Lab 09 - More inference

Due: Thursday, Apr 16 at 11:59pm EST

Team gitData - Blossom Mojekwu, Avanti Shah, Nathan Kim

Packages

```
library(tidyverse)
library(infer)
library(broom)
```

Data

```
no <- read_delim(file = "http://lib.stat.cmu.edu/datasets/NO2.dat",
                 delim = "\t", col_names = FALSE)
```

Setting seed

```
set.seed(8675309)
```

Exercises

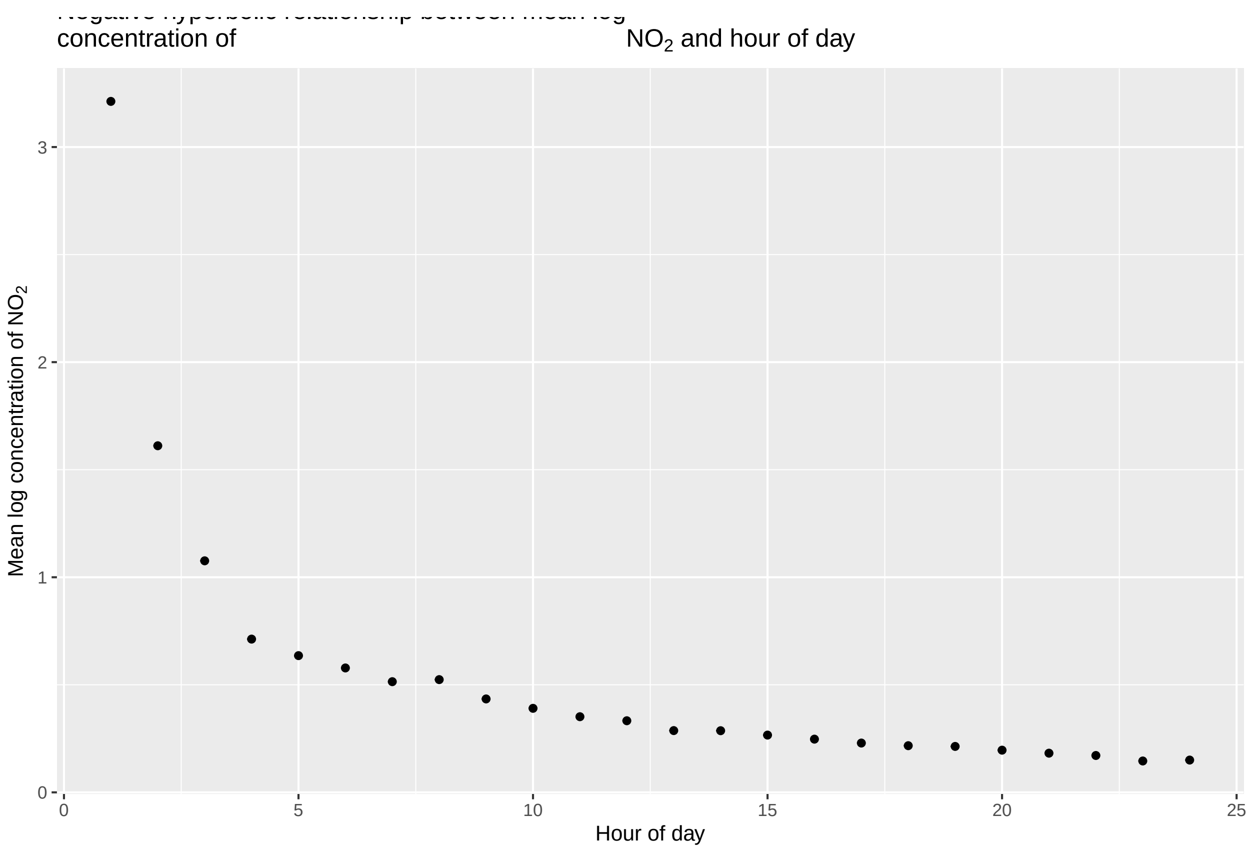
Exercise 1

```
no <- no %>%
  rename(log.NO2 = X1,
         log.cars_per_hr = X2,
         temp_in_C = X3,
         windspeed = X4,
         temp_diff = X5,
         wind_direction = X6,
         hour_of_day = X7,
         day_number = X8)
```

Exercise 2

```r
mean_no <- no %>%
  group_by(hour_of_day) %>%
  select(hour_of_day, log.NO2) %>%
  summarise(total.NO2 = sum(log.NO2),
            num = sum(hour_of_day),
            mean_NO2 = total.NO2 / num)

ggplot(data = mean_no, aes(x = hour_of_day, y = mean_NO2)) +
  geom_point() +
  labs(title = expression("Negative hyperbolic relationship between mean log
concentration of " * NO[2] * " and hour of day"),
       x = "Hour of day", y = expression("Mean log concentration of " * NO[2]))
```



Exercise 3

In the point plot, it is observed that the mean logarithm concentration of NO_2 is highest within hour 1 of the day and reduces to its lowest concentration at hour 24. Thus, as the day goes on and hours pass, the average logarithm concentration of NO_2 particles decreases.

Exercise 4

```r
ex_4_t_test <-
  infer::t_test(x = no, response = log.NO2, conf_int = TRUE, conf_level = 0.95)
```

```
ex_4_t_test %>%
  select(lower_ci, upper_ci)
```

```
# A tibble: 1 x 2
  lower_ci upper_ci
     <dbl>    <dbl>
1     3.63     3.76
```

The 95% confidence interval for the mean log concentration of NO2 is (3.63, 3.76). Thus, we are 95% confident that the interval (3.63, 3.76) captures the true mean log concentration on NO_2 in Oslo, Norway.

The assumptions we must make are that the original sample is representative of the population, and that the sample size is large enough to capture the population. Moreover, we must assume the distribution of the sample statistic is nearly normal, the distribution is centered at the unknown population parameter, and the variability of the distribution is inversely proportional to the square root of the sample size.

We expect that if we took many samples from the original population and built a 95% confidence interval based on each sample, about 95% of the intervals would contain the true population parameter.

Exercise 5

$H_0 : \mu >= 2000$

$H_1 : \mu < 2000$

where μ represents the true mean number of cars per hour

```
no <- no %>%
  mutate(num.cars_per_hr = exp(log.cars_per_hr))

ex_5_t_test <-
  infer::t_test(x = no, response = num.cars_per_hr, mu = 2000,
                alternative = "less")

ex_5_t_test %>%
  select(p_value)
```

```
# A tibble: 1 x 1
   p_value
     <dbl>
1 4.72e-13
```

At the $\alpha = 0.05$ level, the results are statistically significant and we reject the null hypothesis because the p-value (approaching 0.0) is less than α. It is very unlikely to observe our data or more extreme if the null hypothesis were actually true. Thus, there is insufficient evidence to suggest that the true mean number of cars per hour is greater than or equal to 2000.

Exercise 6

```
qt(p = 0.01, df = 31, lower.tail = FALSE)
```

```
[1] 2.452824
```

In this case, we will reject the null hypothesis if our observation is "extreme enough" to be less probable than 0.01 under the null hypothesis (which states that $\mu = \mu_0$), which corresponds to a t-statistic that is extreme enough. Here, we set `p = 0.01` since that is the probability whose corresponding t-statistic we are seeking.

We set `df = 31`, which is one less than the number of observations we have. We set `lower.tail = FALSE`, because the probabilities correspond to $\mu > \mu_0$.

Therefore, $\frac{\bar{x}-\mu_0}{s/\sqrt{n}} > 2.453824$, and `2.452825` is the smallest t-statistic we could obtain and still reject the null hypothesis.

Exercise 7

This claim is false. The p-value of a distribution represents the probability that, if the null hypothesis were in fact true, we would see our observed data or more extreme. In order for us to reject the null hypothesis, the p-value must be less than the α-value for significance level. If we have a p-value of less than 0.02, then at the $\alpha = 0.02$ significance level, there is less than a 2% chance that we see this data under the null hypothesis. However, if we have a p-value of less than 0.01, then at the $\alpha = 0.01$ significance level, there is less than a 1% chance that we see this data under the null hypothesis.

For example - let us assume that we perform bootstrapping with the statistic set to mean and the hypothesis set to a null point hypothesis. We then calculate the p-value using a two-sided test compared to the observed mean of our sample, and we find `p = 0.01568`. At the $\alpha = 0.02$ significance level, we can reject the null hypothesis, because our p-value is less than our predetermined significance level of 0.02, making our results very unlikely to have occurred if the null hypothesis were true. However, we **cannot** reject the null hypothesis at the $\alpha = 0.01$ significance level, because our p-value is greater than 0.01, hence our data is more likely than we want it to be to reject the null hypothesis. As a result, for a distribution with a p-value between 0.01 and 0.02, this statement does not hold true.

An alternative modified claim which is true is that if you **fail to reject** the null hypothesis at the $\alpha = 0.02$ significance level, then you will also **fail to reject** the null hypothesis at the $\alpha = 0.01$ significance level.

Exercise 8

```r
dist_ex_8 <- no %>%
  specify(response = log.NO2, explanatory = temp_in_C) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "slope")

dist_ex_8 %>%
  get_ci(0.90)
```

```
# A tibble: 1 x 2
     `5%`   `95%`
    <dbl>   <dbl>
1 -0.0271 -0.0120
```

```r
lm_ex_8 <-
  lm(log.NO2 ~ temp_in_C, data = no)

lm_ex_8 %>%
  tidy() %>%
  select(term, estimate)
```

```
# A tibble: 2 x 2
  term        estimate
  <chr>          <dbl>
1 (Intercept)   3.71
2 temp_in_C    -0.0193
```

```
tidy(lm_ex_8, conf.int = TRUE, conf.level = 0.90) %>%
  select(conf.low, conf.high) %>%
  slice(2)
```

```
# A tibble: 1 x 2
  conf.low conf.high
     <dbl>     <dbl>
1  -0.0277   -0.0110
```

Here, we can see that the simulation-based inference 90% confidence interval for the slope is (-0.02700996, -0.01172501), while the 90% confidence interval for the slope derived from the regression output is (-0.02771929, -0.01097091). The 5% and 95% values for these are very similar, however the interval for the regression output appears to be more left-skewed since each of the values is smaller than that of the simulation-based confidence interval. Finally, the simulation-based confidence interval is smaller (it has a width of 0.0153), while the regression output has a slightly larger confidence interval of width 0.0167.

Exercise 9

```
no <- no %>%
  arrange(day_number, hour_of_day)

full_lm <- lm(log.NO2 ~ log.cars_per_hr + temp_in_C + windspeed + temp_diff +
              wind_direction + hour_of_day, data = no)

full_lm_step <- step(full_lm, direction = "backward")
```

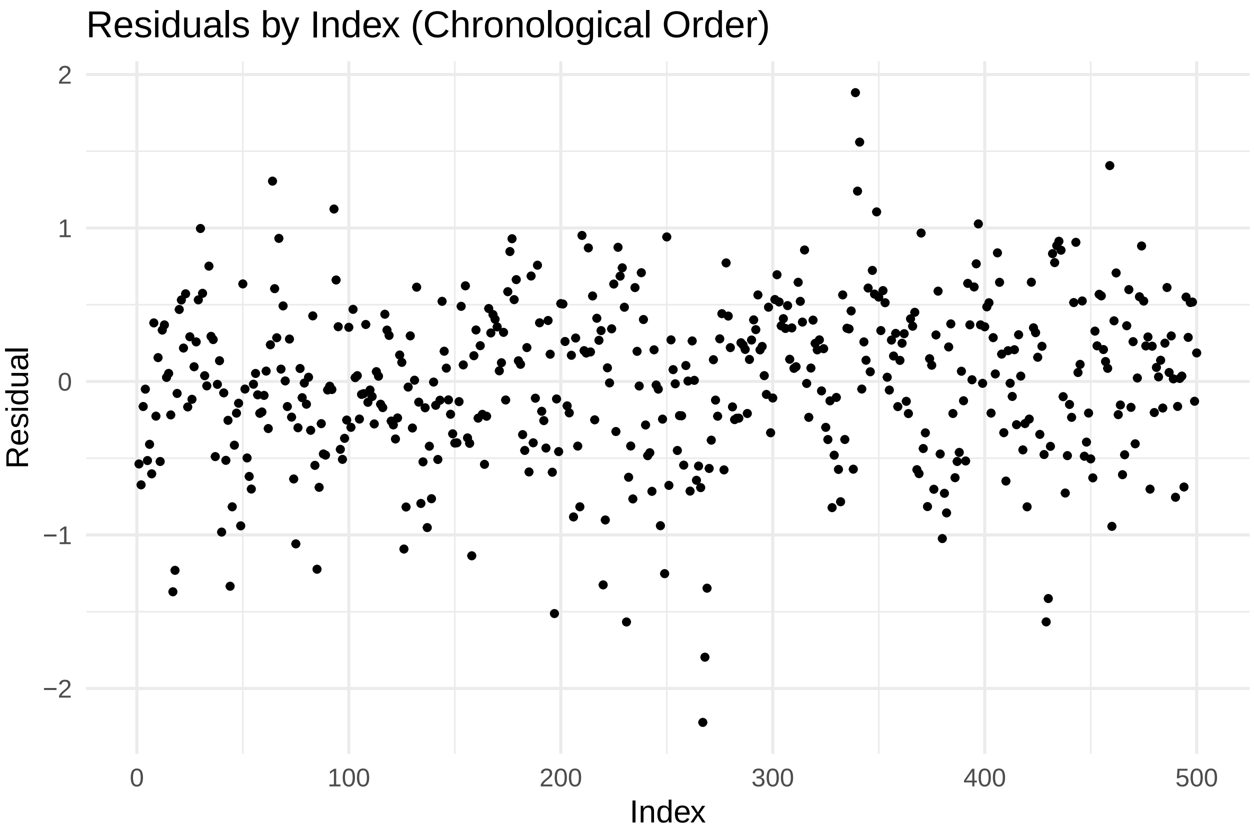
```
tidy(full_lm_step)
```

```
# A tibble: 7 x 5
  term             estimate std.error statistic  p.value
  <chr>               <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)      0.740     0.181         4.09 5.01e- 5
2 log.cars_per_hr  0.499     0.0285       17.5  6.99e-54
3 temp_in_C       -0.0220    0.00427      -5.16 3.66e- 7
4 windspeed       -0.129     0.0140       -9.18 1.15e-18
5 temp_diff        0.151     0.0258        5.84 9.60e- 9
6 wind_direction   0.000674  0.000298      2.26 2.43e- 2
7 hour_of_day     -0.0187    0.00439      -4.26 2.49e- 5
```

Exercise 10

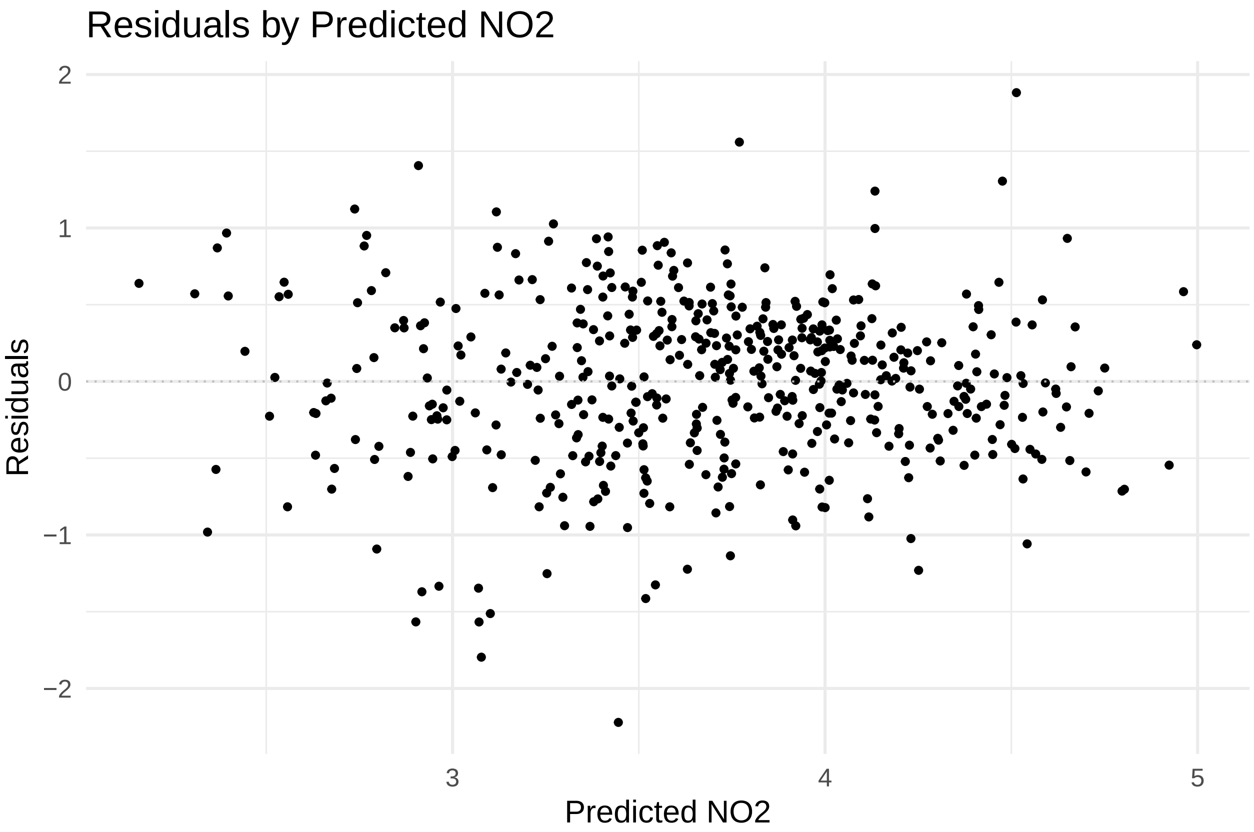
```
full_aug <- augment(full_lm_step)

ggplot(data = full_aug, aes(x = 1:nrow(full_aug), y = .resid)) +
  geom_point() +
  labs(title = "Residuals by Index (Chronological Order)",
       x = "Index", y = "Residual") +
  theme_minimal(base_size = 16)
```



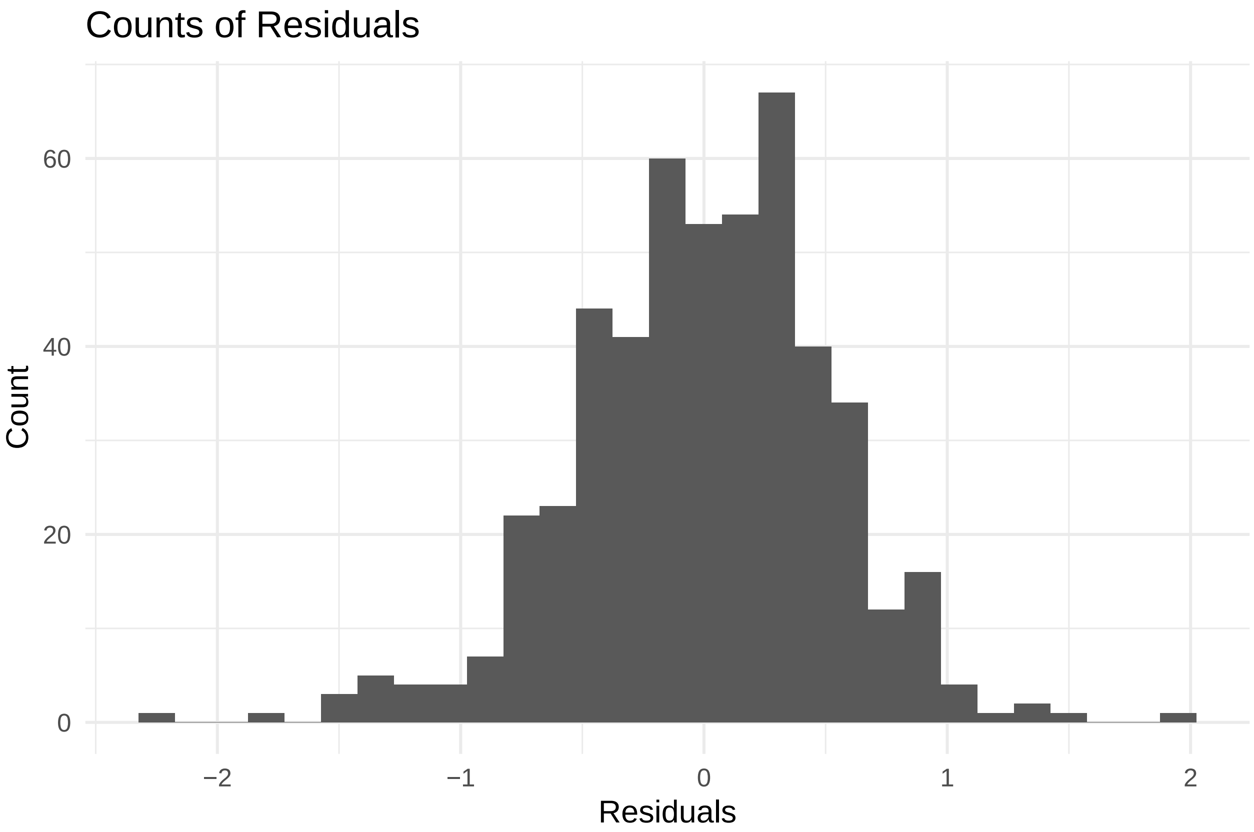
There doesn't appear to be any trend as we increase the index from left to right (in chronological order). Therefore, we can assume that these observations are independent.

```r
ggplot(data = full_aug, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, lty = 3, color = "gray") +
  labs(title = "Residuals by Predicted NO2",
       x = "Predicted NO2", y = "Residuals") +
  theme_minimal(base_size = 16)
```



While there does appear to be some sort of cluster, we can still see that the residuals are mostly randomly distributed around 0 and there is constant variance.

```r
ggplot(data = full_aug, aes(x = .resid)) +
  geom_histogram(binwidth = 0.15) +
  labs(title = "Counts of Residuals", x = "Residuals", y = "Count") +
  theme_minimal(base_size = 16)
```



While the direct mode is not directly over 0. this distribution still appears to be nearly normally ditributed with its center at 0. Therefore, we have satisfied all four conditions for the linear model, and we are thus allowed to use inference for regression.

Exercise 11

```
full_lm_step %>%
  tidy(conf.int = TRUE, conf.level = 0.95) %>%
  select(term, conf.low, conf.high) %>%
  filter(term == "windspeed" | term == "temp_in_C")
```

```
# A tibble: 2 x 3
  term       conf.low conf.high
  <chr>         <dbl>     <dbl>
1 temp_in_C   -0.0304   -0.0136
2 windspeed   -0.156    -0.101
```

We are 95% confident that the true change in the logarithm of the concentration of NO2 as wind speed increases by 1 m/s is between -0.15621575 and -0.10115354.

We are 95% confident that the true change in the logarithm of the concentration of NO2 as the temperature 2 meters above ground increases by 1 degree Celsius is between -0.03043471 and -0.01363948.