Case Study 1 - Legal Advertising - Does It Pay?

Fall 2020 - STAT 214 - Project 1

Nursima Donuk

10/11/2020

Summary: One partner (A) sued the other (B) over who should pay what share of the expenses of their former partnership. Partner A handled personal injury (PI) cases, while partner B handled only workers' compensation (WC) cases. The firm's advertising was focused only on personal injury, but partner A claimed that the ads resulted in the firm getting more workers' compensation cases for partner B, and therefore that partner B should share the advertising expenses.

Setting Up

```
# Set working directory
# setwd("Desktop")

# Load data
load("LEGALADV.Rdata")
head(LEGALADV)
```

```
##   MONTH   TOTADV NEWPI NEWWC ADVEXP6
## 1     1  9221.55     7    26      NA
## 2     2  6684.00     9    33      NA
## 3     3   200.00    12    18      NA
## 4     4 14546.75    27    15      NA
## 5     5  5170.14     9    19      NA
## 6     6  5810.30    13    26      NA
```

We see that our data has some missing values in the ADVEXP6 column.

Handling Missing Data

```
newdata <- LEGALADV[7:48,]

# See first few entries of data
head(newdata)
```

```
##    MONTH   TOTADV NEWPI NEWWC ADVEXP6
## 7      7  5816.20    11    24  41.633
## 8      8  8236.38     7    22  38.227
## 9      9 -2089.55    13    12  39.780
## 10    10 29282.24     7    15  37.490
## 11    11  9193.58     9    21  52.226
## 12    12  9499.18     8    24  56.249
```

Getting Familiar with the Data

```
str(newdata)
```

```
## 'data.frame':    42 obs. of  5 variables:
##  $ MONTH  : num  7 8 9 10 11 12 13 14 15 16 ...
##  $ TOTADV : num  5816 8236 -2090 29282 9194 ...
##  $ NEWPI  : num  11 7 13 7 9 8 18 9 25 26 ...
##  $ NEWWC  : num  24 22 12 15 21 24 25 19 12 33 ...
##  $ ADVEXP6: num  41.6 38.2 39.8 37.5 52.2 ...
```

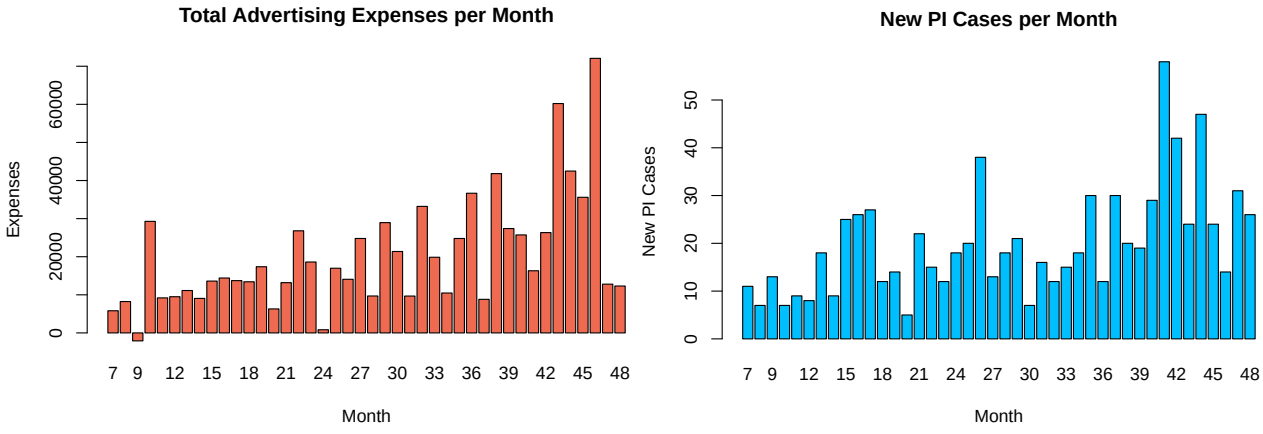
```
summary(newdata)
```

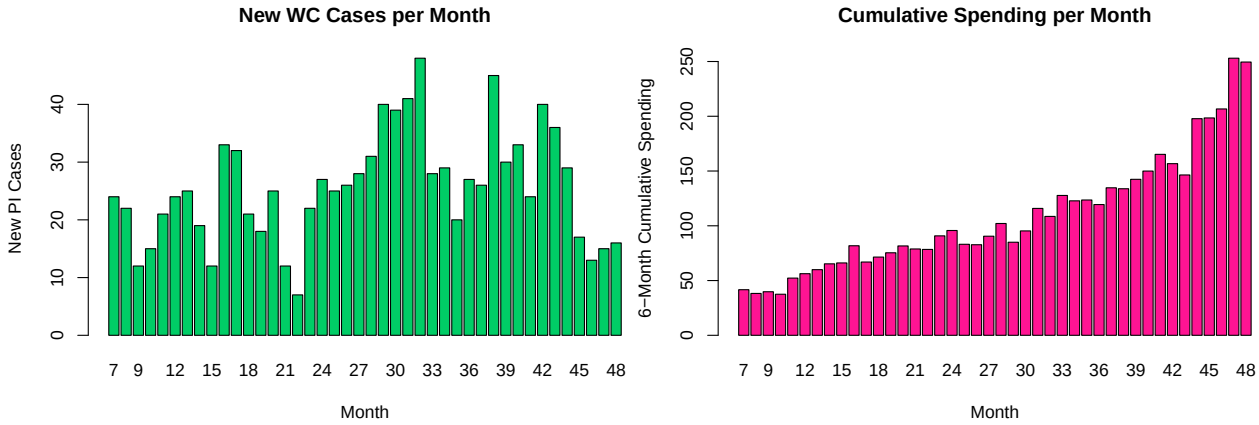
```
##      MONTH           TOTADV          NEWPI           NEWWC
##  Min.   : 7.00   Min.   :-2090   Min.   : 5.00   Min.   : 7.00
##  1st Qu.:17.25   1st Qu.: 9887   1st Qu.:12.00   1st Qu.:19.25
##  Median :27.50   Median :15362   Median :18.00   Median :25.00
##  Mean   :27.50   Mean   :20257   Mean   :20.05   Mean   :25.64
##  3rd Qu.:37.75   3rd Qu.:26682   3rd Qu.:25.75   3rd Qu.:30.75
##  Max.   :48.00   Max.   :72072   Max.   :58.00   Max.   :48.00
##     ADVEXP6
##  Min.   : 37.49
##  1st Qu.: 72.41
##  Median : 93.06
##  Mean   :108.78
##  3rd Qu.:134.46
##  Max.   :253.01
```

Packages

```
library(ggplot2)
library(tidyverse)
```

Bar Plots





The Models

Linear model for new personal injury cases vs cumulative 6-month advertising expenditures. Result: **NEWPI = 7.7675 + 0.1129(ADVEXP6)**

```
PI_lm <- lm(formula = NEWPI ~ ADVEXP6, data = newdata)
coefficients(PI_lm)
```

```
## (Intercept)     ADVEXP6
##   7.7674651   0.1128911
```

Linear model for new workers compensation cases vs cumulative 6-month advertising expenditures. Result: **NEWWC = 24.5741 + 0.0098(ADVEXP6)**

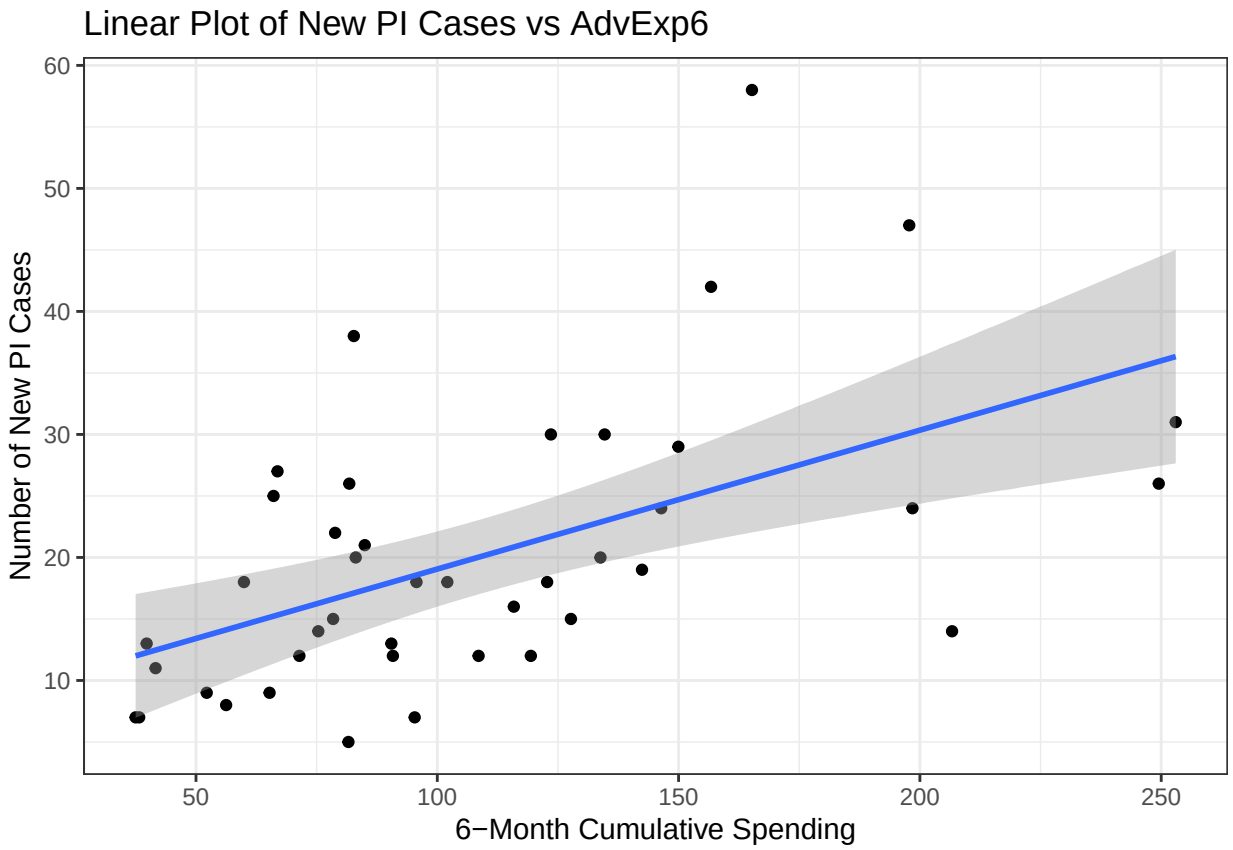
```
WC_lm <- lm(formula = NEWWC ~ ADVEXP6, data = newdata)
coefficients(WC_lm)
```

```
## (Intercept)     ADVEXP6
## 24.57412303  0.00982484
```

Descriptive Analysis: NEWPI vs ADVEXP6

Below is a scatter-plot for new personal injury cases vs. cumulative 6-month advertising expenditures.

```
ggplot(newdata, aes(ADVEXP6, NEWPI)) +
  geom_point() +
  geom_smooth(method = "lm") +
  coord_cartesian() +
  theme_bw() +
  ggtitle("Linear Plot of New PI Cases vs AdvExp6") +
  xlab("6-Month Cumulative Spending") +
  ylab("Number of New PI Cases")
```




```
cor(newdata$ADVEXP6, newdata$NEWPI, method = "pearson")
```

```
## [1] 0.538532
```

```
summary(PI_lm)
```

```
##
## Call:
## lm(formula = NEWPI ~ ADVEXP6, data = newdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.098  -5.846  -2.574   4.138  31.582
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.76747    3.38499   2.295 0.027078 *
## ADVEXP6      0.11289    0.02793   4.042 0.000234 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.675 on 40 degrees of freedom
## Multiple R-squared:   0.29,  Adjusted R-squared:  0.2723
## F-statistic: 16.34 on 1 and 40 DF,  p-value: 0.0002341
```

```
confint(PI_lm, level = 0.95)
```

```
##                   2.5 %    97.5 %
## (Intercept) 0.92613744 14.608793
```

```
## ADVEXP6     0.05644618  0.169336
```

Summary of Results:

Pearson Correlation: 0.538532, implies a moderate positive linear relation between cumulative 6-month advertising expenditures and new personal injury cases.

Coefficient of Determination: $r^2 = 0.29$, means that the sum of squares deviations of the y values (new PI cases) about their predicted values has been reduced 29% by using the least squares line equation, instead of $\bar{y}$, to predict y. A more practical interpretation can be, 29% of the sample variation in new PI cases can be explained by the linear relationship between cumulative 6-month advertising expenditures and new PI cases.

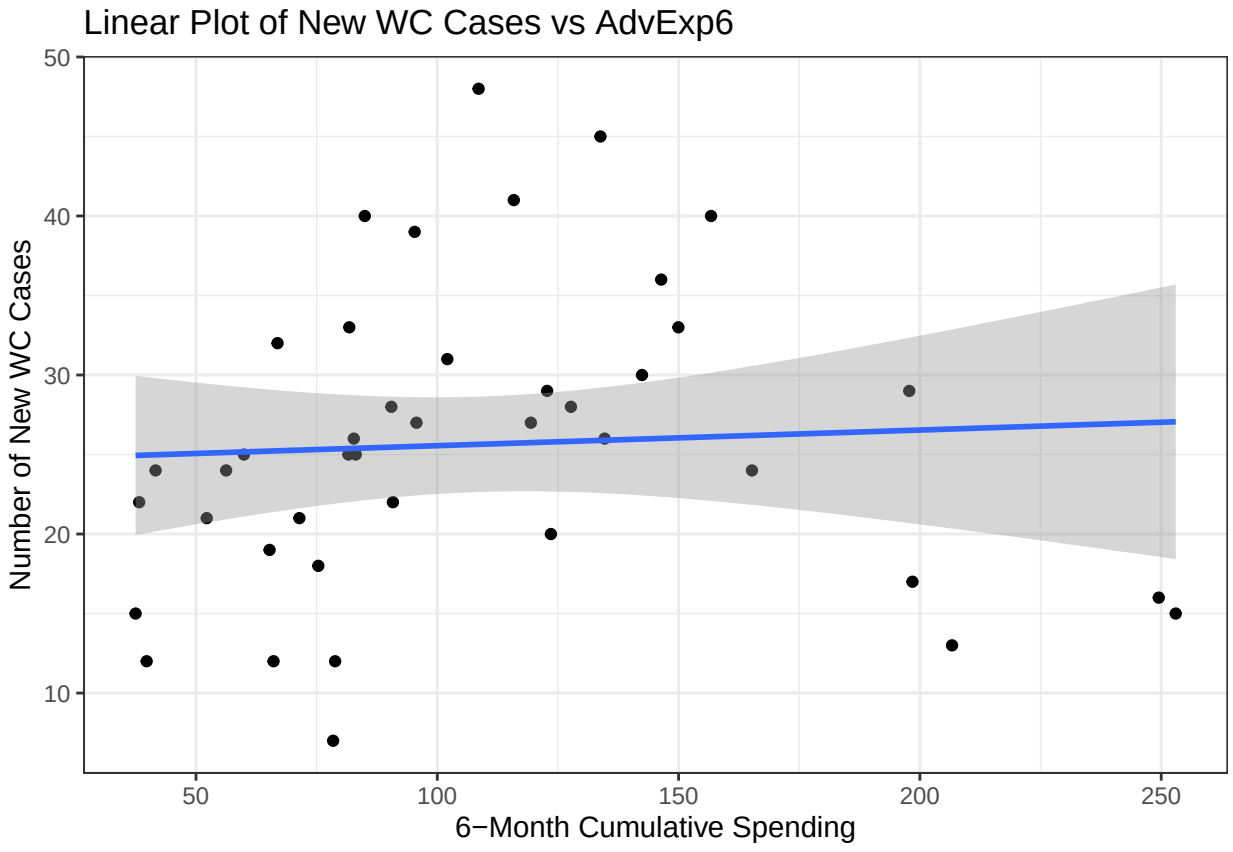
Estimated Slope: $\beta_1 = 0.11289$

95% Confidence Interval for Slope: [0.05644618, 0.169336], we can observe that this confidence interval for the slope of the regression line ranges from two positive numbers. Therefore does not include 0.

Descriptive Analysis: NEWWC vs ADVEXP6

Below is a scatter-plot for new workers compensation cases vs. cumulative 6-month advertising expenditures.

```
ggplot(newdata, aes(ADVEXP6, NEWWC)) +
  geom_point() +
  geom_smooth(method = "lm") +
  coord_cartesian() +
  theme_bw() +
  ggtitle("Linear Plot of New WC Cases vs AdvExp6") +
  xlab("6-Month Cumulative Spending") +
  ylab("Number of New WC Cases")
```



```r
cor(newdata$ADVEXP6, newdata$NEWWC, method = "pearson")
```

```
## [1] 0.05583758
```

```r
summary(WC_lm)
```

```
##
## Call:
## lm(formula = NEWWC ~ ADVEXP6, data = newdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3445  -6.1084  -0.2694   5.0738  22.3593
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.574123   3.366717   7.299 7.21e-09 ***
## ADVEXP6      0.009825   0.027777   0.354    0.725
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.623 on 40 degrees of freedom
## Multiple R-squared:  0.003118,   Adjusted R-squared:  -0.0218
## F-statistic: 0.1251 on 1 and 40 DF,  p-value: 0.7254
```

```r
confint(WC_lm, level = 0.95)
```

```
##                    2.5 %      97.5 %
## (Intercept) 17.76973470 31.37851136
## ADVEXP6     -0.04631528  0.06596496
```

Summary of Results:

Pearson Correlation: 0.05583758, implies that a linear relationship does not exist between cumulative 6-month advertising expenditures and new workers compensation cases.

Coefficient of Determination: $r^2 = 0.003118$, means that the sum of squares deviations of the y values (new WC cases) about their predicted values has been reduced 0.3% by using the least squares line equation, instead of $\bar{y}$, to predict y. A more practical interpretation can be, 0.3% of the sample variation in new WC cases can be explained by the linear relationship between cumulative 6-month advertising expenditures and new WC cases.

Estimated Slope: $\beta_1 = 0.009825$

95% Confidence Interval for Slope: [-0.04631528, 0.06596496], we can observe that this confidence interval for the slope of the regression line ranges from a negative number to a positive number, which means that the interval contains 0. Therefore, there is no statistical evidence of a linear relationship.

Testing the Models

To formally test the models, we conduct the hypothesis tests for the slopes of the regression lines:

```r
pval_PI <- summary(PI_lm)$coefficient[,"Pr(>|t|)"][2]
pval_PI
```

```
##      ADVEXP6
## 0.0002341227
```

```r
pval_WC <- summary(WC_lm)$coefficient[,"Pr(>|t|)"][2]
pval_WC
```

```
##   ADVEXP6
## 0.7254215
```

The two-tailed p-values for testing the null hypothesis, $H_0 : \beta_1 = 0$, are p-value = .0002 for number of new PI cases and p-value = .725 for number of new WC cases. For the number of new personal injury cases, there is sufficient evidence to reject H_0 (at $\alpha = .01$) and conclude that number of new PI cases is linearly related to cumulative 6-month advertising expenditures. In contrast, for the number of worker's compensation cases, there is insufficient evidence to reject H_0 (at $\alpha = .01$); thus, there is no evidence of a linear relationship between number of new WC cases and cumulative 6-month advertising expenditures.

Conclusion

This statistical analysis supports the conclusion that cumulative 6-month advertising expenditures is a statistically useful linear predictor of number of new personal injury cases, but not a useful linear predictor of number of new workers' compensation cases.

In court, a statistician presented the above results in support of the defendant (partner B). Clearly, the descriptive and inferential statistics provide support for the hypothesis that increased advertising expenditures are associated with more personal injury cases, but not with more workers' compensation cases. Ultimately, the court ruled that partner A (not partner B) should bear the brunt of the advertising expenditures.

Research Questions

Do these data provide support for the hypothesis that increased advertising expenditures are associated with more personal injury cases?

Yes, these data do not provide support for the hypothesis that increased advertising expenditures are associated with more personal injury cases. We can see from the analysis that there is evidence of a positive linear relationship between advertising expenditures and personal injury cases.

With more workers' compensation cases?

No, these data do not provide support for the hypothesis that increased advertising expenditures are associated with more workers' compensation cases. We can see from the analysis that there is no evidence of positive linear relationship between advertising expenditures and workers' compensation cases.

If advertising expenditures have a statistically significant association with the number of cases, does this necessarily mean that there is a causal relationship, that is, that spending more on advertising causes an increase in the number of cases?

No, correlation does not mean causation. Just because two variables are correlated does not mean one causes the other. To look for a causal relationship a thorough experiment has to be conducted.

Based on these data, should partner A or partner B bear the brunt of the advertising expenditures?

These data do not support partner A's claim that more advertising caused more workers' compensation cases. There seems to be a more linear and positive relationship between personal injury cases and advertising expenditures. That is the new personal injury cases seem to have increased as the advertising expenditures increased. This is why one may say that partner A should cover the advertising expenditures.

Follow up Questions

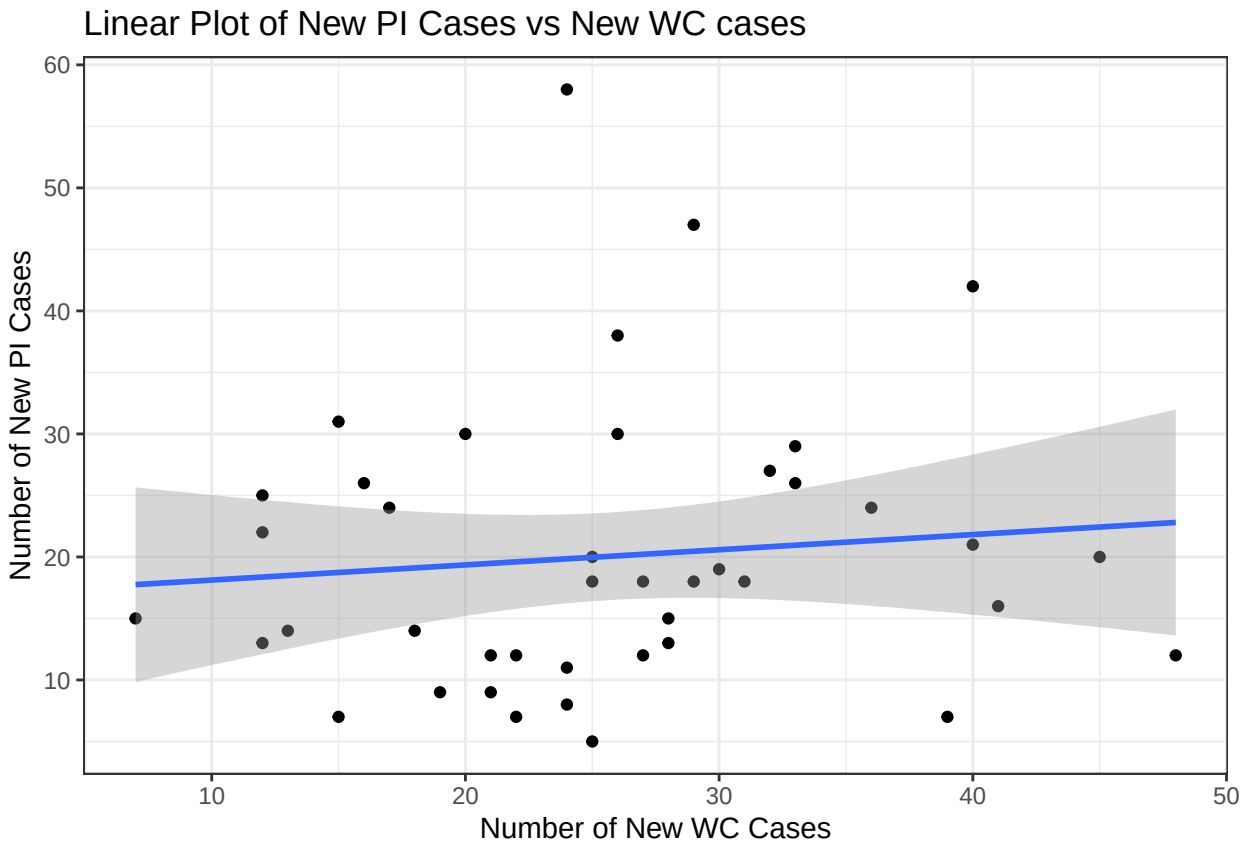
1) Access the data and find the correlation between number of new personal injury cases (y1) and number of new worker's compensation cases (y2).Which partner (A or B) would benefit

from reporting this correlation as evidence in the case? Explain.

```
cor(newdata$NEWWC, newdata$NEWPI, method = "pearson")
```

```
## [1] 0.1033978
```

```
ggplot(newdata, aes(NEWWC, NEWPI)) +
  geom_point() +
  geom_smooth(method = "lm") +
  coord_cartesian() +
  theme_bw() +
  ggtitle("Linear Plot of New PI Cases vs New WC cases") +
  xlab("Number of New WC Cases") +
  ylab("Number of New PI Cases")
```



We see from the correlation of 0.1033978 and observe from the scatter-plot that there is not enough evidence for a linear relationship for NEWWC and NEWPI. This can be beneficial for partner B, since we know that NEWPI has a moderate positive linear relationship with AdvExp6, knowing that there is no evidence supporting linear relationship between NEWPI and NEWWC supports the claim of no positive linear relationship between NEWWC and AdvExp6.

2)Compare the standard deviations for the simple linear regression models of number of new personal injury cases (y1) and number of new worker's compensation cases (y2). Which partner (A or B) would benefit from reporting only these standard deviations as evidence in the case? Explain.

The standard deviations for each simple linear regression model is: 9.675 for NEWPI and 9.623 for NEWWC. We see that the standard deviations for the models are close. This could benefit partner A because since we see a positive linear relationship between NEWPI and AdvExp6, the close correlation can support A's claim.

3)Access the data and find the standard deviation for the number of new personal injury cases (y1) and the standard deviation for the number of new worker's compensation cases (y2). Compare these standard deviations to those you found in question 2. Which partner (A or B) would benefit from reporting this additional information as evidence in the case? Explain.

```
sqrt(var(newdata$NEWWC))
```

```
## [1] 9.519739
```

```
sqrt(var(newdata$NEWPI))
```

```
## [1] 11.3416
```

We can see that NEWPI has a greater standard deviation, which is the result of greater variability. The simple linear regression model for NEWPI also had a greater standard deviation. This will benefit partner B since it shows greater variability in the new cases of personal injury, implying that the new personal injury cases increased at a greater amount as the cumulative advertising expenditures increased.

The End