Classwork 11: Last Notes on Multiple Regression

Part 1: Exact Multicollinearity

In classwork 10, we modeled the price of houses using multiple regression. We saw that in the simple regression context, `bedrooms` seems to be a very important factor for determining the price of a house, but when we control for the other variables in the dataset, `bedrooms` seems to be much less important. The reason why the estimated effect of `bedrooms` on `price` changed so much is that `bedrooms` was highly correlated with many of the other explanatory variables.

What if two explanatory variables are perfectly correlated (correlation of 1 or -1)? As we'll see in this question, OLS will not be able to separately identify the effect of one of the variables versus the other. For example, consider a dataset with an explanatory variable `x` and another explanatory variable `z`, which is a linear function of `x`. If we try to estimate the model $y = \beta_0 + \beta_1 x + \beta_2 z + u$, because `x` and `z` are perfectly correlated, `lm` will not be able to decide whether to attribute explanatory power to `x` or `z`, and it will give us NA's:

```
library(tidyverse)

tibble(
  x = 1:10,
  z = 2 * x + 3,
  y = 1 + 2 * x + 3 * z + rnorm(n = 10)
  ) %>%
  lm(y ~ x + z, data = .) %>%
  broom::tidy()
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)     9.10    0.620       14.7  4.53e- 7
## 2 x               8.15    0.0998      81.6  5.68e-13
## 3 z              NA      NA           NA   NA
```

1) In workbook chapter 10, we learned that for a multiple regression with two explanatory variables, $\hat{\beta}_1 = \frac{(\sum_i x_i y_i) - \bar{x}\bar{y}n - \hat{\beta}_2(\sum_i x_i z_i) + \hat{\beta}_2 \bar{x}\bar{z}n}{(\sum_i x_i^2) - n\bar{x}^2}$. It's possible (with lots of algebra) to eliminate $\hat{\beta}_2$ to show that $\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y}) \sum_i (z_i - \bar{z})^2 - \sum_i (z_i - \bar{z})(y_i - \bar{y}) \sum_i (x_i - \bar{x})(z_i - \bar{z})}{\sum_i (x_i - \bar{x})^2 \sum_i (z_i - \bar{z})^2 - \left(\sum_i (x_i - \bar{x})(z_i - \bar{z})\right)^2}$: call this EQ1. In this problem, consider the case where $z_i = a x_i + b$ for constants a and b.

1a) Show that $z_i - \bar{z} = a(x_i - \bar{x})$.

1b) Show that $\sum_i (z_i - \bar{z})^2 = a^2 \sum_i (x_i - \bar{x})^2$.

1c) Show that $\sum_i (x_i - \bar{x})(z_i - \bar{z}) = a \sum_i (x_i - \bar{x})^2$.

1d) Show that $\sum_i (z_i - \bar{z})(y_i - \bar{y}) = a \sum_i (x_i - \bar{x})(y_i - \bar{y})$.

1e) Use the facts you proved in 1a-1d along with EQ1 to show that $\hat{\beta}_1 = \frac{0}{0}$ when one explanatory variable is a linear function of another explanatory variable.

Because of this, `lm()` considers it to be a mistake to include two variables that are perfectly correlated. It will eliminate any variable that is a linear function of other variables, and it will return `NA`'s for that variable's estimates. In R, an `NA` indicates missing data.

2) Prove that if z_i is a linear function of x_i (that is, $z_i = ax_i + b$ for constants a and b), then x and z will have an estimated correlation of 1 or -1. That is, they will be perfectly correlated (as long as x has variation and a is nonzero).

Hint: recall that the estimate for the correlation between two random variables is $\hat{\rho}_{xz} = \frac{\hat{\sigma}_{xz}}{\sqrt{\hat{\sigma}_x^2 \hat{\sigma}_z^2}}$, and recall that the estimate for the covariance between two random variables is: $\hat{\sigma}_{xz} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})$ and recall that the estimate for the variance of a random variable is: $\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Part 2: F-Tests for Joint Explanatory Power

You can use something called an F test to test the *joint* explanatory power of a multiple regression model. Here's how it works:

A variable has the F distribution if it's the ratio of some sums of squared normals, scaled by degrees of freedom. That is, $F = \frac{Q_1/d1}{Q_2/d2}$, where Q_1 and Q_2 are distributed chi-square (sums of squared normals).

For a regression, we have $F = \frac{\text{explained variation in y}/(k-1)}{\text{unexplained variation in y}/(n-k)} = \frac{\sum_i(\hat{y}_i - \bar{y})^2/(k-1)}{\sum_i(y_i - \hat{y}_i)^2/(n-k)}$. Notice the F statistic is the ratio of sums of squares, scaled by degrees of freedom.

The denominator degrees of freedom are just like the t test. They're $n-k$ where k is the number of parameters in the model. The idea is that you start with n degrees of freedom available and you have to use one up for each parameter (β) you estimate in your model. You can use the leftover $n-k$ to improve your confidence in your estimate (in the t test, the more degrees of freedom, the thinner the tails of the t-distribution).

The numerator degrees of freedom are $k-1$. They have to do with the null and alternative hypotheses of the F test for explanatory power:

H_0: $\beta_1 = \beta_2 = ... = \beta_{k-1} = 0$
H_A: at least one coefficient is nonzero.

The numerator of the F statistic has to do with comparing the "full" model with all k parameters to the "null" model with only the intercept. That is, if none of the explanatory variables have an effect on y, we're left with the null model of $y_i = \beta_0$. The F test is about testing whether the additional variance explained by the full model is significantly greater than what could be expected by random chance, so the numerator degrees of freedom are $k-1$: the extra pieces of information the full model provides compared to the null.

3) Show how the F statistic and its p-value was calculated in the example below (the last line of the summary).

This information might be helpful:

```
example_lm_object <- tibble(
  x = 1:5,
  y = c(3, 4, 6, 6, 7)
) %>%
  lm(y ~ x, data = .)

example_lm_object %>% summary()
```

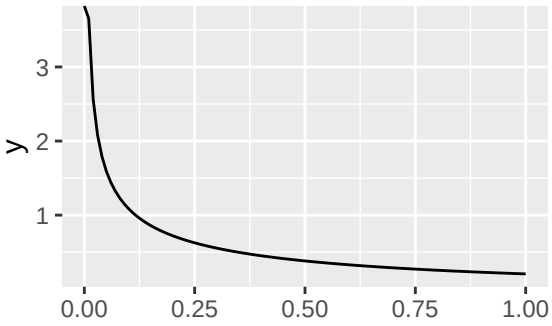
```
##
## Call:
## lm(formula = y ~ x, data = .)
##
## Residuals:
##    1    2    3    4    5
## -0.2 -0.2  0.8 -0.2 -0.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2000     0.5416   4.062  0.02690 *
## x             1.0000     0.1633   6.124  0.00875 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5164 on 3 degrees of freedom
## Multiple R-squared:  0.9259, Adjusted R-squared:  0.9012
## F-statistic:  37.5 on 1 and 3 DF,  p-value: 0.008754
```

```
example_lm_object %>% fitted.values()
```

```
##   1   2   3   4   5
## 3.2 4.2 5.2 6.2 7.2
```

```
# This is what the F distribution looks like with 1 numerator degree of freedom
# and and 3 denominator degrees of freedom:
```

```
ggplot() +
  geom_function(fun = ~ df(., df1 = 1, df2 = 3))
```




```
# Here are some potentially useful statistics about the F distribution:
qf(p = .95, df1 = 1, df2 = 3)
```

```
## [1] 10.12796
```

```
pf(q = 37.5, df1 = 1, df2 = 3)
```

```
## [1] 0.9912456
```

Note that the p-value for the F statistic is the same as the p-value for the t statistic for the hypothesis test for β_1. It should make sense that for the simple regression case, the F test of the explanatory power of the model collapses to being the same as the t test of the significance of the only (non-intercept) parameter of the model.