Inference Verification

C. Durso

Inference Formulas

One purpose of the following code is to connect the inference formulas from the theory of linear regression to the output of the model-fitting code. A second purpose is to provide visualizations for some of the values in the inference formulas.

Data and Model

Remain with the model of Wealth as a linear function of Commerce in central France around 1830.

```
data("Guerry")
dat<-Guerry

dat.c<-filter(dat,Region=="C")
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
m.c.w<-lm(Wealth~Commerce,data=dat.c)
summary(m.c.w)
```

```
##
## Call:
## lm(formula = Wealth ~ Commerce, data = dat.c)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38.54 -10.99   4.56  10.92  35.76
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.4250     9.3684   2.073  0.05577 .
## Commerce      0.5457     0.1682   3.244  0.00545 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.05 on 15 degrees of freedom
## Multiple R-squared:  0.4123, Adjusted R-squared:  0.3731
## F-statistic: 10.52 on 1 and 15 DF,  p-value: 0.005451
```

Some model sums of squares

```
# SSY
(SSY<-sum((dat.c$Wealth-mean(dat.c$Wealth))^2))
```

```
## [1] 8313.529
```

```
# SSE
(SSE<-sum((dat.c$Wealth-m.c.w$fitted)^2))
```

```
## [1] 4885.904
```

```
# SSR
(SSR<-sum((m.c.w$fitted-mean(dat.c$Wealth))^2))
```

```
## [1] 3427.626
```

```
SSY
```

```
## [1] 8313.529
```

```
SSE+SSR
```

```
## [1] 8313.529
```

$R^2 = (SSY - SSE)/SSY = SSR/SSY$, the percent of the variation in the response variable accounted for by the fitted values. Recall that R^2 equals the square of the correlation of the response variable and the explanatory variable.

Practice

In the code block below, please compute $R^2 = SSR/SSY$. Also, please compute the square of the correlation of "$dat.cWealth$"and"$dat.cCommerce$". Confirm that both results equal the value of R^2 given in the summary of the model "(m.c.w)".

```
(SSR/SSY)
```

```
## [1] 0.4122949
```

```
cor(dat.c$Wealth,dat.c$Commerce)
```

```
## [1] 0.6421019
```

```
cor(dat.c$Wealth,dat.c$Commerce)^2
```

```
## [1] 0.4122949
```

```
summary(m.c.w)$r.squared
```

```
## [1] 0.4122949
```

p-value of regression

The p-value of the regression can be calculated from the statistic $\frac{\frac{SSY-SSE}{(n-1)-(n-2)}}{\frac{SSE}{n-2}}$. Under the null hypothesis that $m = 0$, this has an F-distribution with numerator degrees of freedom equal to $(n-1)-(n-2) = 1$ and denominator degrees of freedom equal to $n-2$.

```
n<-nrow(dat.c)
(s2<-SSE/(n-2))
```

```
## [1] 325.7269
```

```
(f.stat<-SSR/s2)
```

```
## [1] 10.523
```

```
(pf(f.stat,1,n-2,lower.tail=FALSE))
```

```
## [1] 0.005450814
```

```r
(pf((SSY-SSE)/((n-1)-(n-2))/(SSE/(n-2)),1,n-2,lower.tail=FALSE ))
```

```
## [1] 0.005450814
```

```r
summary(m.c.w)
```

```
##
## Call:
## lm(formula = Wealth ~ Commerce, data = dat.c)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38.54 -10.99   4.56  10.92  35.76
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.4250     9.3684   2.073  0.05577 .
## Commerce      0.5457     0.1682   3.244  0.00545 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.05 on 15 degrees of freedom
## Multiple R-squared:  0.4123, Adjusted R-squared:  0.3731
## F-statistic: 10.52 on 1 and 15 DF,  p-value: 0.005451
```

Standard error of slope

Practice

Compute the standard error of the slope. Confirm the p-value for the slope. Recall that $s^2\frac{1}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$ *is a formula for the standard error of the slope. This is equivalent to* $s^2\frac{1}{\sum_{i=1}^{n}X_i^2-n\bar{X}^2}$*. The value* s^2 *is stored in the variable "s2" above. Confirm that the standard error computed this way equals the standard error from the summary of "m.c.w".*

```r
sqrt(s2/sum((dat.c$Commerce-mean(dat.c$Commerce))^2))
```

```
## [1] 0.1682313
```

```r
summary(m.c.w)$coefficients[2,2]
```

```
## [1] 0.1682313
```

```r
# Calculate the p-value of the slope from the definition.
2*pt(-abs(summary(m.c.w)$coefficients[2,1]/summary(m.c.w)$coefficients[2,2]),df=n-2)
```

```
## [1] 0.005450814
```

```r
# Verify that this is the p-value from the summary.
summary(m.c.w)$coefficients[2,4]
```

```
## [1] 0.005450814
```

Standard error of intercept

Recall that

$$s^2\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\right)$$

is a formula for the standard error of the intercept. This is equivalent to $s^2\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n}X_i^2-n\bar{X}^2}\right)$.

```
sqrt(s2*(1/n+mean(dat.c$Commerce)^2/sum((dat.c$Commerce-mean(dat.c$Commerce))^2)))
```

```
## [1] 9.368413
```

```
summary(m.c.w)$coefficients[1,2] # Verify this is the Std. Error from the summary.
```

```
## [1] 9.368413
```

```
# Calculate the p-value of the intercept from the definition.
2*pt(-abs(summary(m.c.w)$coefficients[1,1]/summary(m.c.w)$coefficients[1,2]),df=n-2)
```

```
## [1] 0.05576651
```

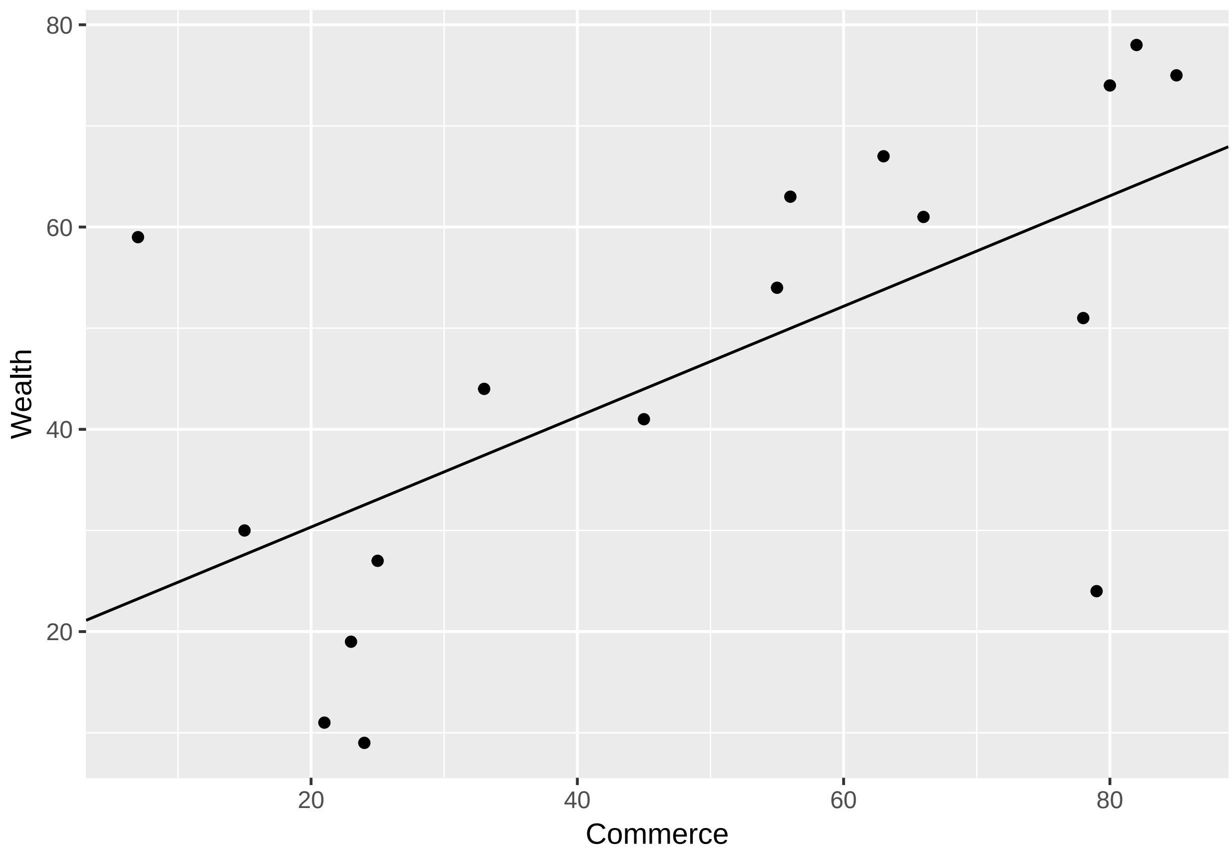
```
# Verify that this is the p-value from the summary.
summary(m.c.w)$coefficients[1,4]
```

```
## [1] 0.05576651
```

Predicted Y Random Variables

In the model that Y is a linear function of X plus iid Normal errors ε, the maximum likelihood slope and intercept are random variables, functions of ε. Thus the predicted value of $\hat{Y}_h$ at X_h is a random variable distributed as $\hat{Y}_h \sim Normal\left(mX_h+b,\sigma^2\left(\frac{1}{n}+\frac{\left(X_h-\bar{X}\right)^2}{\sum\left(X_i-\bar{X}\right)^2}\right)\right)$. The confidence interval based on this uses the corresponding Student's t distribution with $n-2$ degrees of freedom, replacing σ^2 by s^2, that is $\frac{SSE}{n-2}$.

```
g<-ggplot(data=dat.c,aes(x=Commerce,y=Wealth))+geom_point()
g<-g+geom_abline(slope=m.c.w$coefficients[2],intercept=m.c.w$coefficients[1])
g
```



Look at 95% conf. interval for Y.pred

```
x.new<-seq(min(dat.c$Commerce),max(dat.c$Commerce),length.out=50)
temp<-data.frame(Commerce=x.new)
y.pred<-predict(m.c.w,newdata=temp)
```

function to calculate variance for the mean of a new observation

```
s2.new<-function(newx,x,s2){
    n<-length(x)
    return(s2*(1/n+(newx-mean(x))^2/sum((x-mean(x))^2)))
}
```

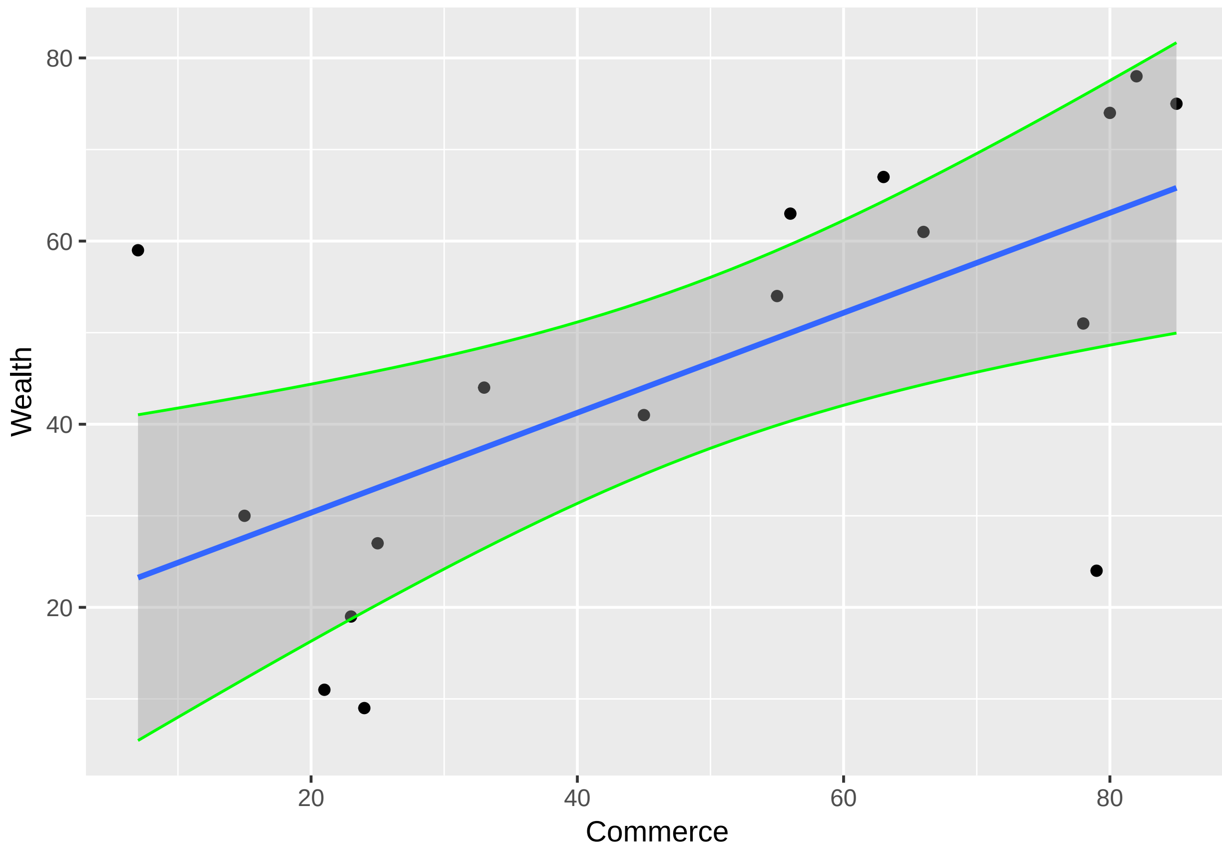
Generate vector of upper and lower bounds for the means

Use the 95% confidence interval. Plot the results.

```
s2s<-s2.new(x.new,dat.c$Commerce,s2) # variance of the mean of an observation at x.new
a<-qt(.975,n-2) # .975 quantile for Student's t with n-2 degrees of freedom
upper<-y.pred+a*sqrt(s2s) # upper bound of 95% CI on y.pred
lower<-y.pred-a*sqrt(s2s) # lower bound of 95% CI on y.pred

dat.new<-data.frame(x=x.new,lower=lower,upper=upper)
```

```r
g<-ggplot(dat.c,aes(x=Commerce,y=Wealth))+geom_point()+geom_smooth(method="lm")+
  geom_line(data=dat.new,aes(x=x,y=lower),color="green")+
  geom_line(data=dat.new,aes(x=x,y=upper),color="green")
g
```

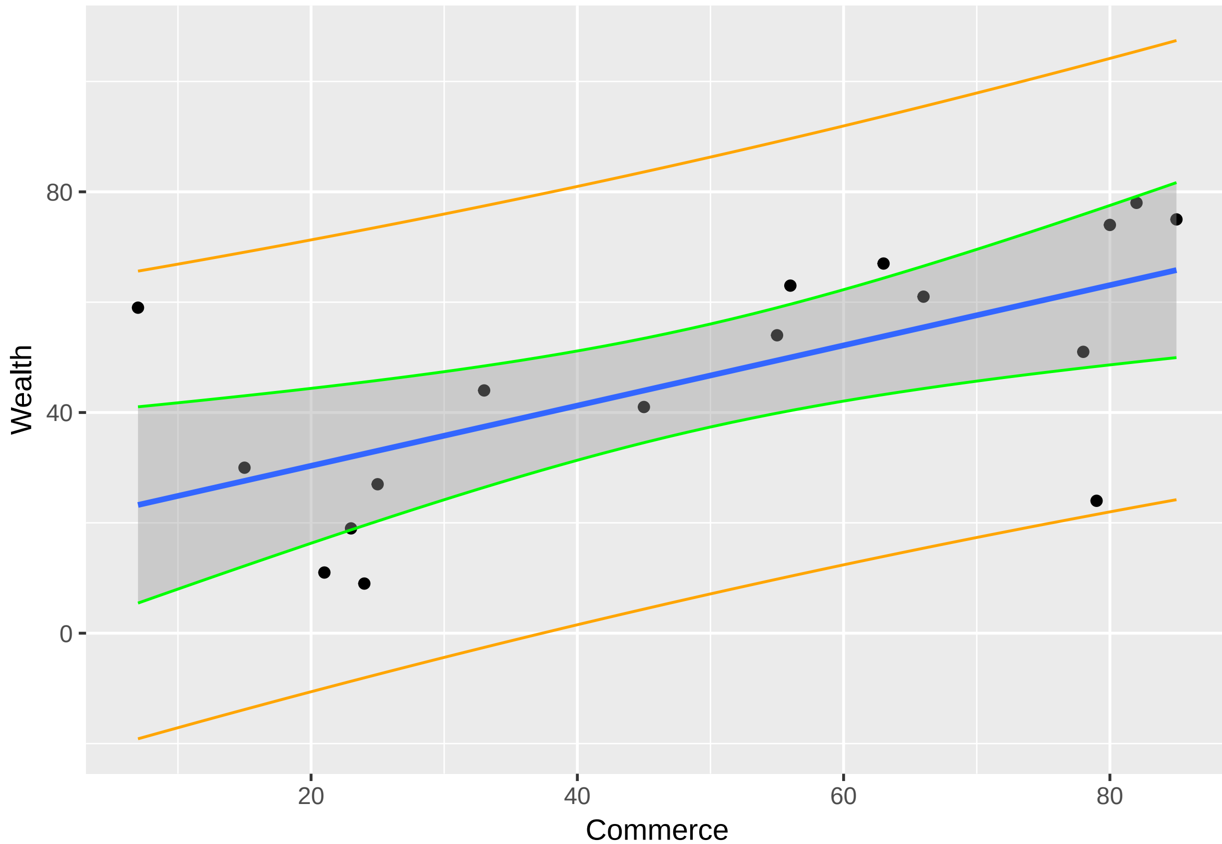


The distribution of a new $\hat{Y}_h new$ is given by $\hat{Y}_h new \sim Normal\left(mX_h + b, \sigma^2\left(\frac{n+1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)\right)$. Again, the confidence intervals corresponding to this use a Student's t distribution with $n-2$ degrees of freedom, replacing σ^2 by s^2.

Practice

Please supply the upper and lower bounds on the 95% confidence intervals for new observations as the indicated columns in dat.new. Run the plotting commands to view these bounds.

```r
# y.pred.new, with error
dat.new$upper.w.error<-y.pred+a*sqrt(s2s+s2)
dat.new$lower.w.error<-y.pred-a*sqrt(s2s+s2)
g<-g+geom_line(data=dat.new,aes(x=x,y=lower.w.error),color="orange")+
  geom_line(data=dat.new,aes(x=x,y=upper.w.error),color="orange")
g
```


Wealth
80
40
0
20
40
60
80
Commerce