Module E: Advanced Inference

Slide Deck E9:

OLS Linear Regression

The section in which we learn how to extend the correlation test, allowing us to model dependent variables using one (or more) independent variables. We will be able to estimate and predict values, which could not be done with correlation.

Today's Objectives

By the end of this slidedeck, you should

1. understand the theory behind testing…
   - the relationship between two numeric variables
2. estimate a value of the dependent variable (and provide a confidence interval)
3. predict a value of the dependent variable (and provide a prediction interval)

Framing Example

Example

What is the relationship between the violent crime rate in 1990 and in 2000?

Note that we want to know *more* than just if there is a significant relationship.

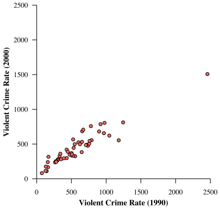
We want to

- specify how much one influences the other
- estimate violent crime rates in 2000, given the value in 1990
- predict violent crime rates in 2000, given the value in 1990

Question: How could we do these things?

Framing Example

What is the relationship between these two variables?



The Theory

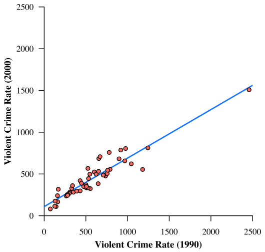
To summarize the relationship, we use a line of “best” fit:

$$y = \beta_0 + \beta_1\ x$$

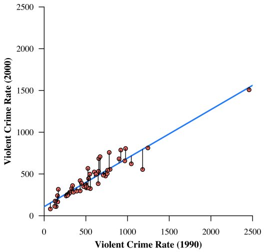
However, this raises the following question: What do we mean by “best” fit?

- This is the most important question we can answer. Different answers lead to different fitting methods.
- MATH/STAT 225 covers a few of the different methods. It also covers the easiest method in great detail.
- In this course, we will only look at the easiest method: ordinary least squares (OLS).
- It is based on **minimizing the sum of the squared residual values**.

The Theory: The Line



The Theory: The Residuals



The Theory

Recall that we are defining “best” fit as the line that minimizes the sum of the squared errors (residuals). The process requires differential calculus. So, if you’ve not had calculus, space out. If you have had it, here is a great use for it.

The Theory (Space-Out)

The first step in minimization is to determine the objective (target) function that needs to be minimized.

$$\begin{aligned} Q &= \sum_i e_i^2 \\ &= \sum_i (y_i - \hat{y}_i)^2 \\ &= \sum_i \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 \end{aligned}$$

This is our objective function. To minimize it, we take its derivative with respect to each parameter, set each equal to 0, and solve for the parameter.

This, we start on the next slide. . .

Start of Lecture Material | A Criminal Framing Example
The Theory of OLS | The Theory of OLS Regression
The Three Examples | Some R Help for Calculations
End of Section Material | A Return to the Framing Example

The Theory (Space-Out)

$$\begin{aligned} \frac{\partial}{\partial \beta_0} Q &= \frac{\partial}{\partial \beta_0} \sum_i \left(y_i - \beta_0 - \beta_1 x_i\right)^2 \\ &= \sum_i 2\left(y_i - \beta_0 - \beta_1 x_i\right)^1 (-1) \\ 0 &\stackrel{\text{set}}{=} -2\left(\sum_i y_i - \sum_i b_0 - \sum_i b_1 x_i\right) \\ &= n\bar{y} - nb_0 - n\bar{x}b_1 \\ \implies \qquad b_0 &= \bar{y} - b_1\bar{x} \end{aligned}$$

Do not forget the definition of the sample mean:

$$\bar{x} = \frac{1}{n}\sum x_i \quad \iff \quad \sum x_i = n\bar{x}$$

The Theory (Space-Out)

$$\begin{aligned}
\frac{\partial}{\partial \beta_1} Q &= \frac{\partial}{\partial \beta_1} \sum_i \left(y_i - \beta_0 - \beta_1 x_i\right)^2 \\
&= \sum_i 2\left(y_i - \beta_0 - \beta_1 x_i\right)(-x_i) \\
0 &\stackrel{\text{set}}{=} -2\left(\sum_i x_i y_i - \sum_i b_0 x_i - \sum_i b_1 x_i^2\right) \\
&= \sum_i x_i y_i - (\bar{y} - b_1 \bar{x}) \sum_i x_i - \sum_i b_1 x_i^2 \\
&= \sum_i x_i y_i - (\bar{y} - b_1 \bar{x}) n\bar{x} - b_1 \sum_i x_i^2 \\
&= \sum_i x_i y_i - n\bar{x}\bar{y} + b_1 n\bar{x}^2 - b_1 \sum_i x_i^2
\end{aligned}$$

The Theory (Space-Out)

$$\begin{aligned}
0 &= \sum_i x_i y_i - n\bar{x}\bar{y} + b_1 n\bar{x}^2 - b_1 \sum_i x_i^2 \\
b_1 \sum_i x_i^2 - b_1 n\bar{x}^2 &= \sum_i x_i y_i - n\bar{x}\bar{y}
\end{aligned}$$

$$b_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

The Theory (Wake Up)

Thus, our estimators of the y-intercept and slope (effect) are

$$b_0 = \bar{y} - b_1\bar{x}$$
$$b_1 = \frac{\sum\ x_i y_i - n\bar{x}\bar{y}}{\sum\ x_i^2 - n\bar{x}^2}$$

By the way, the second formula can also be written as

$$\begin{aligned} b_1 &= \frac{\sum\ (x_i - \bar{x})(y_i - \bar{y})}{\sum\ (x_i - \bar{x})^2} \\ &= r\ \frac{s_y}{s_x} \end{aligned}$$

This gives some insight into the slope, the correlation, and the relation between them.

The Theory: Slope

It can be shown that the slope estimate has the following distribution.

$$b_1 \sim \mathcal{N}\left(\beta_1,\ \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$$

With this distribution, we are able to formulate confidence intervals and p-values for hypotheses about the slope (effect of x on y).

The proof of this is rather elementary and follows from the assumptions we make on the residuals. The three-line proof is provided in STAT 225. Also in this course is the proof of the confidence intervals and the p-values based on this distribution (these are also three-line proofs).

Note: Here, σ^2 is the population variance of the residuals.

The Theory: ANOVA

Sometimes, we want to draw conclusions about *the model, as a whole.*

To do this, we rely on ANOVA (does the model give information about the dependent variable). As such, we need to calculate the following variations:

$$SSM = \sum_i (y_i - \hat{y}_i)^2 \qquad \text{Explained by Model}$$

$$SSE = \sum_i (y_i - \hat{y}_i)^2 \qquad \text{Remaining}$$

$$TSS = \sum_i (y_i - \bar{y})^2 \qquad \text{Total Initial}$$

The Theory: ANOVA

The ANOVA table for the linear regression model with 1 independent variable:

Source	**SS**	**df**	**MS**	**f**	**p-value**
Model	$\sum_i \ (\bar{y}_i - \hat{y}_i)^2$	1	$\frac{SSM}{1}$	$\frac{MSM}{MSE}$	$\mathbb{P}[\ F \geq f\ ]$
Error	$\sum_i \ (y_i - \hat{y}_i)^2$	$n-2$	$\frac{SSE}{n-2}$		
Total	$\sum_i \ (y_i - \bar{y})^2$	$n-1$			

The ANOVA table is useful for drawing conclusions about the model *as a whole.*

The Theory: ANOVA

Finally, we have a measure of how well the model fits the data. This is the so-called "R-squared" value. There are two ways of calculating it:

$$R^2 = \frac{\text{Variation Explained}}{\text{Variation Remaining}} = \frac{TSS - SSE}{TSS} = 1 - \frac{SSE}{TSS}$$

and

$$R^2 = r^2$$

Here, r is the correlation we learned about in the past.

The Theory: R Help

Of course, for any real data set, it is unreasonable to do these calculations by hand.

- The calculations can be done in R, of course.
  1. The first step is to fit the model.
  2. The second step is to summarize the results.

A Return to the Framing Example

This is the code to start to answer the questions raised at the start of this section:

```
### Preamble
source("http://rfs.kvasaheim.com/stat200.R")

# Load and attach the data
dt = read.csv("http://rfs.kvasaheim.com/data/crime.csv")
attach(dt)


### Create the model
modOLS = lm(vcrime00~vcrime90)
```


The Framing Example: ANOVA

To perform the model analysis, you need to run this line:

```
summary.aov(modOLS)
```

Running that line produces the following output:

```
              Df    Sum Sq    Mean Sq    F value    Pr(>F)
vcrime90       1   2556238    2556238      349.3    <2e-16 ***
Residuals     49    358600       7318
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1
```

The Interpretation
Because the p-value is less than our usual $\alpha = 0.05$, we reject the null hypothesis that the model does not offer significant understanding of the relationship between the violent crime rates in 1990 and 2000.

The Framing Example: LM

The ANOVA output tells us if the individual independent variables are statistically significant in describing the dependent variable. That is all the ANOVA output tell us.

If we want to determine things like the effect estimates (which we usually do), we use the "linear model" summary:

```
summary.lm(modOLS)
```

Side Note: If you fit using the `lm` function, the `summary` function is equivalent to the `summary.lm` function. That is,

```
summary(modOLS)
```

gives the same results.

The Framing Example: LM

Here is the resulting output. Make sure you can interpret the important parts.

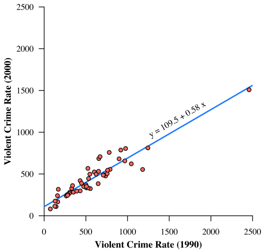
```
Call:
lm(formula = vcrime00 ~ vcrime90)

Residuals:
    Min      1Q  Median      3Q     Max
-241.32  -42.84  -18.04   40.97  208.41

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.52716   21.42679   5.112 5.27e-06 ***
vcrime90      0.58065    0.03107  18.689  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual standard error: 85.55 on 49 degrees of freedom
Multiple R-squared:  0.877, Adjusted R-squared:  0.8745
F-statistic: 349.3 on 1 and 49 DF,  p-value: < 2.2e-16
```

The Framing Example: LM

From this output, we at least know the following:

- The effect of the 1990 violent crime rate on that of 2000 is 0.58065
- For every 1 increase in the violent crime rate in 1990, the estimated violent crime rate in 2000 increases by 0.58065
- Those states with no violent crime in 1990 have an expected violent crime rate of 109.52716 in 2000
- The effect of the 1990 violent crime rate on the 2000 is not 0 (p-value $\ll 0.0001$)

The Framing Example: The Graphic



The Framing Example: Confidence Intervals for a Parameter

The previous code provides the point estimates for the intercept and slope. However, as we already know, we also need a confidence interval to communicate the precision of our estimates.

```
confint(modOLS)
```

```
                2.5 %      97.5 %
(Intercept) 66.4684173 152.5859064
vcrime90     0.5182157   0.6430849
```

Partial Conclusion:
We are 95% confident that the true effect of the 1990 violent crime rate on the 2000 is between 0.518 and 0.643.

Start of Lecture Material
The Theory of OLS
The Three Examples
End of Section Material
A Criminal Framing Example
The Theory of OLS Regression
Some R Help for Calculations
A Return to the Framing Example

The Framing Example: Confidence Intervals for a y-Value

Since we can estimate a value of the dependent variable, we also need to calculate a confidence interval to indicate the precision of our estimate. Recall that the original problem had us

- estimate the 2000 violent crime rate for a state with a 1990 violent crime rate of 1500.

```
predict(modOLS, newdata=data.frame(vcrime90=1500), interval="confidence")
```

```
       fit      lwr      upr
1 980.5026 917.7507 1043.255
```

Partial Conclusion:
We are 95% confident that the *expected* violent crime rate in 2000 for a state with a rate of 1500 in 1990 is between 918 and 1043, with a point estimate of 981.

The Framing Example: Prediction Intervals for a Future y-Value

Finally, we can predict a value of the dependent variable for the next state with that value of the independent variable. Recall that the original problem had us
We can take *yet another* step:

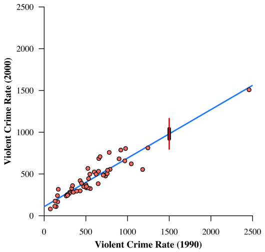
- predict the 2000 violent crime rate for a state with a 1990 violent crime rate of 1500.

```
predict(modOLS, newdata=data.frame(vcrime90=1500), interval="prediction")
```

```
       fit      lwr      upr
1 980.5026 797.4939 1163.511
```

Partial Conclusion:
We are 95% sure that the *actual* violent crime rate in 2000 for a state with a rate of 1500 in 1990 is between 797 and 1164, with a point estimate of 981.

The Two Intervals



Ex 1: Violent Crime against Education

Example

What is the relationship between the violent crime rate in 2000 and the school enrollment in 1990?

The code for this analysis is:

```
mod1 = lm(vcrime00 ~ enroll90)
summary(mod1)
```


Ex 1: Violent Crime against Education

The output is:

```
Call:
lm(formula = vcrime00 ~ enroll90)

Residuals:
    Min      1Q  Median      3Q     Max
-406.58 -160.05  -55.30   98.19 1005.73

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -211.613    804.596  -0.263    0.794
enroll90       7.095      8.732   0.813    0.420

Residual standard error: 242.3 on 49 degrees of freedom
Multiple R-squared:  0.01329, Adjusted R-squared:  -0.006843
F-statistic: 0.6602 on 1 and 49 DF,  p-value: 0.4204
```

Ex 1: Violent Crime against Education

The *important* output is:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -211.613    804.596  -0.263    0.794
enroll90       7.095      8.732   0.813    0.420
```

Brief Conclusion:
Because the p-value of 0.4204 is greater than our usual $\alpha = 0.05$, we cannot reject the null hypothesis. We did not detect a statistically significant relationship between the school enrollment in 1990 with the violent crime rate in 2000.

Ex 2: Violent Crime against Wealth

Example

What is the relationship between the 2000 violent crime rate and the average wealth in 1990? Also, what is the predicted 2000 violent crime rate for a state with average wealth $50,000?

The code for this analysis is:

```
mod2 = lm(vcrime00 ~ gspcap90)
summary(mod2)

confint(mod2)

predict(mod2, newdata=data.frame(gspcap90=50000), interval="prediction")
```

Ex 2: Violent Crime against Wealth

The regression output is:

```
Call:
lm(formula = vcrime00 ~ gspcap90)

Residuals:
   Min     1Q Median     3Q    Max
-413.6 -123.9  -44.5  126.1  427.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 74.32013   88.97655   0.835    0.408
gspcap90     0.01598    0.00366   4.365 6.55e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual standard error: 207 on 49 degrees of freedom
Multiple R-squared:   0.28, Adjusted R-squared:  0.2653
F-statistic: 19.06 on 1 and 49 DF,  p-value: 6.545e-05
```


Ex 2: Violent Crime against Wealth

Brief Conclusion:
Because the p-value is much less than our usual $\alpha = 0.05$, we can reject the null hypothesis. We did detect a statistically significant relationship between the average state wealth in 1990 with the violent crime rate in 2000.

```
                    2.5 %       97.5 %
(Intercept) -1.044849e+02 253.12520261
gspcap90     8.622233e-03   0.02333281
```

For every \$1000 increase in average wealth in the state, the 2000 violent crime rate increases by an average of between 8.6 and 23.3.

```
       fit      lwr      upr
1 873.1962 408.6089 1337.784
```

We predict that a state with a GSP per capita of \$50,000 in 1990 will have a violent crime rate in 2000 between 409 and 1338, with a prediction of 873.

Ex 3: Violent Crime against 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:

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


Ex 3: Violent Crime against the Unemployed

The *important* output is:

```
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
```

Interpretation:
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.

Today's Objectives

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

1. understand the theory behind testing...
   - the relationship between two numeric variables
2. estimate a value of the dependent variable (and provide a confidence interval)
3. predict a value of the dependent variable (and provide 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(mod)`
  This provides a regression table on the model previously run
- `confint(mod)`
  This provides a confidence interval for the intercept and slope (effect) parameters, β_0 and β_1
- `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 7, 24
- 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.