

BIOSTATS 640
Spring 2021
 Illustration
 R for Multiple Linear Regression

Introduction to Example

Source:

Matthews et al. Parity Induced Protection Against Breast Cancer 2007.

Research Question:

What is the relationship of **Y=p53 expression** to **parity** and **age at first pregnancy**, after adjustment for the potentially confounding effects of **current age** and **menopausal status**. Age at first pregnancy has been grouped and is either ≤ 24 years or > 24 years.

Dataset Used:

p53paper.Rdata

Packages Used:

GGally ggfortify
 ggplot2 Hmisc
 stargazer car
 psych

Note: In the pages that follow, package names are enclosed in curly brackets; e.g., {psych}

Data Dictionary:

Variable	Variable Label	Type	Units/Codes
id	Study id	character	e.g., "J-05-15"
p53	p53 Expression	double	
agepreg1	Age at 1 st pregnancy	double	years
agecurr	Current age	double	years
menop	Menopausal status	integer	0=other; 1=Post-Menopausal

Launch R-Studio, Set Working directory. Input R dataset p53data.Rdata

```
setwd("/Users/cbigelow/Desktop/")
load(file="p53data.Rdata")
```

Quick look at all the variables

```
library(stargazer)
stargazer::stargazer(p53data,type="text",median=TRUE)

##
## =====
## Statistic N      Mean  St. Dev.  Min   Pctl(25) Median Pctl(75)  Max
## -----
## p53           67  3.251    1.054    1.000  2.500   3.000   4.000   6.000
## agepreg1      51 23.441    6.163   15.000 19.000  23.000  27.000  40.500
## agecurr       68 39.279   13.894     15     27    39.5    49.2    75
## menop         68 0.279    0.452     0      0      0      1      1
## -----
```

Clean data

```
library(Hmisc)
library(tidyverse)

# Create factor version of menop
p53data$menopf <- factor(p53data$menop,
                        levels = c(0,1),
                        labels = c("0 = other", "1=post-menopausal"))

# Create 0/1 indicator of age at first pregnancy <= 24.
p53data$agefirst_le24 <- NA
p53data$agefirst_le24[p53data$agepreg1 <= 24] <- 1
p53data$agefirst_le24[p53data$agepreg1 > 24] <- 0
p53data$agefirst_le24f <- factor(p53data$agefirst_le24, # (optional)create factor var
                                levels = c(0,1),
                                labels = c("0 = age first > 24", "1= age first <= 24"))

# label variables
Hmisc::label(p53data$p53) <- "p53: p53 expression"
Hmisc::label(p53data$agepreg1) <- "agepreg1: Age at 1st pregnancy (years)"
Hmisc::label(p53data$agecurr) <- "agecurr: Current age (years)"
Hmisc::label(p53data$menop) <- "menop: 0/1 post-menopausal"
Hmisc::label(p53data$menopf) <- "menopf: 0/1 post-menopausal"
Hmisc::label(p53data$agefirst_le24) <- "agefirst_le24: 1st pregnancy at age <= 24"
Hmisc::label(p53data$agefirst_le24f) <- "agefirst_le24f: 1st pregnancy at age <= 24"

# Retain complete data only
p53_complete <- na.omit(p53data)
print("p53_complete: Cleaned and Complete")
```

```
## [1] "p53_complete: Cleaned and Complete"

glimpse(p53_complete)

## Observations: 51
## Variables: 8
## $ id <chr> "J-05-15", "J-05-17", "J-05-18", "J-05-31", "J-05...
## $ p53 <labelled> 3.00, 2.40, 2.20, 5.50, 6.00, 4.00, 4.00, 5...
## $ agepreg1 <labelled> 26.0, 30.0, 29.0, 27.0, 23.0, 19.0, 31.0, 22...
## $ agecurr <labelled> 43, 53, 39, 30, 35, 54, 40, 57, 25, 60, 26, ...
## $ menop <labelled> 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0,...
## $ menopf <fct> 1=post-menopausal, 1=post-menopausal, 0 = other, ...
## $ agefirst_le24 <labelled> 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0,...
## $ agefirst_le24f <fct> 0 = age first > 24, 0 = age first > 24, 0 = age f...
```

Look at the data.

```
library(stargazer)
library(ggplot2)
library(GGally)
library(psych)

# Quick descriptives on every variable in a nice table
stargazer::stargazer(p53_complete, type="text", median=TRUE, title="Data: p53")

##
## Data: p53
## =====
## Statistic      N   Mean  St. Dev.  Min  Pctl(25)  Median  Pctl(75)  Max
## -----
## p53             51  3.465   1.042     1    2.9     3.5     4         6
## agepreg1       51 23.441   6.163    15    19     23     27        40
## agecurr        51 44.216  11.436    25    36     43     53        75
## menop          51  0.373   0.488     0     0     0     1         1
## agefirst_le24  51  0.627   0.488     0     0     1     1         1
## -----
```

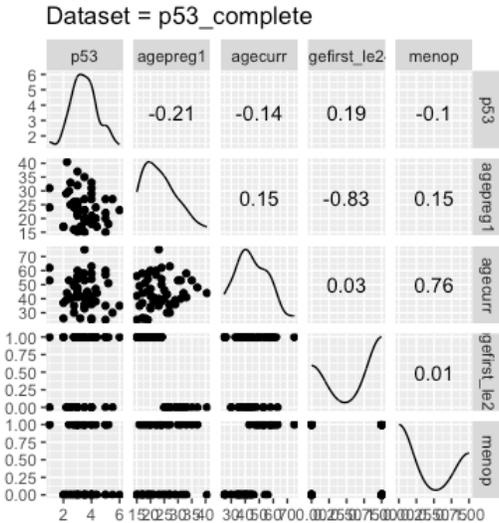
Note – stargazer() has options for choosing which statistics you want to report.

```
# Pairwise correlation of numeric variables only
keepvars <- c("p53", "agepreg1", "agecurr", "agefirst_le24", "menop")
cor(p53_complete[keepvars])

##
##           p53   agepreg1   agecurr  agefirst_le24   menop
## p53      1.0000000 -0.2079093 -0.1355181  0.186324331 -0.103753379
## agepreg1 -0.2079093  1.0000000  0.1507204 -0.831481121  0.150307706
## agecurr  -0.1355181  0.1507204  1.0000000  0.029004800  0.758970642
## agefirst_le24 0.1863243 -0.8314811  0.0290048  1.000000000  0.006578947
## menop    -0.1037534  0.1503077  0.7589706  0.006578947  1.000000000
```

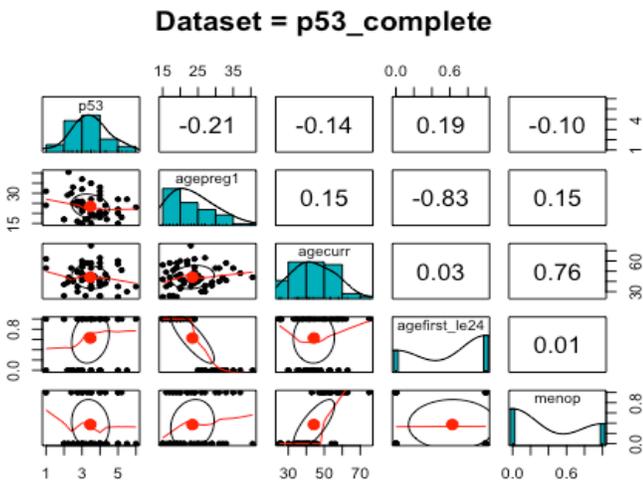
Note – I’m not a huge fan of matrices of correlations like this. But now you know how to produce one. Perhaps this is useful as a “quick look”?

```
# Pairwise scatterplot w correlations using {GGally} function ggscatmat()
GGally::ggscatmat(p53_complete[keepvars]) +
  ggtitle("Dataset = p53_complete") +
  labs(x=" ", y=" ")
```



Note – This approach is nice.

```
# Pairwise scatterplot w correlations using {psych} function pairs.panel()
psych::pairs.panels(p53_complete[keepvars],
  method = "pearson", # correlation method
  hist.col = "#00AFBB",
  density = TRUE, # show density plots
  ellipses = TRUE, # show correlation ellipses
  main="Dataset = p53_complete")
```



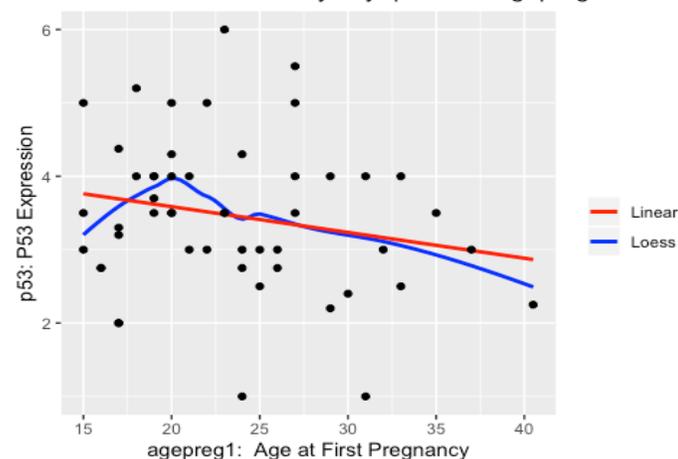
Note – This approach is also nice! Plus, you get to see the ellipses of covariability. What we see here is: (1) Distribution of y=53 is probably reasonably normal; (2) if there is a linear relationship of y=53 with x=agepreg1, it's pretty weak; and (3) ditto with x=agecurr. If you square the correlations, you get values of R^2 . For example, R^2 for y=p53 and x=agepreg1 = $(-0.21)^2 = .04$, which says that we estimate that 4% of the variability in y=p53 is explained by a linear model in agepreg1. Not much!

Simple Linear Regression of Y=p53 on X=agepreg1: Estimation

```
library(ggplot2)
library(stargazer)

# Assess linearity via overlay plot of observed, linear, and loess
ggplot(data=p53_complete) +
  aes(x=agepreg1, y=p53) +
  geom_smooth(method="loess", aes(color="Loess"), se=FALSE) +
  geom_smooth(method="lm", aes(color="Linear"), se=FALSE) +
  geom_point() +
  scale_colour_manual(name="", values=c("red", "blue")) +
  ggtitle("Assessment of Linearity of y=p53 in x=agepreg1") +
  xlab("agepreg1: Age at First Pregnancy") +
  ylab("p53: P53 Expression")
```

Assessment of Linearity of y=p53 in x=agepreg1



Note – In loess smoothing (approximately), at each value of the predictor, a linear model is fit to a local percent of the data set (e.g., 80%). Then (magically!), all these local lines are smoothed out. The result (quite nice, actually) is a rough idea of the nature of the actual curviness of the relationship. Then it’s all smoothed. Here, we see a bit of departure from linearity.

```
# Fit simple linear regression
m1 <- lm(p53 ~ agepreg1, data=p53_complete)

# Show fit using summary()
print("Simple Linear Regression: p53 ~ agepreg1")

## [1] "Simple Linear Regression: p53 ~ agepreg1"

summary(m1) # summary( ) to show results of model fit

##
## Call:
## lm(formula = p53 ~ agepreg1, data = p53_complete)
##
## Residuals:
## p53: p53 expression
##      Min       1Q   Median       3Q      Max
## -2.44556 -0.62052 -0.08612  0.67166  2.51930
##
```

```
## Coefficients:
##           Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  4.28892    0.57207    7.497 0.00000000113 ***
## agepreg1   -0.03514    0.02362   -1.488    0.143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.029 on 49 degrees of freedom
## Multiple R-squared:  0.04323,    Adjusted R-squared:  0.0237
## F-statistic: 2.214 on 1 and 49 DF,  p-value: 0.1432

# Alternative approach is to show fit uses {stargazer} command stargazer()
stargazer::stargazer(m1,type="text",font.size="small", align=TRUE,
                    ci=TRUE, intercept.bottom=FALSE,
                    covariate.labels=c("Intercept", "Age at First Pregnancy"),
                    dep.var.labels=c("p53: p53 Expression"),
                    title="Single Predictor agepreg1: Beta (95% CI)")

##
## Single Predictor agepreg1: Beta (95% CI)
## =====
##                               Dependent variable:
##                               -----
##                               p53: p53 Expression
## -----
## Intercept                      4.289***
##                               (3.168, 5.410)
##
## Age at First Pregnancy          -0.035
##                               (-0.081, 0.011)
##
## -----
## Observations                    51
## R2                              0.043
## Adjusted R2                     0.024
## Residual Std. Error             1.029 (df = 49)
## F Statistic                     2.214 (df = 1; 49)
## =====
## Note:                          *p<0.1; **p<0.05; ***p<0.01

Notes:
(1)  $\hat{p}_{53} = 4.289 - 0.035 \cdot \text{agepreg1}$  says "Each 1 year increase in age is associated with an estimated 0.035 unit reduction in p53 expression"
(2)  $R^2 = .042$  says 4% of the variability in outcome (p53) is explained by the fitted model

# Analysis of variance
anova(m1)

## Analysis of Variance Table
##
## Response: p53
##           Df Sum Sq Mean Sq F value Pr(>F)
## agepreg1  1  2.345  2.3453  2.2138 0.1432
## Residuals 49 51.911  1.0594
```

Notes:

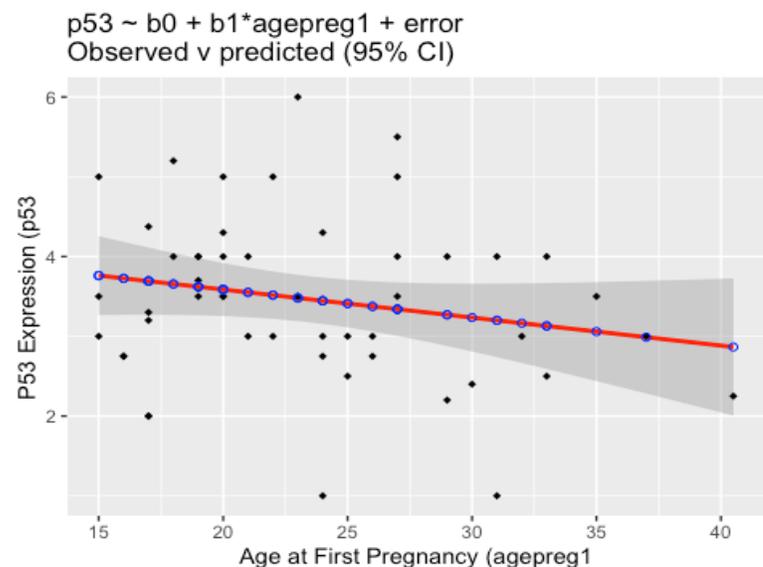
(1) Just confirming for you: $R^2 = \frac{\text{Regression SSQ}}{\text{Total SSQ}} = \frac{2.345}{(2.345 + 51.911)} \cdot 100\% = 4.32\%$ which matches above.

Simple Linear Regression of Y=p53 on X=agepreg1: Graph

```
library(ggplot2)

# Preliminary: save predicted values to dataset
p53_complete$yhat <- predict(m1)

# overlay plot with: observed, fitted, line
ggplot(data=p53_complete) +
  aes(x=agepreg1, y=p53) +
  geom_smooth(method="lm", color="red") +
  geom_point(aes(x=agepreg1, y=p53), color="black", shape=18) +
  geom_point(aes(x=agepreg1, y=yhat), color="blue", shape=21) +
  ggtitle("p53 ~ b0 + b1*agepreg1 + error\nObserved v predicted (95% CI)") +
  xlab("Age at First Pregnancy (agepreg1)") +
  ylab("P53 Expression (p53)")
```



Simple Linear Regression of Y=p53 on X=agepreg1: Confidence and Prediction Intervals

```
# New values of X=agepreg1
mydata <- data.frame(agepreg1 = c(15, 20, 25, 30, 35, 40))

# Confidence Interval Estimates of Means
estimated_mean <- predict(m1, newdata = mydata, interval = "confidence")
out_mean <- cbind(mydata, estimated_mean) # cbind( ) appends agepreg1 for readability

print("Estimated Means and 95% CI")
## [1] "Estimated Means and 95% CI"

out_mean

##   agepreg1      fit      lwr      upr
## 1      15 3.761818 3.267459 4.256176
## 2      20 3.586118 3.253609 3.918628
## 3      25 3.410419 3.111485 3.709354
## 4      30 3.234720 2.809528 3.659913
## 5      35 3.059021 2.438664 3.679378
## 6      40 2.883322 2.045752 3.720892
```

```
# Predictions w 95% CI
estimated_individual <- predict(m1, newdata=mydata, interval = "prediction")
out_individual <- cbind(mydata,estimated_individual)
```

```
print("Predictions and 95% CI")
## [1] "Predictions and 95% CI"
```

out_individual

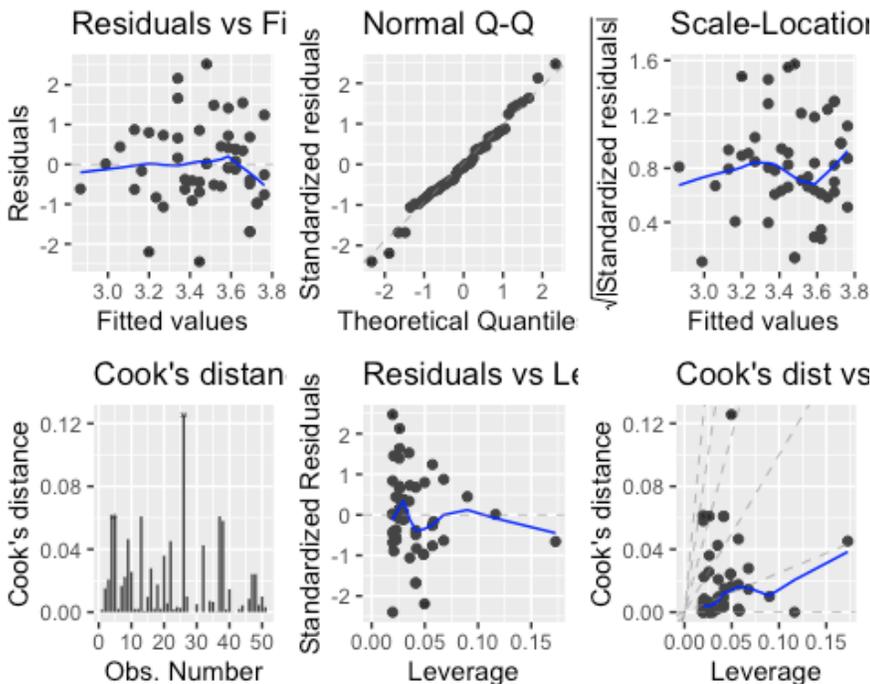
```
##   agepreg1    fit      lwr      upr
## 1      15 3.761818 1.6351535 5.888482
## 2      20 3.586118 1.4911549 5.681082
## 3      25 3.410419 1.3205217 5.500317
## 4      30 3.234720 1.1230623 5.346378
## 5      35 3.059021 0.8995875 5.218454
## 6      40 2.883322 0.6517676 5.114876
```

Hack:
 When I first did the coding to produce the predictions together with the CI, the resulting table of output had only the columns "fit", "lwr", and "upr". So, you didn't know what you were looking at! So, this is why I added another command that "combined" the values of x ("mydata") with the results of obtaining the predictions.

Simple Linear Regression: Diagnostic Plots

```
library(ggplot2)
library(ggfortify)

# model checks using {ggfortify} function autoplot()
autoplot(m1, which = 1:6, ncol = 3, label.size = 1)
```



Notes:
 See course notes

Simple Linear Regression: Other Model Checks

```
library(lmtest)
library(car)

# save residuals to dataset
p53_complete$residuals <- residuals(m1)

# Shapiro-Wilk Test of Normality (Null: residuals ~ Normal)
shapiro.test(p53_complete$residuals)

##
## Shapiro-Wilk normality test
##
## data: p53_complete$residuals
## W = 0.98616, p-value = 0.8117
Null is NOT rejected (p-value = .81). No statistically significant departure from normality.

# Test of Omitted Variables (Null: no important predictors omitted)
lmtest::resettest(m1, power=2, type="regressor")

##
## RESET test
##
## data: m1
## RESET = 1.0449, df1 = 1, df2 = 48, p-value = 0.3118
Null is NOT rejected (p-value = .31). No statistically significant evidence that we forgot to
include a predictor (not surprising, given that this dataset is limited).

# Cook-Weisberg Test of Variance of Residuals (Null: Variance is constant)
car::ncvTest(m1)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.003144433, Df = 1, p = 0.95528
Null is NOT rejected (p-value = .95). No statistically significant departure from constant variance.
```

Multiple Predictor Models

```
library(stargazer)

# Set I: Which is the better predictor: agepreg1 or agefirst_le24
m1 <- lm(p53 ~ agepreg1, data=p53_complete)
m2 <- lm(p53 ~ agefirst_le24, data=p53_complete)
stargazer::stargazer(m1,m2, type="text",font.size="small", align=TRUE,
                    ci=TRUE, intercept.bottom=FALSE,
                    dep.var.labels=c("p53: p53 Expression"),
                    title="Single Predictor Models: Beta (95% CI)")
```

Note:
 I confess. I played around with the options of stargazer()so as to obtain the selection of statistics that I wanted to report plus the layout that I wanted.

```
##
## Single Predictor Models: Beta (95% CI)
## =====
##                               Dependent variable:
##                               -----
##                               p53: p53 Expression
##                               (1)           (2)
## -----
## Constant                      4.289***      3.216***
##                               (3.168, 5.410) (2.751, 3.681)
##
## agepreg1                       -0.035
##                               (-0.081, 0.011)
##
## agefirst_le24                                0.397
##                                               (-0.189, 0.984)
## -----
## Observations                    51           51
## R2                               0.043       0.035
## Adjusted R2                     0.024       0.015
## Residual Std. Error (df = 49)   1.029     1.034
## F Statistic (df = 1; 49)        2.214     1.762
## =====
## Note:                            *p<0.1; **p<0.05; ***p<0.01
```

Note: Whether you model “age at first pregnancy” using the continuous predictor agepreg1 or using the 0/1 indicator variable agefirst_le24, it doesn’t seem to make much difference. We don’t have a lot to go on here. But I wanted to show you the idea of first doing the modeling needed to decide the “best” way to model the primary predictor of interest.

```
# Set II: Comparison of several models
m1 <- lm(p53 ~ agepreg1 + agecurr + menop, data=p53_complete)
m2 <- lm(p53 ~ agepreg1 + agecurr, data=p53_complete)
m3 <- lm(p53 ~ agepreg1 + menop, data=p53_complete)
m4 <- lm(p53 ~ agepreg1, data=p53_complete)
stargazer::stargazer(m1,m2,m3, m4, type="text",font.size="small", align=TRUE,
  report = "vcp*", intercept.bottom=FALSE,
  dep.var.labels=c("p53: p53 Expression"),
  title="Three-, Two- and Single Predictor Models: Beta (95% CI)")
```

Note: (1) My goal here is to show you an efficient way of considering several models. In a real life situation, your analysis plan might have you proceed differently. But I wanted to show you the coding. (2) Here, I begin with fitting a “full” model that contains all the predictors that might be of interest. From there, I fit some smaller models to see if a more parsimonious model performs adequately: sensible and with no appreciable loss of variance explained.

```
##
## Three-, Two- and Single Predictor Models: Beta (95% CI)
## =====
##                               Dependent variable:
## -----
##                               p53: p53 Expression
##                               (1)           (2)           (3)           (4)
## -----
## Constant                4.688                4.655                4.304                4.289
##                          p = 0.00001***        p = 0.00000***        p = 0.000***        p = 0.000***
## agepreg1                 -0.033                -0.032                -0.033                -0.035
##                          p = 0.188                p = 0.183                p = 0.174                p = 0.144
## agecurr                  -0.011                -0.010
##                          p = 0.594                p = 0.457
## menop                    0.030
##                          p = 0.949
##                               -0.158
##                               p = 0.605
## -----
## Observations              51                51                51                51
## R2                        0.054                0.054                0.049                0.043
## Adjusted R2              -0.006                0.015                0.009                0.024
## Residual Std. Error    1.045 (df = 47)    1.034 (df = 48)    1.037 (df = 48)    1.029 (df = 49)
## F Statistic             0.902 (df = 3; 47)  1.379 (df = 2; 48)  1.226 (df = 2; 48)  2.214 (df = 1; 49)
## =====
## Note:                                *p<0.1; **p<0.05; ***p<0.01
Note: Okay. Not a lot to go on here. In a real life situation, we might actually decide that model
4 is the way to go. But to illustrate the thinking: We might decide that models 3 and 4, compared
to models 1 and 2 have associated with them a bit of a drop in R2. So we might decide that our
choice is between models 1 and 2. But which? At face value, they look awfully similar! And the slope
on our primary predictor (-.033 versus -0.032) is unchanged. So we would go with model 2 (no need
to have menop in the model). Nevertheless, so that I can show you the code, the following is
how to do a partial F test.

# Partial F-test
anova(m2,m1)

## Analysis of Variance Table
##
## Model 1: p53 ~ agepreg1 + agecurr
## Model 2: p53 ~ agepreg1 + agecurr + menop
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1      48 51.308
## 2      47 51.304  1 0.0046075 0.0042 0.9485
Null is NOT rejected (p-value = .95). Controlling for agepreg1 and agecurr, the extra predictor
menop is NOT statistically significantly associated with y=p53.
```