MULTIPLE REGRESSION ANALYSIS: FURTHER ISSUES

Hüseyin Taştan[1]

[1]Yıldız Technical University
Department of Economics

These presentation notes are based on
Introductory Econometrics: A Modern Approach (2nd ed.)
by J. Wooldridge.

25 Kasım 2012

Further Issues in MLR: Outline

- Data scaling
- Standardized regression
- Additional topics in functional form: quadratic models, models with interaction terms
- Goodness of fit: Adjusted R^2
- Prediction

Effects of Data Scaling on OLS Statistics

- Changing the units of measurements changes the OLS intercept and slope estimates.
- Why interested in changing the units of measurements: cosmetic purposes, such as reducing the number of zeros on coefficient estimates, easier interpretation.
- Rescaling data does not change the testing outcomes.
- Rescaling data does not change the significance of coefficient estimates. t statistics do not change.
- R^2 remains the same.
- SSR and SER would change if we rescale the data.
- F test statistic remains the same.

Standardized Regression

- If x_j changes 1 standard deviation, instead of 1 unit, how would y change?
- Answer: Standardize all variables in the regression model (y and all xs) and estimate the model using standardized variables.
- Standardization: subtract the arithmetic mean and divide by sample standard deviation:

$$z_y = \frac{y - \bar{y}}{\hat{\sigma}_y},$$

$$z_1 = \frac{x_1 - \bar{x}_1}{\hat{\sigma}_1}, z_2 = \frac{x_2 - \bar{x}_2}{\hat{\sigma}_2}, \ldots, z_k = \frac{x_k - \bar{x}_k}{\hat{\sigma}_k},$$

- $\hat{\sigma}_j$ is the sample standard deviation of x_j.

Standardized Regression

We wand to standardize the following model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} + \hat{u}_i$$

Subtracting the sample averages from the model we obtain:

$$y_i - \bar{y} = \hat{\beta}_1(x_{i1} - \bar{x}_1) + \hat{\beta}_2(x_{i2} - \bar{x}_2) + \ldots + \hat{\beta}_k(x_{ik} - \bar{x}_k) + \hat{u}_i$$

since the sample average of $\hat{u}$ is zero. Notice that there is no intercept term in the model. Dividing by the sample standard deviations and using simple algebra gives:

$$\frac{y_i - \bar{y}}{\hat{\sigma}_y} = \frac{\hat{\sigma}_1}{\hat{\sigma}_y}\hat{\beta}_1\frac{(x_{i1} - \bar{x}_1)}{\hat{\sigma}_1} + \frac{\hat{\sigma}_2}{\hat{\sigma}_y}\hat{\beta}_2\frac{(x_{i2} - \bar{x}_2)}{\hat{\sigma}_2} + \ldots + \frac{\hat{\sigma}_k}{\hat{\sigma}_y}\hat{\beta}_k\frac{(x_{ik} - \bar{x}_k)}{\hat{\sigma}_k} + \frac{\hat{u}_i}{\hat{\sigma}_y}$$

Standardized Regression

- Rewrite the model as follows:

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

where

$$z_y = \frac{y - \bar{y}}{\hat{\sigma}_y}, \quad z_j = \frac{x_j - \bar{x}_j}{\hat{\sigma}_j}, \ j = 1, 2, \ldots, k$$

- Slope coefficients: known as standardized coefficients or beta coefficients

$$\hat{b}_j = \frac{\hat{\sigma}_j}{\hat{\sigma}_y}\hat{\beta}_j, \ j = 1, 2, \ldots, k$$

- Interpretation: In response to a one standard deviation in x_j, y is predicted to change by $\hat{b}_j$ standard deviations.
- Original units of measurements are irrelevant. They are now measured in terms of standard deviation and they can be compared.

Standardized Regression: Example

Air pollution and house prices: hprice2.gdt

Dependent variable: median house prices in the region (price)
Explanatory Variables:
nox: measure of air pollution,
dist: distance to city business centers,
crime: crime rate in the community,
rooms: average number of rooms in the community,
stratio: average student-teacher ratio in the community

In levels:

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

Standardized model:

$$zprice = b_1 znox + b_2 zcrime + b_3 zrooms + b_4 zdist + b_5 zstratio + zu$$

Standardized Regression: Example

Standardized model results

$$\widehat{zprice} = -0.340\ znox - 0.143\ zcrime + 0.514\ zrooms - 0.235\ zdist - 0.270\ zstratio$$

- One standard deviation increase in air pollution decreases price by 0.34 standard deviation.
- One standard deviation increase in crime reduces price by 0.143 standard deviation.
- The same relevant movement of pollution in the population has a larger effect on housing prices than crime does.
- Size of the house (measured by the number of rooms) has the largest standardized effect.

Example cont.: Unstandardized regression results

$$\widehat{\text{price}} = \underset{(5054.6)}{20871.1} - \underset{(354.09)}{2706.43}\,\text{nox} - \underset{(32.929)}{153.601}\,\text{crime} + \underset{(393.60)}{6735.50}\,\text{rooms}$$
$$- \underset{(188.11)}{1026.81}\,\text{dist} - \underset{(127.43)}{1149.20}\,\text{stratio}$$
$$n = 506 \quad \bar{R}^2 = 0.6320 \quad F(5, 500) = 174.47 \quad \hat{\sigma} = 5586.2$$

(standard errors in parentheses)

More on Functional Form: Logarithmic specifications

- In our previous lectures, we learned how to allow for nonlinear relationships between variables using logarithmic transformation.
- Example: house price model

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_3 rooms + u$$

- β_1: elasticity of prices with respect to air pollution
- $100\beta_2$: approximate percentage change in price in response to a one unit increase in rooms (semi-elasticity)

Example

House prices: hprice2.gdt

$$\widehat{\log(\text{price})} = \underset{(0.188)}{9.234} - \underset{(0.066)}{0.718}\,\log(\text{nox}) + \underset{(0.019)}{0.306}\,\text{rooms}$$
$$n = 506 \quad \bar{R}^2 = 0.512 \quad F(2, 503) = 265.69 \quad \hat{\sigma} = 0.28596$$

(standard errors in parentheses)

- Holdings rooms fixed, 1% increase in nox price falls by 0.718%. The elasticity of price with respect to air pollution is 0.718.
- Holding nox fixed, when rooms increases by one, price increases by 30.6% (100×0.306).
- The approximation $\%\Delta y \approx 100 \times \Delta \log(y)$ becomes inaccurate as the change in $\log(y)$ becomes larger and larger.

Approximation Error in Logarithmic Changes

- We can use the following formula for big changes in logarithmic dependent variable:

$$\widehat{\%\Delta y} = 100 \times [\exp(\hat{\beta}_2) - 1]$$

- In the previous example we obtained $\hat{\beta}_2 = 0.306$:

$$\widehat{\%\Delta y} = 100 \times [\exp(0.306) - 1] = \%35.8$$

- Now the semi-elasticity is larger.
- $\exp(\hat{\beta}_2)$ is a biased but consistent estimator (why?)

Advantages of Logarithmic Transformation

- There are many advantages of using logarithms of strictly positive variables ($y > 0$).
- Interpretation of coefficients is easier: independent of the units of measurements of xs (elasticity or semi-elasticity).
- When $y > 0$, $\log(y)$ often satisfies CLM assumptions more closely than y in levels. Strictly positive variables (prices, income, etc.) often have heteroscedastic or skewed distributions. Taking logs can mitigate these problems.
- Log transformation reduces or eliminates skewness and reduces variance.
- Taking logs narrows the range of the variable leading to less sensitive estimates to outliers (extreme observations).

Some Rules of Thumb for Taking Logs

- Strictly positive variables such as wage, income, population, production, sales etc. are generally included in the model using log transformation.
- Proportions or rates such as unemployment rate, interest rate, etc. usually appear in their original form. But sometimes they may be included in log form if strictly positive.
- If rates or proportions are included in levels: **a percentage point increase** or change.
- If logarithms of rates are taken (e.g. log(unemployment rate)): 1% (**a percentage increase**) or change.
- This distinction is important: if unemployment rate increases from 8% to 9% the increase is 1 percentage point.But in log form there is 100×(log(9)-log(8))=100×0.1177=11.77% increase in unemployment.

Log Transformation

- If the variable takes nonnegative values (≥ 0), i.e. it is 0 for some observations, we cannot use log transformation because $\log(0)$ is not defined.
- In this case we can use $\log(1 + y)$ transformation instead of $\log(y)$.
- If the data contain relatively few 0 values we can use this approach. The interpretation is the same (except for the changes beginning at 0)
- We cannot compare the R^2s from two models in which we have $\log(y)$ as the dependent variable in one of the models and y in the other.

Functional Form: Quadratic Models

- Quadratic functions are generally used to capture decreasing or increasing marginal effects.
- In quadratic models slope coefficient is not constant. It depends on the value of x:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

- The slope between x and y can be approximated as follows:

$$\Delta\hat{y} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x)\Delta x$$

- Or,

$$\frac{\Delta\hat{y}}{\Delta x} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x)$$

- If $x = 0$ then $\hat{\beta}_1$ is the slope estimated for the change from $x = 0$ to $x = 1$. For values larger than $x = 1$ we need to consider the second term.

Quadratic Models: Example

$$\widehat{wage} = 3.73 + 0.298\ exper - 0.0061\ exper^2$$

- If $\beta_1 > 0, \beta_2 < 0$ then the relationship is $\frown$-shaped.
- If $\beta_1 < 0, \beta_2 > 0$ then the relationship is $\smile$-shaped.
- The regression above implies that exper has a diminishing marginal effect on wage.
- Slope estimate is

$$\frac{\Delta\widehat{wage}}{\Delta exper} \approx 0.298 - (2 \times 0.0061) exper$$

- The first year of experience is worth approximately \$0.298. The second year of experience is worth less:

$$\frac{\Delta\widehat{wage}}{\Delta exper} = 0.298 - 0.0122(1) = 0.286$$

Quadratic Models: Example

$$\widehat{wage} = 3.73 + 0.298\ exper - 0.0061\ exper^2$$

- If exper changes from 10 to 11, wage is predicted to change by:

$$\frac{\Delta\widehat{wage}}{\Delta exper} = 0.298 - 0.0122(10) = 0.176$$

- Turning point:

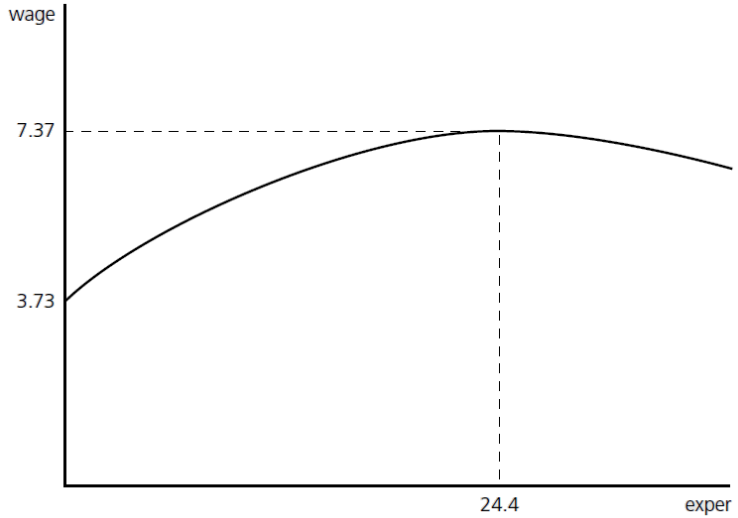
$$\frac{\Delta\hat{y}}{\Delta x} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x) = 0 \Rightarrow x^* = \left|\frac{\hat{\beta}_1}{2\hat{\beta}_2}\right|$$

- Estimated turning point for the wage-exper relationship:

$$exper^* = 0.298/0.0122 = 24.4$$

Quadratic Models: Wage-Experience

$$\widehat{wage} = 3.73 + 0.298\ exper - 0.0061\ exper^2$$



Quadratic Models: Example

$$\widehat{\log(\text{price})} = \underset{(0.566)}{13.386} - \underset{(0.115)}{0.902}\log(\text{nox}) - \underset{(0.043)}{0.0868}\log(\text{dist}) - \underset{(0.0059)}{0.0476}\,\text{stratio} - \underset{(0.1655)}{0.5451}\,\text{rooms} + \underset{(0.0128)}{0.0623}\,\text{rooms}^2$$

$$n = 506 \quad \bar{R}^2 = 0.5988 \quad F(5, 500) = 151.77 \quad \hat{\sigma} = 0.25921$$

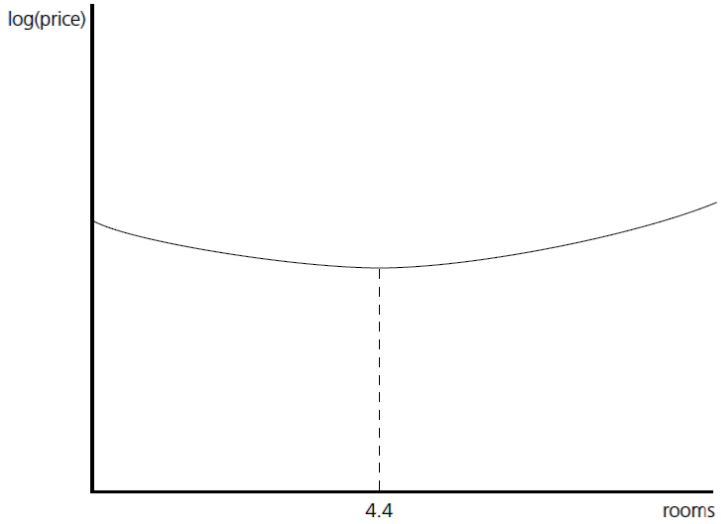
- House value and rooms: First decreasing then increasing.
- As the number of rooms changes from 3 to 4, price is predicted to change by:

$$\frac{\Delta\widehat{\log(price)}}{\Delta rooms} = -0.5451 + 0.1246(3) = -0.1713 \approx -17.13\%$$

- At Rooms=3, an additional room leads to approximately 17.13% decrease in price.
- Turning point:

$$rooms^* = 0.5451/0.1246 = 4.37 \approx 4.4$$

Quadratic Models: House Price



Quadratic Models: House Price

- The impact of an additional room on price:

$$\Delta\widehat{\log(price)} = [-0.545 + 2(0.062)rooms]\Delta rooms$$

$$\%\Delta\widehat{price} = 100 \times [-0.545 + 2(0.062)rooms]\Delta rooms$$
$$= (-54.5 + 12.4rooms)\Delta rooms$$

- For example as the number of rooms changes from 5 to 6, price increases by $-54.5 + 12.4 \times 5 = 7.5\%$. Notice that here, $\Delta rooms = 1$.
- Going from 6 to 7: $-54.5 + 12.4 \times 6 = 19.9\%$.
- Going from 5 to 7: $(-54.5 + 12.4 \times 5)2 = 15\%$. Notice that in this case $\Delta rooms = 2$.

Models with Interaction Terms

- In some cases, the partial impact of one variable may depend on the magnitude of another explanatory variable.
- To capture this we add interaction terms into the regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{x_1 \times x_2}_{interaction} + \beta_4 x_3 + u$$

- Interaction variables: x_1 and x_2. The partial impact of x_1 on y depends on x_2:

$$\frac{\Delta y}{\Delta x_1} = \beta_1 + \beta_3 x_2$$

- To compute this interaction effect we need to plug in a value for x_2. In practice, we generally use mean or median of x_2.
- Similarly, the partial impact of x_2 depends on x_1:

$$\frac{\Delta y}{\Delta x_2} = \beta_2 + \beta_3 x_1$$

Models with Interaction Effects

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{x_1 \times x_2}_{interaction} + \beta_4 x_3 + u$$

$$\frac{\Delta y}{\Delta x_1} = \beta_1 + \beta_3 x_2$$

- Let the sample mean of x_2 be $\bar{x}_2$. Using this value:

$$\frac{\Delta y}{\Delta x_1} = \beta_1 + \beta_3 \bar{x}_2$$

- This gives us the interaction effect at $x_2 = \bar{x}_2$. Is this effect statistically significant?
- To test this we rewrite the model using $x_1 \times (x_2 - \bar{x}_2)$ instead of $x_1 \times x_2$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{x_1 \times (x_2 - \bar{x}_2)}_{interaction} + \beta_4 x_3 + u$$

Models with Interaction Terms

- The model now is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{x_1 \times (x_2 - \bar{x}_2)}_{interaction} + \beta_4 x_3 + u$$

- Simple significance t-test

$$H_0 : \beta_1 = 0$$

- Other effects can be tested similarly.

Interaction Effects: Example, attend.gdt

Variable Definitions:
stndfnl: Standardized final score; **atndrte**: attendance rate (%);
priGPA: cumulative GPA in the previous semester (out of 4);
ACT: achievement test score,

$$\widehat{stndfnl} = \underset{(1.36)}{2.05} - \underset{(.0102)}{.0067}\ atndrte - \underset{(0.48)}{1.63}\ priGPA - \underset{(.098)}{.128}\ ACT$$
$$+ \underset{(.101)}{.296}\ priGPA^2 + \underset{(.0022)}{.0045}\ ACT^2 + \underset{(.0043)}{.0056}\ priGPA{\cdot}atndrte$$
$$n = 680,\ R^2 = .229,\ \bar{R}^2 = .222.$$

The coefficient estimate on $atndrte$ (-0.0067) measures the impact when $priGPA = 0$. Since there is no 0 in $priGPA$ its sign is unimportant. This coefficient alone does not measure the impact of attendance rate because there is interaction term with priGPA.

Interaction Effects: Example, attend.gdt

$$\widehat{stndfnl} = \underset{(1.36)}{2.05} - \underset{(.0102)}{.0067}\ atndrte - \underset{(0.48)}{1.63}\ priGPA - \underset{(.098)}{.128}\ ACT$$
$$+ \underset{(.101)}{.296}\ priGPA^2 + \underset{(.0022)}{.0045}\ ACT^2 + \underset{(.0043)}{.0056}\ priGPA{\cdot}atndrte$$
$$n = 680,\ R^2 = .229,\ \bar{R}^2 = .222.$$

- We need to take into account the interaction term (β_6). Note that β_1 and β_6 cannot pass individual t-statistics but they are jointly significant (The null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ can be rejected using F test with p-value=0.014).
- The sample mean of $priGPA$ is 2.59. Using this:

$$\Delta\widehat{stndfnl} = -0.0067 + (0.0056)(\mathbf{2.59}) = 0.0078$$

- Interpretation: At the mean GPA, $priGPA$ 2.59, a 10 percentage point increase in atndrte increases $\widehat{stndfnl}$ by 0.078 standard deviations from the mean final score.

Interaction Effects: Example, attend.gdt

- The partial effect of the attendance rate at the mean GPA is estimated as 0.0078. Is this effect statistically different from zero?
- To test this we will re-estimate the model using $(priGPA - 2.59) \times atndrte$ instead of $priGPA \times atndrte$.
- In this regression, the coefficient estimate on $atndrte$ (ie., $\hat{\beta}_1$) will measure the predicted partial effect when $priGPA = 2.59$, its sample mean.
- This can easily be tested using the standard t-test.

Interaction Effects: Example, attend.gdt

$$\widehat{\text{stndfnl}} = \underset{(1.36)}{2.05} + \underset{(0.0026)}{\mathbf{0.0078}}\,\mathbf{atndrte} - \underset{(0.481)}{1.6285}\,\text{priGPA} + \underset{(0.101)}{0.2959}\,\text{priGPA}^2$$
$$- \underset{(0.098)}{0.1280}\,\text{ACT} + \underset{(0.0022)}{0.0045}\,\text{ACT}^2 + \underset{(0.004)}{0.0056}\,\mathbf{(priGPA-2.59)\cdot atndrte}$$
$$n = 680 \quad \bar{R}^2 = 0.2218 \quad F(6, 673) = 33.250 \quad \hat{\sigma} = 0.87287$$

- Test:
- $t = 0.0078/0.0026 = 3$, Therefore we reject $H_0 : \beta_1 = 0$, the effect is significant $(p - value = 0.003)$.

Goodness-of-Fit: R-Squared

- The coefficient of determination, R^2, is simply an estimate of "how much variation in y is explained by $x_1, x_2, \ldots, x_k$ in the population".
- A low R^2 value does not automatically imply that the MLR assumptions fail.
- As the number of explanatory variables (k) increases, R^2 always increases (it never decreases). Thus, R^2 has a limited role in choosing between alternative models.
- The relative change in the R-squared when variables added to an equation may be very helpful (e.g. F-statistic for exclusion restrictions depends on the difference in R^2s).

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

- Recall the definition of R^2:

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

- Dividing the numerator and denominator by n:

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

- Since SST/n and SSR/n are biased estimators of respective population variances we will instead use:

$$\frac{SST}{n-1}, \quad \frac{SSR}{n-k-1}$$

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

- Adjusted R-squared is defined as

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

- Adjusted R^2, or R-bar squared may increase or decrease when a new variable is added to the regression. Recall that, in contrast, R^2 never decreases.
- The reason is that when a new variable is added, while SSR decreases, the degrees of freedom $(n-k-1)$ also decreases.
- Basically, it imposes a penalty for adding additional variables to a model. $SSR/(n-k-1)$ can go up or down.
- When a new x variable is added, R-bar square increases if, and only if, the t statistic on the new variable is greater than one in absolute value.
- Extension: when a group of x variables is added, $\bar{R}^2$ increases if, and only if, the F statistic for joint significance of the new variables is greater than 1.

Example

Model 1: OLS, using observations 1–506
Dependent variable: lprice

	Coefficient	Std. Error	t-ratio	p-value
const	8.95348	0.181147	49.4266	0.0000
lnox	−0.304841	0.0821638	−3.7102	0.0002
proptax	−0.00760708	0.000977765	−7.7801	0.0000
rooms	0.288707	0.0181186	15.9343	0.0000

Mean dependent var	9.941057	S.D. dependent var	0.409255
Sum squared resid	36.70511	S.E. of regression	0.270403
R^2	0.566042	Adjusted R^2	0.563449
$F(3, 502)$	218.2650	P-value(F)	1.35e–90

Example

Model 2: OLS, using observations 1–506
Dependent variable: lprice

	Coefficient	Std. Error	t-ratio	p-value
const	8.85532	0.172131	51.4452	0.0000
lnox	−0.275421	0.0779513	−3.5332	0.0004
proptax	−0.00422185	0.00102745	−4.1090	0.0000
rooms	0.281587	0.0171939	16.3771	0.0000
crime	−0.0124893	0.00163861	−7.6219	0.0000

Mean dependent var	9.941057	S.D. dependent var	0.409255
Sum squared resid	32.89123	S.E. of regression	0.256225
R^2	0.611133	Adjusted R^2	**0.608028**
$F(4, 501)$	196.8397	P-value(F)	2.7e–101

Example

Model 3: OLS, using observations 1–506
Dependent variable: lprice

	Coefficient	Std. Error	t-ratio	p-value
const	9.76749	0.222071	43.9837	0.0000
lnox	−0.355701	0.0763150	−4.6610	0.0000
proptax	−0.00185202	0.00106268	−1.7428	0.0820
rooms	0.251409	0.0172902	14.5405	0.0000
crime	−0.0122323	0.00158140	−7.7351	0.0000
stratio	−0.0370699	0.00599178	−6.1868	0.0000

Mean dependent var	9.941057	S.D. dependent var	0.409255
Sum squared resid	30.55237	S.E. of regression	0.247194
R^2	0.638785	Adjusted R^2	**0.635173**
$F(5, 500)$	176.8435	P-value(F)	4.2e–108

Exclusion test for **crime** and **stratio**: $F = 50.35$, $pval < 0.0001$

Adjusted R^2

- When comparing two models using $\bar{R}^2$s the dependent variables must be the same.
- Adjusted R^2 can especially be useful when comparing non-nested models (if the dependent variables are the same)
- For example consider the following non-nested models:

$$y = \beta_0 + \beta_1 \log(x), \quad \bar{R}^2_A$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2, \quad \bar{R}^2_B$$

- F statistic can only be used to test nested models.
- We can choose the model with larger $\bar{R}^2$.

Controlling for Too Many Factors in Regression Analysis

- In general, adding too many variables to a regression, in order to obtain a high R-squared may not be a good idea. We should always consider the ceteris paribus notion together with appropriate statistical tests.
- For example, suppose that we want to examine the relationship between alcohol consumption and traffic accidents. More specifically, we are interested in whether a higher tax on beer will reduce alcohol consumption, thus reduce drunk driving resulting in fewer traffic fatalities.
- Dependent variable: fatalities (number of traffic fatalities); explanatory variable: tax (tax on alcohol)
- We should not add beer consumption (beercons) into this regression model. If we add beercons, its coefficient will measure the difference in traffic fatalities due to a 1% point increase in tax for two regions (or countries) with the same level of beer consumption.
- But, we are not interested in this.

Controlling for Too Many Factors

- We are interested in the partial effect of a one percentage point increase in tax on traffic fatalities.
- To measure this we can add demographic characteristics of regions, or variables that reflect individual tastes and preferences.
- For example, we may use the percentage of males in the population, the percentage of population between 16-21, etc.
- In short, we should not add unnecessary variables just to obtain a high R-squared.

Adding New Variables

- Adding a new variable decreases the error variance, σ_u^2, but at the same time, may increase the degree of multi-collinearity if highly correlated with the variables in the model.
- We should always include independent variables that affect y and decrease σ_u^2 and are uncorrelated with all of the variables in the model.
- For example, we want to estimate the demand for beer:

$$\log(beercons) = \beta_0 + \beta_1 \log(price) + u$$

- Personal characteristics that reflect tastes and preferences (e.g., age, education, gender, etc.) can be added to the model above as they will decrease the error variance significantly. At the same time, these variables are uncorrelated with price.

Prediction

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

- When we plug in particular x values into the model above we obtain a prediction for y which is an estimate of the expected value of y given the particular values for the explanatory variables, $\mathrm{E}(y|x)$.
- Let particular values be $x_1 = c_1$, $x_2 = c_2, \ldots, x_k = c_k$. Also let the prediction value for y be θ_0:

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

- The OLS estimator of θ_0 is:

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

Predicting E$(y|x)$

- 95% confidence interval for θ_0:

$$\hat{\theta}_0 \pm 2\ se(\hat{\theta}_0)$$

- To compute this we need the standard error of $\hat{\theta}_0$.
- This standard error can easily be calculated using an auxiliary regression. By definiton

$$\beta_0 = \theta_0 - \beta_1 c_1 - \beta_2 c_2 - \ldots - \beta_k c_k$$

- Substituting into the model and rearranging we get

$$y = \theta_0 + \beta_1(x_1 - c_1) + \beta_2(x_2 - c_2) + \ldots + \beta_k(x_k - c_k) + u$$

- The standard error on the intercept estimate will give us the standard error of the prediction.

Prediction: Example, gpa2.gdt

$$\widehat{colgpa} = \underset{(0.075)}{1.493} + \underset{(.00007)}{.00149}\ sat - \underset{(.00056)}{.01386}\ hsperc - \underset{(.01650)}{.06088}\ hsize + \underset{(.00227)}{.00546}\ hsize^2$$

$$n = 4{,}137,\ R^2 = .278,\ \bar{R}^2 = .277,\ \hat{\sigma} = .560,$$

- Prediction points: $sat = 1200, hsperc = 30, hsize = 5$ (hsrank:rank in class; hsize:size of class; hsper:100*(hsrank/hssize)
- Plugging into the estimated regression we get $colGPA = 2.70$.
- To compute the standard error of this prediction we define: $sat0 = sat - 1200$, $hsperc0 = hsperc - 30$, $hsize0 = hsize - 5$, $hsizesq0 = hsize - 25$. Then, we regress colGPA on these variables.

Prediction: Example, gpa2.gdt

sat0 = *sat* − 1,200, *hsperc0* = *hsperc* − 30, *hsize0* = *hsize* − 5, and *hsizesq0* = *hsize*^{2} − 25.

$$\widehat{colgpa} = \underset{(0.020)}{2.700} + \underset{(.00007)}{.00149}\ sat0 - \underset{(.00056)}{.01386}\ hsperc0 - \underset{(.01650)}{.06088}\ hsize0 + \underset{(.00227)}{.00546}\ hsizesq0$$

$$n = 4{,}137,\ R^2 = .278,\ \bar{R}^2 = .277,\ \hat{\sigma} = .560.$$

- 95% Confidence Interval: $2.70 \pm 1.96(0.020) = [2.66, 2.74]$.
- The variance of $\hat{\theta}_0$ reaches its smallest value at the arithmetic means of x variables ($c_j = \bar{x}_j$).
- Thus, as the values of c_j get farther away from the $\bar{x}_j$, Var$(\hat{y})$ gets larger and larger.
- The standard error and the confidence interval computed above are for the average value of y for the subpopulation with a given set of covariates.
- This is not the same as the confidence interval for the **individual** predictions of y.

Confidence Interval (CI) for Individual Predictions

- In forming a CI for an unknown outcome on y, we must account for another source of variation: the variance in the unobserved error u in addition to the variance in $\hat{y}$.
- Let y^o represent a new cross-sectional unit (individual, firm, region, country, etc.) not in our original sample:

$$y^o = \beta_0 + \beta_1 x_1^o + \beta_2 x_2^o + \ldots + \beta_k x_k^o + u^o$$

- The OLS prediction of y^o at the values x_j^o:

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

- The prediction error is

$$\hat{e}^o = y^o - \hat{y}^o = \beta_0 + \beta_1 x_1^o + \beta_2 x_2^o + \ldots + \beta_k x_k^o + u^o - \hat{y}^o$$

- Taking expectations we obtain

$$\mathrm{E}(\hat{e}^o) = 0$$

Confidence Interval (CI) for Individual Predictions

- The variance of the prediction error

$$\text{Var}(\hat{e}^o) = \text{Var}(\hat{y}^o) + \text{Var}(u^o) = \text{Var}(\hat{y}^o) + \sigma^2$$

- $\text{Var}(\hat{y}^o)$ is inversely related to the sample size n. It gets smaller as n increases.
- σ^2 is the variance of the unobserved error term. It does not decrease as n increases.
- Thus, σ^2 is the dominant term in the variance of the prediction error.
- The standard error of the prediction error

$$se(\hat{e}^o) = \sqrt{\text{Var}(\hat{y}^o) + \hat{\sigma}^2}$$

- 95% CI is

$$[\hat{y}^o \pm t_{0.025} \cdot se(\hat{e}^o)]$$

Confidence Interval (CI) for Individual Predictions: Example

- The confidence intervals for the individual predictions will be much larger than the CI for the conditional average of y. The reason is that $\hat{\sigma}^2$ is much larger than $\text{Var}(\hat{y}^o)$.
- For example, suppose that we want to construct a 95% CI for the colGPA of a high school student with $sat = 1200, hsperc = 30, hsize = 5$.
- Plugging these values in the regression model we obtain $colGPA = 2.70$ $(\hat{y}^o)$ as before.
- From our earlier calculations we know $se(\hat{y}^o) = 0.02$ and $\hat{\sigma} = 0.56$. Thus, $se(\hat{e}^o) = \sqrt{0.02^2 + 0.56^2} = 0.56$ and the 95% CI is

$$2.70 \pm 1.96 \cdot (0.56) = [1.6,\ 3.8]$$

- This is a very wide confidence interval. It is so large that it is almost impossible to accurately pin down an individual's future college grade point average.