Lab 2: Multiple Regression in the Wild

Partial Solutions

Social Media and Web Analytics @ TiSEM

Last updated: 02 May, 2021

Motivation

Linear regression is a workhorse model of a Marketing Analyst's toolkit. This is because it gives them the ability to describe data patterns, predict the value of marketing metrics in data and potentially make causal claims about the relationships between multiple variables.

In this tutorial you will apply linear regression to get first hand experience with these tools. We will focus both on how to linear regression in `R` and how to correctly interpret the results. You will use linear regression to evaluate the association between product characteristics and product price in an internet mediated market.

Learning Goals

By the end of this tutorial you will be able to:

1. Estimate Single and Multiple Regression models with R.
2. Interpret regression coefficients.
3. Discuss likely biases in regression coefficients due to omitted variable bias.
4. Discuss why regression standard errors may need to be adjusted for heteroskedasticity or clustering.
5. Estimate Fixed Effect regressions with and without clustered standard errors.
6. Present regression coefficients in a table and in a plot.

Instructions to Students

These tutorials are **not graded**, but we encourage you to invest time and effort into working through them from start to finish. Add your solutions to the `lab-02_answer.Rmd` file as you work through the exercises so that you have a record of the work you have done.

Obtain a copy of both the question and answer files using Git. To clone a copy of this repository to your own PC, use the following command:

```
$ git clone https://github.com/tisem-digital-marketing/smwa-lab-02.git
```

Once you have your copy, open the answer document in RStudio as an RStudio project and work through the questions.

The goal of the tutorials is to explore how to "do" the technical side of social media analytics. Use this as an opportunity to push your limits and develop new skills. When you are uncertain or do not know what to do next - ask questions of your peers and the instructors on the class Slack channel `#lab02-discussion`.

Multiple Regression Analysis

The advent of the internet, and the rise in user generated content has had a large effect on sex markets. In 2008 and 2009, Scott Cunningham and Todd Kendall surveyed approximately 700 US internet mediated sex workers. The questions they asked included information about their illicit and legal labor market experiences and their demographics. Part of the survey asked respondents to share information about each of the previous four sessions with clients.

To gain access to the data, run the following code to download it and save it in the file `data/sasp_panel.dta`:

```
url <- "https://github.com/scunning1975/mixtape/raw/master/sasp_panel.dta"
# where to save data
out_file <- "data/sasp_panel.dta"
# download it!
download.file(url, destfile = out_file, mode = "wb")
```

The data include the log hourly price, the log of the session length (in hours), characteristics of the client (such as whether he was a regular), whether a condom was used, and some characteristics of the provider (such as they race, marital status and education level). The goal of this exercise is to estimate the price premium of unsafe sex and think through any bias in the coefficients within the regression models we estimate.

You might need to use the following `R` libraries throughout this exercise:[1]

```
library(haven) # to read stata datasets
library(dplyr)
library(tidyr)
library(fixest)
library(broom)
library(ggplot2)
library(modelsummary)
```

1. Load the data. The data is stored as a Stata dataset, so it can be loaded with the `read_dta()` function from `haven`.

solution

```
sasp <- read_dta('data/sasp_panel.dta')
```

2. Some rows of the data have missing values. Let's drop these.[2] Write a short command to drop any rows which have missing values from the data.

solution

```
sasp <-
    sasp %>%
    drop_na()
```

As mentioned above, the focus for the rest of this exercise is the price premium for unprotected sex. In the `sasp` data, there is a variable `lnw` which is the log of the hourly wage and a variable `unsafe` which takes the value 1 if there was unsafe sex during the client's appointment and 0 otherwise.

[1] If you haven't installed one or more of these packages, do so by entering `install.packages("PKG_NAME")` into the R console and pressing ENTER.

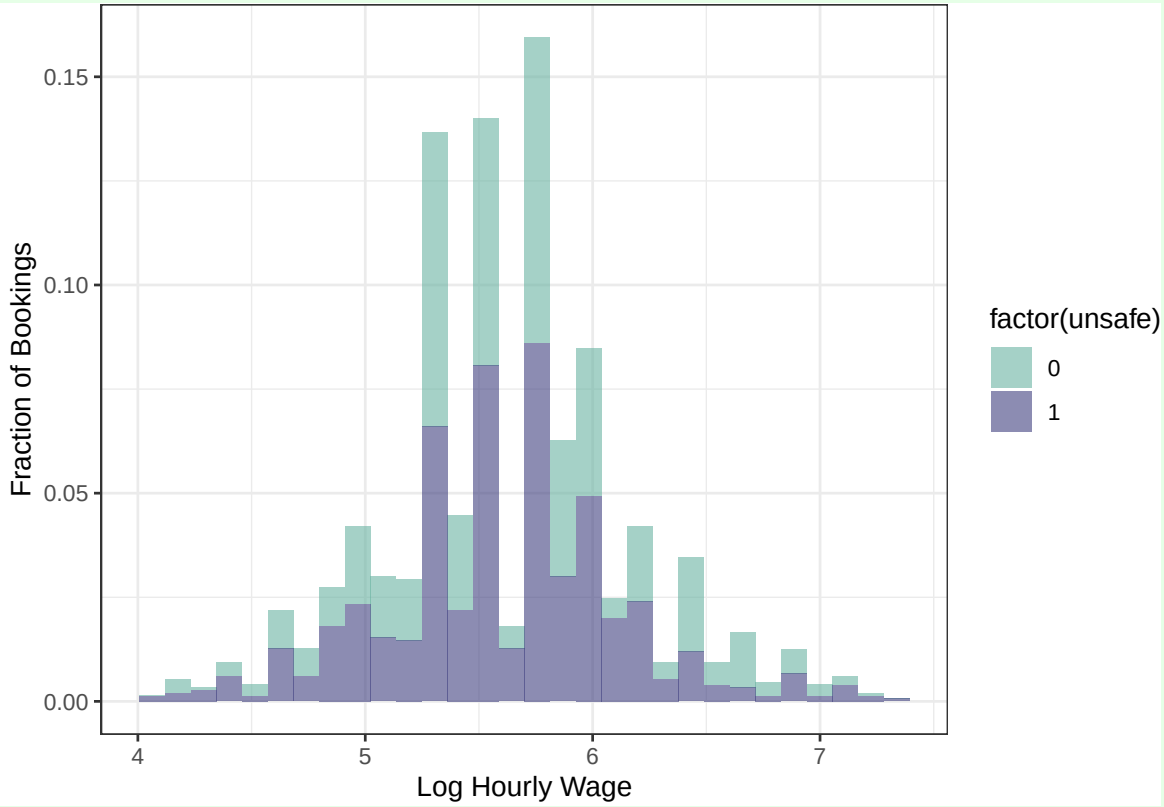
[2] Generally, we need to be quite careful when we make decisions about dropping rows of data, and think through what the consequences of it might be. We've not done this here because our goal was to illustrate how to estimate and interpret regression estimates, but we would encourage you to be careful when you do this in your own work. At a minimum, you should mention why you've dropped rows, and whether there is likely to be selection bias in your subsequent results.

3. Produce a diagram that plots a histogram of log hourly wage, `lnw`, for sessions featuring either unsafe and safe sex. Your plot should therefore have two histograms, potentially overlaying each other. Does there appear to be a difference in price between safe and unsafe sex?

solution

```
sasp %>%
    ggplot(aes(x = lnw, fill = factor(unsafe))) +
    # i plot proportions, you don't need to
    geom_histogram(aes(y = stat(count / sum(count))), alpha=0.6) +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    ylab("Fraction of Bookings") +
    xlab("Log Hourly Wage") +
    theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



4. Let's formalize this idea with a regression. Run a single variable regression of log hourly wage, `lnw` on the variable `unsafe`. Report the results.

solution

```
simple_reg <- lm(lnw ~ unsafe, data = sasp)

tidy(simple_reg, conf.int = TRUE)
```

```
## # A tibble: 2 x 7
##   term        estimate std.error statistic p.value conf.low conf.high
```

```
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 (Intercept)   5.63      0.0199    283.     0       5.60      5.67
## 2 unsafe       -0.0351    0.0273     -1.29   0.198  -0.0886    0.0184
```

5. Interpret the coefficient on `unsafe`. Is it statistically significant?

solution

Interpretation (1): On average, unsafe sex decreases the log hourly wage by 0.035

Interpretation (2): On average, unsafe sex decreases the the hourly wage by approximately 3.5 percent.

Interpretation 2 utilizes the log-level interpretation of the regression. Technically, the size of the effect is $(\exp \hat{\beta}_1 - 1) * 100$ percent, for small values of β_1, $\exp \hat{\beta}_1 - 1 \approx \hat{\beta}$.

Statistical Significance: the p-value is $0.197 > 0.05$, so the effect is not statistically significant at the 5 percent level of significance.

6. A single variable regression most likely suffers from omitted variable bias. Explain what omitted variable bias is, and why it might impact your regression estimates.

solution

Omitted Variable Bias: the effect of leaving out one or more relevant variables on the regression coefficients in the "misspecified" regression.

For omitted variable bias to occur we need:

1. The included X variable(s) to be correlated with the omitted variable
2. The omitted variable to be a relevant determinant of y

(1) and (2) leave to a violation of the exogeneity assumption $E(u_i|x_i) = 0$. When we don't have exogeneity,

$$E[\hat{\beta}] = \beta + \text{bias}$$

which means that our estimated coefficient cannot accurately estimate the true population parameter, and thus can't be interpreted causally.

7. Add the log of the length of the session, `llength`, as a second variable to your regression. Report the results. Did the coefficient on `unsafe` change?

solution

```
twovar_reg <- lm(lnw ~ unsafe + llength, data = sasp)

tidy(twovar_reg, conf.int = TRUE)
```

```
## # A tibble: 3 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)  6.74       0.0730    92.3   0          6.59      6.88
## 2 unsafe       0.00296    0.0254     0.117 9.07e- 1  -0.0469    0.0528
## 3 llength     -0.265      0.0169   -15.6   5.13e-51  -0.298    -0.231
```

7. Explain why ignoring `llength` in your regression led to the coefficient on `unsafe` to be different in sign

in the single variable regression than in the two variable regression.

solution

The formula for Omitted Variable Bias (assuming omitted variable, x_2 has coefficient β_2)

$$E[\hat{\beta}_1] = \beta_1 + \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{std dev}(x_2)}$$

One would reason that:

- $\beta_2 < 0$ ... longer sessions lead to quantity discounts
- $\text{Cov}(x_1, x_2) > 0$ ... longer sessions more likely to feature unsafe sex

$\implies$ bias is negative, so that

$$\begin{aligned} E[\hat{\beta}_1] &= \beta_1 + \text{something negative} \\ &< \beta_1 \end{aligned}$$

Remark: In case it was not clear from above, then:

- $\text{Cov}(x_1, x_2)$ is the covariance between x_1 and x_2
- std dev(x_2) is the standard deviation of x_2

8. Add a third variable to the regression, whether the client is a regular or not (`reg` in the data). Report your results and comment on any change in the regression estimate of `unsafe`.

solution

```
threevar_reg <- lm(lnw ~ unsafe + reg + llength, data = sasp)

tidy(threevar_reg, conf.int = TRUE)
```

```
## # A tibble: 4 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)  6.75       0.0731    92.3   0          6.61      6.89
## 2 unsafe       0.00418    0.0254     0.165 8.69e- 1  -0.0456    0.0539
## 3 reg         -0.0608     0.0256    -2.38  1.75e- 2  -0.111    -0.0107
## 4 llength     -0.260      0.0170   -15.3   4.62e-49  -0.293    -0.227
```

```
# I'll leave the comments out...
```

9. When discussing your interim results with a friend who is a bit of a statistical whiz they make the following remark: "I think you're not getting the expected results due to unobserved heterogeneity. Try adding fixed effects for each provider." What is unobserved heterogeneity? Why might it matter?

solution

Unobserved heterogeneity: unmeasured (typically time invariant) differences between (in this case) providers.

Think as follows: we have not included any variable about the provider so far - and there might be something about them that influences the prices they charge *and* their willingness to engage in unsafe sex.

Omitting unobserved heterogeneity - which in what follows is provider fixed effects - leads to omitted

variable bias. We leave it to you to think through the likely direction of that bias (do it! This kind of thinking is heavily valued in analytics work).

10. The data has a unique identifier for each provider in the `id` column. Use the `feols()` command from the `fixest` package to re-estimate your regression in (8) adding the provider ID fixed effects. Report your results with 'normal' standard errors (i.e. no clustering).

solution

```
fixedeff <- feols(lnw ~ unsafe + reg + llength
                  |
                      id,
                  data = sasp)

tidy(fixedeff, se = 'standard', conf.int = TRUE)
```

```
## # A tibble: 3 x 7
##   term    estimate std.error statistic   p.value conf.low conf.high
##   <chr>      <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 unsafe    0.0477    0.0286      1.67 9.57e-  2 -0.00836   0.104
## 2 reg      -0.0426    0.0176     -2.42 1.55e-  2 -0.0770   -0.00815
## 3 llength  -0.428     0.0138    -30.9  2.13e-149 -0.455    -0.401
```

11. Interpret your new results from (10). Is the coefficient on `unsafe` now statistically significant? Is the coefficient large from a 'marketing' viewpoint?

solution

On average, unsafe sex increases the hourly wage by approx 4.7 percent (holding other variables constant).

Our effect is statistically significant at the 10 percent level of significance (p value < 0.1) but not at the 5 percent level (p value > 0.05).

Is this big? It's approximately a 5 percent increase. The mean hourly wage is `r round(exp(mean(sasp$lnw)) -1,0)`, so a five percent increase is '$round(0.05 * (exp(mean(sasp$lnw)) -1),0)$' per hour. That really isn't *that* much of a premium.

Your next concern should be the standard errors - and whether we have 'correctly' adjusted for heteroskedasticity and/or clustering.

13. Produce a plot that visualizes the relationship between the predicted values of `lnw` from your regression on the horizontal axis and the residuals from the regression on the vertical axis.[3] Does there appear to be evidence of heteroskedasticity?

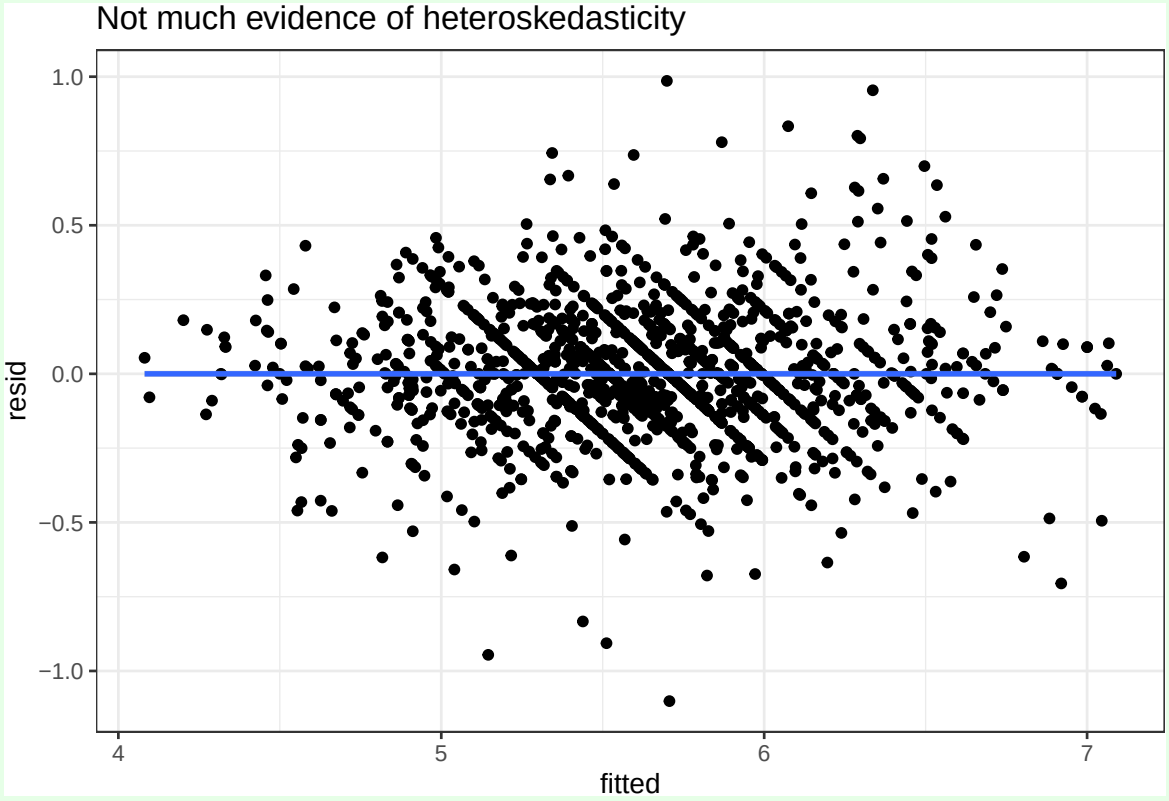
solution

[3]The function `predict(MODEL_NAME)` will create a column of predicted values from a regression stored as `MODEL_NAME`. The function `residuals(MODEL_NAME)` will create a column of residual values from a regression stored as `MODEL_NAME`.

```
sasp <- sasp %>%
  mutate(resid = residuals(fixedeff),
         fitted = predict(fixedeff))

sasp %>%
  ggplot(aes(x = fitted, y = resid)) +
  geom_point() +
  geom_smooth() +
  theme_bw() +
  ggtitle("Not much evidence of heteroskedasticity")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



14. Report regression results that use heteroskedasticity robust standard errors. You might be able to do this **without** re-estimating the regression model in (10). Does the standard error on `unsafe` change by much? Is this consistent with what you found graphically above?

solution

```r
# Doesn't change much...
tidy(fixedeff, se = 'hetero', conf.int = TRUE)
```

```
## Warning: In vcov.fixest(object, se = se, cluster = cluster, d...:
##  Asked for 4 threads while the maximum is 1. Set to 1 threads instead.

## Warning: In vcov.fixest(object, se = se, cluster = cluster, d...:
##  Asked for 4 threads while the maximum is 1. Set to 1 threads instead.

## # A tibble: 3 x 7
```

```
##   term    estimate std.error statistic   p.value conf.low conf.high
##   <chr>      <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 unsafe    0.0477    0.0261      1.83 6.83e-  2 -0.00352   0.0988
## 2 reg      -0.0426    0.0175     -2.44 1.50e-  2 -0.0769   -0.00833
## 3 llength  -0.428     0.0180    -23.8  6.57e-100 -0.463    -0.393
```

15. Report results that allow the standard errors to be clustered by `id` (i.e. clustered at the provider level). Again, you might be able to do this **without** re-estimating the regression model in (10). Why might you want to cluster the standard errors this way?

solution

```
tidy(fixedeff, se = 'cluster', conf.int = TRUE)
```

```
## Warning: In vcov.fixest(object, se = se, cluster = cluster, d...:
##  Asked for 4 threads while the maximum is 1. Set to 1 threads instead.

## Warning: In vcov.fixest(object, se = se, cluster = cluster, d...:
##  Asked for 4 threads while the maximum is 1. Set to 1 threads instead.
```

```
## # A tibble: 3 x 7
##   term    estimate std.error statistic  p.value conf.low conf.high
##   <chr>      <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 unsafe    0.0477    0.0257      1.86 6.39e- 2 -0.00263   0.0980
## 2 reg      -0.0426    0.0172     -2.48 1.34e- 2 -0.0762   -0.00895
## 3 llength  -0.428     0.0199    -21.5  1.20e-71 -0.467    -0.389
```

Marketers are generally interested in whether effects they find are heterogeneous, i.e. whether the reported coefficients vary across different observable characteristics.

16. Estimate a regression model that allows the price effect of unsafe sex to differ for customers who are regulars to those who aren't. Do this by modifying your regression command from (10). Report your results and discuss your findings.

solution

```
fixedeff_het <- feols(lnw ~ unsafe:reg + unsafe + reg + llength
                      |
                          id,
                      cluster = ~id,
                      data = sasp)
```

```
## Warning: In fixest_env(fml = fml, data = data, weights = weig...:
##  Asked for 4 threads while the maximum is 1. Set to 1 threads instead.
```

```
## Warning: In vcov.fixest(object, se = se, cluster = cluster, d...:
##  Asked for 4 threads while the maximum is 1. Set to 1 threads instead.
```

```
tidy(fixedeff_het, conf.int = TRUE)
```

```
## Warning: In vcov.fixest(object, se = se, cluster = cluster, d...:
##  Asked for 4 threads while the maximum is 1. Set to 1 threads instead.
```

```
## # A tibble: 4 x 7
##   term       estimate std.error statistic  p.value conf.low conf.high
##   <chr>         <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 unsafe       0.0218    0.0315     0.693 4.88e- 1  -0.0398    0.0835
```

```
## 2 reg          -0.0680    0.0249    -2.73  6.58e- 3  -0.117    -0.0192
## 3 llength      -0.428     0.0198   -21.6   4.77e-72  -0.467    -0.389
## 4 unsafe:reg    0.0460    0.0326     1.41  1.60e- 1  -0.0180    0.110
```

17. Interpret the results you found in (16).

solution

First, notice that neither of the terms `unsafe` or `unsafe:reg` are statistically significant - so we don't find overwhelming evidence for differences.

If we want to take the point estimates seriously (ignoring the std errors - purely for the sake of interpretation practice), we'd see that there seems to be evidence of price discrimination. Providers charge a higher price for unsafe sex with clients who are are regulars than those who aren't. A potential reason could be that regulars are less likely to switch to a different provider, so they're taken advantage of and charged a higher premium. I wouldn't want to push this argument too hard.

18. Are the effects you documented *causal*, *descriptive* or *predictive*? Explain your answer.

solution

For the heterogeneity results - definitely descriptive. There's a bunch of "selection on unobservables" issues and potentially omitted variables that would make me nervous about causal interpretation.

For the earlier regressions - The authors of the survey would probably argue towards causal interpretation after adding the fixed effects for the provider. Essentially they'd argue that the coefficient on unsafe is being estimated by differences in wages between unsafe and safe sex within each provider.

Now that you have run a series of regressions, you want to present the results in a way that you could use in a report or a presentation.

19. Take your regression estimates and produce a regression table to summarize four of them in one place. You can choose any of the estimates you like to produce the table, but we encourage you to think about how each column adds something to a story you could tell to explain your findings. The final result should look similar to a regression table you see in academic publications.

solution

```r
# a simple table - minimum customization
mods <- list(
  simple_reg,
  threevar_reg,
  fixedeff,
  fixedeff_het
)

msummary(mods,
         coef_omit = "Interc",
         gof_omit = "AIC|BIC|Log|Pseudo|F"
         )
```

```
## Warning: In vcov.fixest(object, se = se, cluster = cluster, d...:
##  Asked for 4 threads while the maximum is 1. Set to 1 threads instead.
```

	Model 1	Model 2	Model 3
unsafe	-0.035	0.004	0.048
	(0.027)	(0.025)	(0.026)
reg		-0.061	-0.043
		(0.026)	(0.017)
llength		-0.260	-0.428
		(0.017)	(0.020)
Num.Obs.	1499	1499	1499
R2	0.001	0.144	0.846
R2 Adj.	0.000	0.143	0.777
R2 Within			0.484
Std. Errors			Clustered (id)
FE: id			X

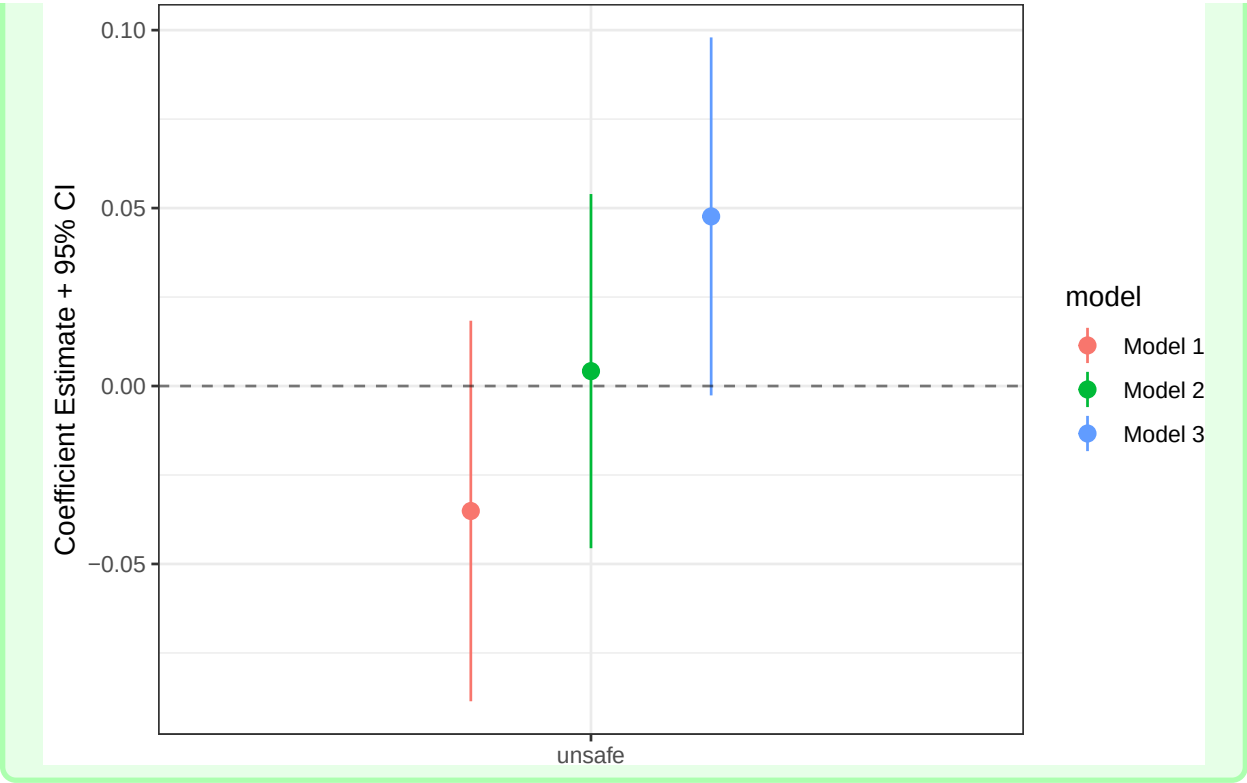
20. Take your regression estimates and produce a coefficient plot to summarize four of them in one place. You can choose any of the estimates you like to produce the plot, but we encourage you to think about the plot you produce can be used as part of a story you could tell to explain your findings.

solution

```
# heterog is more difficult to plot, so I am going to ignore it
mods <- list(
  simple_reg,
  threevar_reg,
  fixedeff
)


modelplot(mods,
          coef_omit = "Interc|reg|ll") +
  geom_vline(xintercept = 0,
             alpha = 0.5,
             linetype = "dashed") +
  xlab("Coefficient Estimate + 95% CI") +
  coord_flip() +
  theme_bw()
```

```
## Warning: In vcov.fixest(object, se = se, cluster = cluster, d...:
##  Asked for 4 threads while the maximum is 1. Set to 1 threads instead.
```



License

Suggested Citation

Deer, Lachlan and de With, Hendrik. 2021. Social Media and Web Analytics: Lab 2 - Multiple Regression in the Wild. Tilburg University. url = “https://github.com/tisem-digital-marketing/smwa-lab-02”