Chapter 6.3 - Inference in Observation

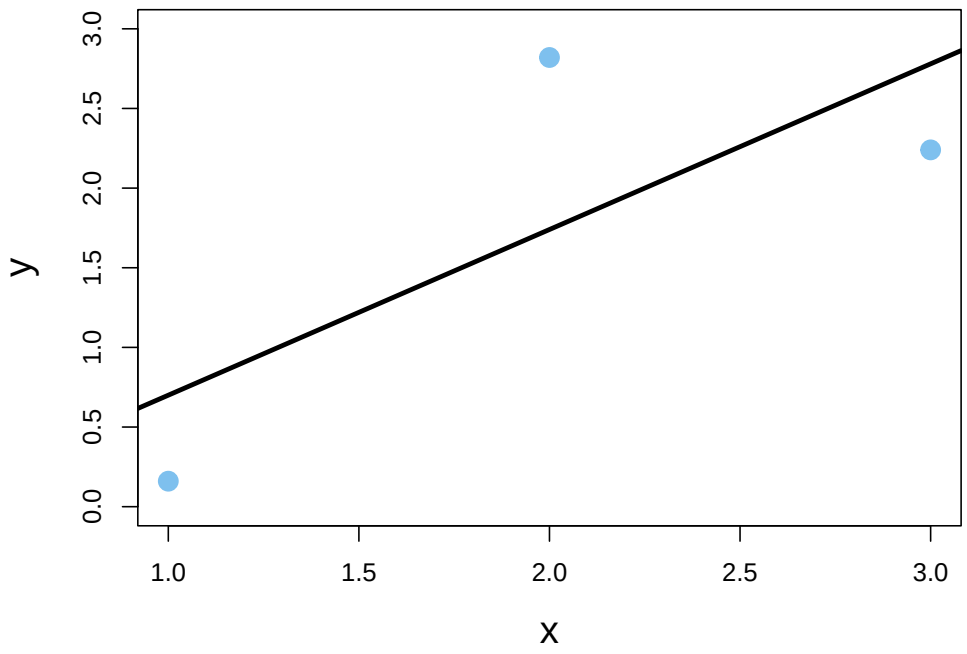
> "Causal interpretation of the results of regression analysis of observational data is a risky business. The responsibility rests entirely on the shoulders of the researcher, because the shoulders of the statistical technique cannot carry such strong inferences." - Jan de Leeuw

It is at this point that we can finally call ourselves statisticians. Everything prior was an introduction to principles, methods, and historical context. Statisticians use these as tools but our actual *profession* involves taking those methods and pulling **statistical inference** out of them.

Standard Error

Through the lens of summary statistics, confidence intervals can feel awkward. Many can agree that testing the efficacy of the *method* of calculating a sample mean is odd. This is natural since the scientific context doesn't make sense until we see a *real* use case, which is best shown with regression. As we've previously addressed we can place confidence intervals on anything— regression coefficients are no exception.

Recall our favorite $n = 3$ dataset:



We calculated a variety of summary statistics in chapter 5 and tossed them in tables:

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	0.16	-1	-1.58	1.58	1
2	2.82	0	1.08	0	0
3	2.24	1	0.5	0.5	1

We calculated the regression coefficients β_0 and β_1:

$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2} = \frac{2.08}{2} = 1.04$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 1.74 - 1.04 \times 2 = -0.34$$

We went through *even more* summary statistics to derive R^2:

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$	$y - \bar{y}$	$(y - \bar{y})^2$
1	0.16	0.7	-0.54	0.2916	-1.58	2.4964
2	2.82	1.74	1.08	1.1664	1.08	1.1664
3	2.24	2.78	-0.54	0.2916	0.50	0.2500

One of those summary statistics, *residual sum of squares*, is especially useful for getting an important piece of the puzzle in constructing regression confidence intervals.

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

$$\text{RSS} = 1.7496$$

We need to look at the error associated with the parameter estimates themselves. This quantity is referred to as **mean squared error**. While we would normally go through all the steps to derive the formula we're about to use; this specific set of derivations would be fairly advanced to work through. What we should note is that mean squared error is the result of dividing RSS by the degrees of freedom of the *regression parameters*. In this case since we have 2 regression parameters the degrees of freedom are $df_r = n - 2$, which is a simple enough calculation given that $n = 3$:

$$\text{MSE} = \frac{\text{RSS}}{n - 2}$$

$$\text{MSE} = \frac{1.7496}{3 - 2} = 1.7496$$

With MSE we have everything we need to "plug-and-chug" the formulas for the **standard errors** of the regression coefficients. Standard error is the analog for standard deviation in the β estimates. We'll once again skip derivations for the sake of brevity and simplicity.

$$SE(\hat{\beta}_0) = \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i + \bar{x})^2}\right)}$$

$$SE(\hat{\beta}_1) = \sqrt{\frac{\text{MSE}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$SE(\hat{\beta}_0) = \sqrt{1.7496\left(\frac{1}{3} + \frac{4}{2}\right)}$$

$$SE(\hat{\beta}_0) = \sqrt{4.0824} = 2.0205$$

$$SE(\hat{\beta}_1) = \sqrt{\frac{1.7496}{2}}$$

$$SE(\hat{\beta}_1) = \sqrt{0.8748} = 0.9353$$

When referencing the t-table we'll use df_r to isolate the t^* value (sticking to the 95% interval convention of course):

df	0.500	0.250	0.200	0.150	0.100	0.050	0.025	0.010	0.001
1	0.000	1.000	1.376	1.963	3.078	6.314	**12.706**	31.821	318.309
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	22.327
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	10.215

The rest follows that same format of Point estimate ± Margin of error, except we replace the standard deviation component with the standard error.

$$CI_{\hat{\beta}_0} = \hat{\beta}_0 \pm t^* \times SE(\hat{\beta}_0)$$

$$CI_{\hat{\beta}_0} = \begin{cases} -0.34 - 12.706 \times 2.0205 \\ -0.34 + 12.706 \times 2.0205 \end{cases}$$

$$CI_{\hat{\beta}_1} = \hat{\beta}_1 \pm t^* \times SE(\hat{\beta}_1)$$

$$CI_{\hat{\beta}_1} = \begin{cases} 1.04 - 12.706 \times 0.9353 \\ 1.04 + 12.706 \times 0.9353 \end{cases}$$

$$CI_{\hat{\beta}_0} = (-26.0125, 25.3325)$$

$$CI_{\hat{\beta}_1} = (-10.8439, 12.9239)$$

This is where the *inference* comes into play. Most statisticians would say two things about these confidence intervals:

1. The intervals are *uninformative*
2. The parameters *lack significance*

The first statement refers to the fact that these intervals are *very* wide. There's no hard rule to define when an interval is too wide. We should instead use our *scientific* inferential skills here to decide if the intervals are reasonable or not. A range of 20 might mean very little if we're talking about the number of fruit flies in an acre of land but it's fairly unhelpful if we're referring to the number of years before a piece of equipment fails.

The second statement refers to the fact that each interval crosses zero; this is the *statistical* inference component of the problem. If the interval is crossing zero what we're saying is that the estimated effects aren't reasonably difference from no effect whatsoever. There's more concrete methods of quantifying how significant an estimate is but even after applying those methods there's a level of critical thinking we have to apply.

Consider the following:

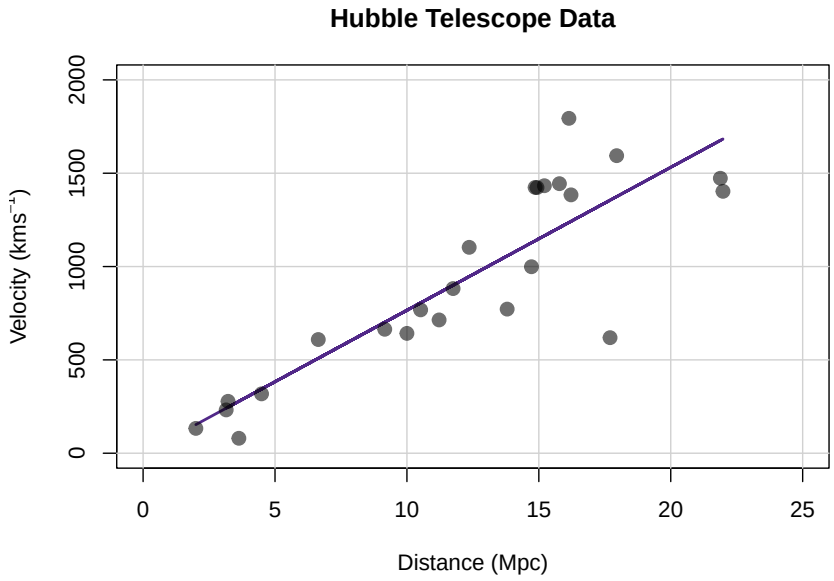
> You have two boxes of cash in front of you and are asked to choose one. Box A typically has \$25 in it, give or take \$5 on occasion. Box B typically has \$50 in it, give or take \$50 on occasion.

If you choose box A you're guaranteed to walk away with *at least* \$20 but if you choose box B you could walk away with nothing. When interpreting confidence intervals we apply the same logic.

It's not the end of the world if an interval crosses zero, usually we can fix the problem by collecting more data. In practice we rarely gather data on the effects of some process or treatment if we don't think there's any effect at all. The magnitude of that effect is another question— one we can answer with sufficient data.

Intervals on Regression

In chapter 5.1 we fit a regression line to some data from the Hubble space telescope:



I introduced a formula that allows us to use the value of $\hat{\beta}_1$ to estimate the age of the universe:

$$\frac{979.708}{\text{Distance}}$$

A long time ago statisticians did all of these calculations by hand. Fortunately it's no longer a long time ago and we can use the `lm()` function (or analogous functions in other programs) to calculate $\hat{\beta}_1$

```
# Y~X-1 removes the intercept term in the lm() function
m=lm(hubble$y~hubble$x-1)
summary(m)[4]
```

```
## $coefficients
##          Estimate Std. Error t value     Pr(>|t|)
## hubble$x 76.58117   3.964794 19.3153 1.031907e-15
```

The answer we'll get from this formula will be in *billions of years*:

$$\frac{979.708}{76.58} = 12.79326$$

Which is a little off from what the robot that always lies (Google Gemini) tell us, but we should expect some error given that this is a dataset of size $n = 24$. What we can do is use the standard error from that same R output to quantify the uncertainty in $\hat{\beta}_1$ and compute a confidence interval for the age of the universe.

We'll use $df_r = n - 1$ since we only have one parameter, giving us $df_r = 23$:

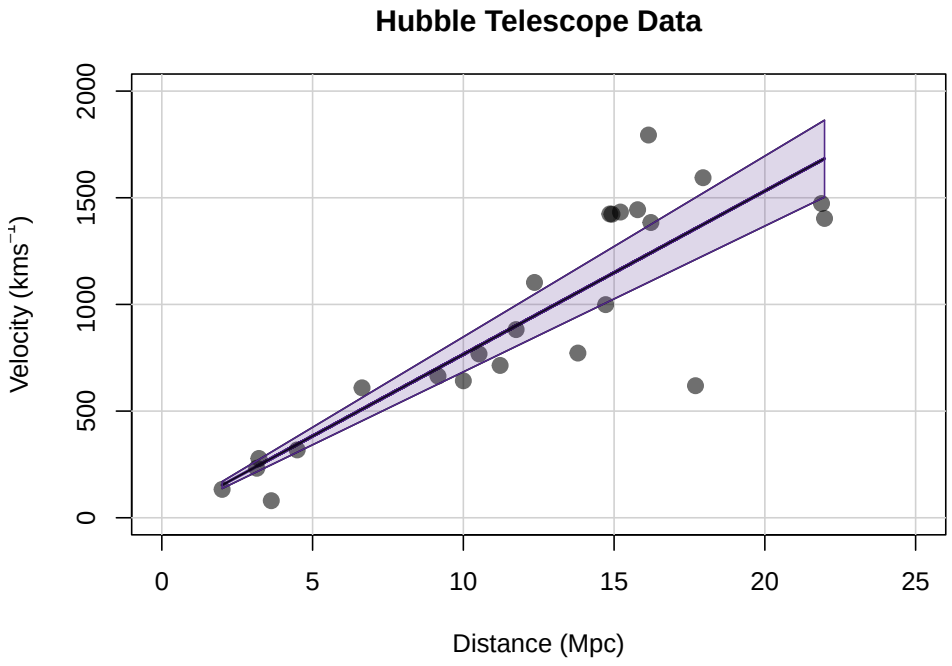
df	0.500	0.250	0.200	0.150	0.100	0.050	0.025	0.010	0.001
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	3.527
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	3.505
23	0.000	0.685	0.858	1.060	1.319	1.714	**2.069**	2.500	3.485
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	3.467

$$CI_{\hat{\beta}_1} = \hat{\beta}_1 \pm t^* \times SE(\hat{\beta}_1)$$

$$CI_{\hat{\beta}_1} = \begin{cases} 76.58117 - 2.069 \times 3.964794 \\ 76.58117 + 2.069 \times 3.964794 \end{cases}$$

$$CI_{\hat{\beta}_1} = (68.39, 84.79)$$

It's good form (tradition) to show confidence intervals graphically whenever you can. For regression we display these in the form of *confidence bands* that wrap around the regression line to show the uncertainty surrounding the fitted line:



A common misconception with confidence bands is that they should contain a certain percentage of the data since they're "95% confident". This is simply untrue; confidence bands have no expectation or requirement to cross any of the data whatsoever. There are intervals specifically designed to contain $X\%$ of the data, **prediction intervals**, but those require a deeper understanding of sampling distributions than this book will cover for quite a while.

Plugging in the lower and upper bound for $\hat{\beta}_1$ will produce an interesting result, in that the lower bound will be the higher value and the upper bound will be the lower value:

$$\frac{979.708}{68.36} = 14.3316$$

$$\frac{979.708}{84.80} = 11.55316$$

But this is a product of two oversights: (1) The equation for the derived quantity (the age of the universe) is inherently going to produce higher value outputs for lower value denominators (2) this isn't the best method for calculating confidence intervals of derived quantities (a discussion for another time).

You might have noticed that the value of t^* is close to 2, which is the approximation we use for $z_{0.05/2} = 1.96$. We'll compute the confidence interval with this value as well and check the differences.

$$CI_{\hat{\beta}_1} = \hat{\beta}_1 \pm z^* \times SE(\hat{\beta}_1)$$

$$CI_{\hat{\beta}_1} = \begin{cases} 76.58117 - 2 \times 3.964794 \\ 76.58117 + 2 \times 3.964794 \end{cases}$$

$$CI_{\hat{\beta}_1} = (68.65, 84.51)$$

$$\frac{979.708}{68.65} = 14.27106$$

$$\frac{979.708}{84.51} = 11.59281$$

While this seems like an inconsequential difference we should recall that the units of these estimates is *billions* of years. So while the above interval is easy to toss together in a pinch, it's about $50,000,000$ years off on average from the "more accurate" interval. Does this make the interval invalid? I won't be the one to say.

The theme of every chapter (and book) about statistical inference should be "use critical thinking skills". It's perfectly rational to use 1.96 or 2.00 to develop your confidence intervals, Fisher did it his entire career and he's worshiped by a *large* proportion of statisticians. It's equally rational to want a higher degree of accuracy and instead use a t^* value. You could even abandon the 95% interval *if the situation calls for it.* The point is that we shouldn't blindly apply techniques and be surprised when they don't work 100% of the time. Use some degree of scientific intuition in all things you do and your analyses will be (shockingly) sound.