MA478 - Lesson3

Clark

For a linear model with full rank $\boldsymbol{X}$ and projection matrix $\boldsymbol{P_x}$, show that $\boldsymbol{P_x X} = \boldsymbol{X}$ and show that the column space of $\boldsymbol{X}$ is equal to the columns pace of $\boldsymbol{P_x}$.

Show that $\boldsymbol{P}y$ and $y - \boldsymbol{P}y$ are orthogonal.

Multivariate Normal Distribution Review:

Recall that our *residuals* capture the unexplained variation in our linear regression model and are calculated through $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$. Using the properties the distribution of $\boldsymbol{y}$, we know that $E[\boldsymbol{e}]$ is:

Geometrically, our residuals are orthogonal to the column space of $\boldsymbol{X}$. Further, the correlation between the residuals and $\hat{\boldsymbol{y}}$ is zero (which makes sense as $\hat{\boldsymbol{y}}$ is in the column space of X). This means, to examine the residuals, we should see no relationship between $\boldsymbol{e}$ and the columns of X or $\hat{\boldsymbol{y}}$.

We also typically assume $\boldsymbol{\epsilon} \sim MVN(\boldsymbol{0}, \sigma^2 I)$, which means we should have constant variance. This can be relaxed if we make additional choices for the random component of our GLM though. However, even if we don't relax this assumption, this stipulation does NOT mean that our residuals have constant variance.

While the covariance of $\boldsymbol{y}$ is clearly $\sigma^2 I$, this isn't true for covariance of $\hat{\boldsymbol{y}}$.

We can compute:

What this means, then, is that the covariance of our **residiual** is NOT $\sigma^2 I$. In fact, we can compute:

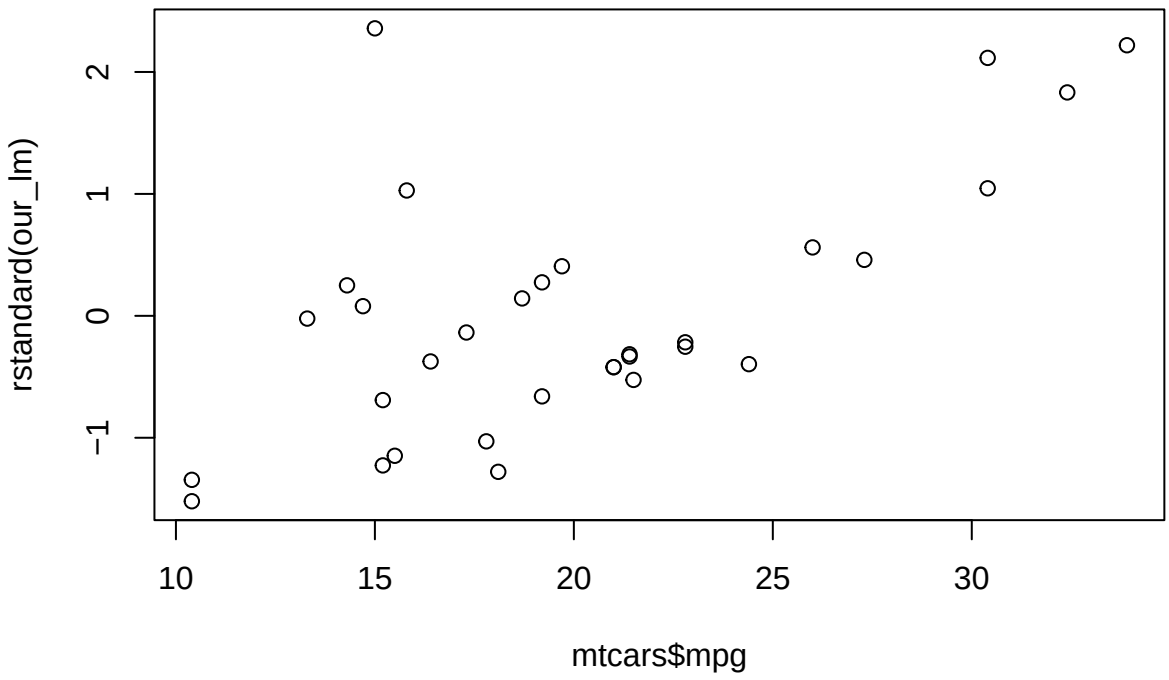
What this means is that our residuals are correlated and they don't necessarily have constant variance. Thus, to actually find residuals and examine their variance we should standardize them first:

So, what should we be doing?

- Plot standardized residuals against estimated Y
- Plot residuals against actual Y
- Plot fitted Y against actual Y

Look for patterns (there should be none)

```
library(faraway)
our_lm = lm(mpg~hp,data=mtcars)
plot(mtcars$mpg,rstandard(our_lm))
```



How do we fix this?

Another issue with our model that residuals can help us detect are outliers. Now, we typically think about outliers as being atypical observations, but do outliers always impact our results?

Outliers tend to be an issue if they have high *leverage* and they are *influential*. Recall that the $var\hat{y}_i = \sigma^2 h_{ii}$ where h_{ii} are the diagonal entires for the hat matrix. If h_{ii} larger (close to 1), then $var(e) \approx 0$. Meaning, y_i is highly correlated wtih $\hat{y}_i$. Therefore, the data point exactly determines where the regression like is. We say for these points the have high leverage.

However, must having high leverage isn't enough to cause concern. Consider the two examples:

If we look at the first, if the high leverage point is in our dataset the slope of the line is drastically different

than if the high leverage point is not in our dataset.

Conversely, in the second dataset, the presence (or absence) of our high leverage point does not influence the slope of our best fit line at all.

To quantify this, we compute the *Cook's distance* for each of our observations

```
our_lm = lm(mpg~hp,data=mtcars)
cooks.distance(our_lm)[which.max(cooks.distance(our_lm))]
```

```
## Maserati Bora
##      1.052231
```

What's a 'big' cooks distance? I dunno. Some people use 1 as a cut off. We can see why this point was chosen though by looking at a plot of the data.

Multicollinearity

I think the last thing I want to talk about with linear regression is multicollinarity. Colliearity means that there is a correlation between any two predictors. Multicollinearity is a relationship among several predictors. Why do we think this might be a problem?

Let's consider the following. Let's say I want to determine if weight impacts IOCT time. Consider the following two models:

```
APFT_dat = read.csv("APFT.csv")

IOCT_lm1 = lm(IOCT_Time~weight,data=APFT_dat)

IOCT_lm2 = lm(IOCT_Time~height+weight,data=APFT_dat)

summary(IOCT_lm1)
```

```
##
## Call:
## lm(formula = IOCT_Time ~ weight, data = APFT_dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -80.66 -32.12 -14.56  16.57 424.59
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 300.8224    17.6099  17.083  < 2e-16 ***
## weight       -0.5739     0.1033  -5.555 5.22e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.77 on 382 degrees of freedom
## Multiple R-squared:  0.07474,    Adjusted R-squared:  0.07231
## F-statistic: 30.86 on 1 and 382 DF,  p-value: 5.218e-08
```

```r
summary(IOCT_lm2)
```

```
##
## Call:
## lm(formula = IOCT_Time ~ height + weight, data = APFT_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -104.80  -28.32  -11.57   15.35  383.62
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 730.4184    55.5637  13.146  < 2e-16 ***
## height       -8.1424     1.0069  -8.087 8.24e-15 ***
## weight        0.2147     0.1365   1.573    0.117
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.82 on 381 degrees of freedom
## Multiple R-squared:  0.2103, Adjusted R-squared:  0.2061
## F-statistic: 50.73 on 2 and 381 DF,  p-value: < 2.2e-16
```

What happened?

What can we do about it?

Let me summarize this a bit. When we are fitting a linear regression model, prior to looking at any p value or conducting any statistical test, we want to make sure the model is appropriate for the data. If it's a simple linear regression model, or there are no interaction terms, we can just plot our X columns vs Y. If there's curvature between X and Y then we likely have an interaction or we need to transform our X column in some way (If it's appropriate for the problem).

I always check the standardized residuals and plot them against both y as well as $\hat{y}$. I don't usually bother checking against X. I typically don't worry too much about the normality of the residuals, but I do like to check for constant variance as well as checking to ensure there's no curvature. If we do not have constant variance I might consider fitting a different model such as an additive error model. I don't like transforming y unless I absolutely have to.

I sometimes check Cook's distance, especially if I potentially have outliers. Sometimes I forget to do this.

I then, if appropriate, conduct my statistical analysis. If I'm comparing two models I'll conduct an F-test (likelihood ratio test), if my question is about the impact of a covariate I'll conduct the associated t-test.

When presenting my final model, I want to ensure I address the statistical question I started with. If my model violates some of the validity conditions, I typically address it, but unless it's a really bad violation, I

usually am not too concerned. If we were conducting an analysis for a client, what types of things would we want to ensure we contain in our report? What is probably not necessary?

Homework

Quiz