ECON 3818: Introduction to Statistics with Computer Applications

---

Kyle Butts

November 30, 2020

Chapter 26: Regression Inference

Introduction

Chapter 4 and 5 discussed how scatterplots and lines of best fit show us linear relationships, but there are remaining questions

- Is there really a linear relationship between x and y, or is the pattern just by chance
    - Spurious correlations – think Nick Cage and number of swimming pool drownings
- What is the slope that explains how y responds to x *in the population*, what is the margin of error for that estimate?
- If we use the least-squares line to predict y for a given x, how accurate is that prediction?

Regression Review

We can model the linear relationship between X and Y by thinking of a conditional expectation:

$$E(Y|X) = a + bX$$

We want estimates for a and b, $\hat{\alpha}$ and $\hat{\beta}$, and we find these estimates by minimizing the sum of squared residuals

$$\varepsilon_i = Y_i - \widehat{Y}_i$$

where,

$$\widehat{Y}_i = \hat{\alpha} + \hat{\beta}X$$

OLS Estimators

Once we minimize the sum of least squares, $\sum_{i=1}^{n} \varepsilon_i^2$, we can obtain the Ordinary Least Squares estimators

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\hat{\beta} = r_{XY}\frac{s_Y}{s_X}$$

Next Steps

This chapter will answer

- How can I interpret $\hat{\alpha}$ and $\hat{\beta}$?
- Inference from a Regression
- What conditions are necessary for those interpretations?

Interpreting α and β

Example: We are studyin a species of coral and how sea surface temperature affects the calcification rate. The equation of the regression line is shown below:

$$\hat{Y} = -12.103 + 0.4615x$$

We can now predict how temperature, x, affects the calcification rate, y. The R^2 will tell us how much of the variation in calcification rate is due to temperature, but it will not tell us whether this relationship is statistically significant.

In order for this regression to be meaningful, we must determine whether the results are statistically significant

Estimating the Parameters

When the conditions for the regression are met:

- The slope $\hat{\beta}$ of the least-squares line is an unbiased estimator of the population slope β
- The intercept $\hat{\alpha}$ of the least-squares line is an unbiased estimator of the population intercept α

Now we only need to estimate the remaining parameters, σ

Regression Standard Error

Our regression model is:

$$y = \alpha + X\beta + \varepsilon$$

ε is the error term that describes why an individual doesn't fall directly on regression line $\alpha + X\beta$.

We denote the variance of ε as σ^2. σ describes variability of response variable y about the population regression line.

- The least-squares line estimates the population regression line
- The residuals are the deviations of data points from the least-squares line

$$\hat{\varepsilon} \equiv \text{residual} = \text{observed } y - \text{predicted } y = y - \hat{y}$$

Therefore we estimate σ by the sample standard deviation of the residuals, known as the regression standard error

Regression Standard Error

$$s = \sqrt{\frac{1}{n-2}\sum \text{residual}^2}$$

$$s = \sqrt{\frac{1}{n-2}\sum (y - \hat{y})^2}$$

We use s to estimate the standard deviation, σ, of responses about the mean given by the population regression line

We will use this error to determine whether our predictions are statistically significant

Testing the Hypothesis of No Linear Relationship

To answer questions about whether associations between two variables are statistically significant, we must test a hypothesis about the slope β:

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

If we fail to reject H_0:

- Regression line with slope 0 is horizontal – meaning y does not change at all when x changes
- H_0 says that there is no linear relationship between X and Y

If we reject H_0, and accept H_1:

- There is some linear relationship between X and Y

Sampling Distribution of $\hat{\beta}$

Since $\hat{\beta}$ is a function of our data, it has a sampling distribution.

The sampling distribution of $\hat{\beta}$ is:

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sigma_X^2}\right)$$

σ^2 is the variance of ε and σ_X^2 is the variance of X.

▸ Visualization of Sampling Distribution

Significance Test for Regression Slope

To test the hypothesis, $H_0 : \beta = 0$, compute the t-statistic:

$$t_{n-2} = \frac{\hat{\beta}}{SE_{\hat{\beta}}}$$

Important to note that the degrees of freedom for the t-statistic for testing a regression slope is $n - 2$

In this formula, the standard error of the least-squares slope is:

$$SE_{\hat{\beta}} = \frac{s}{\sqrt{\sum(x - \bar{x}^2)}}$$

Example

We fit a least-squares line to the model, Price $= \alpha + \beta$(age) with 28 observations from items sold at antiques show. A summary of the output is below:

Parameter	Parameter Estimate	St. Error of Estimate
α	27.73	34.84
β	1.893	0.267

Suppose we want to test the hypothesis, $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$. The value of this t-statistic is:

$$t_{26} = \frac{b}{SE_b} = \frac{1.893}{0.267} = 7.09$$

Using t-table $\implies$ $p < 0.001$

Clicker Question

In the previous example we rejected the null hypothesis of $\beta = 0$, meaning we claim there is sufficient evidence to say there is a linear relationship between age and price sold of items at a antiques road show.

What type of error would we have committed if it turned out there was no relationship between age and price?

(a) Type I, reject the null even though its true

(b) Type II, reject the null even though its true

(c) Type I, fail to reject a false null

(d) Type II, fail to reject a false null

Additional Example – Exam Style

My budtender friend Eric did a study on marijuana consumption and hot cheeto consumption. He surveyed 25 of his friends and collected the following regression results. Assume $\alpha = 0.05$

Cheeto Consumption	**Estimate**	**Std. Error**	**t-statistic**	**p-value**
Intercept	21	12.3		
Joints Smoked	4.20	1.80		

(a) Fill in the rest of the table

(b) Is the intercept statistically significant? Why?

(c) Is the slope coefficient statistically significant? Why?

(d) Interpret slope coefficient

Hypothesis Testing Example

Example:

Regression analysis provides estimates on the relationship between daily wine consumption on risk of breast cancer. The estimated slope was $\hat{\beta} = 0.009$ with a standard error of $SE_{\hat{\beta}} = 0.001$ based off 25 observations.

We want to test whether these results are strong enough to reject the null hypothesis $H_0 : \beta = 0$ in favor or the alternative hypothesis $H_1 : \beta > 0$

Hypothesis Testing Example

So we have $\hat{\beta}$=0.009 and $SE_{\hat{\beta}}$=0.001. Solving hypothesis test:

- Find t-stat

$$t = \frac{0.009}{0.001} = 9$$

- Use t-table to find p-value

$$\text{25 observations} \implies t_{n-2} = t_{23}$$

$$t_{23}^{0.0005} = 3.8 \implies p < 0.0005$$

- Interpret p-value

$$p < 0.0005 \implies p < 0.05 \implies \textbf{Reject } H_0$$

Regression Results

```
Call:
lm(formula = inc95 ~ yeared, data = nlsy)

Residuals:
   Min     1Q Median     3Q    Max
-51516 -13417  -4417   6583 115004

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36230.8     2674.8  -13.54   <2e-16 ***
yeared        5137.3      194.5   26.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22750 on 3612 degrees of freedom
Multiple R-squared:  0.1618,    Adjusted R-squared:  0.1616
F-statistic: 697.4 on 1 and 3612 DF,  p-value: < 2.2e-16
```

Confidence Interval for Regression Slope

The slope, β, of the population regression is usually the most important parameter in a regression problem

- The slope is the rate of change of the mean response as the explanatory variable increases
- The slope explains how changes in x affect outcome variable y

A confidence interval is useful because it shows us *how accurate the estimate of β is likely to be*.

Confidence Interval for Regression Slope

A level C confidence interval for the slope β of the population regression line is

$$\hat{\beta} \pm t^* SE_{\hat{\beta}}, \text{ where } t^* = t_{n-2}^{\frac{1-C}{2}}$$

Confidence Interval for Regression Slope

Example:

Recall our regression results looking at the relationship of temperature on coral calcification. The estimated slope was $\hat{\beta} = 0.4615$ and a standard error $SE_{\hat{\beta}} = 0.07394$. Note this was based off a sample of 12 observations.

If we want to construct a 95% confidence interval:

$$\hat{\beta} \pm t^* SE_{\hat{\beta}} = 0.4615 \pm (2.23)(0.07394)$$

This is because our 12 observations, mean our t_{n-2} distribution has 12-2=10 degrees of freedom and that critical t-stat is 2.23 when $(1 - C)/2 = 0.05/2 = 0.025$

The 95% confidence interval for population slope β is 0.297 to 0.626.

Clicker Question

A random sample of 19 companies were selected and the relationship between sales (in hundreds of thousands of dollars) and profits (in hundreds of thousands of dollars) was investigated by a regression, $\textit{profits} = \alpha + \beta \textit{sales}$. The following results were obtained from statistical software:

Parameter	Parameter Estimate	Std. Error
α	-176.644	61.16
β	0.0925	0.0075

An approximate 90% confidence interval for the slope β is:

(a) -176.66 to -176.63

(b) 0.079 to 0.106

(c) 0.071 to 0.114

Confidence Intervals

Some software programs will automatically spit out a 95% confidence interval associated with slope estimates, like STATA for example:

```
. reg inc95 yeared
```

Source	SS	df	MS
Model	3.6100e+11	1	3.6100e+11
Residual	1.8696e+12	3612	517597115
Total	2.2306e+12	3613	617369553

Number of obs	3614
F(1, 3612)	697.44
Prob > F	0.0000
R-squared	0.1618
Adj R-squared	0.1616
Root MSE	22751

inc95	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
yeared	5137.333	194.5283	26.41	0.000	4755.937	5518.729
_cons	-36230.81	2674.839	-13.55	0.000	-41475.16	-30986.47

95% confident that an additional year of schooling increases predicted income by \$4,755.9 - \$5,518.7

Significance and Margin of Error

- Conducting a hypothesis test on $\hat{\beta}$ tells you about the significance of your result
  - p-value $< \alpha$, we can say our coefficient is statistically different from zero
- A confidence interval says something about the precision of the coefficient
  - What are the ranges of coefficient values we expect the true-value to be in between

Significance and Margin of Error

income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
phd	27331.35	12423.74	2.20	0.028	2973.103	51689.6
_cons	33668.65	413.3218	81.46	0.000	32858.28	34479.02

What we're looking at is the effect of having a PhD on predicted income

- p-value=0.028, which is < 0.05 (standard α)
  - Enough evidence to overturn null hypothesis that a PhD has no affect on income
- Confidence interval $= [2973.1, 51689.6]$
  - Suggests that the effect of PhD varies greatly. 95% confident the true mean is between about 3,000 and 50,000.

Categorical Variable inside Regression

In that previous example, the explanatory variable was categorical. Let's see how that changes interpretation.

income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
phd	27331.35	12423.74	2.20	0.028	2973.103	51689.6
_cons	33668.65	413.3218	81.46	0.000	32858.28	34479.02

This regression implies the relationship between PhD and income is:

$$Income = 33,668.65 + 27,331.35 PhD$$

The takeaways here would be:

- Without a PhD, predicted income is \$33,668.65
- With a PhD, predicted income is \$33,668.65 + \$27,331.35 = \$61,000

Conditions for Regression Inference

Say we have n observations regarding explanatory variable x and response variable y.

- The mean response $E(Y|X)$ has a straight-line relationship with x, given by a population regression line

$$\mu = \alpha + \beta X$$

- For any fixed value of x, the response variable y varies according to a normal distribution
- Repeated responses y are independent of each other
- The standard deviation of ε, σ, is the same for all values of x. The value of σ is unknown.

Intuition about Conditions

The mean response $E(Y \mid X)$ has a straight-line relationship with x, given by a population regression line

- In practice, we observe y for many different values of x. Eventually we see an overall linear pattern formed by points scattered about the population line.

Intuition about Conditions

For any fixed value of x, the response variable y varies according to a normal distribution

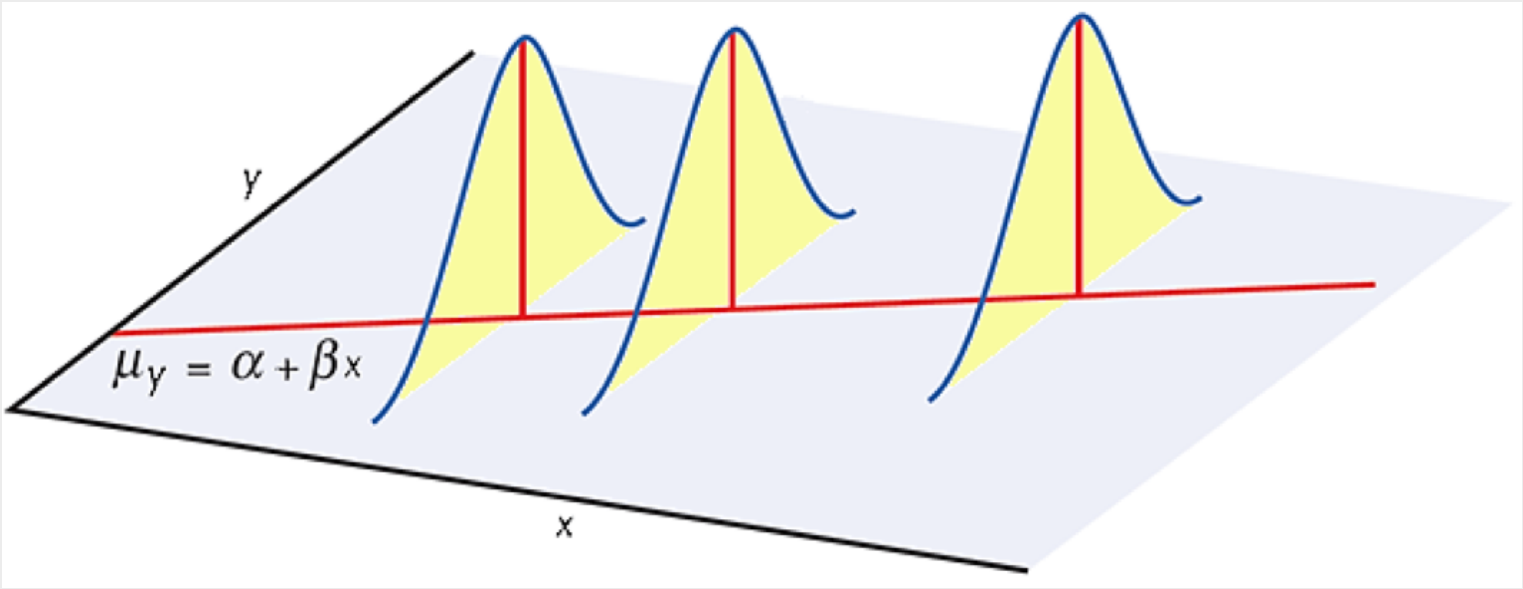
- We cannot observe the entire population regression line. The values of y that we do observe vary about their means according to a normal distribution. If we hold x constant and take many observations of y, the Normal pattern will eventually appear in a histogram.

Intuition about Conditions

The standard deviation of ε, σ, is the same for all values of x. The value of σ is unknown.

- The standard deviation determines whether the points fall close to the population regression line (small σ) or are widely scattered (large σ)
- If σ changes depending on x, then our sample distribution would be wrong.

Intuition about Conditions



- for each possible value of x, the mean of the responses moves along the population regression line
- For a fixed x, the responses y follow a normal distribution with std. dev σ
    - value of σ determines whether points fall close to the line
- the normal curve shows how y will vary when x is held constant

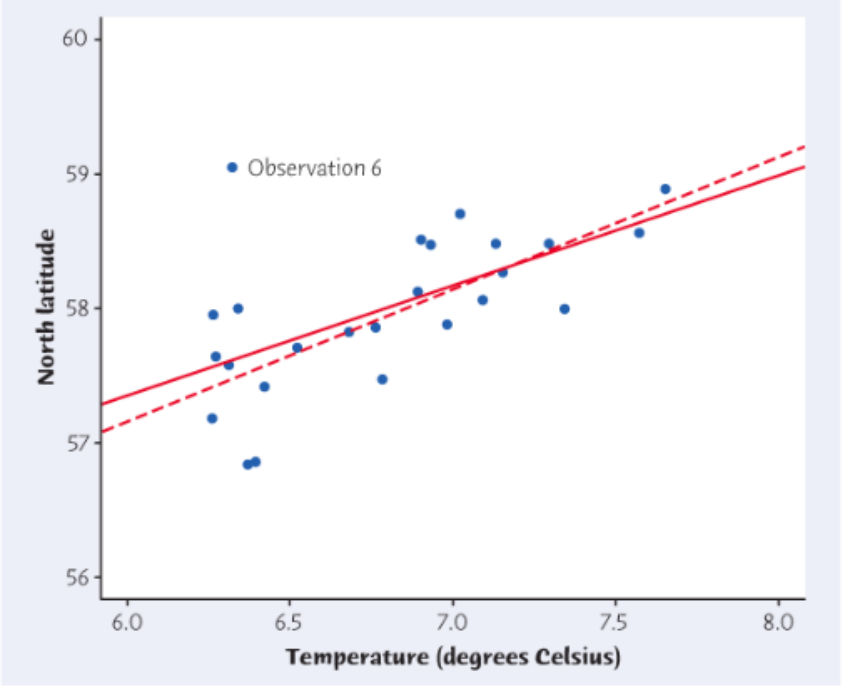
Checking Conditions for Inference

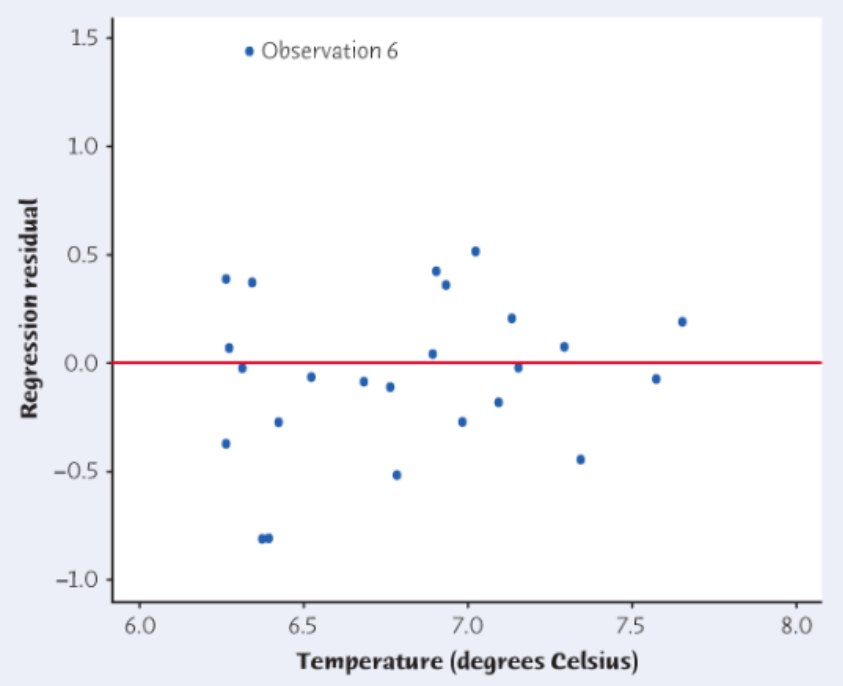
Remember, all of this discussion about inferences hinges on the data meeting certain conditions.

- The relationship is linear in the population
- The response varies normally about the regression line
- Observations are independent
- The standard deviation of the responses is the same for all values of x

Checking Conditions for Inference

In order to check these conditions, it can be helpful to look at a residual plot. A residual plot plots the residuals against the explanatory variable x, with a horizontal line at the "residual =0" position. The "residual =0" line represents the position of the least-squares line in the scatterplot of y against x.





Checking Conditions for Inference

- The relationship is linear. Look for curved patterns or other deviations from an overall straight line pattern in residual plot
- The response varies normally about regression line. Check for departures from normality in your stemplot or histogram of residuals.
- Observations are independent. Signs of dependence in the residual plot are subtle, so usually use common sense.
- Standard deviation of responses is same for all values of x. Look at the scatter of residuals above and below the "residual =0" line. The scatter should be roughly the same from one end to the other.

Residual Histogram