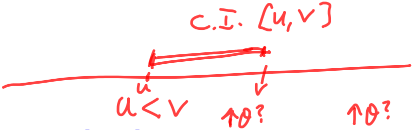
Chapter 6. Confidence intervals and hypothesis testing



- An interval $[u, v]$ constructed using the data $\mathbf{y}$ is said to **cover** a parameter θ if $u \leq \theta \leq v$.
- $[u, v]$ is a 95% **confidence interval** (CI) for θ if the same construction, applied to a large number of draws from the model, would cover θ 95% of the time.
- A **parameter** is a name for any unknown constant in a model. In linear models,each component $\beta_1, \ldots, \beta_p$ of the **coefficient vector** $\boldsymbol{\beta}$ is a parameter. The only other parameter is σ, the standard deviation of the measurement error.

- A confidence interval is the usual way to represent the amount of uncertainty in an estimated parameter.
- The parameter is not random. According to the model, it has a fixed but unknown value.
- The observed interval $[u, v]$ is also not random.
- An interval $[U, V]$ constructed using a vector of random variables $\mathbf{Y}$ defined in a probability model is random. i.e., U and V are random variables.
- If the model is appropriate, then it is reasonable to treat the observed confidence interval $[u, v]$ like a realization from the probability model.
- Call $[u,v]$ the sample confidence interval and $[U,V]$ is a model-generated confidence interval

Not quite a confidence interval for a linear model

- Consider estimating β_1 in the linear model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathrm{MVN}(\mathbf{0}, \sigma^2 \mathbb{I})$. $\hat{\underset{\sim}{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \underset{\sim}{Y}$, a random variable.

- Recall that $\mathrm{E}[\hat{\beta}_1] = \beta_1$ and $\mathrm{SD}(\hat{\beta}_1) = \sigma\sqrt{\left[\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\right]_{11}}$.

note: E[ ] applies to random variables.

this looks like a CI

Question 6.1. Find $\mathrm{P}\left(\hat{\beta}_1 - 1.96\,\mathrm{SD}(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + 1.96\,\mathrm{SD}(\hat{\beta}_1)\right)$

Recall that $\hat{\beta}_1$ has a normal distribution.

Notice that the event $\{\hat{\beta}_1 - 1.96\,\mathrm{SD}(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + 1.96\,\mathrm{SD}(\hat{\beta}_1)\}$

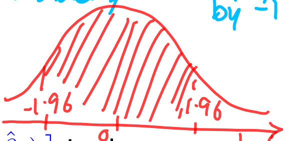
is the same as $\{\beta_1 - 1.96\,\mathrm{SD}(\hat{\beta}_1) \le \hat{\beta}_1 \le \beta_1 + 1.96\,\mathrm{SD}(\hat{\beta}_1)\}$

this defines a region for $\hat{\beta}_1$.

to see they are the same, ~~add~~ subtract β_1 and $\hat{\beta}_1$ from both sides, then multiply by -1.

Thinking in standard units,

$\mathrm{P}\left[\beta_1 - 1.96\,\mathrm{SD}(\hat{\beta}_1) \le \hat{\beta}_1 \le \beta_1 + 1.96\,\mathrm{SD}(\hat{\beta}_1)\right] = 0.95$



standard units

- The interval $\left[\,\hat{\beta}_1 - 1.96\,\mathrm{SD}(\hat{\beta}_1), \hat{\beta}_1 + 1.96\,\mathrm{SD}(\hat{\beta}_1)\,\right]$ is almost a confidence interval. Sadly, we don't know σ.

An approximate confidence interval for a linear model

s is the sample standard deviation of the residuals → $s^2 = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$n-p$: residual degrees of freedom from fitting p parameters.

$\hat{y}_i$: fitted values.

- An approximate 95% CI for β_1 is

$$\boxed{\left[\, b_1 - 1.96\,\mathrm{SE}(b_1)\,,\; b_1 + 1.96\,\mathrm{SE}(b_1) \,\right]}$$

where $\mathbf{y} = \mathbb{X}\mathbf{b} + \mathbf{e}$ with $\mathrm{SE}(b_1) = s\sqrt{\left[\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\right]_{11}}$.

in words, qnorm(p) returns the p^{th} quantile of the standard normal.

- The **standard error** $\mathrm{SE}(b_1)$ is an estimated standard deviation of $\hat{\beta}_1$ under the linear model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathrm{MVN}(\mathbf{0}, \sigma^2\mathbb{I})$.

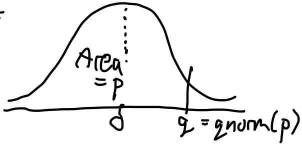
---

qnorm(p) returns a value q such that

$$\int_{-\infty}^{q} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = p$$

e.g.

qnorm(0.5) = 0, qnorm(0.025) = −1.96



A CI for association between unemployment and mortality

pnorm(qnorm(0.975)) = 0.975 ✓

1.96

the z-score of an estimate is its value divided by its SE. Here, this is called a t-value instead, since R compares it to the t distribution.

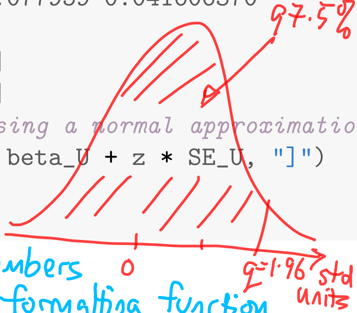
```
c1 <- summary(lm(L_detrended~U_detrended))$coefficients ; c1

##                Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 0.2899928 0.09343146 3.103802 0.002812739
## U_detrended 0.1313673 0.06321939 2.077959 0.041606370

beta_U <- c1["U_detrended","Estimate"]
SE_U <- c1["U_detrended","Std. Error"]
z <- qnorm(1-0.05/2) # for a 95% CI using a normal approximation
cat("CI = [", beta_U - z * SE_U, ",", beta_U + z * SE_U, "]")

## CI = [ 0.0074596 , 0.2552751 ]
```

0.975



cat() is useful for combining numbers & text in output. It is a formatting function.

Interpreting and criticizing a ~~p-value~~ confidence interval.

Question 6.2. We appear to have found evidence that each percentage point of unemployment above trend is associated with about 0.13 years of additional life expectancy, since the 95% CI doesn't include zero. Do you believe this discovery? How could you criticize it?

A CI depends on a probability model. Any assumption in the probability model can be questioned to bring the CI into dispute. Assumptions: errors are independent random variables; relationship is linear. In equation form,

$$\boxed{\text{Assumption: } \underset{\sim}{Y} = X \underset{\sim}{\beta} + \underset{\sim}{\epsilon}, \quad \underset{\sim}{\epsilon} \sim MVN[0, \sigma^2 \mathbb{I}]}$$

The assumption is that we can treat the data as being generated by the probability model. We can think about checking this assumption

"Exogeneity": "Confounding"

econometrics term

statistics term

Even if the model is a good statistical explanation of the data, we must be cautious about causal interpretations

Association is not causation

Association

"Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular manner, is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation." *(John Stuart Mill, A System of Logic, Vol. 1. 1843. p. 470.*

- Put differently: If A and B are associated statistically beyond reasonable statistical doubt we can infer that either A causes B, or B causes A, or both have some common cause C.
- A useful mantra: **Association is not causation.**
- Writing a linear model where A depends on B can show association but we need extra work to argue B causes A. We need to rule out A causing B and the possibility of any common cause C.

Association is not causation: a case study

Question 6.3. Discuss the extent to which the observed association between detrended unemployment and life expectancy in our data can and cannot be interpreted causally.

Unemployment is one of many variables cycling in the boom/bust cycles. We could add more variables (?) But, if we add many variables, it will be hard to distinguish statistically which ones are most relevant.

If we think of unemployment as a measure of the economic cycle, an argument "economic cycle causes life expectancy fluctuations" seems stronger. Is there any plausible variable that explains both the boom/bust cycles and aviation life expectancy fluctuations?

Hypothesis tests

- We try to see patterns in our data. We hope to discover phenomena that will advance science, or help the environment, or reduce sickness and poverty, or make us rich, . . .
- How can we tell whether our new theory is like seeing animals or faces in the clouds?
- From Wikipedia: "**Pareidolia** is a psychological phenomenon in which the mind responds to a stimulus ... by perceiving a familiar pattern where none exists (e.g. in random data)".
- The research community has set a standard: The evidence presented to support a new theory should be unlikely under a **null hypothesis** that the new theory is false. To quantify *unlikely* we need a probability model.

Hypothesis tests and the scientific method

- From a different perspective, a standard view of scientific progress holds that scientific theories cannot be proved correct, they can only be falsified (`https://en.wikipedia.org/wiki/Falsifiability`).
- Accordingly, scientists look for evidence to refute the **null hypothesis** that data can be explained by current scientific understanding.
- If the null hypothesis is inadequate to explain data, the scientist may propose an **alternative hypothesis** which better explains these data.
- The alternative hypothesis will subsequently be challenged with new data.

The scientific method in statistical language

1. **Ask a question**
2. **Obtain relevant data**.
3. **Write a null and alternative hypothesis to represent your question in a probability model**. This may involve writing a linear model so that $\beta_1 = 0$ corresponds to the null hypothesis of "no effect" and $\beta_1 \neq 0$ is a discovered "effect."
4. **Choose a test statistic**. The **sample test statistic** is a quantity computed using the data summarizing the evidence against the null hypothesis. For our linear model example, the least squares coefficient b_1 is a natural sample test statistic for the hypothesis $\beta_1 = 0$.
5. **Calculate the p-value**, the probability that a **model-generated test statistic** is at least as extreme as that observed. For our linear model example, the p-value is $\mathrm{P}\big(|\hat{\beta}_1| > |b_1|\big)$. We can find this probability, when $\beta_1 = 0$, using a normal approximation. b_1 is a constant
6. **Conclusions**. A small p-value (often, < 0.05) is evidence favoring **rejection** of the null hypothesis. The data analysis may suggest new questions: **Return to Step 1**.

$\hat{\beta}_1$ is a random variable.

Using confidence intervals to construct a hypothesis test

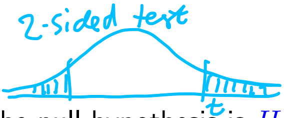
- It is often convenient to use the confidence interval (as a sample test statistic.) to construct a hypothesis test.
- If the confidence interval doesn't cover the null hypothesis, then we have evidence to reject that null hypothesis.
- If we do this test using a 95% confidence interval, we have a 5% chance that we reject the null hypothesis if it is true. This follows from the definition of a confidence interval: whatever the true unknown value of a parameter θ, a model-generated confidence interval covers θ with probability 0.95.

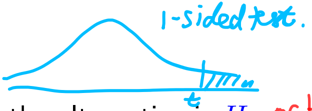
$1-\alpha$

$\mathbb{P}[\text{CI covers } \theta_0 \text{ when } \theta_0 \text{ is true}] = 0.95$

θ_0

Some notation for hypothesis tests





- The null hypothesis is H_0 and the alternative is H_a. or H_1.
- We write t for the sample test statistic calculated using the data $\mathbf{y}$. We write T for the model-generated test statistic, which is a random variable constructed by calculating the test statistic using a random vector $\mathbf{Y}$ drawn from the probability model under H_0.

  |T| is a R.V. |t| is a constant
- The p-value is $\mathrm{pval} = \mathrm{P}\big(|T| \geq |t|\big)$. Here, we are assuming "extreme" means "large in magnitude." Occasionally, it may make more sense to use $\mathrm{pval} = \mathrm{P}\big(T \geq t\big)$.
- We reject H_0 at **significance level** α if $\mathrm{pval} < \alpha$. Common choices of α are $\alpha = 0.05$, $\alpha = 0.01$, $\alpha = 0.001$.

Should we do $P(T>t)$ or $P(|T|>|t|)$? Why? When?

1-sided   2-sided test

1-sided test is more specific; make a 2-sided test unless you have strong scientific reasons for thinking only 1 alternative is possible.

Alternative ways to report a hypothesis test

Question 6.4. When we report the results of a hypothesis test, we can either (i) give the p-value, or (ii) say whether H_0 is rejected at a particular significance level. What are the advantages and disadvantages of each?

Give both! It is almost always good to report the p-value.

(i) Is better if you may want to keep analyzing the data — you can test again later at any level you want.

(ii) The significance level alone is more compact. People sometimes write * for significant at 0.05, ** for significant at 0.01, *** significant at 0.001. When you make a lot of tests, this is convenient, † significant at 0.1.

(iii) p-value alone doesn't reach a conclusion; the test level adds interpretation.

0.05 is the most usual level required for scientific publication. In fields where lots of data are available, stronger evidence is required.

Terminology for test statistics

- Recall that a **sample test statistic** is a summary of the data, constructed to test a hypothesis.
- A **model-generated test statistic** is the same summary applied to random variables drawn from a probability model. Usually, this probability model represents the null hypothesis. We can say "model-generated test statistic under H_0" to make this explicit.
- Distinguishing between sample test statistics and model-generated ones under a null hypothesis is critical to the logic of hypothesis testing.

Example: testing whether $\beta_1 = 0$ in the linear model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$,

- The sample test statistic is $b_1 = \left[\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}\right]_1$.
- A model-generated test statistic is $\hat{\beta}_1 = \left[\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{Y}\right]_1$.

A hypothesis test for unemployment and mortality

$X\ x,\ U\ u,\ V\ v$

Question 6.5. Write a formal hypothesis test of the null hypothesis that there is no association between unemployment and mortality. Compute a p-value using a normal approximation. What do you think is an appropriate significance level α for deciding whether to reject the null hypothesis? **Steps: (1) write the probability model; (2) write the null hypothesis; (3) specify your test statistic; (4) find the distribution of the test statistic under the null hypothesis; (5) calculate the p-value; (6) draw conclusions.**

1. Probability model in subscript form: $Y_i = \beta_1 x_i + \beta_2 + \varepsilon_i$ for $i=1,\dots,n$ with $n=68$ where x_i is detrended unemployment for the i^{th} year and $\varepsilon_i \sim$ iid normal$(0,\sigma)$. β_1 and β_2 are unknown constants. Y_i is a probability model for the data y_i, the detrended life expectancy for the i^{th} year.

$$\mathbb{X} = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$$

for simple linear regression

2. Null hypothesis: $H_0: \beta_1 = 0$, so any observed association between unemployment and life expectancy is just chance variation.

A hypothesis test: continued

3. Test statistic. We use b_1, the sample least squares regression coefficient.

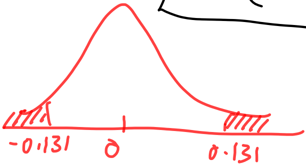
note: this doesn't depend on the value of β_1.

4. Under the null hypothesis, $\hat{\beta}_1 \sim \text{normal}(0, SD(\hat{\beta}_1))$ since $\beta_1 = 0$ under H_0.
where $SD(\hat{\beta}_1) = \sigma\sqrt{[(X^TX)^{-1}]_{11}}$. Since we don't know σ, we use instead $SE(b_1) = s\sqrt{[(X^TX)^{-1}]_{11}}$

Recall
$Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$

5. From the R output above,
$b_1 = 0.131$
$SE(b_1) = 0.063$
p-value $= 2\,\text{pnorm}(-0.131, \text{mean}=0, \text{sd}=0.063)$

z-test: normal approximation. $= 0.0376$.



6. A significance level of 0.05 is typical for this kind of social science data. We can't collect more data — there are only so many recessions to look at. We reject the null, H_0.

Normal approximations versus Student's t distribution

- Notice that `summary(lm(...))` gives `tvalue` and `Pr(>|t|)`.
- The `tvalue` is the estimated coefficient divided by its standard error. This measures how many standard error units the estimated coefficient is from zero.
- `Pr(>|t|)` is similar, but slightly larger, than the p-value coming from the normal approximation.
- R is using Student's t distribution, which makes allowance for chance variation from using s as an approximation to σ when we compute the standard error.
- R uses a t random variable to model the distribution of the statistic t. Giving the full name (Student's t distribution) may add clarity.
- With sophisticated statistical methods, it is often hard to see if they work well just by reading about them. Fortunately, it is often relatively easy to do a **simulation study** to see what is going on.

Simulating from Student's t distribution

- Suppose X and $X_1, \ldots, X_d$ are independent identially distributed (iid) normal random variables with mean zero and standard deviation σ.
- Student's t distribution on d degrees of freedom is defined to be the distribution of $\underbrace{T = X/\hat{\sigma}}_{①}$ where $\hat{\sigma} = \sqrt{\frac{1}{d}\sum_{i=1}^{d} X_i^2}$.
- A normal approximation would say T is approximately $\mathrm{normal}(0, 1)$ since $\hat{\sigma}$ is an estimate of σ.
- With a computer, we can simulate T many times, plot a histogram, and compare it to the probability density function of the normal distribution and Student's t distribution.

Question 6.6. This is almost the same representation of the t distribution as HW4. What is the difference? Why does it not matter?

In HW4, we simulated $\underbrace{T = \frac{Y}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} Z_i^2}}}_{②}$, $Y, Z_1, \ldots, Z_n \sim$ iid normal(0,1)

① rescales the numerator & denominator of ② by the (unknown) σ, so both are equal.

- Here is a different way from HW4 to do the simulation experiment.
- We start by simulating a matrix `X` of iid normal random variables.

```
N <- 50000 ; sigma <- 1 ; d <- 10 ; set.seed(23)
X <- matrix(rnorm(N*(d+1),mean=0,sd=sigma),nrow=N)
```

- Now, we write a function that computes T given $X_1, \ldots, X_d, X$

```
T_evaluator <- function(x) x[d+1] / sqrt(sum(x[1:d]^2)/d)
```

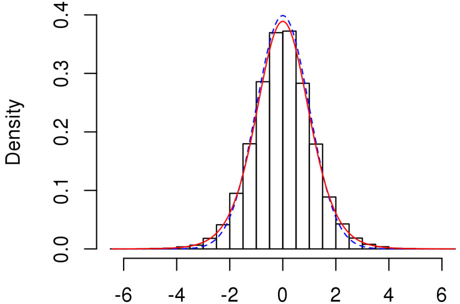
- Then, use `apply()` to evaluate T on each row of 'X'.

```
Tsim <- apply(X,1,T_evaluator)
```

[the HW4 method using replicate() is probably easier]

- We add the normal and t densities to a histogram of the simulations.

```
hist(Tsim,freq=F,main="",
  breaks=30,ylim=c(0,0.4))
x <- seq(length=200,
  min(Tsim),max(Tsim))
lines(x,dnorm(x),
  col="blue",
  lty="dashed")
lines(x,dt(x,df=d),
  col="red")
```



Comparing the normal and t distributions

- Even with as few as $d = 10$ degees of freedom to estimate σ, the Student's t density looks similar to the normal density.
- Student's t has fatter tails. This is important for the probability of rare extreme outcomes.
- Here, the largest and smallest of the $N = 5 \times 10^4$ simulations are

```
range(Tsim)

## [1] -6.438830  6.480262
```

- Let's check the chance of an outcome more than 5 (or 6) standard deviations from the mean for the normal distribution and the t on 10 degrees of freedom.

```
2*(1-pnorm(5))

## [1] 5.733031e-07

2*(1-pnorm(6))

## [1] 1.973175e-09
```

```
2*(1-pt(5,df=d))

## [1] 0.0005373336

2*(1-pt(6,df=d))

## [1] 0.0001321089
```

Hypotheses about predictions from a linear model

- Consider the sample linear model $\mathbf{y} = \mathbb{X}\mathbf{b} + \mathbf{e}$, where $\mathbb{X} = [x_{ij}]_{n\times p}$.
- We might be interested in predicting outcomes at some new set of explanatory variables $\mathbf{x}^* = (x_1^*, \ldots, x_p^*)$, treated as a $1 \times p$ **row vector**.
- Making a prediction involves estimating (i) the expected value of a new outcome; (ii) its variability. In addition, we must make allowance for the statistical uncertainty in these estimates.
- To do inference, we need a probability model. As usual, consider $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\epsilon_1, \ldots, \epsilon_n \sim \text{iid normal}(0, \sigma)$. Also, model a new measurement at $\mathbf{x}^*$ as

$$Y^* = \mathbf{x}^*\boldsymbol{\beta} + \epsilon^*$$

where ϵ^* is another independent draw from the measurement model.

Question 6.7. (a) Why do we want $\mathbf{x}^*$ to be a row vector not a column vector? (b) What is the dimension of $\mathbf{x}^*\boldsymbol{\beta}$?

We have already decided $\underset{\sim}{\beta}$ is a column vector. Then, $\underset{\sim}{x}^*$ must be a row vector to make $\underset{1\times p}{\underset{\sim}{x}^*}\,\underset{p\times 1}{\underset{\sim}{\beta}}$ a scalar

The expected value of a new outcome and its uncertainty

this is a constant.

- According to the model, the expected value of a new outcome at $\mathbf{x}^*$ is

$$\mathrm{E}[Y^*] = \mathbf{x}^* \boldsymbol{\beta}.$$

by linearity of E, $E[\underset{\sim}{x}^* \underset{\sim}{\beta} + \varepsilon] = \underset{\sim}{x}^* \underset{\sim}{\beta} + \underbrace{E(\varepsilon)}_{=0}$

- But, we don't know $\boldsymbol{\beta}$. We estimate $\boldsymbol{\beta}$ by the sample least squares coefficient $\mathbf{b} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}$, which is modeled as a realization of the model-generated least squares coefficient $\hat{\boldsymbol{\beta}} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{Y}$.

this is a random variable.

- A **sample estimate of the expected value** is the **fitted value** at $\mathbf{x}^*$

$\hat{y}^*$ depends on $y_1, \ldots, y_n$ but not on y^*, which has not been collected yet.

$$\hat{y}^* = \mathbf{x}^* \mathbf{b} = \sum_{j=1}^{p} x_j^* b_j.$$

constant, depending on the data

- The **model-generated estimate of the expected value** is

$$\hat{Y}^* = \mathbf{x}^* \hat{\boldsymbol{\beta}} = \sum_{j=1}^{p} x_j^* \hat{\beta}_j.$$

random variable, depends on hypothetical random draws from the model.

- We can find the mean and variance of $\hat{Y}^*$. We can use these (together with a normal approximation) to find a confidence interval for $\mathrm{E}[Y^*]$. If the model is reasonable, this will tell us the uncertainty in using $\hat{y}^*$ to estimate the sample average of many new outcomes collected at $\mathbf{x}^*$.

Q6.8. Use linearity of expectation to show that $E[\hat{Y}^*] = \underset{\sim}{x}^* \underset{\sim}{\beta}$.

previously calculated property of $\hat{\beta}$.

$$E[\hat{Y}^*] = E[\underset{\sim}{x}^* \hat{\underset{\sim}{\beta}}] = \underset{\sim}{x}^* E[\hat{\underset{\sim}{\beta}}] = \underset{\sim}{x}^* \underset{\sim}{\beta}$$

Question 6.9. Use the formula $\mathrm{Var}(\mathbb{A}\mathbf{X}) = \mathbb{A}\mathrm{Var}(\mathbf{X})\mathbb{A}^{\mathrm{T}}$ to show that $\mathrm{Var}[\hat{Y}^*] = \sigma^2 \mathbf{x}^* (\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1} \mathbf{x}^{*\mathrm{T}}$

$$\mathrm{Var}(\hat{Y}^*) = \mathrm{Var}(\underset{\sim}{x}^* \hat{\underset{\sim}{\beta}}) = \underset{\sim}{x}^* \mathrm{Var}(\hat{\underset{\sim}{\beta}}) \underset{\sim}{x}^{*T}$$

$$= \underset{\sim}{x}^* . \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1} . \underset{\sim}{x}^{*T}$$

matrix of x values used for fitting the model

new set of x values for prediction

$$= \sigma^2 \underset{\sim}{x}^* (\underbrace{\mathbb{X}^T}_{p\times n} \underbrace{\mathbb{X}}_{n\times p})^{-1} \underset{\sim}{x}^{*T}$$

$$(1\times p)\times(p\times p)\times(p\times 1)$$

Question 6.10. Check the dimension of $\mathrm{Var}[\hat{Y}^*]$. Is this correct?

$\mathrm{Var}(\hat{Y}^*)$ is 1×1, as required.

note: since $\mathbb{X}$ is $n\times p$, $\underset{p\times n}{\mathbb{X}^T} \underset{n\times p}{\mathbb{X}}$ is $p\times p$ so $(\mathbb{X}^T\mathbb{X})^{-1}$ is $p\times p$.

A CI for the expected value of a new outcome

- We can get a confidence interval (CI) for the **linear combination of coefficients** $\mathbf{x}^*\boldsymbol{\beta}$ in a similar way to what we did for a single coefficient.
- A standard error is $\mathrm{SE}(\mathbf{x}^*\mathbf{b}) = s\sqrt{\mathbf{x}^*\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbf{x}^{*\mathrm{T}}}$.
- Then, making a normal approximation, a 95% CI is $[\mathbf{x}^*\mathbf{b} - 1.96\,\mathrm{SE}(\mathbf{x}^*\mathbf{b})\,,\,\mathbf{x}^*\mathbf{b} + 1.96\,\mathrm{SE}(\mathbf{x}^*\mathbf{b})]$.

Example. We consider again the data on freshman GPA, ACT exam scores and percentile ranking of each student within their high school for 705 students at a large state university. We seek to predict using the probability model considered in the midterm exam, where freshman GPA is modeled to depend linearly on ACT score and high school ranking.

```
gpa <- read.table("gpa.txt",header=T); gpa[1,]

##   ID  GPA High_School ACT Year
## 1  1 0.98          61  20 1996
```

Worked example 6.1. Find a 95% confidence interval for the expected freshman GPA among students with an ACT score of 20 ranking at the 40th percentile in his/her high school.

V is $(X^T X)^{-1}$

```
lm1 <- lm(GPA~ACT+High_School,data=gpa)
x <- c(1,20,40)
pred <- x%*%coef(lm1)
V <- summary(lm1)$cov.unscaled
s <- summary(lm1)$sigma
SE_pred <-sqrt(x%*%V%*%x)*s
c <- qnorm(0.975)
cat("CI = [", round(pred-c*SE_pred,3),
  ",", round(pred+c*SE_pred,3), "]")

## CI = [ 2.344 , 2.532 ]
```

what is the order of the coefficients in this model? we need to use the same order for x here.

e.g. summary(lm1)
names(lm1$coef)
model.matrix(lm1)

we find that the 1st coefficient is called "(intercept)". This needs a value of 1, since x^* is like a new row in the design matrix.

Question 6.11. How would you check whether your answer is plausible? How would you check the R calculation has done what you want it to do?

Sanity check: it should be between 1 and 4.

Sanity check: the predicted value should be in the center of the interval.

$\boxed{\text{Look at the data}}$

A prediction interval for a new outcome

- A 95% **prediction interval** for a new outcome of a linear model with explanatory variables $\mathbf{x}^*$ covers the outcome with probability 95%.

  with new, independent measurement

- The prediction interval allows for the uncertainty around the mean, modeled as **measurement error** in the outcome.

  error

- The prediction interval aims to cover $Y^* = \mathbf{x}^*\boldsymbol{\beta} + \epsilon^*$ whereas the confidence interval for the mean only aims to cover $\mathrm{E}[Y^*] = \mathbf{x}^*\boldsymbol{\beta}$.

- Since ϵ^* is independent of $\mathbf{x}^*\hat{\boldsymbol{\beta}}$ we have

  $Y^* - \underset{\sim}{x}^*\underset{\sim}{\hat\beta} = \underbrace{Y^* - \underset{\sim}{x}^*\underset{\sim}{\beta}}_{\epsilon^*} + \underset{\sim}{x}^*\underset{\sim}{\beta} - \underset{\sim}{x}^*\underset{\sim}{\hat\beta}$

$$\begin{aligned}
\mathrm{Var}[Y^* - \mathbf{x}^*\hat{\boldsymbol{\beta}}] &= \mathrm{Var}[Y^* - \mathbf{x}^*\boldsymbol{\beta}] + \mathrm{Var}[\mathbf{x}^*\boldsymbol{\beta} - \mathbf{x}^*\hat{\boldsymbol{\beta}}] \\
&= \sigma^2 + \sigma^2\mathbf{x}^*\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbf{x}^{*\mathrm{T}}
\end{aligned}$$

Random random

constant, so $\mathrm{Var}(\underset{\sim}{x}^*\underset{\sim}{\beta} - \underset{\sim}{x}^*\underset{\sim}{\hat\beta}) = \mathrm{Var}(\underset{\sim}{x}^*\underset{\sim}{\hat\beta})$.

- This suggests using a standard error for prediction of

  best available estimate of σ

$$\mathrm{SE}_{\mathrm{pred}} = s\sqrt{1 + \mathbf{x}^*\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbf{x}^{*\mathrm{T}}}$$

- A 95% prediction interval, using a normal approximation, is

  two independent terms, the new measurement error & estimation error on β

$$\left[\mathbf{x}^*\mathbf{b} - 1.96\,\mathrm{SE}_{\mathrm{pred}}\,,\ \mathbf{x}^*\mathbf{b} + 1.96\,\mathrm{SE}_{\mathrm{pred}}\right]$$

Using the t distribution for predictions

- We could use a t quantile instead of a normal approximation.
- Just as for parameter confidence intervals, since we use the sample standard deviation s in place of the true standard deviation σ, a t distribution is more accurate.
- With 705 observations, the normal quantile `1.96=qnorm(0.975)` is identical to `1.96=qt(0.975,df=702)` up to 3 significant figures.

3 parameters in the model
705 datapoints
702 residual degrees of freedom.

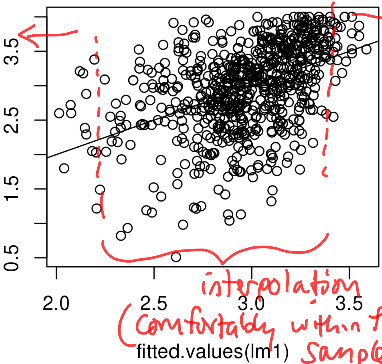
```
plot(x=fitted.values(lm1),y=gpa$GPA,ylab="GPA")
abline(a=0,b=1)
```

model generated fitted value.

$\hat{Y}_i = \mathbf{X}_i \hat{\beta}$

ith row of $\mathbb{X}$

central limit theorem helps us.



extrapolation.

interpolation (comfortably within the range of sampled explanatory variables)

Question 6.12. Is the linear model a good fit for the data? What cautions do you recommend when using this model for prediction?

Use caution at lower GPA, and at higher GPA.

The normal approximation is more important for prediction than for CIs for parameters. The distribution of parameters has a central limit theorem, since $\underset{\sim}{b}$ is a sum of contributions of many data points, via $\underset{\sim}{b} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \underset{\sim}{y}$.

Worked example 6.2. Find a 95% prediction interval for the freshman GPA of an incoming student with an ACT score of 20 ranking at the 40th percentile in his/her high school.

```
lm1 <- lm(GPA~ACT+High_School,data=gpa)
x <- c(1,20,40)
pred <- x%*%coef(lm1)
V <- summary(lm1)$cov.unscaled
s <- summary(lm1)$sigma
SE_pred <-sqrt(x%*%V%*%x + 1)*s
c <- qnorm(0.975)
cat("prediction interval = [", round(pred-c*SE_pred,3),
  ",", round(pred+c*SE_pred,3), "]")

## prediction interval = [ 1.322 , 3.553 ]
```

this is the only difference.

Question 6.13. Where does this calculation differ from the confidence interval for the expected value?

becomes small as n increases

remains the same as n increases.