# S&DS 220: Homework 11

Due Friday April 26

Braeden

## Instructions

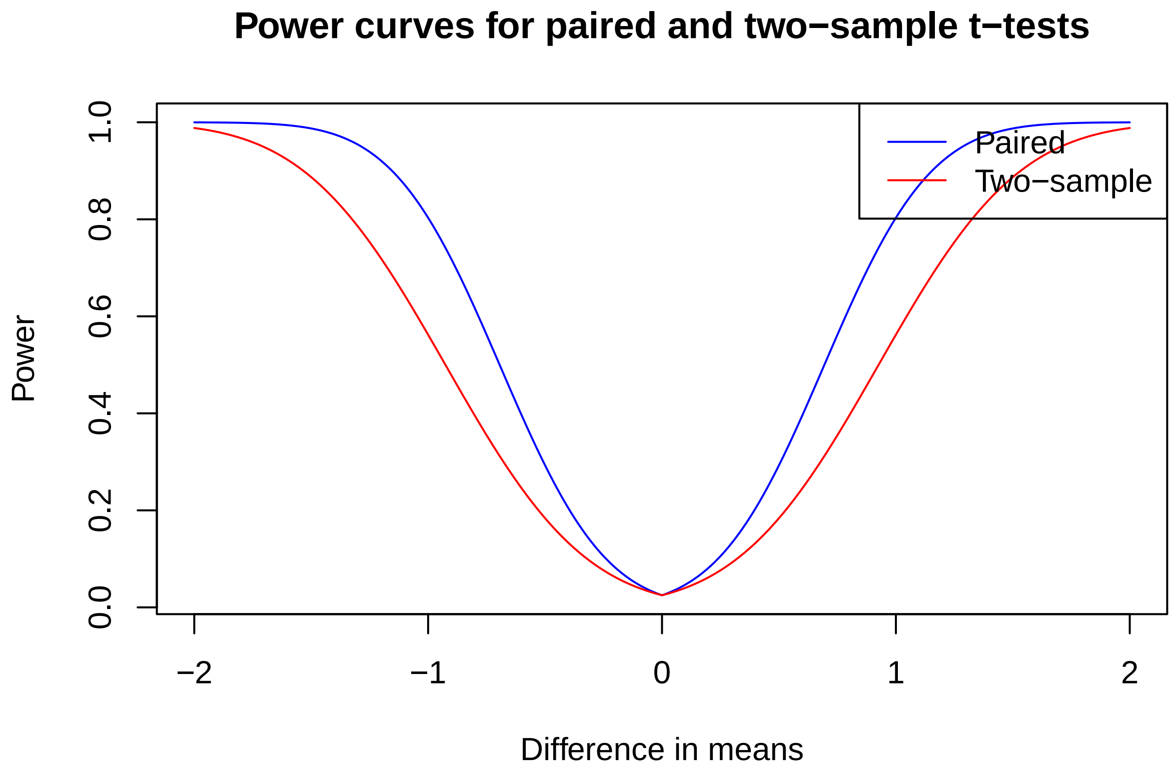
1. Complete the questions below. Upload your knitted PDF solutions to Gradescope by the due date.
2. Your solutions should be a combination of writing and R code. When writing, use complete sentences.
3. Previous homework assignments already had code chunks created for you. Now it is up to you to insert R code chunks within each problem as needed.
4. You should aim for clear and concise communication (in both words and R code).

## Problem set questions

### Question 1: Comparing powers for a paired sample

Consider a *paired* random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ where the $X_i \sim \mathcal{N}(\mu_x, \sigma_x)$ and $Y_i \sim \mathcal{N}(\mu_y, \sigma_y)$. We wish to test $H_0 : \mu_x = \mu_y$ versus the alternative $H_a : \mu_x \neq \mu_y$. For a given significance level $\alpha$, which test has more power: a two-sample $t$-test or a paired $t$-test? Experiment using `power.t.test`. Plot a power curve for each.

```
d <- seq(-2, 2, by = 0.01)

paired <- rep(NA_real_, length(d))
two_sample <- rep(NA_real_, length(d))

for(i in 1:length(d)) {
  paired[i] <- power.t.test(n = 10, delta = d[i], sd = 1, sig.level = 0.05, type = "paired")$power
  two_sample[i] <- power.t.test(n = 10, delta = d[i], sd = 1, sig.level = 0.05, type = "two.sample")$pow
}

plot(d, paired, type = "l", col = "blue", xlab = "Difference in means", ylab = "Power", main = "Power cu

lines(d, two_sample, col = "red")

legend("topright", legend = c("Paired", "Two-sample"), col = c("blue", "red"), lty = 1)
```


Power curves for paired and two−sample t−tests
Paired
Two−sample
Power
0.0
0.2
0.4
0.6
0.8
1.0
−2
−1
0
1
2
Difference in means

**Question 2: (11.5) Correlation**

For each of the following four plots, indicate whether the sample correlation coefficient is strongly positive (greater than 0.3), weak (between $-0.3$ and 0.3), or strongly negative (less than $-0.3$). (See the textbook for the plots).

*Solution.*

PLOT A: STRONGLY negative PLOT B: WEAK PLOT C: STRONGLY positive PLOT D: STRONGLY negative

**Question 3: (11.9) Slope-intercept form of regression line**

Suppose you have 100 data points, and $\overline{x} = 3$, $s_x = 1$, $\overline{y} = 2$, $s_y = 2$, and the correlation coefficient is $r = 0.7$. Find the equation of the least squares regression line in slope-intercept form.

*Solution.*

beta = r * (s_y / s_x) = 0.7 * (2 / 1) = 1.4 beta2 = $\overline{y}$ - beta * $\overline{x}$ = 2 - 1.4 * 3 = -2.2

Equation of the line = y = 1.4x - 2.2

**Question 4: (11.14) Residual plots**

For each of the following eight residual plots, indicate whether the residual plot is evidence against the linear model being satisfied or not. (See the textbook for the plots).

*Solution.*

PLOT 1: not satisfied, U-shape PLOT 2: not satisfied, evidenced by positive trend PLOT 3: satisfied PLOT 4: not satisfied, due to high leverage outliner PLOT 5: not statisfied, resituals skewed PLOT 6: satisfied PLOT 7: not satisfied, evidenced by negative trend PLOT 8: satisfied

**Question 5: (11.16) Regression and transformations**

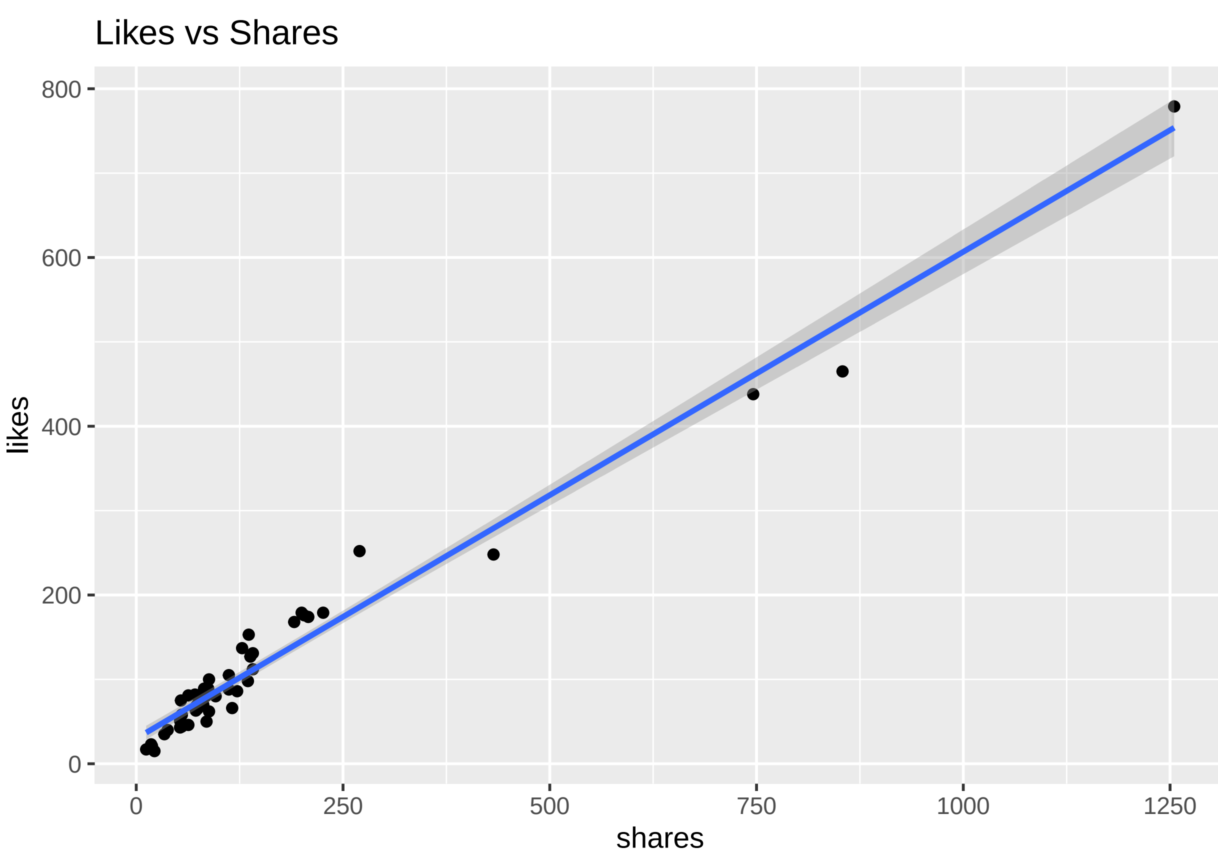
Consider the `cern` data set in the `fosdata` package. This data contains information on social media interactions of CERN. For the purposes of this problem, restrict to the `platform` Twitter.

(a) Create a linear model of `likes` on `shares`, and examine the residuals.

*Solution.*

```
data_twt <- filter(cern, platform == "Twitter")

model <- lm(likes ~ shares, data = data_twt)

ggplot(data_twt, aes(x = shares, y = likes)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Likes vs Shares")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

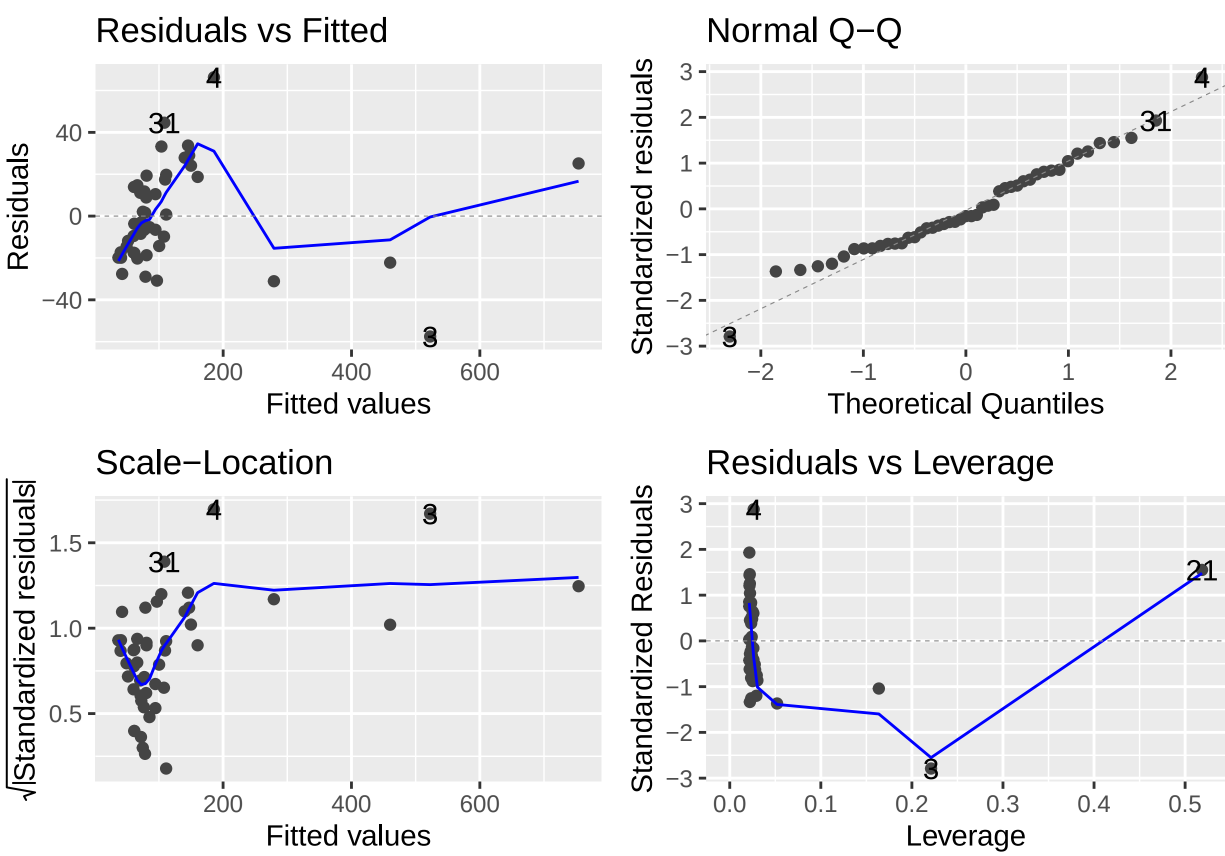



```
summary(model)
```

```
##
## Call:
## lm(formula = likes ~ shares, data = data_twt)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.536 -17.411  -3.659  16.096  66.325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.93437    4.17245   7.174 5.62e-09 ***
## shares       0.57682    0.01505  38.336  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.36 on 45 degrees of freedom
## Multiple R-squared:  0.9703, Adjusted R-squared:  0.9696
## F-statistic:  1470 on 1 and 45 DF,  p-value: < 2.2e-16
```

```r
autoplot(model)
```



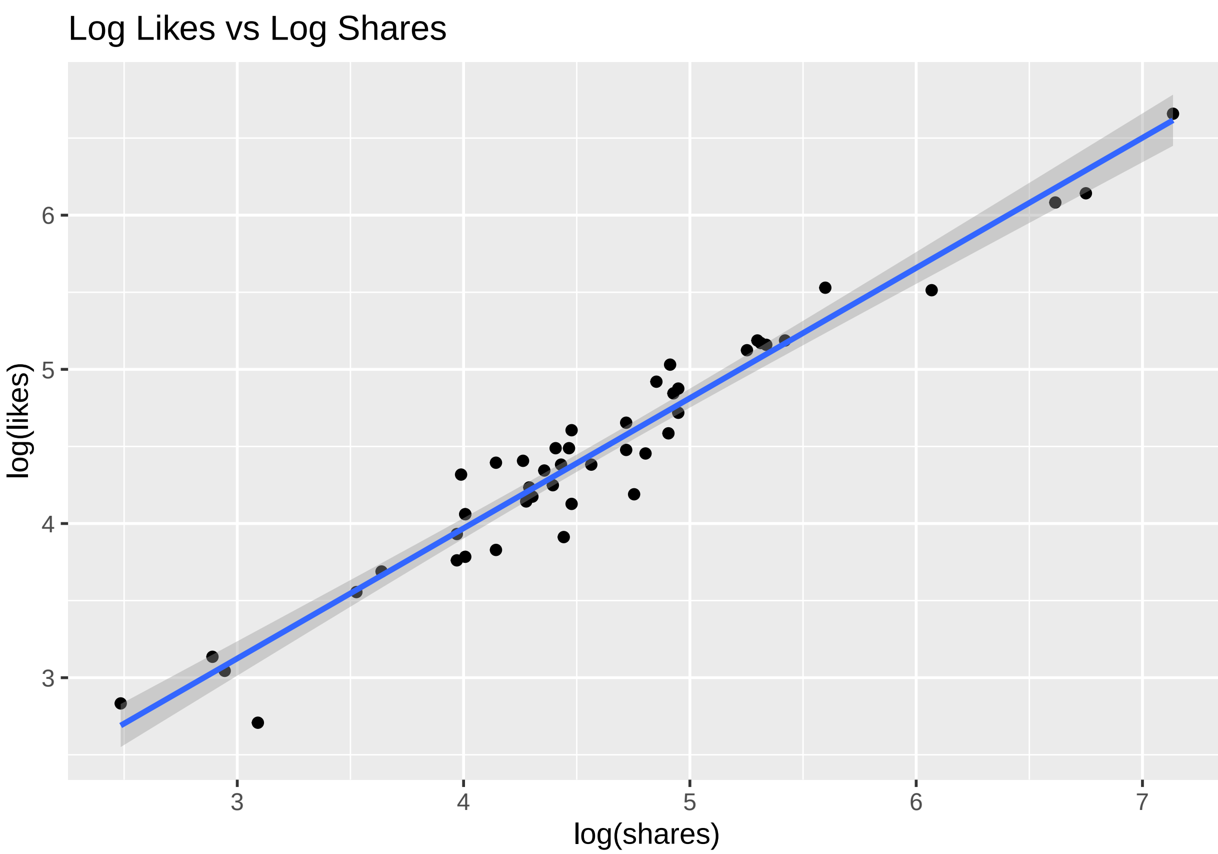

After examining the residuals of this model, we can see that there are outliers and a high leverage point w/ a large residual. Notice that the assumptions of linear regression are not satisfied.

(b) Create a linear model of `log(likes)` on `log(shares)` and examine the residuals.

*Solution.*

```r
model2 <- lm(log(likes) ~ log(shares), data = data_twt)

ggplot(data_twt, aes(x = log(shares), y = log(likes))) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Log Likes vs Log Shares")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

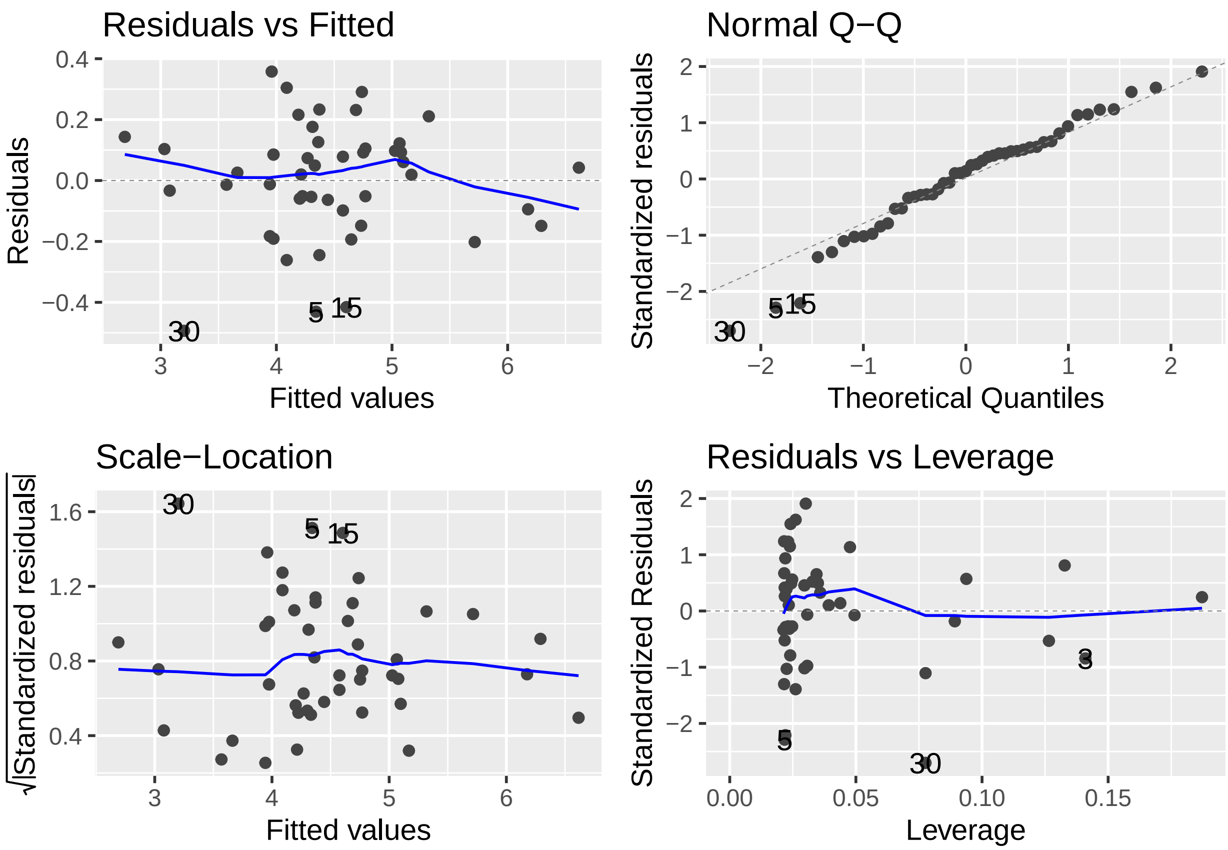



```r
summary(model2)
```

```
##
## Call:
## lm(formula = log(likes) ~ log(shares), data = data_twt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49350 -0.09637  0.02587  0.10426  0.35779
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.59173    0.14163   4.178 0.000134 ***
## log(shares)  0.84432    0.03033  27.840  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.1903 on 45 degrees of freedom
## Multiple R-squared:  0.9451, Adjusted R-squared:  0.9439
## F-statistic:   775 on 1 and 45 DF,  p-value: < 2.2e-16
```

```r
autoplot(model2)
```




The assumptions of linear regression are satisfied in this model. The plots look much better than the previous model.

(c) Which model seems to better match the assumptions of linear regression?

The model transformed by the natural logarithm seems to better match the assumptions of linear regression.

### Question 6: (11.21) Confidence intervals for regression coefficients

Consider the `cern` data set in the `fosdata` package. Create a linear model of `log(likes)` on `log(shares)` for interactions in the Twitter platform (see Question 4 (11.16)). Find 95 percent confidence intervals for the slope and intercept for the model if the residuals are acceptable.

*Solution.*

We can simply reuse the log-log model from the previous question. The residuals, from the previous question in parts b and c, look solid. 95% confidence intervals for the slope and intercept are:

```r
confint(model2)
```

```
##                 2.5 %    97.5 %
## (Intercept) 0.3064638 0.8769978
## log(shares) 0.7832336 0.9054010
```