4.1 Square one

The simplest setting for the linear model is where there aren’t any \( x \)'s to predict with. The linear model reduces to

\[
Y_i = \mu + \varepsilon_i. \tag{4.1}
\]

If you like, \( Z \in \mathbb{R}^{n \times 1} \) is a column of ones and \( \beta \in \mathbb{R}^{1 \times 1} \) is simply the scalar \( \mu \). The random variables \( \varepsilon_i \) are IID with mean zero and variance \( \sigma^2 > 0 \). We are mainly interested in estimating \( \mu \), but of course learning \( \sigma \) would also have value.

The least squares estimate of \( \mu \) is obviously \( \bar{Y} = (1/n) \sum_{i=1}^{n} Y_i \). The idea of averaging disparate numbers is second nature now but, according to Stephen Stigler (The Seven Pillars of Statistical Wisdom) was once a radical idea. He describes contexts where a physical quantity was measured by multiple observers who all get different values. It seemed more natural then to many people to just reason about which observer was the most careful and use the one best value.

Hold the following thought for later: where it was once controversial to view distinct observations as being sampled from a common random mechanism, now we often suppose that seemingly unrelated data sets are sampled from a random mechanism that makes whole data sets one at time. Then we might learn something about an unseen data set from the ones we have seen.

In class, I used an example of the average age of abalone. It would not have been a radical idea to think that all abalone have different ages.

4.2 Staircase of inference

Suppose we measure 17 abalone ages, thinking that the population is possibly being overharvested if the average age is below 10 years. Maybe the sample average turns out to be 11.4. There is reasonable uncertainty about how that generalizes to the population we sampled from. We know that \( \text{Var}(\bar{Y}) = \sigma^2/n \) where \( \sigma^2 \) is the common variance of \( Y_i \). That very well describes the uncertainty in \( \mu \), but unfortunately we don’t know \( \sigma^2 \). We can however estimate \( \sigma^2 \) by

\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \quad \text{or} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2.
\]

Disclaimer: These notes have not been subjected to the usual scrutiny reserved for formal publications. For instance, citations are mostly omitted or imprecisely made. The notes are meant as a memory aid for students who took Stat 305A at Stanford University. They appear in http://statweb.stanford.edu/~owen/courses/305a/ where you may find the most recent version. Stanford University holds the copyright.

© Stanford University 2019
The first choice is a natural plug-in estimator and the second is the very popular unbiased estimator. Now we might wonder whether \( s^2 \) is a good estimator of \( \sigma^2 \). We can estimate the variance of \( s^2 \) but it will depend on and unknown fourth moment quantity. We end up on Tukey’s staircase of inference: each step is farther from the data using a harder to estimate \( 2^k \)-th moment of the distribution. E.g.,

\[
\text{Var}(s^2) = \sigma^4 \left( \frac{2}{n-1} + \frac{\kappa}{n} \right),
\]

as can be found by grinding it out.

### 4.3 One sample \( t \) test

If \( Y_i \sim (\mu, \sigma^2) \) we might want to test whether \( \mu = \mu_0 \) is plausible for some specific value \( \mu_0 \). Usually the value we want to test is 0. A zero parameter value in a more general regression means that some feature is perfectly unrelated to \( Y \).

In the abalone example, we might test \( \mu_0 = 10 \).

For this test we set up a null hypothesis \( H_0 : \mu = \mu_0 \). We will test this hypothesis either rejecting it or not rejecting it. Not rejecting it is called accepting but it does not mean that we believe \( H_0 \), much less have proved \( H_0 \). I suppose that acceptance might mean that \( H_0 \) is acceptable. At most.

There are exceptional cases where we somehow know that \( \mu \) can only be one of two or a small list of values and there we might be able establish one of those values by rejecting all the others.

We juxtapose the null with an alternative hypothesis. It could be \( H_A : \mu \neq \mu_0 \). In class I mentioned one-sided alternatives such as \( H_A : \mu > \mu_0 \). That is sometimes justified by reasoning that the only possible way for \( \mu \neq \mu_0 \) to arise is \( \mu > \mu_0 \). E.g., even if eating more kale does not help you, there is no way it could be bad. A second justification is that only \( \mu > \mu_0 \) is consequential and worth discovering. Maybe \( \mu < \mu_0 \) has no possible benefit or harm. It is possible to be wrong about both assertions and while that probability is not on the same footing as that driving random \( Y_i \), the chance of being wrong about these things may not be negligible compared to the statistical uncertainty we are working with. Also a common way that one-sided tests come up is post hoc. After failing to reject versus \( \mu \neq \mu_0 \) somebody might notice that one can reject versus \( \mu > \mu_0 \) and then reason that after all that was the only possible/consequential direction.

Using our distribution theory we know that, if \( Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2) \) and if \( \mu = \mu_0 \), then

\[
t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} \sim t_{(n-1)}.
\]

Suppose for \( n - 1 = 16 \) degrees of freedom that we get the value 7.64 shown in Figure 4.1. That value looks to be incompatible with a \( t \) distribution. For the abalone data we might get \( t = 1.438 \), also marked on the figure. That one looks compatible.

From the \( t \) distribution \( \Pr(t = 1.438) \) and \( \Pr(t = 7.64) \) are both zero. The difference is that \( \Pr(|t| \geq 7.64) \approx 10^{-6} \) while \( \Pr(|t| \geq 1.438) \approx 0.17 \). We are using tail areas. The reason to use a tail area is to think of what would happen under \( H_A \). Under \( H_A \), \( |t| \) would be much larger than it would be under \( H_0 \). Larger \( |t| \) is stronger evidence against \( H_0 \).

For this problem we define a \( p \)-value as

\[
p = \Pr\left( |t_{(n-1)}| \geq \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} ; H_0 \right)
\]
Figure 4.1: The curve is the $t_{(16)}$ density. The red mark is an unusually large value for this distribution. The solid point with a dotted line shows $t = 1.438$.

The $t_{(n-1)}$ there is a random variable with the $t$ distribution and the algebraic expression is the observed value (not the random variable that turned out to be the observed value). Ok, let’s write it as

$$p = \Pr \left( |t_{(n-1)}| \geq \frac{|\bar{Y} - \mu_0|}{s/\sqrt{n}} |Y_1, \ldots, Y_n; H_0 \right),$$

but just for now.

When $p$ is small, like $10^{-6}$ we can reason that either $H_0$ is wrong or something of low probability happened. A super common error is to think that $p$ is a probability that $H_0$ is false.

If $H_0$ is false then maybe $\mu \neq \mu_0$. Or maybe the data were not normal or not independent or not identically distributed.

For $t = 1.438$ we get $p = 0.17$. If $\mu = 10$, then there is about a 17% chance that $|t| \geq 1.438$. That casts no material doubt on $H_0$.

When $p$ is small, say $p < \alpha$ for some threshold $\alpha$, we might say that the result is statistically significant at level $\alpha$. The original wording was about whether what we see in the data signifies anything at all about the world.

### 4.4 Statistical vs practical significance

A tiny $p$-value like $10^{-88}$ can seem impressive. It might not be. It could arise from a small effect $\mu - \mu_0$ with an enormous sample size $n$. Practical significance is about the magnitude of $|\mu - \mu_0|$. Our rough idea of it comes from $|\bar{Y} - \mu_0|$. If $|\bar{Y} - \mu_0|$ is large enough to matter for the application area then what we see is practically significant. In class we considered the $2 \times 2$ table of yes or no for practical and statistical significance. Each of those four corners has a different consequence for us.
4.5 Confidence intervals

Equation (4.2) is amazing. We have a formula with our desired unknown $\mu$ surrounded by quantities that we can compute. The combination is of course unknown but its distribution is known. It is called a **pivotal** quantity for this reason. Using the $1 - \alpha/2$ quantile of the $t$ distribution we get

$$1 - \alpha = \Pr\left(\frac{|\bar{Y} - \mu|}{s/\sqrt{n}} \leq t_{(n-1)}^{1-\alpha/2}\right).$$

We can manipulate this algebraically to get

$$1 - \alpha = \Pr\left(\bar{Y} - t_{(n-1)}^{1-\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{Y} + t_{(n-1)}^{1-\alpha/2} \frac{s}{\sqrt{n}}\right).$$

This is a $100(1 - \alpha)\%$ confidence interval for $\mu$.

This is exact. It gets us out of the staircase of inference problem. But only if the data were normally distributed. We can of course test for normality or measure the amount of non-normality, but then that measure would have some kind of error. We’re back on another staircase. Fortunately, the central limit theorem will be of use. The value $\bar{Y}$ will be approximately normally distributed for large enough $n$, without requiring the $Y_i$ to be. Also, bootstrap confidence intervals coming later are effective.

A confidence interval takes the form

$$1 - \alpha = \Pr(L \leq \mu \leq U).$$

(4.3)

[I think I had this under $H_0$ on the board but it is not just under $H_0$ that this probability holds. It holds for any $\mu$. The distribution of $(L, U)$ depends on that $\mu$.] In (4.3), it is the interval $(L, U)$ that is random not the parameter $\mu$. In a Bayesian analysis, coming later, this is reversed. There $L$ and $U$ are fixed (because we have observed them and conditioned on them) and the randomness models our knowledge of $\mu$ given the data.

Confidence intervals allow us to consider statistical and practical significance together. Plot $L$, $U$ and $\mu_0$ on the same number line. If the practical consequences of $L$ and $U$ are the same, then the remaining uncertainty about $\mu$ is not very important. If the interval $[L, U]$ spans a wide range

Confidence intervals and $p$-values have a dual property. We can construct a confidence set out of all of the values of $\mu_0$ that would not have been rejected. Usually that set is an interval. Also the points outside of the confidence set are the ones that are rejected.

4.6 Confidence intervals in practice

Youden looked at confidence intervals for the astronomical unit (the distance from Earth to the sun). Figure 4.2 show them versus the year they came out. No interval contained the following center point. The confidence intervals quantified the uncertainty that the users knew about. There may have been some remaining bias in their estimates.
Figure 4.2: Successive confidence intervals for the astronomical unit (in millions of miles) collected by Youden.