**Inference**

Inference for regression parameters

Is x linearly related to y?

Can you use x to predict y?

Construct a CI and/or perform a hypothesis test for β1 to answer these questions!

Suppose β1 = 0. The population model becomes

Y = β0 + β1x + ε­ = β0 + 0x + ε= ­β0 + ε

Example plot where β0 = 3:



As x changes, E(Y) does not change; thus, x is not linearly related to Y. Our inferences about β1 are typically centered on whether or not β1 = 0.

One can show that



has a t distribution with ν = n – 2 degrees of freedom, where we use  as a random variable in the above quantity and

.

Thus,



Rearranging some terms in the probability produces,



This leads to the (1-α)100% CI for β1:



where we use  and  as observed values.

Similarly, the hypothesis test for β1 = 0 can use the following procedures.

CI method:

1. Ho: β1 = 0 (no linear relationship)
Ha: β1 ≠ 0 (linear relationship)
2. Calculate the (1-α)100% CI
3. Decide whether or not to reject Ho by checking if 0 is in the interval
4. State a conclusion in terms of the problem

Reject H­o – There is sufficient evidence to show that x is linearly related to Y

Don’t Reject Ho – There is not sufficient evidence to show that x is linearly related to Y

where \_\_\_\_ means to put in what x and Y represent in the problem

Test statistic method:

1. Ho: β1 = 0 (no linear relationship)
Ha: β1 ≠ 0 (linear relationship)
2. Calculate the test statistic:



1. State the critical value: ±tα/2, n-2
2. Decide whether or not to reject Ho
3. State a conclusion in terms of the problem

P-value method:

1. Ho: β1 = 0 (no linear relationship)
Ha: β1 ≠ 0 (linear relationship)
2. Calculate the p-value: p-value = 2×P(T>|t|) where T is a random variable with a t-distribution with ν=n-2.
3. State α
4. Decide whether or not to reject Ho
5. State a conclusion in terms of the problem

## Example: Sales and Advertising

Is advertising linearly related to sales? Use α = 0.05.

1. Ho: β1 = 0
Ha: β1 ≠ 0
2. We previously calculated . Note that  from using the calculations below.

|  |  |  |
| --- | --- | --- |
| x | y | (x-)2 |
| 1 | 1 | 4 |
| 2 | 1 | 1 |
| 3 | 2 | 0 |
| 4 | 2 | 1 |
| 5 | 4 | 4 |
| Σ |  | **10** |

Then



1. ± t0.05/2, 5-2 = ±3.182

Because 3.6556 > 3.182, reject Ho­.

1. There is sufficient evidence to show that advertising is linearly related to sales.

Example: HS and College GPA (gpa\_regression.R, gpa.csv)

Is HS GPA linearly related to college GPA? Use α = 0.05.

From earlier,

> summary(object = mod.fit)

Call:

lm(formula = College.GPA ~ HS.GPA, data = gpa)

Residuals:

 Min 1Q Median 3Q Max

-0.55074 -0.25086 0.01633 0.24242 0.77976

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.0869 0.3666 2.965 0.008299 \*\*

HS.GPA 0.6125 0.1237 4.953 0.000103 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3437 on 18 degrees of freedom

Multiple R-squared: 0.5768, Adjusted R-squared: 0.5533

F-statistic: 24.54 on 1 and 18 DF, p-value: 0.0001027

and also

> confint(object = mod.fit, level = 0.95)

 2.5 % 97.5 %

(Intercept) 0.3166459 1.8571131

HS.GPA 0.3527077 0.8722805

1. Ho: β1 = 0
Ha: β1 ≠ 0
2. p-value = 0.0001
3. α = 0.05
4. Because 0.0001 < 0.05, reject Ho­
5. There is sufficient evidence to show that HS GPA is linearly related to college GPA

P-value interpretation: If β1 is really 0 in the population, then a test statistic value, t, at least this large in absolute value (4.95) would occur approximately once if the hypothesis test process (take a new sample and perform a new hypothesis test) is repeated 1000 times. In other words, this is very unlikely to occur if β1 = 0. Thus, β1 is not 0 and Ho is rejected.

Examine the scatter plot again to see why this conclusion makes intuitive sense.

Example: Pizza and college GPA (gpa\_regression.R, College\_GPA\_pizza.csv)

Suppose we also want to know if there is a linear relationship between the number of times a student ate pizza during their freshman year of college and their college GPA. The same 20 students are in this sample.

> gpa2 <- read.csv(file = "College\_GPA\_pizza.csv")

> head(gpa2)

 pizza College.GPA

1 2 3.10

2 9 2.30

3 3 3.00

4 5 2.45

5 3 2.50

6 1 3.70

> mod.fit2 <- lm(formula = College.GPA ~ pizza, data =

 gpa2)

> summary(mod.fit2)

Call:

lm(formula = College.GPA ~ pizza, data = gpa2)

Residuals:

 Min 1Q Median 3Q Max

-0.92519 -0.39645 -0.06745 0.41992 0.94080

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.9417 0.2082 14.132 3.48e-11 \*\*\*

pizza -0.0165 0.0358 -0.461 0.651

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5252 on 18 degrees of freedom

Multiple R-squared: 0.01166, Adjusted R-squared: -0.04325

F-statistic: 0.2123 on 1 and 18 DF, p-value: 0.6505

> plot(x = gpa2$pizza, y = gpa2$College.GPA, xlab = "Times

 ate pizza", ylab = "College GPA", main = "College GPA

 vs. Pizza consumption", ylim = c(0,4), col = "black",

 pch = 1, lwd = 2, panel.first = grid(col = "gray", lty

 = "dotted"))

> curve(expr = mod.fit2$coefficients[1] +

 mod.fit2$coefficients[2]\*x, col = "red", lty = "solid",

 lwd = 2, add = TRUE, xlim = c(min(gpa2$pizza),

 max(gpa2$pizza)))

Use α = 0.01 for the hypothesis test.

1) Ho: β1 = 0

Ha: β1 ≠ 0

2) p-value = 0.65

3) α = 0.01

4) Because 0.65 > 0.01, don’t reject Ho­

5)There is not sufficient evidence to show that the number of times a student ate pizza during their freshman year is linearly related to college GPA.

Examine the scatter plot with the sample model plotted upon it to see why this conclusion makes intuitive sense.



Making inferences about β0

This is much less frequently done, so we will not discuss it in our course.

CIs for E(Y)

We have seen earlier how the sample regression model can be used to estimate E(Y). Simply, use



for a particular x to obtain a . Of course, this does not say how good the estimate is. We can again use probability distributions to develop an interval estimate instead that shows how good an estimate is!

Suppose we want to estimate the average value of Y at a value of x denoted by xh. This value does NOT need to be within our original sample! Some books will use xn+1 instead to emphasize this aspect. One can show that



has a t distribution with ν = n – 2 degrees of freedom, where

.

Thus,



Rearranging some terms in the probability produces,



This leads to the (1-α)100% CI for E(Yh):



Example: HS and College GPA (gpa\_regression.R, gpa.csv)

> more.gpa <- data.frame(HS.GPA = c(2, 3, 4))

> predict(object = mod.fit, newdata = more.gpa, se.fit =

 TRUE, interval = "confidence", level = 0.95)

$fit

 fit lwr upr

1 2.311868 2.027943 2.595793

2 2.924362 2.760785 3.087939

3 3.536856 3.208407 3.865306

$se.fit

 1 2 3

0.13514317 0.07785951 0.15633590

$df

[1] 18

$residual.scale

[1] 0.3436896

For example, at x = 2, the estimated value of E(Yh) is 2.31 and the 95% confidence interval for E(Yh) is

2.03 < E(Yh) < 2.60

PIs for Y

The last sub-section found an interval for E(Yh), which is a constant (like a parameter) value. How about an interval for the random variable Yh itself? As you might expect, there will be more variability because Yh is not a constant value – it can vary for multiple observations with the same xh. The difference can be seen in the plot below:



Intuitively, the line is easier to contain within an interval than the observations around it. Thus, a larger variance is needed for an interval to include Yh in order to maintain a particular level of confidence.

One can show that



has a t distribution with ν = n – 2 degrees of freedom, where



Notice that . Using the t distribution, we have



Rearranging some terms in the probability produces,



This leads to the (1-α)100% prediction interval (PI) for Yh:



Notice that we call this a “prediction interval” now rather than a confidence interval. This is commonly done to help differentiate it from the confidence interval for E(Yh) shown earlier.

Example: HS and College GPA (gpa\_regression.R, gpa.csv)

> predict(object = mod.fit, newdata = more.gpa, se.fit =

 TRUE, interval = "prediction", level = 0.95)

$fit

 fit lwr upr

1 2.311868 1.535987 3.087749

2 2.924362 2.184000 3.664724

3 3.536856 2.743599 4.330113

$se.fit

 1 2 3

0.13514317 0.07785951 0.15633590

$df

[1] 18

$residual.scale

[1] 0.3436896

For example, at x = 2, the 95% prediction interval for Yh is

1.54 < Yh < 3.09

Notice that this interval is wider than the previous interval for E(Yh).

A useful plot to summarize the CIs and PIs is to put all possible CIs and PIs within the range of observed x’s. These plotted intervals are often called “bands” because they appear on both sides of the sample regression model. Below is the plot:



At xh = 2, you can see the CI for E(Yh) and the PI for Yh. In order to understand the code, let’s first examine how to save results from the predict() function into an object:

> #Notice no se.fit or interval arguments

> save.pred1 <- predict(object = mod.fit, newdata =

 data.frame(HS.GPA = 2))

> save.pred1

 1

2.106894

> #Notice no se.fit argument

> save.pred2 <- predict(object = mod.fit, newdata =

 data.frame(HS.GPA = 2), interval = "confidence", level

 = 0.95)

> save.pred2

 fit lwr upr

1 2.106894 1.942197 2.271591

> save.pred2[,2]

[1] 1.942197

Using this method to access the predictions and intervals, the curve() function plots the interval bands:

> plot(x = gpa$HS.GPA, y = gpa$College.GPA, xlab = "HS

 GPA", ylab = "College GPA", main = "College GPA vs. HS

 GPA", xlim = c(0,4.5), ylim = c(0,4.5), col = "black",

 pch = 1, cex = 1.0, panel.first = grid(col = "gray",

 lty = "dotted"))

> curve(expr = predict(object = mod.fit, newdata =

 data.frame(HS.GPA = x)), col = "red", lty = "solid",

 lwd = 1, add = TRUE, xlim = c(min(gpa$HS.GPA),

 max(gpa$HS.GPA)))

> curve(expr = predict(object = mod.fit, newdata =

 data.frame(HS.GPA = x), interval = "confidence", level

 = 0.95)[,2], col = "darkgreen", lty = "dashed", lwd =

 1, add = TRUE, xlim = c(min(gpa$HS.GPA),

 max(gpa$HS.GPA)))

> curve(expr = predict(object = mod.fit, newdata =

 data.frame(HS.GPA = x), interval = "confidence", level

 = 0.95)[,3], col = "darkgreen", lty = "dashed", lwd =

 1, add = TRUE, xlim = c(min(gpa$HS.GPA), to =

 max(gpa$HS.GPA))

> curve(expr = predict(object = mod.fit, newdata =

 data.frame(HS.GPA = x), interval = "prediction", level

 = 0.95)[,2], col = "blue", lty = "dashed", lwd = 1, add

 = TRUE, xlim = c(min(gpa$HS.GPA), to = max(gpa$HS.GPA))

> curve(expr = predict(object = mod.fit, newdata =

 data.frame(HS.GPA = x), interval = "prediction", level =

 0.95)[,3], col = "blue", lty = "dashed", lwd = 1, add =

 TRUE, xlim = c(min(gpa$HS.GPA), to = max(gpa$HS.GPA))

> legend(locator(1), legend = c("Sample model", "95% C.I.",

 "95% P.I."), col = c("red", "darkgreen", "blue"), lty =

 c("solid", "dashed", "dashed"), bty = "n", cex = 0.75)