Final Project

Andrew Shapero

BST 260

Overview, Motivation, and Related Work

Particulate matter is associated with a variety of health outcomes.

For example Dr. Douglas Dockery's 1993 article *Association Between Air Pollution and Mortality in Six U.S. Cities* demonstrated that fine particles were associated with increased mortality. Similarly, Dr. C. Arden Pope III's 1991 article *Respiratory Health and PM10 Pollution* demonstrated that elevated levels of particulate matter with an aerodynamic diameter less than ten microns (PM_{10}) pollution were associated with increases in reported symptoms of respiratory disease and use of asthma medication.

However, answers remain regarding the biological mechanisms in the relationship between particulate matter air pollution and adverse health outcomes. Examining the components of particulate matter might help explain this relationship.

As such, it is essential to characterize human exposures to the different components of particulate matter.

Recently, there has been increased interest in the metal components of particulate matter with an aerodynamic diameter less than 2.5 microns ($PM_{2.5}$). As such, for this study, I examined the relationship between particulate metal exposures among workers in the trucking industry and their biomarkers of inflammation.

Data Description

The data used for this analysis was collected from 140 terminal-based workers from trucking terminals in Carlisle, PA and South Chicago, IL. In March 2007 and June 2007, blood samples were collected from the workers. A health and exposure questionnaire was administered when blood was collected.

Air pollution exposures were collected via ($PM_{2.5}$) filters, which were then examined for metals with an energy dispersive X-ray fluorescence (EDXRF) spectrometer.

For this project, biomarkers and covariate data were in separate files from the exposure data. The files then had to be merged by `sampleid` . More details are available in the following sections. Not much data cleaning was required, although to examine each metal I do convert the data frame from a *wide* to a *long* format, as explained in later sections.

Data is included in my GitHub repository. https://github.com/pomegranate511/BST260_Final_Project (https://github.com/pomegranate511/BST260_Final_Project)

The GitHub repository can also be accessed through my project website. https://sites.google.com/view/metalscrp/ (https://sites.google.com/view/metalscrp/)

Initial Questions

My initial question is whether lead exposure is associated with increased C-Reactive Protein (CRP) blood levels. CRP is a biomarker of systemic inflammation.

However, I also wanted to explore the relationships between other metals exposures and CRP blood levels. As such, I explored the relationship between each of the 51 metals examined in this dataset and CRP blood levels.

To avoid issues of multiple testing, I used a Bonferroni correction.

$$\alpha^* = \alpha/m$$

where m is the number of tests being made.

Exploratory Analysis

Read In Data

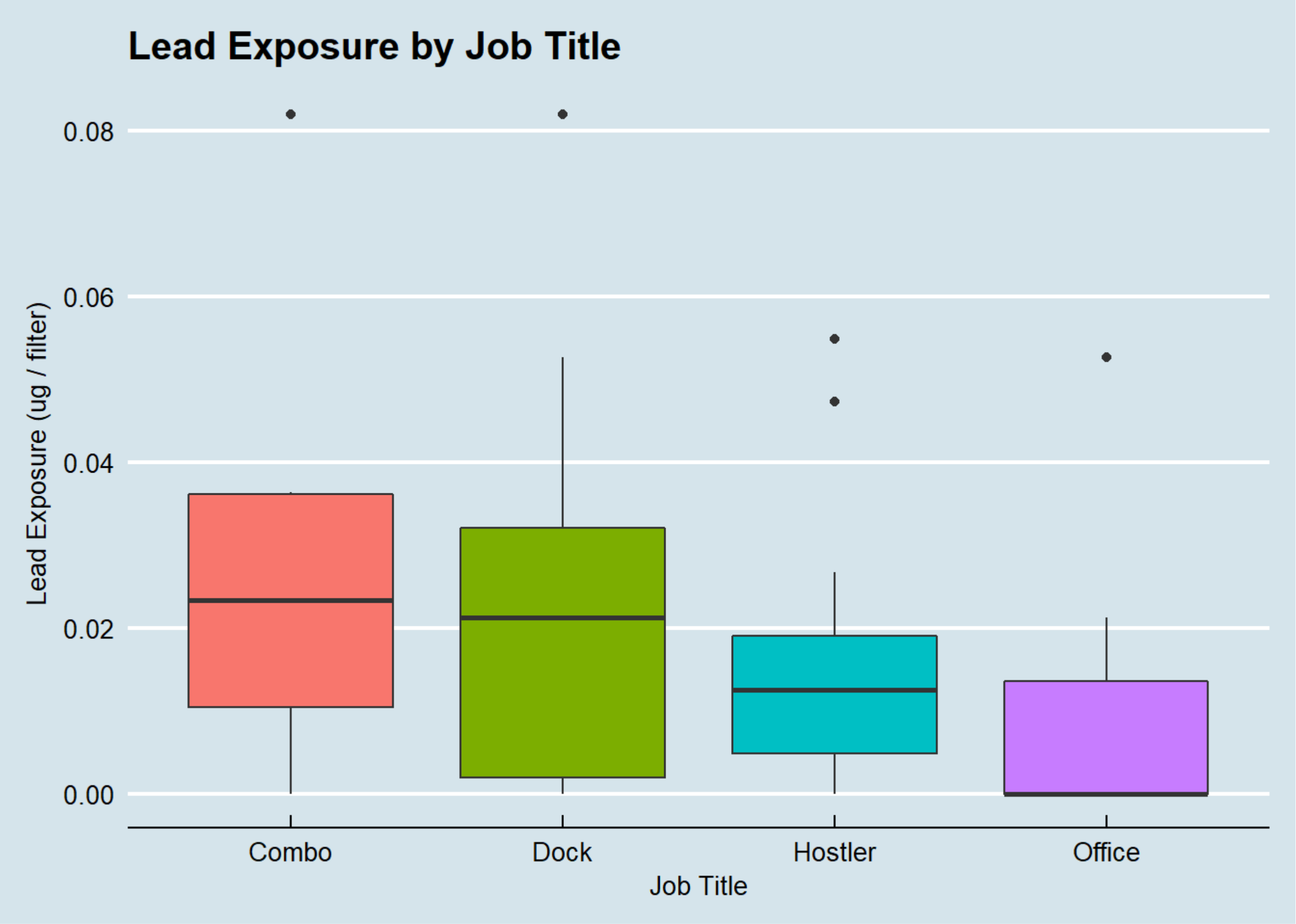
Let’s read in our biomarkers data and then our metals and covariates data. And then we can merge those two datasets.

```
biomarkers <- read_excel ("xfr_with_inflamm.xls")
metals <- read_excel ("XRF_results_HVELX63X.xls", sheet = 3)
data <- left_join (biomarkers, metals, by = "sampleid")
```

Single-Variable Analyses

Now let’s take a look at lead levels by each standardized job title. Are certain occupations more likely to be exposed to lead?

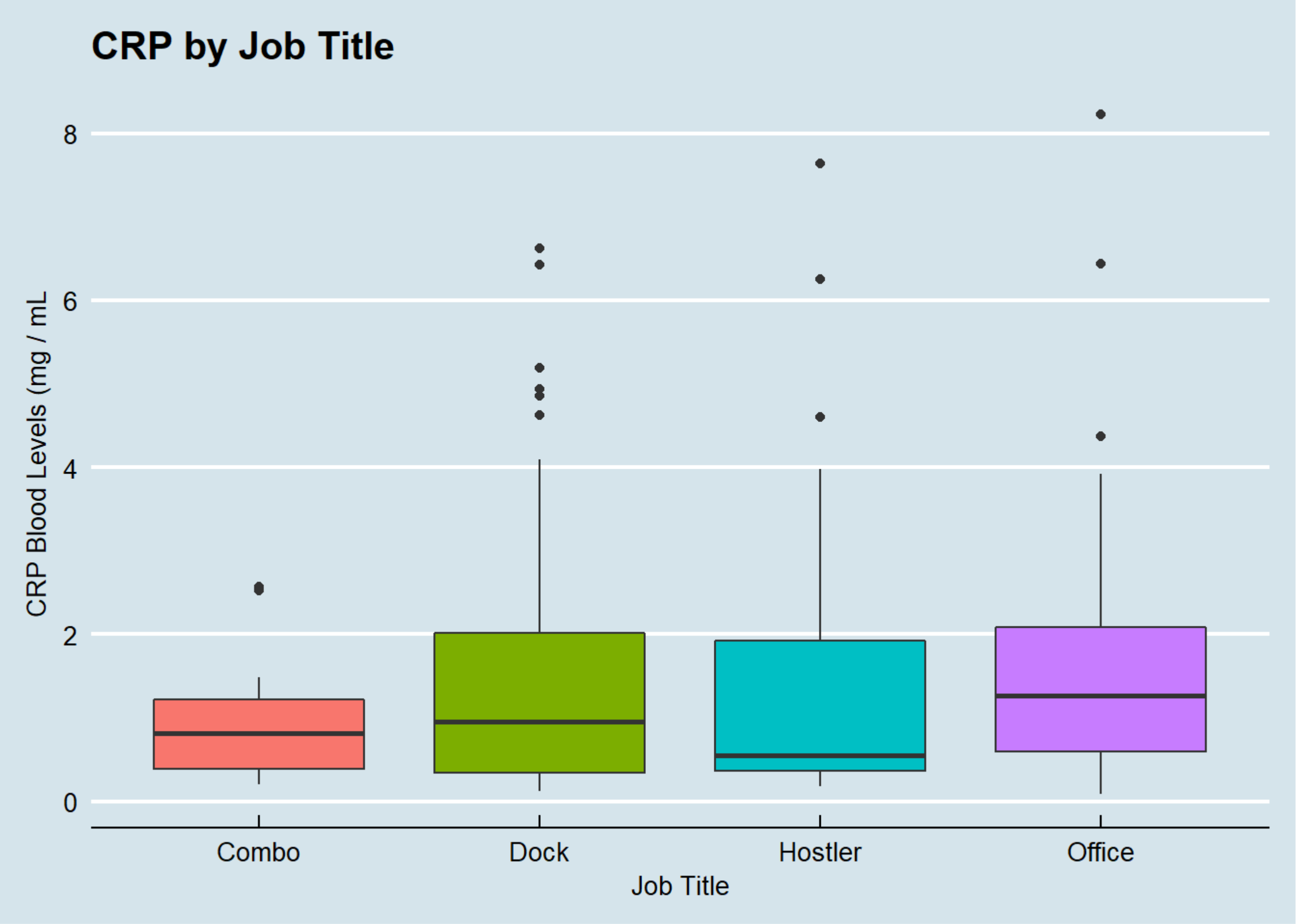
```
data %>% ggplot (aes (x = standard_job_title, y = PBXC, fill = standard_job_title)) +
          geom_boxplot () +
          xlab ("Job Title") +
          ylab ("Lead Exposure (ug / filter)") +
          ggtitle ("Lead Exposure by Job Title") +
          guides (fill = FALSE) +
          theme_economist ()
```



It appears that office jobs likely have the lowest lead exposure. This makes sense. We shouldn't expect office workers to have significant exposures.

Now let's look at CRP levels by each standardized job title. Are certain occupations more likely to show biomarkers of inflammation?

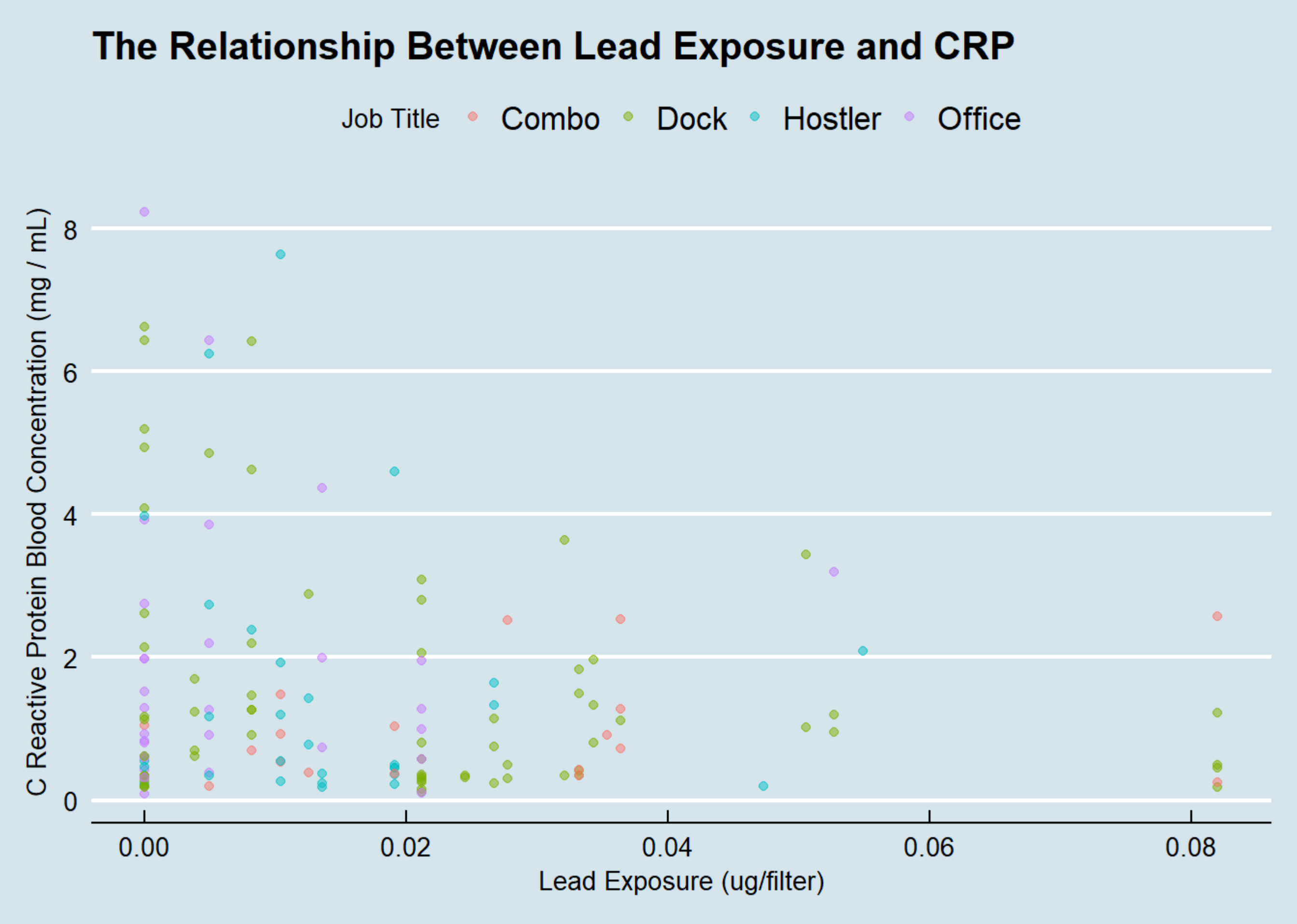
```
data %>% ggplot (aes (x = standard_job_title, y = CRP, fill = standard_job_title)) +
          geom_boxplot () +
          xlab ("Job Title") +
          ylab ("CRP Blood Levels (mg / mL") +
          ggtitle ("CRP by Job Title") +
          guides (fill = FALSE) +
          theme_economist ()
```



We pretty much see the opposite from the previous plot. Here we see that the office workers have the higher biomarkers of inflammation.

Now let’s look at the association between lead exposure and CRP levels. I’ll also use color to show each observation’s standardized job title.

```
p<- data %>% ggplot (aes (x = PBXC , y = CRP, col = standard_job_title)) +
          geom_point(alpha = 0.5) +
          xlab ("Lead Exposure (ug/filter)") +
          ylab ("C Reactive Protein Blood Concentration (mg / mL)") +
          ggtitle ("The Relationship Between Lead Exposure and CRP") +
          labs (col = "Job Title") +
          theme_economist()
p
```



From the above plot, it seems that there is an inverse relationship between lead exposure and CRP blood concentration. This is the opposite of what I expected. However, there could be confounding factors in this relationship. For example, the workers who are generally healthier might work in more physically demanding jobs that have increased exposures. In this case of reverse causation, the healthier workers (those with lower CRP) are exposed to more pollution.

Now let’s run a simple unadjusted linear regression to check if there truly is an inverse relationship between lead exposure and CRP. This is CRP as a function of lead exposure.

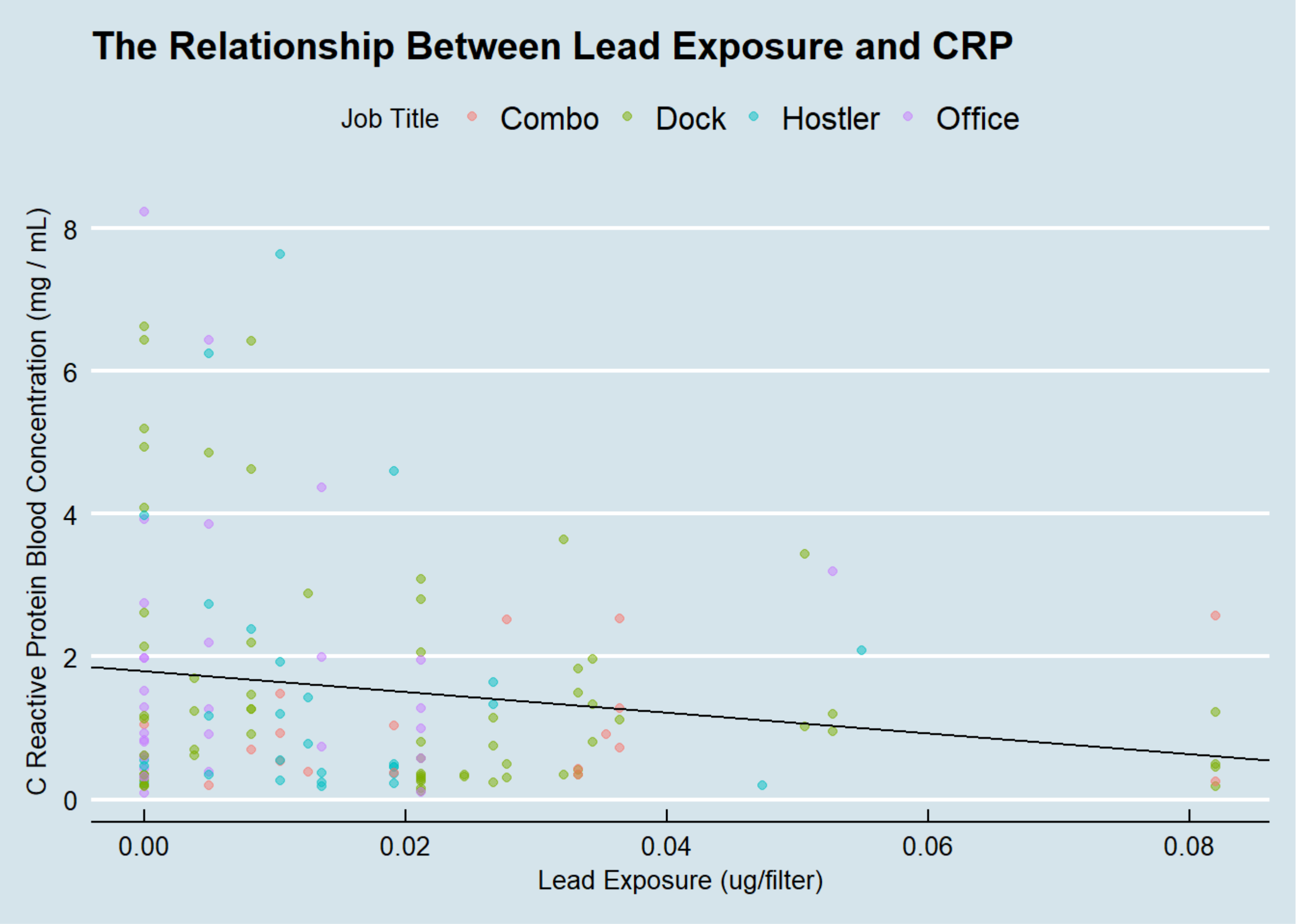
```
fit <- lm (CRP ~ PBXC, data = data, na.rm = TRUE)
fit <- tidy(fit)
fit
```

```
##          term   estimate std.error statistic      p.value
## 1 (Intercept)   1.795326 0.1871858  9.591147 3.972930e-17
## 2        PBXC -14.505816 7.1653893 -2.024428 4.478785e-02
```

```
int <- fit$estimate [1]
m <- fit$estimate [2]
```

From the regression we see that for every 1 unit increase in lead exposure is associated with a -14.5 mg/mL decrease in CRP. We can now add the best fit line from the simple linear regression to our graph.

```
p + geom_abline ( intercept = int, slope = m)
```



Data Cleaning

So there's the data for one metal. We also want to account for potential confounders. But we also want to make sure we can look at other metals in the dataset. To do that, I'm going to make a *long* dataset instead of a *wide* dataset. In this case, we can look at all metals at the same time.

Here we convert the data into a *long* format.

```
tidy_data <- data %>% gather (Code, concentration, `NAXC`:`URXU`)
```

Each metal reading has a concentration and an error estimate. Let's get rid of the error estimates for now, as the actual readings are our best estimates of exposure. Each of the estimates ends in "XC". These are the data points we want to keep in our data frame.

```
tidy_data <- tidy_data %>% filter (str_sub (Code, -2) == "XC")
```

Let's also rename the metals, so that they correspond to actual metal names. I made an Excel sheet to decipher each of the codes. Let's read that in and then translate the metal codes to the actual metal names. We'll then join the metal codes to the `tidy_data` data frame.

```
metal_codes <- read_csv ("metal_codes.csv")
tidy_data <- left_join(tidy_data, metal_codes, by = "Code")
```

Further Single-Variable Analyses

Now let’s run a regression for every metal using the `do` function. We’re doing the same simple unadjusted regression we did earlier for lead, but now for all 51 metals.

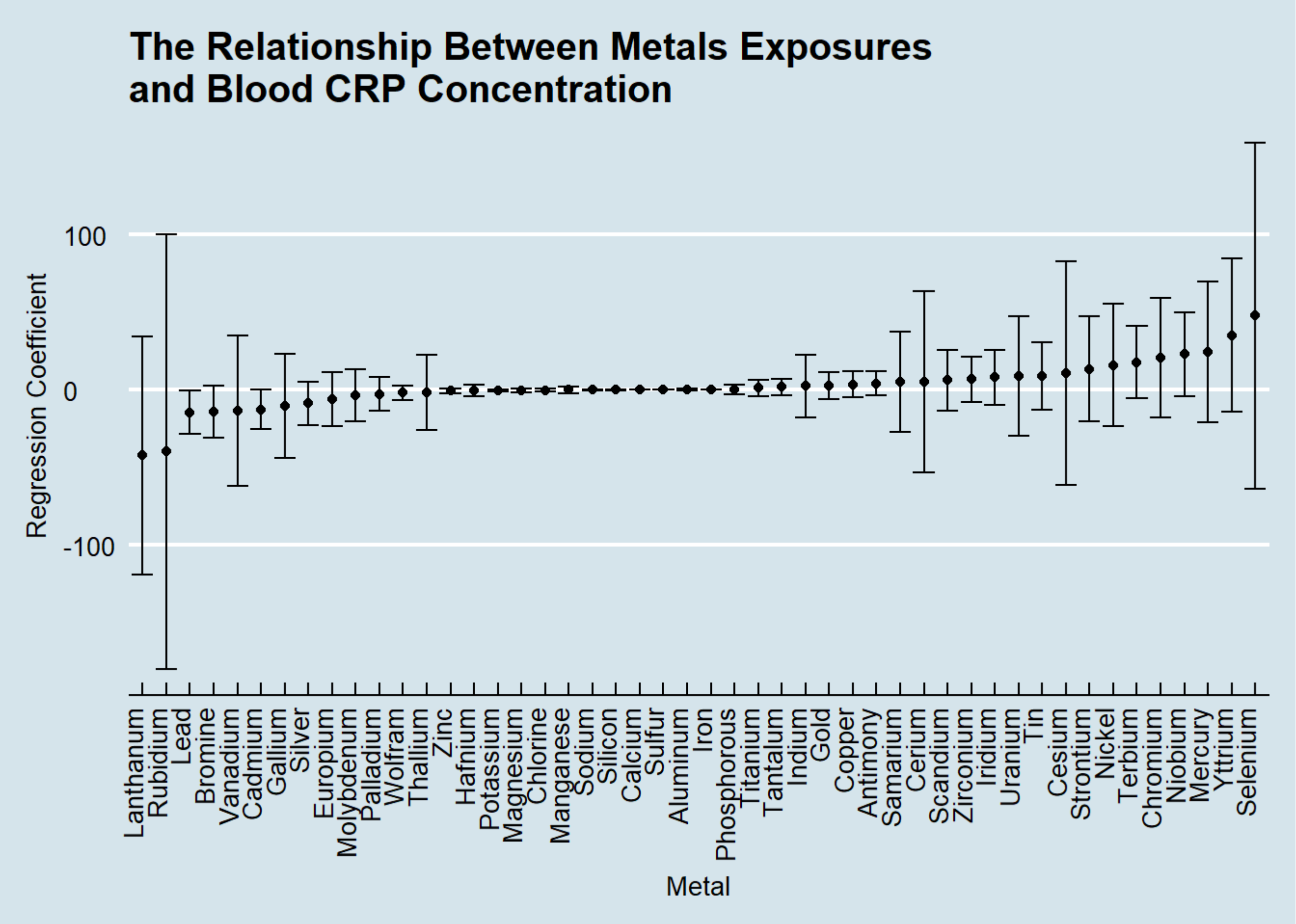
```
reg <- tidy_data %>% group_by (Metal) %>%
  do (tidy (lm (CRP ~ concentration, data = .), conf.int = TRUE))
```

The table that we just generated has estimates for the intercepts and the coefficients for 51 different regression models. Let’s filter out all the intercepts. We’re not necessarily interested in those.

```
reg <- reg %>% filter (term != "(Intercept)")
```

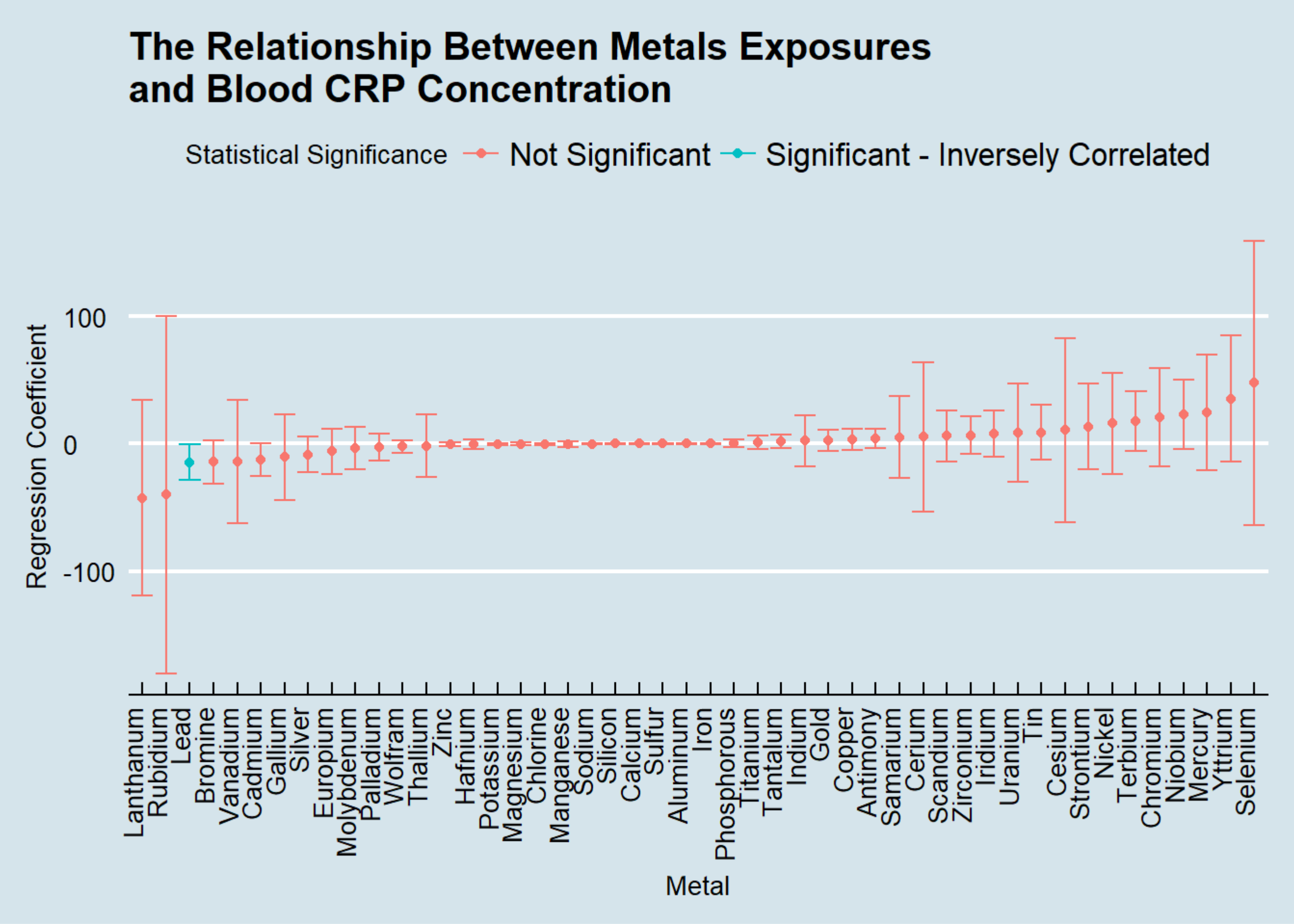
Now we can show the estimates and confidence intervals for each of the metals. But the confidence interval for barium is way too wide. Let’s filter that out, as it is obscuring the other confidence intervals.

```
reg %>% filter (Metal != "Barium") %>%
  ggplot ( aes ( x = reorder (Metal, estimate), y = estimate, ymin = conf.low, ymax = conf.high)
) +
    geom_errorbar () +
    geom_point () +
    xlab ("Metal") +
    ylab ("Regression Coefficient") +
    ggtitle ("The Relationship Between Metals Exposures\nand Blood CRP Concentration") +
    theme_economist () +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Let’s now create a color code so we can see if any of the relationships are statistically significant. This is the same plot as above but with a color code for statistical significance.

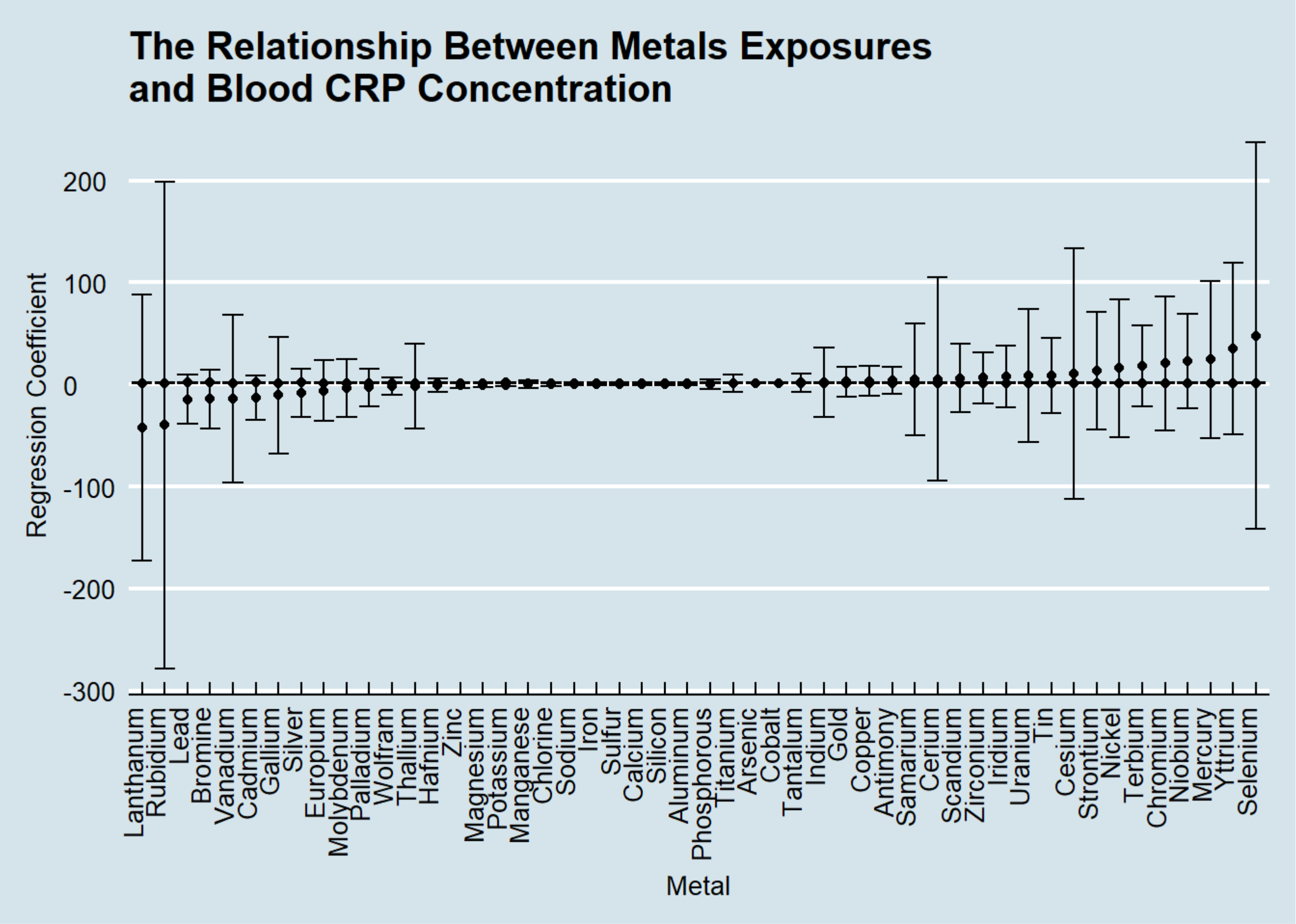
```
reg <- reg %>% mutate (
  sig = ifelse(conf.high < 0 , "Significant - Inversely Correlated",
               ifelse(conf.high >0 | conf.low <0, "Not Significant",
                      ifelse(conf.low > 0, "Significant - Positively Correlated" , NA)))
)

reg %>% filter (Metal != "Barium") %>%
  ggplot (aes (x = reorder (Metal, estimate), y = estimate, ymin = conf.low, ymax = conf.high, c
ol = sig)) +
    geom_errorbar () +
    geom_point () +
    xlab ("Metal") +
    ylab ("Regression Coefficient") +
    ggtitle ("The Relationship Between Metals Exposures\nand Blood CRP Concentration") +
    theme_economist () +
    theme (axis.text.x = element_text (angle = 90, hjust = 1)) +
    guides (col = guide_legend (title = "Statistical Significance"))
```



Here we see that only lead is statistically significantly associated with CRP. But we did 51 different analyses. Generally, we use a p-value of 0.05 for statistical significance. So if there is truly no relationship, there is a 5% chance that we observe a significant relationship. When we do one test, we are willing to accept this 5% chance for what’s called a Type I error. But when we do 51 different tests, we would expect to find approximately 2.5 significant relationships even if there were truly no significant relationships.

So let’s apply a Bonferroni correction to account for multiple testing, given that we are looking at 51 different metals. We’ll use $m = 51$ since we are performing 51 different tests.

```
#Establish adjusted alpha and new confidence level.
alpha <- 0.05 / 51
conf_level <- 1 - alpha

#Re-run simple unadjusted regression for each metal with the new confidence level.
reg2 <- tidy_data %>% group_by (Metal) %>%
  do (tidy (lm (CRP ~ concentration, data = .), conf.int = TRUE, conf.level = conf_level))

#Remove intercept terms.
reg <- reg %>% filter (term != "(Intercept)")

#Add indicators of significance.
reg2 <- reg2 %>% mutate (
  sig = ifelse(conf.high < 0 , "Significant - Inversely Correlated",
               ifelse(conf.high >0 | conf.low <0, "Not Significant",
                      ifelse(conf.low > 0, "Significant - Positively Correlated" , NA)))
)

#Plot estimates and confidence intervals, again removing barium from the plot because its confid
ence intervals are very wide.
reg2 %>% filter (Metal != "Barium") %>%
  ggplot ( aes ( x = reorder (Metal, estimate), y = estimate, ymin = conf.low, ymax = conf.high)
) +
    geom_errorbar () +
    geom_point () +
    xlab ("Metal") +
    ylab ("Regression Coefficient") +
    ggtitle ("The Relationship Between Metals Exposures\nand Blood CRP Concentration") +
    theme_economist () +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

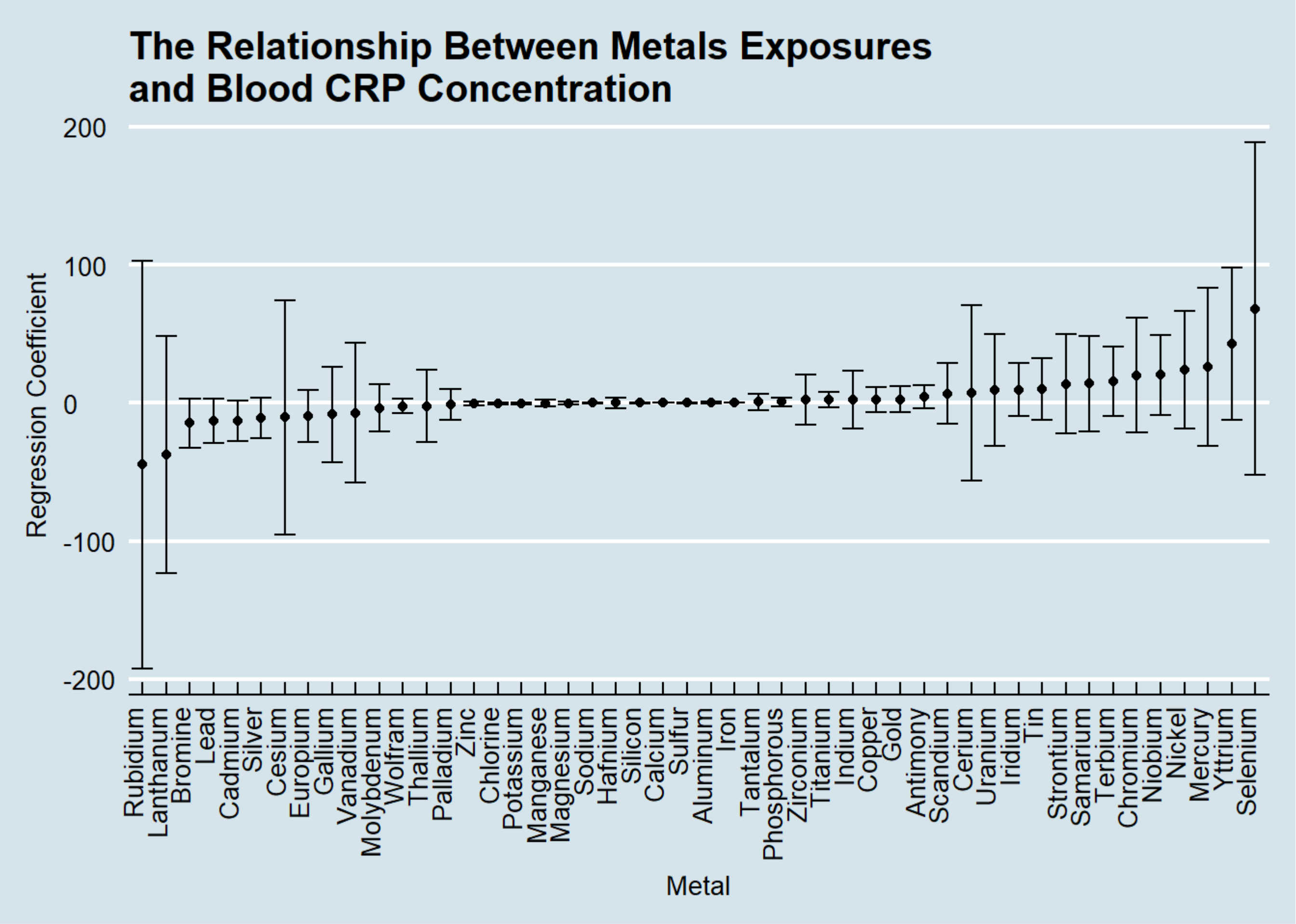


Now we see that none of metals are statistically significantly associated with CRP. However, we have already acknowledged that there could be confounding.

Adjusting for Covariates

Let’s include sex, race, education, and smoking status, as these are pretty typical covariates to include in an analysis. Let’s also include job, as this is an occupational study, and job might be associated with inflammatory biomarkers and with metals exposures. We’ll only end up focusing on the job title and concentration variables, as we are not necessarily interested in the effects of these covariates on CRP.

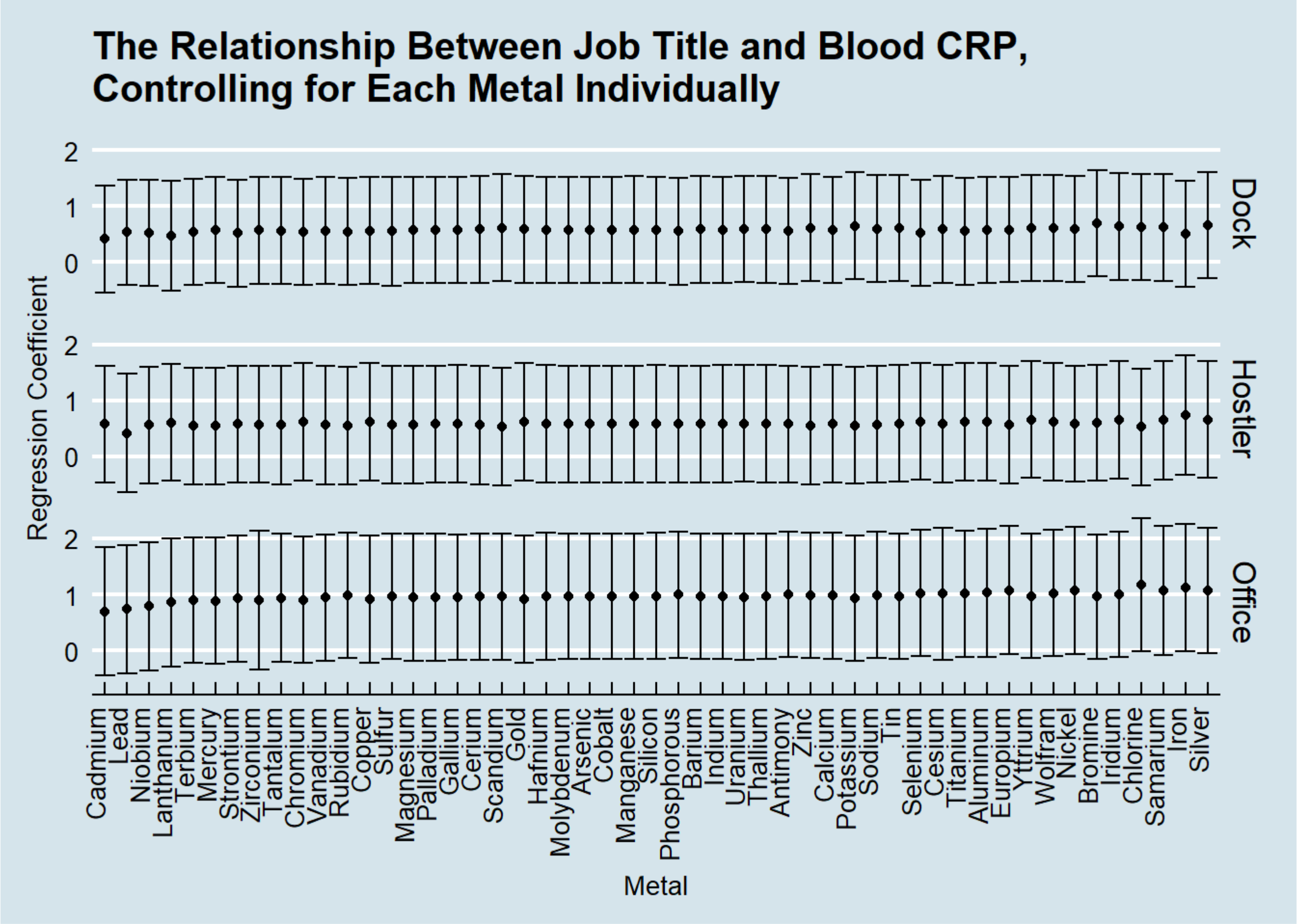
```
#Run regression for each metal. Include covariates and job title variables this time.
reg <- tidy_data %>% group_by (Metal) %>%
  do (tidy (lm (CRP ~ concentration + Gender + White + Education + Avg_cigarettes + standard_job
_title , data = .,  na.rm = TRUE, conf.int = TRUE)))

#Filter out all the intercepts and coefficients that do not correspond with the concentration te
rm. Note that I also filter out the job title variables here. But we'll look at those later.
conc_coeffs <- reg %>% filter (term != "(Intercept)" & term != "Gendermale" & term != "Whiteyes"
  & term != "Educationhigh school or GED" & term != "Educationless than high school" & term !=
"Avg_cigarettes" & term != "standard_job_titleDock" & term != "standard_job_titleHostler" & term
 != "standard_job_titleOffice")

#Add the confidence intervals.
conc_coeffs <- conc_coeffs %>% mutate (
  conf.high = estimate + qnorm(0.975) * std.error,
  conf.low = estimate - qnorm(0.975) * std.error
)

#Add the indicators of statistical significance.
conc_coeffs <- conc_coeffs %>% mutate (
  sig = ifelse(conf.high < 0 , "Significant - Inversely Correlated",
               ifelse(conf.high >0 | conf.low <0, "Not Significant",
                      ifelse(conf.low > 0, "Significant - Positively Correlated" , NA)))
)

#Plot the concentration coefficients for each metal. Again, we remove the coefficient for barium
 because it has a very wide confidence interval.
conc_coeffs  %>% filter (Metal != "Barium") %>%
  ggplot ( aes ( x = reorder (Metal, estimate), y = estimate, ymin = conf.low, ymax = conf.high)
) +
    geom_errorbar () +
    geom_point () +
    xlab ("Metal") +
    ylab ("Regression Coefficient") +
    ggtitle ("The Relationship Between Metals Exposures\nand Blood CRP Concentration") +
    theme_economist () +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Above we see that none of the relationships is statistically significant. I hadn't accounted for multiple testing here. But obviously, if we were to adjust, we would still have no statistically significant coefficients.

Now let's look at the coefficients on the job title variables.

```
#We're using the same regression model as in the previous code chunk. But now we are filtering s
o that we only keep the job title variables.
job_coeffs <- reg %>% filter (term == "standard_job_titleDock" | term == "standard_job_titleHost
ler" | term == "standard_job_titleOffice") %>% mutate (term = ifelse(term == "standard_job_title
Dock", "Dock", ifelse(term == "standard_job_titleHostler", "Hostler", ifelse(term == "standard_j
ob_titleOffice", "Office", NA))))

#Add the confidence intervals.
job_coeffs <- job_coeffs %>% mutate (
  conf.high = estimate + qnorm(0.975) * std.error,
  conf.low = estimate - qnorm(0.975) * std.error
)

#Add the indicators of statistical significance.
job_coeffs <- job_coeffs %>% mutate (
  sig = ifelse(conf.high < 0 , "Significant - Inversely Correlated",
               ifelse(conf.high >0 | conf.low <0, "Not Significant",
                      ifelse(conf.low > 0, "Significant - Positively Correlated" , NA)))
)

#Plot the job title coefficients from each metal model.
job_coeffs  %>%
  ggplot ( aes ( x = reorder (Metal, estimate), y = estimate, ymin = conf.low, ymax = conf.high)
) +
    geom_errorbar () +
    geom_point () +
    xlab ("Metal") +
    ylab ("Regression Coefficient") +
    ggtitle ("The Relationship Between Job Title and Blood CRP,\nControlling for Each Metal Indi
vidually") +
    theme_economist () +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) + facet_grid(term~.)
```



Here we see no statistically significant relationships.Again, I didn’t account for multiple testing here because no relationships were significant in the first place.

Let’s look at the RMSE of our models. Even if none of the occupation or metal variables is statistically significant, we can still assess the predictive power of our model. We can use RMSE to estimate the predictive power of our models. Of course, prediction is not the purpose of this projection. I’m just calculating RMSE as a side note.

We’ll define the following function, which calculates RMSE.

```
RMSE <- function(true_ratings, predicted_ratings){
    sqrt(mean((true_ratings - predicted_ratings)^2))
}
```

Here we make all of our predictions, and then we’ll compare . We can use the `augment` function in the `broom` package. It uses basically the same syntax as the `tidy` and `lm` function in the same package.

```
tidy_data[is.na(tidy_data)] <- 0 #replace all NAs in the dataframe with zero.

predictions <- tidy_data %>% group_by (Metal) %>% do (augment (lm (CRP ~ concentration + Gender
+ White + Education + Avg_cigarettes + standard_job_title , data = .))) #save predictions from r
egression in new `predictions` data frame.

predictions$sampleid <- tidy_data$sampleid #add sampleid's to `predictions` data frame.
```

Now that we’ve calculated predictions for each metal model, we can see which model best predicts CRP levels. So we’ll set up a blank data frame for storing the RMSEs.

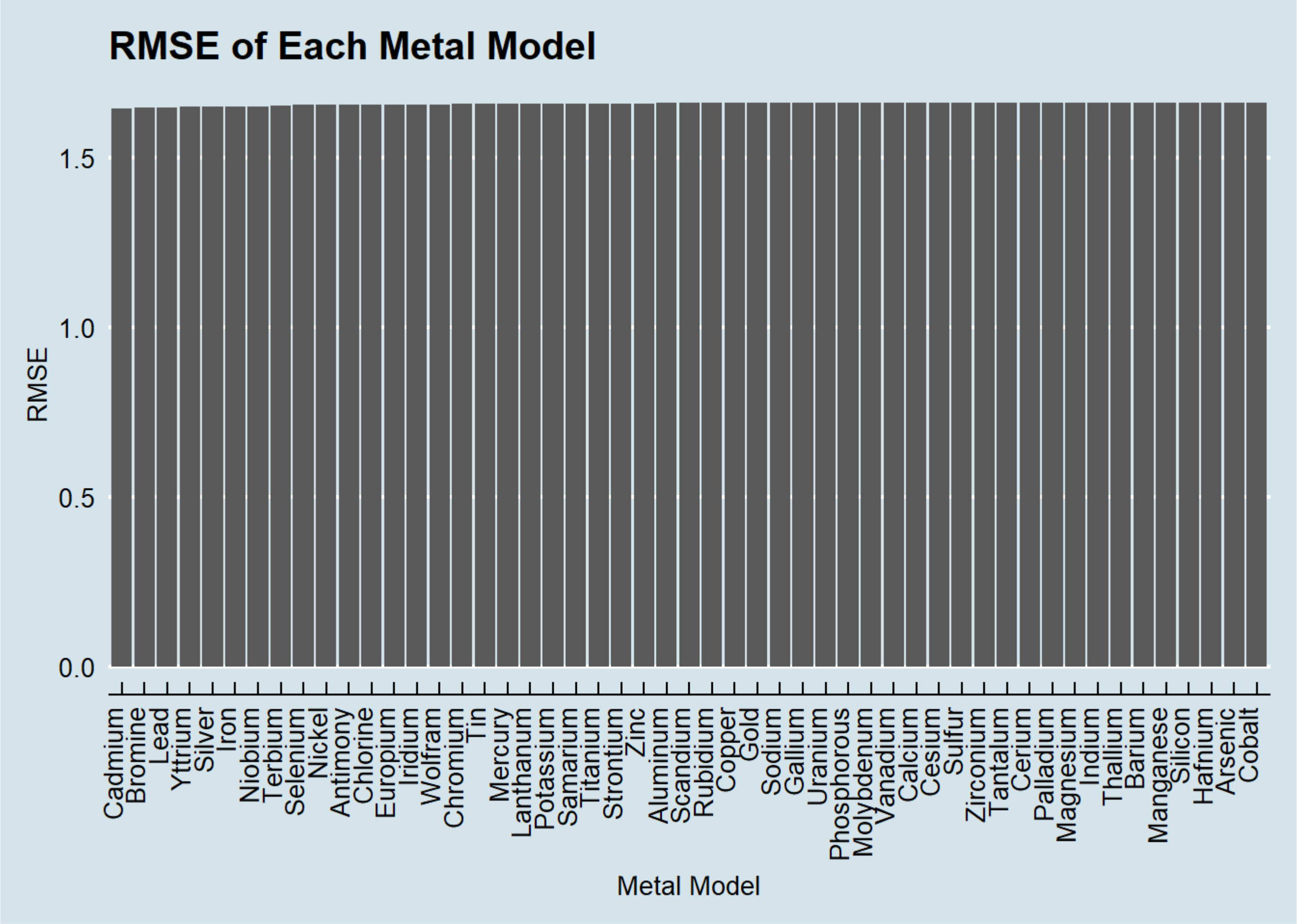
```
metal_RMSEs <- data.frame(
  Metal = metal_codes$Metal,
  RMSE = NA
)
metal_RMSEs$Metal <- as.character (metal_RMSEs$Metal)
```

No we are going to use a loop to populate the `metal_RMSEs` data frame that we just set up.

```
for( i in  1:51){ #run through the loop 51 times for each metal
CRP_prediction <- predictions %>% filter (Metal == metal_RMSEs[i, 1]) #In each loop, we loop at
 just one metal from the `predictions` data frame.
CRP_actual <- tidy_data %>% filter(Metal == metal_RMSEs[i, 1]) #filter out the tidy data so that
 it shows only the metal of interest for the loop number.

metal_RMSEs[i, 2] <- RMSE(CRP_actual$CRP, CRP_prediction$.fitted) #input actual CRP and predicte
d CRP to calculate RMSE with the function we established earlier. Store RMSE in the `metal_RMSEs
` data frame we established earlier.
}
```

Now we can see which metal model best predicts CRP.

```
metal_RMSEs %>%
  ggplot(aes(x=reorder(Metal, RMSE),y = RMSE)) +
  geom_bar(stat = "identity")+
  ggtitle ("RMSE of Each Metal Model" ) +
  xlab ("Metal Model") +
  theme_economist ()+
  theme (axis.text.x = element_text(angle = 90, hjust = 1))
```



Here we see that each metal model performs about equally well.

Final Analysis

For the final analysis we are going to use backwards stepwise covariate selection. Basically we are going to start with all covariates in the model and then we'll work backwards and remove statistically insignificant variables.

Here we set up a blank data frame that will store each of our estimates and t-statistics from each of the 51 regressions we are going to run.

```
metal = metal_codes$Metal
estimate = rep(NA, 51)
t_stat = rep(NA, 51)

metal_step = data.frame (metal, estimate, t_stat)
```

When we run each of the regressions, we might get a situation where the concentration variable is not included in the regression. We will use the follow function to identify if the regression does not include the concentration variable.

```
is.integer0 <- function(x)
 {
     is.numeric(x) && length(x) == 0L
 }
```

When we run each of the regressions, we are going to calculate RMSE. So let’s set up our blank data frame to store those later on.

```
step_RMSEs <- data.frame(
  Metal = metal_codes$Metal,
  RMSE = NA
)
step_RMSEs$Metal <- as.character (metal_RMSEs$Metal)
```

Here’s where we run the 51 backwards selection models. Be careful with this code chunk; it takes over an hour to run. In the loop we are storing the coefficients and t-scores from each regression. And we are also making predictions from each regression and comparing those predictions to the actual CRP values. We then calculate RMSE from each prediction and store those values.

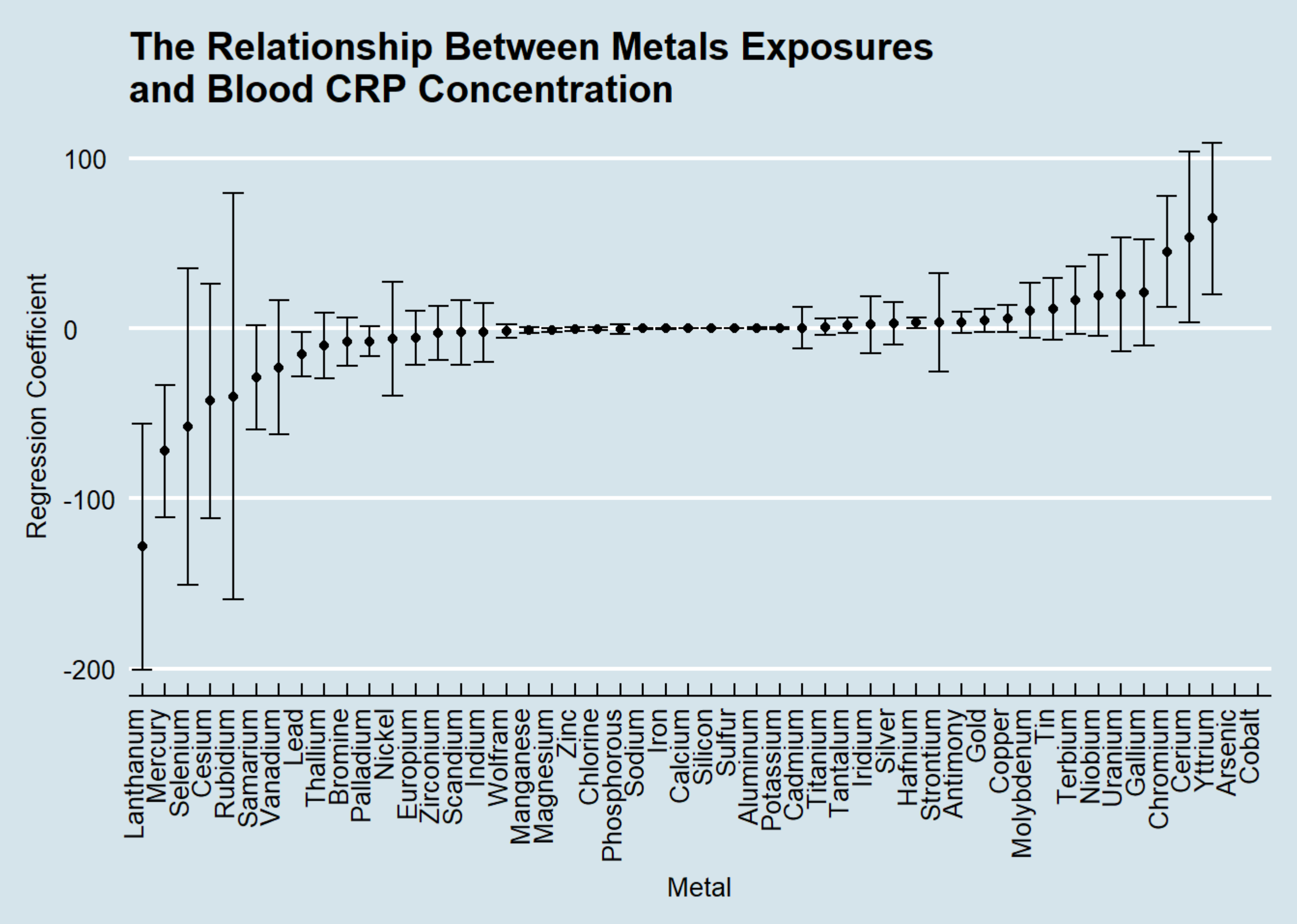
```
for(i in 1:51){
step_data <- tidy_data %>% filter (Metal == metal_step[i, 1]) #filter out the tidy_data so that
 it includes the metal of focus for the loop.

step<-stepAIC(lm(CRP~., data = step_data[, c(1, 6:23, 25:101, 103:111, 116:129, 134)]),
              scope = list(lower=as.formula(CRP ~ concentration), upper=as.formula(CRP ~ .))) #
 I'm including almost all of the variables. But I'm excluding a few that just don't really make
 sense to include.
s = summary(step)
sc = as.data.frame(s$coefficients) #store the coefficients.

sc[with(sc, order(-Estimate)), ]
sc$variable <- rownames(sc)
metal_step[i,2] <- ifelse(is.integer0(sc$Estimate[sc$variable == "concentration"]) == TRUE, NA,
sc$Estimate[sc$variable == "concentration"]) #Store just the concentration coefficient. Use the
 is.integer0 function so that if a model does not have a concentration variable, an NA will be s
tored for that metal.

metal_step[i, 3] <- ifelse(is.integer0(sc$`t value`[sc$variable == "concentration"]) == TRUE, NA
, sc$`t value`[sc$variable == "concentration"]) #save the t-score for the concentration coefficn
e.t Again, if the model does not have a concentration variable, store an NA for that metal.

step_prediction <- predict(step) #Predict CRP from regression model
step_RMSEs[i, 2] <- RMSE(CRP_actual$CRP, step_prediction) #Use actual CRP values and the predict
ion values from the model to calculate RMSE. Store RMSE value.
}
```

Now we will calculate the standard error and the confidence interval for each concentration coefficient. Note that we could have done this in the previous code chunk. But it’s a little more straightforward to do it here just because the above code chunk takes so long to run.

```
metal_step <- metal_step %>% mutate (
  std_err = estimate / t_stat,
  upper = estimate + qnorm(0.975)*std_err,
  lower = estimate - qnorm(0.975)*std_err
  )
```

Now let’s plot each of the coefficients and their confidence intervals. Again, we exclude barium because it has such a wide confidence interval.

```
metal_step %>% filter(metal != "Barium" ) %>%
  ggplot ( aes ( x = reorder (metal, estimate), y = estimate, ymin = lower, ymax = upper)) +
    geom_errorbar () +
    geom_point () +
    xlab ("Metal") +
    ylab ("Regression Coefficient") +
    ggtitle ("The Relationship Between Metals Exposures\nand Blood CRP Concentration") +
    theme_economist () +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

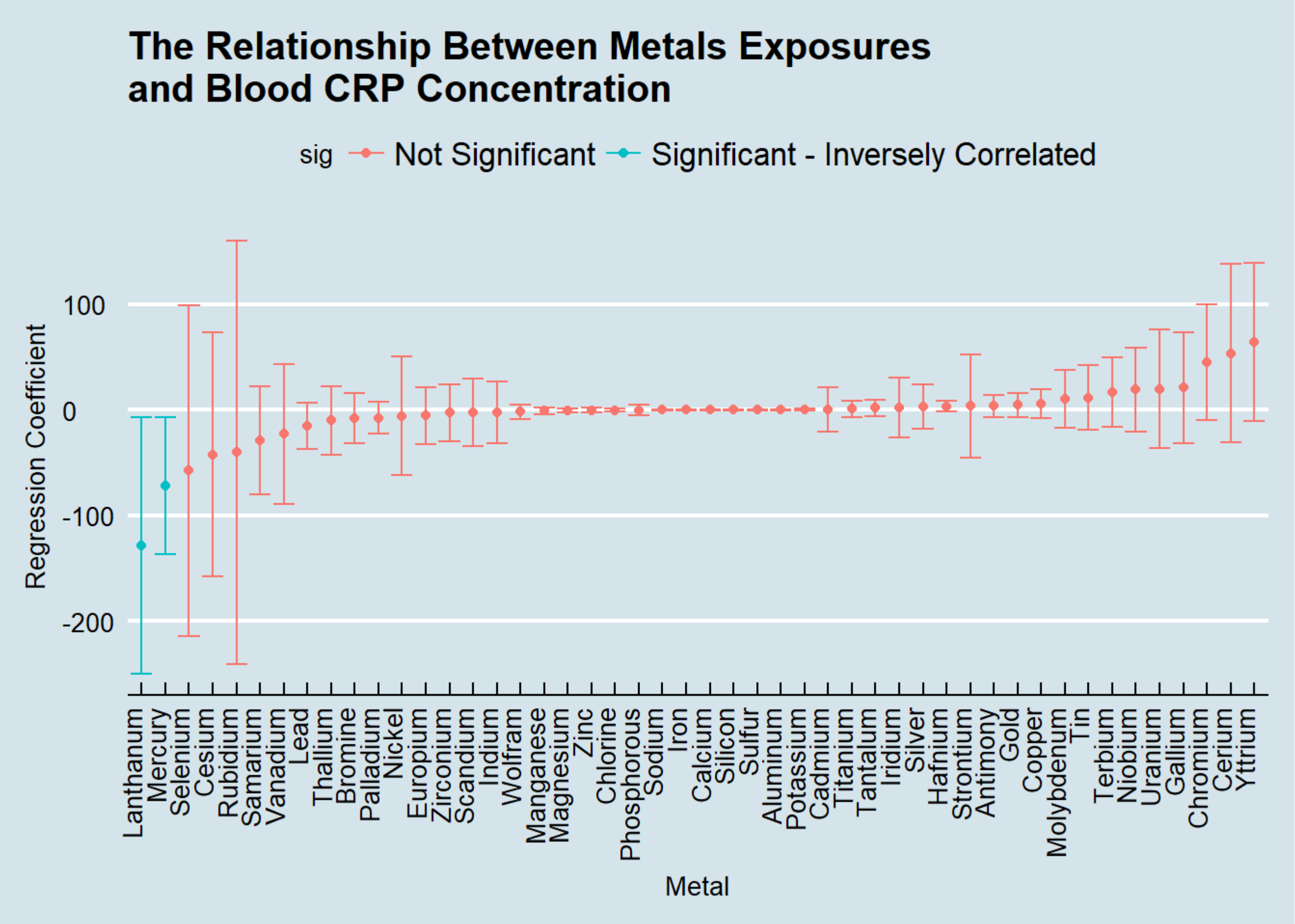


Now let’s apply a Bonferroni correction. And see if that affects the significance of any of our estimates. We’ll also add indicators of significance.

```
alpha <- 0.05 / 51
z <- 1 - alpha / 2
metal_step <- metal_step %>% mutate (
  upper_corrected = estimate + qnorm(z)*std_err,
  lower_corrected = estimate - qnorm(z)*std_err,
  sig = ifelse(upper_corrected < 0 , "Significant - Inversely Correlated",
               ifelse(upper_corrected >0 | upper_corrected <0, "Not Significant",
                      ifelse(lower_corrected > 0, "Significant - Positively Correlated" , NA)))
  )
```

Here we plot the concentration coefficients for each metal and the the Bonferroni adjusted confidence intervals.

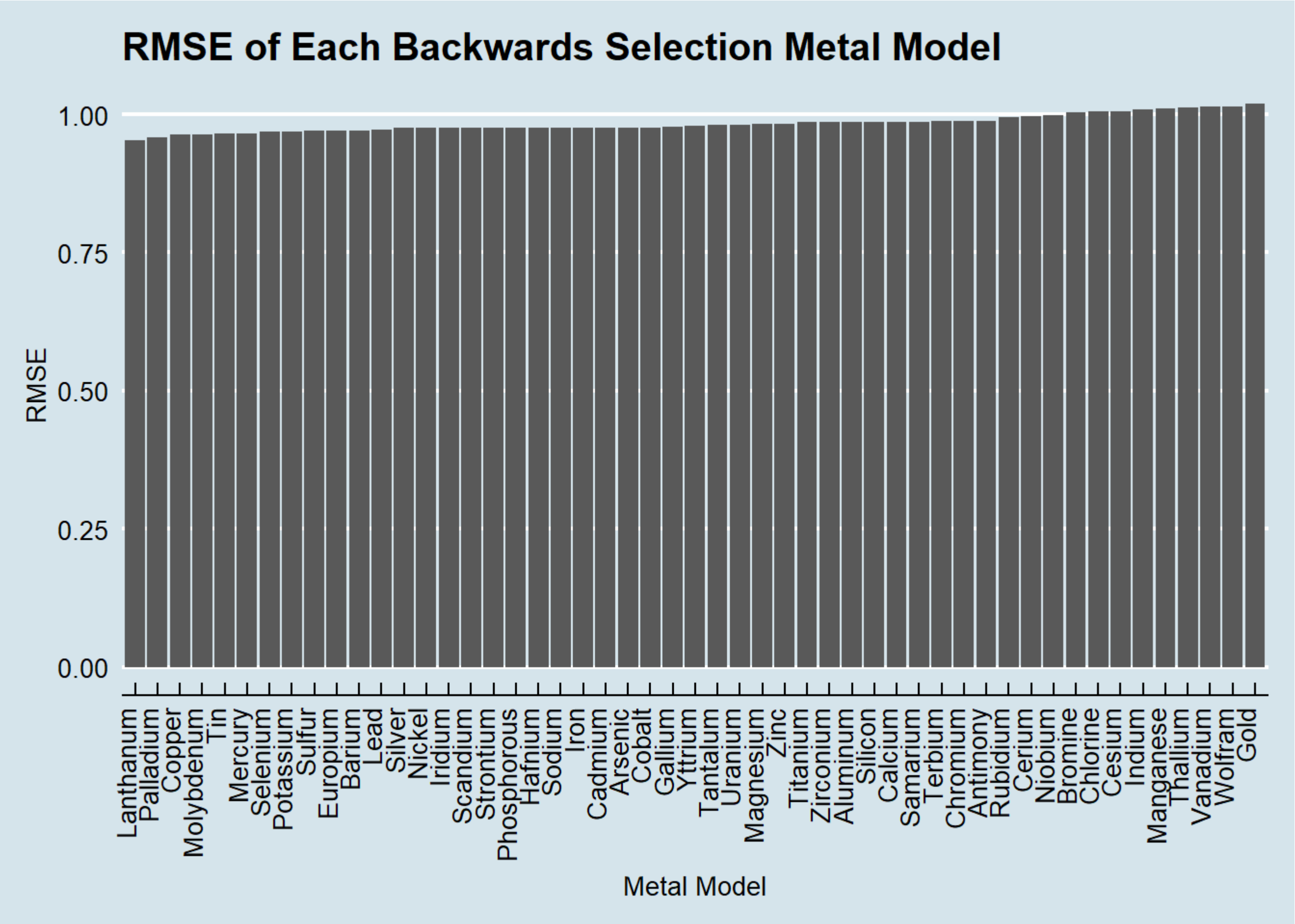
```
metal_step <- na.omit(metal_step)
metal_step %>% filter(metal != "Barium" ) %>%
  ggplot ( aes ( x = reorder (metal, estimate), y = estimate, ymin = lower_corrected, ymax = upper_corrected, col = sig  )) +
    geom_errorbar () +
    geom_point () +
    xlab ("Metal") +
    ylab ("Regression Coefficient") +
    ggtitle ("The Relationship Between Metals Exposures\nand Blood CRP Concentration") +
    theme_economist () +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Here we see that lanthanum and mercury are the only two metals that are statistically significantly associated with CRP. But they are inversely associated with CRP, which is the opposite of expected. I'll discuss this further in the conclusion.

Finally, let's look at the RMSEs for each of the regression models that we developed with the backwards selection models.

```
step_RMSEs %>%
  ggplot(aes(x=reorder(Metal, RMSE),y = RMSE)) +
  geom_bar(stat = "identity")+
  ggtitle ("RMSE of Each Backwards Selection Metal Model" ) +
  xlab ("Metal Model") +
  theme_economist ()+
  theme (axis.text.x = element_text(angle = 90, hjust = 1))
```



Here we see that the backwards selection models have much lower RMSEs than the previous models.

Conclusion

Running backward selection models for each metal, we found that lanthanum and mercury exposures are statistically inversely associated with CRP blood levels, a biomarker of inflammation. Even though the results are the opposite of the expected, I think that these results makes sense under considerations of reverse causality. It is possible that only the healthiest workers are exposed to metals. Metals exposure likely occurs during outside work, which might be more labor intensive, whereas the less healthy workers remain inside and unexposed.This is also consistent with what we saw in the first two plots, where we saw the office workers had the highest CRP but also the lowest lead exposure.

Nonetheless, I did expect to see some statistically significant positive associations. It is possible that some of the true adverse effects of metals were attenuated in the study by measurement error. The study only accounts for a snapshot of an individual worker’s exposure; it does not necessarily represent their true exposures. As such, the measurements that we have in our data set are not necessarily equal to the workers’ true exposures. Therefore, there is measurement error.

Non-differential measurement error (that is, measurement not associated with the outcome of interest) causes a downward bias towards the null. As such, for any metals that truly do have a significant positive association with CRP, that association may be attenuated towards the null.

Overall, I do not think that the results shown in this analysis should be taken to represent the true effects of metals exposure on CRP blood levels. However, I do think that the analysis shows how the program `R` can be used to explore the relationships between environmental exposures and adverse health outcomes. In my backwards selection models, `R` ran thousands of regression models and picked the best models. Such analysis could not be performed without data science skills. And I hope that this analysis provides insight on possible data exploration paths in the future.

Works Cited

1. Pope, C. A., Dockery, D. W., Spengler, J. D., & Raizenne, M. E. (1991). Respiratory Health and PM 10 Pollution. American Review of Respiratory Disease.
2. Dockery, D. W. et al. (1993). An Association Between Air Pollution and Mortality in Six U.S. Cities. New England Journal of Medicine.