4211 Homework 8

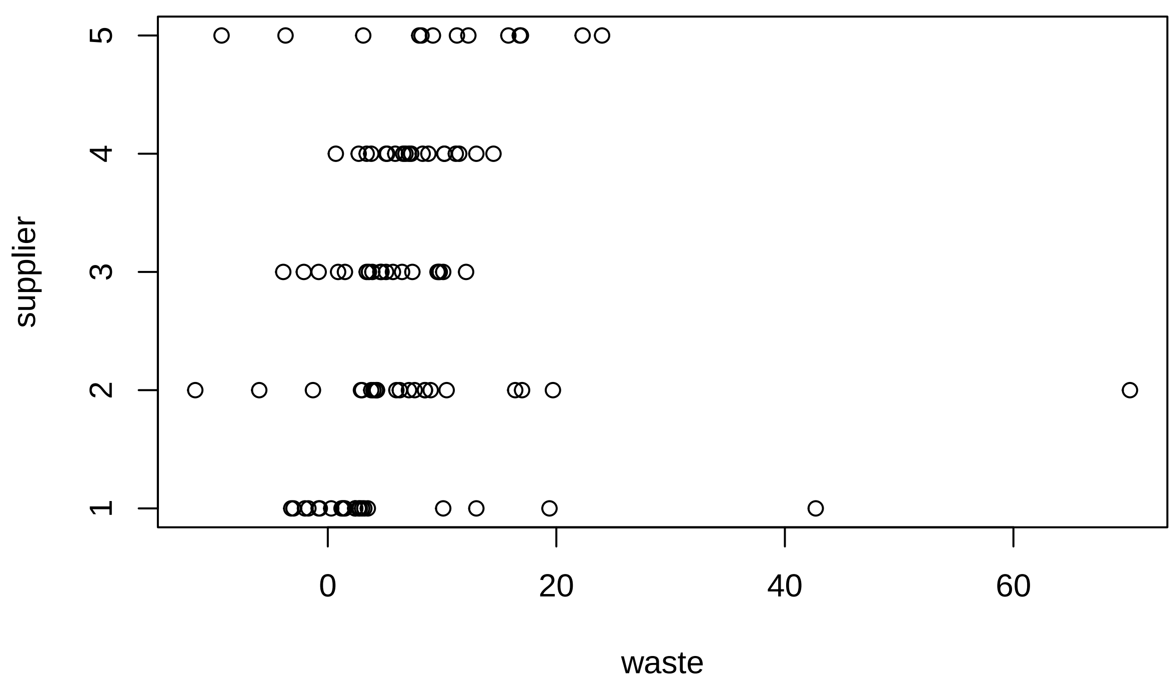
Matthew DeSantis

2023-04-14

1 (10.1)

(a)

```r
data("denim")
plot(denim)
```



Based off of the plot, it seems that supplier 4 may waste the most on average, and that supplier 1 seems to be the least wasteful. There also appear to be two outlying points, one around 70 in for supplier 2, and one close to 40 for supplier 1.

(b)

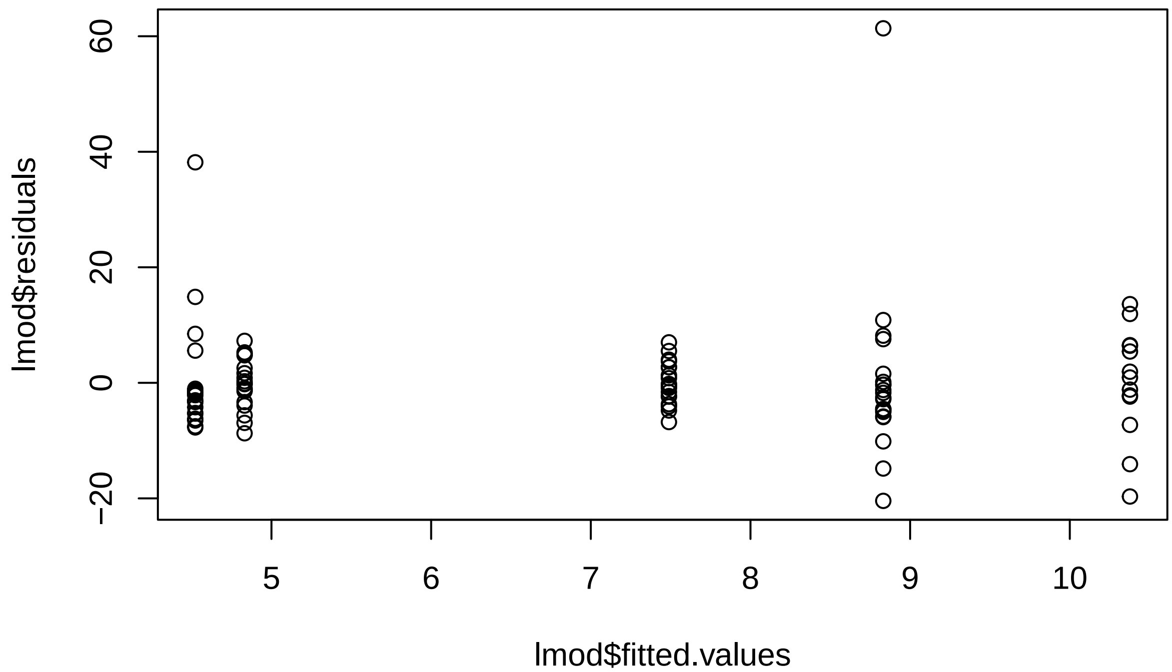
```r
lmod = lm(waste~supplier, data = denim)
summary(lmod)
```

```
##
## Call:
## lm(formula = waste ~ supplier, data = denim)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.432  -4.377  -1.323   2.639  61.368
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.5227     2.1021   2.152   0.0341 *
## supplier2     4.3091     2.9728   1.450   0.1507
## supplier3     0.3089     3.0879   0.100   0.9206
## supplier4     2.9667     3.0879   0.961   0.3392
## supplier5     5.8542     3.4491   1.697   0.0931 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.86 on 90 degrees of freedom
## Multiple R-squared:  0.04901,    Adjusted R-squared:  0.006747
## F-statistic:  1.16 on 4 and 90 DF,  p-value: 0.334
```

According to the fixed effect model, the supplier is not significant.

(c)

```r
plot(lmod$fitted.values,lmod$residuals)
```



Here I plot the residuals vs the fitted values. From the graph, there does not appear to be a constant variance, which indicates that the samples are not independent. In cases like this, we should consider fitting the supplier as a random effect.

(d)

```
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 4.2.2

## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 4.2.2
```

```
rem = lmer(waste~(1|supplier), data = denim)
summary(rem)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: waste ~ (1 | supplier)
##    Data: denim
##
## REML criterion at convergence: 702.1
##
```

```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9095 -0.4363 -0.1669  0.3142  6.3817
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  supplier (Intercept)  0.6711  0.8192
##  Residual             97.3350  9.8658
## Number of obs: 95, groups:  supplier, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)    6.997      1.078    6.49
```

(e)

```
nullrem = lm(waste~1, data = denim)
lrtstat = as.numeric(2*(logLik(rem)-logLik(nullrem)))
pvalue = pchisq(lrtstat, 1, lower.tail = FALSE)
data.frame(lrtstat, pvalue)
```

```
##    lrtstat    pvalue
## 1 1.891748 0.1690049
```

(f)

```
confint(rem, method="boot")
```

```
## Computing bootstrap confidence intervals ...
```

```
##
## 266 message(s): boundary (singular) fit: see help('isSingular')
```

```
##                2.5 %    97.5 %
## .sig01      0.000000  3.433268
## .sigma      8.333369 11.164856
## (Intercept) 4.834909  9.196028
```

(g)

```
denim2 = denim[-c(82, 87),]
rem2 = lmer(waste~(1|supplier), data = denim2)
summary(rem2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: waste ~ (1 | supplier)
```

```
##    Data: denim2
##
## REML criterion at convergence: 603.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.99119 -0.48597 -0.08981  0.49970  2.60002
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  supplier (Intercept)  5.718   2.391
##  Residual             37.292   6.107
## Number of obs: 93, groups:  supplier, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)    6.155      1.246   4.938
```

```
nullrem2 = lm(waste~1, data = denim2)
lrtstat2 = as.numeric(2*(logLik(rem2)-logLik(nullrem2)))
pvalue2 = pchisq(lrtstat2, 1, lower.tail = FALSE)
data.frame(lrtstat2, pvalue2)
```

```
##   lrtstat2    pvalue2
## 1 5.593837 0.01802377
```

```
confint(rem2, method="boot")
```

```
## Computing bootstrap confidence intervals ...
```

```
##
## 52 message(s): boundary (singular) fit: see help('isSingular')
```

```
##                 2.5 %   97.5 %
## .sig01      0.000000 4.476467
## .sigma      5.204829 6.965835
## (Intercept) 3.372327 8.710342
```

After removing the outliers, we can now see that the supplier is significant.

(h)

```
ranef(rem2)
```

```
## $supplier
##   (Intercept)
## 1  -2.6325749
## 2  -0.1872530
## 3  -0.9851799
```

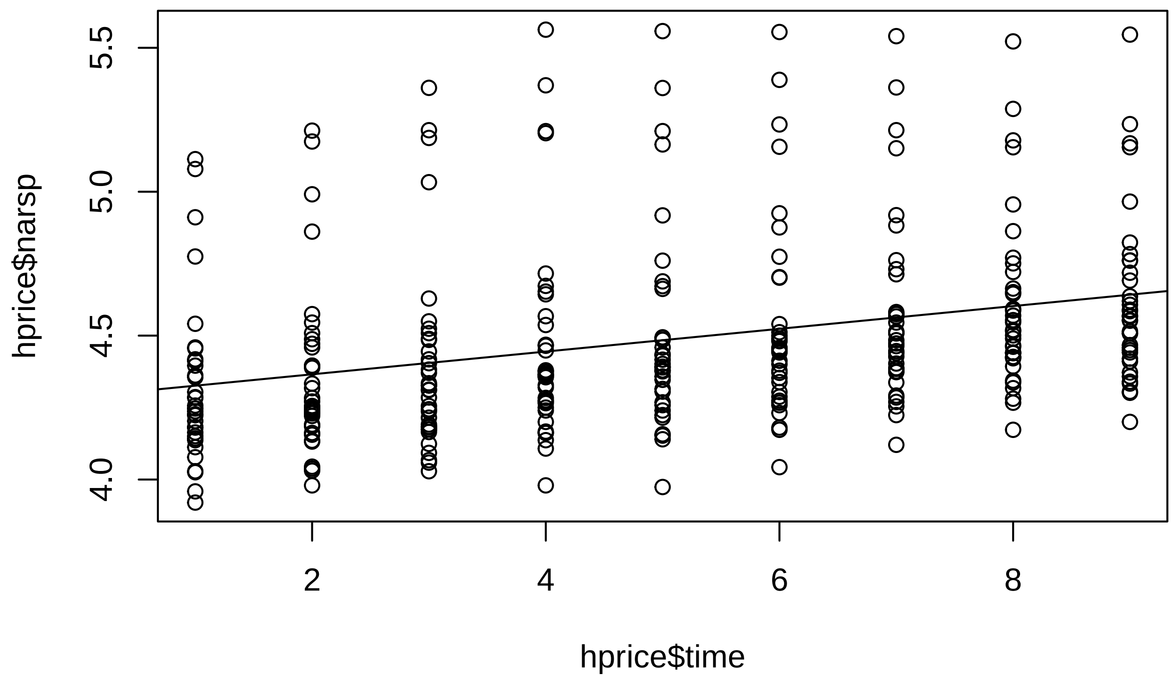
```
## 4   0.9935099
## 5   2.8114979
##
## with conditional variances for "supplier"
```

From the output, we can see that the best supplier is 1, because it has the least waste.

2 (11.2)

(a)

```
data(hprice)
plot(hprice$time, hprice$narsp)
abline(lm(hprice$narsp~hprice$time))
```



The trendline shows (log) prices increasing over time.

(b)

```
summary(lm(narsp~ypc+perypc+regtest+rcdum+ajwtr+time, data = hprice))
```

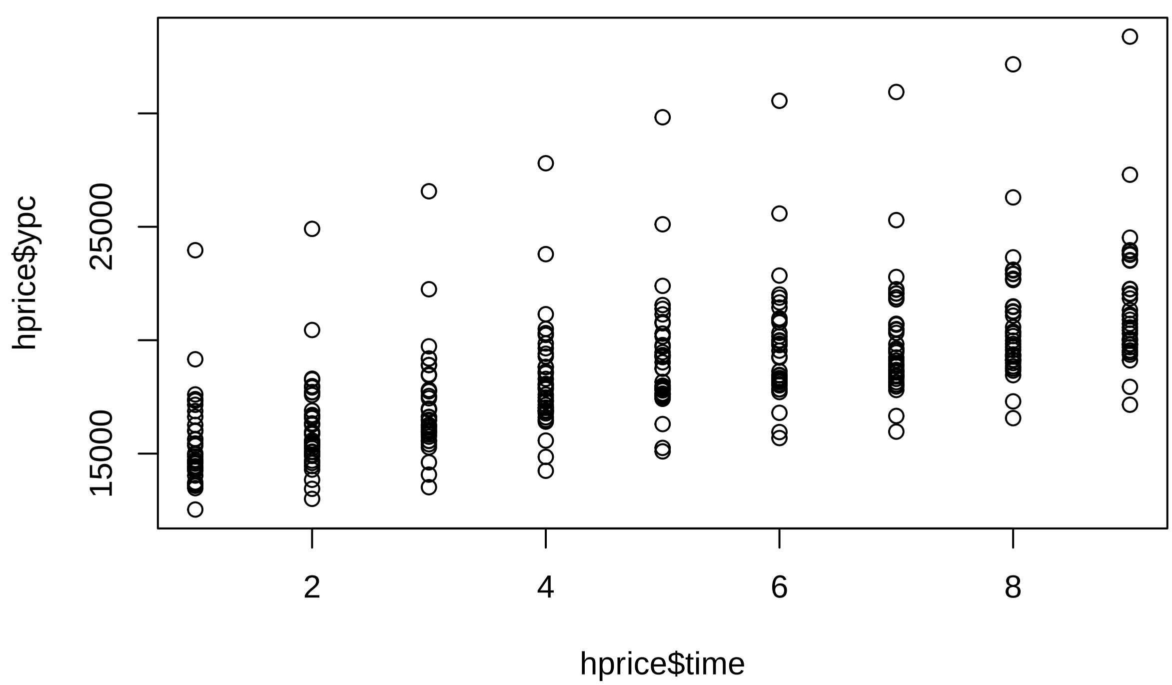
```
##
## Call:
## lm(formula = narsp ~ ypc + perypc + regtest + rcdum + ajwtr +
##     time, data = hprice)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31386 -0.10810 -0.01525  0.08547  0.55594
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.673e+00  8.456e-02  31.612  < 2e-16 ***
## ypc          7.029e-05  4.358e-06  16.128  < 2e-16 ***
## perypc      -1.372e-02  5.074e-03  -2.704 0.007216 **
## regtest      2.954e-02  3.103e-03   9.520  < 2e-16 ***
## rcdum1       1.488e-01  3.235e-02   4.599 6.15e-06 ***
## ajwtr1       3.593e-02  2.001e-02   1.796 0.073482 .
## time        -1.767e-02  5.128e-03  -3.445 0.000647 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1651 on 317 degrees of freedom
## Multiple R-squared:  0.7572, Adjusted R-squared:  0.7526
## F-statistic: 164.7 on 6 and 317 DF,  p-value: < 2.2e-16
```

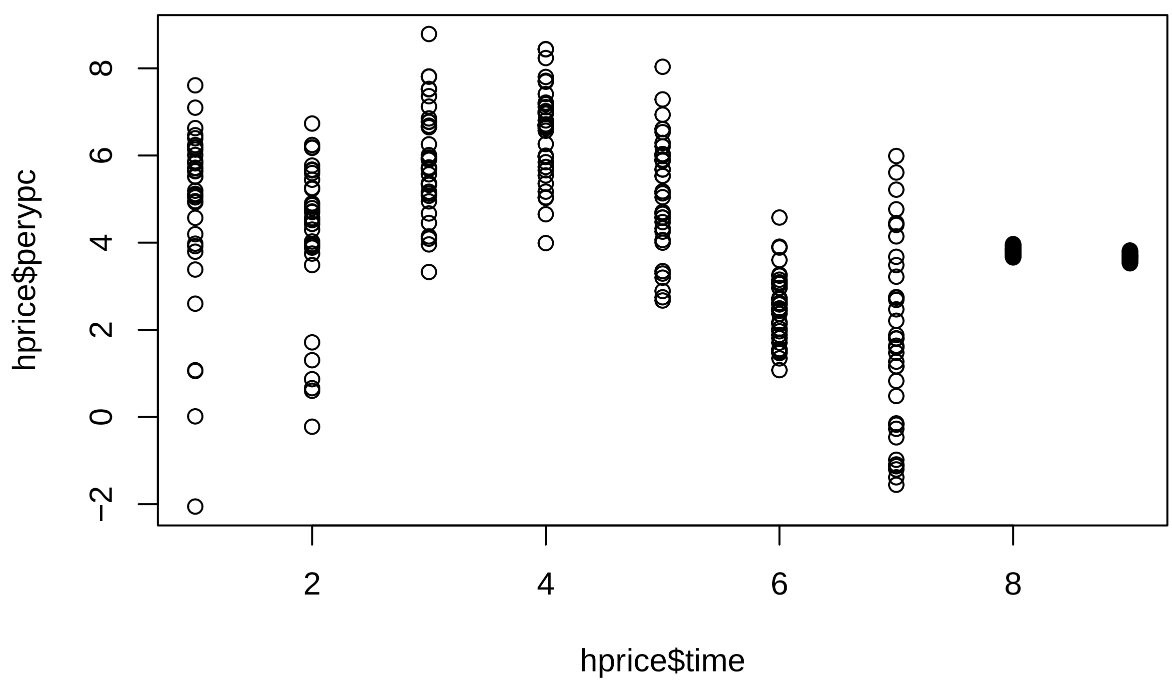
All factors are significant besides being adjacent to a coastline (ajwtr). The coefficient estimate for time is negative, which is unintuitive because earlier we observed prices to be going up with time.

(c)

```
plot(hprice$time, hprice$ypc)
```



```r
plot(hprice$time, hprice$perypc)
```



Per capita income seems to be increasing linearly over time. The percent growth in per capita income, on the other hand, does not display a linear pattern at all. It both increases and decreases. Interestingly, there is a drastic reduction in spread for time periods 8 and 9 when compared to other time periods.

(d)

```r
hprice$start = NA
for(x in 1:nrow(hprice)){
  hprice$start[x] = hprice[(hprice$msa==hprice[x,]$msa) & (hprice$time==1),]$ypc
}

summary(lm(narsp~start+perypc+regtest+rcdum+ajwtr+time, data = hprice))
```

```
##
## Call:
## lm(formula = narsp ~ start + perypc + regtest + rcdum + ajwtr +
##     time, data = hprice)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37278 -0.10546 -0.01540  0.07898  0.53906
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.300e+00  9.666e-02  23.798  < 2e-16 ***
## start        8.865e-05  5.275e-06  16.806  < 2e-16 ***
## perypc      -8.128e-03  4.971e-03  -1.635    0.103
## regtest      3.014e-02  3.036e-03   9.929  < 2e-16 ***
## rcdum1       1.512e-01  3.162e-02   4.782 2.66e-06 ***
## ajwtr1       3.854e-02  1.959e-02   1.967    0.050 .
## time         3.711e-02  3.785e-03   9.803  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1619 on 317 degrees of freedom
## Multiple R-squared:  0.7662, Adjusted R-squared:  0.7618
## F-statistic: 173.2 on 6 and 317 DF,  p-value: < 2.2e-16
```

There are two main points of interest in this new model. First, perypc is no longer significant. Second, the sign on the coefficient for time has changed from negative to positive, which more aligns with our intuition.

(e)

```r
remh = lmer(narsp~start+perypc+regtest+rcdum+ajwtr+time+(1|msa), data = hprice)
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```r
summary(remh)
```

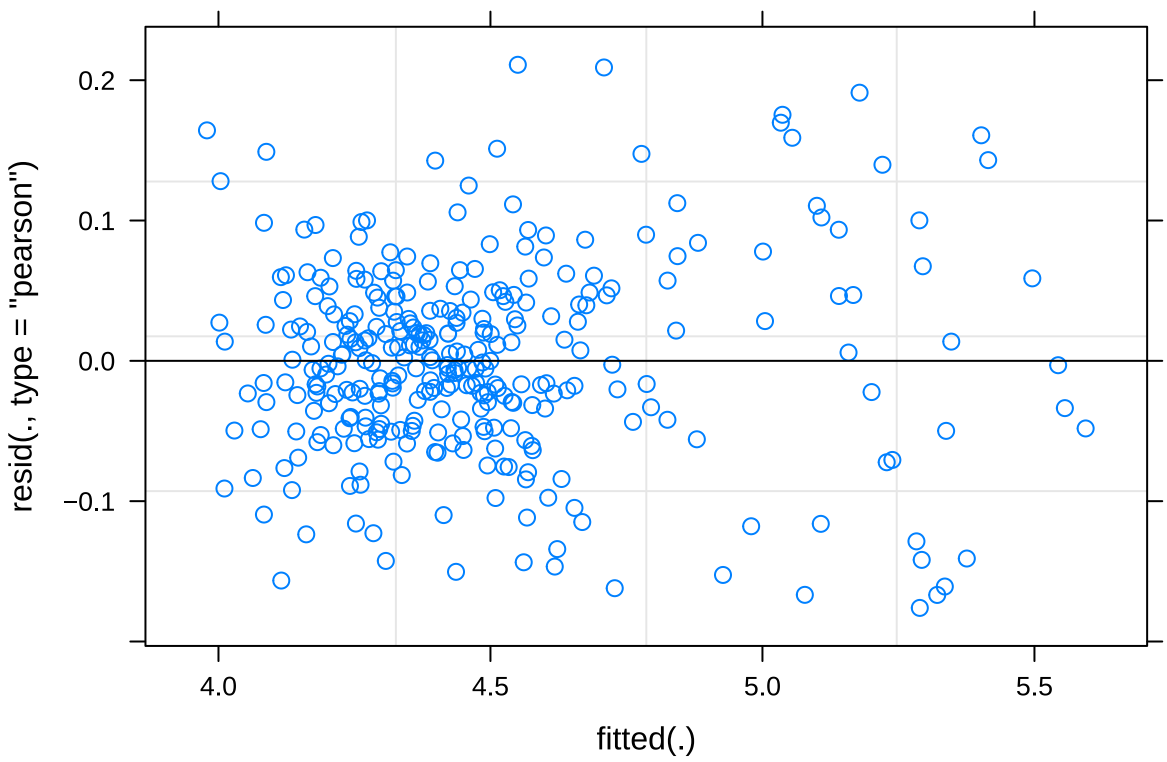
```
## Linear mixed model fit by REML ['lmerMod']
## Formula: narsp ~ start + perypc + regtest + rcdum + ajwtr + time + (1 |
##     msa)
##    Data: hprice
##
## REML criterion at convergence: -581.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.38462 -0.59378 -0.02502  0.57411  2.85776
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  msa      (Intercept) 0.023611 0.15366
##  Residual             0.005451 0.07383
## Number of obs: 324, groups:  msa, 36
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept)  2.306e+00  2.627e-01   8.779
## start        8.864e-05  1.521e-05   5.830
## perypc      -9.148e-03  2.298e-03  -3.981
## regtest      3.016e-02  8.747e-03   3.448
```

```
## rcdum1       1.514e-01  9.111e-02   1.661
## ajwtr1       3.853e-02  5.647e-02   0.682
## time         3.680e-02  1.729e-03  21.282
##
## Correlation of Fixed Effects:
##         (Intr) start  perypc regtst rcdum1 ajwtr1
## start   -0.751
## perypc  -0.049  0.002
## regtest -0.504 -0.174 -0.005
## rcdum1   0.478 -0.332 -0.004 -0.310
## ajwtr1   0.110 -0.182  0.001 -0.040 -0.178
## time    -0.047  0.001  0.395 -0.002 -0.002  0.000
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```

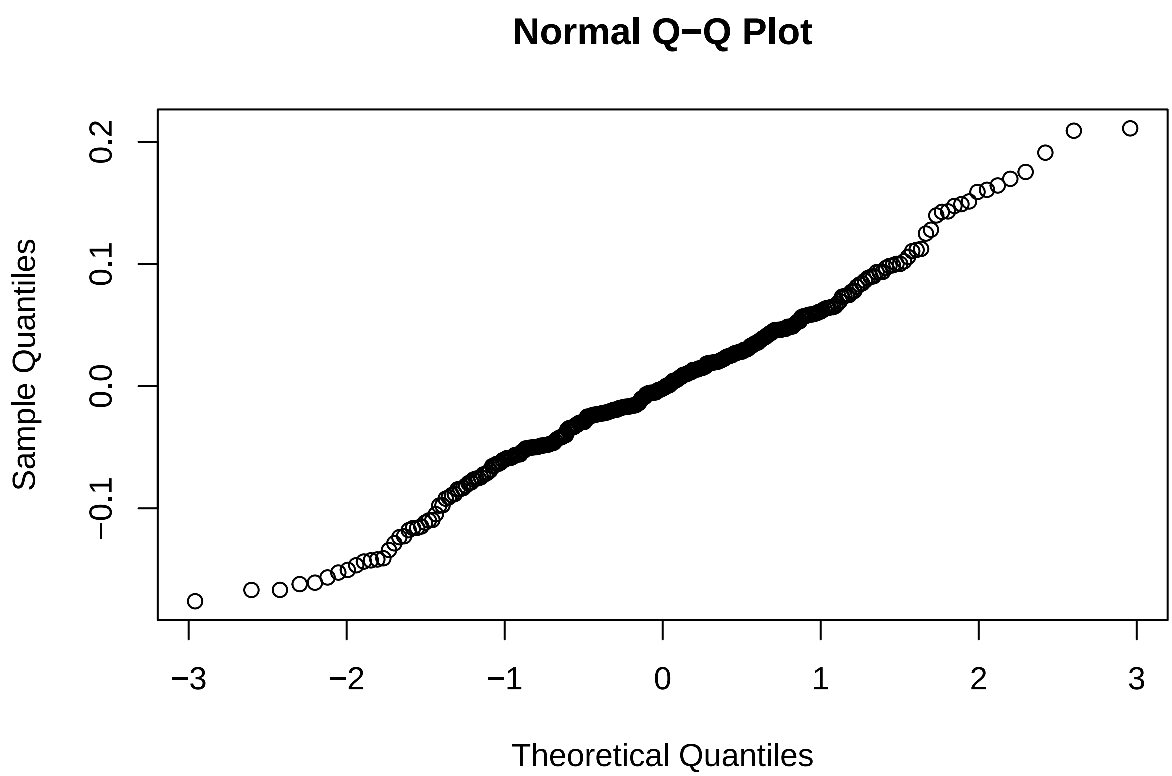
For every year after 1985, the log price of housing (in thousands USD) goes up by 0.0368. This is equivalent to saying that every year, the price of housing is increased by a factor of $\exp(0.0368) = 1.037486$ (about 4% price increase per year).

(f)

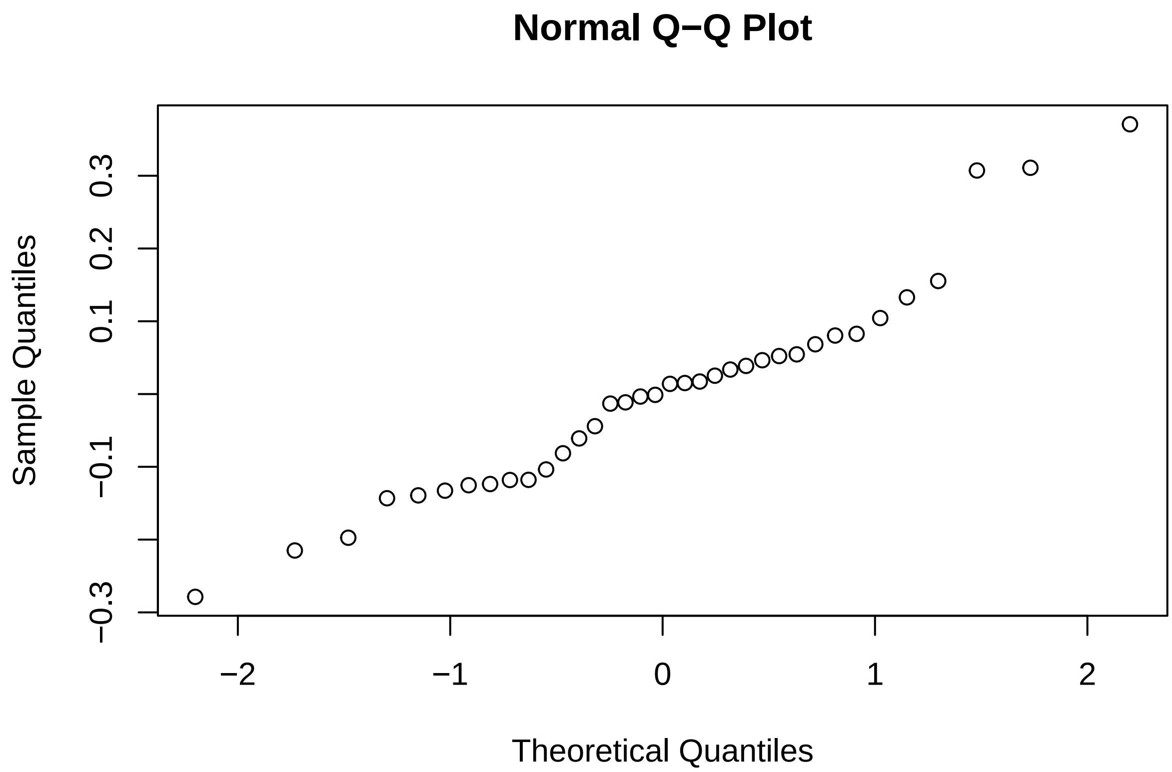
```
plot(remh)
```



```r
qqnorm(residuals(remh))
```




```r
rans = ranef(remh)
qqnorm(rans$msa$`(Intercept)`)
```



The residual plot indicates a linear effect, and the two QQplots show that both our errors and random effects are roughly normally distributed. All this together is good, and indicates that there are no major issues with our model.

(g)

```
remh2 = lmer(narsp~start+perypc+regtest+time+(1|msa), data = hprice)
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
anova(remh, remh2)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: hprice
## Models:
## remh2: narsp ~ start + perypc + regtest + time + (1 | msa)
## remh: narsp ~ start + perypc + regtest + rcdum + ajwtr + time + (1 | msa)
##       npar     AIC     BIC logLik deviance  Chisq Df Pr(>Chisq)
## remh2    7 -625.69 -599.22 319.84  -639.69
## remh     9 -625.80 -591.77 321.90  -643.80 4.1081  2     0.1282
```

An insignificant test result indicates that we can, indeed, drop the terms.

(h)

```r
remh3 = lmer(narsp~start+perypc+regtest+factor(time)+(1|msa), data = hprice)
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```r
anova(remh2, remh3)
```

```
## refitting model(s) with ML (instead of REML)
```

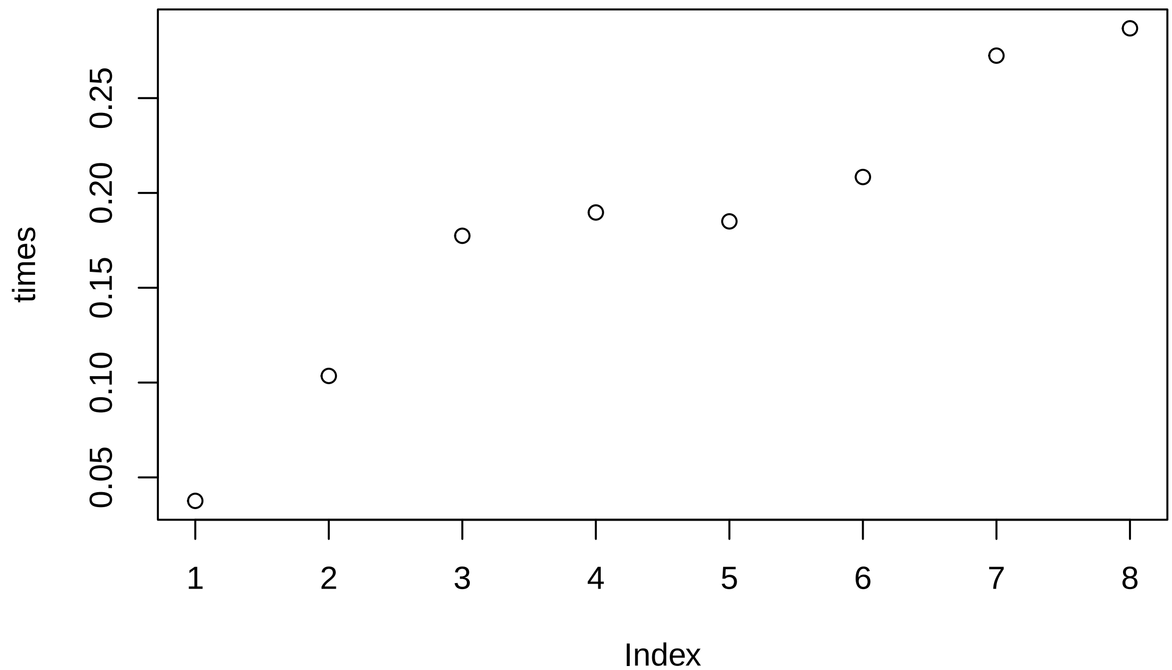
```
## Data: hprice
## Models:
## remh2: narsp ~ start + perypc + regtest + time + (1 | msa)
## remh3: narsp ~ start + perypc + regtest + factor(time) + (1 | msa)
##       npar     AIC     BIC logLik deviance  Chisq Df Pr(>Chisq)
## remh2    7 -625.69 -599.22 319.84  -639.69
## remh3   14 -635.92 -582.99 331.96  -663.92 24.232  7   0.001037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(remh3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: narsp ~ start + perypc + regtest + factor(time) + (1 | msa)
##    Data: hprice
##
## REML criterion at convergence: -566.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2167 -0.5523 -0.0592  0.5879  3.1723
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  msa      (Intercept) 0.025029 0.15820
##  Residual             0.005135 0.07166
## Number of obs: 324, groups:  msa, 36
##
## Fixed effects:
##                 Estimate Std. Error t value
## (Intercept)    2.098e+00  2.313e-01   9.067
## start          1.007e-04  1.424e-05   7.069
## perypc        -1.695e-02  3.042e-03  -5.570
## regtest        3.569e-02  8.507e-03   4.195
## factor(time)2  3.762e-02  1.698e-02   2.215
## factor(time)3  1.035e-01  1.726e-02   5.996
## factor(time)4  1.774e-01  1.770e-02  10.023
## factor(time)5  1.897e-01  1.693e-02  11.204
## factor(time)6  1.850e-01  1.820e-02  10.166
## factor(time)7  2.084e-01  1.917e-02  10.873
```

```
## factor(time)8  2.724e-01  1.714e-02  15.894
## factor(time)9  2.868e-01  1.722e-02  16.657
##
## Correlation of Fixed Effects:
##              (Intr) start  perypc regtst fct()2 fct()3 fct()4 fct()5 fct()6
## start        -0.697
## perypc       -0.057  0.001
## regtest      -0.416 -0.349 -0.008
## factor(tm)2 -0.042  0.000  0.103 -0.001
## factor(tm)3 -0.024  0.000 -0.207  0.002  0.465
## factor(tm)4 -0.018  0.000 -0.300  0.002  0.444  0.529
## factor(tm)5 -0.032  0.000 -0.070  0.001  0.489  0.502  0.497
## factor(tm)6 -0.055  0.000  0.372 -0.003  0.500  0.377  0.331  0.437
## factor(tm)7 -0.059  0.000  0.473 -0.004  0.487  0.333  0.279  0.406  0.585
## factor(tm)8 -0.046  0.000  0.171 -0.001  0.508  0.447  0.419  0.480  0.521
## factor(tm)9 -0.047  0.000  0.195 -0.002  0.508  0.440  0.410  0.476  0.528
##              fct()7 fct()8
## start
## perypc
## regtest
## factor(tm)2
## factor(tm)3
## factor(tm)4
## factor(tm)5
## factor(tm)6
## factor(tm)7
## factor(tm)8  0.515
## factor(tm)9  0.524  0.516
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```

```r
times = c(0.03762, 0.1035, 0.1774, 0.1897, 0.185, 0.2084, 0.2724, 0.2868)
plot(times)
```



A significant result from the test shows that we shouldn't prefer this model. The plot shows that while the effect over time may not be perfectly linear, it's close enough to not be worth complicating the model.

(i)

The log price (in thousands USD) increases by 0.0001007 for each dollar of average per capita income in 1986. The log price (in thousands USD) falls by 0.01695 for each percentage point that the average per capita income increases in a year. The log price (in thousands USD) increases by 0.03569 for each point in the regulatory environment index (higher amounts of regulation mean higher prices).