253Project

Jenny Li, Kristy Ma, Liz Cao

2022/2/12

Contents

Part a & b

Library statements

```
library(dplyr)
library(readr)
library(broom)
library(ggplot2)
library(tidymodels)
tidymodels_prefer()
theme_set(theme_bw())
Sys.setlocale("LC_TIME", "English")
```

```
## [1] ""
```

```
set.seed(74)
```

Read in data

```
breastCa<-read_csv(file = "breast-cancer.csv")
```

Data cleaning

```
breastCa_Re<-breastCa %>%
  drop_na() %>%
  select(radius_mean:fractal_dimension_mean)
```

Creation of cv folds

```
breastCa_Re_CV<-vfold_cv(breastCa_Re, v = 10)
```

Model spec

```
#least square
lm_spec <-
    linear_reg() %>%
    set_engine(engine = 'lm') %>%
    set_mode('regression')

#LASSO
lm_lasso_spec <-
  linear_reg() %>%
  set_args(mixture = 1, penalty = tune()) %>% ## mixture = 1 indicates Lasso
  set_engine(engine = 'glmnet') %>% #note we are using a different engine
  set_mode('regression')
```

Recipes & workflows

```
#least square
least_rec <- recipe(area_mean ~ ., data = breastCa_Re) %>%
    step_corr(all_predictors()) %>%
    step_nzv(all_predictors()) %>% # removes variables with the same value
    step_normalize(all_numeric_predictors()) %>% # important standardization step for LASSO
    step_dummy(all_nominal_predictors())

least_lm_wf <- workflow() %>%
    add_recipe(least_rec) %>%
    add_model(lm_spec)

#LASSO
lasso_wf<- workflow() %>%
  add_recipe(least_rec) %>%
  add_model(lm_lasso_spec)
```

Fit & tune models

```
#least square
least_fit <- fit(least_lm_wf, data = breastCa_Re)
```

```
least_fit %>% tidy()
```

```
## # A tibble: 8 x 5
##   term                    estimate std.error statistic   p.value
##   <chr>                      <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)              655.         2.10   311.    0
## 2 radius_mean              369.         5.07    72.8   4.66e-288
## 3 texture_mean               1.63       2.29     0.713 4.76e-  1
## 4 smoothness_mean            0.746      3.11     0.239 8.11e-  1
## 5 compactness_mean         -69.7        6.51   -10.7   1.82e- 24
## 6 concavity_mean            37.5        5.54     6.76  3.56e- 11
## 7 symmetry_mean              1.65       2.79     0.594 5.53e-  1
## 8 fractal_dimension_mean    41.2        4.93     8.36  5.05e- 16
```

```
#LASSO
#tune
penalty_grid <- grid_regular(
  penalty(range = c(-3, 1)), #log10 transformed
  levels = 30)

tune_output <- tune_grid( # new function for tuning hyperparameters
  lasso_wf, # workflow
  resamples = breastCa_Re_CV, # cv folds
  metrics = metric_set(rmse, mae),
  grid = penalty_grid # penalty grid defined above
)

#fit
best_se_penalty <- select_by_one_std_err(tune_output, metric = 'mae', desc(penalty))
final_wf_se <- finalize_workflow(lasso_wf, best_se_penalty)
lasso_fit <- fit(final_wf_se , data = breastCa_Re)
lasso_fit %>% tidy()
```

```
## # A tibble: 8 x 3
##   term                    estimate penalty
##   <chr>                      <dbl>   <dbl>
## 1 (Intercept)              655.      2.04
## 2 radius_mean              348.      2.04
## 3 texture_mean               0       2.04
## 4 smoothness_mean            0       2.04
## 5 compactness_mean         -26.3     2.04
## 6 concavity_mean            21.9     2.04
## 7 symmetry_mean              0       2.04
## 8 fractal_dimension_mean    14.4     2.04
```

Part c

Calculate and collect CV metrics

```
# Least Square model
least_fit_cv <- fit_resamples(least_lm_wf,
  resamples = breastCa_Re_CV,
  metrics = metric_set(rmse, mae)
```

```
)

least_fit_cv %>% collect_metrics(summarize = TRUE)
```

```
## # A tibble: 2 x 6
##   .metric .estimator  mean     n std_err .config
##   <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 mae     standard    33.6    10    1.56 Preprocessor1_Model1
## 2 rmse    standard    49.4    10    4.90 Preprocessor1_Model1
```

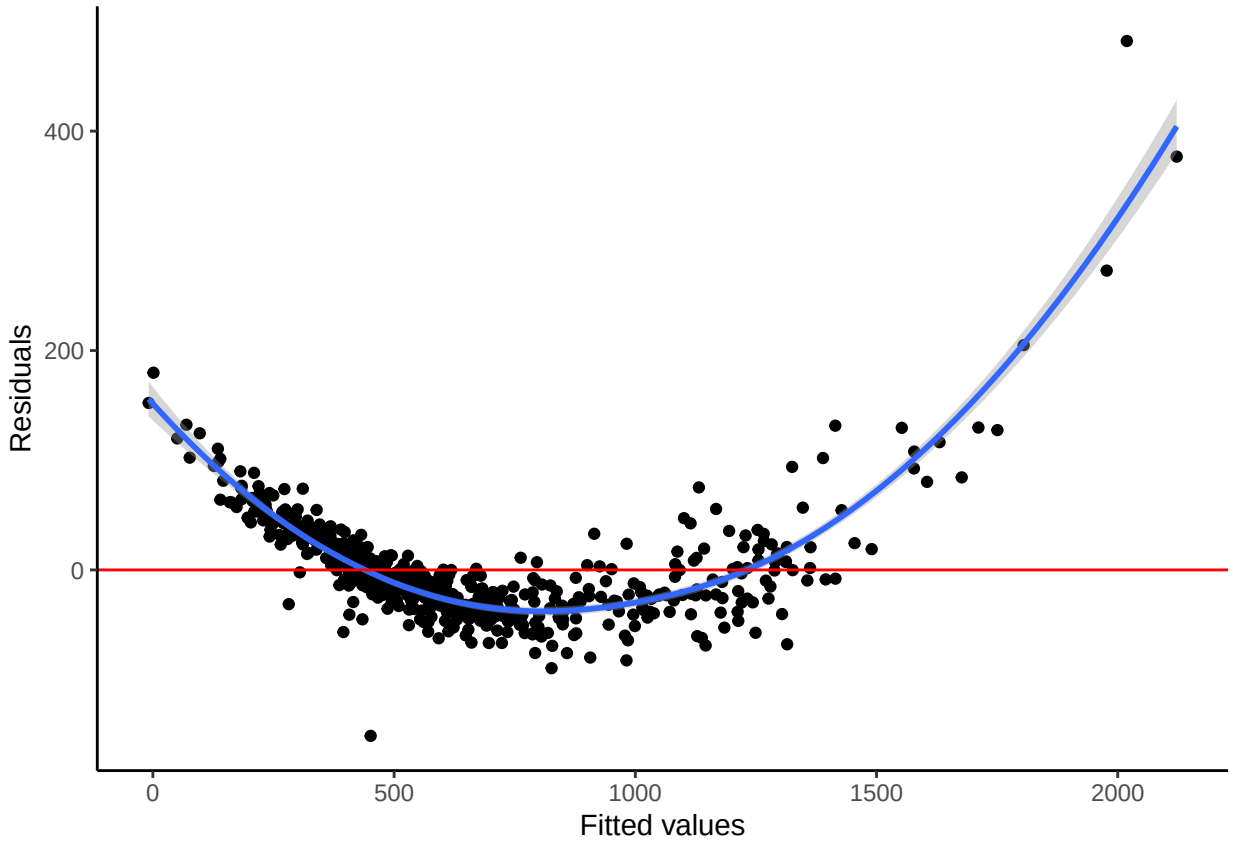
```
# LASSO model
tune_output %>%
  collect_metrics() %>%
  filter(penalty == (best_se_penalty
                     %>% pull(penalty)))
```

```
## # A tibble: 2 x 7
##   penalty .metric .estimator  mean     n std_err .config
##     <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1    2.04 mae     standard    34.2    10    1.49 Preprocessor1_Model25
## 2    2.04 rmse    standard    51.1    10    5.26 Preprocessor1_Model25
```

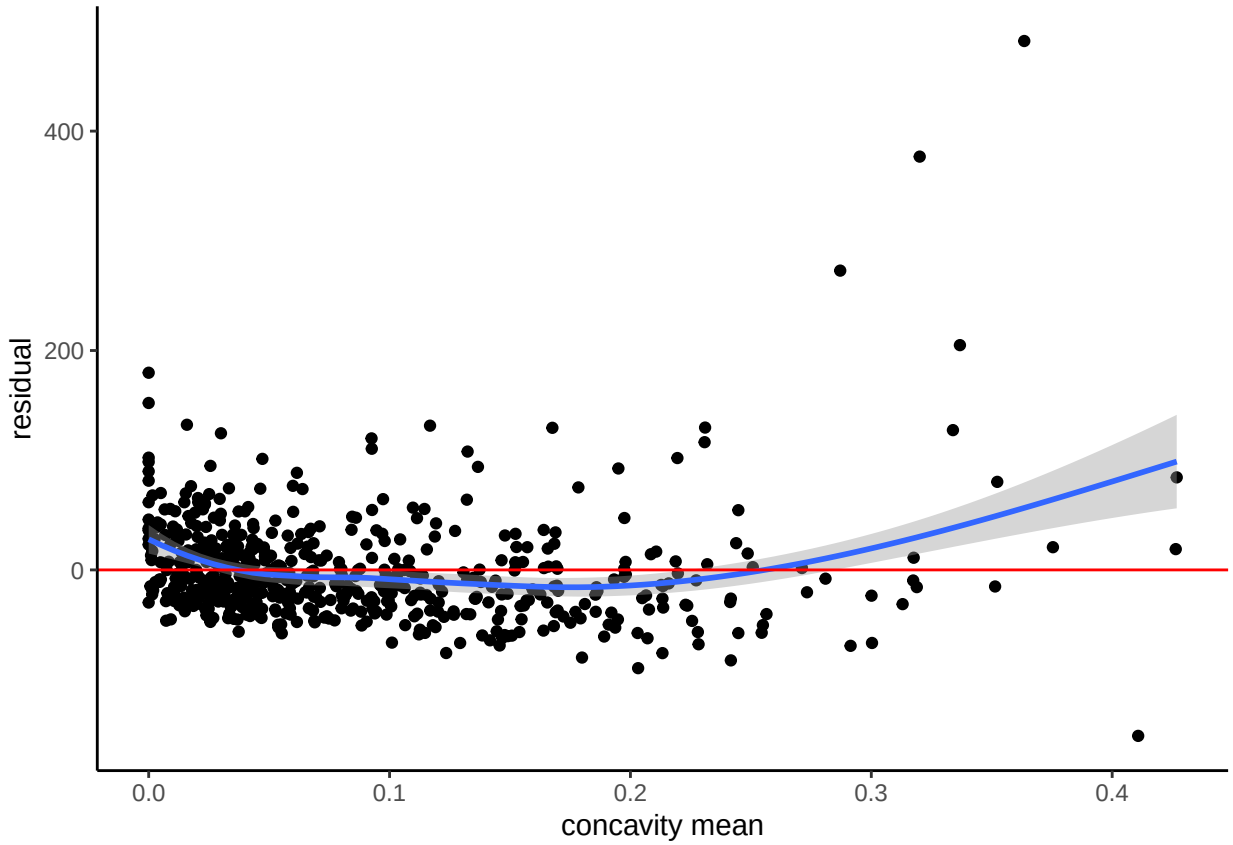
Part d

Residual Plots

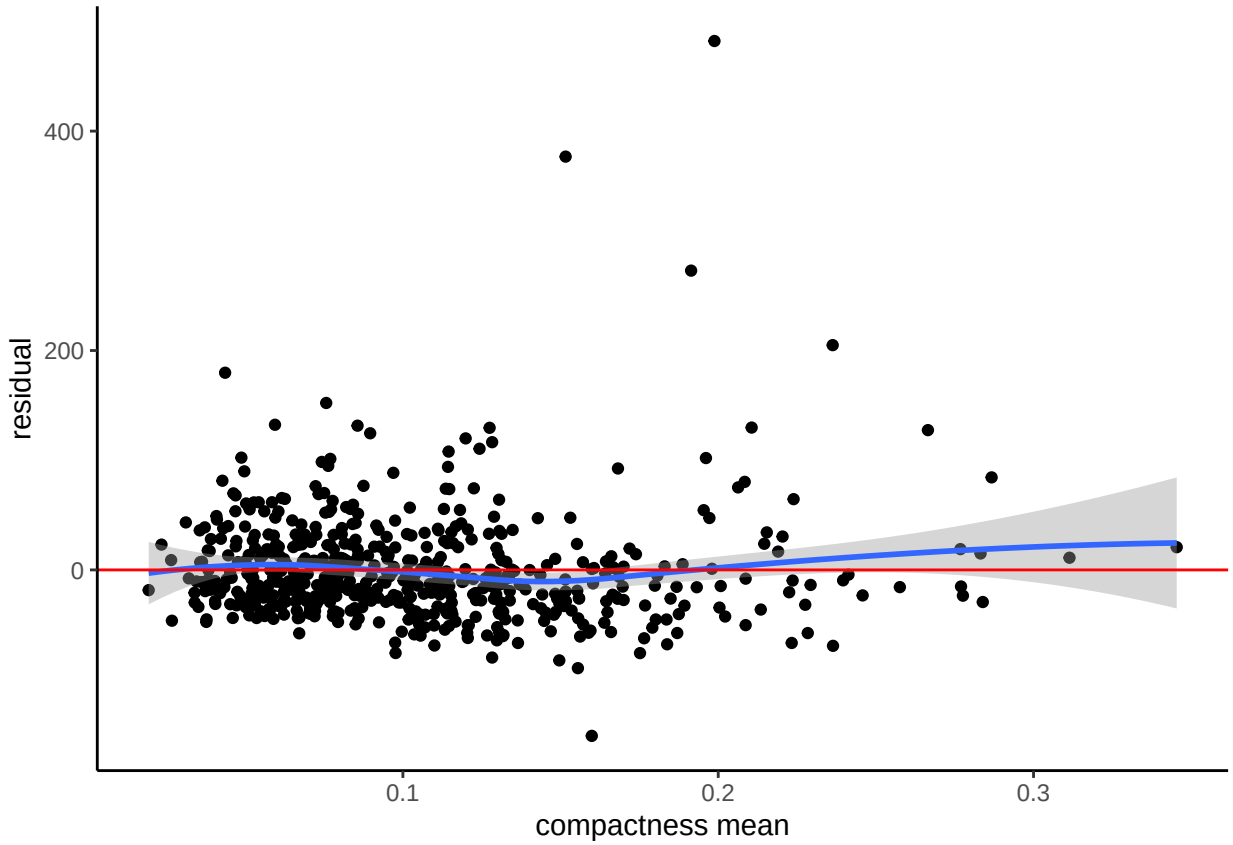
```
least_fit_output <- least_fit %>%
  predict(new_data = breastCa_Re) %>%
  bind_cols(breastCa_Re) %>%
  mutate(resid = area_mean - .pred)

ggplot(least_fit_output, aes(x = .pred, y = resid)) +
  geom_point() +
  geom_smooth() +
  geom_hline(yintercept = 0, color = "red") +
  labs(x = "Fitted values", y = "Residuals") +
  theme_classic()
```



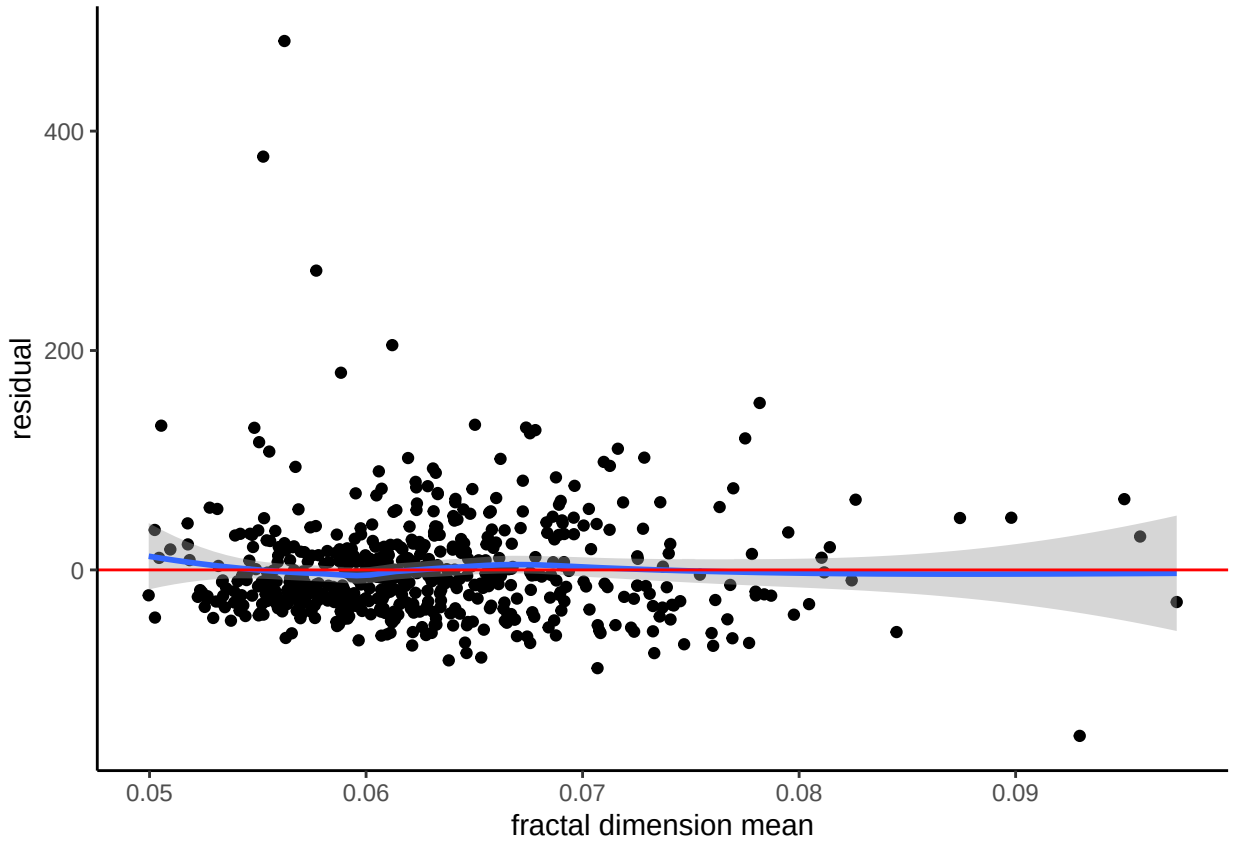

```
# Residuals vs. predictors (x's)
ggplot(least_fit_output, aes(x = concavity_mean, y = resid)) +
  geom_point() +
  geom_smooth() +
  labs(x = "concavity mean", y = "residual") +
  geom_hline(yintercept = 0, color = "red") +
  theme_classic()
```



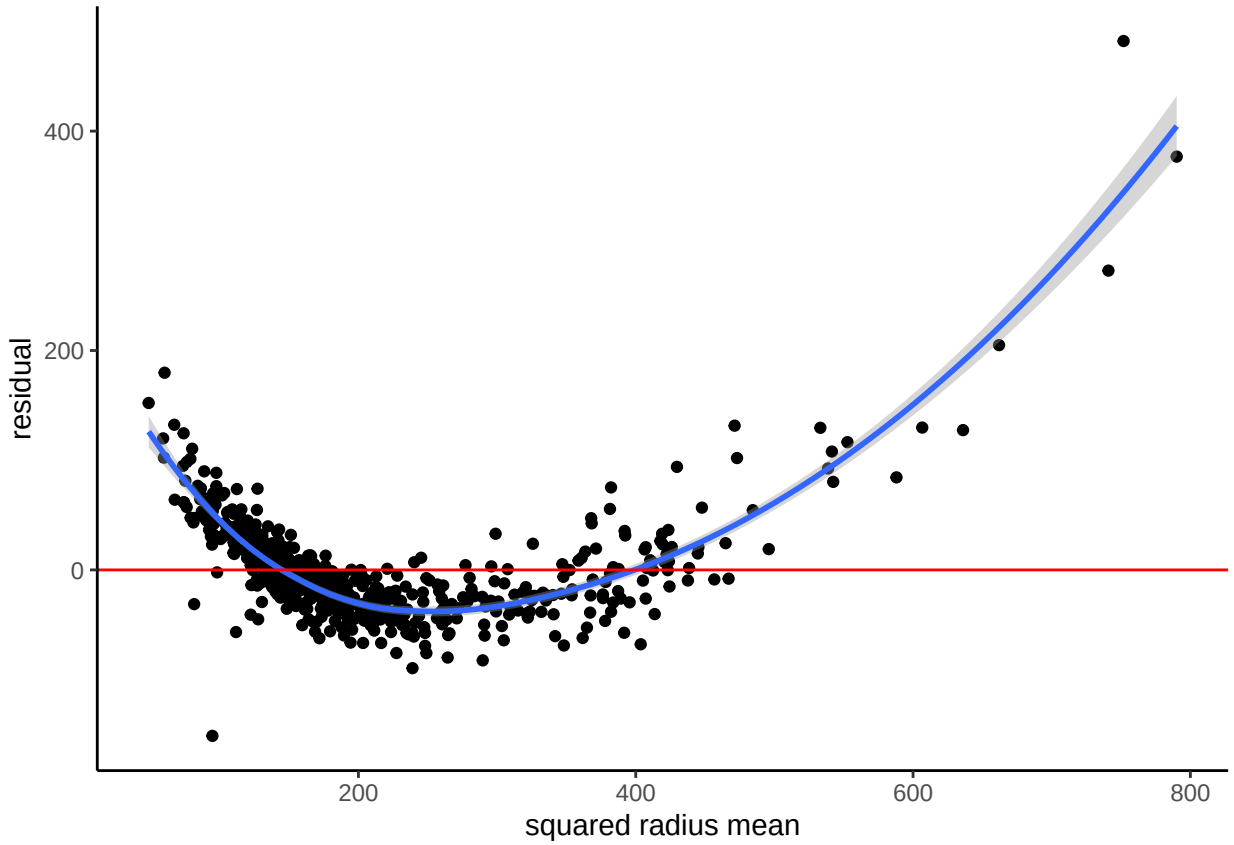

```
# Residuals vs. predictors (x's)
ggplot(least_fit_output, aes(x = compactness_mean, y = resid)) +
  geom_point() +
  geom_smooth() +
  labs(x = "compactness mean", y = "residual") +
  geom_hline(yintercept = 0, color = "red") +
  theme_classic()
```



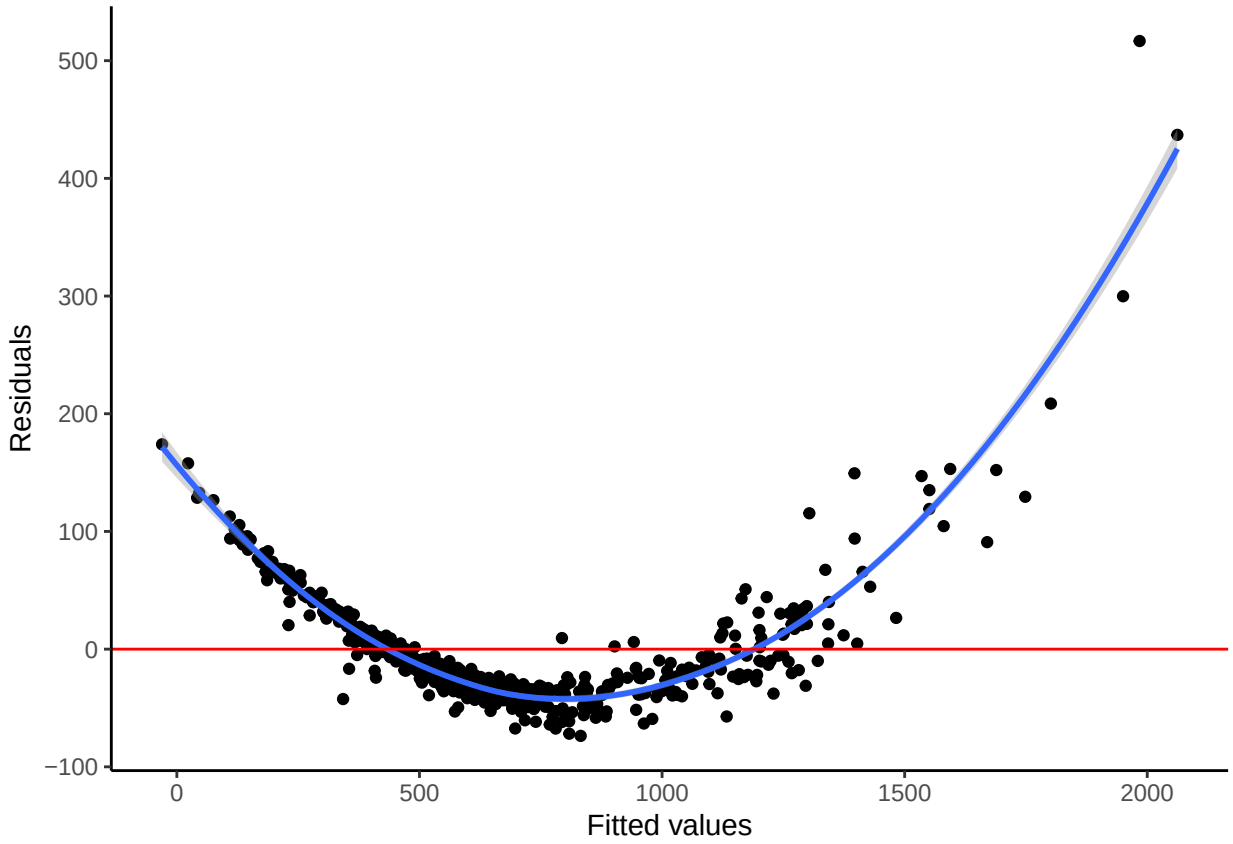

```
# Residuals vs. predictors (x's)
ggplot(least_fit_output, aes(x = fractal_dimension_mean, y = resid)) +
  geom_point() +
  geom_smooth() +
  labs(x = "fractal dimension mean", y = "residual") +
  geom_hline(yintercept = 0, color = "red") +
  theme_classic()
```



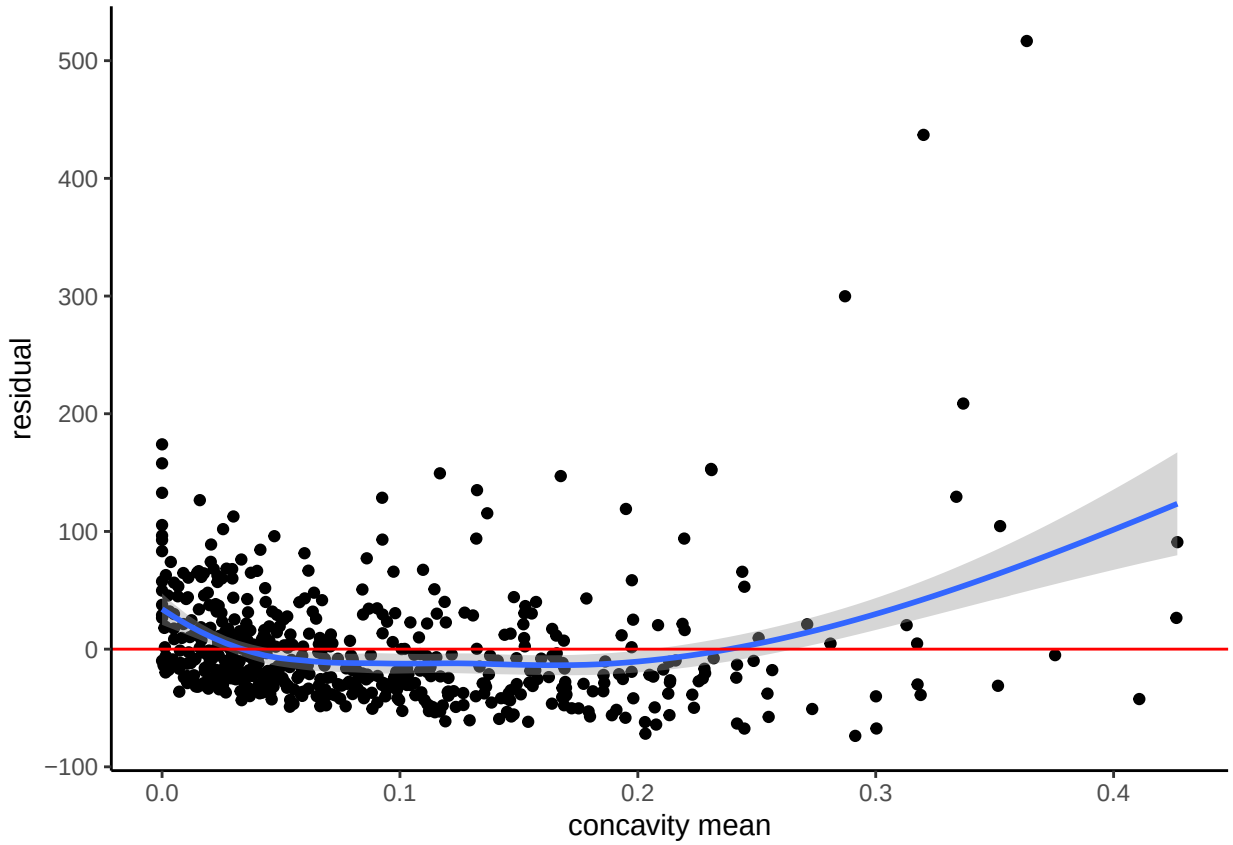

```
# Residuals vs. predictors (x's)
ggplot(least_fit_output, aes(x = radius_mean*radius_mean, y = resid)) +
  geom_point() +
  geom_smooth() +
  labs(x = "squared radius mean", y = "residual") +
  geom_hline(yintercept = 0, color = "red") +
  theme_classic()
```



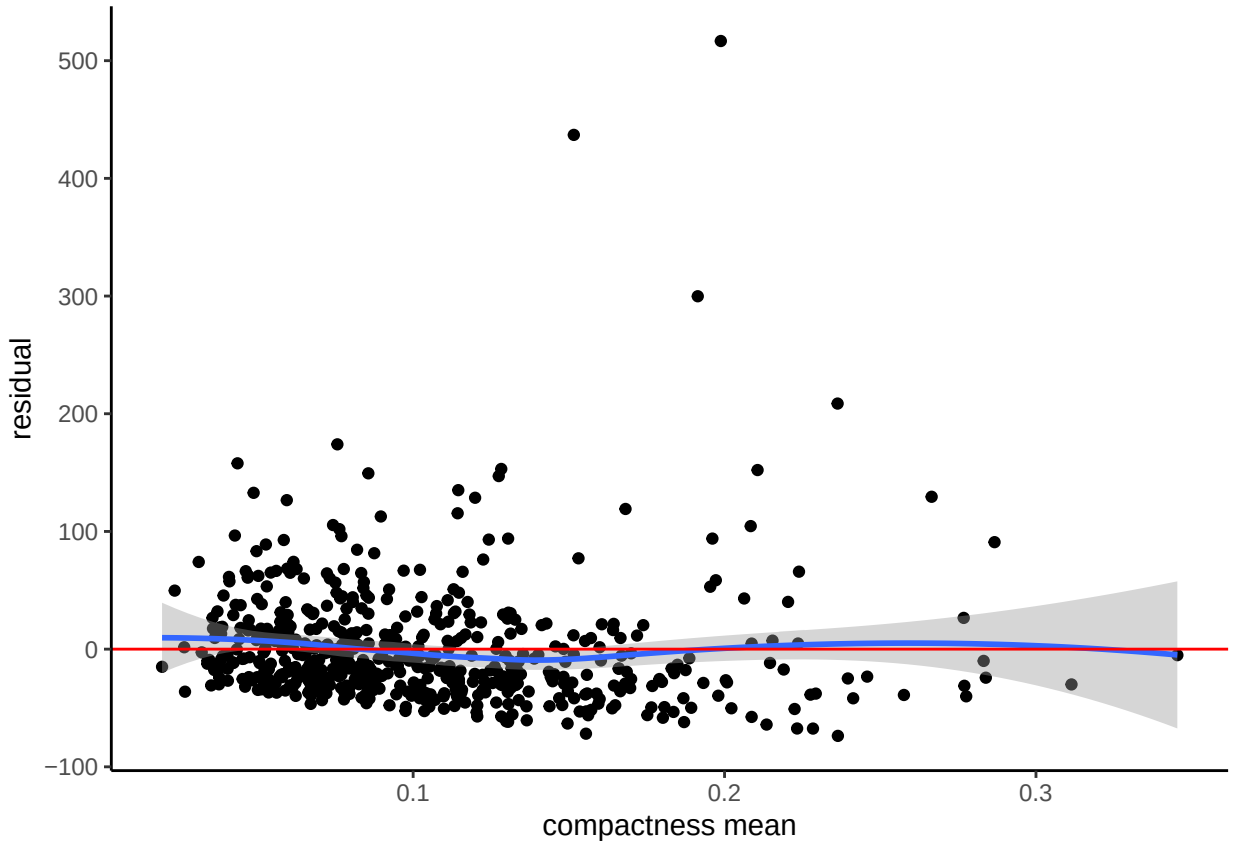

```
lasso_fit_output <- lasso_fit %>%
  predict(new_data = breastCa_Re) %>%
  bind_cols(breastCa_Re) %>%
  mutate(resid = area_mean - .pred)

ggplot(lasso_fit_output, aes(x = .pred, y = resid)) +
  geom_point() +
  geom_smooth() +
  geom_hline(yintercept = 0, color = "red") +
  labs(x = "Fitted values", y = "Residuals") +
  theme_classic()
```



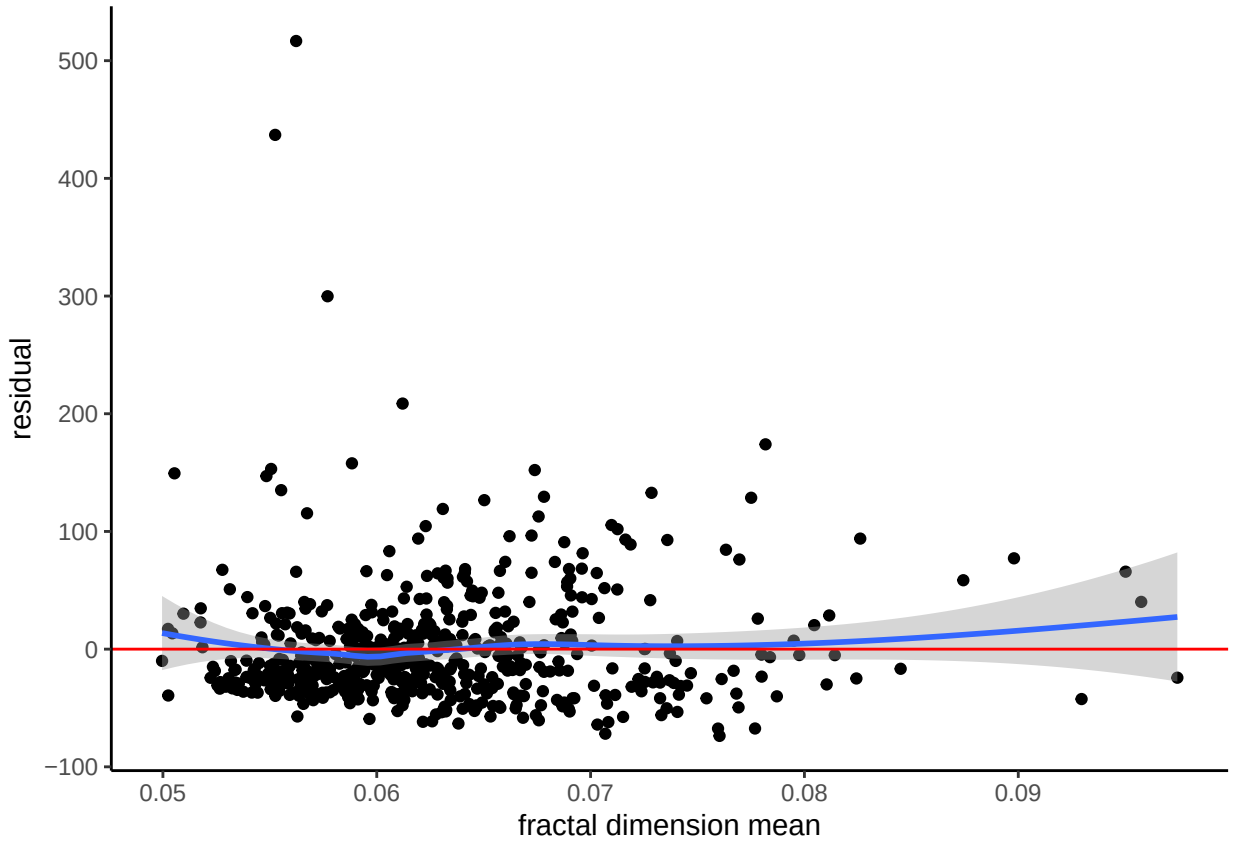

```
# Residuals vs. predictors (x's)
ggplot(lasso_fit_output, aes(x = concavity_mean, y = resid)) +
  geom_point() +
  geom_smooth() +
  labs(x = "concavity mean", y = "residual") +
  geom_hline(yintercept = 0, color = "red") +
  theme_classic()
```



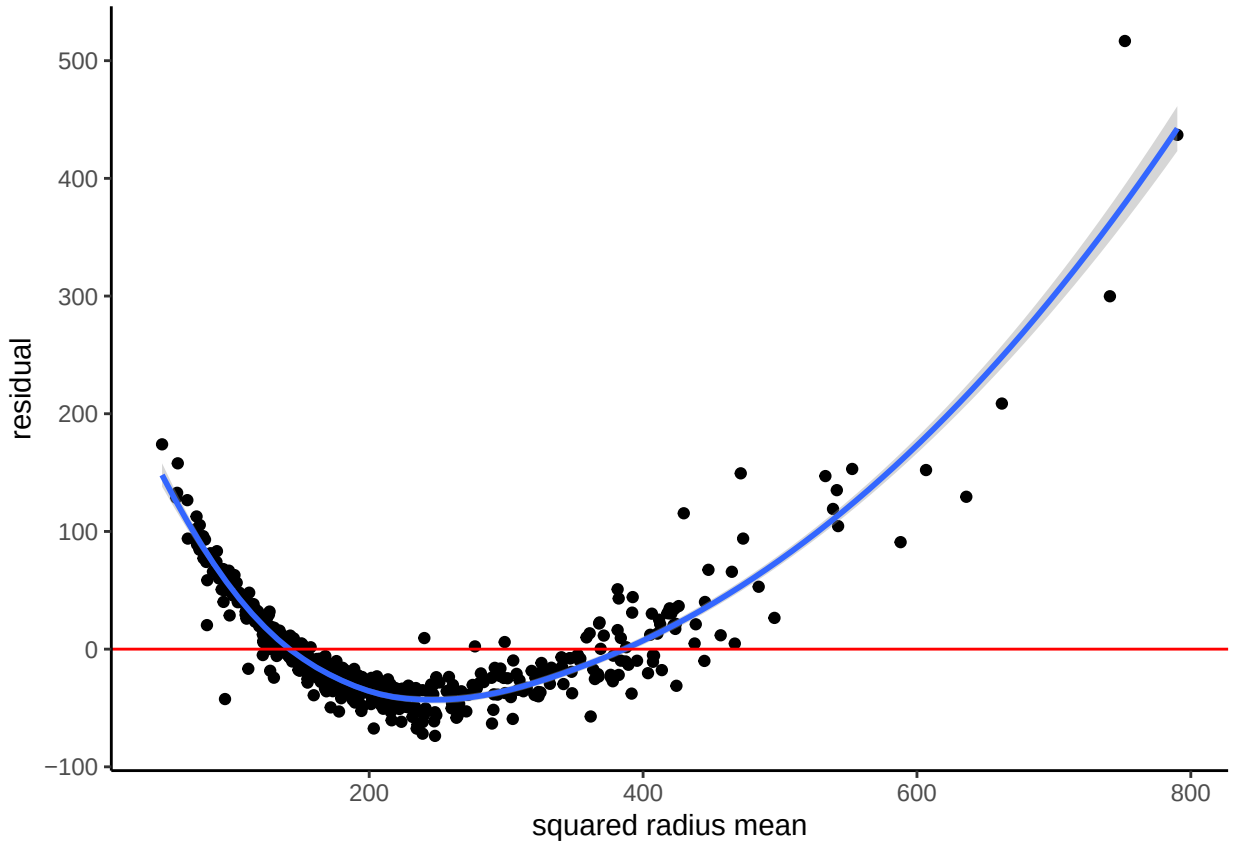

```
# Residuals vs. predictors (x's)
ggplot(lasso_fit_output, aes(x = compactness_mean, y = resid)) +
  geom_point() +
  geom_smooth() +
  labs(x = "compactness mean", y = "residual") +
  geom_hline(yintercept = 0, color = "red") +
  theme_classic()
```




```
# Residuals vs. predictors (x's)
ggplot(lasso_fit_output, aes(x = fractal_dimension_mean, y = resid)) +
  geom_point() +
  geom_smooth() +
  labs(x = "fractal dimension mean", y = "residual") +
  geom_hline(yintercept = 0, color = "red") +
  theme_classic()
```




```
# Residuals vs. predictors (x's)
ggplot(lasso_fit_output, aes(x = radius_mean*radius_mean, y = resid)) +
  geom_point() +
  geom_smooth() +
  labs(x = "squared radius mean", y = "residual") +
  geom_hline(yintercept = 0, color = "red") +
  theme_classic()
```



2. Summarize investigations: Decide on an overall best model based on your investigations so far. To do this, make clear your analysis goals. Predictive accuracy? Interpretability? A combination of both?

Answer: Our goal is to have both accurate predictions and keep interpretability. Our proposed best model: area_mean~ radius_mean + compactness_mean + concavity_mean + fractal_dimension_mean.(based on the p_value and the coefficents from LASSO. We prefer to keep all the variables that contain cofficients because the total of 4 variabls for a model is interpretable.) However, we are using the mean of radius to predict the area which does not really make sense in implementation, but for this assignment we will just leave it there. Also, this is coherent with the residual plot of radius which is a non-linear relationship.

3. Societal impact: Are there any harms that may come from your analyses and/or how the data were collected? What cautions do you want to keep in mind when communicating your work?

Answer: There are some harms that may occur, such as some patients may not be willing to provide their personal information or records for the case study. We need to protect the information safety of the patient. Also, we should know the number of observations may not be able to be generalized.