Lecture 14: Exam 2 review

Criminology 250

Prof Maria Cuellar

University of Pennslyvania

How is a correlation different from linear regression?

Similarities

- **Strength of relationship**: The two give you different pieces of information about the same question - they each tell you the strength of the linear relationship between x and y.

Differences

- **Symmetry**: Correlation is symmetric, but regression is not: cor(x,y)=cor(y,x), but lm(y~x) $\neq$ lm(x~y). (Note x only moves 1 unit, so y is the thing that's varying.)
- **Point vs. line**: Correlation is not a line, it's only a number that describes the relationship. You could plot a line with the correlation as slope, but you'd have to figure out the units. It won't look like a line of best fit.

Correlation coefficient vs. linear regression slope coefficient

- The estimated slope from a linear regression is defined as

$$\hat{\beta} = cor(Y_i, X_i) \cdot \frac{SD(Y_i)}{SD(X_i)}.$$

  The slope coefficient is the correlation *scaled* by the variability in both x and y.

- Therefore, the two are only equal when $SD(Y_i) = SD(X_i)$, that is, they only coincide when the two variables are on the same scale. This can be achieved by standardizing the data.

- Note: This is mostly an issue of unit differences. Also, in the regression we're

Correlation vs. regression

Let's simulate data that have a linear relationship, where x has a normal distribution with mean 50 and standard deviation 9, and y is defined by the equation of a line, with error (let b_0=10 and b_1=4).

```
x=rnorm(100,50,9)
error=rnorm(100,0,16)
y=10+(4*x)+error
dat <- data.frame(x,y)
```

Scatterplot looks like:

What is the correlation?

```
## [1] 0.91
```

What is the estimated slope coefficient?

```
## [1] 4.27
```

Correlation vs. regression

This is the full regression output:

```
##
## Call:
## lm(formula = y ~ x, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.682 -11.457  -1.769   7.274  33.346
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.5943     9.3699    0.17    0.865
## x             4.1399     0.1849   22.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.17 on 98 degrees of freedom
## Multiple R-squared:  0.8364,    Adjusted R-squared:  0.8348
```

Correlation vs. regression

What does the regression line look like?

Probability and statistics

Probability allows us to describe how likely something is to happen.

How are statistics and probability related?

- Suppose I know the exact distribution of car purchases in Pennsylvania. Then I can find the probability that the first car I see on the road is a Ford. This is probabilistic reasoning since I know the population and predict a sample.
- Suppose I do not know the distribution, but I would like to estimate it. I can observe a random sample of cars in the street and use it to estimate the proportions of the population. This is statistical inference.

Linear regression in R

Usual process is:

1. You assume the variables have a linear relationship.
2. You fit a model.
3. You draw diagnostic plots (and read R^2) to see if the model had a good fit.
4. If it did then you can interpret the coefficients and do inference. If it didn't you try something else (e.g. transforming y or x, adding polynomial terms), and start again in step 1.

Assumptions

Why are these important?

When you use linear regression, you are making four assumptions.

1. Linearity: The relationship between X and the mean of Y is linear
2. Independence: Observations are independent of each other.
3. Homoscedasticity: The variance of residual is the same for any value of X.
4. Normality: For any fixed value of X, Y is normally distributed.

Residuals vs. x plot

- For a simple linear regression model, if the predictor on the x axis is the same predictor that is used in the regression model, the residuals vs. predictor plot offers no new information to that which is already learned by the residuals vs. fits plot.
- On the other hand, if the predictor on the x axis is a new and different predictor, the residuals vs. predictor plot can help to determine whether the predictor should be added to the model (and hence a multiple regression model used instead).
- The interpretation of a "residuals vs. predictor plot" is identical to that for a "residuals vs. fits plot." That is, a well-behaved plot will bounce randomly and form a roughly horizontal band around the residual = 0 line. And, no data points will stand out from the basic random pattern of the other residuals.

Interpret coefficients after a transformation

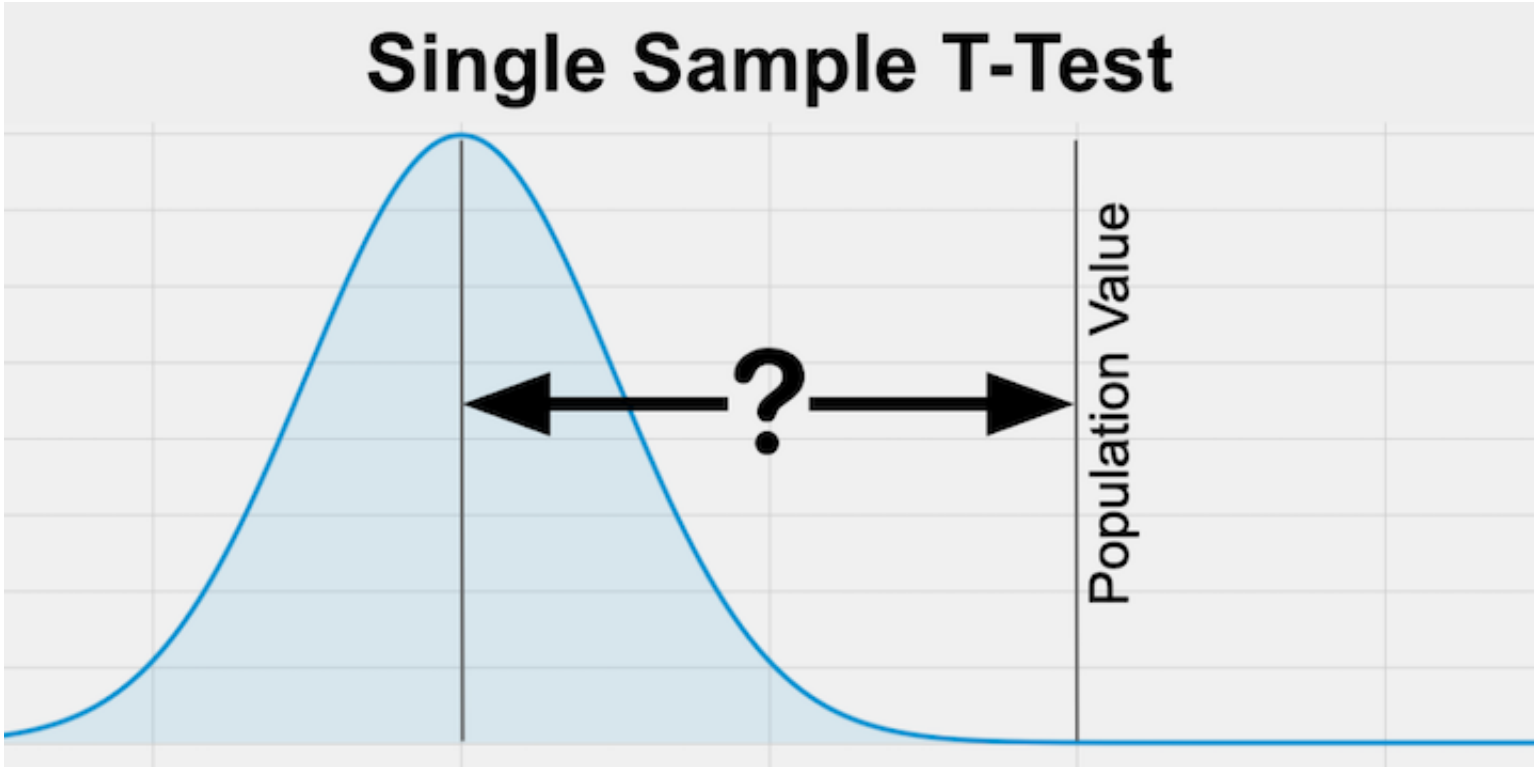
TABLE 2.3 Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

- **Inference**: How to read this table? e.g. log-level: A 1 unit higher in x is associated with a 100 β_1 percent increase in y.
- **Prediction**: To make a prediction you just solve for y and plug in the value of x you want. e.g. If you used $log(y)$, then you solve for

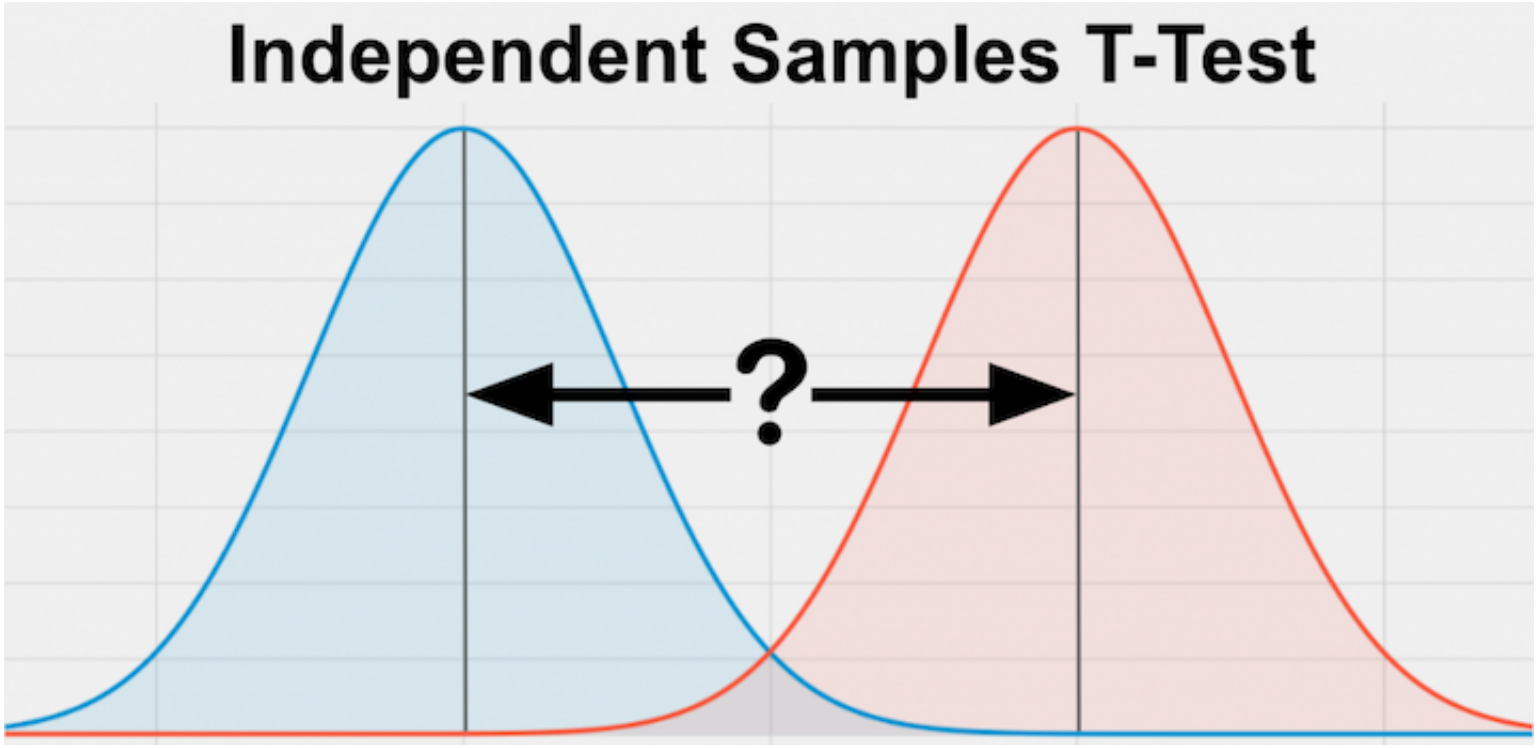
Hypothesis testing

Why do we need to do this? Why can't we just find the slope (after making sure the diagnostics look good) and be done with it?



Hypothesis testing

Why do we need to do this? Why can't we just find the slope (after making sure the diagnostics look good) and be done with it?



p-value

p-value: How likely is it that, if the null hypothesis were true and there really is no relationship between x and y, by chance I were to observe a statistic as extreme as this one? We want this probability to be small. How small? It depends on the significance value you've selected (usually $\alpha = 0.05$).

We want error rates to be low

Type I and Type II Error

Null hypothesis is...	True	False
Rejected	**Type I error** False positive Probability = α	**Correct decision** True positive Probability = 1−β
Not rejected	**Correct decision** True negative Probability = 1−α	**Type II error** False negative Probability = β