Module E: Advanced Inference

Slide Deck E10:

Beyond Simple Linear Regression

The section in which we go beyond the simple model that uses one independent variable. Here, we examine how to estimate and interpret models with multiple independent variables of different types.

Today's Objectives

By the end of this slidedeck, you should

1. model a single numeric variable using...
   - more than just one numeric variable
   - a mixture of numeric and categorical variables
2. estimating a value of the dependent variable (with a confidence interval)
3. predicting a value of the dependent variable (with a prediction interval)

Note that we are moving beyond the general theory of confidence intervals and hypothesis testing. We are looking at how to specifically perform the procedures. Make sure you pay attention to the statistical process we follow.

Previously

Last time, in terms of *simple* linear regression, we...

- estimated the parameters of the classical linear model using the ordinary least squares method
- estimated the expected value of y, given x
- predicted the value of a new y, given x

Today, we will rely on the computer to perform the calculations necessary to perform *multiple* regression... regression with multiple independent variables.

- Note that there *are* assumptions/requirements that need to be met in these analyses, but they are reserved for another course (STAT 225).

Recall: Violent Crime and the Unemployed

Example

What is the relationship between the 2000 violent crime rate and the unemployment rate in 1990? If there is a relationship, estimate the 2000 violent crime rate given the 1990 unemployment rate is 10%.

The code for this analysis is:

```
dt = read.csv("http://rfs.kvasaheim.com/data/crime.csv")
attach(dt)

mod3 = lm(vcrime00 ~ unemp1990)
summary(mod3)
```

Violent Crime and the Unemployed

This the important parts of the regression output:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   293.51     168.04   1.747    0.087 .
unemp1990      27.06      30.09   0.899    0.373
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual standard error: 241.9 on 49 degrees of freedom
Multiple R-squared:  0.01624, Adjusted R-squared:  -0.003835
F-statistic: 0.809 on 1 and 49 DF,  p-value: 0.3728
```

Again, you should be able to interpret the important parts of this output.

Violent Crime and the Unemployed

The interpretation is:

Because the p-value of 0.373 is greater than our usual $\alpha = 0.05$, we cannot reject the null hypothesis. We did not detect a statistically significant relationship between the unemployment rate in 1990 with the violent crime rate in 2000.

Beyond Simple Linear Regression

The previous example had one numeric dependent variable and *one* numeric independent variable, hence **Simple Linear Regression**. However, research usually posits that multiple variables affect the variable of interest (response variable). For example:

- state wealth depends on political, economic, and geographical factors
- propensity for terrorist activity depends on minority concentration, economic strengths, and past activity
- football power depends on the quality of the quarterback, the quality of the league, and the quality of the coach
- success probability in STAT 200 depends on professor, preparation, study habits, and events happening in your life

Nothing we have discussed so far can handle this many independent variables... this *reality*.

Ex 1: Violent Crime using Urban Percent *and* Region

Example

As a part of her research, a former student of mine hypothesized that there was a relationship between the violent crime rate and the percent of the population that is urban. She also hypothesized that this relationship differed across the four census regions.

Her actual research focused on policing policies across the United States and their effects on crime rate. This statistical work was just a small part of her effort. In fact, it was a very small part once she had the data.

From her literature review, she further hypothesized that there was a positive relationship between the violent crime rate and the urban percent, and that the relationship would be strongest in the Southern census region of the United States.

Ex 1: Violent Crime using Urban Percent *and* Region

Let us perform the analysis part, the easy part. The code for the *first* part of this analysis is:

```
mod1 = lm(vcrime00 ~ urbanp1990 * census4)
summary.aov(mod1)
```

Why `summary.aov`? There is a categorical independent variable and we want to determine if the *variable* and/or its *interaction* is important.

Ex 1: Violent Crime using Urban Percent *and* Region

The output from this part of the analysis is:

```
                   Df  Sum Sq Mean Sq F value    Pr(>F)
urbanp1990          1  178031  178031   5.402    0.0249 *
census4             3  950724  316908   9.615 5.61e-05 ***
urbanp1990:census4  3  368879  122960   3.731    0.0181 *
Residuals          43 1417204   32958
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1
```

Note that the interaction between the two variables is significant ($p = 0.0181 < 0.05 = \alpha$). As such, both variables are important in the understanding of the violent crime rate in 2000 (in this model).

Furthermore, this result supports her second hypothesis, that the effect of urbanness on the violent crime rate varies from region to region.

Ex 1: Violent Crime using Urban Percent *and* Region

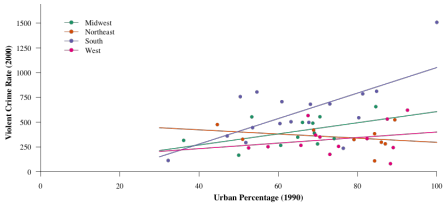
The important parts of the **regression table** are:

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                       45.660    282.048   0.162    0.872
urbanp1990                         5.611      4.331   1.296    0.202
census4Northeast                 461.768    409.146   1.129    0.265
census4South                    -281.398    336.494  -0.836    0.408
census4West                       75.278    422.510   0.178    0.859
urbanp1990:census4Northeast       -7.718      5.809  -1.329    0.191
urbanp1990:census4South            7.263      5.124   1.417    0.164
urbanp1990:census4West            -2.811      6.003  -0.468    0.642
```

Note: Only three of the four regions are listed here. This is because all of the effects (estimates) are measured with respect to this "base" category (Midwest). Thus, the region effect of the South is 281.398 less than for the Midwest. Also, the interaction effect for the South is 7.263 greater than for the Midwest.

Ex 1: Violent Crime using Urban Percent *and* Region

Because the table on the previous slide is difficult to understand (for everyone), here is a graphic of the results.



Ex 1: Violent Crime using Urban Percent *and* Region

Because it is important to illustrate your results with specific examples, let us *estimate* the 2000 violent crime rate for a specific southern state with an urban population of 90%:

```
predict(mod1,data.frame(urbanp1990=90,census4="South"),interval="confidence")
```

```
  fit         lwr         upr
  922.8865   758.8959   1086.877
```

From these results, a 95% confidence interval for the 2000 violent crime rate for a southern state with a 90% urban population is from 759 to 1087, with a point estimate of 923 violent crimes per 100,000 people.

Ex 1: Violent Crime using Urban Percent *and* Region

Again, because it is important to illustrate your results with specific examples, let us *predict* the 2000 violent crime rate for a southern state with an urban population of 90%:

```
predict(mod1,data.frame(urbanp1990=90,census4="South"),interval="predict")
```

```
  fit         lwr         upr
  922.8865   521.7188   1324.054
```

From these results, a 95% prediction interval for the 2000 violent crime rate for a southern state with a 90% urban population is from 522 to 1324, with a point estimate of 923 violent crimes per 100,000 people.

Ex 2: Violent Crime using Property Crime and Region

Example

What is the relationship between the 2000 violent crime rate and the 1990 property crime rate, when controlling for the four census regions?

The code for the *first* part of this analysis is:

```
mod2 = lm(vcrime00 ~ pcrime90 * census4)
summary.aov(mod2)
```

Again, why `summary.aov`?

There is a categorical independent variable and we want to determine if the *variable* and its interaction is important.

Ex 2: Violent Crime using Property Crime Rate *and* Region

The ANOVA output for *this* model is:

```
                Df    Sum Sq   Mean Sq  F value     Pr(>F)
pcrime90         1   1322873   1322873   52.352  5.88e-09 ***
census4          3    423136    141045    5.582   0.00252 **
pcrime90:census4 3     82263     27421    1.085   0.36556
Residuals       43   1086565     25269
```

Note that the interaction between the two variables is *not* statistically significant ($p = 0.36556 > 0.05 = \alpha$). As such, we need to modify our model by removing this interaction.

This decision arises from Occam's Razor: All things being equal, the simpler model is the more useful. This translates as one should *generally* remove non-significant variables.

Ex 2: Violent Crime using Property Crime Rate *and* Region

This leads to the following model:

```
mod2a = lm(vcrime00 ~ pcrime90 + census4)
summary.aov(mod2a)
```

The ANOVA output for the *additive* model is:

```
             Df  Sum Sq Mean Sq F value    Pr(>F)
pcrime90      1 1322873 1322873  52.063 4.33e-09 ***
census4       3  423136  141045   5.551   0.00245 **
Residuals    46 1168828   25409
```

Note: Both variables have a detectable effect. As such, this is our model.

Ex 2: Violent Crime using Property Crime Rate *and* Region

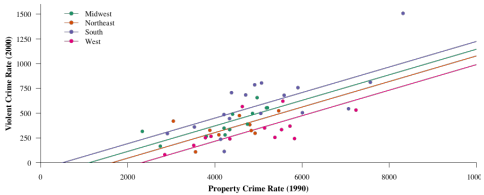
The important parts of the **regression table** are:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -143.37481   96.65541  -1.483   0.1448
pcrime90           0.12889    0.01999   6.448 6.15e-08 ***
census4Northeast  -67.79721   70.31720  -0.964   0.3400
census4South       78.62681   62.83648   1.251   0.2172
census4West      -155.24230   65.02444  -2.387   0.0211 *
```

Again: Note that only three of the four regions are listed here. This is because all of the effects (estimates) are measured with respect to this "base" category (Midwest). Thus, for instance, the region effect of the South is 78.63 greater than for the Midwest.

Ex 2: Violent Crime using Property Crime Rate *and* Region

The following is a graphic of the relationships between the property and the violent crime rate across the four census regions.



Ex 3: Violent Crime using Property Crime and Education Level

Example

What is the relationship between the 2000 violent crime rate and the 1990 property crime rate, when controlling for the level of education level of the state?

Note that this is a very different analysis than the previous multivariate analyses. The two independent variables are *both* numeric. The analysis will be very similar, but the graphic will be quite different.

The code for starting this analysis is:

```
mod3 = lm(vcrime00 ~ pcrime90 * waea90)
summary.aov(mod3)
```

Ex 3: Violent Crime using Property Crime and Education Level

From this first part, the interaction term is not significant. Thus, we should use the additive model.

```
mod3a = lm(vcrime00 ~ pcrime90 + waea90)
summary.aov(mod3a)
```

The above code results in the following ANOVA table:

```
             Df    Sum Sq   Mean Sq   F value      Pr(>F)
pcrime90      1   1322873   1322873     45.03   2.04e-08 ***
waea90        1    181846    181846      6.19      0.0164 *
Residuals    48   1410119     29377
```

Since both are statistically significant (p-value $< \alpha = 0.05$), this is an appropriate model. It is what we will use.

Ex 3: Violent Crime using Property Crime and Education Level

The important parts of the **regression table** are:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 555.94745  321.06193   1.732   0.0898 .
pcrime90      0.14060    0.02039   6.897 1.06e-08 ***
waea90      -12.54722    5.04317  -2.488   0.0164 *
```

Interpretation: For every one increase in the property crime rate, the violent crime rate tends to increase by 0.14 points, all things being equal. Similarly, for every one increase in the weighted average educational achievement, the violent crime rate drops by 12.55 points.

Ex 3: Violent Crime using Property Crime and Education Level

So, how do we illustrate the relationship between/among the two independent *numeric* variables and the dependent?

This is **quite difficult**.

Ex 3: Violent Crime using Property Crime and Education Level

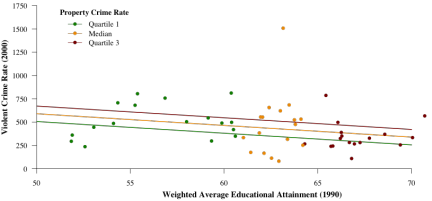
The key is to determine which variable *you want to emphasize* in your graphic. We shall designate this variable as the "x-variable." *Your* x-variable will be assigned a range of values. All other variables need to be set to a specific value (or values).

Depending on how complicated your graphic, those specific values may be means, medians, or quartiles. This is where a conversation with others (*and* experience) is very useful. Remember that education is also a *community activity*. The best work is done with others.

Also, this is where acknowledging "good" is good enough is helpful. No graphic is perfect. They are (usually) good enough. They always have weaknesses... especially because different people look for different features in a graphic.

Ex 3: Violent Crime using Property Crime and Education Level

The following is a graphic of the relationships between the weighted average educational achievement and the violent crime rate for three exemplar property crime rates.



Beyond STAT 200

At this point in the course, we are *technically* able to model the following:

- a Normally-distributed response variable
- that's all

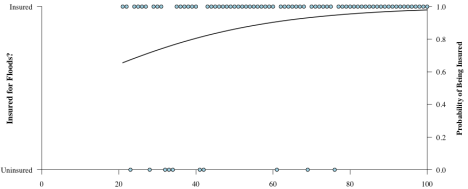
In the future, you may wish to model dependent variables that are not Normal, that are

- just successes or failures (Bernoulli)
- counts of successes over a specified number of trials (Binomial)
- counts of successes over a specified time or region (Poisson)

These are covered in STAT 225. Here are some results that we will be able to perform by the end of *that* course.

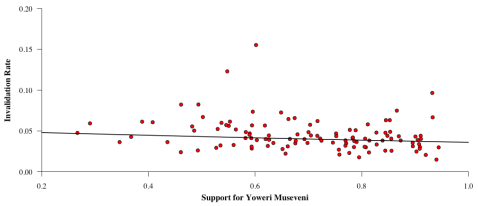
Dichotomous (Bernoulli) Response

The following is a graphic of the relationship between a homeowner's age and the likelihood that the homeowner has purchased flood insurance.



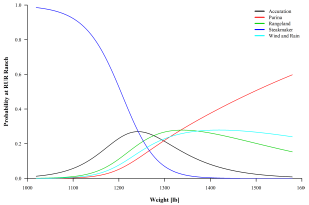
Binomial Response

The following is a graphic of the relationship between a vote and the likelihood it was invalidated in the 2011 presidential election in Uganda.



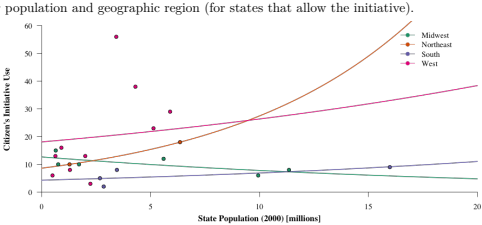
Categorical Response

The following is a graphic estimating the likelihood of a cow being from the RUR ranch, given the cow's weight and the brand of feed.



Count (Rate) Response

The following is a graphic estimating the number of citizen's initiatives in states based on their population and geographic region (for states that allow the initiative).



Today's Objectives

Now that we have concluded this lecture, you should be able to

1. model a single numeric variable using...
   - two or more numeric variables
   - a mixture of numeric and categorical variables
2. estimating a value of the dependent variable (with a confidence interval)
3. predicting a value of the dependent variable (with a prediction interval)

Today's R Functions

Here are the primary R functions we used in exploring linear regression:

- `mod = lm(y~ x)`
  This performs regression and stores the results in the variable `mod`
- `summary.aov(mod)`
  This provides an ANOVA table on the model previously run, which is useful for model selection
- `summary(mod)`
  This provides a regression table on the model previously run, which is useful in determining the effect of the independent variable on the dependent
- `predict(mod, newdata=data.frame(x=n), interval="confidence")`
  This calculates the expected value of y when $x = n$ and provides a **confidence** interval
- `predict(mod, newdata=data.frame(x=n), interval="prediction")`
  This calculates the expected value of y when $x = n$ and provides a **prediction** interval

Supplemental Activities

The following activities are currently available from the STAT 200 website to give you some practice in performing linear regression.

- SCA 42a
- SCA 42b

Source: `https://www.kvasaheim.com/courses/stat200/sca/`

Supplemental Readings

The following are some readings that may be of interest to you in terms of understanding how to perform linear regression:

- Hawkes Learning: Chapter 12
- Intro to Modern Statistics: Chapters 8 and 25
- R for Starters: Chapter 12

- Wikipedia: Ordinary Least Squares

Please do not forget to use the `allProcedures` document that lists all of the procedures we will use in R.