6. Hypothesis testing and confidence intervals

We have the following goals:

- Understand how to construct confidence intervals for parameters in a linear model.
- Understand how to test statistical hypotheses about a linear model.
- In particular, to ask and answer the question: "Are the data consistent with a hypothesis that a covariate, or a collection of covariates, are unimportant?" (What is the fundamental scientific importance of the slightly contorted logical reasoning in this question?)
- Learn to use R to carry out these tasks.
- See how the linear model includes and extends basic tests for means of one and two samples.

Confidence intervals

- 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. So is the variance σ^2 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.
- If the model is appropriate, then it is reasonable to treat the data $\mathbf{y}$ like a realization from the probability model.

A confidence interval for the coefficient of a linear model

- Consider estimating β_1 in the linear model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- Recall that $\mathrm{E}[\hat{\beta}_1] = \beta_1$ and $\mathrm{SD}(\hat{\beta}_1) = \sigma \sqrt{\left[\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\right]_{11}}$.

Question 6.1. Supposing we can make a normal approximation, show that $\mathrm{P}\big[\, \hat{\beta}_1 - 1.96\,\mathrm{SD}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + 1.96\,\mathrm{SD}(\hat{\beta}_1) \,\big] = 0.95$

- Therefore, an approximate 95% CI for β_1 is

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

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

A CI for association between unemployment and mortality

```
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 ]
```

Interpretation. 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. The 95% CI doesn't include zero.

Question 6.2. Do you believe this discovery? How could you criticize it?

Association is not causation

"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, 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.

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

A review of progress so far in this course

Producing and understanding this confidence interval for a linear model brought together all the things we've done so far in this course.

- We needed to get the data into a computer and run statistical software.
- To understand what the computer was doing for us, and help us to command it correctly, we needed to know about:
  1. matrices
  2. writing a linear model and fitting it by least squares
  3. probability models
  4. expectation and variance
  5. the normal distribution

You could run computer code by learning to follow line-by-line instructions without understanding what the instructions do. But then you wouldn't be in control of your own data analysis.

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 statistics to test 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 $\mathrm{P}\left[|\hat{\beta}_1| > |b_1|\right]$. We can find this probability, when $\beta_1 = 0$, using a normal approximation.
6. **Conclusions**. A small p-value (often, < 0.05) is evidence for **rejecting** the null hypothesis. The data analysis may suggest new questions: **Return to Step 1**.

Using confidence intervals to construct a hypothesis test

- It is often convenient to use the confidence interval as a sample test statistic.
- 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.

Some notation for hypothesis tests

- The null hypothesis is H_0 and the alternative is H_a.
- 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.
- The p-value is $\text{pval} = \text{P}\big[|T| \geq |t|\big]$. Here, we are assuming "extreme" means "large in magnitude." Occasionally, it may make more sense to use $\text{pval} = \text{P}\big[T \geq t\big]$.
- We reject H_0 at **significance level** α if $\text{pval} < \alpha$. Common choices of α are $\alpha = 0.05$, $\alpha = 0.01$, $\alpha = 0.001$.

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?

Careful 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.
- When we just say **test statistic** we are talking about the procedure used to obtain the summary.
- Data analysts don't always explicitly distinguish between sample test statistics and model-generated test statistics. However, the difference 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 $\hat{\beta}_1 = \left[\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{Y}\right]_1$.
- The test statistic is least-squares estimation of the coefficient.

A hypothesis test for unemployment and mortality

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?

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 $X, X_1, \ldots, X_d$ are $d+1$ independent identially distributed (iid) normal random variables with mean zero and standard deviation σ.
- We write $X, X_1, \ldots, X_d \sim \text{iid } N[0, \sigma]$.
- Student's t distribution on d degrees of freedom is defined to be the distribution of $T = X/\hat{\sigma}$ where $\hat{\sigma} = \sqrt{\frac{1}{d}\sum_{i=1}^{d} X_i^2}$.
- A normal approximation would say T is approximately $N[0,1]$ 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.
- The goals in doing this:
  1. Some practice working with Student's t distribution.
  2. Finding how the t distribution compares to the normal distribution as d varies.
  3. Practice the skill of designing a simulation experiment.

- Let's 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 $X_1, \ldots, X_d, X$

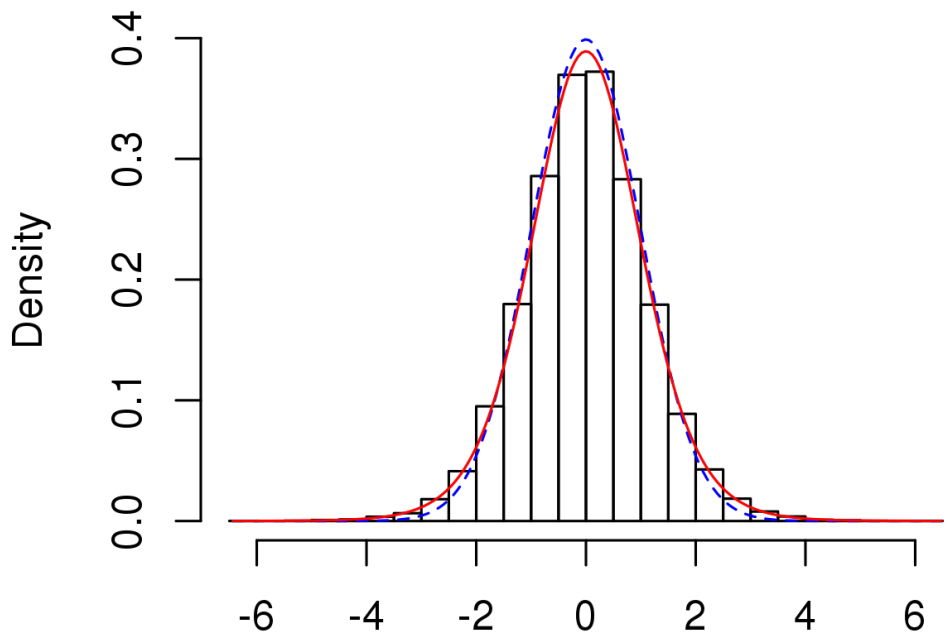
```
T_evaluator <- function(x) x[d+1] / sqrt(sum(x[1:d]^2)/d)
```

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

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

- A histogram of these simulations can be compared with the normal density and the t density

```
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
```

Hypothesis tests for groups of parameters

- We've seen how the least squares coefficient can be used as a test statistic for the null hypothesis that a parameter in a linear model is zero.
- Sometimes we want to test many parameters at the same time. For example, when analyzing the field goal kicking data, we must decide whether to have a separate intercept for each player.

Question 6.6. There are 19 kickers in the dataset. How many extra parameters are needed if we add an intercept for each player?

- This type of question is called **model selection**. Our test statistic should compare **goodness of fit** with and without the additional parameters.
- We need to know the distribution of the model-generated test statistic under the null hypothesis to find the p-value for the test.

Residual sum of squares to quantify goodness of fit

Let $\mathbf{y}$ be the data. Let H_0 be a linear model, $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Let H_a extend H_0 by adding d additional explanatory variables.

- Let RSS_0 be the residual sum of squares for H_0. The residual errors are $\mathbf{e} = \mathbf{y} - \mathbb{X}\mathbf{b}$ where $\mathbf{b} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}$. So, $\mathrm{RSS}_0 = \sum_{i=1}^{n} e_i^2$.
- Let RSS_a be the residual sum of squares for H_a.
- Residual sum of squares is a measure of goodness of fit. A small residual sum of squares suggests a model that fits the data well.

Question 6.7. It is always true that $\mathrm{RSS}_a \leq \mathrm{RSS}_0$. Why?

- We want to know how much smaller RSS_a has to be than RSS_0 to give satisfactory evidence in support of adding the extra explanatory variables into our model. In other words, when should we reject H_0 in favor of H_a?

The f statistic for adding groups of parameters

Formally, we have $H_0 : \mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \epsilon$ and $H_a : \mathbf{Y} = \mathbb{X}_a\boldsymbol{\beta}_a + \epsilon$, where $\mathbb{X}$ is an $n \times p$ matrix and $\mathbb{X}_a = [\,\mathbb{X}\,\mathbb{Z}\,]$ is an $n \times q$ matrix with $q = p + d$. Here, $\mathbb{Z}$ is a $n \times d$ matrix of additional explanatory variables for H_a. As usual, we model $\epsilon_1, \ldots, \epsilon_n$ as iid $N[0, \sigma]$.

- Consider the following sample test statistic:

$$f = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_a)/d}{\mathrm{RSS}_a/(n-q)}.$$

- The denominator is an estimate of σ^2 under H_a. Using this denominator **standardizes** the test statistic.
- The numerator $(\mathrm{RSS}_0 - \mathrm{RSS}_a)/d$ is the **change in RSS per degree of freedom**. Parameters in linear models are often interpreted as degrees of freedom of the model.
- Let F be a model-generated version of f, with the data $\mathbf{y}$ replaced by a random vector $\mathbf{Y}$. If H_0 is true, then the RSS per degree of freedom should be about the same on the numerator and the denominator, so $F \approx 1$. Large values, $f \gg 1$, are therefore evidence against H_0.

The F test for model selection

- Under H_0, the model-generated F statistic has an F distribution on d and $n - q$ degrees of freedom.
- Because of the way we constructed the F statistic, its distribution under H_0 doesn't depend on σ. It only depends on the dimension of $\mathbb{X}$ and $\mathbb{X}_a$.
- We can obtain p-values for the F distribution in R using `pf()`. Try `?pf`.
- Testing H_0 verus H_a using this p-value is called the F test.
- When we add a single parameter, so $d = 1$ and $q = p + 1$, the F test is equivalent to carrying out Student's t test using the estimated coefficient as the test statistic. As homework, you are asked to check this using `pt()` and `pf()` in R.
- Degrees of freedom are mysterious. The mathematics for how they work involves matrix algebra beyond this course. An intuition is that fitting a parameter that is not in the model "explains" a share of the residual sum of squares; in an extreme case, fitting n parameters to n data points may give a perfect fit (residual sum of squares $=$ zero) even if none of these parameters are in the true model.

The F test is called "analysis of variance"

- The F test was invented before computers existed.
- Working out the sums of squares efficiently, by hand, was a big deal!
- Sums of squares of residuals are relevant for estimating variance.
- Building F tests is historically called **analysis of variance** or abbreviated to **ANOVA**.
- The sums of squares and corresponding F tests are presented in an **ANOVA table**. We will see one in the following data analysis.

An F test for kickers. (i) Reviewing the data

```
goals <- read.table("FieldGoals2003to2006.csv",header=T,sep=",")
goals[1:5,c("Name","Teamt","FGt","FGtM1")]
```

```
##              Name Teamt  FGt FGtM1
## 1 Adam Vinatieri    NE 73.5  90.0
## 2 Adam Vinatieri    NE 93.9  73.5
## 3 Adam Vinatieri    NE 80.0  93.9
## 4 Adam Vinatieri   IND 89.4  80.0
## 5    David Akers   PHI 82.7  88.2
```

```
lm0 <- lm(FGt~FGtM1+Name,data=goals)
```

- This is model syntax we have not seen before.
- `Name` is a **factor**

```
class(goals$Name)
```

```
## [1] "factor"
```

- A factor is a vector with **levels**. Here, the levels are the kicker names.

An F test for kickers. (ii) Checking the design matrix

```
X <- model.matrix(lm0)
dim(X)
```

```
## [1] 76 20
```

```
unname(X[c(1,5,9,13,17),1:8])
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]    1 90.0    0    0    0    0    0    0
## [2,]    1 88.2    1    0    0    0    0    0
## [3,]    1 72.2    0    1    0    0    0    0
## [4,]    1 82.1    0    0    1    0    0    0
## [5,]    1 80.0    0    0    0    1    0    0
```

Question 6.8. Is this the design matrix that you want? Can we use our experience working with design matrices to understand what R is doing?

An F test for kickers. (ii) Interpreting the ANOVA table

```
anova(lm0)


## Analysis of Variance Table
##
## Response: FGt
##           Df Sum Sq Mean Sq F value    Pr(>F)
## FGtM1      1   87.2  87.199  2.2597 0.1383978
## Name      18 2252.5 125.137  3.2429 0.0003858 ***
## Residuals 56 2161.0  38.589
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 6.9. Focus on the row labeled `Name`. Explain what is being tested, how it is being tested, and what you conclude.

Predicting future outcomes using 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**.

Question 6.10. Why do we want $\mathbf{x}^*$ to be a row vector not a column 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 $\epsilon_1, \ldots, \epsilon_n$ are iid $N[0, \sigma]$. 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.
- Note that the dimension of $\mathbf{x}^*\boldsymbol{\beta}$ is $(1 \times p) \times (p \times 1) = 1 \times 1$.

The expected value of a new outcome and its uncertainty

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

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

- 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 $\hat{\boldsymbol{\beta}} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{Y}$.
- A **sample estimate of the expected value** is the **fitted value** at $\mathbf{x}^*$

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

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

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

- 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 $\mathbf{x}^*$.

Question 6.11. Show that $\mathrm{E}[\hat{Y}^*] = \mathbf{x}^*\boldsymbol{\beta}$

Question 6.12. Show that $\mathrm{Var}[\hat{Y}^*] = \sigma^2\mathbf{x}^*\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbf{x}^{*\mathrm{T}}$

Question 6.13. Check the dimension of $\mathrm{Var}[\hat{Y}^*]$. Is this correct?

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
```

Question. 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.

Solution

```
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 ]
```

- Notice how R guesses whether to interpret a vector as a row or column, depending on the situation.

Question 6.14. How would you check whether your answer is plausible? How would you check the R calculation has done what you want it to do?

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%.
- The prediction interval allows for the uncertainty around the mean, usually called **measurement error**, in the outcome.
- Formally, 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 $\mathbf{x}^*\hat{\boldsymbol{\beta}}$ (why?), we have

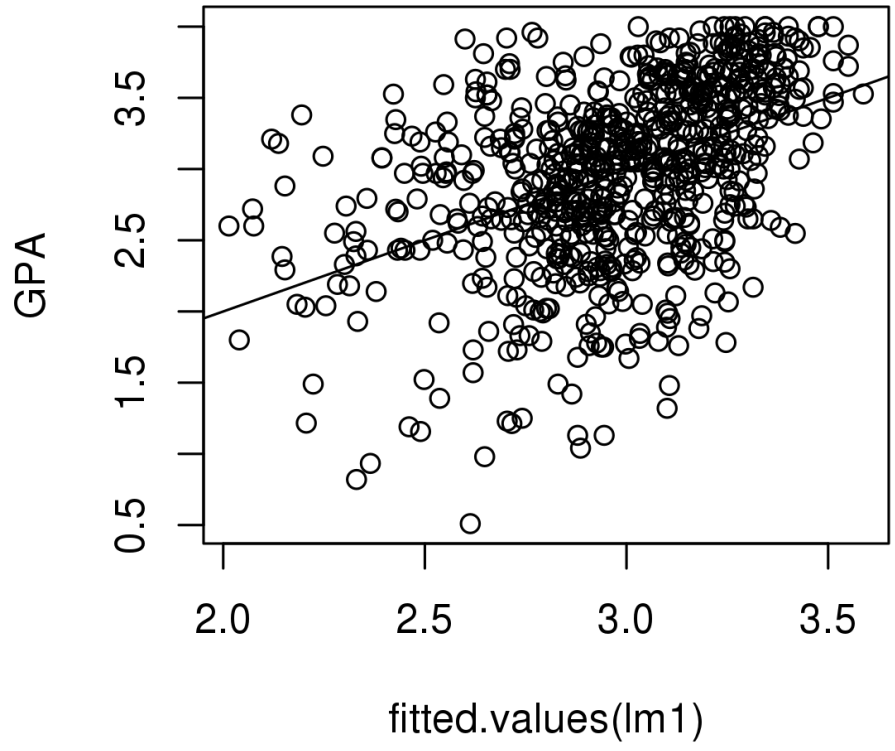
$$\begin{aligned}\mathrm{Var}[Y^* - \mathbf{x}^*\hat{\boldsymbol{\beta}}] &= \mathrm{Var}[Y^* - \mathbf{x}^*\boldsymbol{\beta}] + \mathrm{Var}[\mathbf{x}^*\boldsymbol{\beta} - \mathbf{x}^*\hat{\boldsymbol{\beta}}] \\ &= \sigma^2 + \sigma^2\mathbf{x}^*\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbf{x}^{*\mathrm{T}}\end{aligned}$$

- This suggests using a standard error for prediction of

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

- A 95% prediction interval, using a normal approximation, is $[\mathbf{x}^*\mathbf{b} - 1.96\,\mathrm{SE}_{\mathrm{pred}}\,,\ \mathbf{x}^*\mathbf{b} + 1.96\,\mathrm{SE}_{\mathrm{pred}}]$.
- We could use a t quantile. With 705 observations, the normal quantile `1.96=qnorm(0.975)` is equivalent to `1.96=qt(0.975,df=702)`

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



Question 6.15. Is the linear model a good fit for the data? What cautions do you recommend when using this model for prediction?

Question. 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.

Solution

```
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 ]
```

Question 6.16. Where does this calculation differ from the confidence interval?