stat_final.R

benjamin

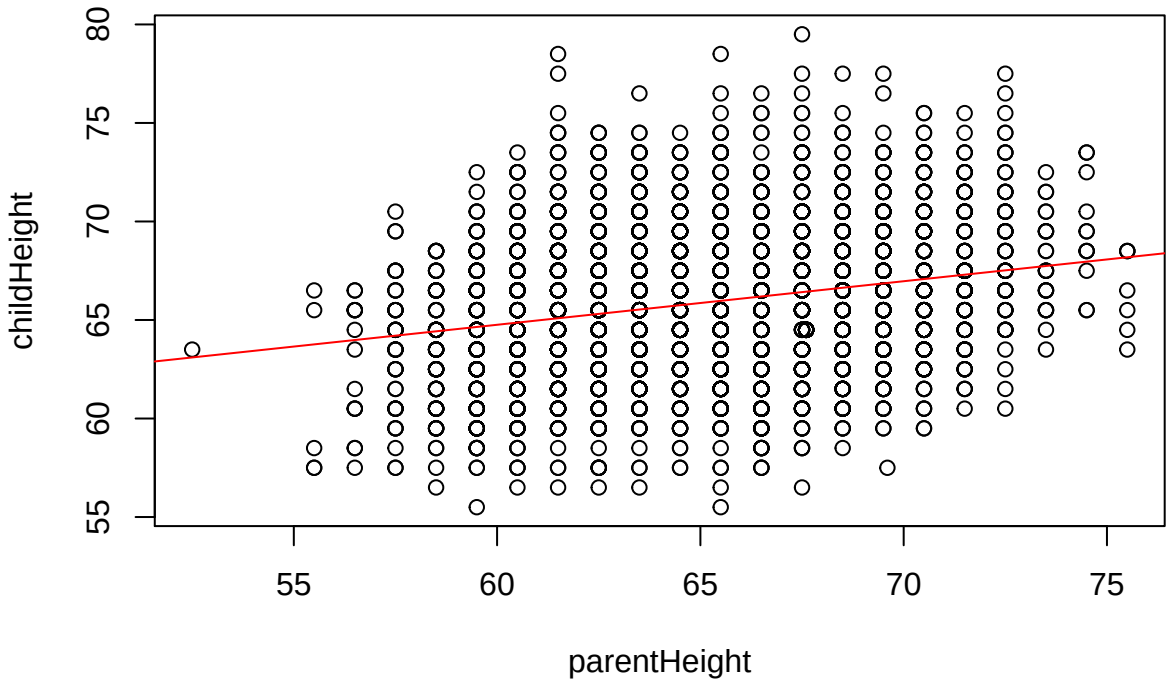
2020-12-01

```r
### Ben Reichert
### Final


## 1
# a)
girls <- c(rep(0, 23236), rep(1,58529), rep(2,53908), rep(3, 18770))
n <- length(girls)
mean <- mean(girls)
var <- var(girls)
# mean = 1.441665
# variance = 0.7905698

# b)
t <- t.test(girls)
# 95% confidence interval of (1.437230, 1.446099)

# c)
# width/2 = Z * sd(n)/sqrt(n)
# 0.02/2 = 1.96 * sd(n) / sqrt(n)
# 0.01 = 1.96 * sd(n) / sqrt(n)
# sqrt(n) = 1.96 * sd(n) / 0.01
# n = (1.96 * sd(n) / 0.01)
n <- (qnorm(0.975)*sd(girls)/0.01)^2
# Number of families to sample (n) when we want 95% confidence and width of 0.02: n = 30370 (rounded up


## 2 ##
# a)
# We need to compare the means of two independent groups of data, so we either do students t-test or we
# Because the standard deviations are not the same we use Welch's t-test.
# Even though the samples are right skewed, they are large enough so that this will not matter much whe

# b)
t <- (16.8 - 24.3) / sqrt( 15.9^2/592 + 17.3^2/154 )
# T-statistic = -4.87
p <- 2*pt(-abs(t),df=225)
# P-value = 0.000002

# c)
m <- 16.8-24.3
diff <- qnorm(0.975)*sqrt( 15.9^2/592 + 17.3^2/154 )
```

```r
min <- m-diff
max <- m+diff
# Confidence interval: (-10.51764, -4.482365)


## 3 ##
pears <- read.table("PearsonLee.txt", header=TRUE)

# a)
lin <- lm(childHeight ~ parentHeight, data = pears)
# Linear regression formula: y = 0.221x + 51.497
# Where y is the predicted child's height and x is parent's height.
library(ggplot2)
print(plot(pears[,1], pears[,2], xlab="parentHeight", ylab="childHeight"))
```

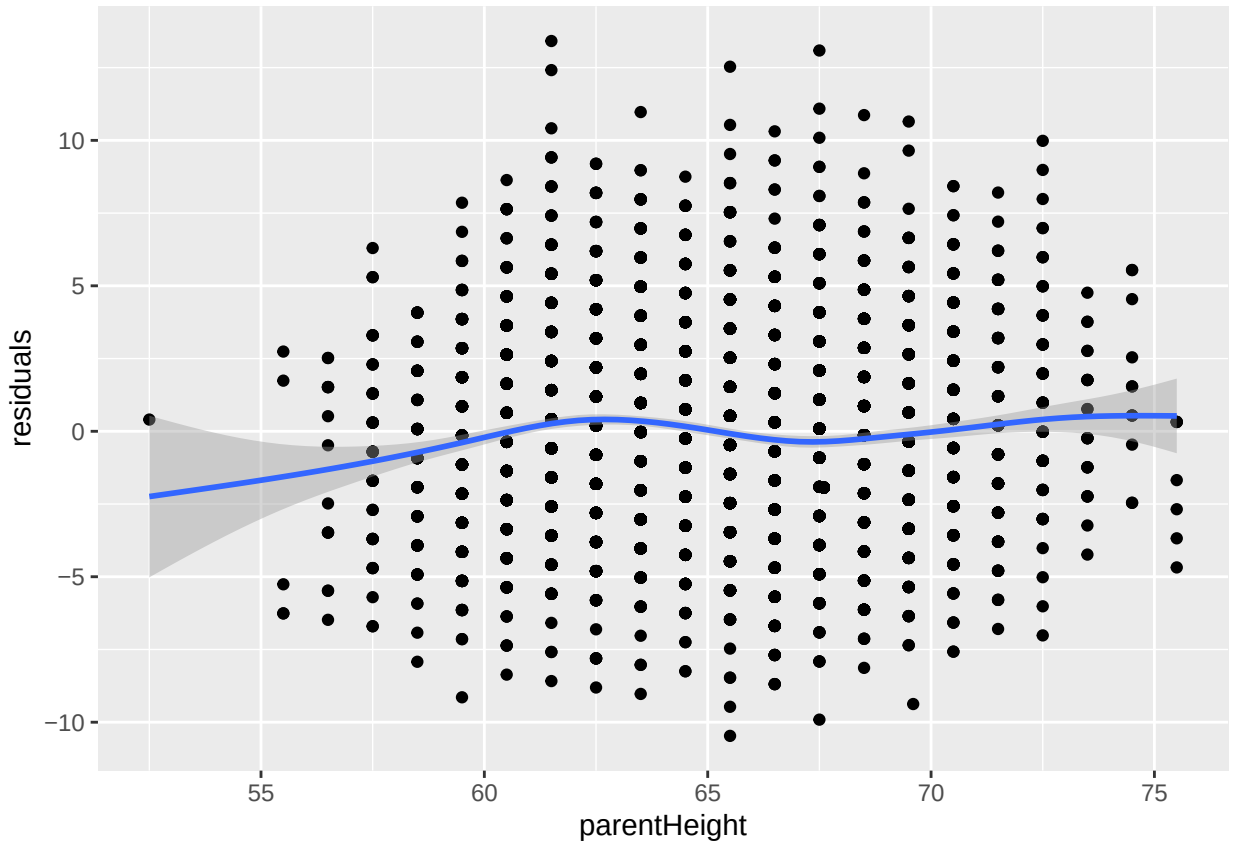
```
## NULL
```

```r
abline(a=51.497, b=0.221, col=2)
```



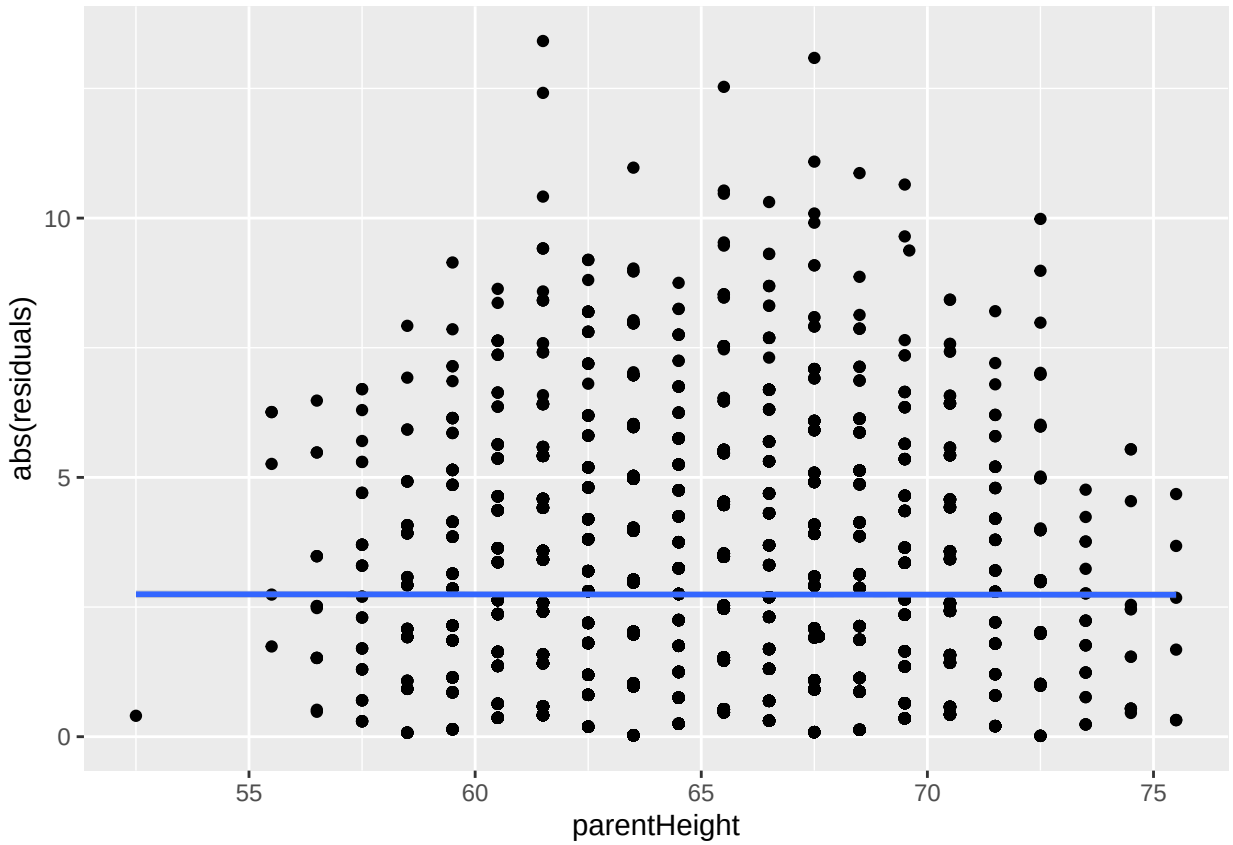

```r
# b)
# linearity
pears.df <- data.frame(pears,
                       fitted = fitted.values(lin),
                       residuals = residuals(lin))
print(ggplot(pears.df, aes(x=parentHeight, y=residuals)) + geom_point() + geom_smooth())
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

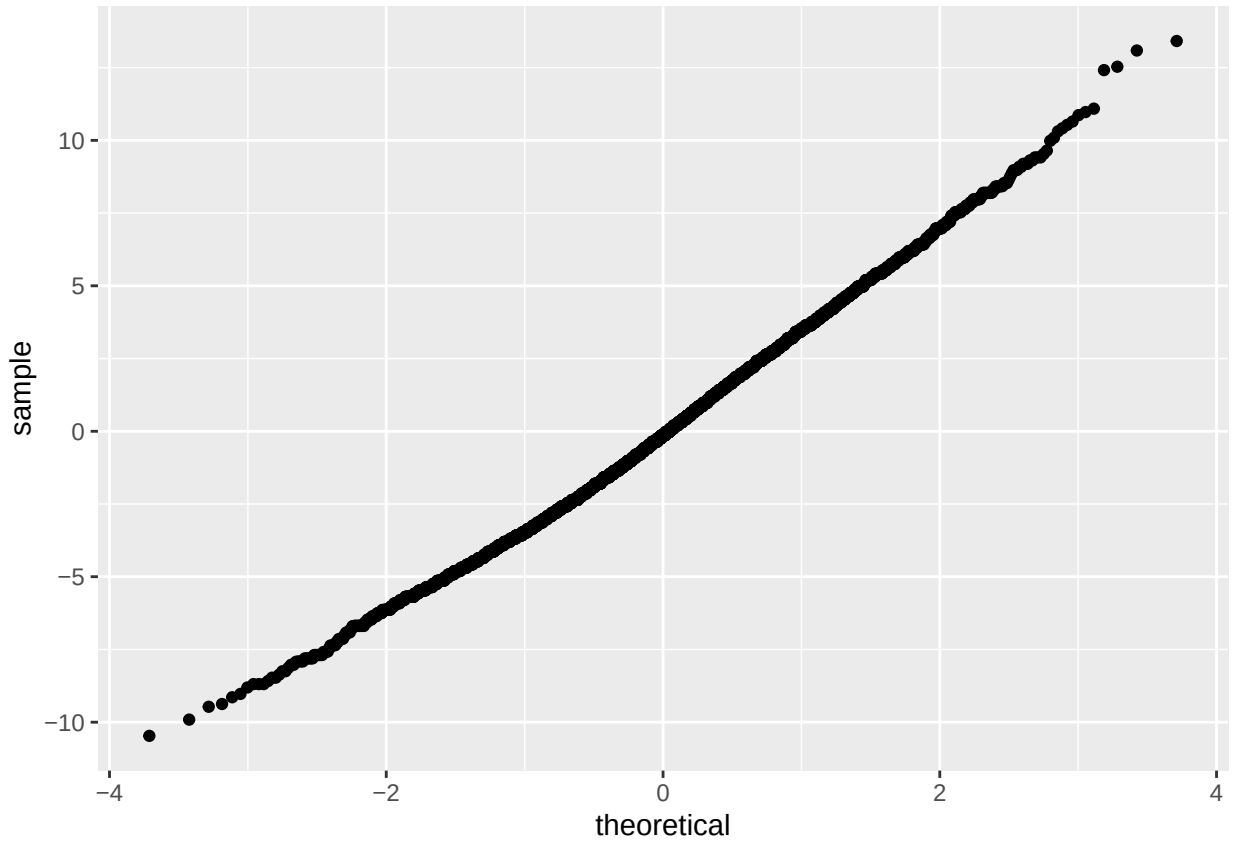



```
# this data looks relatively linear. The regression line on the graph follows an upward trend of
# taller parents produce taller children on average, which is what we would expect. However, the risidu
# graph shaded region does not entirely contain the line y=0, we see an upward trend. Because we see
# an upward trend in the residuals it is hard to confirm that this data is in fact linear.

# independence
# the two sets of data are not independent. The height of the child is affected by
# the height their parent, and is thus not independent of it, or is dependent on it.

# homoskedasticity
print(ggplot(pears.df, aes(x = parentHeight, y = abs(residuals))) + geom_point() + geom_smooth())
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
# Equal variance of errors seems to hold. The magnitude of residuals (blue line) holds constant.

# normality
print(ggplot(pears.df, aes(sample = residuals)) + stat_qq())
```

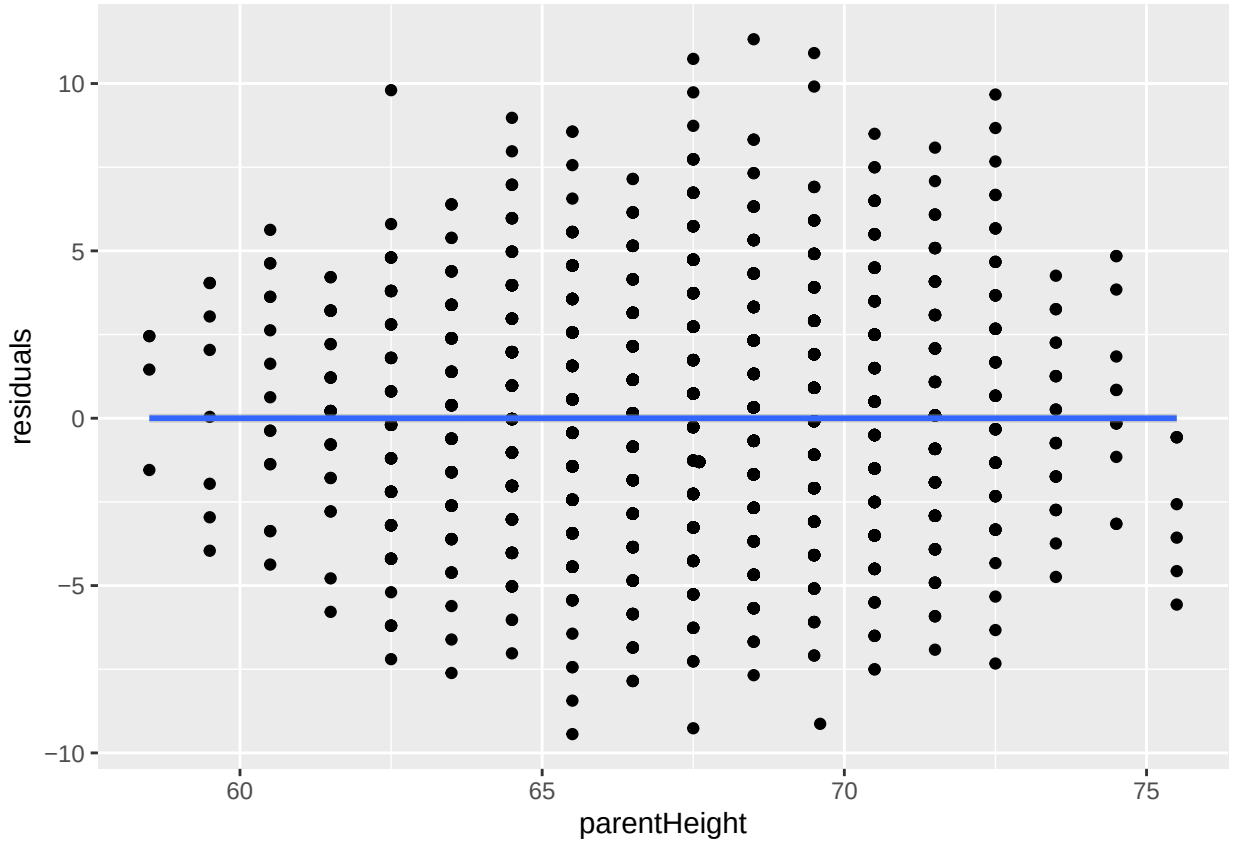



```r
# the normal QQ plot is very straight so its safe to say the residuals are normal.

# c)
# linear model for dads
dads <- subset(pears, parentGender == "Father")
lin <- lm(childHeight ~ parentHeight, data = dads)
print(summary(lin))
```

```
##
## Call:
## lm(formula = childHeight ~ parentHeight, data = dads)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4377 -2.2638 -0.0899  2.3232 11.3232
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.88347    1.58648   23.88   <2e-16 ***
## parentHeight  0.41304    0.02348   17.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.196 on 2444 degrees of freedom
## Multiple R-squared:  0.1124, Adjusted R-squared:  0.112
## F-statistic: 309.5 on 1 and 2444 DF,  p-value: < 2.2e-16
```

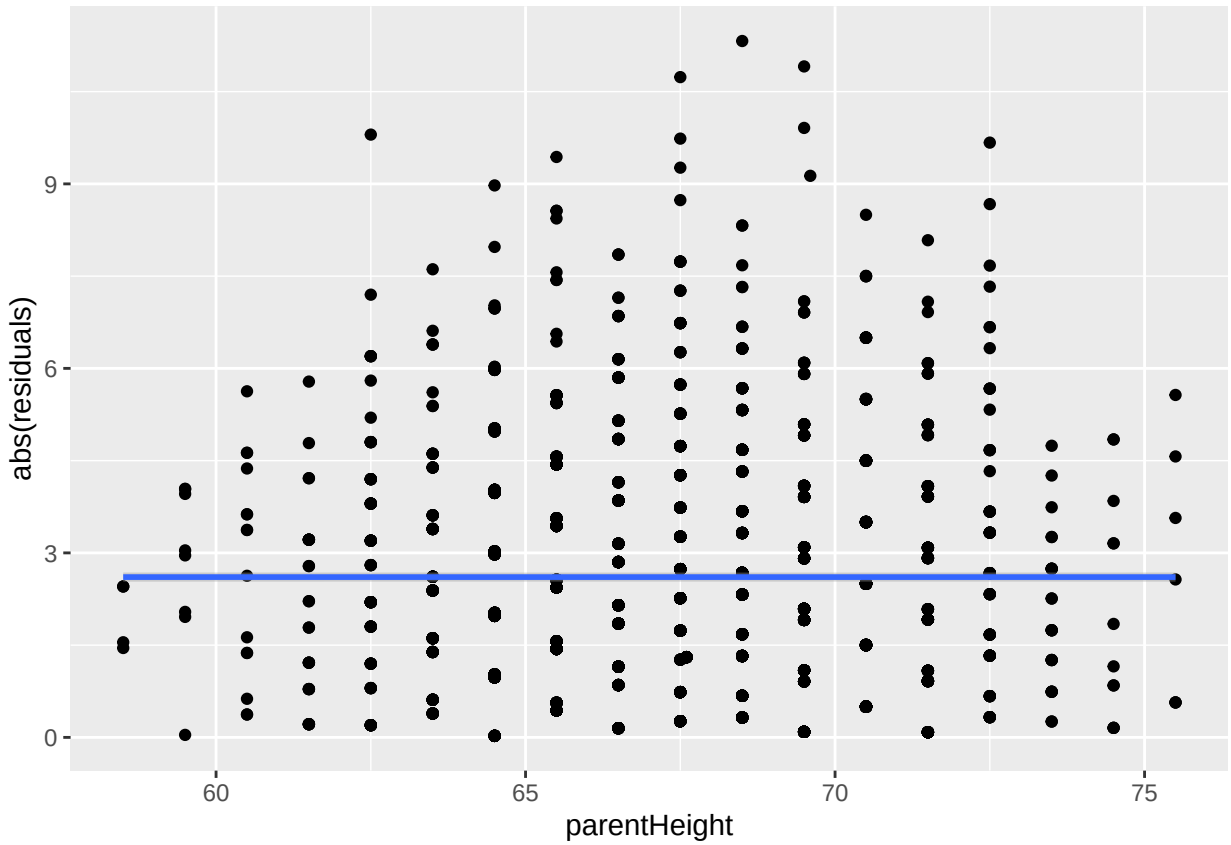
```
# y = 0.41304x + 37.88347
dads.df <- data.frame(dads,
                      fitted = fitted.values(lin),
                      residuals = residuals(lin))
print(ggplot(dads.df, aes(x=parentHeight, y=residuals)) + geom_point() + geom_smooth())
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

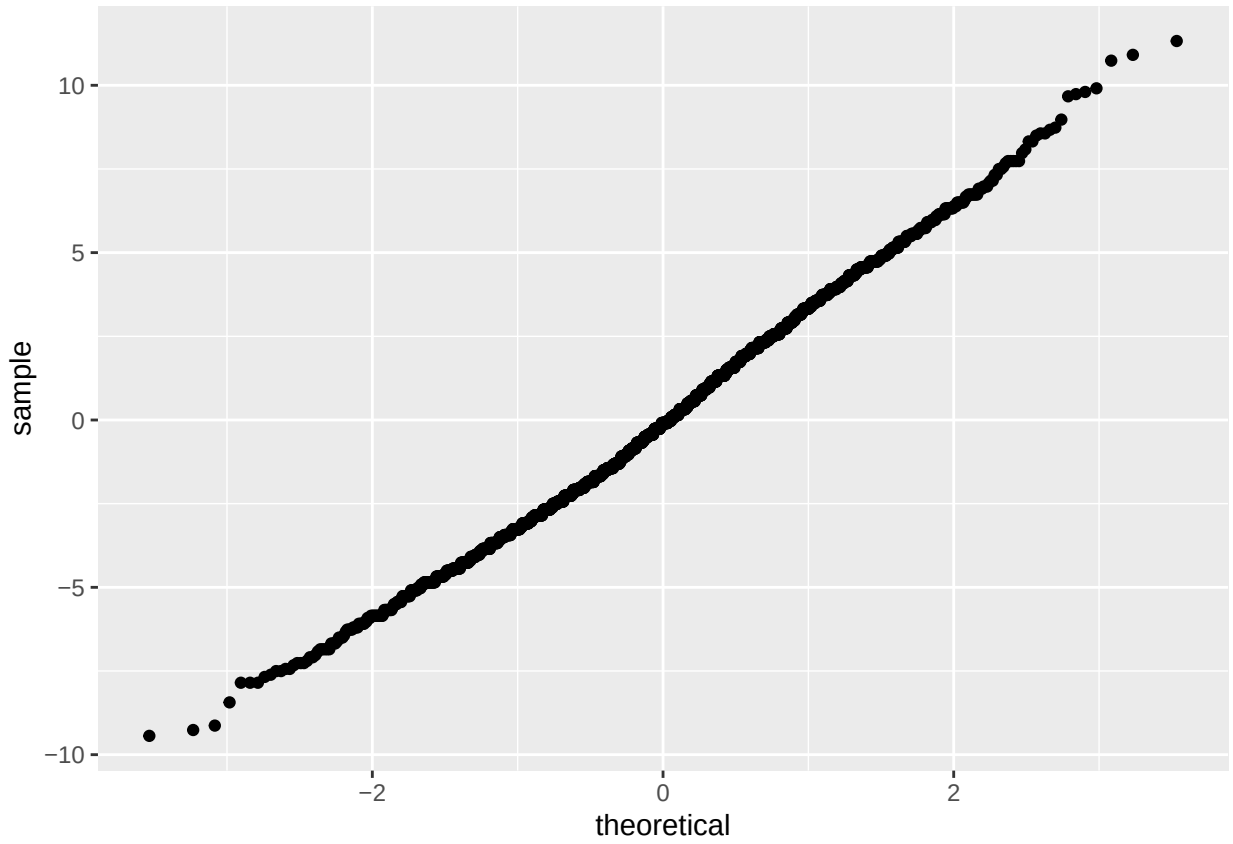



```
print(ggplot(dads.df, aes(x = parentHeight, y = abs(residuals))) + geom_point() + geom_smooth())
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
print(ggplot(dads.df, aes(sample = residuals)) + stat_qq())
```



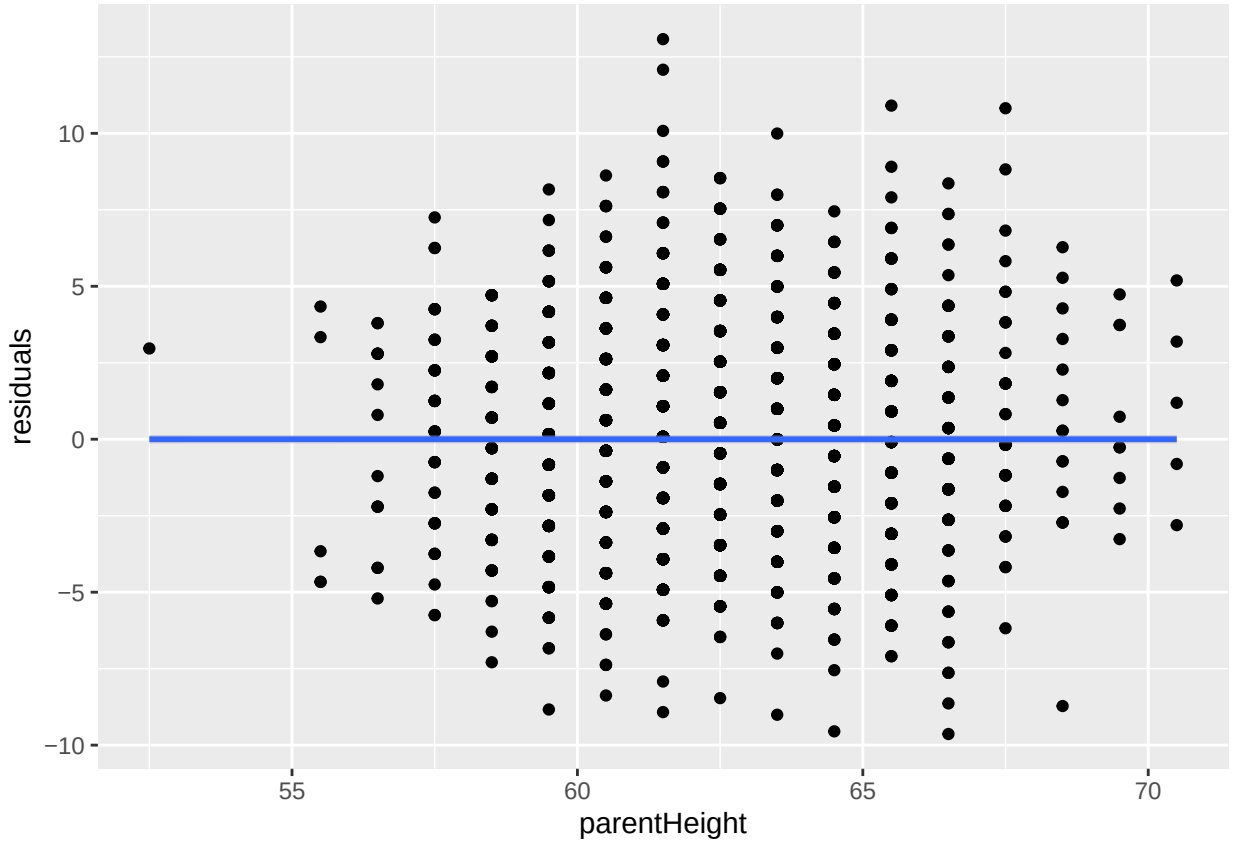

```r
# the assumptions of linear regression are much easier to confirm with the dads data alone.

moms <- subset(pears, parentGender == "Mother")
lin <- lm(childHeight ~ parentHeight, data = moms)
print(summary(lin))
```

```
##
## Call:
## lm(formula = childHeight ~ parentHeight, data = moms)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.636 -2.463 -0.179  2.450 13.080
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   32.0195     1.7755   18.03   <2e-16 ***
## parentHeight   0.5431     0.0284   19.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.31 on 2419 degrees of freedom
## Multiple R-squared:  0.1314, Adjusted R-squared:  0.131
## F-statistic: 365.8 on 1 and 2419 DF,  p-value: < 2.2e-16
```

```r
# y = 0.5431x + 32.0195
moms.df <- data.frame(moms,
                      fitted = fitted.values(lin),
```

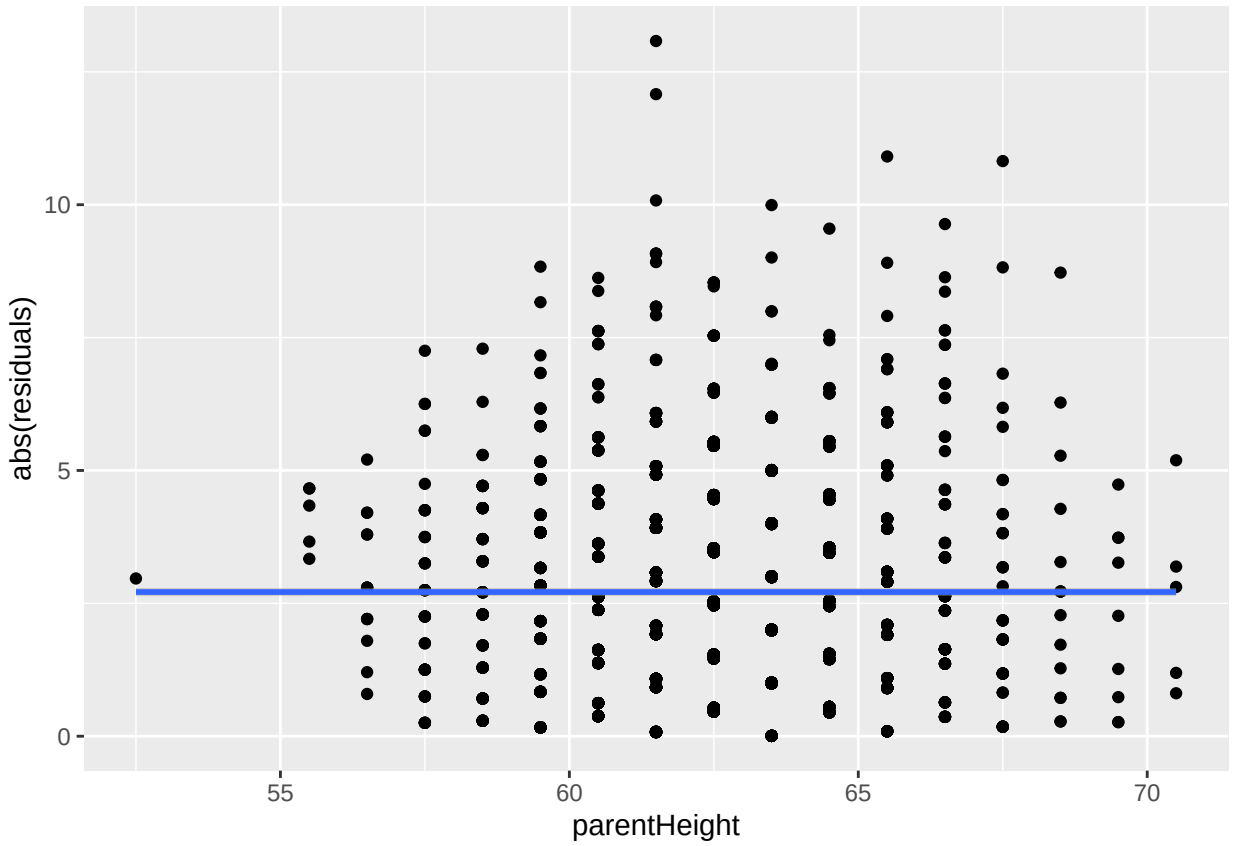
```
                    residuals = residuals(lin))
print(ggplot(moms.df, aes(x=parentHeight, y=residuals)) + geom_point() + geom_smooth())
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

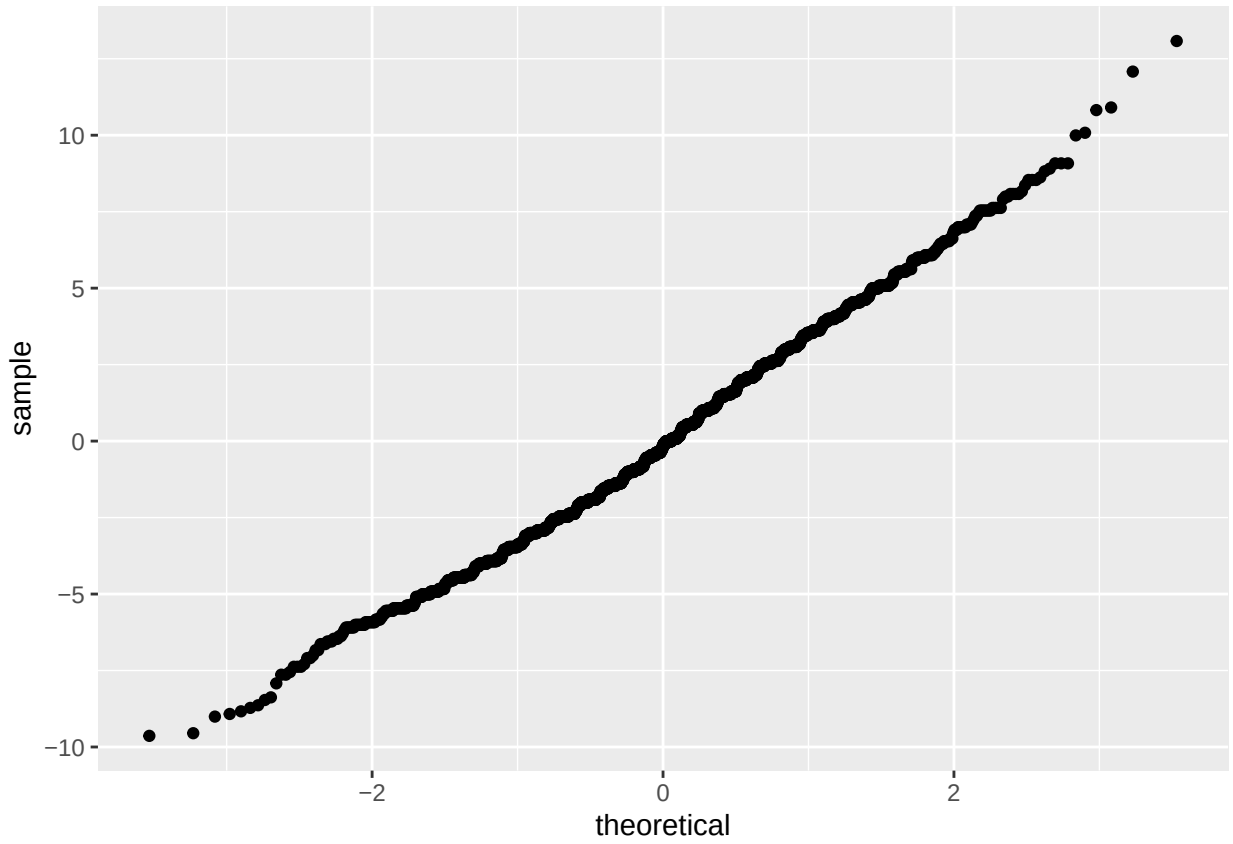



```
print(ggplot(moms.df, aes(x = parentHeight, y = abs(residuals))) + geom_point() + geom_smooth())
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
print(ggplot(moms.df, aes(sample = residuals)) + stat_qq())
```




```
# the assumptions for linear regression are also much easier to confirm with the moms data by itself.


## 4 ##
birds <- read.table("flowersandbirds.txt", header=TRUE)

# a)
m <- mean(birds[,2])
# The mean is 44.64, which is much lower than 57.
# Null: the mean of the difference between flowerGuess and birdGuess is 0.
# Alternative: the mean of the difference between flowerGuess and birdGuess is not 0.
print(t.test(birds[,2], birds[,3], paired = TRUE))
```

```
##
##  Paired t-test
##
## data:  birds[, 2] and birds[, 3]
## t = 0.58631, df = 76, p-value = 0.5594
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.525338  2.798065
## sample estimates:
## mean of the differences
##               0.6363636
```

```
# Test used: Paired t-test
# Test statistic t = 0.58631
# P-value = 0.5594
```

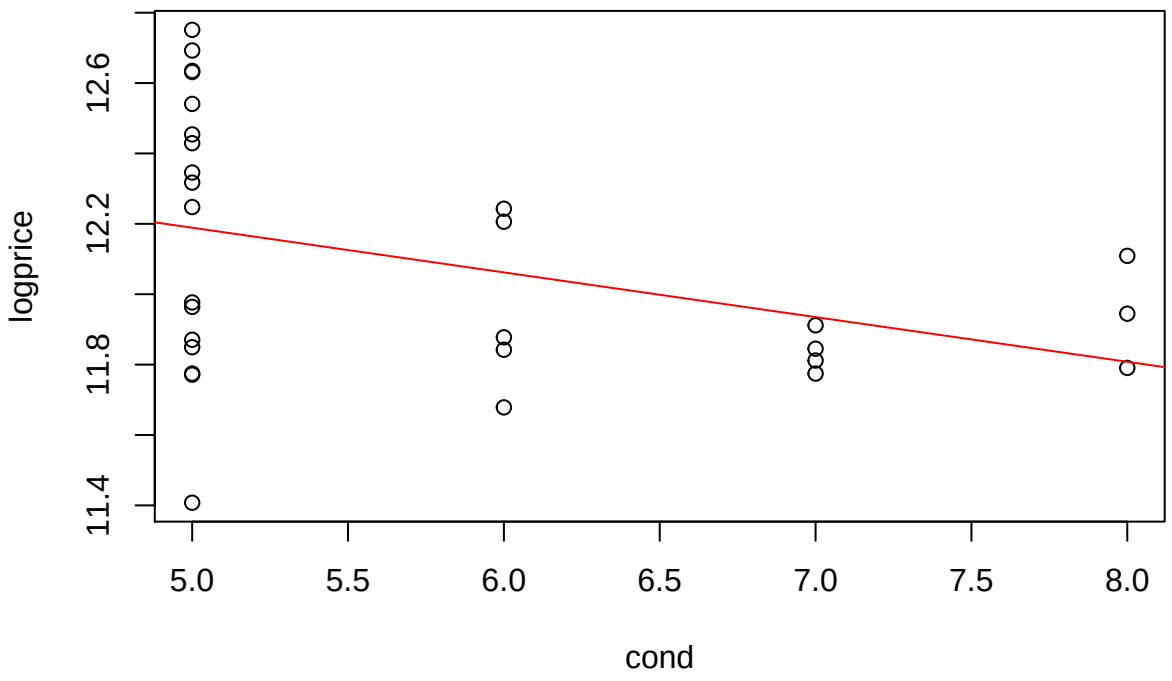
```r
# Substantive conclusion: since p > 0.05, we fail to reject the null and resort to the conclusion that

# b)
print(plot(birds[,2], birds[,3], main="flowerGuess v birdGuess", xlab="flowerGuess", ylab="birdGuess"))
```


```
## NULL
```

```r
# The graph above makes the data look like a strong positive linear relationship between flowerGuess an

# c)
heads <- subset(birds, Group == "Heads")
tails <- subset(birds, Group == "Tails")
head <- heads[,2]-heads[,3]
tail <- tails[,2]-tails[,3]
# Null: expected value for 'head' and 'tail' data sets above are the same
# Alternative: expected value for diff in flowerGuess and birdGuess for heads is stastically different
# the expected value for diff in flowerGuess and birdGuess for tails.
print(t.test(head,tail))
```

```
##
##  Welch Two Sample t-test
##
## data:  head and tail
## t = 1.475, df = 54.284, p-value = 0.146
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.112415  7.308361
## sample estimates:
## mean of x mean of y
##  2.125000 -0.972973
```

```r
# Test used: Welch's two-sample t-test
# t = 1.475
# df = 54.284
# p-value = 0.146
# Conclusion: because p-value > 0.05, we fail to reject the null and assert that the expected values fo
# in flowerGuess and birdGuess between the heads population and tails population are the same.


## 5 ##
iowa <- read.table("IowaHouses.txt", header=TRUE)
price <- iowa[,1]
logprice <- log(price)
cond <- iowa[,2]

# a)
lin <- lm(logprice ~ cond, data = iowa)
# Formula: y = -0.127 + 12.824
plot(cond, logprice)
abline(a = 12.824, b = -0.127, col=2)
```

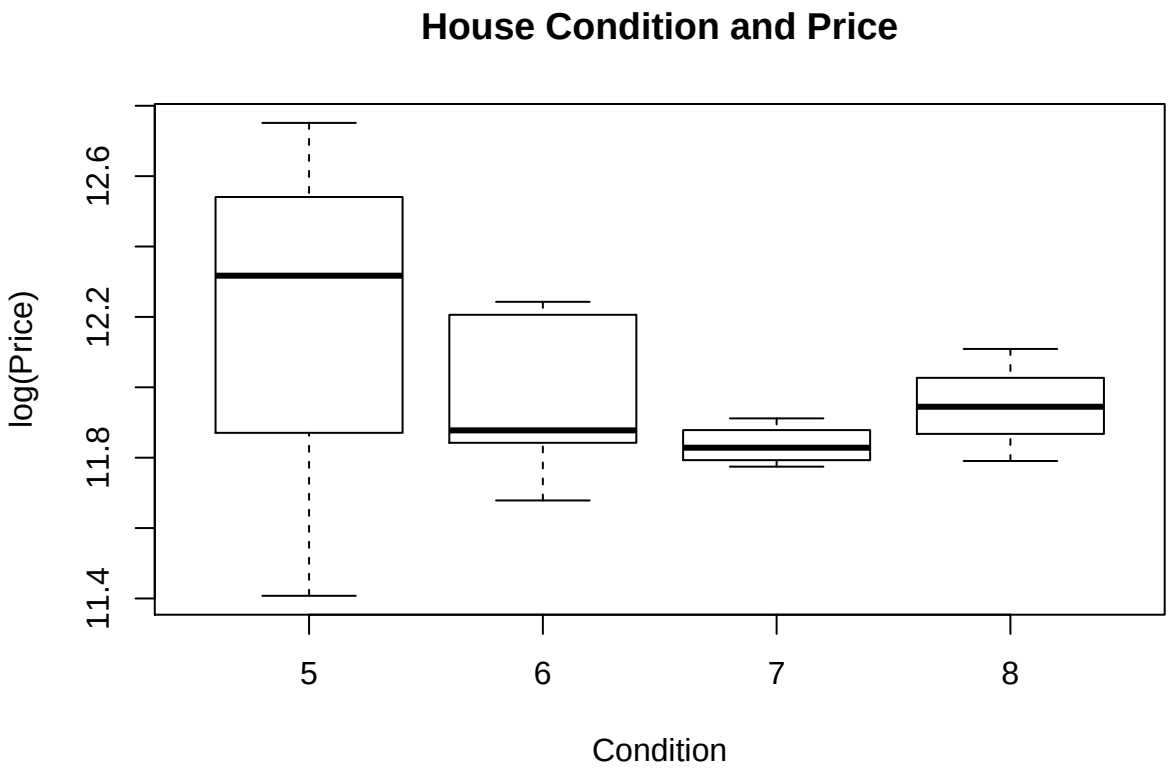



```r
print(summary(lin))
```

```
##
## Call:
## lm(formula = logprice ~ cond, data = iowa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78166 -0.21997 -0.01759  0.23999  0.56207
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.82436    0.34692  36.966   <2e-16 ***
```

```
## cond        -0.12703    0.05929  -2.143   0.0413 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3317 on 27 degrees of freedom
## Multiple R-squared:  0.1453, Adjusted R-squared:  0.1137
## F-statistic: 4.591 on 1 and 27 DF,  p-value: 0.04132
```

```
# Since p-value: 0.04132 < 0.05, we reject the null hypothesis and assert the alternative that house pr
# is dependent on house condition.

# b)
boxplot(logprice~cond,data=iowa, main="House Condition and Price",
        xlab="Condition", ylab="log(Price)")
```




```
# Referencing the boxplot, it appears that
cond <- as.factor(cond)
lin <- lm(logprice ~ cond, data = iowa)
print(anova(lin))
```

```
## Analysis of Variance Table
##
## Response: logprice
##           Df  Sum Sq Mean Sq F value Pr(>F)
## cond       3 0.65804 0.21935  1.9455 0.1481
## Residuals 25 2.81863 0.11274
```

```
# The p-value of the variance analysis: 0.1481 > 0.05 so we fail to reject the null hypothesis and conc
# that we do not have enough evidence to conclude that price depends on condition.

# c)
```

```
# (a) The problems with doing linear regression here is that a house can be in good condition, yet smal
# And a house can be in bad condition, but much larger and better location. Obviously the homes that ar
# locales will sell for more than a smaller house that's in better condition. So the problem with this
# we cannot accurately use house condition as a numerical predictor. We are not confident that the 4 as
# regression are met, namely the Homoskedasticity or normalty of errors.
# (b) The problem with doing variance analysis here is that we don't know if we can check off all of th
# of the ANOVA test. For one, it is hard to say if the population follows a normal trend. Also, hard to
# is independent of itself, as real estate is often a bubble sort of a market that follows local trends

# I think despite both of their flaws that the ANOVA variance test is the best to run on this data. I t
# of real estate prices is a better statistic to study because, especially in real estate, there is goi
# of variance for a number of reasons. The ANOVA test builds that into the strategy of analysis whereas
# does not. I also think that it is more valuable to separate the house conditions into seperate catego
# instead of making it numerical, as linear regression does. The condition of a house on a scale of 1-1
# than objective, it should be in a category and not serve as an actual value of measurement. Plus, the
# are slightly more easily met, as we don't have to assume that the data is linear... which it does not
# agree more with the results of the ANOVA test than the linear regression model.
```