Announcements

- Don't forget about Problem Set 4
- Midterm is Thursday, February 24 in class
- Midterm 2 covers chapters 5 through 8, lectures 1-20-11 through 2-10-11
- Don't forget a scantron sheet and a calculator
- Office hours next week: no Monday office hours (building is locked for Presidents' Day), Tuesday 2pm-5pm, Wednesday 9am-noon

Review: Multivariate Regression

- Our model is now:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_K X_K + \varepsilon$$

- We want to estimate a 'best-fit' line:

$$\hat{y}_i = b_1 + b_2 x_{2i} + b_3 x_{3i} + ... + b_K x_{Ki}$$

  - $\hat{y}_i$: predicted value of Y for individual i
  - $x_{2i}, ..., x_{Ki}$: values of $X_2, ..., X_K$ for individual i
  - b_1: intercept
  - b_k: predicted ΔY for a one unit increase in X_k *holding all other X's constant*

Review: Multivariate Regression: Goodness of Fit

- The adjusted R^2:

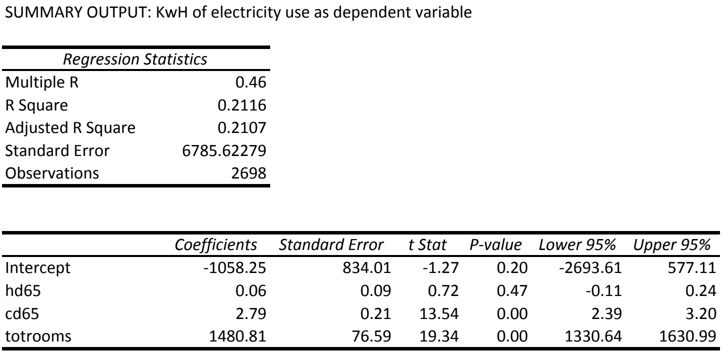
$$\bar{R}^2 = 1 - \frac{n-1}{n-K}\frac{ESS}{TSS}$$

$$ESS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- The adjusted R^2 will be between 0 and 1 and will be closer to 1 the better the fit is
- Adding a regressor will raise the adjusted R^2 if it lowers the error sum of squares enough to offset the penalty for increasing k
- To Excel for an example with using energy consumption data (energy-use-w-calculations.xlsx)...

Review: Multivariate Regression



SUMMARY OUTPUT: KwH of electricity use as dependent variable

Regression Statistics	
Multiple R	0.46
R Square	0.2116
Adjusted R Square	0.2107
Standard Error	6785.62279
Observations	2698

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-1058.25	834.01	-1.27	0.20	-2693.61	577.11
hd65	0.06	0.09	0.72	0.47	-0.11	0.24
cd65	2.79	0.21	13.54	0.00	2.39	3.20
totrooms	1480.81	76.59	19.34	0.00	1330.64	1630.99

Statistical Inference with Multivariate Data

- Now it's time to do statistical inference with multivariate analysis
- Recall that we are still using a single dependent variable (y) but we now have multiple regressors ($x_2, x_3, ..., x_K$)
- We can use ordinary least squares to estimate an intercept (b_1) and a slope coefficient for each regressor ($b_2, ..., b_K$)
- Now our task is to use these results to infer properties of the population relationships between y and $x_2, ..., x_K$.

Assumptions for the Multivariate Population Model

Just like with bivariate data, we need to make a set of assumptions about how multivariate regressors are related to y at the population level. We will make the following assumptions for multivariate data:

1. The population model is: $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k + \varepsilon$
2. The error has mean zero and is unrelated with the regressor
3. The errors for different observations have the same variance, σ_ε^2
4. The errors for different observations are unrelated
5. The errors are normally distributed
6. The regressors are not perfectly correlated with each other

Assumptions for the Multivariate Population Model

- Why do we need this last assumption?
- If two regressors are perfectly correlated, it won't be possible to distinguish the effect of one from the effect of the other on y
- Suppose $x_2 = 2x_3$ and consider the regression equation:

$$y = 10 + 5x_2 + 10x_3$$

- If $x_2 = 2x_3$ this is the same as:

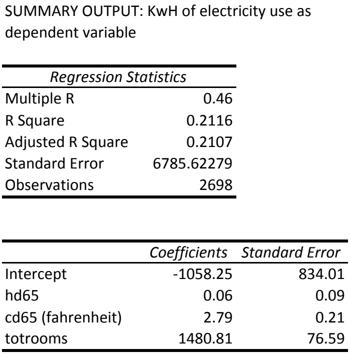
$$y = 10 + 5 \cdot 2x_3 + 10x_3 = 10 + 0 \cdot x_2 + 20x_3$$

- So $b_2 = 5$ and $b_3 = 10$ actually gives the same fit as $b_2 = 0$ and $b_3 = 20$

Assumptions for the Multivariate Population Model

- If two regressors are perfectly correlated, we don't have a unique set of coefficients and we won't be able to run our regressions
- When might we run into perfect correlation? Think about trying to predict how people vote based on how much money they earn and how much they pay in taxes. If taxes are just a percentage of income, we've got perfectly correlated regressors.
- What do we do? Drop one of the regressors (if we don't very bad things happen)
- Now back to Excel to see very bad things in action...

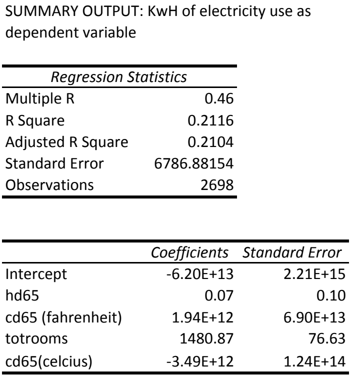
Perfectly Correlated Variables in a Regression



SUMMARY OUTPUT: KwH of electricity use as dependent variable

Regression Statistics	
Multiple R	0.46
R Square	0.2116
Adjusted R Square	0.2107
Standard Error	6785.62279
Observations	2698

	Coefficients	*Standard Error*
Intercept	-1058.25	834.01
hd65	0.06	0.09
cd65 (fahrenheit)	2.79	0.21
totrooms	1480.81	76.59



SUMMARY OUTPUT: KwH of electricity use as dependent variable

Regression Statistics	
Multiple R	0.46
R Square	0.2116
Adjusted R Square	0.2104
Standard Error	6786.88154
Observations	2698

	Coefficients	*Standard Error*
Intercept	-6.20E+13	2.21E+15
hd65	0.07	0.10
cd65 (fahrenheit)	1.94E+12	6.90E+13
totrooms	1480.87	76.63
cd65(celcius)	-3.49E+12	1.24E+14

Properties of our Coefficient Estimates

- If all of the assumptions hold, the coefficient estimates have very useful properties
- First, they are unbiased estimates of the population coefficients ($E(b_j) = \beta_j$)
- Second, each estimated coefficient is normally distributed ($b_j \sim N(\beta_j, \sigma_j^2)$)
- Finally, the test statistic:

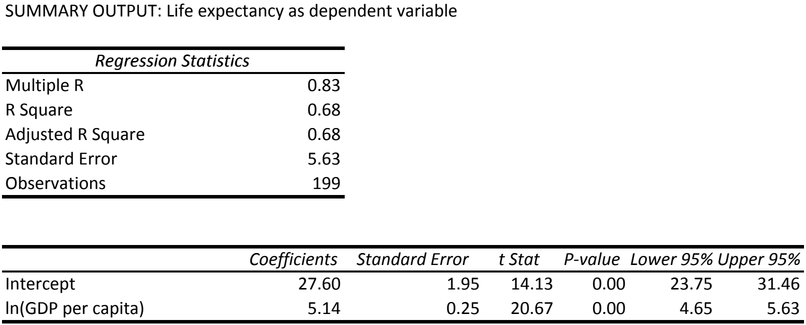
$$t = \frac{b_j - \beta_j}{s_{b_j}}$$

is t distributed with $n - k$ degrees of freedom (s_{b_j} is the standard error of the estimated slope coefficient)

Hypothesis Testing on Individual Coefficients

- If you want to do hypothesis testing for an individual coefficient, everything works the same as the bivariate case except the degrees of freedom are different now ($n - k$ instead of $n - 2$)
- The one big difference is in interpretation
- If we're testing b_j, we're testing the relationship between x_j and y holding the values of all of our other regressors constant
- To see the difference, let's look at an example in Excel (life-expectancy-data.xlsx)...

Hypothesis Testing on Individual Coefficients

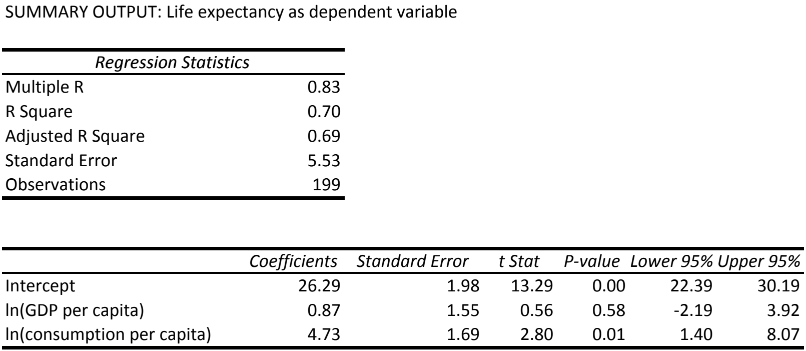


SUMMARY OUTPUT: Life expectancy as dependent variable

Regression Statistics	
Multiple R	0.83
R Square	0.68
Adjusted R Square	0.68
Standard Error	5.63
Observations	199

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	27.60	1.95	14.13	0.00	23.75	31.46
ln(GDP per capita)	5.14	0.25	20.67	0.00	4.65	5.63

Hypothesis Testing on Individual Coefficients



SUMMARY OUTPUT: Life expectancy as dependent variable

Regression Statistics	
Multiple R	0.83
R Square	0.70
Adjusted R Square	0.69
Standard Error	5.53
Observations	199

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	26.29	1.98	13.29	0.00	22.39	30.19
ln(GDP per capita)	0.87	1.55	0.56	0.58	-2.19	3.92
ln(consumption per capita)	4.73	1.69	2.80	0.01	1.40	8.07

Hypothesis Testing on Multiple Coefficients: Overall Significance

- We have a new option available to us, hypothesis tests relating to multiple coefficients
- We'll start by looking at all of the coefficients at once in a test of *overall significance*
- Formally, we want to test the following set of hypotheses:

$$H_o : \ \beta_2 = 0, \beta_3 = 0, ..., \beta_k = 0$$

$$H_a : \ \text{at least one of } \beta_2, ..., \beta_k \neq 0$$

- Another way of stating the null hypothesis is that the regressors explain none of the variation in y
- To test this hypothesis, we need a new kind of test statistic

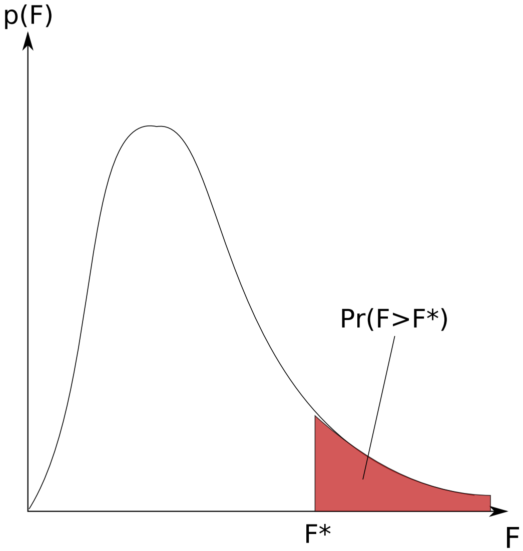
Testing Overall Significance

- Our test statistic for testing overall significance is:

$$F^* = \frac{R^2}{1 - R^2} \frac{n - k}{k - 1}$$

- Notice that the value of the test statistic will get larger as R^2 gets larger
- So bigger values of F should make us more likely to reject the null hypothesis
- How do we know when a value is big?

The F Distribution



Testing Overall Significance

- It turns out that our test statistic F^* is distributed according to an F distribution with $n - k$ and $k - 1$ degrees of freedom
- We can use the F distribution to determine the probability of observing our value of F or something larger if the null hypothesis is true
- If this probability is too low, we'll reject the null hypothesis
- We can do this with either a p-value approach or a critical value approach

Testing Overall Significance

- Using the p-value approach:

$$p = Pr(F_{k-1,n-k} > F^*)$$

- If p is less than our significance level α we reject the null hypothesis
- Using the critical value approach:

$$c = F_{\alpha,k-1,n-k}$$

- If F^* is greater than c we reject the null hypothesis

Testing Overall Significance

- In Excel, we can calculate the p-value as:

$$p = FDIST(F^*, k-1, n-k)$$

- We can calculate the critical value as:

$$c = FINV(\alpha, k-1, n-k)$$

- Alternatively, we can use the F statistic and p-value reported in Excel's regression output
- Back to our Excel regression...

Testing the Significance of a Subset of Regressors

- Sometimes we don't want to test the overall significance of a regression, instead we want to test the significance of a particular subset of regressors
- For example, suppose we had a wage regression with lots of information on education, demographics, etc.
- We might be interested in testing whether including information on an individual's parents can improve our model
- Our hypotheses in this case are:

$$H_o : \ \beta_{g+1} = 0, ..., \beta_k = 0$$

$$H_a : \ \text{at least one of } \beta_{g+1}, ..., \beta_k \neq 0$$

Testing the Significance of a Subset of Regressors

- We call our model with all of the regressors in it the **unrestricted model**:

$$y = \beta_1 + \beta_2 x_2 + ... + \beta_g x_g + \beta_{g+1} x_{g+1} + ... + \beta_k x_k + \varepsilon$$

- We call our model without the subset of regressors we are interested in the **restricted model**:

$$y = \beta_1 + \beta_2 x_2 + ... + \beta_g x_g + \varepsilon$$

- We basically want to test whether the fit is significantly better for the unrestricted model compared to the restricted model

Testing the Significance of a Subset of Regressors

- To do that, we use the following test statistic:

$$F^* = \frac{ESS_r - ESS_u}{ESS_u} \frac{n-k}{k-g}$$

  where ESS_r is the error sum of squares for the restricted model and ESS_u is the error sum of squares for the unrestricted model

- We can also write this test statistic in terms of the R^2 of the two models:

$$F^* = \frac{R_u^2 - R_r^2}{1 - R_u^2} \frac{n-k}{k-g}$$

- Either way, it is clear that F^* is larger the bigger the improvement in fit is when switching from the restricted to unrestricted model

Testing the Significance of a Subset of Regressors

- The test statistic is distributed according to an F distribution with $k - g$ and $n - k$ degrees of freedom
- To test the hypothesis, we can take either the p-value approach ($p = Pr(F_{k-g,n-k} > F^*)$) or the critical value approach ($c = F_{\alpha,k-g,n-k}$)
- If p is less than α or if F^* is greater than c, we will reject the null hypothesis
- Back to Excel...