Lab 3

2/8/2022

Variable transformations

The World Bank provides valuable data on a number of public health and economic indicators for countries across the globe[1]. Today, we will be looking indicators which might predict infant mortality, which is the number of children (per 1000 births) who die before the age of 1.

Questions

- What factors do you think might affect or correlate with infant mortality?

In particular, we will be looking at 2 specific factors which might correlate well with infant mortality (measured in 2015) - GDP per capita (roughly how much income does the average individual produce) as measured in 2013 and the proportion of the population with access to electricity (as measured in 2012). I have removed countries which were missing data for any of the variables.

```
fileName <- "https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lab2/world_bank_data.csv"
wb.data <- read.csv(fileName)
head(wb.data)
```

```
##                country  elec_acc inf_mort gdp_capita
## 1              Andorra 100.00000      2.1 42806.5226
## 2          Afghanistan  43.00000     66.3   666.7951
## 3               Angola  37.00000     96.0  5900.5296
## 4              Albania 100.00000     12.5  4411.2582
## 5 United Arab Emirates  97.69783      5.9 42831.0891
## 6            Argentina  99.80000     11.1 14443.0657
```

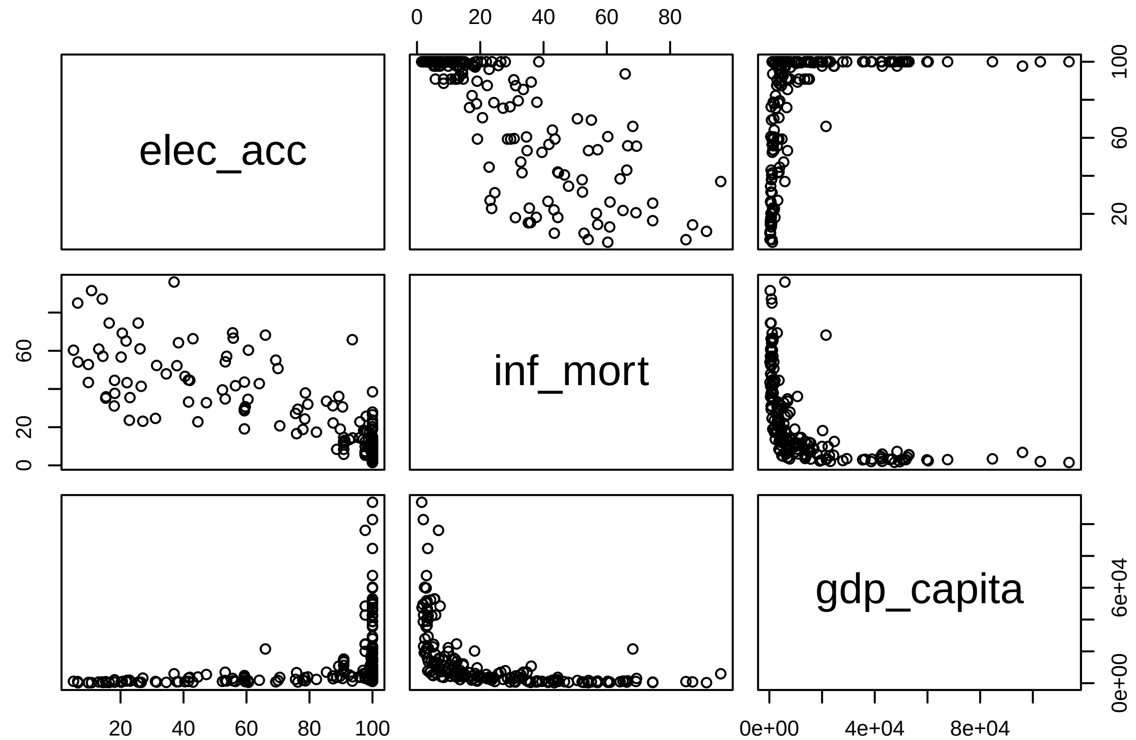
Questions

- What direction do you think the association is between each of these variables?
- What strength do you think the association is between each of these variables?

We can use the `pairs` command to plot the many pairs of variables at once. Note that we've excluded the first column here, since that's just the name of countries

[1] You can access the data at http://data.worldbank.org/

```r
pairs(wb.data[, -1])
```



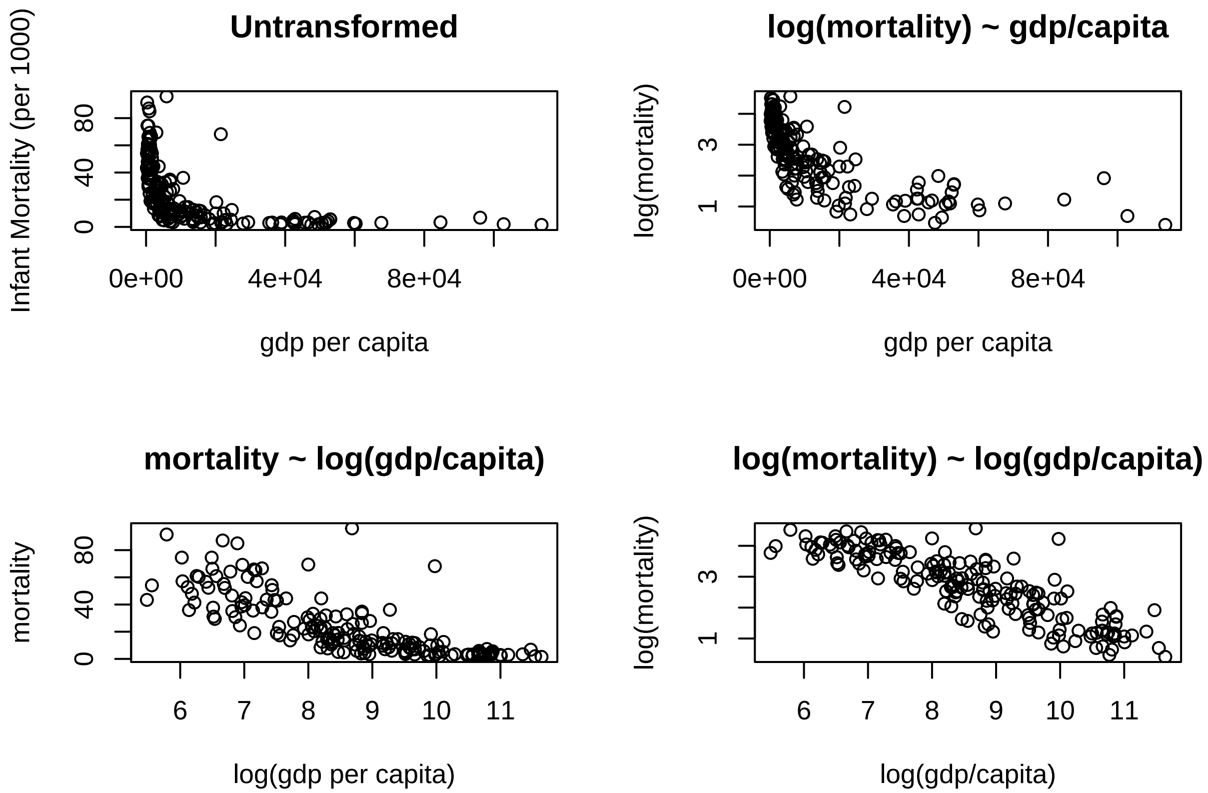
Questions

- Does this look like what you might expect?
- What sticks out?
- Do the relationships look linear?

The relationship between electricity and infant mortality looks roughly linear, but the relationship between GDP per capita and infant mortality does not. Let's see how we might transform the data. The `log` function by default returns the natural log (base e). Let's plot a few transformations and see what makes the relationship linear.

```r
# using the par(mfrow = c(r, c)) puts multiple
# plots together. The plots are arranged so
# that there are r rows and c columns

par(mfrow = c(2,2))

# first argument is the X variable, second argument is the Y variable
# main specifies the title, xlab specifies the x axis label
# and ylab specifies the y axis label
plot(wb.data$gdp_capita, wb.data$inf_mort, main = "Untransformed",
     xlab = "gdp per capita", ylab = "Infant Mortality (per 1000)")

plot(wb.data$gdp_capita, log(wb.data$inf_mort),
     main = "log(mortality) ~ gdp/capita",
```

```
     xlab = "gdp per capita", ylab = "log(mortality)")

plot(log(wb.data$gdp_capita), wb.data$inf_mort,
     main = "mortality ~ log(gdp/capita)",
     xlab = "log(gdp per capita)", ylab = "mortality")

plot(log(wb.data$gdp_capita), log(wb.data$inf_mort),
     main = "log(mortality) ~ log(gdp/capita)",
     xlab = "log(gdp/capita)", ylab = "log(mortality)")
```



The plots correspond to the models:

$$E(\text{mortality} \mid \text{gdp/capita}) = b_0 + b_1 \text{gdp/capita}$$

$$E(\log(\text{mortality}) \mid \text{gdp/capita}) = b_0 + b_1 \text{gdp/capita}$$

$$E(\text{mortality} \mid \text{gdp/capita}) = b_0 + b_1 \log(\text{gdp/capita})$$

$$E(\log(\text{mortality}) \mid \text{gdp/capita}) = b_0 + b_1 \log(\text{gdp/capita})$$

Questions

- Which transformation looks most linear?
- How do we interpret the b_1 parameter in each model?

The transformation that looks most linear takes the log of both mortality and gdp per capita. We can estimate the transformed and untransformed models now using the `lm` command.

```r
# Untransformed data
untransformed.reg <- lm(inf_mort ~ gdp_capita, data = wb.data)

summary(untransformed.reg)
```

```
##
## Call:
## lm(formula = inf_mort ~ gdp_capita, data = wb.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.011 -14.633  -5.749   8.625  67.583
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.168e+01  1.743e+00  18.171  < 2e-16 ***
## gdp_capita  -5.523e-04  7.093e-05  -7.787 5.68e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.07 on 176 degrees of freedom
## Multiple R-squared:  0.2562, Adjusted R-squared:  0.252
## F-statistic: 60.63 on 1 and 176 DF,  p-value: 5.678e-13
```

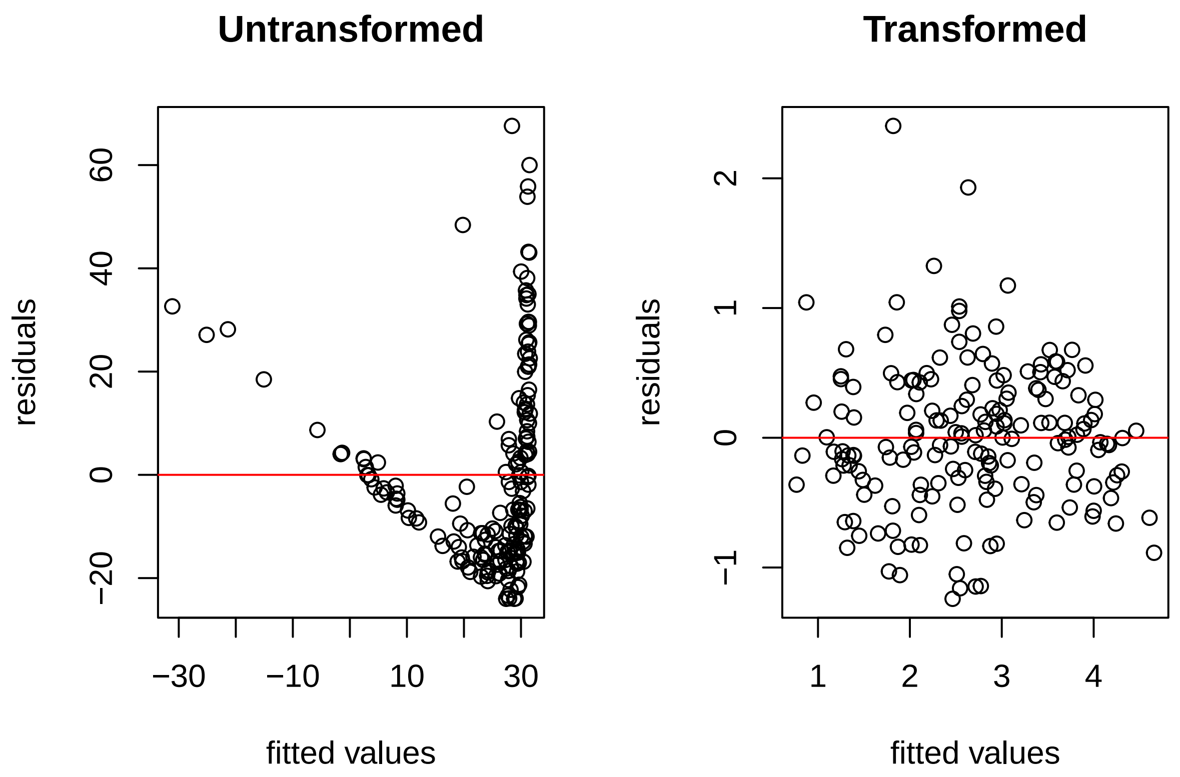
```r
# regression with transformed data
transformed.reg <- lm(log(inf_mort) ~ log(gdp_capita), data = wb.data)

summary(transformed.reg)
```

```
##
## Call:
## lm(formula = log(inf_mort) ~ log(gdp_capita), data = wb.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24132 -0.34865 -0.00525  0.34525  2.40377
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.11682    0.24882   32.62   <2e-16 ***
## log(gdp_capita) -0.63135    0.02848  -22.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5554 on 176 degrees of freedom
## Multiple R-squared:  0.7363, Adjusted R-squared:  0.7348
## F-statistic: 491.3 on 1 and 176 DF,  p-value: < 2.2e-16
```

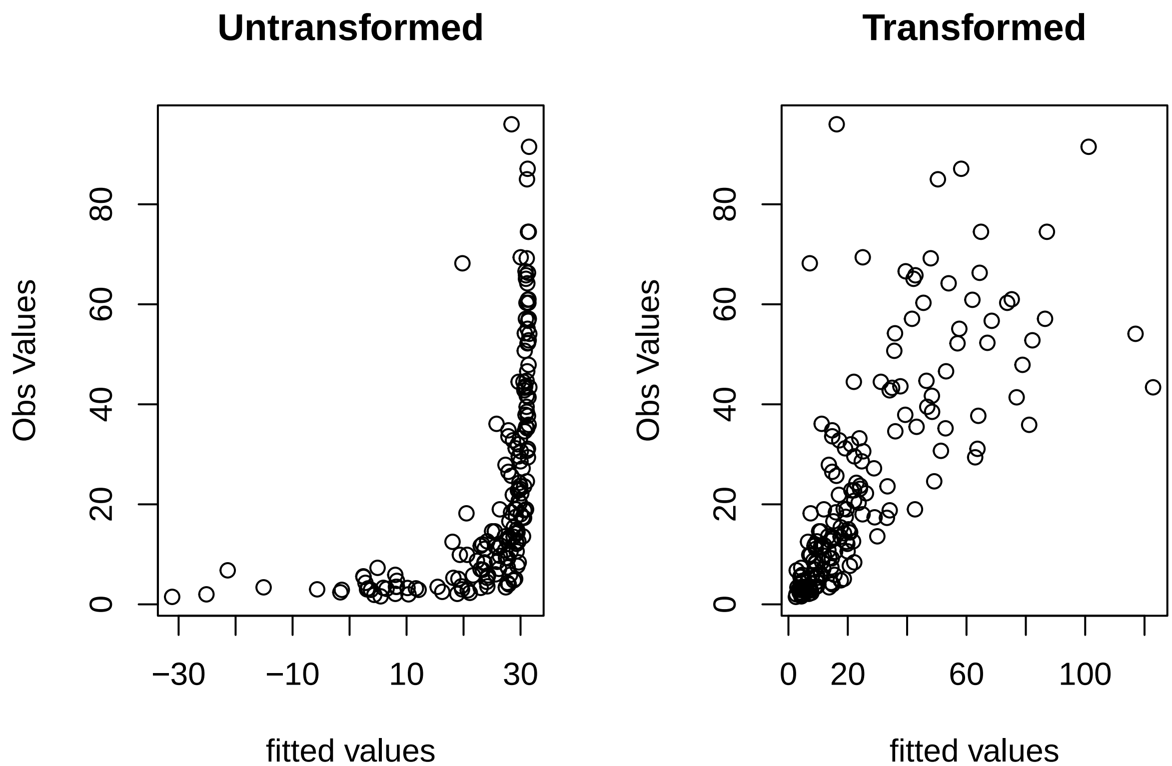
We can also look at the residuals plotted against fitted values and fitted values vs observed values for both models. What does this suggest about how each model fits our data?

```r
par(mfrow = c(1,2))
plot(untransformed.reg$fitted.values, untransformed.reg$residuals, main = "Untransformed",
     xlab = "fitted values", ylab = "residuals")
abline(h=0,col="red")
```

```
plot(transformed.reg$fitted.values, transformed.reg$residuals, main = "Transformed",
     xlab = "fitted values", ylab = "residuals")
abline(h=0,col="red")
```




```
par(mfrow = c(1,2))
plot(untransformed.reg$fitted.values, wb.data$inf_mort, main = "Untransformed",
     xlab = "fitted values", ylab = "Obs Values")


fitted.values.log <- exp(transformed.reg$fitted.values + summary(transformed.reg)$sigma^2/2)
plot(fitted.values.log,wb.data$inf_mort, main = "Transformed",
     xlab = "fitted values", ylab = "Obs Values")
```



Questions

- What do you notice about the fitted values for the untransformed data? Hint: What is the range of fitted values, and does it make sense given the variable we are predicting?
- Compare the R^2 from both regressions. What does this suggest about which explanatory variable is a better predictor of infant mortality?
- Why do you think this is true?
- Note that we aren't exactly comparing apples to apples here because one regression has log(mortality) as the response while the other uses mortality untransformed. Is there a way you could make the comparison more fair?
- Which model would you use if you are trying to predict infant mortality for a country not in the data set? Which model would you use if you are trying to explain to a collaborator? Which model would you use if you are trying to test if infant mortality is associated with gdp/capita?
- Repeat the exercise but with electricity access? Which model would you select when using electricity access? What about when you include both electricity access and gdp per capita?

Housing Data

In class, we've been discussing data about housing prices and in last week's lab, we considered modeling the home prices with polynomial regression. As a quick refresher, recall that there are 522 observations with the following variables:

- price: in 2002 dollars
- area: Square footage
- bed: number of bedrooms
- bath: number of bathrooms
- ac: central AC (yes/no)
- garage: number of garage spaces
- pool: yes/no
- year: year of construction
- quality: high/medium/low
- home style: coded 1 through 7
- lot size: sq ft
- highway: near a highway (yes/no)

```
fileName <- url("https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lectureData/estate.csv")
housing_data <- read.csv(fileName)

head(housing_data)
```

```
##   id  price area bed bath  ac garage pool year quality style   lot highway
## 1  1 360000 3032   4    4 yes      2   no 1972  medium     1 22221      no
## 2  2 340000 2058   4    2 yes      2   no 1976  medium     1 22912      no
## 3  3 250000 1780   4    3 yes      2   no 1980  medium     1 21345      no
## 4  4 205500 1638   4    2 yes      2   no 1963  medium     1 17342      no
## 5  5 275500 2196   4    3 yes      2   no 1968  medium     7 21786      no
## 6  6 248000 1966   4    3 yes      5  yes 1972  medium     1 18902      no
```

```
housing_data$age <- 2002 - housing_data$year
```

Categorical variables

In our data, Housing Style is coded 1 through 7

```
table(housing_data$style)
```

```
##
##   1   2   3   4   5   6   7   9  10  11
## 214  58  64  11  18  18 136   1   1   1
```

In class, we described how to include categorical variables in a regression by picking a reference category and then including binary variables for the other categories. R does this entire process for us inside the lm command.

```
###
# Include style
model1 <- lm(price ~ area + style, data = housing_data)
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ area + style, data = housing_data)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -271624  -34852   -5465   28660  312589
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.030e+05  1.126e+04  -9.152  < 2e-16 ***
## area         1.875e+02  5.857e+00  32.021  < 2e-16 ***
## style       -1.286e+04  1.625e+03  -7.912 1.54e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74820 on 519 degrees of freedom
## Multiple R-squared:  0.7069, Adjusted R-squared:  0.7058
## F-statistic: 625.8 on 2 and 519 DF,  p-value: < 2.2e-16
```

```
###
# Include style as a factor (i.e., make sure R knows it is categorical data)
model2 <- lm(price ~ area + as.factor(style), data = housing_data)
summary(model2)
```

```
##
## Call:
## lm(formula = price ~ area + as.factor(style), data = housing_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -273461  -34602   -4571   28259  310176
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -1.162e+05  1.286e+04  -9.034  < 2e-16 ***
## area                1.882e+02  6.094e+00  30.886  < 2e-16 ***
## as.factor(style)2  -2.040e+04  1.107e+04  -1.843   0.0659 .
## as.factor(style)3  -1.785e+04  1.066e+04  -1.674   0.0948 .
## as.factor(style)4  -3.446e+04  2.311e+04  -1.491   0.1366
## as.factor(style)5  -8.499e+04  1.856e+04  -4.578 5.90e-06 ***
## as.factor(style)6  -7.597e+04  1.867e+04  -4.068 5.49e-05 ***
## as.factor(style)7  -7.854e+04  1.043e+04  -7.528 2.35e-13 ***
## as.factor(style)9   2.033e+04  7.504e+04   0.271   0.7866
## as.factor(style)10 -8.684e+04  7.597e+04  -1.143   0.2535
## as.factor(style)11 -6.179e+04  7.493e+04  -0.825   0.4100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74750 on 511 degrees of freedom
## Multiple R-squared:  0.7119, Adjusted R-squared:  0.7063
## F-statistic: 126.3 on 10 and 511 DF,  p-value: < 2.2e-16
```

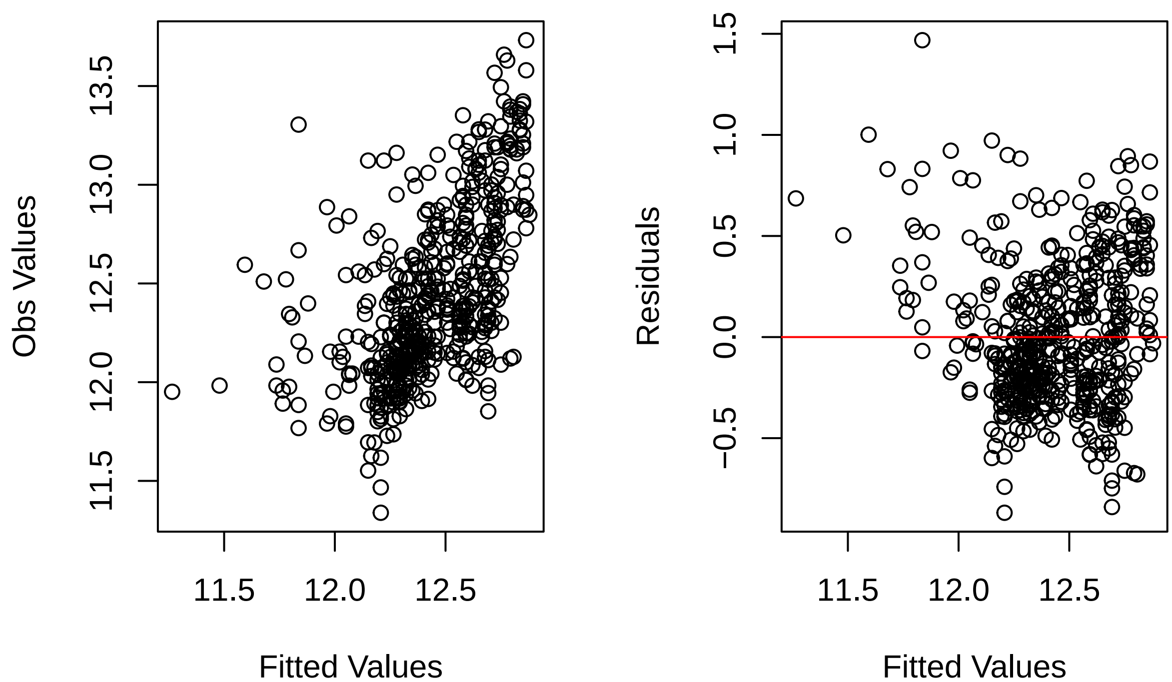
Questions

- What is the reference category that R is using?
- How would you interpret the estimated coefficients?
- What is the estimated difference in home price when comparing a house which is style 2 against a house which is style 4?

Interaction terms

Last week, we examined how home prices were associated with age and modeled the relationship with polynomial regressions. If you recall, none of the models fit particularly well. Turns out, using a log transformation on housing price seems to make the relationship more linear.

```
model3 <- lm(log(price) ~ age, data = housing_data)

par(mfrow = c(1,2))
plot(model3$fitted.values, log(housing_data$price), xlab = "Fitted Values", ylab = "Obs Values")
plot(model3$fitted.values, model3$residuals, xlab = "Fitted Values", ylab = "Residuals")
abline(h=0,col="red")
```



We see from the estimated coefficients that an older home is typically less expensive than a newer home.

```
summary(model3)
```

```
##
## Call:
## lm(formula = log(price) ~ age, data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86899 -0.24789 -0.07036  0.22367  1.46833
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.9356873  0.0342333  377.87   <2e-16 ***
```

```
## age         -0.0142770  0.0008717  -16.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3509 on 520 degrees of freedom
## Multiple R-squared:  0.3403, Adjusted R-squared:  0.339
## F-statistic: 268.2 on 1 and 520 DF,  p-value: < 2.2e-16
```

However, as we discussed in class, we might also expect that the association of price and age depends on the quality of the home. We can fit a model with the interaction between age and quality to see

```r
# We can include each covariate and the interaction term in the lm formula
model4 <- lm(log(price) ~ age + quality + age * quality, data = housing_data)
summary(model4)
```

```
##
## Call:
## lm(formula = log(price) ~ age + quality + age * quality, data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70937 -0.16798 -0.00132  0.14146  0.82006
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       13.2543198  0.0518526 255.615   <2e-16 ***
## age               -0.0045991  0.0026210  -1.755   0.0799 .
## qualitylow        -1.0931907  0.0888987 -12.297   <2e-16 ***
## qualitymedium     -0.6222432  0.0631323  -9.856   <2e-16 ***
## age:qualitylow     0.0023796  0.0029750   0.800   0.4241
## age:qualitymedium -0.0003452  0.0028204  -0.122   0.9026
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2524 on 516 degrees of freedom
## Multiple R-squared:  0.6615, Adjusted R-squared:  0.6582
## F-statistic: 201.7 on 5 and 516 DF,  p-value: < 2.2e-16
```

```r
# Alternatively, if we only explicitly specify the interaction term, the main
# effects are automatically included
model5 <- lm(log(price) ~ age * quality, data = housing_data)
summary(model5)
```

```
##
## Call:
## lm(formula = log(price) ~ age * quality, data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70937 -0.16798 -0.00132  0.14146  0.82006
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       13.2543198  0.0518526 255.615   <2e-16 ***
## age               -0.0045991  0.0026210  -1.755   0.0799 .
## qualitylow        -1.0931907  0.0888987 -12.297   <2e-16 ***
```

```
## qualitymedium     -0.6222432  0.0631323  -9.856   <2e-16 ***
## age:qualitylow     0.0023796  0.0029750   0.800   0.4241
## age:qualitymedium -0.0003452  0.0028204  -0.122   0.9026
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2524 on 516 degrees of freedom
## Multiple R-squared:  0.6615, Adjusted R-squared:  0.6582
## F-statistic: 201.7 on 5 and 516 DF,  p-value: < 2.2e-16
```

Questions

- Write out the form of the model that is being estimated
- Looking at the estimated coefficients, are you surprised by the results?
- Do you think the relationship between age and price differs depending on quality?
- What are some reasons you might include the interaction term in your model?
- What are some reasons you might choose to not include the interaction term in your model?

```r
#A question was asked in class: if we omit the interaction, will the coefficient estimate for the effec
model6 <- lm(log(price) ~ age + quality, data = housing_data)
summary(model6)
```

```
##
## Call:
## lm(formula = log(price) ~ age + quality, data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70638 -0.16989 -0.00538  0.15353  0.84320
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.2453484  0.0331623  399.41  < 2e-16 ***
## age           -0.0040373  0.0007979   -5.06 5.83e-07 ***
## qualitylow    -0.9944879  0.0451194  -22.04  < 2e-16 ***
## qualitymedium -0.6418531  0.0362071  -17.73  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2525 on 518 degrees of freedom
## Multiple R-squared:  0.6599, Adjusted R-squared:  0.6579
## F-statistic:   335 on 3 and 518 DF,  p-value: < 2.2e-16
```

```r
#We see that the coefficients are different. This is because in the model with the interaction, the coe
```