3,    80) =    0.17
       Model |    14923253     3  4974417.67           Prob > F      =  0.9170
    Residual |  2.3551e+09    80  29439336.7           R-squared     =  0.0063
-------------+------------------------------           Adj R-squared = -0.0310
       Total |  2.3701e+09    83  28555062.5           Root MSE      =  5425.8

------------------------------------------------------------------------------
        u_sq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lotsize |  -.0485008   .2086501    -0.23   0.817    -.4637276    .3667261
       sqrft |   -.005961   1.380002    -0.00   0.997    -2.752253    2.740331
       bdrms |    509.292   856.6127     0.59   0.554    -1195.422    2214.006
       _cons |   1193.495   2968.212     0.40   0.689    -4713.435    7100.425
------------------------------------------------------------------------------
```

$$F = \frac{0.0063/3}{(1-0.0063)/80} = 0.17 < \text{critical value of } 2.7 \Rightarrow \text{Fail to reject } \mathrm{H}_0$$

Heteroskedasticity in house price equations with log($price$)

- Benefit of using log form is that heteroskedasticity is often reduced:

```
. reg u_sq lotsize lsqrft bdrms

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  3,    84) =    1.34
       Model |  .021478783     3  .007159594           Prob > F      =  0.2679
    Residual |  .449858578    84  .005355459           R-squared     =  0.0456
-------------+------------------------------           Adj R-squared =  0.0115
       Total |  .471337362    87  .005417671           Root MSE      =  .07318

------------------------------------------------------------------------------
        u_sq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lotsize |   1.90e-08   7.83e-07     0.02   0.981    -1.54e-06    1.58e-06
      lsqrft |   -.067345   .0357179    -1.89   0.063     -.138374    .0036841
       bdrms |   .0167766   .0109334     1.53   0.129    -.0049656    .0385188
       _cons |   .4824728   .2521839     1.91   0.059    -.0190225    .9839681
------------------------------------------------------------------------------
```

$$F = \frac{0.046/3}{(1-0.046)/84} = 1.34 < \text{critical value of } 2.7 \Rightarrow \text{Fail to reject } \mathrm{H}_0$$

Testing for heteroskedasticity: White test

- An alternative test that allows for a more flexible pattern of heteroskedasticity
  - Idea: Test whether u^2 is correlated with the right-hand side variables (x_i), the squares of the variables (x_i^2), and the cross products (e.g. $x_1 x_2$)
- When $k = 3$, White test based on estimation of:

$$\begin{aligned}\hat{u}^2 = &\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 + \delta_7 x_1 x_2 \\ &+ \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + error\end{aligned}$$

- Perform F test to see whether coefficients $\delta_1 \ldots \delta_9$ are jointly equal to zero

Testing for heteroskedasticity: White test

- With a large number of regressors (k), the White test consumes many degrees of freedom
  - With 5 independent variables, the White test requires $5 + 5 + 10 = 20$ regressors
- To preserve degrees of freedom, one could regress $\hat{u}^2$ on the fitted values instead:

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + error$$

- With $k = 2$,

$$\begin{aligned}\hat{y}^2 =& (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2) \times (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2) \\ =& \ldots + \hat{\beta}_1^2 x_1^2 + \ldots + 2\hat{\beta}_1 \hat{\beta}_2 x_1 x_2 + \ldots + \hat{\beta}_2^2 x_2^2\end{aligned}$$

so this approaches captures the spirit of the White test

Effects of data scaling on OLS statistics

- When variables are rescaled (e.g. ounces to pounds), the coefficients, standard errors, confidence intervals, t statistics, and F statistics change in ways to preserved measured effects and testing outcomes
- Consider our standard birth weight equation:

$$\widehat{bwght} = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 faminc$$

What is the effect of:

1. Measuring birth weight (left-hand side variable) in pounds rather than ounces? Note: 1 pound = 16 ounces
2. Measuring cigarettes (right-hand side variable) in packets rather than individual cigarettes? Note: Assume 1 packet has 20 cigarettes

Effects of data scaling

	(1) birth weight, ounces	(2) birth weight, pounds	(3) birth weight, ounces
cigs	-0.463**	-0.029**	
	(0.092)	(0.006)	
faminc	0.093**	0.006**	0.093**
	(0.029)	(0.002)	(0.029)
packs			-9.268**
			(1.832)
Constant	116.974**	7.311**	116.974**
	(1.049)	(0.066)	(1.049)
Observations	1388	1388	1388
R^2	0.030	0.030	0.030
SSR	557,485.51	2,177.6778	557,485.51

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$

Effect of measuring birth weight in pounds

- Divide the original equation by 16:

$$\widehat{bwght}/16 = \hat{\beta}_0/16 + (\hat{\beta}_1/16)cigs + (\hat{\beta}_2/16)faminc$$

  so that each new coefficient is simply the original coefficient divided by 16

  - e.g. $\frac{\hat{\beta}_1}{16} = \frac{-0.463}{16} = -0.029$

- R^2 and statistical significance equivalent across models
- SSR is different since $\hat{u}$ is now $\frac{\hat{u}}{16}$ so SSR $= (\frac{\hat{u}}{16})^2 = \frac{\hat{u}^2}{256}$
  - Therefore $\hat{\sigma}^2 = SSR/(n-k-1) = SSR/1,385$ is 256 times smaller, i.e. 557,485.51/256 = 2,177.6778
  - Small SSR simply reflects a difference in units of measurement

Effect of measuring cigarettes in terms of packets

- We can rewrite the original equation as:

$$\begin{aligned}\widehat{bwght} =& \hat{\beta}_0 + (20\hat{\beta}_1)(cigs/20) + \hat{\beta}_2 faminc \\ =& \hat{\beta}_0 + (20\hat{\beta}_1) packs + \hat{\beta}_2 faminc\end{aligned}$$

  so the coefficient on *packs* is now 20 times the original coefficient on *cigs*
- The standard error is also 20 times larger so the t statistic for testing the statistical significance of smoking is unchanged
  - $t = \frac{20\hat{\beta}_j}{20 \cdot se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$
- Note you would not want to include both *cigs* and *packs* on the right-hand side since they are perfectly linearly related

Units of measurement and logarithmic functional form

- If a dependent (left-hand side) variable is in logarithmic form then changing the unit of measurement will not affect any of the slope coefficients since

$$\log(c_1 y_i) = \log(c_1) + \log(y_i)$$

- This means the new intercept will be $\hat{\beta}_0 + \log(c_1)$
- Intercept will also be the only thing to change if an x_j that is represented in logarithmic form has its units of measurement changed

Beta coefficients

- Often difficult to understand the scale that right-hand side variables are measured in
  - This makes it difficult to compare the size of coefficients across variables and infer which variables are "most important"
- Often helpful to talk about what effect a right-hand side variable has in terms of a **one standard deviation** change
- This can be achieved by standardizing each variable by computing its *z-score*, which involves subtracting its sample mean and dividing by its standard deviation

Beta coefficients

- Begin with original OLS equation:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} + \hat{u} \qquad (1)$$

- Recall that

$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}_1 + \hat{\beta}_2 \overline{x}_2 + \ldots + \hat{\beta}_k \overline{x}_k \qquad (2)$$

So subtracting (2) from (1) gives:

$$\begin{aligned} y_i - \overline{y} =& \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} + \hat{u}_i - \hat{\beta}_0 - \hat{\beta}_1 \overline{x}_1 \\ & - \hat{\beta}_2 \overline{x}_2 - \ldots - \hat{\beta}_k \overline{x}_k \\ \Rightarrow y_i - \overline{y} =& \hat{\beta}_1 (x_{i1} - \overline{x}_1) + \hat{\beta}_2 (x_{i2} - \overline{x}_2) + \ldots + \hat{\beta}_k (x_{ik} - \overline{x}_k) + \hat{u}_i \end{aligned}$$

Beta coefficients

- Let $\hat{\sigma}_y$ be the sample standard deviation of the left-hand side variable and $\hat{\sigma}_1$ be the sample standard deviation of x_1. Therefore:

$$(y_i - \overline{y})/\hat{\sigma}_y = (\hat{\sigma}_1/\hat{\sigma}_y)\hat{\beta}_1[(x_{i1} - \overline{x}_1)/\hat{\sigma}_1] + \ldots$$
$$\ldots + (\hat{\sigma}_k/\hat{\sigma}_y)\hat{\beta}_k[(x_{ik} - \overline{x}_k)/\hat{\sigma}_k] + \hat{u}_i$$

- Each variable has now been replaced by its z-score and the new slope coefficient for x_j is $\hat{b}_j = (\hat{\sigma}_j/\hat{\sigma}_y)\hat{\beta}_j$:

$$z_y = \hat{b}_1 z_1 + \hat{b}_2 z_2 + \ldots + \hat{b}_k z_k + error$$

- Intercept has been dropped
- These coefficients are known as **standardized coefficients** or **beta coefficients**
- Interpretation: If x_j increases by one standard deviation then $\hat{y}$ increases by $\hat{b}_j$ standard deviations

Example: Effects of pollution on house prices
One s.d. increase in *nox* decreases *price* by 0.34 s.d.

$$price = \beta_0 + \beta_1 nox + \beta_2 crime + \beta_3 rooms + \beta_4 dist + \beta_5 stratio + u$$

	Regular coeff. $(\hat{\beta}_j)$	Standardized coeff. $(\hat{b}_j)$
nox	-2706.433**	-0.340**
crime	-153.601**	-0.143**
rooms	6735.498**	0.514**
dist	-1026.806**	-0.234**
stratio	-1149.204**	-0.270**
Constant	20871.127**	
Observations	506	506
R^2	0.636	0.636

* $p < 0.05$, ** $p < 0.01$

General remarks on R-squared

- A high R-squared does not imply that there is a causal interpretation
    - Incorrectly estimated time-series regressions with lags of key variables often produces unrealistically high R-squared statistics
    - If I regressed *drowning deaths* on *ice-cream sales* I might obtain a large R^2 but ice-cream consumption should not cause drowning!
- A low R-squared does not preclude precise estimation of partial effects
    - All that matters is the zero conditional mean assumption!
    - There is nothing in the classical linear model assumptions that requires R^2 to be above some value

Adjusted R-squared

- Ordinary R-squared can be rewritten as

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{SSR/n}{SST/n}$$

- Define the **population R-squared** as

$$\rho^2 = 1 - \frac{\sigma_u^2}{\sigma_y^2}$$

  which is the proportion of the variation in y in the population explained by the independent variables
- The ordinary R^2 should be estimating this!

Adjusted R-squared

- Ordinary R-squared estimates σ_u^2 by $\frac{SSR}{n}$, and estimates σ_y^2 by $\frac{SST}{n}$, which are both biased
- Replace them with unbiased estimates, $SSR/(n-k-1)$ and $SST/(n-1)$, to obtain the adjusted R-squared statistic:

$$\overline{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$

- Primary appeal: can impose penalty for adding additional variables to a model, unlike with the ordinary R-squared
  - SSR will fall but k will rise
  - Adjusted R-squared will rise if, and only if, the t-statistic of a newly added regressor is greater than 1 in absolute value
- Relationship between two statistics: $\overline{R}^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$

Using adjusted R-squared to choose between non-nested models

- Models are non-nested if neither is a special case of the other
- Consider the following two models relating R&D intensity to sales:

$$
\begin{aligned}
rdintens =& \beta_0 + \beta_1 \log(sales) + u, \qquad R^2 = 0.06, \overline{R}^2 = 0.03 \\
rdintens =& \beta_0 + \beta_1 sales + \beta_2 sales^2 + u, \quad R^2 = 0.15, \overline{R}^2 = 0.09
\end{aligned}
$$

- Comparison between the R^2 of both models is unfair since the first model contains fewer parameters
- In the above example, even after adjusting for the difference in degrees of freedom, the quadratic model is preferred

Prediction

- Using model estimates to obtain predictions is useful but are subject to sampling variation
- How do we obtain confidence intervals for a prediction from the OLS regression line?
    - Note that the interval is for the *average* y, not a particular y
- Suppose we have estimated the equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

- Let $c_1, c_2, \ldots, c_k$ denote particular values for each of the k independent variables
- We seek to estimate:

$$\begin{aligned} \theta_0 =& \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \ldots + \beta_k c_k \\ =& E(y | x_1 = c_1, x_2 = c_2, \ldots, x_k = c_k) \end{aligned}$$

- The estimator of θ_0 is

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \ldots + \hat{\beta}_k c_k$$

Prediction

- This is easy to compute by hand or with Stata "predict" command
- How can we construct a confidence interval for θ_0, which is centered around $\hat{\theta}_0$?
  - Let's use our trick again!
- Rearrange the above equation

$$\beta_0 = \theta_0 - \beta_1 c_1 - \ldots - \beta_k c_k$$

and plug it into the original equation to obtain:

$$\begin{aligned} y =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \\ =& \theta_0 + \beta_1(x_1 - c_1) + \beta_2(x_2 - c_2) + \ldots + \beta_k(x_k - c_k) + u \end{aligned}$$

- Predicted value and its standard error obtained from the intercept of the above regression!

Prediction example: College GPA

- Suppose we predict college GPA using results from estimated equation:

$$\begin{aligned}\widehat{colgpa} =& 1.493 + 0.00149 sat - 0.01386 hsperc - 0.06088 hsize \\ &+ 0.00546 hsize^2\end{aligned}$$

- To obtain a prediction when $sat = 1,200$, $hsperc = 30$, and $hsize = 5$ we plug these values in to obtain $\widehat{colgpa} = 2.70$
- To obtain a standard error, we generate new variables by subtracting the values of the right-hand side variables, e.g. $sat0 = sat - 1200$, $hsperc0 = hsperc - 30$ etc. and re-run the regression with the new variables
- The coefficient on the intercept (2.7) is our prediction, and the standard error is 0.02
- 95% confidence interval for expected college GPA is $2.70 \pm 1.96(0.020)$ or (2.66, 2.74)

Prediction example: College GPA

```
. gen sat0=sat-1200

. gen hsperc0=hsperc-30

. gen hsize0=hsize-5

. gen hsizesq0=hsizesq-25

. reg colgpa sat0 hsperc0 hsize0 hsizesq0

      Source |       SS       df       MS              Number of obs =    4137
-------------+------------------------------           F(  4,  4132) =  398.02
       Model |  499.030503     4  124.757626           Prob > F      =  0.0000
    Residual |  1295.16517  4132  .313447524           R-squared     =  0.2781
-------------+------------------------------           Adj R-squared =  0.2774
       Total |  1794.19567  4136  .433799728           Root MSE      =  .55986

------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sat0 |   .0014925   .0000652    22.89   0.000     .0013646    .0016204
     hsperc0 |  -.0138558    .000561   -24.70   0.000    -.0149557   -.0127559
      hsize0 |  -.0608815   .0165012    -3.69   0.000    -.0932328   -.0285302
    hsizesq0 |   .0054603   .0022698     2.41   0.016     .0010102    .0099104
       _cons |   2.700075   .0198778   135.83   0.000     2.661104    2.739047
------------------------------------------------------------------------------
```

Next lecture

- Final exam revision