SSR, MSE, F-Statistic

Isabella Chittumuri

4/12/2021

```
setwd("~/Documents/Hunter College/Spring 2021/Stat 707/HW")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

7.1 State the number of degrees of freedom that are associated with each of the following extra sums of squares:

(1) $SSR(X_1|X_2)$ has 1 degree of freedom associated with it
(2) $SSR(X_2|X_1, X_3)$ has 1 degrees of freedom associated with it
(3) $SSR(X_1, X_2|X_3, X_4$ has 2 degrees of freedom associated with it

7.5 Refer to Patient satisfaction Problem 6.15.

A hospital administrator wished to study the relation between patient satisfaction (Y) and patient's age (X_1, in years), severity of illness (X_2, an index), and anxiety level (X_3, an index). The administrator randomly selected 46 patients and collected the data presented below, where larger values of Y, X_2, and X_3 are, respectively, associated with more satisfaction, increased severity of illness, and more anxiety.

(a)

Obtain the analysis of variance (ANOVA) table that decomposes the regression sum of squares into extra sums of squares associated with X_2; with X_1, given X_2; and with X_3, given X_2 and X_1.

```r
# import patient satisfaction (PS)
PS <- read.csv("Patient_Satisfaction.csv", header = F)

# names(PS) <- c("satisfaction", "age", "illness_severity", "anxiety_level")
names(PS) <- c("Y", "x1", "x2", "x3")
```

```r
# regression model, have to switch x1 and x2 to obtain anova that decomposes with x2
fit <- lm(Y ~ x2 + x1 + x3, data = PS)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ x2 + x1 + x3, data = PS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
## x2           -0.4420     0.4920  -0.898   0.3741
## x1           -1.1416     0.2148  -5.315 3.81e-06 ***
## x3          -13.4702     7.0997  -1.897   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 42 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6595
## F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

Regression Equation: $Y = 158.5 - 1.142X_1 - 0.442X_2 - 13.47X_3$

```r
fit.aov <- anova(fit)
tab <- as.table(cbind(
  'SS' = c("SSR(x2, x1, x3)" = sum(fit.aov[1:3, 2]),
           "SSR(x2)"         = fit.aov[1, 2],
           "SSR(x1|x2)"      = fit.aov[2, 2],
           "SSR(x3|x2, x1)"  = fit.aov[3, 2],
           "SSE"             = fit.aov[4, 2],
           "Total"           = sum(fit.aov[, 2])),

  'Df' = c(                    sum(fit.aov[1:3, 1]),
                               fit.aov[1, 1],
                               fit.aov[2, 1],
                               fit.aov[3, 1],
                               fit.aov[4, 1],
                               sum(fit.aov$Df)),

  'MS' = c(                    sum(fit.aov[1:3, 2]) / sum(fit.aov[1:3, 1]),
                               fit.aov[1, 3],
                               fit.aov[2, 3],
```

```
                              fit.aov[3, 3],
                              fit.aov[4, 3],
                              NA)
))

round(tab, 2)
```

```
##                        SS       Df       MS
## SSR(x2, x1, x3)  9120.46     3.00  3040.15
## SSR(x2)          4860.26     1.00  4860.26
## SSR(x1|x2)       3896.04     1.00  3896.04
## SSR(x3|x2, x1)    364.16     1.00   364.16
## SSE              4248.84    42.00   101.16
## Total           13369.30    45.00
```

The extra sum of squares associated with $X_2 = 4860.26$, with 1 degree of freedom The extra sum of squares associated with $X_1|X_2 = 3896.04$, with 1 degree of freedom The extra sum of squares associated with $X_3|X_2, X_1 = 364.16$, with 1 degree of freedom

(b)

Test whether X_3 can be dropped from the regression model given that X_1, and X_2 are retained. Use the F* test statistic and level of significance .025. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

Alternatives:

$H_0 : B_3 = 0$ $H_a : B_3 \neq 0$

General Decision Rule

$$F^* \leq F(1-\alpha, df_R - df_F, df_F), \text{ fail to reject } H_0$$
$$F^* > F(1-\alpha, df_R - df_F, df_F), \text{ reject } H_0$$

General F* Test Statistic Formula

$$F^* = \frac{\frac{SSE(R)-SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}}$$

F* test statistic for whether or not $B_3 = 0$ is a marginal test, given X_1, X_2 is in the model

$$F^* = \frac{\frac{SSR(X_3|X_1,X_2)}{1}}{\frac{SSE(X_1,X_2,X_3)}{n-4}}$$
$$F^* = \frac{MSR(X_3|X_1,X_2)}{MSE(X_1,X_2,X_3)}$$
$$F^* = \frac{\frac{SSR(X_3|X_1,X_2)}{1}}{MSE(X_1,X_2,X_3)}$$

```
fit.aov
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x2         1 4860.3  4860.3 48.0439 1.822e-08 ***
## x1         1 3896.0  3896.0 38.5126 2.008e-07 ***
## x3         1  364.2   364.2  3.5997   0.06468 .
## Residuals 42 4248.8   101.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F^* = \frac{\frac{364.16}{1}}{101.16} = 3.599$$

```
# look at the f distribution
# 1-a, df, n-p (46-2)
qf(1-0.025, 1, 42)
```

```
## [1] 5.403859
```

$$3.599 \leq 5.404, \text{ fail to reject } H_0$$

```
# another way to test full and reduced model
f_mod <- lm(Y ~ x2 + x1 + x3, data = PS)
r_mod <- lm(Y ~ x2 + x1, data = PS)
anova(r_mod,f_mod, test="Chisq")
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ x2 + x1
## Model 2: Y ~ x2 + x1 + x3
##   Res.Df    RSS Df Sum of Sq Pr(>Chi)
## 1     43 4613.0
## 2     42 4248.8  1    364.16  0.05779 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Get p-value of test
summary(f_mod)
```

```
##
## Call:
## lm(formula = Y ~ x2 + x1 + x3, data = PS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
## x2           -0.4420     0.4920  -0.898   0.3741
## x1           -1.1416     0.2148  -5.315 3.81e-06 ***
## x3          -13.4702     7.0997  -1.897   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 42 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6595
## F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

The p-value of the F is 0.065, which is greater than 0.05 suggesting that we don't need X_3. In conclusion, based on the F* test statistic and it's p-value, we can omit X_3 from the final model fitting.

Note: the p-value of the test is the same as p-value of the tested parameter in summary function.

7.14 Refer to Patient satisfaction Problem 6.15

(a)

Calculate $R^2_{Y1}, R^2_{Y1|2}$, and $R^2_{Y1|23}$. How is the degree of marginal linear association between Y and X_1 affected, when adjusted for X_2 ? When adjusted for both X_2 and X_3?

```
library(rsq)
x1 <- lm(Y ~ x1, data=PS)
rsq.partial(x1)
```

```
## $adjustment
## [1] FALSE
##
## $variable
## [1] "x1"
##
## $partial.rsq
## [1] 0.6189843
```

```
x2_giv_x1 <- lm(Y ~ x1 + x2, data=PS)
rsq.partial(x2_giv_x1)
```

```
## $adjustment
## [1] FALSE
##
## $variable
## [1] "x1" "x2"
##
## $partial.rsq
## [1] 0.45787094 0.09440975
```

```
x2x3_giv_x1 <- lm(Y ~ x1 + x2 + x3, data=PS)
rsq.partial(x2x3_giv_x1)
```

```
## $adjustment
## [1] FALSE
##
## $variable
## [1] "x1" "x2" "x3"
##
## $partial.rsq
## [1] 0.40211018 0.01885673 0.07894201
```

$$R^2_{Y1} = 0.6189$$
$$R^2_{Y1|2} = 0.4578$$
$$R^2_{Y1|23} = 0.4021$$

The degree of marginal linear association between Y and X_1 decreases by roughly a third when adjusted for X_2. It decreases only a little more when adjusted for X_2 and X_3

(b)

Make a similar analysis to that in part (a) for the degree of marginal linear association between Y and X_2. Are your findings similar to those in part (a) for Y and X_1?

```
x2 <- lm(Y ~ x2, data=PS)
rsq.partial(x2)
```

```
## $adjustment
## [1] FALSE
##
## $variable
## [1] "x2"
##
## $partial.rsq
## [1] 0.3635387
```

```
x1_giv_x2 <- lm(Y ~ x2 + x1, data=PS)
rsq.partial(x1_giv_x2)
```

```
## $adjustment
## [1] FALSE
##
## $variable
## [1] "x2" "x1"
##
## $partial.rsq
## [1] 0.09440975 0.45787094
```

```
x3_giv_x2x1 <- lm(Y ~ x2 + x1 + x3, data=PS)
rsq.partial(x3_giv_x2x1)
```

```
## $adjustment
## [1] FALSE
##
## $variable
## [1] "x2" "x1" "x3"
##
## $partial.rsq
## [1] 0.01885673 0.40211018 0.07894201
```

$$R^2_{Y2} = 0.3635$$
$$R^2_{Y2|1} = 0.094$$
$$R^2_{Y2|13} = 0.0189$$

The degree of marginal linear association between Y and X_2 decreases roughly to a fourth of its value when adjusted for X_1. It decreases to about a fifth of this value when adjusted for X_1 and X_3.

My findings are similar to those in part (a), because R^2 keeps decreasing when adjusting for more parameters.

7.31 The following regression model is being considered in a water resources study:

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \beta_4 \sqrt{X_{i3}} + \epsilon_i$

State the reduced models for testing whether or not:

(1)

$\beta_3 = \beta_4 = 0$ Reduced model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$

(2)

$\beta_3 = 0$ Reduced model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_4 \sqrt{X_{i3}} + \epsilon_i$

7.37

Refer to the CDI data set in Appendix C.2. For predicting the number of active physicians Y in a county, it has been decided to include total population X_1 and total personal income X_2 as predictor variables. The question now is whether an additional predictor variable would be helpful in the model and, if so, which variable would be most helpful. Assume that a first-order multiple regression model is appropriate.

```
# import county demographic information (CDI)
CDI <- read.csv("CDI_Data.csv", header = F)


names(CDI) <- c("ID", "county", "state", "land_area", "total_pop", "precent_pop_18_34", "percent_pop_65
```

(a)

For each of the following variables, calculate the coefficient of partial determination given that X_1 and X_2 are included in the model: land area X_3, percent of population 65 or older X_4, number of hospital beds X_5, and total serious crimes X_6.

```
x3_mod <- lm(num_physicians ~ total_pop + total_income + land_area, data = CDI)

x4_mod <- lm(num_physicians ~ total_pop + total_income + percent_pop_65, data = CDI)

x5_mod <- lm(num_physicians ~ total_pop + total_income + n_hospital_beds, data = CDI)

x6_mod <- lm(num_physicians ~ total_pop + total_income + total_crimes, data = CDI)

full_mod <- lm(num_physicians ~ total_pop + total_income + land_area + percent_pop_65 + n_hospital_beds

rsq.partial(x3_mod)
```

```
## $adjustment
## [1] FALSE
##
## $variable
## [1] "total_pop"    "total_income" "land_area"
##
## $partial.rsq
## [1] 0.01916009 0.10305593 0.02882495
```

```
rsq.partial(x4_mod)
```

```
## $adjustment
## [1] FALSE
##
## $variable
## [1] "total_pop"      "total_income"   "percent_pop_65"
##
## $partial.rsq
## [1] 0.008789391 0.134259566 0.003842367
```

```
rsq.partial(x5_mod)
```

```
## $adjustment
## [1] FALSE
##
## $variable
## [1] "total_pop"       "total_income"    "n_hospital_beds"
##
## $partial.rsq
## [1] 0.1468062 0.3627428 0.5538182
```

```
rsq.partial(x6_mod)
```

```
## $adjustment
## [1] FALSE
##
## $variable
## [1] "total_pop"    "total_income" "total_crimes"
##
## $partial.rsq
## [1] 0.000333806 0.135840559 0.007323408
```

$$R^2_{Y3|12} = 0.028824$$
$$R^2_{Y4|12} = 0.003842$$
$$R^2_{Y5|12} = 0.5538$$
$$R^2_{Y6|12} = 0.007323$$

b

On the basis of the results in part (a), which of the four additional predictor variables is best? Is the extra sum of squares associated with this variable larger than those for the other three variables?

Based on the results in part (a), number of hospital beds (X_5) is the best because it's has the highest coefficient of partial determination.

c

Using the F* test statistic, test whether or not the variable determined to be best in part (b) is helpful in the regression model when X_1 and X_2 are included in the model; use $\alpha = .01$. State the alternatives, decision rule, and conclusion. Would the F* test statistics for the other three potential predictor variables be as large as the one here? Discuss.

Alternatives:

$H_0 : B_k = 0$ $H_a : B_k \neq 0$

General Decision Rule

$$F^* \leq F(1-\alpha, df_R - df_F, df_F), \text{ fail to reject } H_0$$
$$F^* > F(1-\alpha, df_R - df_F, df_F), \text{ reject } H_0$$

F* Test Statistic Formula

$$F^* = \frac{\frac{SSR(X_k|X_1,X_2)}{1}}{MSE(X_1, X_2, X_k)}$$

```
anova(x3_mod)
```

```
## Analysis of Variance Table
##
## Response: num_physicians
##               Df     Sum Sq    Mean Sq  F value    Pr(>F)
## total_pop      1 1243181164 1243181164 3959.184 < 2.2e-16 ***
## total_income   1   22058054   22058054   70.249 7.271e-16 ***
```

```
## land_area      1    4063370    4063370   12.941 0.0003583 ***
## Residuals    436  136903711     313999
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(x4_mod)
```

```
## Analysis of Variance Table
##
## Response: num_physicians
##                 Df     Sum Sq    Mean Sq   F value    Pr(>F)
## total_pop        1 1243181164 1243181164 3859.8919 < 2.2e-16 ***
## total_income     1   22058054   22058054   68.4870 1.571e-15 ***
## percent_pop_65   1     541647     541647    1.6817    0.1954
## Residuals      436  140425434     322077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(x5_mod)
```

```
## Analysis of Variance Table
##
## Response: num_physicians
##                  Df     Sum Sq    Mean Sq F value    Pr(>F)
## total_pop         1 1243181164 1243181164 8617.70 < 2.2e-16 ***
## total_income      1   22058054   22058054  152.91 < 2.2e-16 ***
## n_hospital_beds   1   78070132   78070132  541.18 < 2.2e-16 ***
## Residuals       436   62896949     144259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(x6_mod)
```

```
## Analysis of Variance Table
##
## Response: num_physicians
##                Df     Sum Sq    Mean Sq   F value    Pr(>F)
## total_pop       1 1243181164 1243181164 3873.4274 < 2.2e-16 ***
## total_income    1   22058054   22058054   68.7271 1.414e-15 ***
## total_crimes    1    1032359    1032359    3.2166   0.07359 .
## Residuals     436  139934722     320951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\text{For } X_3 = F^* = \frac{\frac{4063370}{1}}{313999} = 12.94$$
$$\text{For } X_4 = F^* = \frac{\frac{541647}{1}}{322077} = 1.68$$
$$\text{For } X_5 = F^* = \frac{\frac{78070132}{1}}{144259} = 541.18$$
$$\text{For } X_6 = F^* = \frac{\frac{1032359}{1}}{320951} = 3.216$$

```
# look at the f distribution
qf(1-0.01, 1, 436)
```

```
## [1] 6.693358
```

$$\text{For } X_3 = 12.94 > 6.69, \text{ reject } H_0$$
$$\text{For } X_4 = 1.68 \leq 6.69, \text{ fail to reject } H_0$$
$$\text{For } X_5 = 541.18 > 6.69, \text{ reject } H_0$$
$$\text{For } X_6 = 3.216 \leq 6.69, \text{ fail to reject } H_0$$

Number of hospital beds (X_5), the variable determined to be the best in part (a), is helpful in the regression model when X_1 and X_2 are included in the model. This is because it's F* test statistic is greater than the alpha level F statistic, therefore we can reject the null H_0 that $B_5 = 0$

The other three predictor variables F* test statistic is not as large as that of X_5.

8.6

Steroid level. An endocrinologist was interested in exploring the relationship between the level of a steroid (Y) and age (X) in healthy female subjects whose ages ranged from 8 to 25 years She collected a sample of 27 healthy females in this age range. The data are given below:

```
steroid <- read.table('Steroid_level.txt')
names(steroid) <- c("steroid_level", "age")
```

a. Fit regression model (8.2). Plot the fitted regression function and the data. Does the quadratic regression function appear to be a good fit here? Find R2.

(8.2)

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon_i$$

```
# reference on why we do this: lecture 8 slide 8

# centering predictor age
steroid$center_age <- steroid$age - mean(steroid$age)
steroid$center_age_sq <- (steroid$center_age)^2
```

```
lmod <- lm(steroid_level ~ center_age + center_age_sq, data=steroid)
summary(lmod)
```

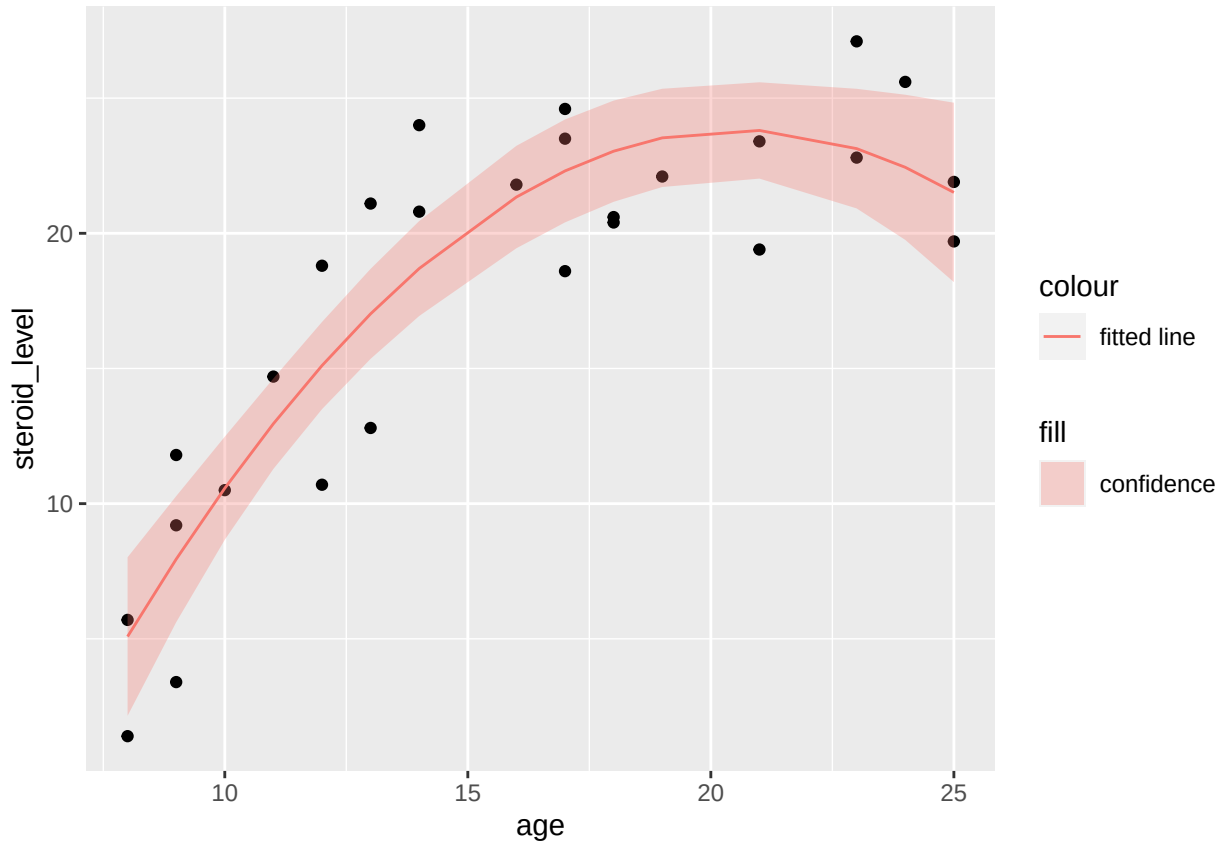
```
##
## Call:
## lm(formula = steroid_level ~ center_age + center_age_sq, data = steroid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5463 -2.5369  0.3868  2.1973  5.3020
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   21.09416    0.91415  23.075  < 2e-16 ***
## center_age     1.13736    0.11546   9.851 6.59e-10 ***
## center_age_sq -0.11840    0.02347  -5.045 3.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.153 on 24 degrees of freedom
## Multiple R-squared:  0.8143, Adjusted R-squared:  0.7989
## F-statistic: 52.63 on 2 and 24 DF,  p-value: 1.678e-09
```

$$Y = 21.0942 + 1.13736x - .118401x^2 + \epsilon_i$$

```
# model fitted values w/ CI
steroid <- bind_cols(
  steroid, as.data.frame(predict.lm(lmod, interval = "confidence"))
) %>% rename(conf_low = lwr, conf_high=upr)
```

```
# plot of actual values and CI of fitted
ggplot(steroid, aes(x=age, y=steroid_level)) +
  geom_point() +
  geom_line(aes(y = fit, color = "fitted line")) +
  geom_ribbon(aes(ymin= conf_low, ymax = conf_high, fill = "confidence"), alpha = 0.3)
```



b. Test whether or not there is a regression relation; use a = .01. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

Alternatives:

$H_0 : B_1 = B_{11} = 0$ $H_a : B_1 = B_{11} \neq 0$

Note: it's β_{11} because it is the square value of β_1

General Decision Rule

$$F^* \leq F(1-\alpha, df_R - df_F, df_F), \text{ fail to reject } H_0$$
$$F^* > F(1-\alpha, df_R - df_F, df_F), \text{ reject } H_0$$

```
# gives F-statistic, df for partial and full
summary(lmod)
```

```
##
## Call:
## lm(formula = steroid_level ~ center_age + center_age_sq, data = steroid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5463 -2.5369  0.3868  2.1973  5.3020
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   21.09416    0.91415  23.075  < 2e-16 ***
## center_age     1.13736    0.11546   9.851 6.59e-10 ***
## center_age_sq -0.11840    0.02347  -5.045 3.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.153 on 24 degrees of freedom
## Multiple R-squared:  0.8143, Adjusted R-squared:  0.7989
## F-statistic: 52.63 on 2 and 24 DF,  p-value: 1.678e-09
```

Note: summary F-stat 24 DF $= 27 - 3 = (n - p)$

```
# get number of obs
nrow(steroid)
```

```
## [1] 27
```

```
# look at the f distribution
# 1-alpha, p-1, n-p
qf(1-0.01, 2, 24)
```

```
## [1] 5.613591
```

$$52.63 > 5.613, \text{ reject } H_0$$

Therefore, we should keep both age and age squared in the model.

d. Predict the steroid levels of females aged 15 using a 99 percent prediction interval. Interpret your interval.

```
# data frame of one value 15
steroid15 <- data.frame(center_age = 15, center_age_sq = (15)^2)
```

```
21.0942 + 1.13736*(15) - .118401*(15)^2
```

```
## [1] 11.51438
```

```
predict.lm(lmod, newdata=steroid15, interval = "prediction", level = 0.99)
```

```
##        fit       lwr      upr
## 1 11.51424 -4.007162 27.03564
```

e. Test whether the quadratic term can be dropped from the model; use a = .01. State the alternatives, decision rule, and conclusion.

Alternatives:

$H_0 : B_{11} = 0$ $H_a : B_{11} \neq 0$

General Decision Rule

$$F^* \leq F(1-\alpha, df_R - df_F, df_F), \text{ fail to reject } H_0$$
$$F^* > F(1-\alpha, df_R - df_F, df_F), \text{ reject } H_0$$

```
# gives F-value, df for partial and full
anova(lmod)
```

```
## Analysis of Variance Table
##
## Response: steroid_level
##               Df Sum Sq Mean Sq F value    Pr(>F)
## center_age     1 793.28  793.28  79.813 4.236e-09 ***
## center_age_sq  1 252.99  252.99  25.453 3.708e-05 ***
## Residuals     24 238.54    9.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# look at the f distribution
# 1-alpha, df of partial determination, df of whole model
qf(1-0.01, 1, 24)
```

```
## [1] 7.822871
```

$$25.453 > 7.822871, \text{ reject } H_0$$

Therefore, we should keep quadratic term in the model.