# COMPLETE EXAMPLE HIERARCHICAL

# MULTIPLE LINEAR REGRESSION

**Research Question:** After **controlling** for demographic variables, does the extroversion of the participant predict how well they take care of their car?

**Data set:** data 2.csv

**IV(s):**

* Sex – gender of the participant (0 = female, 1 = male)
* Age – age of the participant
* Extro – extroversion of the participant, low numbers are introverted, high numbers are extroverted.

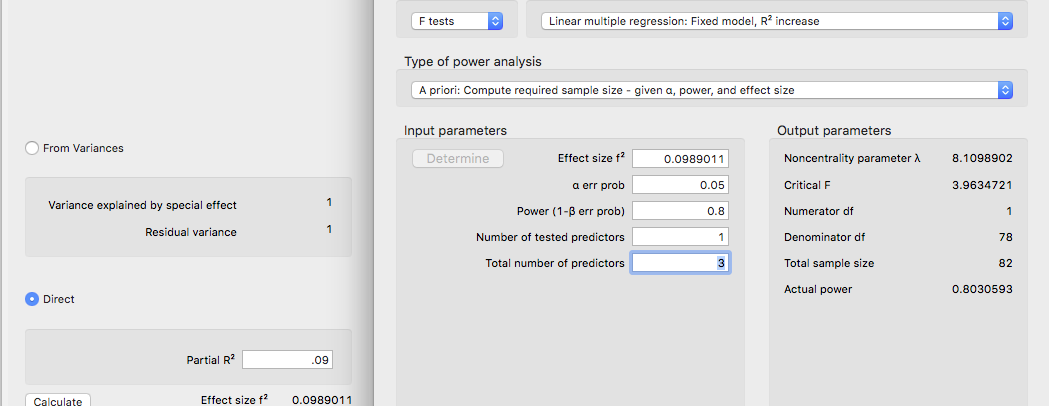
**DV:**

* Car – how well a person takes care of their car (regular washes, cleaned, oil changed, etc.).

**Power:**

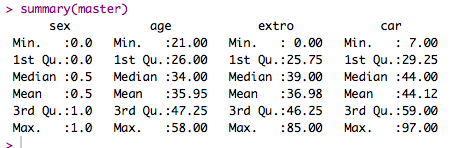
1. Open Gpower!
   1. Test family: F-test
   2. Statistical Test: Linear multiple regression: fixed model, R2 increase
      1. We are using multiple regression because we have more than one predictor.
      2. R2 increase indicates that we are asking if the addition of more predictors to previous model are useful.
   3. Estimate an effect size: click determine 🡪 use R square sizes you think might be accurate, remember small, medium, and large estimates from the notes.
      1. You will be entering partial R2 🡪 what do you think the effect size is for the last step of the model?
   4. Alpha = .05
   5. Power (1-beta .20) = .80
   6. Number of tested predictors: number of IVs/X variables in that step.
   7. Total number of predictors: number of IVs/Xs overall.
2. Let’s estimate the following:
   1. Medium effect size (*R2* = .09)
   2. Number of tested predictors: 1 (extroversion)
   3. Total number of predictors: 3 (extroversion + gender + age)

Says we needed to run 82 people to find a significant effect with a medium effect size.

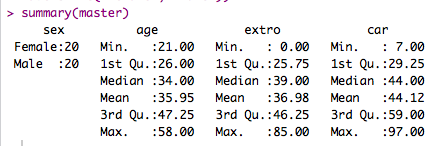


**Assumptions:**

1. Accuracy:
   1. Use the summary(*dataset name*) function to get the basic information for the data.
   2. Let’s check out minimum and maximum:
      1. Gender should be factored.
      2. Age, extro, and car should not be negative.



* 1. Fixed the gender by factoring.

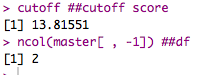


1. Missing:
   1. I can see from my summary function that I do not have missing data. Remember that you will need at least twenty variables to estimate missing data for participants – so mostly you won’t be estimating for regression.
2. Run the lm model for your data.
   1. Seems like a strange place to stop and run the analysis – but we need the regression to calculate values for outliers below. So you will run the FINAL model of your analysis for data screening.
   2. Therefore, if you are running a mediation/moderation/hierarchical be sure to run the model with ALL the variables here.
   3. output = lm(*DV* ~ *IV + IV + IV…,* data = *dataset*)
3. Outliers:
   1. First: Mahalanobis scores:
      1. mahal = mahalanobis(*dataset*,

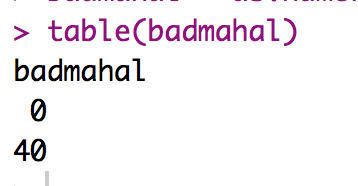
colMeans(*dataset*, na.rm = T),

cov(*dataset*, use = “pairwise.complete.obs”))

* + 1. Create the cut off score:
       1. cutoff = qchisq(1-.001, ncol(*dataset*))
    2. Remember you can use:
       1. cutoff to get the cutoff score
       2. ncol(*dataset*) to get the *df*



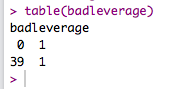
* 1. SAVE the scores:
     1. mahalout = as.numeric(mahal > cutoff) – notice that we have used > …
     2. We are checking if people are greater than the cutoff (that’s bad), and if so, giving them a 1 to mark they are an outlier. The as.numeric changes the TRUE for outlier to 1, while FALSE no outlier is a 0.
     3. This procedure is slightly different than before, because we are not simply going to keep people who are less than the cut off score – we need to keep a total of their bad scores.
     4. Check out the number of outliers (1 is bad!):
        1. table(badmahal)



* + - 1. No outliers!
  1. Leverage scores:
     1. Remember that leverage is the influence of a single person over the slope.
     2. k = number of predictors.
     3. To get leverage values:
        1. leverage = hatvalues(output)
     4. To get the cut off score:
        1. (2\*k+2)/N
        2. cutleverage = (2\*k+2) / nrow(*dataset*)
     5. Run cutleverage to see the cut off score:



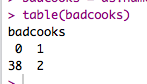
* + 1. Save the scores and see how many outliers:
       1. badleverage = as.numeric(leverage > cutleverage)
       2. table(badleverage)



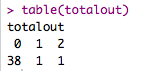
* + 1. We have one outlier.
  1. Cook’s scores:
     1. Remember that Cook’s is a measure of influence and discrepancy.
     2. To get Cook’s values:
        1. cooks = cooks.distance(output)
     3. Get the cutoff score:
        1. 4 / (N-k-1)
        2. cutcooks = 4 / (nrow(*dataset*) - k - 1)
        3. Run cutcooks to see the cut off score.



* + 1. Save the scores and see how many outliers:
       1. badcooks = as.numeric(cooks > cutcooks)
       2. table(badcooks)

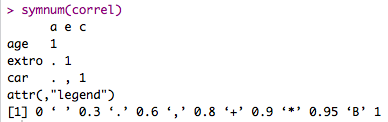


* + 1. We have two outliers!
  1. So, what does that mean overall?
     1. We want to create a total score for each participant of outliers.
     2. So, we add them up for total outlier-ness.
        1. totalout = badmahal + badleverage + badcooks
        2. table(totalout)
        3. Remember that top row = their score: 0, 1, 2, 3
        4. Bottom row is the number of people who have that score.

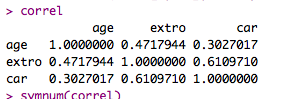


* + 1. Now, any people we have two or more problems need to get excluded:
       1. noout = subset(master, totalout < 2)
       2. We have one overall outlier.

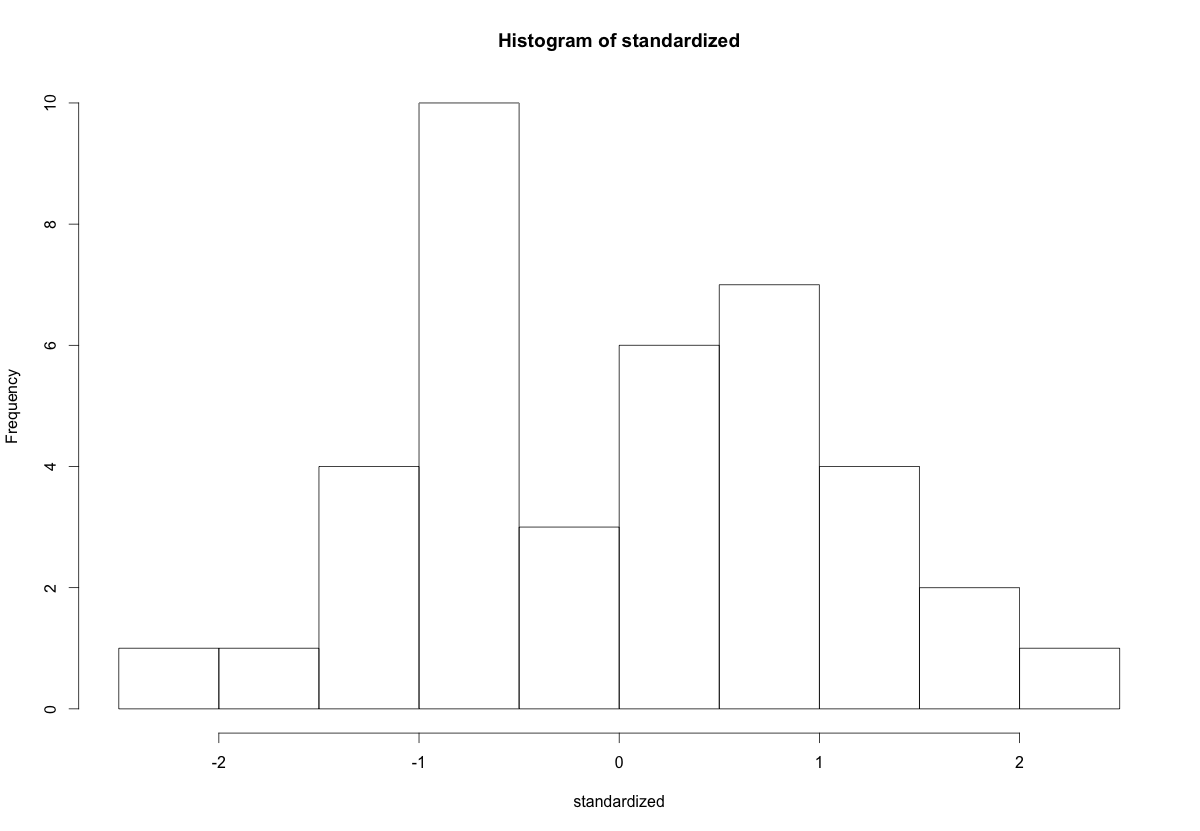
1. Additivity
   1. We do not want correlated IVs! It is a waste of time!
      1. The rule is technically *r* = .90.
      2. However, I would argue you shouldn’t really go over *r* = .70 – at that point the IVs are so correlated, you are risking suppression of one or both variables.
   2. Get the correlations:
      1. correl = cor(*dataset*, use = “pairwise.complete.obs”)
   3. Get the symbols chart:
      1. symnum(correl)
   4. Look for things with a , or higher.
   5. NOTE: I included the DV in this analysis (just to run it easily). You WANT strong correlations with the DV. So don’t exclude an IV if it’s correlated with the DV (doh!) or you’ve excluded your hypothesis!



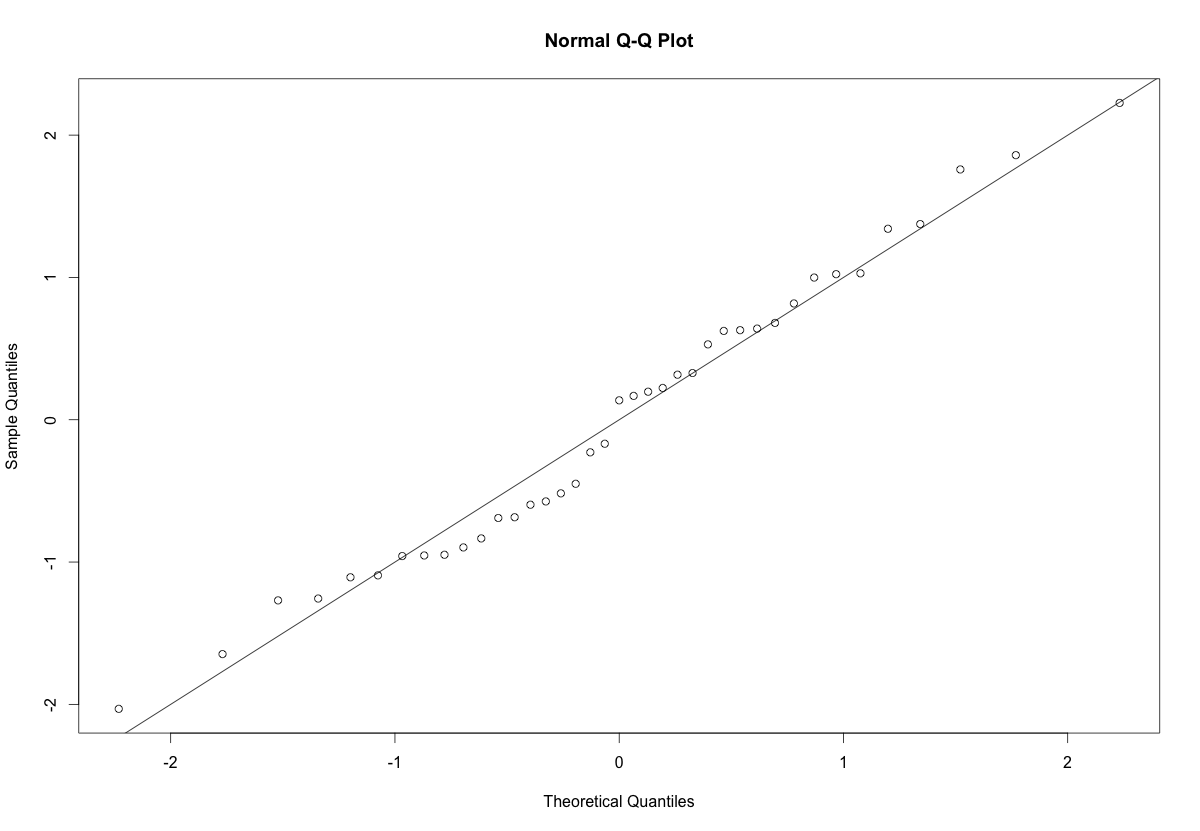
* 1. Also helps to look at the real correlations, since , = > .6 which is a lower cut off than we want, but it’s a good place to start looking for problems.



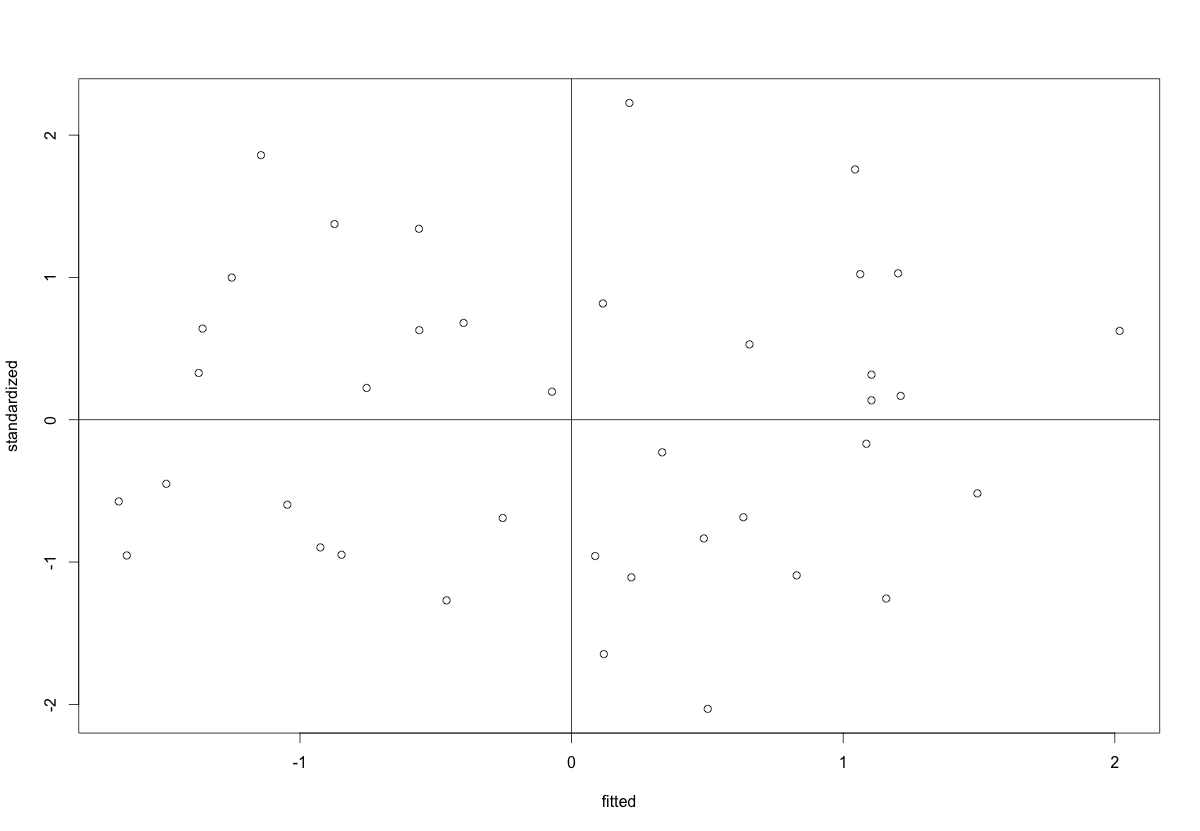
1. Set up the rest of the assumptions:
   1. Run the real analysis again with the no outliers dataset.
   2. No fake or randomness! It’s real regression!
   3. Create the standardized residuals:
      1. standardized = rstudent(output)
   4. Create the fitted values:
      1. fitted = scale(output$fitted.values)
2. Normality:
   1. hist(standardized)
   2. Most of the data is between -2 and 2 but not really centered over zero. Thank goodness we have at least 30 people.



1. Linearity:
   1. qqnorm(standardized)
   2. abline(0,1)
   3. Oh! Pretty!

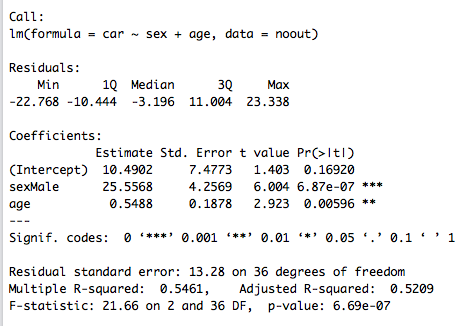


1. Homogeneity:
   1. plot(fitted,standardized)
   2. abline(0,0)
   3. abline(v = 0)
   4. Spread looks ok both vertically and horizontally.
2. Homoscedasticity:
   1. The spread around the line looks fairly blobby / uniform, so I would say this graph is ok.



**Running the real analysis:**

1. Now, we are going to run things in steps.
   1. Be sure each subsequent step **includes** the previous variables. You are adding onto the equation with each step.
2. Get the regression and model:
   1. model1 = lm(*DV* ~ *IV* + *IV*, data = *noout*)
   2. summary(model1)

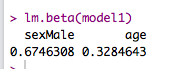


**Interpret the output:**

1. With steps, we are going to interpret each predictor **in the step it was entered.**
2. Is the overall model significant?
   1. Yes, *F*(2, 36) = 21.66, *p* < .001, *R2 =* .55
3. Are our individual predictors significant?
   1. Gender, yes: *b* = 25.56, *t*(36) = 6.00, *p* < .001
   2. Age, yes: *b* = 0.55, *t*(36) = 2.92, *p* = .006
4. Interpretation:
   1. Males have higher car scores than females.
   2. As age increases, we have higher car scores.
   3. We should figure out beta and *pr2* to determine which of those is better though.

**Beta:**

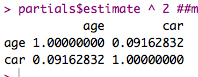
1. We can usually use the *QuantPsy* package and the lm.beta function.
   1. library(QuantPsy)
   2. lm.beta(model)



* 1. Gender is much better predictor.
  2. Remember these are standardized, NOT correlations. They can be negative and be larger than 1. Strength is about distance from zero, not sign (i.e. + or -).

**Effect Size:**

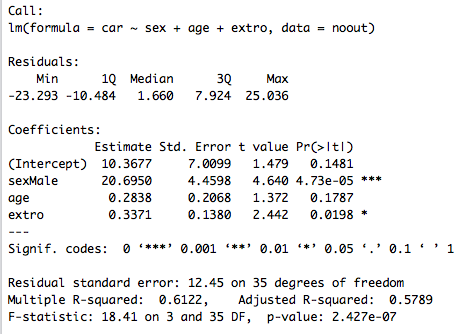
1. Another way to determine which predictor was stronger is to examine their effect sizes. We are going to use *pr2* for predictor effect size in the *ppcor* library with the *pcor* function.
   1. BUT: use only the variables in that step.
   2. BUT: no categorical variables.
      1. So, partials aren’t very helpful right now.
   3. library(ppcor)
   4. partials = pcor(*dataset*, method = "pearson")
   5. partials$estimate ^ 2
      1. The second line calculates and saves the partial correlations – you want to control / exclude the variance for all other variables to see the correlation of the IV and DV.
      2. The third line squares and prints out the correlation table. We square the values because it’s effect size (remember *R* is squared for effect size).
      3. Special NOTE: these effect sizes are partials – they have different denominators and will not add up to *R2*.

****

* 1. Only look at the line with the DV.
  2. Gender, yes: *b* = 25.56, t(36) = 6.00, *p* < .001
  3. Age, yes: *b* = 0.55, t(36) = 2.92, *p* = .001, *pr2 =* .09

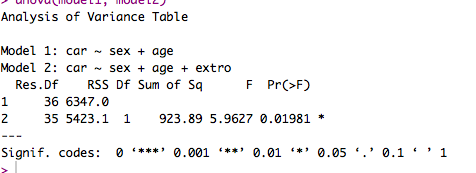
**Running the real analysis part 2:**

1. Be sure each subsequent step **includes** the previous variables. You are adding onto the equation with each step.
   1. So, now we add the rest of the variables.
2. Get the regression and model:
   1. model2 = lm(*DV* ~ *IV* + *IV + IV + IV…*, data = *noout*)
   2. summary(model2)

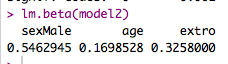


**Interpret the output:**

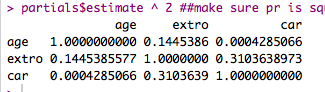
1. Is the overall model significant?
   1. Uh, who cares? (I mean, yes it is). If I find that the overall model is significant, that just tells me if R2 is greater than zero. That includes all the variables. I want to know if the **addition or change between model 1 and 2** is significant.
   2. So, does adding extroversion help us predict car?
   3. We can find that out by comparing model 1 and model 2 with the anova() command.
      1. anova(model1, model2)



1. Is the addition of extroversion significant?
   1. Yes, Δ*F*(1,35) = 5.96, *p* = .02, *ΔR2 =* .06
   2. How did I get change in R2? I subtracted model 1 and model 2.
2. Is extroversion a significant predictor?
   1. Yes, *b* = 0.34, *t*(35) = 2.44, *p* = .02, *pr2 =* .31 (see below)
3. Is it stronger than gender and age?
   1. Better than age (in this step), gender still stronger.



1. Partials indicate the same thing:



1. You would repeat this process until you were done with your different hypothesized steps.

**Make a picture:**

1. Pictures for regression analyses are not the most common thing, because they show the overall predicted values to the actual Y values (and not the IVs individually).
2. However, we will use them for some types of analyses, so let’s make some!
3. Load ggplot2.
   1. library(ggplot2)
4. Run some cleanup coding.
   1. cleanup = theme(panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(),

panel.background = element\_blank(),

axis.line = element\_line(colour = "black"),

legend.key = element\_rect(fill = "white"),

text = element\_text(size = 15))

1. Get the finalized fitted values:
   1. fitted = model$fitted.values
2. Make the plot:
   1. scatter = ggplot(*dataset*, aes(*fitted, y values*))
3. Build that graph:
   1. scatter +

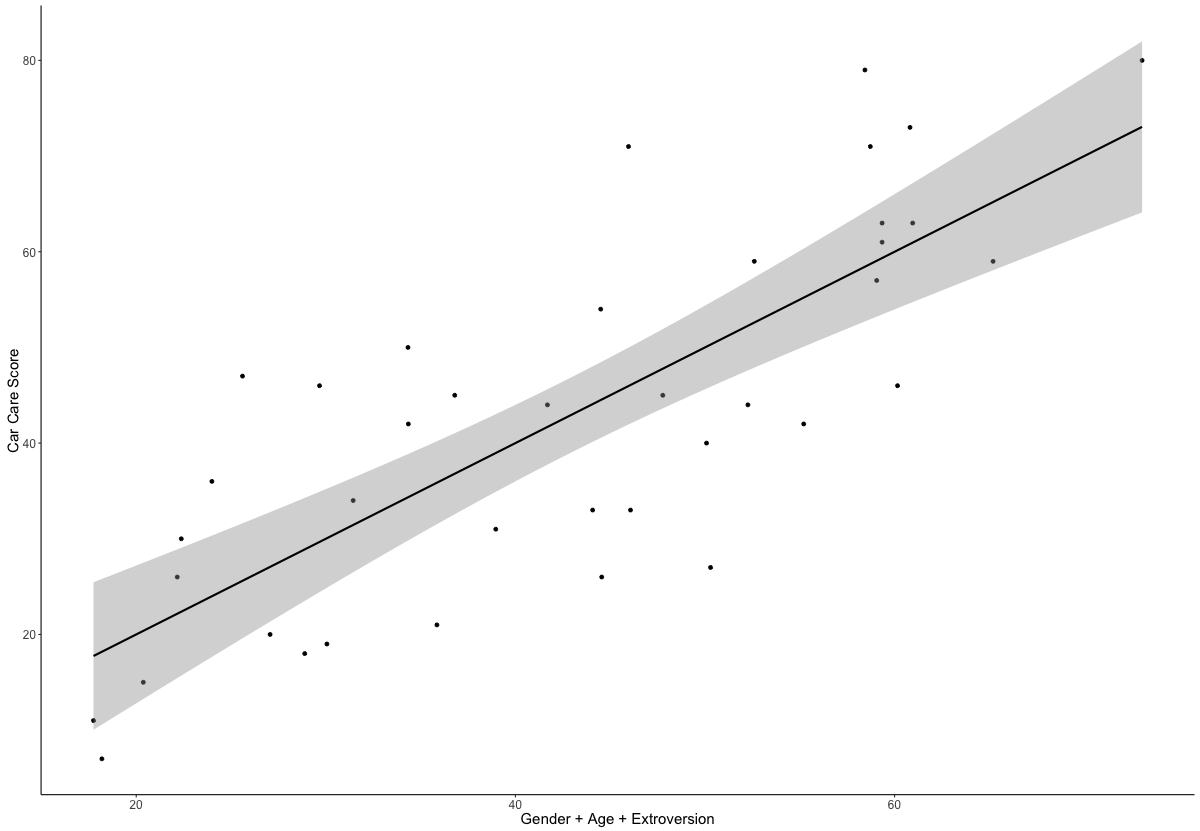
cleanup +

geom\_point() + ##to get the dots

geom\_smooth(method = “lm”, color = “black”) + ##to get the regression line

xlab(“Formula of line”) +

ylab(“DV”)



*Figure 1*. Predicted values of regression equation for gender, age, and extroversion predicting car care scores. Gray band indicates 95% confidence interval.

**Results**

Age, gender, and extroversion were used to predict a person’s overall care for their car. The data were screened for assumptions, and one participant was eliminated due to high Cook’s and Leverage values. Linearity, normality, multicollinearity, homogeneity, and homoscedasticity were all met.

Age and gender were entered first into a hierarchical regression to control for demographic differences in care maintenance. Overall, this model was significant, indicating that demographics predict how much a person takes care of their car, *F*(2, 36) = 21.66, *p* < .001, *R2* = .55. Gender was a stronger predictor of car maintenance, *β* = 0.68, *t*(36) = 6.00, *p* < .001, *pr2* = .50, which showed that males are more likely to take care of their cars by regular maintenance. Age was also positively related to car maintenance, *β* = 0.33, *t*(36) = 2.92, *p* < .01, *pr2* = .19; therefore, older participants indicated better car care. Next, extroversion was added in a second step to examine its predictive value after controlling for demographic variables. The addition of this variable was significant, Δ*F*(1, 35) = 5.96, *p* =.02, *ΔR2*= .06. Participants who were more extroverted took better care of their cars, *β* = 0.33, *t*(35) = 2.44, *p* = .02, *pr2* = .15.