STA 210 Project

Shamikh Hossain et al.

Section 1: Introduction

In this analysis we take a look at an energy efficiency dataset featuring simulated designs of various buildings shapes and properties along with their associated heating and cooling loads, which are important indicators of a building's energy efficiency. We are interested in using a regression model to determine the key physical properties of the building that, independently or in interaction with other characteristics, have an impact on the building's energy efficiency. Based on these findings, we're hoping to be able to provide insight into specifics on the designing of energy-efficient buildings, and the quantitative impact that each significant feature has on energy efficiency. This is important information because of the growing emphasis placed on energy performance of both old and newly built buildings, and the considerations of improved energy conservation techniques in development projects in countries all around the world.

Section 2: The Data

This data set was created by Angeliki Xifara and was processed by Dr. Athanasios Tsanas at the University of Oxford, UK, by means of simulation on the environmental/architectural analysis software program, *Ecotect*. The software allows civil and environmental engineers to design and simulate a building's performance in the earliest stages, using just its conceptual design. The data set contains 768 samples of building shapes, parameterized by six numerical features and two categorical features, and two potential target variables, heating load and cooling load.

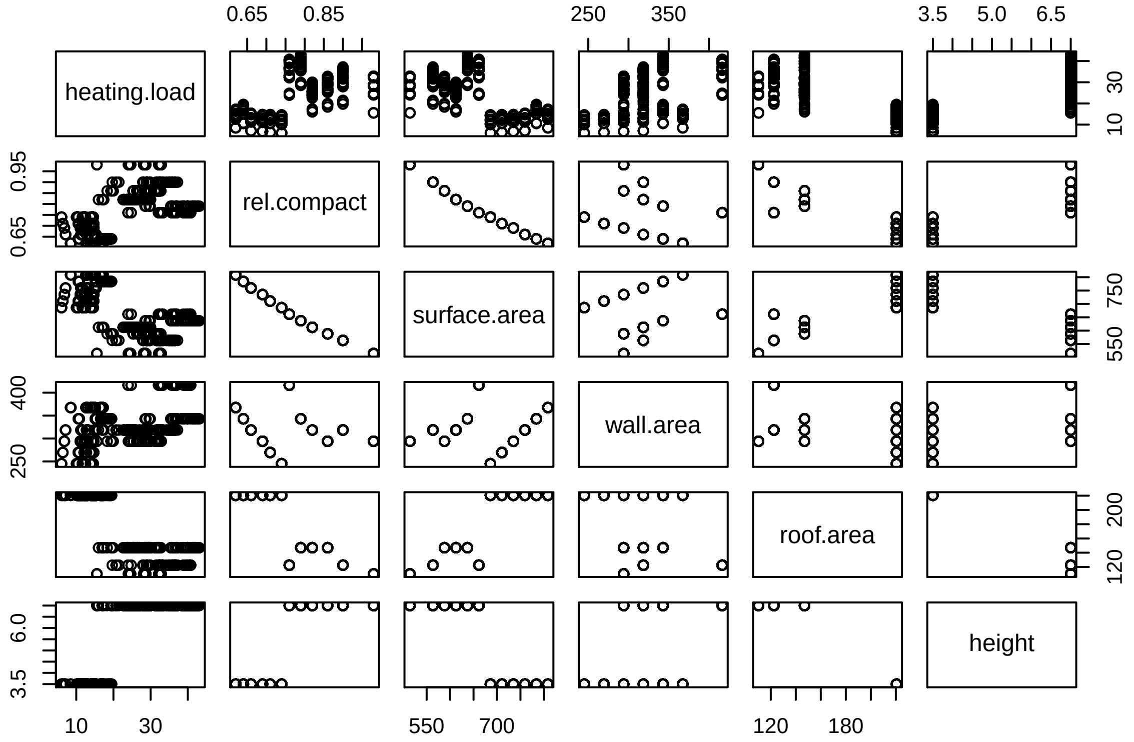
Variables

1. Relative Compactness (`rel.compact`)
2. Surface Area (`surface.area`) - m^2
3. Wall Area (`wall.area`) - m^2
4. Roof Area (`roof.area`) - m^2
5. Overall Height (`height`) - m
6. Orientation (`orientation`) - 2:North, 3:East, 4:South, 5:West
7. Glazing Area (`glazing.area`) - 0%, 10%, 25%, 40% (of floor area)
8. Glazing Area Distribution (`glazing.dist`) - 1:Uniform, 2:North, 3:East, 4:South, 5:West
9. Heating Load (`heating.load`) - kWh/m^2
10. Cooling Load (`cooling.load`) - kWh/m^2

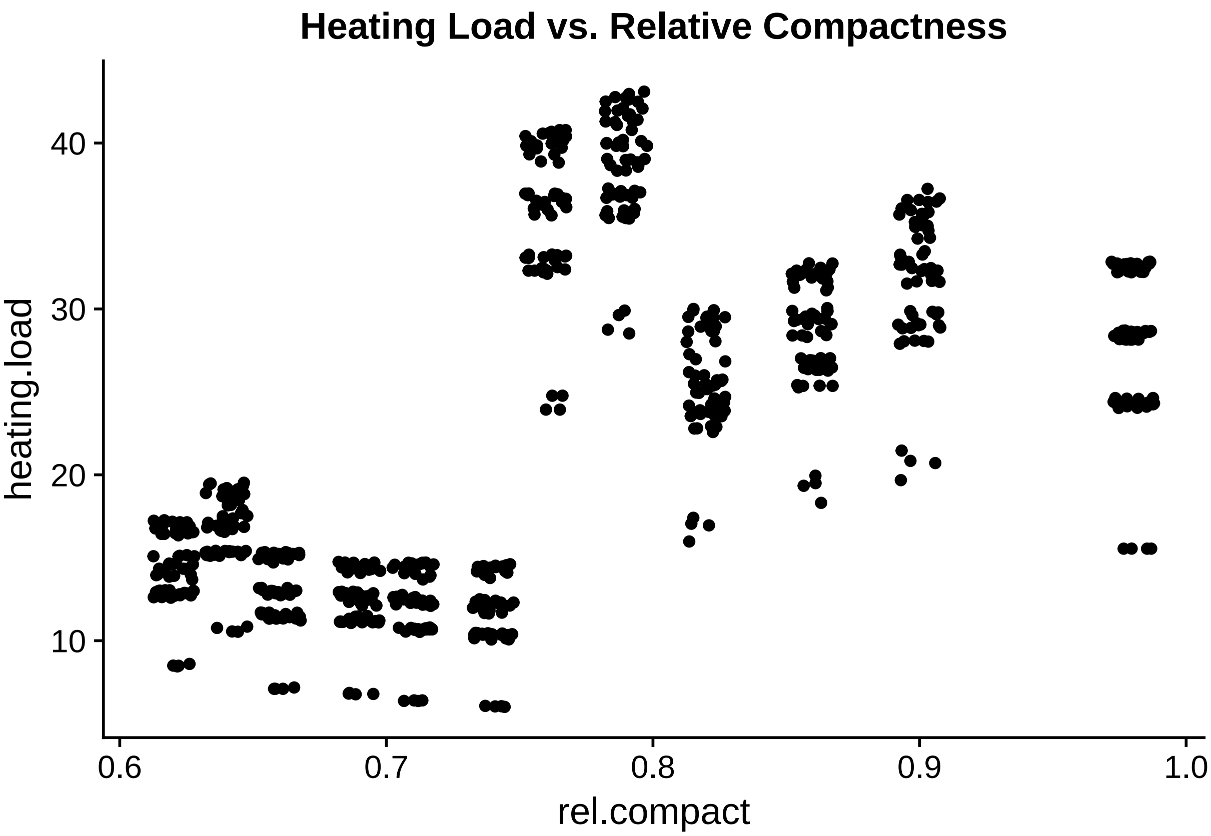
```r
energy <- readxl::read_excel('ENB2012_data.xlsx') %>%
  rename(rel.compact = X1,
         surface.area = X2,
         wall.area = X3,
         roof.area = X4,
         height = X5,
         orientation = X6,
         glazing.area = X7,
         glazing.dist = X8,
         heating.load = Y1,
         cooling.load = Y2) %>%
  mutate(orientation = as.factor(orientation),
         glazing.dist = as.factor(glazing.dist))
glimpse(energy)
```

```
## Observations: 768
## Variables: 10
## $ rel.compact  <dbl> 0.98, 0.98, 0.98, 0.98, 0.90, 0.90, 0.90, 0.90, 0...
## $ surface.area <dbl> 514.5, 514.5, 514.5, 514.5, 563.5, 563.5, 563.5, ...
## $ wall.area    <dbl> 294.0, 294.0, 294.0, 294.0, 318.5, 318.5, 318.5, ...
## $ roof.area    <dbl> 110.25, 110.25, 110.25, 110.25, 122.50, 122.50, 1...
## $ height       <dbl> 7.0, 7.0, 7.0, 7.0, 7.0, 7.0, 7.0, 7.0, 7.0, 7.0,...
## $ orientation  <fct> 2, 3, 4, 5, 2, 3, 4, 5, 2, 3, 4, 5, 2, 3, 4, 5, 2...
## $ glazing.area <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ glazing.dist <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ heating.load <dbl> 15.55, 15.55, 15.55, 15.55, 20.84, 21.46, 20.71, ...
## $ cooling.load <dbl> 21.33, 21.33, 21.33, 21.33, 28.28, 25.38, 25.16, ...
```

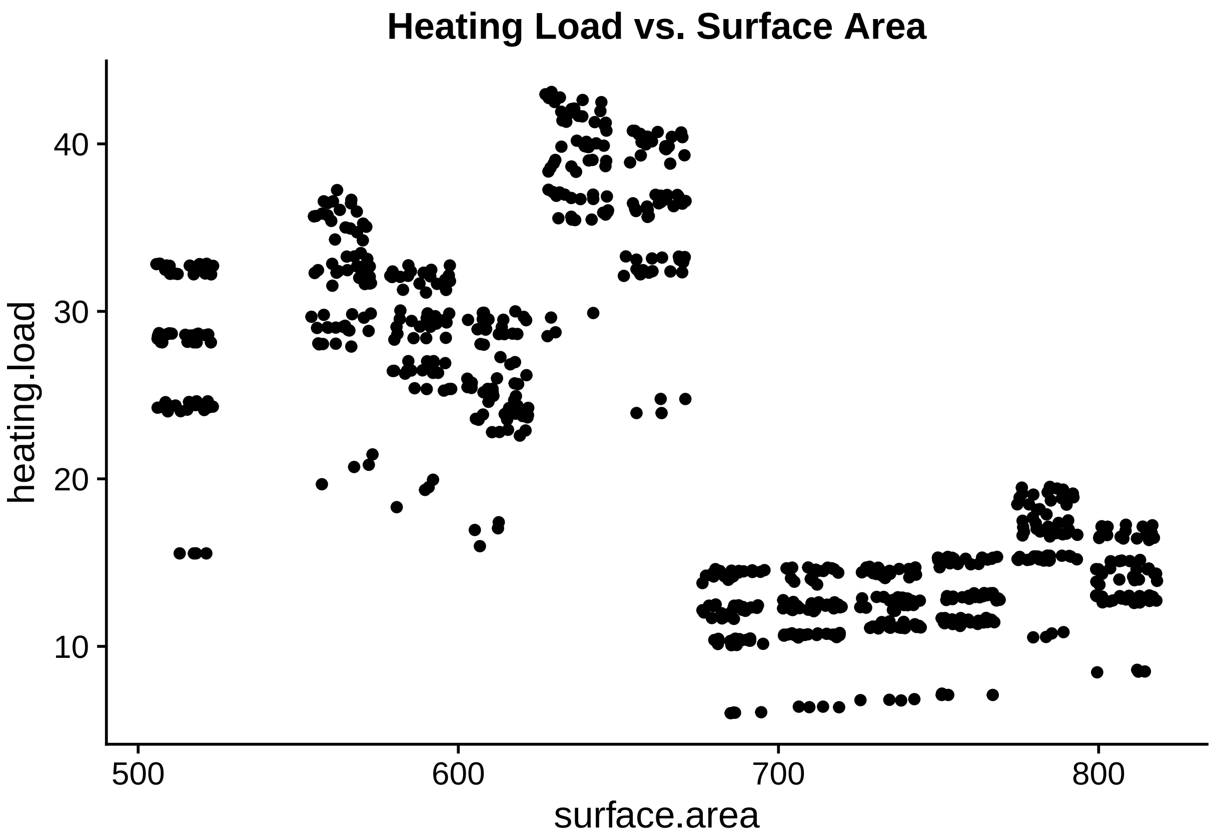
```
# Scatter plot matrix of variables vs. heating load
pairs(heating.load ~ rel.compact + surface.area + wall.area + roof.area + height, data = energy)
```



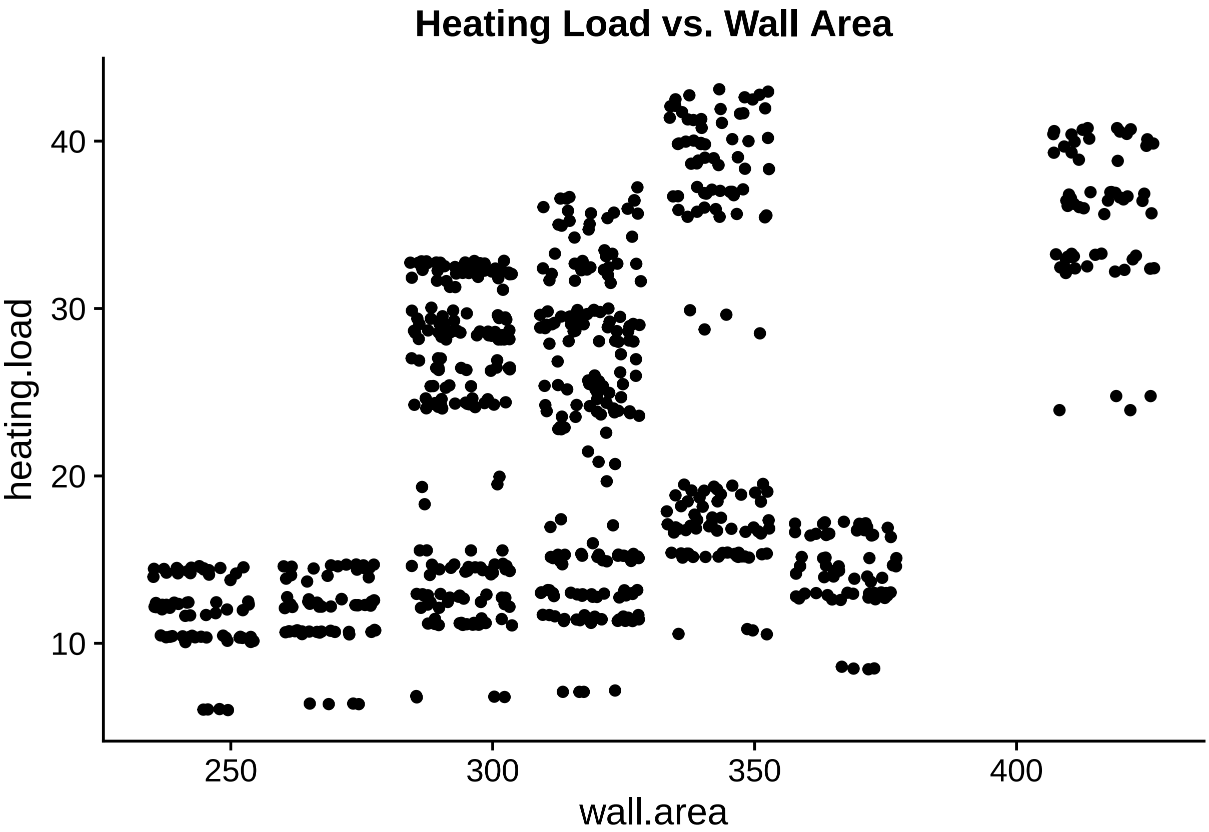
```
# Scatter plot matrix of variables vs. heating load
#energy %>%
#  dplyr::select(heating.load, rel.compact, surface.area, wall.area, roof.area, height) %>%
#  gather(key="key", value="value", -heating.load) %>%
#  ggplot(aes(x=value, y=heating.load)) + geom_jitter()  + facet_wrap(~ key, scales='free_x')

energy %>%
  dplyr::select(heating.load, rel.compact) %>%
  ggplot(aes(x=rel.compact, y=heating.load)) + geom_jitter() +
  labs(title="Heating Load vs. Relative Compactness")
```



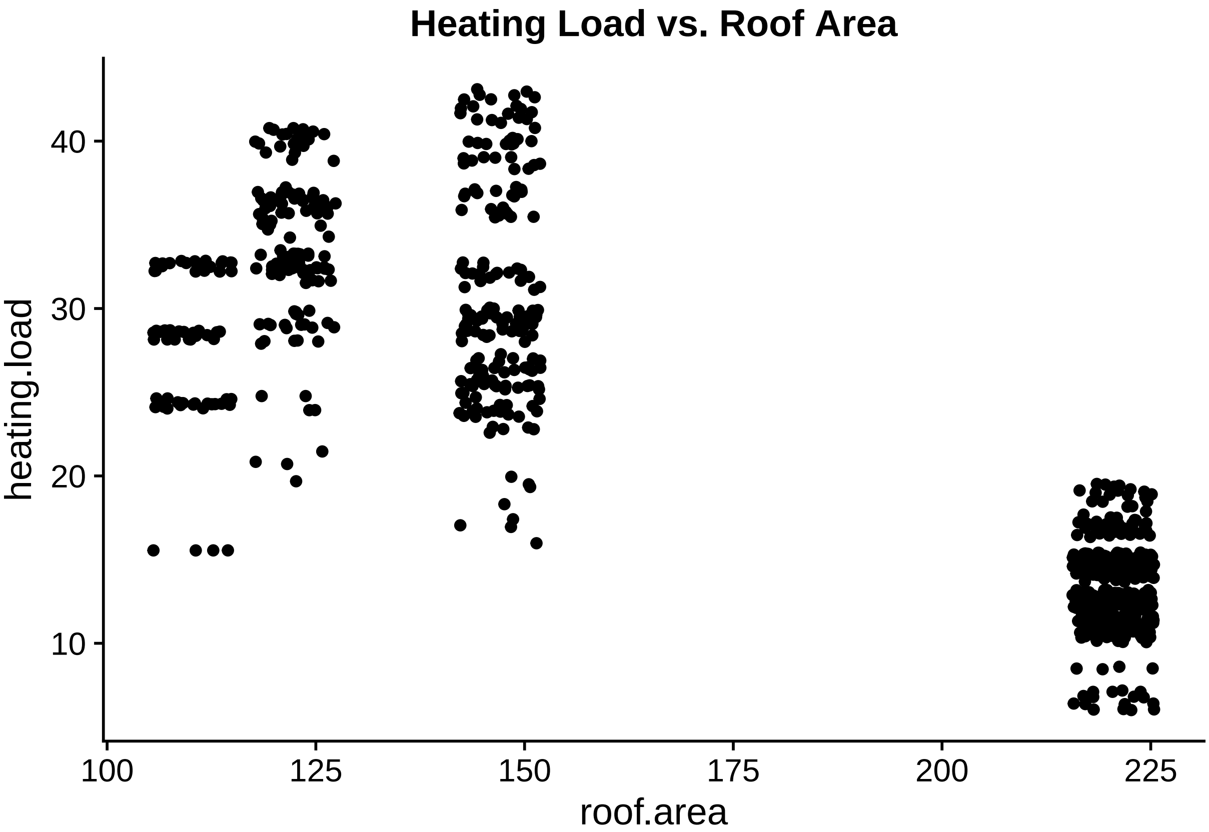
```
energy %>%
  dplyr::select(heating.load, surface.area) %>%
  ggplot(aes(x=surface.area, y=heating.load)) + geom_jitter() +
  labs(title="Heating Load vs. Surface Area")
```



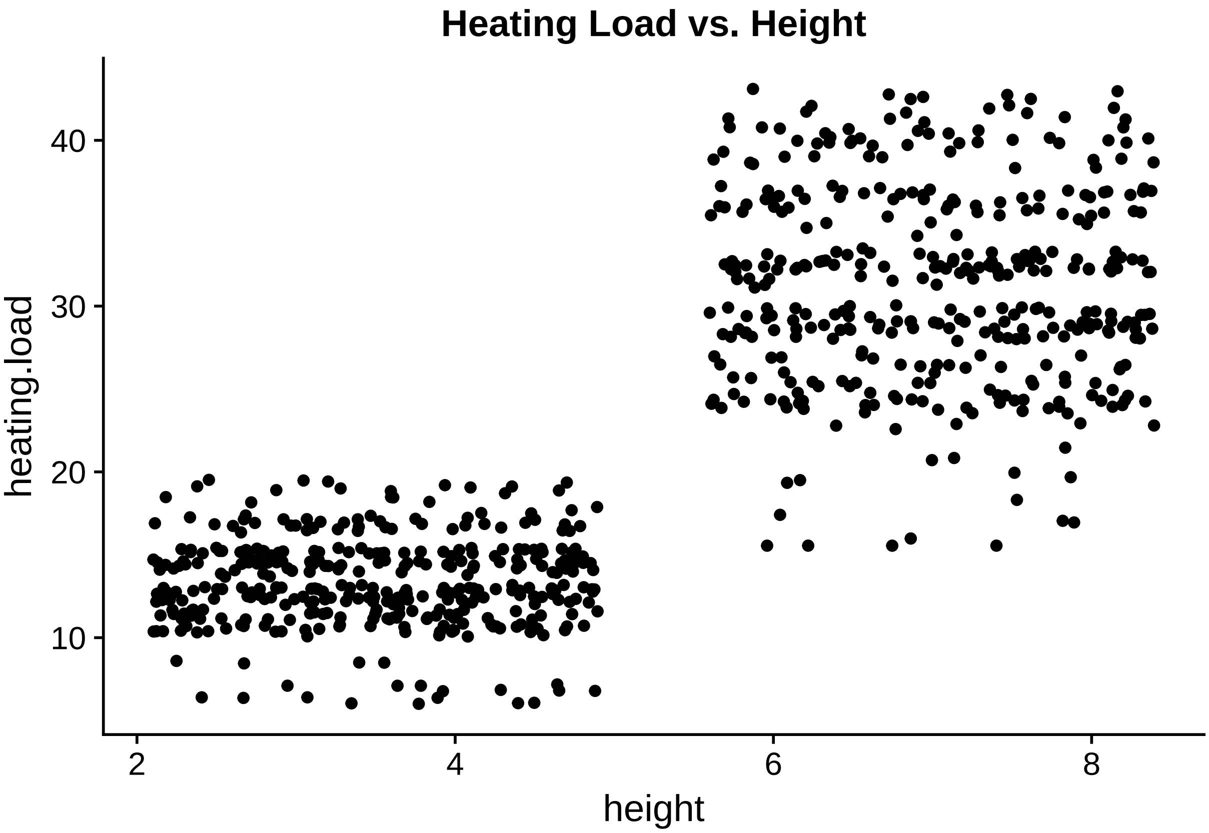
```
energy %>%
  dplyr::select(heating.load, wall.area) %>%
  ggplot(aes(x=wall.area, y=heating.load)) + geom_jitter() +
  labs(title="Heating Load vs. Wall Area")
```



```
energy %>%
  dplyr::select(heating.load, roof.area) %>%
  ggplot(aes(x=roof.area, y=heating.load)) + geom_jitter() +
  labs(title="Heating Load vs. Roof Area")
```

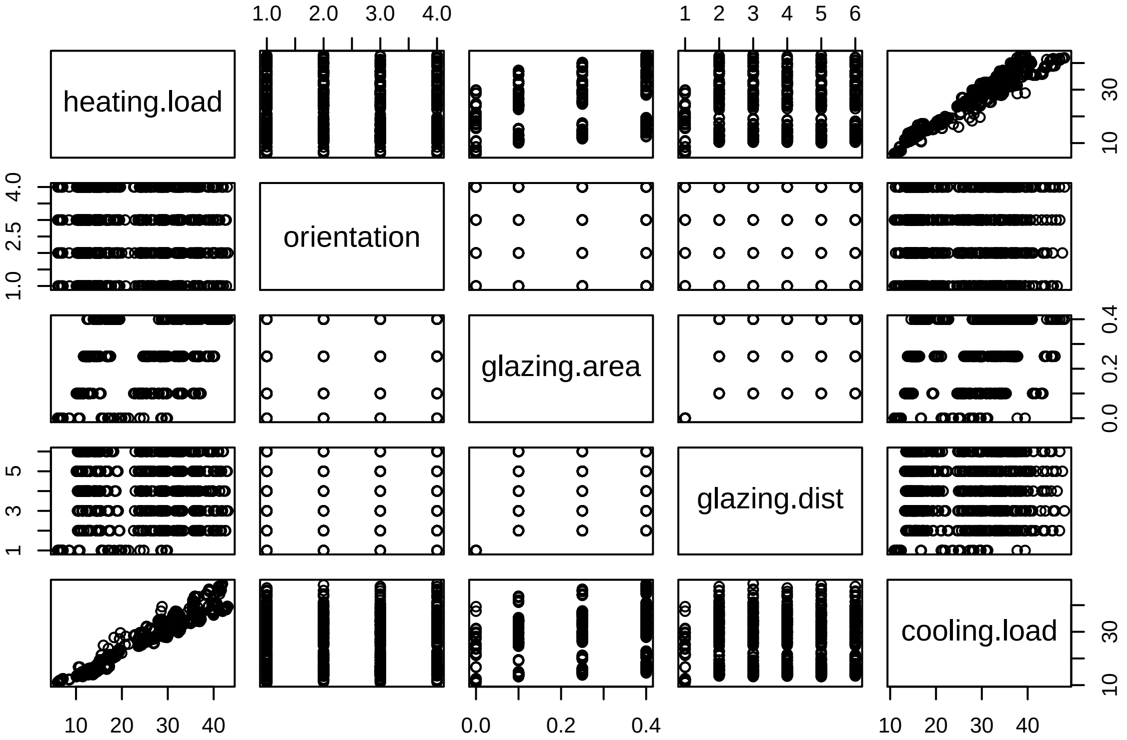


```
energy %>%
  dplyr::select(heating.load, height) %>%
  ggplot(aes(x=height, y=heating.load)) + geom_jitter() +
  labs(title="Heating Load vs. Height")
```



```r
#pairs(heating.load ~ rel.compact + surface.area + wall.area + roof.area + height, data = energy)
```

```r
pairs(heating.load ~ orientation + glazing.area + glazing.dist + cooling.load, data = energy)
```

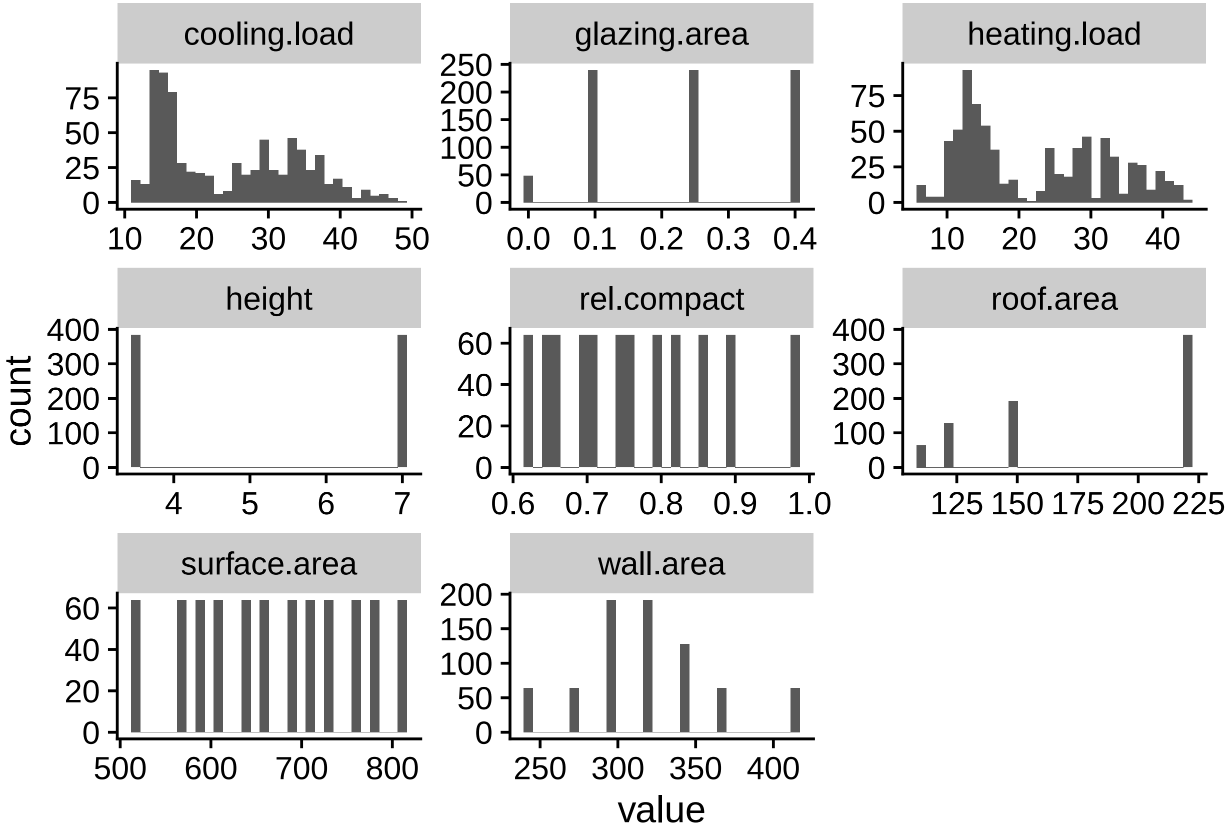


```r
sapply(energy, summary) # Use lapply for list
```

```
## $rel.compact
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.6200  0.6825  0.7500  0.7642  0.8300  0.9800 
## 
## $surface.area
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   514.5   606.4   673.8   671.7   741.1   808.5 
## 
## $wall.area
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   245.0   294.0   318.5   318.5   343.0   416.5 
## 
## $roof.area
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   110.2   140.9   183.8   176.6   220.5   220.5 
## 
## $height
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    3.50    3.50    5.25    5.25    7.00    7.00 
## 
## $orientation
##   2   3   4   5 
## 192 192 192 192 
## 
## $glazing.area
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.1000  0.2500  0.2344  0.4000  0.4000 
## 
## $glazing.dist
##   0   1   2   3   4   5 
##  48 144 144 144 144 144 
## 
## $heating.load
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    6.01   12.99   18.95   22.31   31.67   43.10 
## 
## $cooling.load
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   10.90   15.62   22.08   24.59   33.13   48.03
```

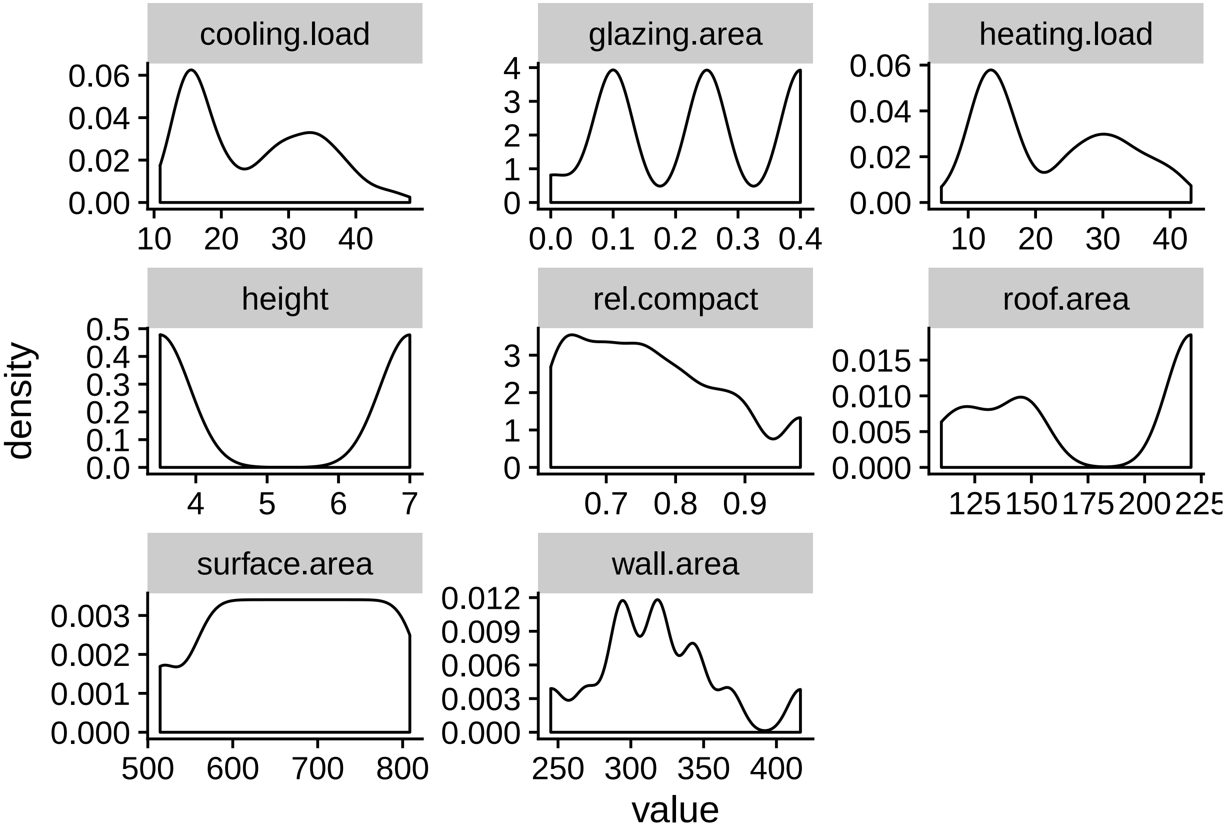
```r
# Plot the distributions of the numerical features
energy %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



INCLUDE

```
energy %>%
  keep(is.numeric) %>%                     # Keep only numeric columns
  gather() %>%                             # Convert to key-value pairs
  ggplot(aes(value)) +                     # Plot the values
    facet_wrap(~ key, scales = "free") +   # In separate panels
    geom_density()                         # as density
```



Miles

```
set.seed(101)

# 768 datapoints
idx <- sample.int(n = nrow(energy),
                  size = floor(0.8*nrow(energy)),
                  replace=FALSE)
train <- energy[idx,]
test <- energy[-idx,]
```

```
df <- train

model <- lm(heating.load ~ ., data=df)

#model2 <- lm(heating.load ~ (rel.compact + surface.area + wall.area + roof.area + height + orientation

#anova(model, model2)

reduced <- step(model)
```

```
## Start:  AIC=609.47
## heating.load ~ rel.compact + surface.area + wall.area + roof.area +
##     height + orientation + glazing.area + glazing.dist + cooling.load
##
##
```

```
## Step:  AIC=609.47
## heating.load ~ rel.compact + surface.area + wall.area + height + 
##     orientation + glazing.area + glazing.dist + cooling.load
## 
##                Df Sum of Sq    RSS     AIC
## <none>                      1577.8  609.47
## - rel.compact   1      7.29 1585.1  610.31
## - surface.area  1      8.31 1586.1  610.70
## - orientation   3     19.64 1597.4  611.07
## - height        1     78.27 1656.0  637.20
## - wall.area     1    108.55 1686.3  648.33
## - glazing.dist  5    260.84 1838.6  693.42
## - glazing.area  1    360.87 1938.6  733.94
## - cooling.load  1   2979.70 4557.5 1258.78
```

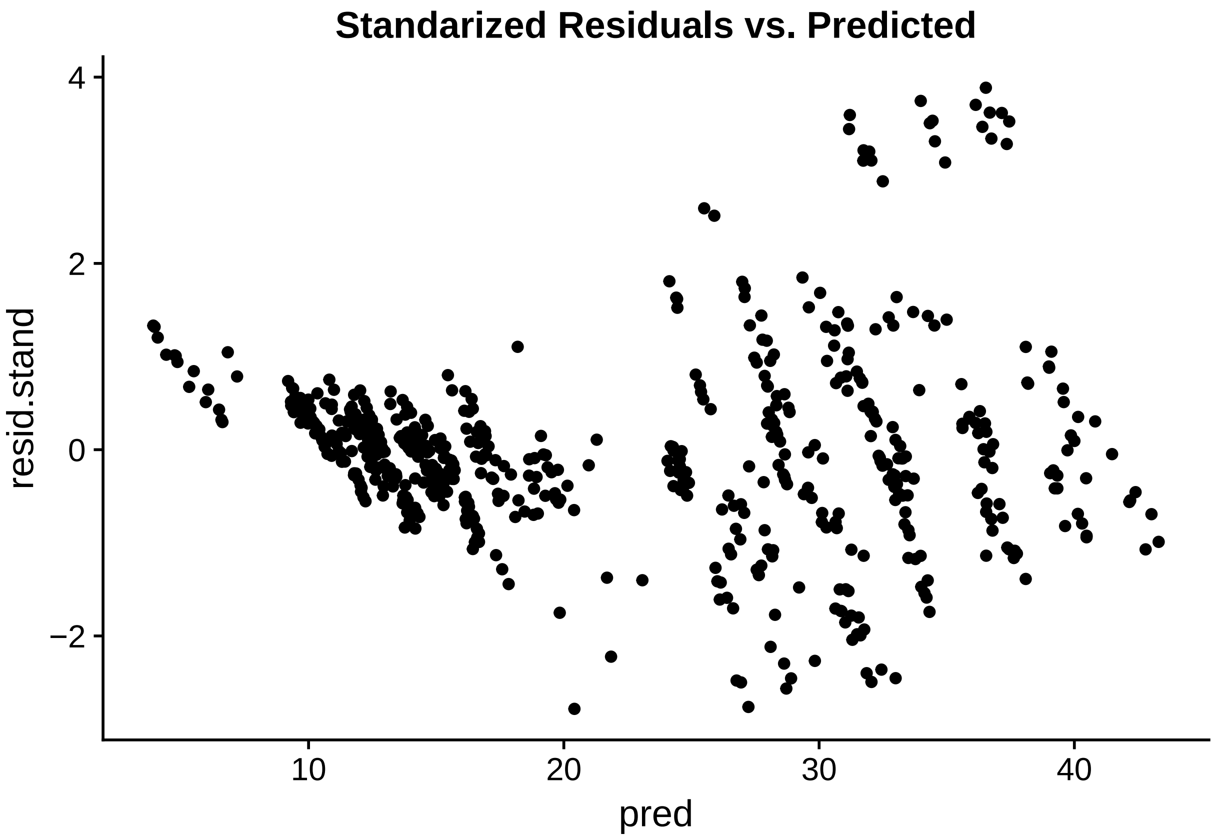
```
summary(reduced)
```

```
## 
## Call:
## lm(formula = heating.load ~ rel.compact + surface.area + wall.area + 
##     height + orientation + glazing.area + glazing.dist + cooling.load, 
##     data = df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.4347 -0.7819 -0.0537  0.6092  6.2372 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      6.207320  11.783605   0.527   0.5985    
## rel.compact    -10.725159   6.447076  -1.664   0.0967 .  
## surface.area    -0.018828   0.010602  -1.776   0.0762 .  
## wall.area        0.027146   0.004229   6.420 2.78e-10 ***
## height           1.226551   0.225003   5.451 7.32e-08 ***
## orientation3     0.252610   0.185026   1.365   0.1727    
## orientation4     0.095293   0.186657   0.511   0.6099    
## orientation5    -0.240646   0.183217  -1.313   0.1895    
## glazing.area     7.294574   0.623205  11.705  < 2e-16 ***
## glazing.dist1    3.050364   0.330905   9.218  < 2e-16 ***
## glazing.dist2    2.907211   0.329521   8.823  < 2e-16 ***
## glazing.dist3    2.918238   0.326857   8.928  < 2e-16 ***
## glazing.dist4    3.004619   0.332328   9.041  < 2e-16 ***
## glazing.dist5    3.059312   0.332438   9.203  < 2e-16 ***
## cooling.load     0.703691   0.020922  33.634  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.623 on 599 degrees of freedom
## Multiple R-squared:  0.9742, Adjusted R-squared:  0.9736 
## F-statistic:  1614 on 14 and 599 DF,  p-value: < 2.2e-16
```

```
train.pred <- train %>%
  mutate(pred = predict.lm(model, train),
         resid.stand = rstandard(model),
         resid = resid(model))
```

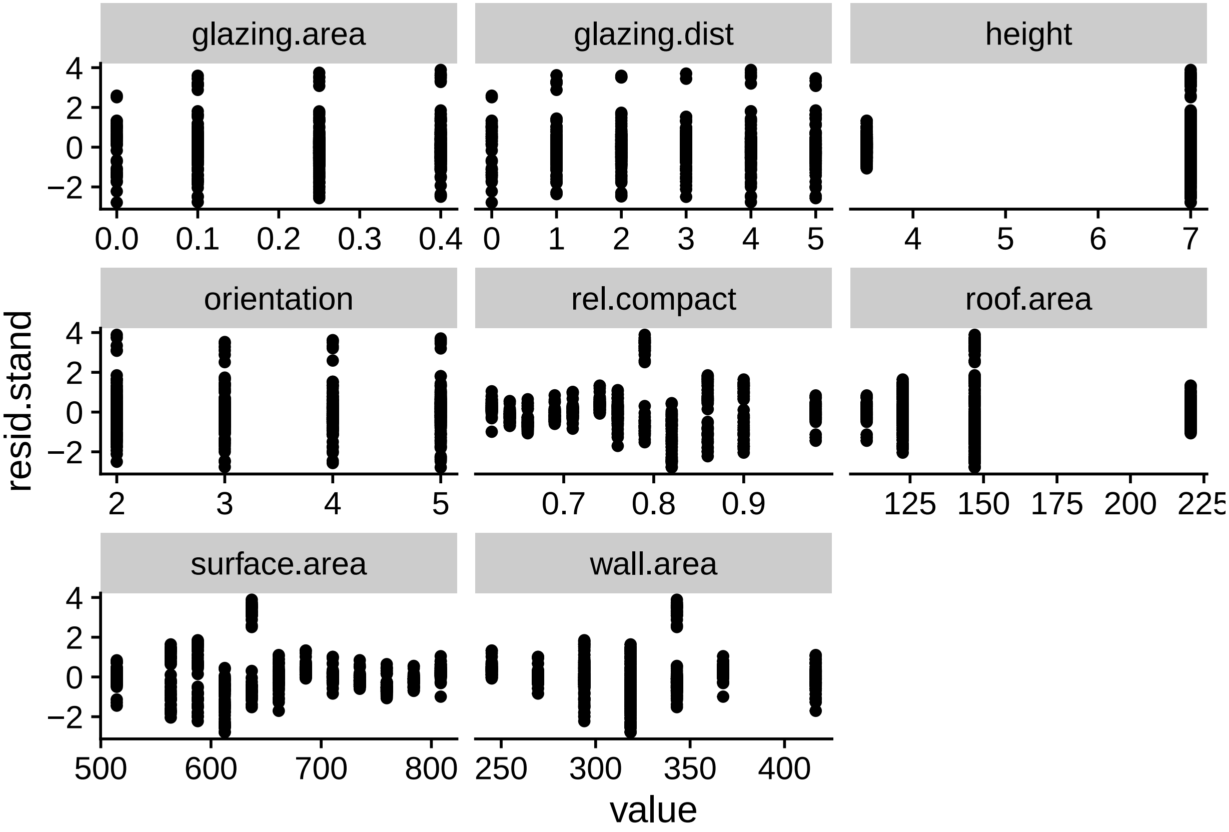
```
## Warning in predict.lm(model, train): prediction from a rank-deficient fit
## may be misleading
```

```
train.pred %>%
  ggplot(aes(x=pred, y=resid.stand)) + geom_point() +
  labs(title = "Standarized Residuals vs. Predicted")
```



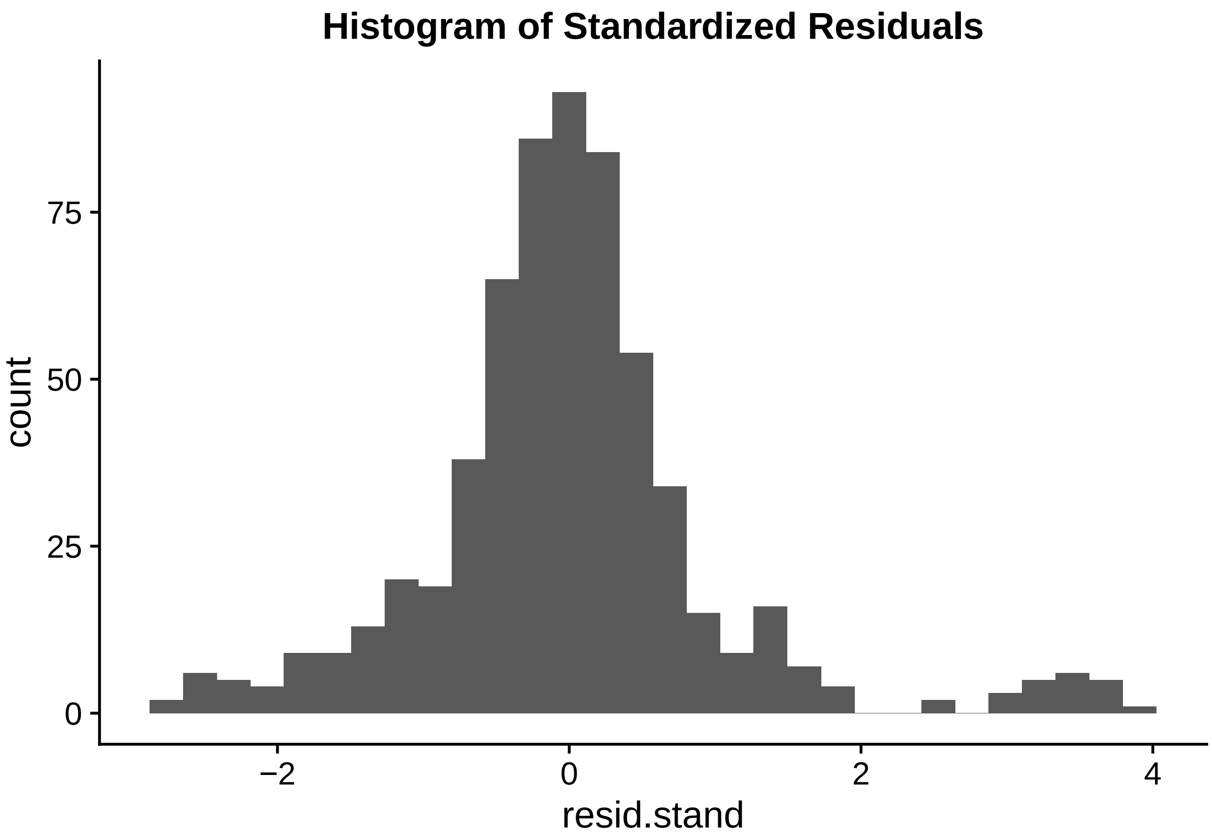
```
train.pred %>%
  #dplyr::select(glazing.dist, rel.compact, surface.area,  resid) %>%
  dplyr::select(-pred, -heating.load, -cooling.load, -resid) %>%
  gather(key="var", value="value", -resid.stand) %>%
  mutate(value = as.numeric(value)) %>%
  ggplot(aes(x=value, y=resid.stand)) +
  geom_point() +
  #geom_jitter(size=1) +
  #geom_boxplot() +
  facet_wrap( ~ var, ncol=3, scales = 'free_x')
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

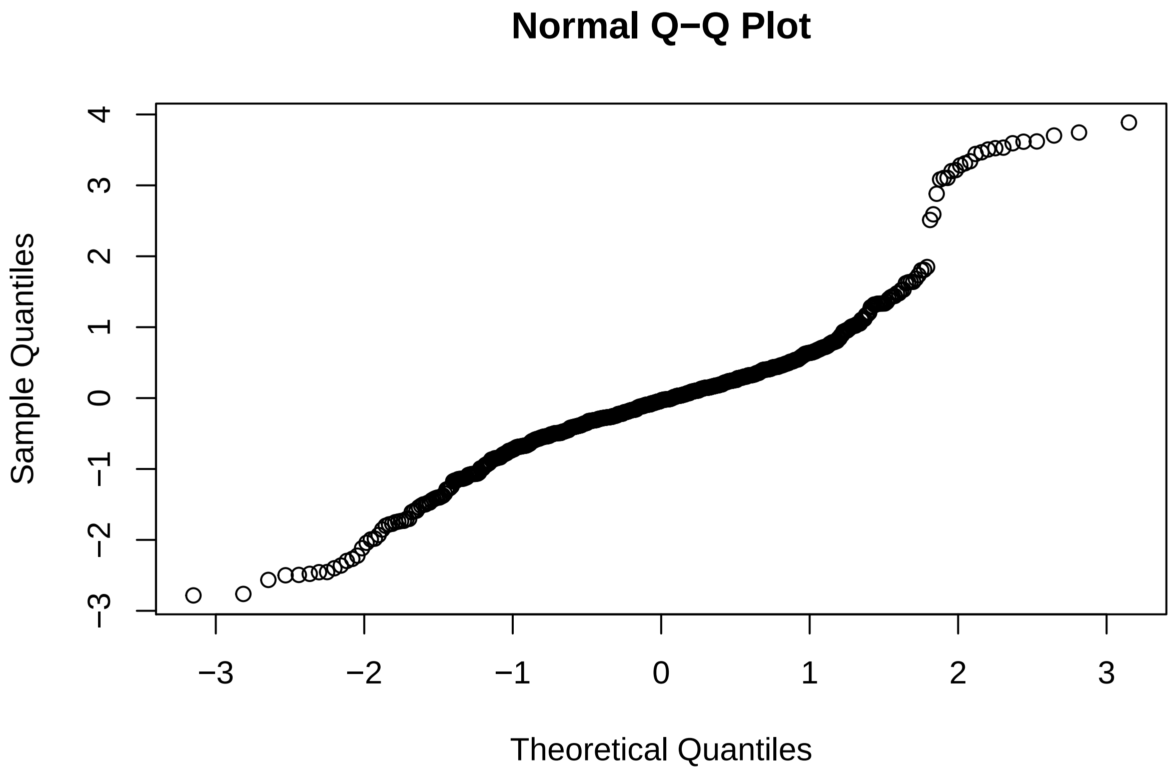


```r
train.pred %>%
  ggplot(aes(x=resid.stand)) + geom_histogram() +
  labs(title='Histogram of Standardized Residuals')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
qqnorm(train.pred$resid.stand)
```



```r
test.pred <- test %>%
  mutate(pred = predict.lm(model, test)) %>%
  mutate(resid = pred - heating.load,
         resid.stand = ((pred - heating.load) - mean(pred - heating.load)) / sd(pred - heating.load))
```

```
## Warning in predict.lm(model, test): prediction from a rank-deficient fit
## may be misleading
```

```r
print('MSE Train')
```

```
## [1] "MSE Train"
```

```r
print(mean(train.pred$resid^2))
```

```
## [1] 2.569651
```

```r
print('MSE Val')
```

```
## [1] "MSE Val"
```

```r
print(mean(test.pred$resid^2))
```

```
## [1] 3.652673
```

```r
print('Train R^2')
```

```
## [1] "Train R^2"
```

```r
RSS = sum(train.pred$resid^2)
TSS = sum((train.pred$heating.load - mean(train.pred$heating.load))^2)
print(1 - RSS/TSS)
```

```
## [1] 0.9741752
```

```r
print('Test R^2')
```

```
## [1] "Test R^2"
```

```r
RSS = sum(test.pred$resid^2)
TSS = sum((test.pred$heating.load - mean(test.pred$heating.load))^2)
print(1 - RSS/TSS)
```

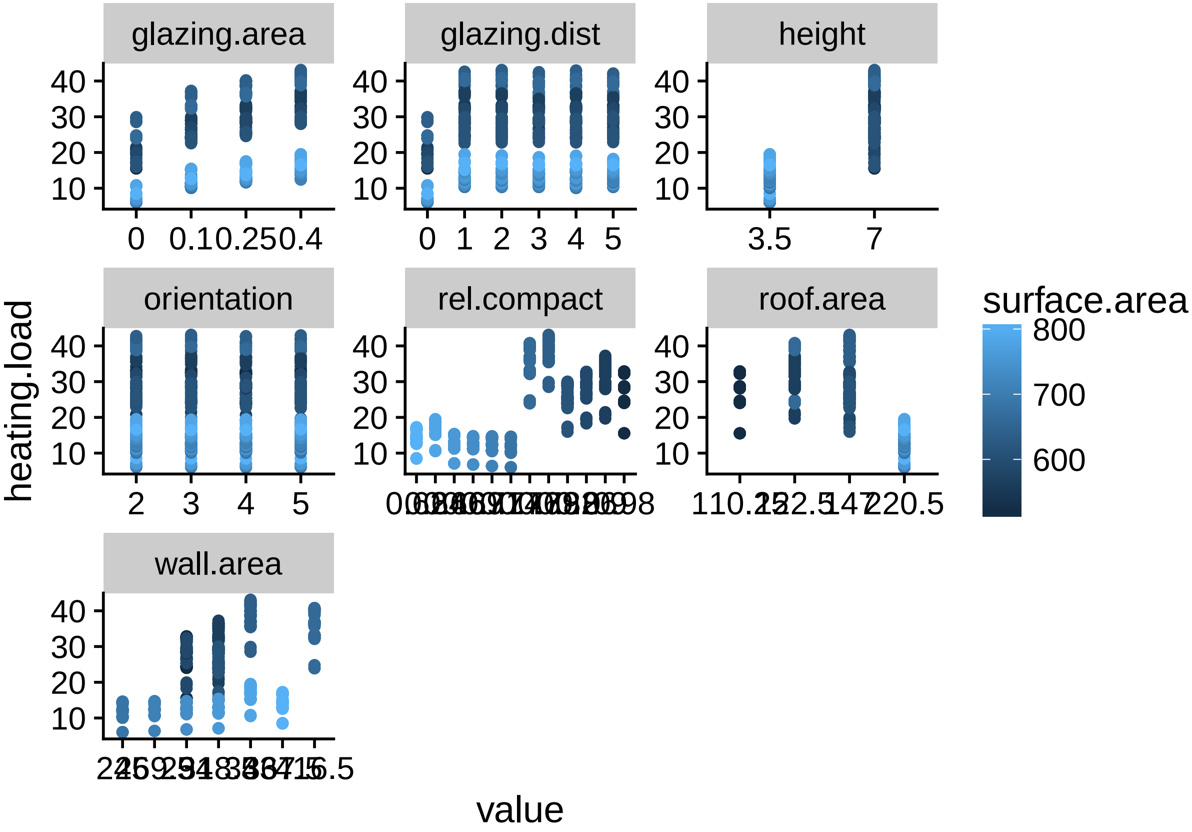
```
## [1] 0.9667713
```

```r
train.pred %>% filter(abs(resid.stand) > 3)
```

```
## # A tibble: 19 x 13
##    rel.compact surface.area wall.area roof.area height orientation
##          <dbl>        <dbl>     <dbl>     <dbl>  <dbl> <fct>
##  1       0.790         637.      343.      147.     7. 3
##  2       0.790         637.      343.      147.     7. 5
##  3       0.790         637.      343.      147.     7. 3
##  4       0.790         637.      343.      147.     7. 4
##  5       0.790         637.      343.      147.     7. 4
##  6       0.790         637.      343.      147.     7. 5
##  7       0.790         637.      343.      147.     7. 4
##  8       0.790         637.      343.      147.     7. 2
##  9       0.790         637.      343.      147.     7. 5
## 10       0.790         637.      343.      147.     7. 3
## 11       0.790         637.      343.      147.     7. 2
## 12       0.790         637.      343.      147.     7. 4
## 13       0.790         637.      343.      147.     7. 2
## 14       0.790         637.      343.      147.     7. 3
## 15       0.790         637.      343.      147.     7. 5
## 16       0.790         637.      343.      147.     7. 5
## 17       0.790         637.      343.      147.     7. 2
## 18       0.790         637.      343.      147.     7. 2
## 19       0.790         637.      343.      147.     7. 4
## # ... with 7 more variables: glazing.area <dbl>, glazing.dist <fct>,
## #   heating.load <dbl>, cooling.load <dbl>, pred <dbl>, resid.stand <dbl>,
## #   resid <dbl>
```

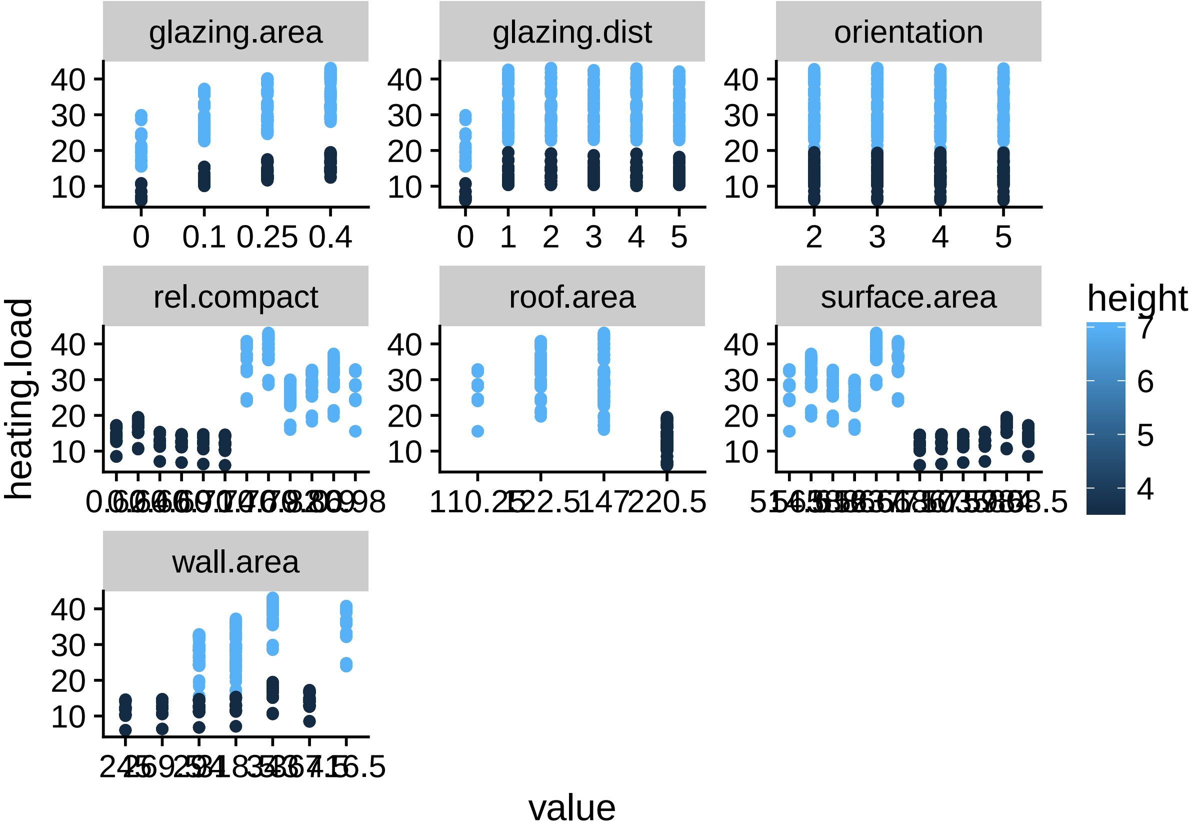
```
#energy %>%
#  filter(height == 3.5) %>%
#  keep(is.numeric) %>%
#  gather() %>%
#  ggplot(aes(value)) +
#    facet_wrap(~ key, scales = "free") +
#    geom_histogram()

energy %>%
  #filter(height == 7) %>%
  #dplyr::select(rel.compact, surface.area, wall.area, roof.area, height, heating.load) %>%
  #keep(is.numeric, height) %>%
  dplyr::select(-cooling.load) %>%
  gather(key='var', value='value',-surface.area, -heating.load) %>%
  ggplot(aes(x=value, y=heating.load, color=surface.area)) +
    facet_wrap(~ var, scales = "free") +
    geom_point()
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```



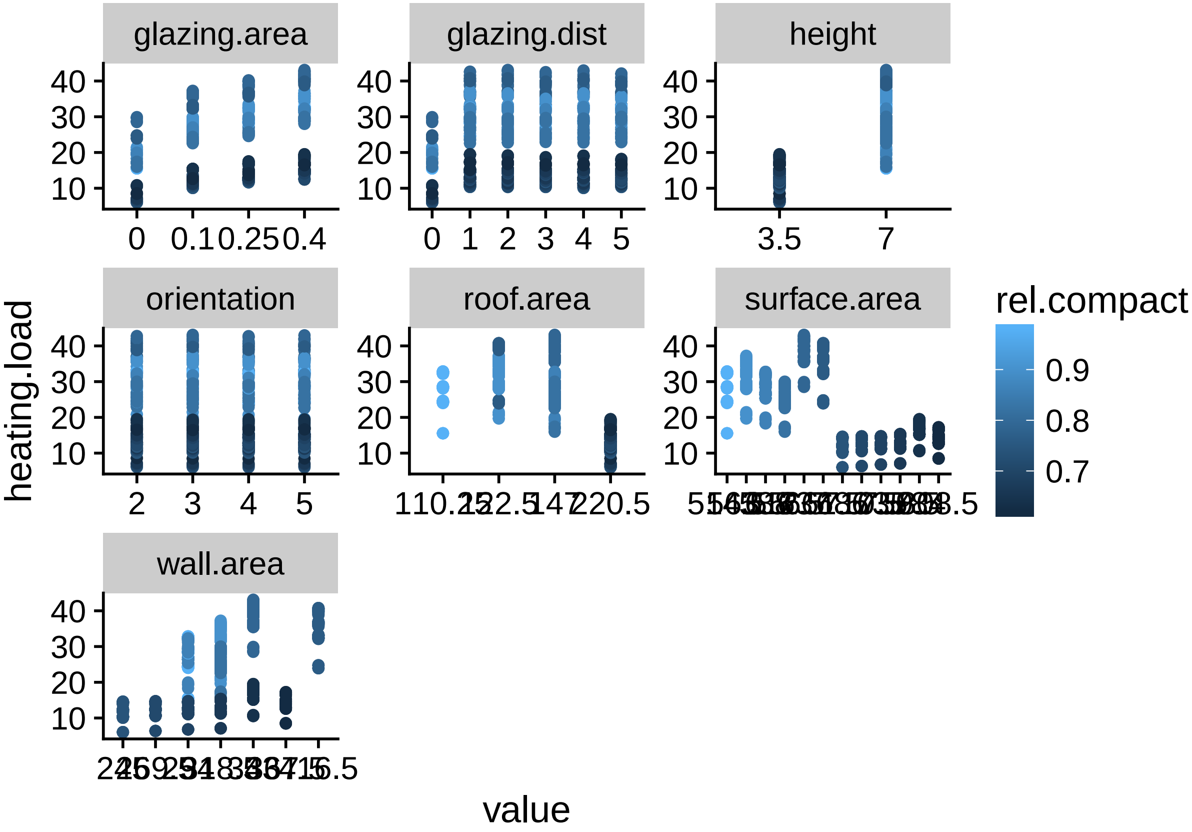
```
energy %>%
  #filter(height == 7) %>%
  #dplyr::select(rel.compact, surface.area, wall.area, roof.area, height, heating.load) %>%
  #keep(is.numeric, height) %>%
  dplyr::select(-cooling.load) %>%
  gather(key='var', value='value',-height, -heating.load) %>%
  ggplot(aes(x=value, y=heating.load, color=height)) +
    facet_wrap(~ var, scales = "free") +
    geom_point()
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```



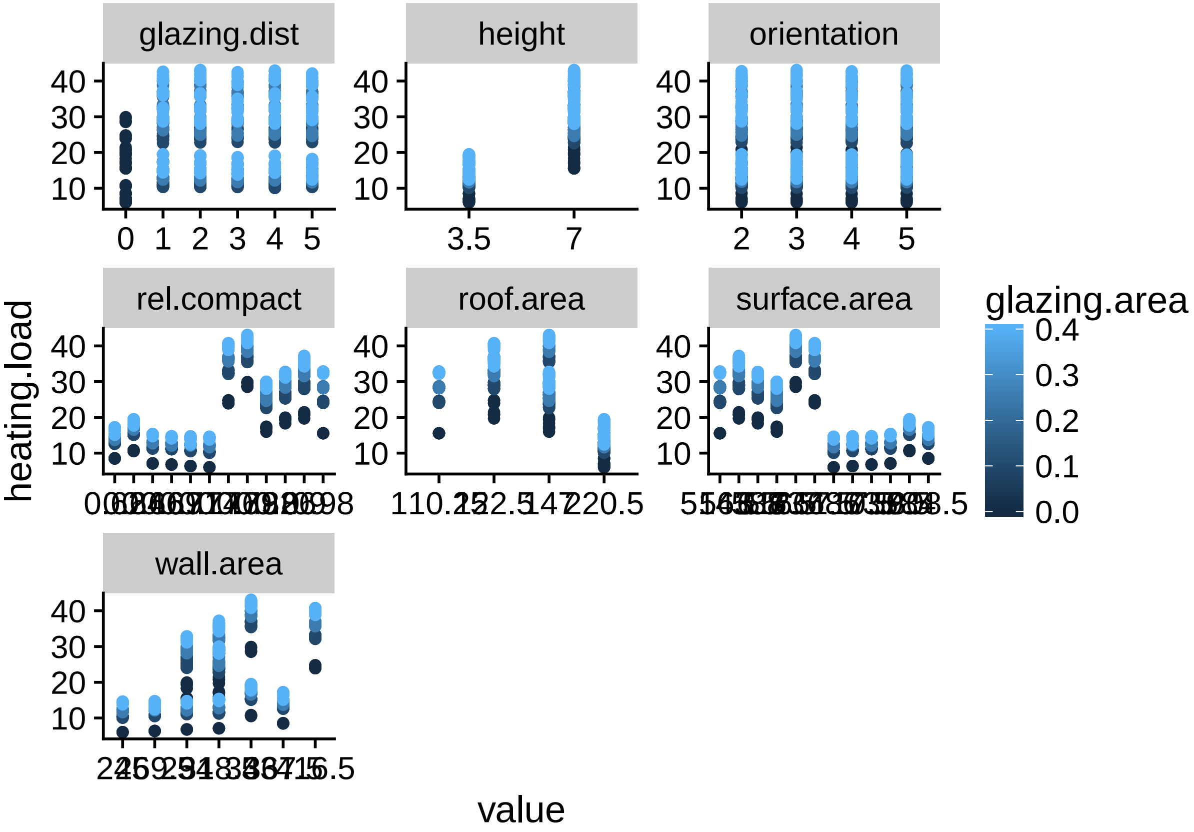
```r
energy %>%
  #filter(height == 7) %>%
  #dplyr::select(rel.compact, surface.area, wall.area, roof.area, height, heating.load) %>%
  #keep(is.numeric, height) %>%
  dplyr::select(-cooling.load) %>%
  gather(key='var', value='value',-rel.compact, -heating.load) %>%
  ggplot(aes(x=value, y=heating.load, color=rel.compact)) +
    facet_wrap(~ var, scales = "free") +
    geom_point()
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```



```
energy %>%
  #filter(height == 7) %>%
  #dplyr::select(rel.compact, surface.area, wall.area, roof.area, height, heating.load) %>%
  #keep(is.numeric, height) %>%
  dplyr::select(-cooling.load) %>%
  gather(key='var', value='value',-glazing.area, -heating.load) %>%
  ggplot(aes(x=value, y=heating.load, color=glazing.area)) +
    facet_wrap(~ var, scales = "free") +
    geom_point()
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```



```
train
```

```
## # A tibble: 614 x 10
##    rel.compact surface.area wall.area roof.area height orientation
##          <dbl>        <dbl>     <dbl>     <dbl>  <dbl> <fct>
##  1       0.620         808.      368.      220.   3.50 3
##  2       0.690         735.      294.      220.   3.50 3
##  3       0.820         612.      318.      147.   7.00 5
##  4       0.760         662.      416.      122.   7.00 5
##  5       0.620         808.      368.      220.   3.50 4
##  6       0.660         760.      318.      220.   3.50 2
##  7       0.820         612.      318.      147.   7.00 3
##  8       0.820         612.      318.      147.   7.00 3
##  9       0.640         784.      343.      220.   3.50 2
## 10       0.710         710.      270.      220.   3.50 4
## # ... with 604 more rows, and 4 more variables: glazing.area <dbl>,
## #   glazing.dist <fct>, heating.load <dbl>, cooling.load <dbl>
```

```
cor(train %>% keep(is.numeric))
```

```
##                rel.compact surface.area   wall.area  roof.area      height
## rel.compact    1.000000000 -0.991883739 -0.21278612 -0.8673495  0.82786322
## surface.area  -0.991883739  1.000000000  0.20579298  0.8787073 -0.85734923
## wall.area     -0.212786118  0.205792980  1.00000000 -0.2863115  0.27377177
## roof.area     -0.867349467  0.878707313 -0.28631149  1.0000000 -0.97297154
## height         0.827863219 -0.857349228  0.27377177 -0.9729715  1.00000000
## glazing.area  -0.007327639  0.008882311 -0.01203682  0.0145682 -0.01158195
## heating.load   0.625207871 -0.658820003  0.44727176 -0.8632265  0.88886607
## cooling.load   0.635337392 -0.672263397  0.41967145 -0.8629254  0.89350522
##              glazing.area heading.load cooling.load
```

```
##              glazing.area heating.load cooling.load
## rel.compact  -0.007327639    0.6252079    0.6353374
## surface.area  0.008882311   -0.6588200   -0.6722634
## wall.area    -0.012036823    0.4472718    0.4196715
## roof.area     0.014568201   -0.8632265   -0.8629254
## height       -0.011581947    0.8888661    0.8935052
## glazing.area  1.000000000    0.2634072    0.2027239
## heating.load  0.263407173    1.0000000    0.9765619
## cooling.load  0.202723944    0.9765619    1.0000000
```

```
df <- train

model.interact <- lm(heating.load ~ . - cooling.load +
                        #wall.area*(glazing.area + glazing.dist) +
                        surface.area*(rel.compact + height + roof.area + wall.area) +
                        rel.compact*(height + wall.area), data=df)

#model2 <- lm(heating.load ~ (rel.compact + surface.area + wall.area + roof.area + height + orientation

anova(model, model.interact)
```

```
## Analysis of Variance Table
##
## Model 1: heating.load ~ rel.compact + surface.area + wall.area + roof.area +
##     height + orientation + glazing.area + glazing.dist + cooling.load
## Model 2: heating.load ~ (rel.compact + surface.area + wall.area + roof.area +
##     height + orientation + glazing.area + glazing.dist + cooling.load) -
##     cooling.load + surface.area * (rel.compact + height + roof.area +
##     wall.area) + rel.compact * (height + wall.area)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    599 1577.77
## 2    594  891.27  5     686.5 91.506 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

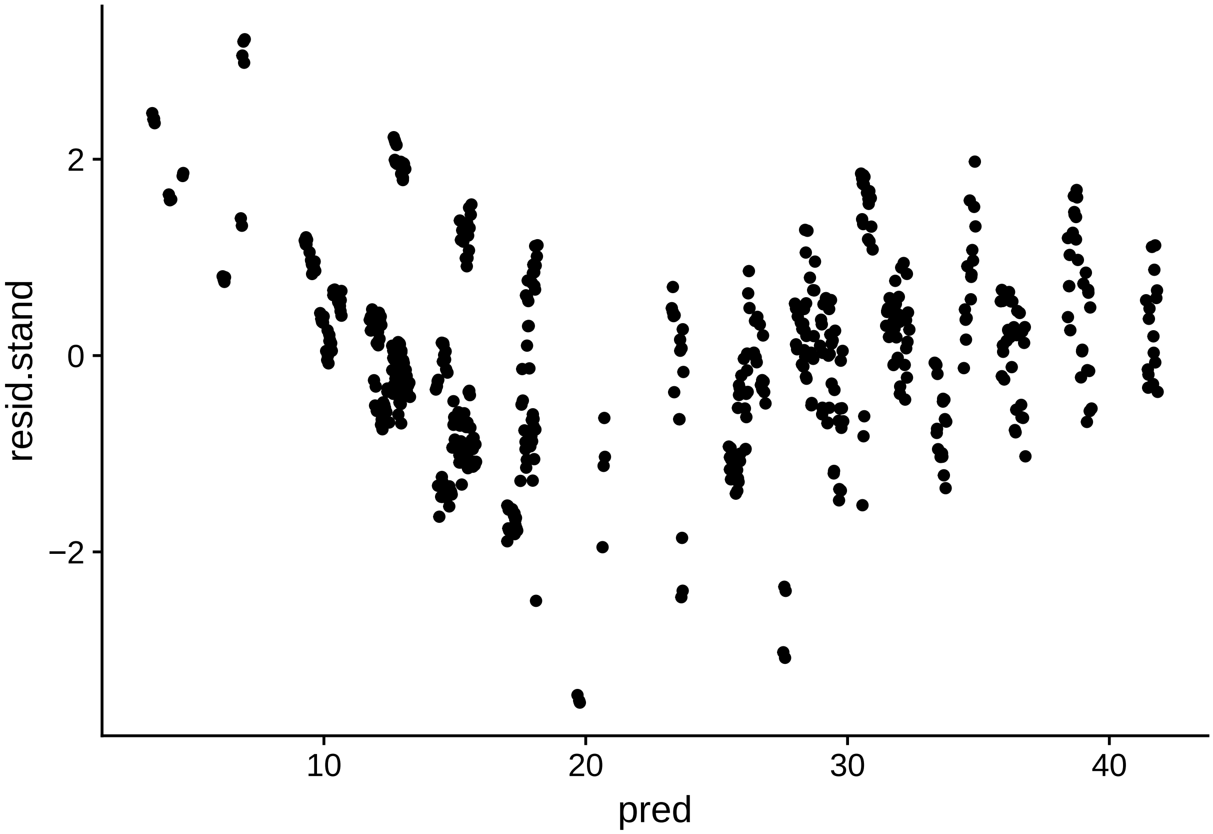
```
#reduced <- step(model.interact)
summary(model.interact)
```

```
##
## Call:
## lm(formula = heating.load ~ . - cooling.load + surface.area *
##     (rel.compact + height + roof.area + wall.area) + rel.compact *
##     (height + wall.area), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2258 -0.7974 -0.0077  0.6890  3.8739
##
## Coefficients: (1 not defined because of singularities)
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                5.100e+04  2.433e+03  20.966  < 2e-16 ***
## rel.compact               -2.849e+04  1.329e+03 -21.434  < 2e-16 ***
## surface.area              -1.308e+02  6.756e+00 -19.358  < 2e-16 ***
## wall.area                  9.328e+01  5.523e+00  16.890  < 2e-16 ***
## roof.area                         NA         NA      NA       NA
## height                   -4.346e+03  1.891e+02 -22.986  < 2e-16 ***
## orientation3             -2.361e-02  1.396e-01  -0.169    0.866
## orientation4             -5.042e-02  1.409e-01  -0.358    0.721
## orientation5             -9.091e-02  1.384e-01  -0.657    0.512
## glazing.area              1.683e+01  4.207e-01  40.016  < 2e-16 ***
## glazing.dist1             4.450e+00  2.472e-01  18.001  < 2e-16 ***
## glazing.dist2             4.380e+00  2.469e-01  17.740  < 2e-16 ***
## glazing.dist3             4.098e+00  2.456e-01  16.686  < 2e-16 ***
## glazing.dist4             4.324e+00  2.490e-01  17.368  < 2e-16 ***
## glazing.dist5             4.139e+00  2.495e-01  16.590  < 2e-16 ***
## rel.compact:surface.area  4.861e+01  2.596e+00  18.723  < 2e-16 ***
## surface.area:height       3.791e+00  1.644e-01  23.056  < 2e-16 ***
## surface.area:roof.area    1.495e-01  7.973e-03  18.752  < 2e-16 ***
## surface.area:wall.area   -4.951e-03  6.823e-04  -7.256 1.25e-12 ***
## rel.compact:height        2.455e+03  1.062e+02  23.113  < 2e-16 ***
## rel.compact:wall.area    -5.384e+01  3.262e+00 -16.508  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.225 on 594 degrees of freedom
## Multiple R-squared:  0.9854, Adjusted R-squared:  0.9849
## F-statistic:  2112 on 19 and 594 DF,  p-value: < 2.2e-16
```

```
train.pred.interact <- train %>%
  mutate(pred = predict.lm(model.interact, train),
         resid.stand = rstandard(model.interact),
         resid = resid(model.interact))
```

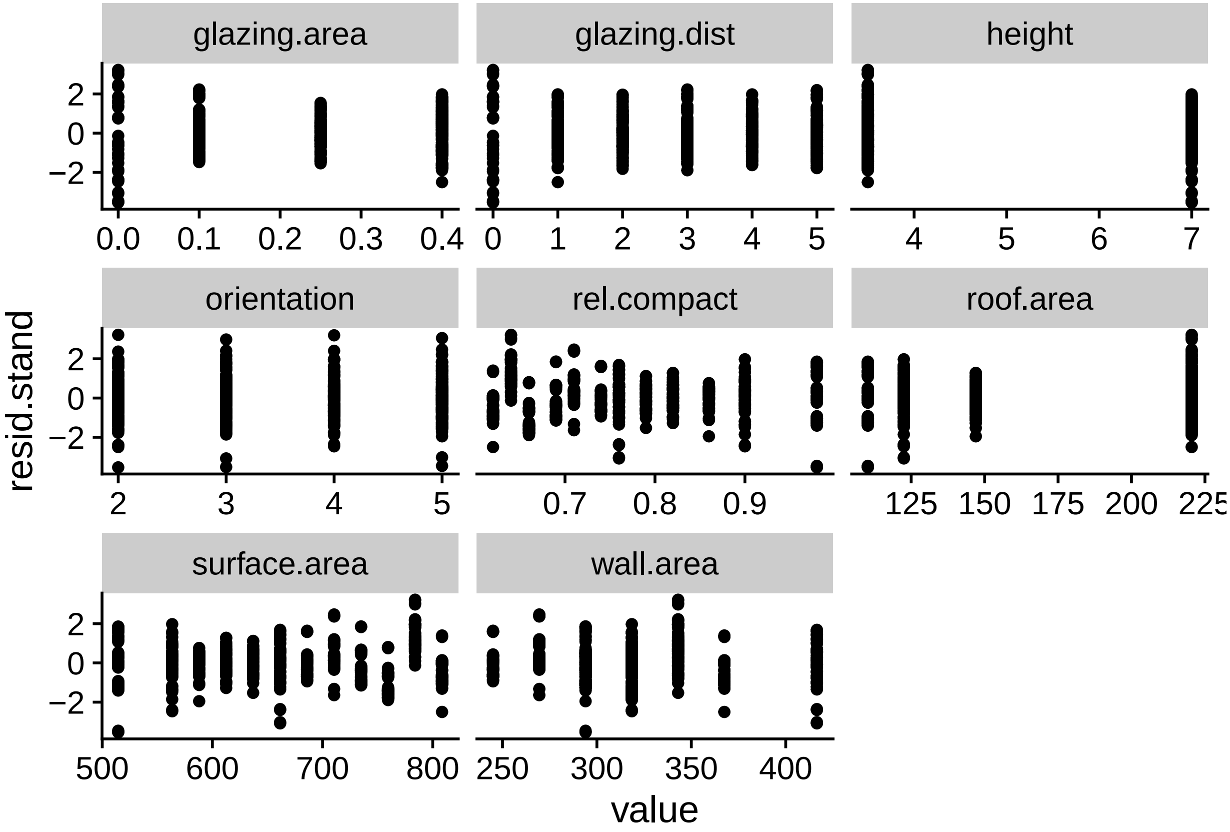
```
## Warning in predict.lm(model.interact, train): prediction from a rank-
## deficient fit may be misleading
```

```
train.pred.interact %>%
  ggplot(aes(x=pred, y=resid.stand)) + geom_point()
```



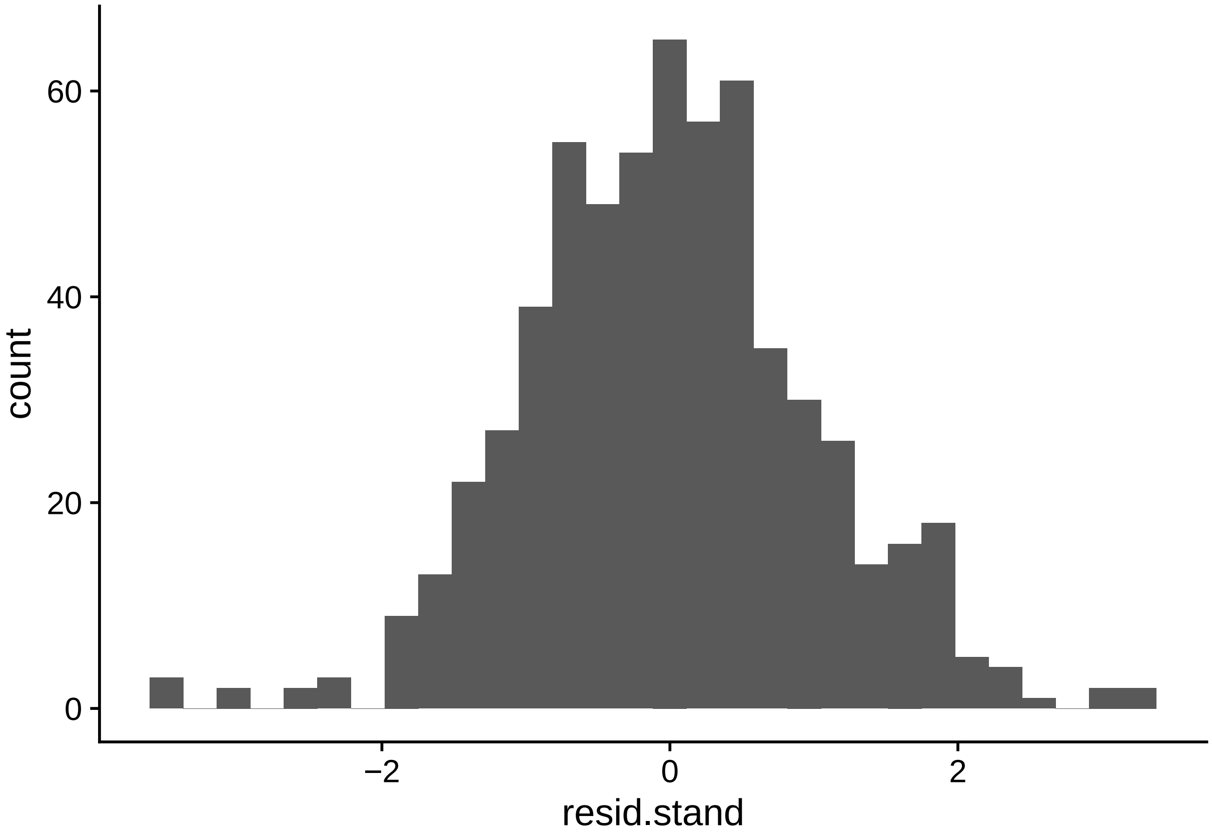
```
train.pred.interact %>%
  #dplyr::select(glazing.dist, rel.compact, surface.area,  resid) %>%
  dplyr::select(-pred, -heating.load, -cooling.load, -resid) %>%
  gather(key="var", value="value", -resid.stand) %>%
  mutate(value = as.numeric(value)) %>%
  ggplot(aes(x=value, y=resid.stand)) +
  geom_point() +
  #geom_jitter(size=1) +
  #geom_boxplot() +
  facet_wrap( ~ var, ncol=3, scales = 'free_x')
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

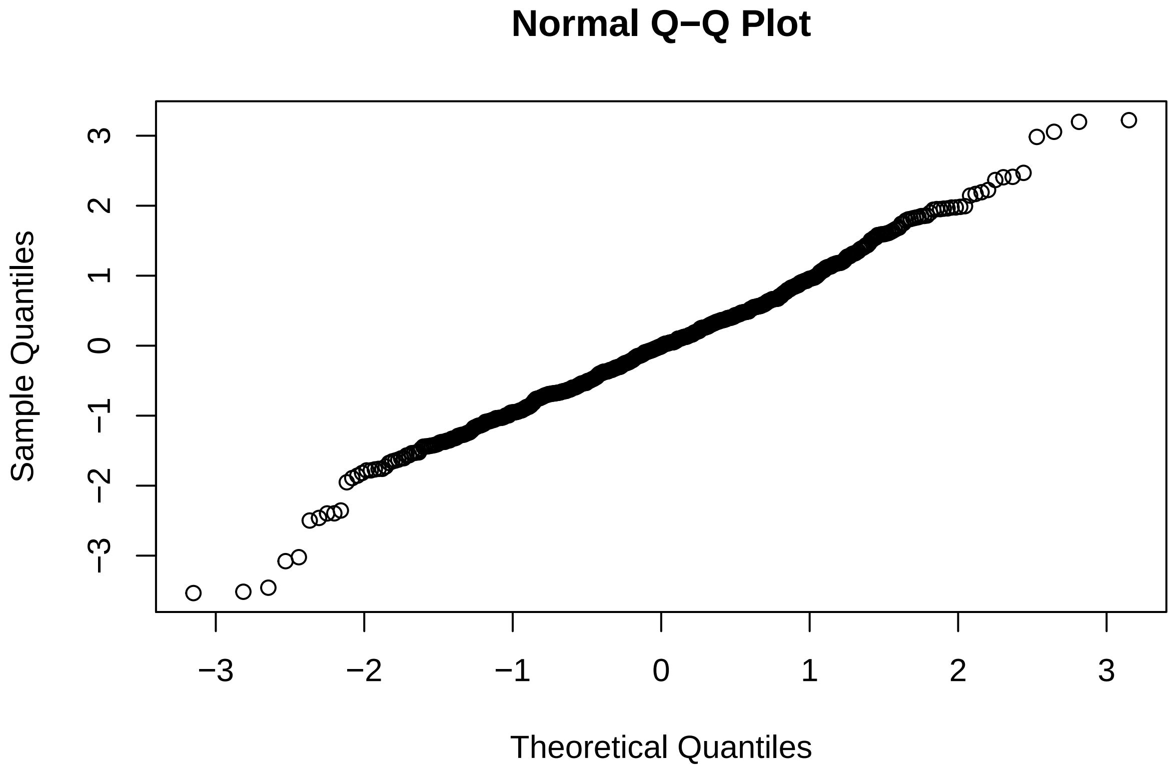


```
train.pred.interact %>%
  ggplot(aes(x=resid.stand)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```r
qqnorm(train.pred.interact$resid.stand)
```



```
test.pred.interact <- test %>%
  mutate(pred = predict.lm(model.interact, test)) %>%
  mutate(resid = pred - heating.load,
         resid.stand = ((pred - heating.load) - mean(pred - heating.load)) / sd(pred - heating.load))
```

```
## Warning in predict.lm(model.interact, test): prediction from a rank-
## deficient fit may be misleading
```

```
print('MSE Train')
```

```
## [1] "MSE Train"
```

```
print(mean(train.pred.interact$resid^2))
```

```
## [1] 1.451573
```

```
print('MSE Val')
```

```
## [1] "MSE Val"
```

```
print(mean(test.pred.interact$resid^2))
```

```
## [1] 1.577603
```

```
print('Train R^2')
```

```
## [1] "Train R^2"
```

```
RSS = sum(train.pred.interact$resid^2)
TSS = sum((train.pred.interact$heating.load - mean(train.pred.interact$heating.load))^2)
print(1 - RSS/TSS)
```

```
## [1] 0.9854118
```

```
print('Test R^2')
```

```
## [1] "Test R^2"
```

```
RSS = sum(test.pred.interact$resid^2)
TSS = sum((test.pred.interact$heating.load - mean(test.pred.interact$heating.load))^2)
print(1 - RSS/TSS)
```

```
## [1] 0.9856484
```

```
train.pred.interact %>% filter(abs(resid.stand) > 3)
```

```
## # A tibble: 8 x 13
##   rel.compact surface.area wall.area roof.area height orientation
##         <dbl>        <dbl>     <dbl>     <dbl>  <dbl> <fct>
## 1       0.760         662.      416.      122.   7.00 3
## 2       0.640         784.      343.      220.   3.50 4
## 3       0.980         514.      294.      110.   7.00 3
## 4       0.640         784.      343.      220.   3.50 2
## 5       0.640         784.      343.      220.   3.50 5
## 6       0.760         662.      416.      122.   7.00 5
## 7       0.980         514.      294.      110.   7.00 2
## 8       0.980         514.      294.      110.   7.00 5
## # ... with 7 more variables: glazing.area <dbl>, glazing.dist <fct>,
## #   heating.load <dbl>, cooling.load <dbl>, pred <dbl>, resid.stand <dbl>,
## #   resid <dbl>
```

Part III

```
df <- train

model.interact2 <- lm(heating.load ~ . - cooling.load +
                        wall.area*(glazing.area + glazing.dist) +
                        surface.area*(rel.compact + height + roof.area + wall.area) +
                        rel.compact*(height + wall.area), data=df)

#model2 <- lm(heating.load ~ (rel.compact + surface.area + wall.area + roof.area + height + orientation

anova(model.interact, model.interact2)
```

```
## Analysis of Variance Table
##
## Model 1: heating.load ~ (rel.compact + surface.area + wall.area + roof.area +
##     height + orientation + glazing.area + glazing.dist + cooling.load) -
##     cooling.load + surface.area * (rel.compact + height + roof.area +
##     wall.area) + rel.compact * (height + wall.area)
## Model 2: heating.load ~ (rel.compact + surface.area + wall.area + roof.area +
##     height + orientation + glazing.area + glazing.dist + cooling.load) -
##     cooling.load + wall.area * (glazing.area + glazing.dist) +
##     surface.area * (rel.compact + height + roof.area + wall.area) +
##     rel.compact * (height + wall.area)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    594 891.27
## 2    588 838.13  6    53.137 6.2131 2.446e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

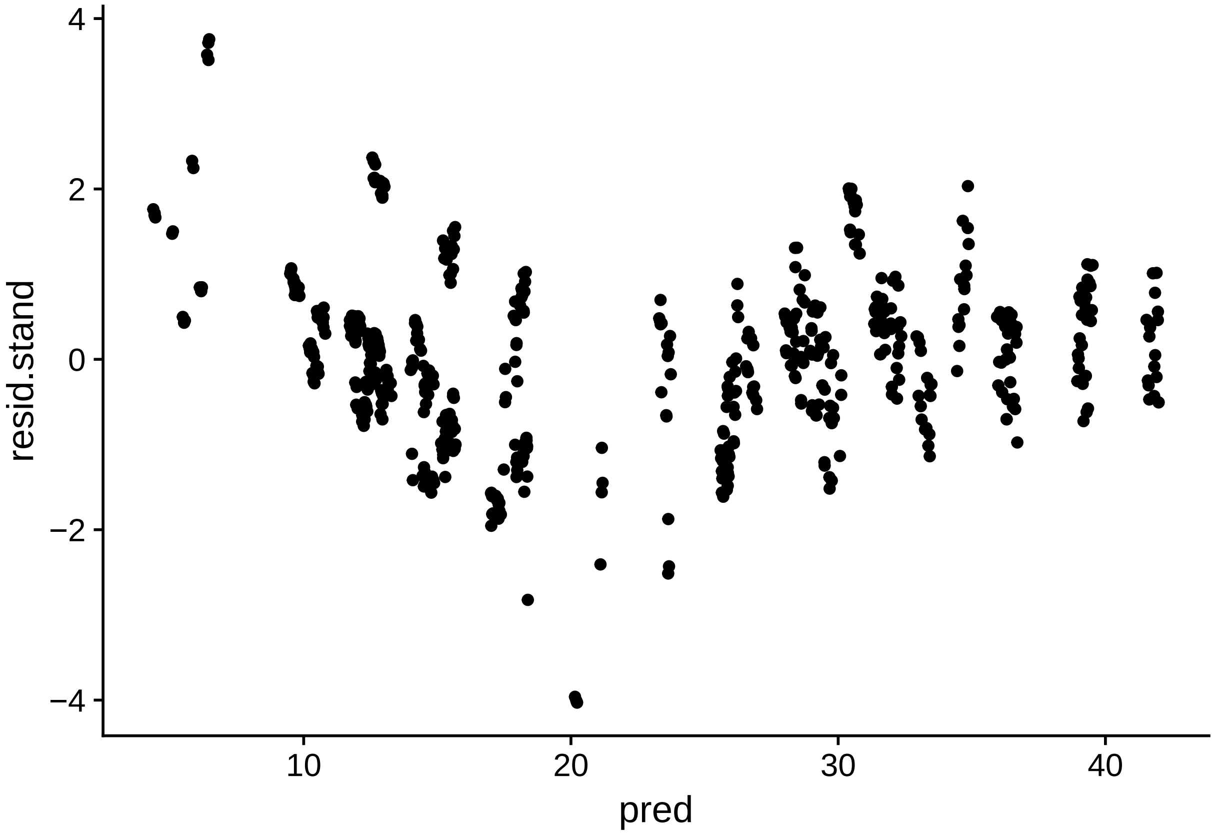
```
#reduced <- step(model.interact)
summary(model.interact2)
```

```
##
## Call:
## lm(formula = heating.load ~ . - cooling.load + wall.area * (glazing.area +
##     glazing.dist) + surface.area * (rel.compact + height + roof.area +
##     wall.area) + rel.compact * (height + wall.area), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6800 -0.7025  0.0435  0.6108  4.3852
##
## Coefficients: (1 not defined because of singularities)
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               5.082e+04  2.374e+03  21.410  < 2e-16 ***
## rel.compact              -2.839e+04  1.297e+03 -21.887  < 2e-16 ***
## surface.area             -1.303e+02  6.592e+00 -19.771  < 2e-16 ***
## wall.area                 9.300e+01  5.388e+00  17.260  < 2e-16 ***
## roof.area                        NA         NA      NA       NA
## height                   -4.332e+03  1.845e+02 -23.485  < 2e-16 ***
## orientation3             -2.703e-02  1.362e-01  -0.198 0.842773
## orientation4             -3.150e-02  1.375e-01  -0.229 0.818832
## orientation5             -7.873e-02  1.350e-01  -0.583 0.559904
## glazing.area              6.135e+00  2.978e+00   2.060 0.039844 *
## glazing.dist1             4.910e-01  1.701e+00   0.289 0.772987
## glazing.dist2             2.626e-02  1.739e+00   0.015 0.987958
## glazing.dist3             4.360e-01  1.712e+00   0.255 0.799081
## glazing.dist4             3.042e-01  1.724e+00   0.176 0.859990
## glazing.dist5             2.306e-01  1.734e+00   0.133 0.894231
## wall.area:glazing.area    3.351e-02  9.241e-03   3.627 0.000312 ***
## wall.area:glazing.dist1   1.255e-02  5.287e-03   2.374 0.017941 *
## wall.area:glazing.dist2   1.376e-02  5.424e-03   2.536 0.011469 *
## wall.area:glazing.dist3   1.162e-02  5.335e-03   2.177 0.029845 *
## wall.area:glazing.dist4   1.270e-02  5.376e-03   2.362 0.018521 *
## wall.area:glazing.dist5   1.238e-02  5.378e-03   2.301 0.021720 *
## rel.compact:surface.area  4.845e+01  2.533e+00  19.128  < 2e-16 ***
## surface.area:height       3.779e+00  1.604e-01  23.556  < 2e-16 ***
## surface.area:roof.area    1.490e-01  7.779e-03  19.154  < 2e-16 ***
## surface.area:wall.area   -4.978e-03  6.656e-04  -7.479 2.73e-13 ***
## rel.compact:height        2.447e+03  1.036e+02  23.616  < 2e-16 ***
## rel.compact:wall.area    -5.370e+01  3.182e+00 -16.874  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.194 on 588 degrees of freedom
## Multiple R-squared:  0.9863, Adjusted R-squared:  0.9857
## F-statistic:  1691 on 25 and 588 DF,  p-value: < 2.2e-16
```

```r
train.pred.interact2 <- train %>%
  mutate(pred = predict.lm(model.interact2, train),
         resid.stand = rstandard(model.interact2),
         resid = resid(model.interact2))
```

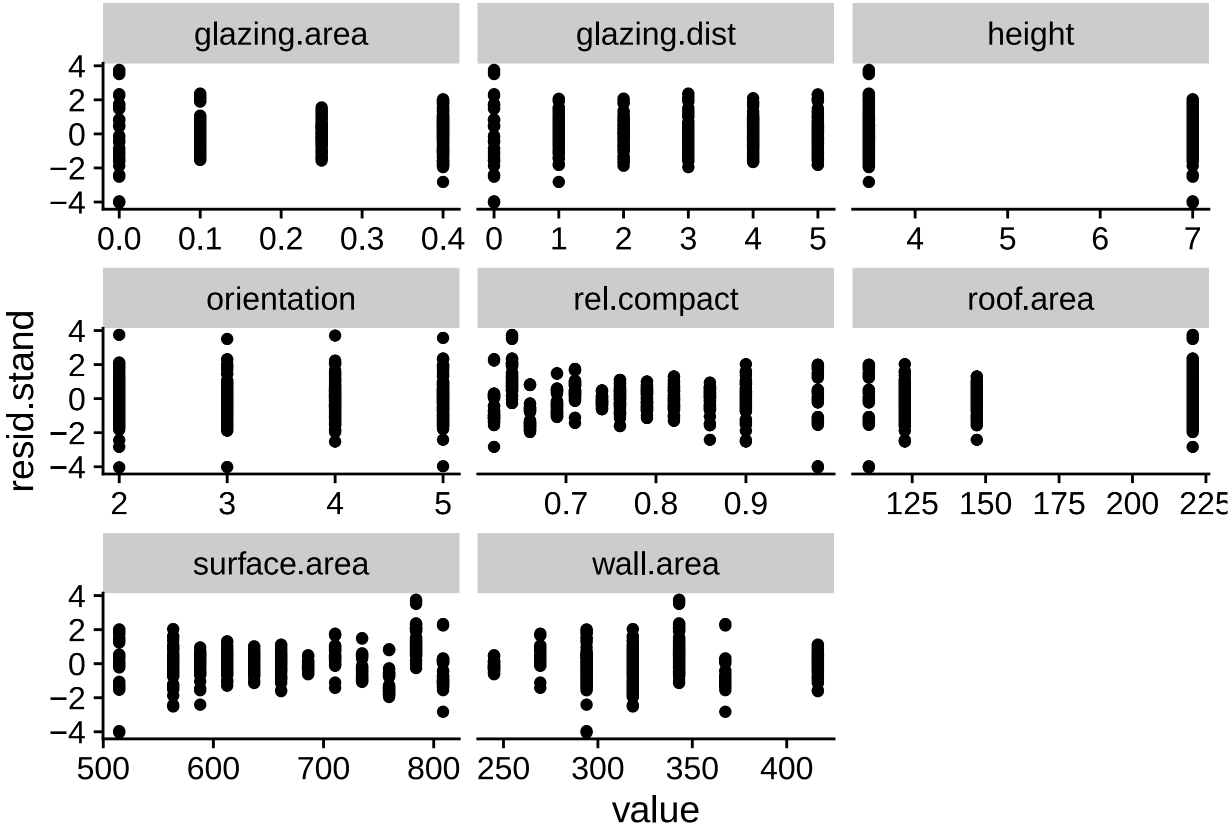
```
## Warning in predict.lm(model.interact2, train): prediction from a rank-
## deficient fit may be misleading
```

```r
train.pred.interact2 %>%
  ggplot(aes(x=pred, y=resid.stand)) + geom_point()
```



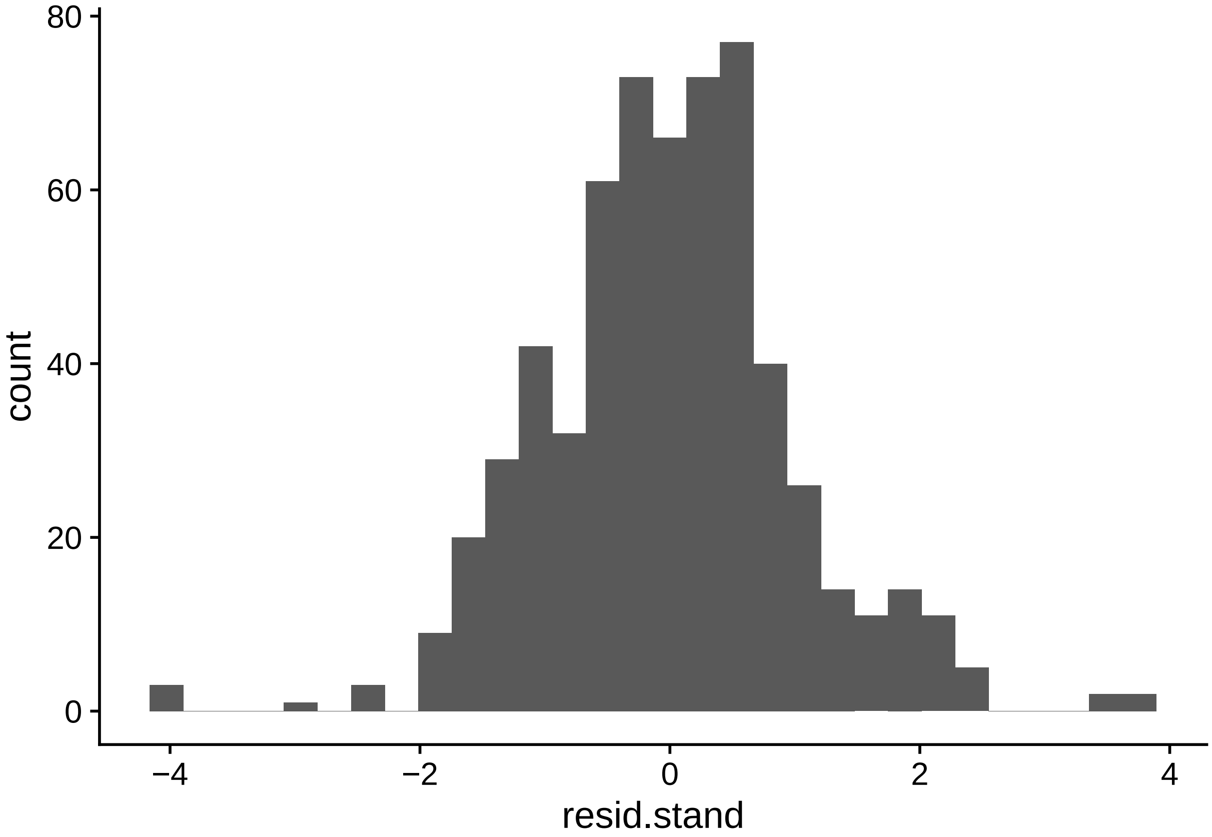
```
train.pred.interact2 %>%
  #dplyr::select(glazing.dist, rel.compact, surface.area,  resid) %>%
  dplyr::select(-pred, -heating.load, -cooling.load, -resid) %>%
  gather(key="var", value="value", -resid.stand) %>%
  mutate(value = as.numeric(value)) %>%
  ggplot(aes(x=value, y=resid.stand)) +
  geom_point() +
  #geom_jitter(size=1) +
  #geom_boxplot() +
  facet_wrap( ~ var, ncol=3, scales = 'free_x')
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

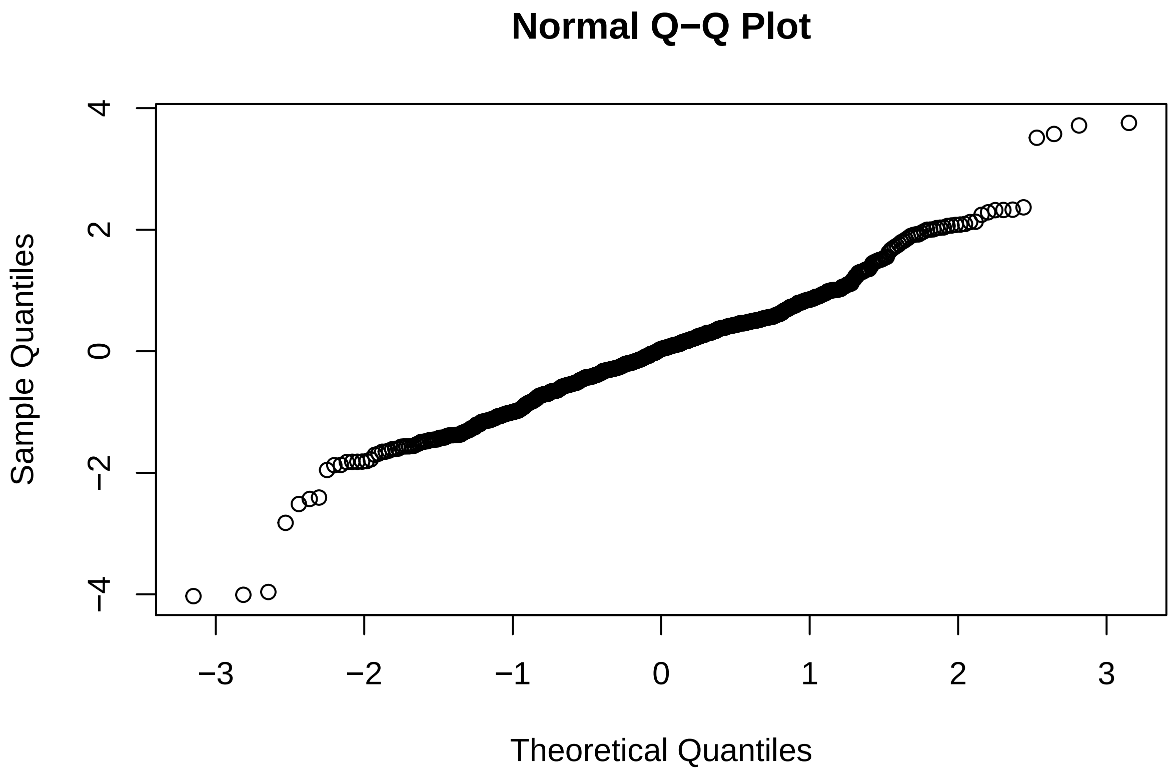


```
train.pred.interact2 %>%
  ggplot(aes(x=resid.stand)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```r
qqnorm(train.pred.interact2$resid.stand)
```



```
test.pred.interact2 <- test %>%
  mutate(pred = predict.lm(model.interact2, test)) %>%
  mutate(resid = pred - heating.load,
         resid.stand = ((pred - heating.load) - mean(pred - heating.load)) / sd(pred - heating.load))
```

```
## Warning in predict.lm(model.interact2, test): prediction from a rank-
## deficient fit may be misleading
```

```
print('MSE Train')
```

```
## [1] "MSE Train"
```

```
print(mean(train.pred.interact2$resid^2))
```

```
## [1] 1.365031
```

```
print('MSE Val')
```

```
## [1] "MSE Val"
```

```
print(mean(test.pred.interact2$resid^2))
```

```
## [1] 1.539967
```

```
print('Train R^2')
```

```
## [1] "Train R^2"
```

```
RSS = sum(train.pred.interact2$resid^2)
TSS = sum((train.pred.interact2$heating.load - mean(train.pred.interact2$heating.load))^2)
print(1 - RSS/TSS)
```

```
## [1] 0.9862815
```

```
print('Test R^2')
```

```
## [1] "Test R^2"
```

```
RSS = sum(test.pred.interact2$resid^2)
TSS = sum((test.pred.interact2$heating.load - mean(test.pred.interact2$heating.load))^2)
print(1 - RSS/TSS)
```

```
## [1] 0.9859908
```

```
train.pred.interact2 %>% filter(abs(resid.stand) > 3)
```

```
## # A tibble: 7 x 13
##   rel.compact surface.area wall.area roof.area height orientation
##         <dbl>        <dbl>     <dbl>     <dbl>  <dbl> <fct>
## 1       0.640         784.      343.      220.   3.50 3
## 2       0.640         784.      343.      220.   3.50 4
## 3       0.980         514.      294.      110.   7.00 3
## 4       0.640         784.      343.      220.   3.50 2
## 5       0.640         784.      343.      220.   3.50 5
## 6       0.980         514.      294.      110.   7.00 2
## 7       0.980         514.      294.      110.   7.00 5
## # ... with 7 more variables: glazing.area <dbl>, glazing.dist <fct>,
## #   heating.load <dbl>, cooling.load <dbl>, pred <dbl>, resid.stand <dbl>,
## #   resid <dbl>
```

Part IV

INCLUDE

```
df <- train %>%
  mutate(glazing.dist = relevel(glazing.dist, ref="5"))

model.interact3 <- lm(heating.load ~ . - cooling.load +
                        wall.area*(roof.area + glazing.area + glazing.dist) +
                        surface.area*(rel.compact + height + roof.area + wall.area  + glazing.area + gla
                        rel.compact*(height + wall.area), data=df)

#model2 <- lm(heating.load ~ (rel.compact + surface.area + wall.area + roof.area + height + orientation

anova(model.interact2, model.interact3)
```

```
## Analysis of Variance Table
##
## Model 1: heating.load ~ (rel.compact + surface.area + wall.area + roof.area +
##     height + orientation + glazing.area + glazing.dist + cooling.load) -
##     cooling.load + wall.area * (glazing.area + glazing.dist) +
##     surface.area * (rel.compact + height + roof.area + wall.area) +
##     rel.compact * (height + wall.area)
## Model 2: heating.load ~ (rel.compact + surface.area + wall.area + roof.area +
##     height + orientation + glazing.area + glazing.dist + cooling.load) -
##     cooling.load + wall.area * (roof.area + glazing.area + glazing.dist) +
##     surface.area * (rel.compact + height + roof.area + wall.area +
##         glazing.area + glazing.dist) + rel.compact * (height +
##     wall.area)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    588 838.13
## 2    582 404.78  6    433.35 103.85 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
reduced <- step(model.interact3)
```

```
## Start:  AIC=-191.83
## heating.load ~ (rel.compact + surface.area + wall.area + roof.area +
##     height + orientation + glazing.area + glazing.dist + cooling.load) -
##     cooling.load + wall.area * (roof.area + glazing.area + glazing.dist) +
##     surface.area * (rel.compact + height + roof.area + wall.area +
##         glazing.area + glazing.dist) + rel.compact * (height +
##     wall.area)
##
##
## Step:  AIC=-191.83
## heating.load ~ rel.compact + surface.area + wall.area + roof.area +
##     height + orientation + glazing.area + glazing.dist + wall.area:roof.area +
##     wall.area:glazing.area + wall.area:glazing.dist + rel.compact:surface.area +
##     surface.area:height + surface.area:roof.area + surface.area:wall.area +
##     surface.area:glazing.area + surface.area:glazing.dist + rel.compact:height
##
##                               Df Sum of Sq    RSS      AIC
## - orientation                3      1.04 405.82 -196.255
## <none>                                   404.78 -191.828
## - wall.area:glazing.dist     5     11.93 416.71 -183.992
## - rel.compact:height         1     34.29 439.07 -143.902
## - surface.area:glazing.dist  5     40.95 445.73 -142.658
## - wall.area:glazing.area     1     55.27 460.05 -115.239
## - surface.area:wall.area     1     78.57 483.35  -84.900
## - rel.compact:surface.area   1     80.35 485.12  -82.653
## - surface.area:roof.area     1    203.07 607.85   55.819
## - surface.area:height        1    215.32 620.09   68.062
## - surface.area:glazing.area  1    218.15 622.92   70.859
## - wall.area:roof.area        1    395.60 800.38  224.768
##
## Step:  AIC=-196.25
## heating.load ~ rel.compact + surface.area + wall.area + roof.area +
##     height + glazing.area + glazing.dist + wall.area:roof.area +
##     wall.area:glazing.area + wall.area:glazing.dist + rel.compact:surface.area +
##     surface.area:height + surface.area:roof.area + surface.area:wall.area +
##     surface.area:glazing.area + surface.area:glazing.dist + rel.compact:height
##
##                             Df Sum of Sq    RSS      AIC
## <none>                                   405.82 -196.255
## - wall.area:glazing.dist     5     11.99 417.80 -188.382
## - rel.compact:height         1     34.53 440.35 -148.113
## - surface.area:glazing.dist  5     40.66 446.48 -147.624
## - wall.area:glazing.area     1     55.52 461.34 -119.519
## - surface.area:wall.area     1     78.53 484.35  -89.638
## - rel.compact:surface.area   1     80.29 486.10  -87.413
## - surface.area:roof.area     1    203.09 608.91   50.887
## - surface.area:height        1    216.06 621.87   63.824
## - surface.area:glazing.area  1    218.16 623.98   65.895
## - wall.area:roof.area        1    395.68 801.49  219.618
```

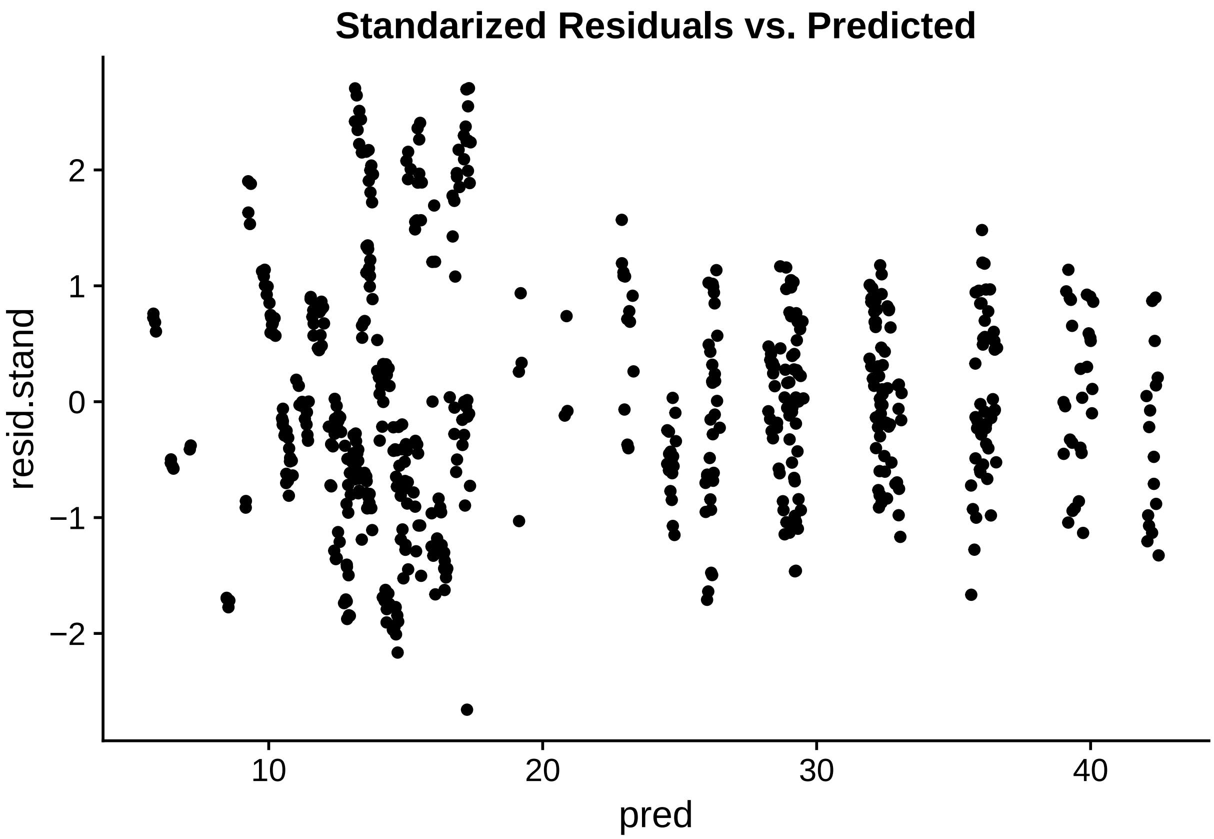
```
summary(reduced)
```

```
##
## Call:
## lm(formula = heating.load ~ rel.compact + surface.area + wall.area +
##     roof.area + height + glazing.area + glazing.dist + wall.area:roof.area +
##     wall.area:glazing.area + wall.area:glazing.dist + rel.compact:surface.area +
##     surface.area:height + surface.area:roof.area + surface.area:wall.area +
##     surface.area:glazing.area + surface.area:glazing.dist + rel.compact:height,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0897 -0.5421 -0.1067  0.5667  2.2698
##
## Coefficients: (1 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -3.307e+03  6.643e+02  -4.978 8.46e-07 ***
## rel.compact                 3.147e+03  4.466e+02   7.046 5.20e-12 ***
## surface.area                1.793e+01  1.635e+00  10.964  < 2e-16 ***
## wall.area                  -1.021e+01  5.177e-01 -19.718  < 2e-16 ***
## roof.area                          NA         NA      NA       NA
## height                      -6.623e+02  3.566e+01 -18.572  < 2e-16 ***
## glazing.area                 3.708e+01  2.714e+00  13.660  < 2e-16 ***
## glazing.dist0               -9.623e+00  1.661e+00  -5.793 1.13e-08 ***
## glazing.dist1               -5.659e-01  1.084e+00  -0.522 0.601978    
## glazing.dist2               -1.611e+00  1.077e+00  -1.496 0.135242    
## glazing.dist3               -6.769e-01  1.058e+00  -0.640 0.522463    
## glazing.dist4               -1.021e+00  1.077e+00  -0.948 0.343498    
## wall.area:roof.area          4.328e-02  1.812e-03  23.883  < 2e-16 ***
## wall.area:glazing.area       5.907e-02  6.603e-03   8.946  < 2e-16 ***
## wall.area:glazing.dist0     -1.387e-02  3.809e-03  -3.642 0.000294 ***
## wall.area:glazing.dist1      1.982e-04  2.489e-03   0.080 0.936555    
## wall.area:glazing.dist2      1.133e-03  2.650e-03   0.428 0.669065    
## wall.area:glazing.dist3     -5.624e-04  2.555e-03  -0.220 0.825873    
## wall.area:glazing.dist4      4.101e-04  2.576e-03   0.159 0.873600    
## rel.compact:surface.area    -5.160e+00  4.796e-01 -10.758  < 2e-16 ***
## surface.area:height          5.532e-01  3.135e-02  17.648  < 2e-16 ***
## surface.area:roof.area      -4.763e-02  2.784e-03 -17.110  < 2e-16 ***
## surface.area:wall.area      -4.940e-03  4.643e-04 -10.640  < 2e-16 ***
## surface.area:glazing.area   -5.800e-02  3.271e-03 -17.734  < 2e-16 ***
## surface.area:glazing.dist0   1.487e-02  2.001e-03   7.429 3.89e-13 ***
## surface.area:glazing.dist1   1.266e-03  1.303e-03   0.972 0.331390    
## surface.area:glazing.dist2   2.269e-03  1.281e-03   1.771 0.077056 .  
## surface.area:glazing.dist3   1.304e-03  1.265e-03   1.031 0.302897    
## surface.area:glazing.dist4   1.646e-03  1.287e-03   1.279 0.201406    
## rel.compact:height           1.916e+02  2.715e+01   7.055 4.88e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8329 on 585 degrees of freedom
## Multiple R-squared:  0.9934, Adjusted R-squared:  0.993 
## F-statistic:  3124 on 28 and 585 DF,  p-value: < 2.2e-16
```

```
train.pred.interact3 <- train %>%
  mutate(pred = predict.lm(model.interact3, train),
         resid.stand = rstandard(model.interact3),
         resid = resid(model.interact3))
```

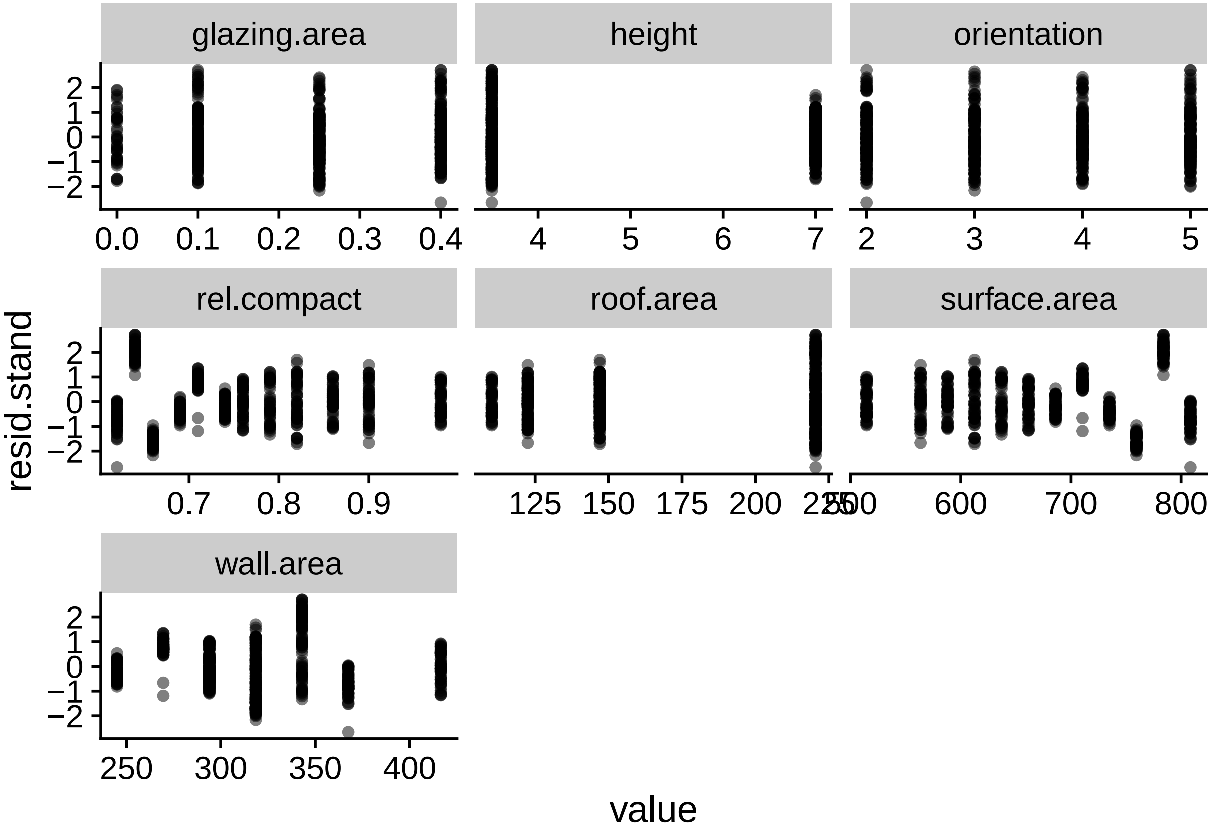
```
## Warning in predict.lm(model.interact3, train): prediction from a rank-
## deficient fit may be misleading
```

```
train.pred.interact3 %>%
  ggplot(aes(x=pred, y=resid.stand)) + geom_point() +
  labs(title = "Standarized Residuals vs. Predicted")
```



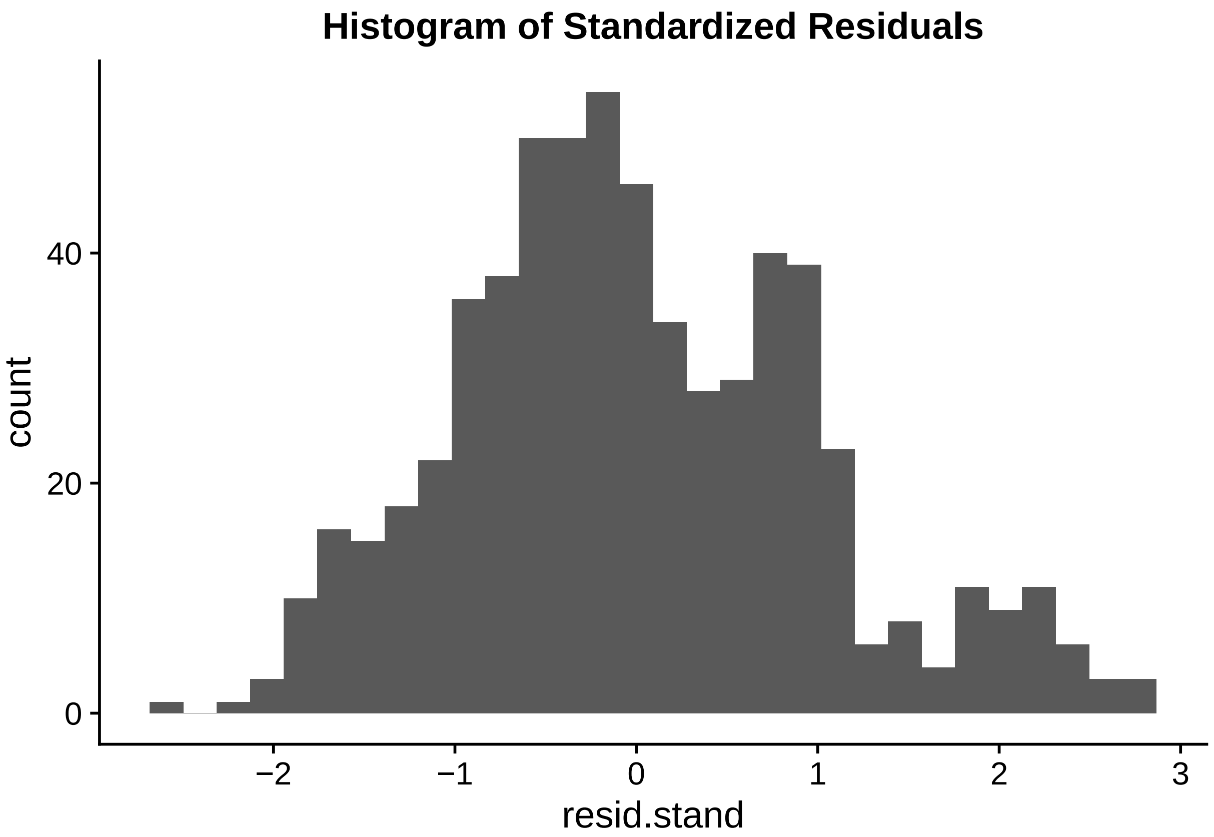
```
train.pred.interact3 %>%
  #dplyr::select(glazing.dist, rel.compact, surface.area,  resid) %>%
  dplyr::select(-pred, -heating.load, -cooling.load, -resid) %>%
  gather(key="var", value="value", -resid.stand, -glazing.dist) %>%
  mutate(value = as.numeric(value)) %>%
  ggplot(aes(x=value, y=resid.stand)) +
  geom_point(alpha=0.5) +
  #geom_jitter(size=1) +
  #geom_boxplot() +
  facet_wrap( ~ var, ncol=3, scales = 'free_x')
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

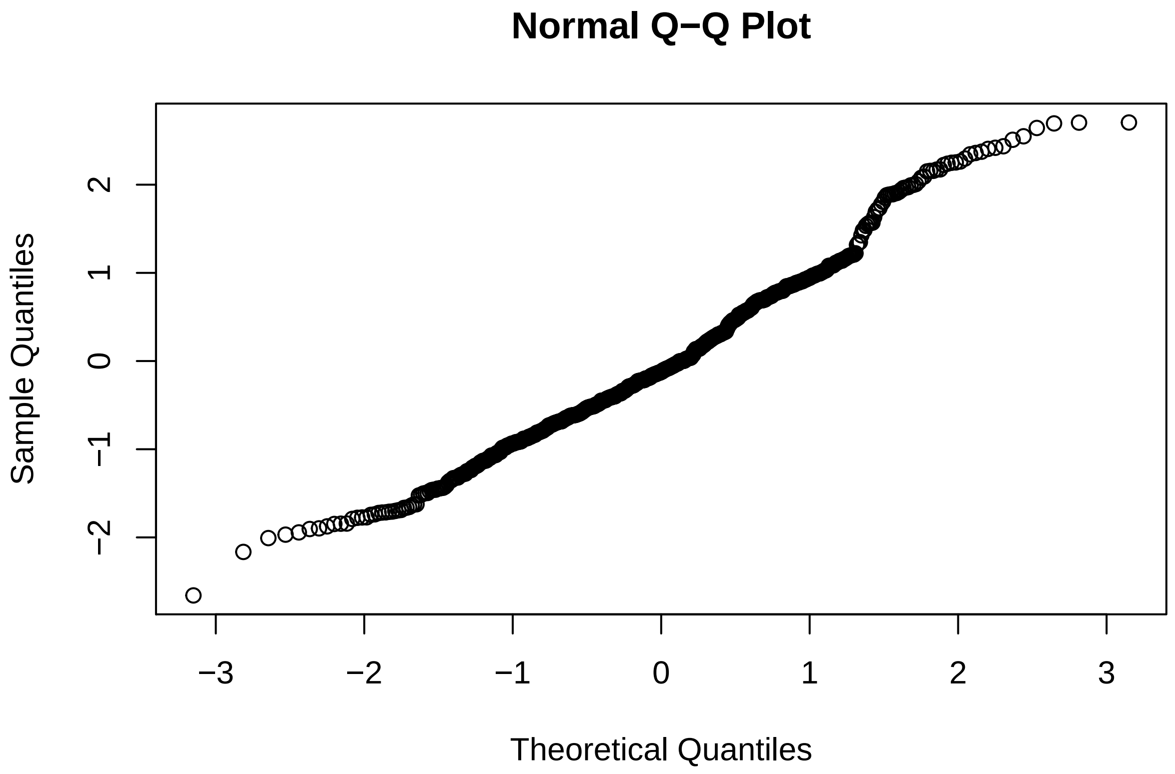


```r
train.pred.interact3 %>%
  ggplot(aes(x=resid.stand)) + geom_histogram() +
  labs(title='Histogram of Standardized Residuals')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
qqnorm(train.pred.interact3$resid.stand)
```



```r
test.pred.interact3 <- test %>%
  mutate(pred = predict.lm(model.interact3, test)) %>%
  mutate(resid = pred - heating.load,
         resid.stand = ((pred - heating.load) - mean(pred - heating.load)) / sd(pred - heating.load))
```

```
## Warning in predict.lm(model.interact3, test): prediction from a rank-
## deficient fit may be misleading
```

```r
print('MAE Train')
```

```
## [1] "MAE Train"
```

```r
print(mean(abs(train.pred.interact3$resid)))
```

```
## [1] 0.6462966
```

```r
print('MAE Val')
```

```
## [1] "MAE Val"
```

```r
print(mean(abs(test.pred.interact3$resid)))
```

```
## [1] 0.6953759
```

```r
print('Train R^2')
```

```
## [1] "Train R^2"
```

```r
RSS = sum(train.pred.interact3$resid^2)
TSS = sum((train.pred.interact3$heating.load - mean(train.pred.interact3$heating.load))^2)
print(1 - RSS/TSS)
```

```
## [1] 0.9933746
```

```r
print('Test R^2')
```

```
## [1] "Test R^2"
```

```r
RSS = sum(test.pred.interact3$resid^2)
TSS = sum((test.pred.interact3$heating.load - mean(test.pred.interact3$heating.load))^2)
print(1 - RSS/TSS)
```

```
## [1] 0.9926262
```

```r
train.pred.interact3 %>% arrange(desc(abs(resid.stand)))
```

```
## # A tibble: 614 x 13
##    rel.compact surface.area wall.area roof.area height orientation
##          <dbl>        <dbl>     <dbl>     <dbl>  <dbl> <fct>
##  1       0.640         784.      343.      220.   3.50 2
##  2       0.640         784.      343.      220.   3.50 5
##  3       0.640         784.      343.      220.   3.50 5
##  4       0.620         808.      368.      220.   3.50 2
##  5       0.640         784.      343.      220.   3.50 3
##  6       0.640         784.      343.      220.   3.50 3
##  7       0.640         784.      343.      220.   3.50 5
##  8       0.640         784.      343.      220.   3.50 3
##  9       0.640         784.      343.      220.   3.50 4
## 10       0.640         784.      343.      220.   3.50 2
## # ... with 604 more rows, and 7 more variables: glazing.area <dbl>,
## #   glazing.dist <fct>, heating.load <dbl>, cooling.load <dbl>,
## #   pred <dbl>, resid.stand <dbl>, resid <dbl>
```

```
train.pred.interact3
```

```
## # A tibble: 614 x 13
##    rel.compact surface.area wall.area roof.area height orientation
##          <dbl>        <dbl>     <dbl>     <dbl>  <dbl> <fct>
##  1       0.620         808.      368.      220.   3.50 3
##  2       0.690         735.      294.      220.   3.50 3
##  3       0.820         612.      318.      147.   7.00 5
##  4       0.760         662.      416.      122.   7.00 5
##  5       0.620         808.      368.      220.   3.50 4
##  6       0.660         760.      318.      220.   3.50 2
##  7       0.820         612.      318.      147.   7.00 3
##  8       0.820         612.      318.      147.   7.00 3
##  9       0.640         784.      343.      220.   3.50 2
## 10       0.710         710.      270.      220.   3.50 4
## # ... with 604 more rows, and 7 more variables: glazing.area <dbl>,
## #   glazing.dist <fct>, heating.load <dbl>, cooling.load <dbl>,
## #   pred <dbl>, resid.stand <dbl>, resid <dbl>
```

Section X: Discussion

- Real world impact of findings
- Balance with the fact that the dataset is simulated