Case Study 1

Jonathan Sneh, Ishani Tarafdar, Georges Durand, Raul Higareda

2023-03-28

Data Explorations and Summary Statistics

```
grades <- read.csv("grades.csv", header=TRUE)
dim(grades)
```

```
## [1] 275   7
```

Model Selection

We want to make a model with 95% confidence (i.e. $\alpha = 0.05$)

```
grades.mlr <- lm(exam2 ~ . ,data=grades)
summary(grades.mlr)
```

```
##
## Call:
## lm(formula = exam2 ~ ., data = grades)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.414  -6.793   0.850   7.831  27.124
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   35.26655    5.46283   6.456 5.03e-10 ***
## exam1          0.34756    0.05853   5.939 8.88e-09 ***
## project        0.01576    0.03971   0.397    0.692
## cs             0.02337    0.05811   0.402    0.688
## hw             0.40359    0.06267   6.440 5.49e-10 ***
## participation -0.02741    0.05210  -0.526    0.599
## semester      -7.51155    0.66993 -11.212  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.06 on 268 degrees of freedom
## Multiple R-squared:  0.5405, Adjusted R-squared:  0.5302
## F-statistic: 52.54 on 6 and 268 DF,  p-value: < 2.2e-16
```

Based on the full model summary above, we may want to look into dropping project, cs, and participation from the dataset since the t-values in the summary output for project, cs, and participation are all small, meaning that they're likely up to chance.

However, the individual t-tests do not tell us enough information to drop multiple predictors from our model at a time.

So, we can start by dropping an individual predictor from our model. We will check `project`.

Our null and alternative hypothesis are as follows.

$$\begin{cases} H_0, & \beta_{project} = 0 \\ H_A, & \beta_{project} \neq 0 \end{cases}$$

By conducting an individual t-test (which can be found in our summary output), we can see that the t-test statistics for project is 0.397. ## TODO PLEASE CHECK IF WE NEED TO BE THOROUGH AND MENTION WHICH DISTRIBUTION THESE ARE FROM This comes from a . Thus, we can see that the p-value for project is 0.692. $p = 0.692 > 0.05 = \alpha$. Thus, we fail to reject the null hypothesis, meaning that it is likely that $\beta_{project} = 0$. In other words, we can drop project from our model.

This leaves us with the following reduced model:

```r
grades.reducedmlr1 = lm(exam2~exam1 + semester + hw + cs + participation, data=grades)
summary(grades.reducedmlr1)
```

```
##
## Call:
## lm(formula = exam2 ~ exam1 + semester + hw + cs + participation,
##     data = grades)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.434  -6.864   0.784   7.872  26.832
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.99965    5.41281   6.466 4.72e-10 ***
## exam1          0.34762    0.05843   5.949 8.37e-09 ***
## semester      -7.50742    0.66880 -11.225  < 2e-16 ***
## hw             0.41028    0.06027   6.808 6.44e-11 ***
## cs             0.03178    0.05403   0.588    0.557
## participation -0.02404    0.05132  -0.468    0.640
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.04 on 269 degrees of freedom
## Multiple R-squared:  0.5402, Adjusted R-squared:  0.5317
## F-statistic: 63.21 on 5 and 269 DF,  p-value: < 2.2e-16
```

```r
anova(grades.mlr, grades.reducedmlr1)
```

```
## Analysis of Variance Table
##
## Model 1: exam2 ~ exam1 + project + cs + hw + participation + semester
## Model 2: exam2 ~ exam1 + semester + hw + cs + participation
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    268 38963
## 2    269 38986 -1   -22.912 0.1576 0.6917
```

Can we reference the overall summary and p-values to give us guidance on where to look next?

From the summary, We can see that the p-values for cs and participation have changed. They are still high, so we can conduct a different test.

$$\begin{cases} H_0, & \beta_{participation} = \beta_{cs} = 0 \\ H_A, & \text{Either } \beta_{participation} \text{ or } \beta_{cs} \text{ is not equal to zero} \end{cases}$$

```
library(ellipse)
```

```
##
## Attaching package: 'ellipse'

## The following object is masked from 'package:graphics':
##
##     pairs
```

```
library(ggplot2)
```

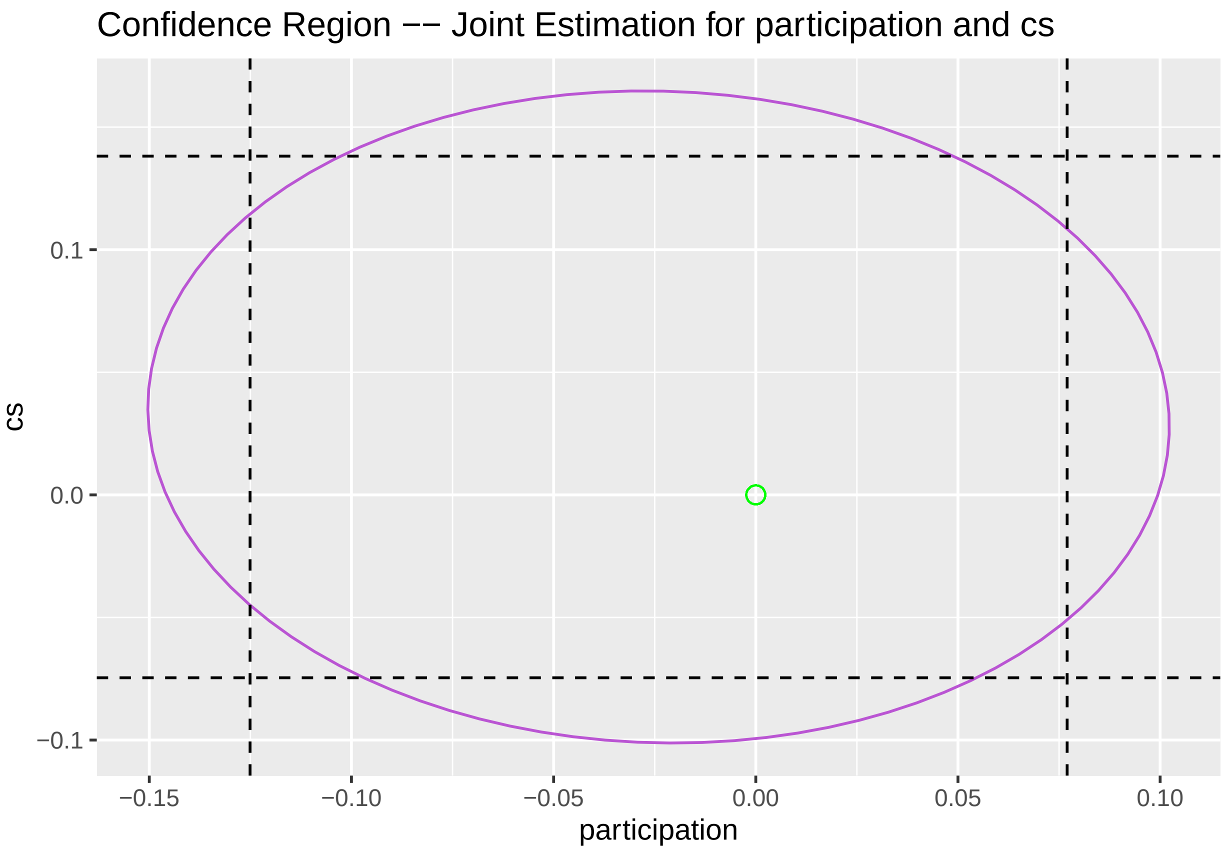
We can draw the confidence region (as an ellipse) for both participation and for cs. If the point (0,0) falls inside of our confidence region, then it is likely that both the coefficients β_{cs} and $\beta_{participation}$ are zero—as according to the null hypothesis.

```
intervals <-confint(grades.reducedmlr1)
cr_ellipse <- ellipse(grades.reducedmlr1, c(5,6), level=0.95)

par_interval <- confint(grades.reducedmlr1, level = 0.95, 'participation')
cs_interval <- confint(grades.reducedmlr1, level = 0.95, 'cs')

cr_df <- as.data.frame(cr_ellipse)
cr_plot <-
ggplot(data=cr_df, aes(x=participation, y=cs)) +
  ggtitle("Confidence Region -- Joint Estimation for participation and cs") +
  geom_path(aes(x=participation,y=cs), colour='mediumorchid') +
  geom_point(x=coef(grades.reducedmlr1)[2], y=coef(grades.reducedmlr1)[3],
             shape=3, size=3, colour='mediumorchid') +
  geom_hline(yintercept = cs_interval[1], lty=2) +
  geom_hline(yintercept = cs_interval[2], lty=2) +
  geom_vline(xintercept = par_interval[1], lty=2) +
  geom_vline(xintercept = par_interval[2], lty=2)+
  geom_point(x=0, y=0, shape=1, size=3, colour='green')

plot(cr_plot)
```



As we can see, the origin—which is the green dot—falls inside the confidence region. Thus, it is likely enough that both β_{cs} and $\beta_{participation}$ are zero. Therefore, we can drop them both from our model.

Our null and alternative hypothesis are as follows:

$$\begin{cases} H_0, & \beta_{project} = \beta_{cs} = \beta_{participation} \\ H_A, & \text{Either } \beta_{project}, \beta_{cs}, \text{ or } \beta_{participation} \text{ is not equal to zero} \end{cases}$$

```
grades.reducedmlr1 = lm(exam2~exam1 + semester + hw + cs + participation, data=grades)
summary(grades.reducedmlr1)
```

```
##
## Call:
## lm(formula = exam2 ~ exam1 + semester + hw + cs + participation,
##     data = grades)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.434  -6.864   0.784   7.872  26.832
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.99965    5.41281   6.466 4.72e-10 ***
## exam1          0.34762    0.05843   5.949 8.37e-09 ***
## semester      -7.50742    0.66880 -11.225  < 2e-16 ***
## hw             0.41028    0.06027   6.808 6.44e-11 ***
## cs             0.03178    0.05403   0.588    0.557
## participation -0.02404    0.05132  -0.468    0.640
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.04 on 269 degrees of freedom
## Multiple R-squared:  0.5402, Adjusted R-squared:  0.5317
## F-statistic: 63.21 on 5 and 269 DF,  p-value: < 2.2e-16
```

```
anova(grades.mlr, grades.reducedmlr1)
```

```
## Analysis of Variance Table
##
## Model 1: exam2 ~ exam1 + project + cs + hw + participation + semester
## Model 2: exam2 ~ exam1 + semester + hw + cs + participation
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    268 38963
## 2    269 38986 -1   -22.912 0.1576 0.6917
```

Since the $p = 0.8717 > \alpha$ (p-value is greater than an alpha level of 0.05) in the anova ouput, we fail to reject the null hypothesis with 95% level of confidence.

So, our final model (before diagonistcs) is blah

```
grades.reducedmlr = lm(exam2 ~ exam1 + semester + hw, data=grades)
summary(grades.reducedmlr)
```

```
##
## Call:
## lm(formula = exam2 ~ exam1 + semester + hw, data = grades)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.029  -7.070   0.881   7.921  28.145
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.02239    4.73371   7.610 4.56e-13 ***
## exam1        0.34759    0.05717   6.080 4.08e-09 ***
## semester    -7.54042    0.66197 -11.391  < 2e-16 ***
## hw           0.40989    0.04775   8.584 7.28e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.01 on 271 degrees of freedom
## Multiple R-squared:  0.5393, Adjusted R-squared:  0.5342
## F-statistic: 105.7 on 3 and 271 DF,  p-value: < 2.2e-16
```

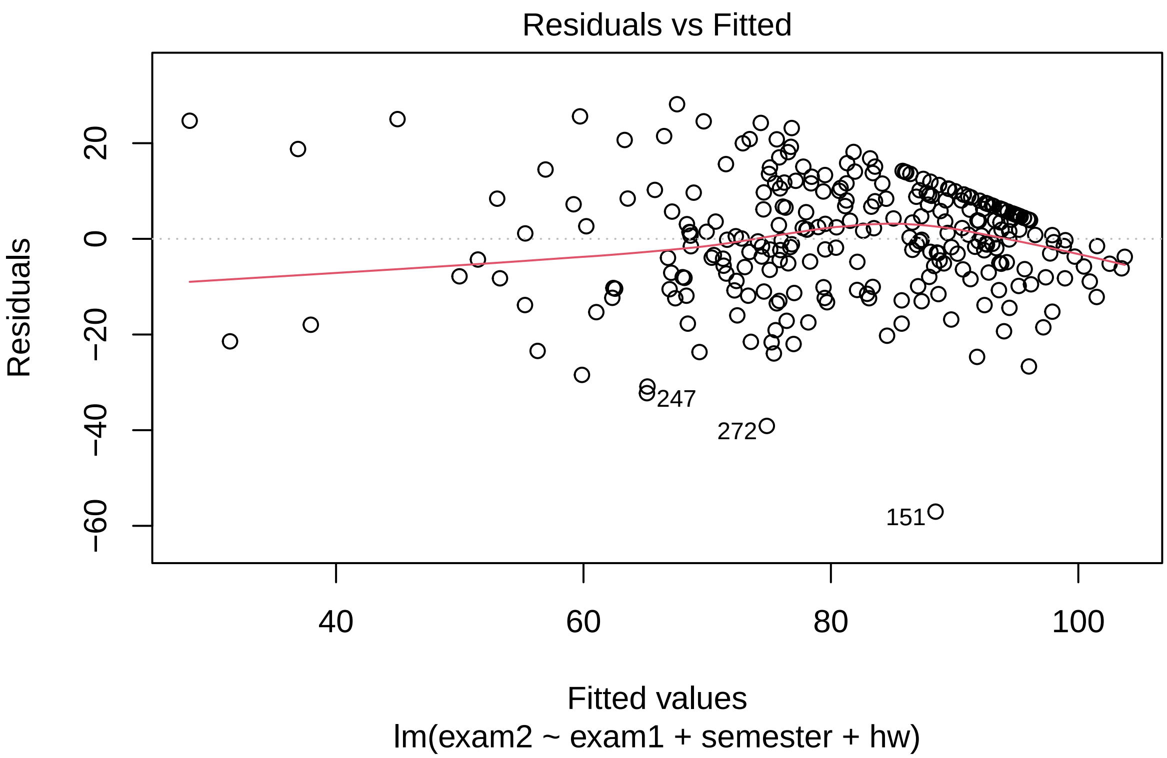
Unusual Observations and Model Assumptions

Now we can analyze the final model for unusual observations and check for deviations from the model assumptions.

Constant variances

First, we can check the model assumption for constant variances by checking the residual vs. fitted plot.

```
plot(grades.reducedmlr, which=1)
```

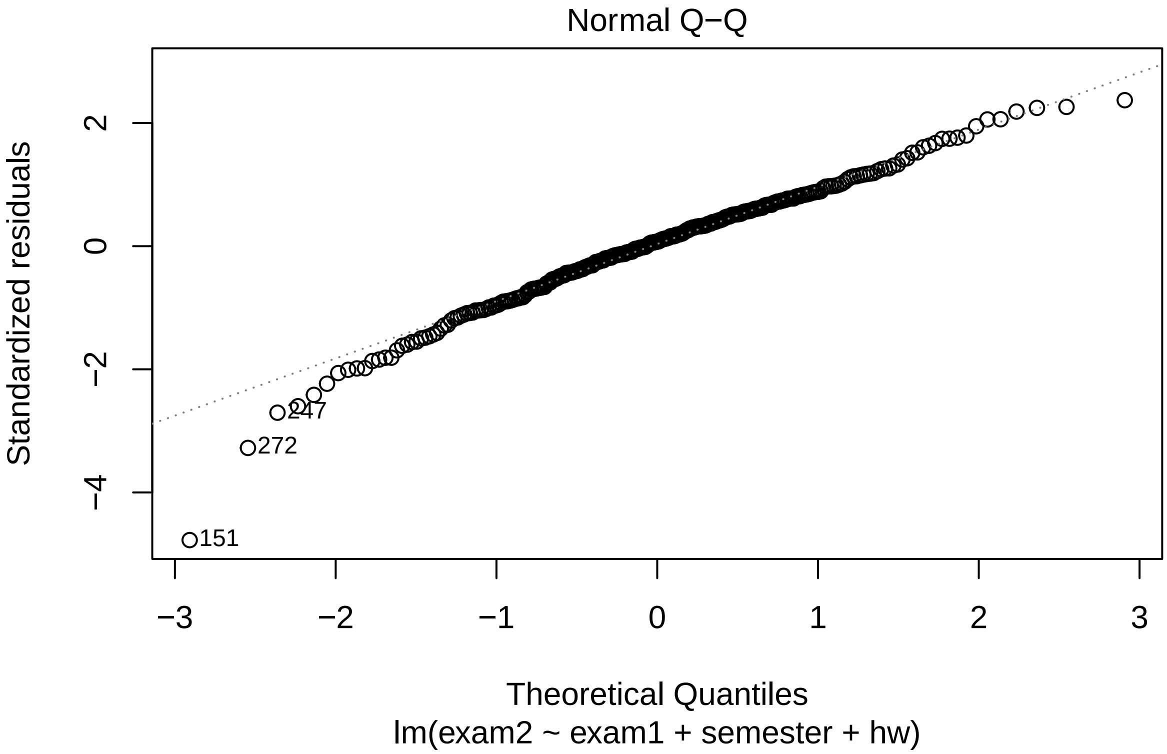


From the residuals vs. fitted plots, we can see the the assumptions for constant variance are not met because the residuals are not evenly disributed around the 0 line, and seem to decrease in magnitude as the residuals increase.

Normality

Next, we can chck for normaltiy of the residuals by creating a QQ plot.

```
plot(grades.reducedmlr, which=2)
```



From the QQ plot, we can see that we seem to have departures from the normality assumption as points along the edges of the plot don't follow the straight line. We can attempt to remedy this and reduce the non-normality of the errors by performing a Box-Cox transformation.

Serial Dependence

It is not possible to check serial dependence for this model because there is no order or time value associated with the data points.

Unusual Observations

```
grades.leverages = lm.influence(grades.reducedmlr)$hat
head(grades.leverages)
```

High Leverage points

```
##          1          2          3          4          5          6
## 0.01564114 0.01553494 0.01452583 0.01434028 0.01372131 0.01219510
```

Outliers

Influential Observations