3/4/2015 Class Notes

HW #6 Review

Exercise #1

$Yi = B_o + B_1 X_i + E_i$

$E_i \sim N(0, var)$

A) lm1 = lm(expensive~cheap, data=shocks)
summary(lm1)

```
Multiple R-squared:  0.9344
```

Cheap model and expensive screen seem to be correlated looking at the plot
R^2 = .93
.93 > .90, so the company will use the cheap model

B) Two criteria:

a. confint(lm1)

```
   2.5 %    97.5 %
(Intercept) -31.7099020 67.306530
cheap         0.8917277  1.076474
```

Slope of “cheap” = 1 is within the 95% confidence interval => test passes

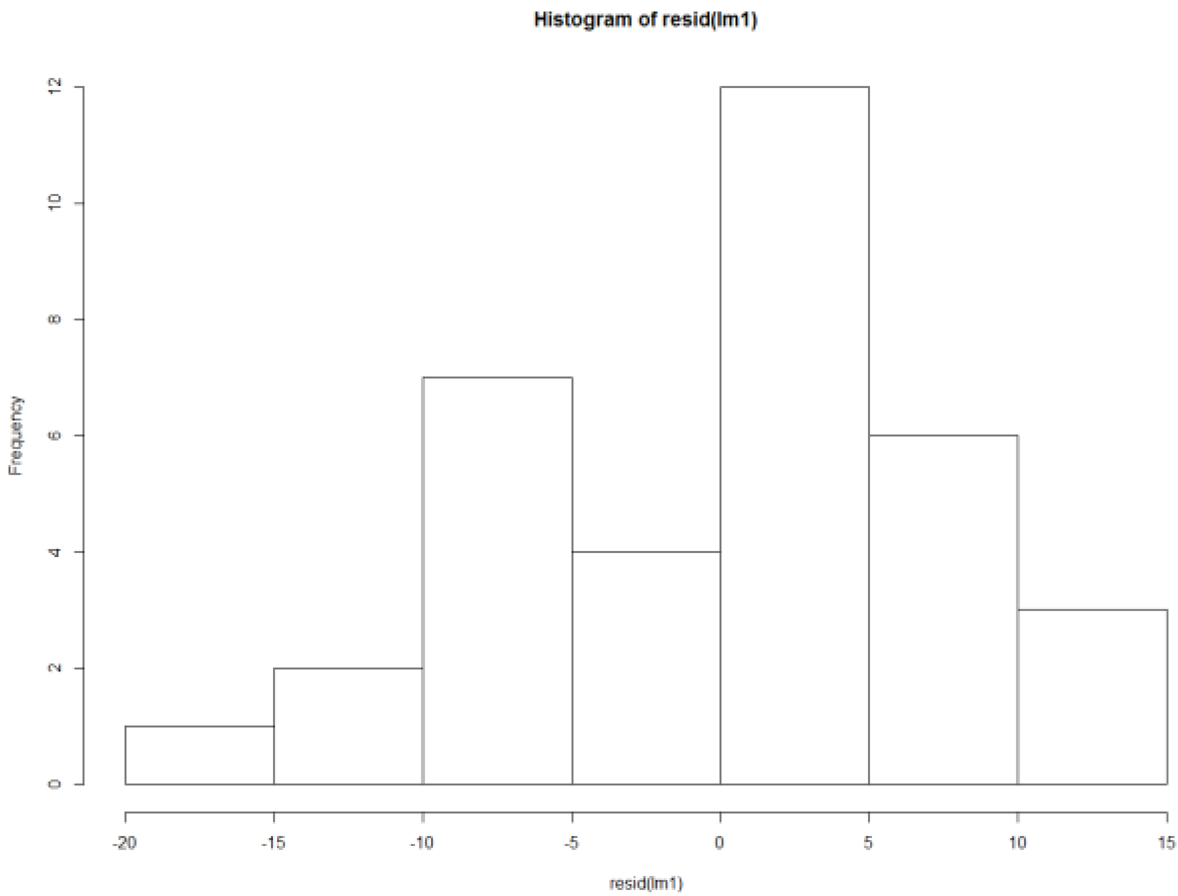
b. new_shocks = data.frame(cheap = c(510, 550, 590))
predict(lm1, new_shocks, interval='prediction', level = 0.95)

```
fit      lwr      upr
1 519.6896 503.4111 535.9682
2 559.0537 542.8839 575.2234
3 598.4177 581.5288 615.3066
```

Make a data frame for the new x values (510, 550, 590)
Predict new shockers using these values
The “fit” refers to plug-in-values
Colum 3 has a width greater than 33 => test fails

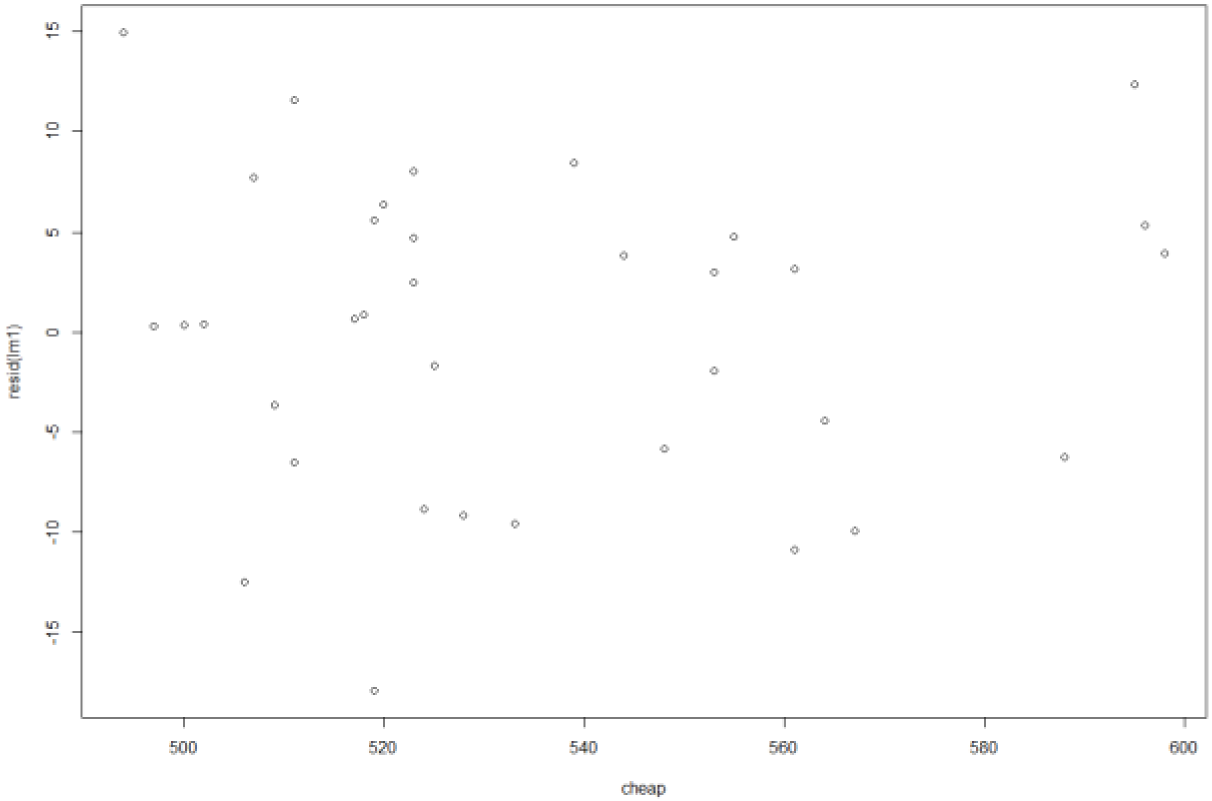
Lastly, check the normality assumptions:

hist(resid(lm1))



The histogram has no extreme outliers so it can be assumed to be normally, especially since there aren't too many data points

plot(resid(lm1) ~ cheap, data=shocks)



Plot of the residuals seems random (no clumps, fan shape, identifiable curve) so the residuals can be assumed to be normal

Exercise #2

Even though undercount isn't a given variable, another variable might be a good proxy for undercount

georgia2000$undercount = 100*(georgia2000$ballots - georgia2000$votes)/georgia2000$ballots

georgia2000$repshare = georgia2000$bush/georgia2000$ballots

These two variables make it easier to calculate undercount percentage and the percent of votes that went towards republicans (and 1 – repshare = votes towards democrats)

Model Building Strategy

1) Make plots (box, scatter) to generate ideas/test preconceived ideas
2) Start somewhere! Use the exploratory data analysis done in step one to find an initial conclusion
3) Play around with the model
    a. Add/delete variables

Step One: Exploratory Analysis

boxplot(undercount ~ equip, data=georgia2000)

boxplot(undercount ~ poor, data=georgia2000)

boxplot(undercount ~ urban, data=georgia2000)

boxplot(undercount ~ atlanta, data=georgia2000)

plot(undercount ~ perAA, data=georgia2000)

plot(undercount ~ repshare, data=georgia2000)

lm1 = lm(undercount ~ poor + urban + atlanta + perAA + repshare + equip, data=georgia2000)

summary(lm1)

anova(lm1)

confint(lm1)

```
Coefficients:
             Estimate Std. Error t value
(Intercept)   13.4347     2.7547   4.877
poor           1.9494     0.4499   4.333
urban         -0.7198     0.4792  -1.502
atlanta       -1.0165     0.6838  -1.487
perAA         -8.0727     2.6211  -3.080
repshare     -15.0330     3.9086  -3.846
equipOPTICAL   1.0831     0.3897   2.779
equipPAPER    -1.3812     1.5303  -0.903
equipPUNCH     1.5578     0.6284   2.479
```

Equipment seems to cause ~1% change either way from the baseline (Lever)

```
Response: undercount
           Df Sum Sq Mean Sq F value    Pr(>F)
poor        1 160.28 160.283 35.6994 1.607e-08
urban       1  13.61  13.605  3.0302  0.083777
atlanta     1   2.54   2.536  0.5648  0.453525
perAA       1   0.66   0.661  0.1473  0.701683
repshare    1  80.89  80.888 18.0159 3.823e-05
equip       3  53.33  17.776  3.9592  0.009444
Residuals 150 673.47   4.490
```

Equipment seems to have the third largest effect on the sum of squares

Check robustness of the initial conclusion by adding/removing variables

lm2 = lm(undercount ~ poor + urban + atlanta + repshare + equip, data=georgia2000)

summary(lm2)

anova(lm2)

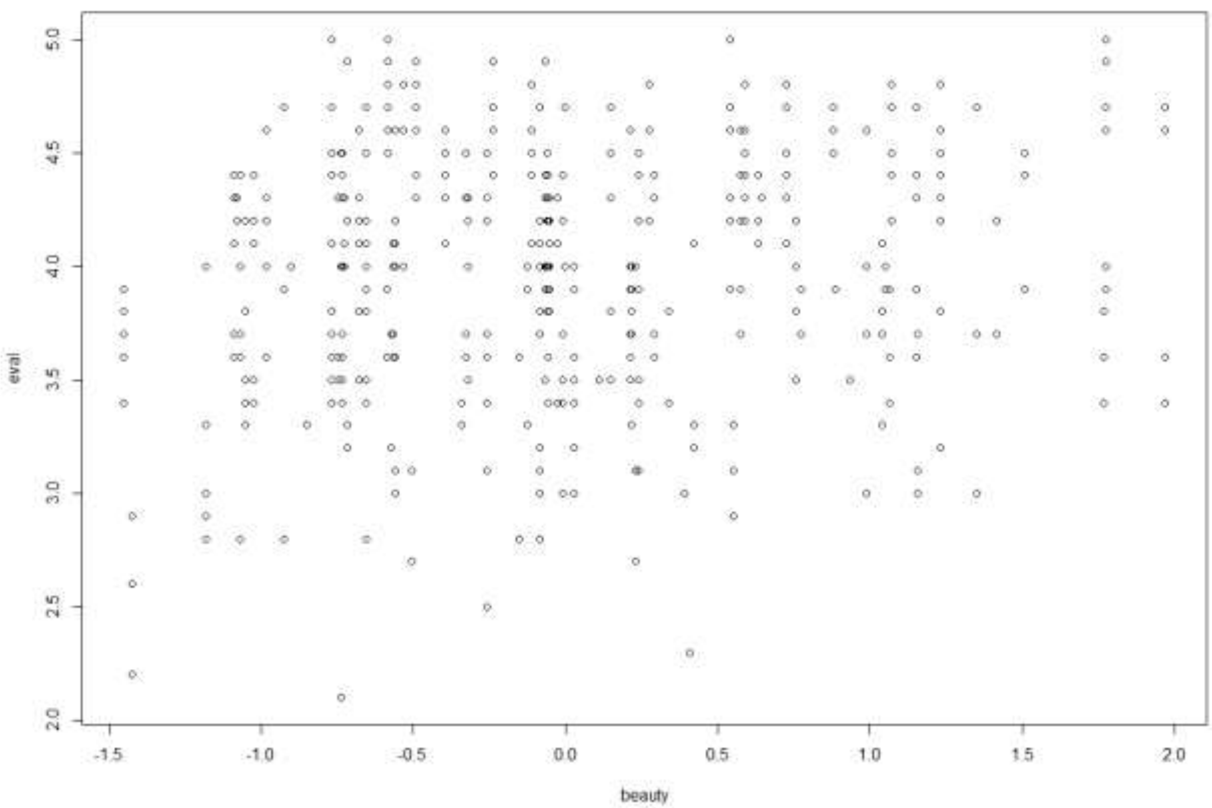
```
(Intercept)     5.7651     1.2104   4.763
poor            1.5587     0.4436   3.514
urban          -0.8420     0.4908  -1.716
atlanta        -0.9857     0.7027  -1.403
repshare       -4.4705     1.9271  -2.320
equipOPTICAL    1.3058     0.3936   3.318
equipPAPER     -1.3389     1.5726  -0.851
equipPUNCH      1.5302     0.6458   2.370
```

```
Response: undercount
           Df Sum Sq Mean Sq F value    Pr(>F)
poor        1 160.28 160.283 33.7999 3.512e-08
urban       1  13.61  13.605  2.8690  0.092363
atlanta     1   2.54   2.536  0.5347  0.465765
repshare    1  25.34  25.345  5.3446  0.022139
equip       3  66.94  22.314  4.7056  0.003604
Residuals 151 716.06   4.742
```

Model seems to stay consistent. But, taking out equipment reduces R^2 by 7%. Main question is, is 7% significant? (come back to this later)

Exercise #3

Scale for this question is the deviation from the average



First, make plots and do an exploratory analysis…

plot(eval ~ beauty, data=profs)

boxplot(eval ~ minority, data=profs)

plot(eval ~ age, data=profs)

boxplot(eval ~ gender, data=profs)

boxplot(eval ~ division, data=profs)

plot(eval ~ log(students), data=profs)

boxplot(eval ~ credits, data=profs)

boxplot(eval ~ tenure, data=profs)

boxplot(eval ~ native, data=profs)

Second, create a model and start somewhere…

lm1 = lm(eval ~ native + tenure + credits + log(students) + gender + minority + beauty, data=profs)

summary(lm1)

anova(lm1)

```
Coefficients:
              Estimate Std. Error t value
(Intercept)    3.82598    0.15043  25.433
nativeyes      0.25356    0.10604   2.391
tenureyes     -0.04420    0.06198  -0.713
creditssingle  0.60035    0.11215   5.353
log(students) -0.04583    0.03182  -1.440
gendermale     0.18003    0.04933   3.649
minorityyes   -0.16074    0.07628  -2.107
beauty         0.17052    0.03087   5.523
```

One added point in beauty increases the evaluation by ~.17

```
Response: eval
               Df  Sum Sq Mean Sq F value
native          1   2.845  2.8453 10.8350
tenure          1   1.767  1.7674  6.7304
credits         1   6.183  6.1832 23.5459
log(students)   1   0.111  0.1111  0.4231
gender          1   2.954  2.9543 11.2502
minority        1   0.882  0.8815  3.3569
beauty          1   8.011  8.0110 30.5061
Residuals     455 119.485  0.2626
```

And beauty seems to be the biggest contributor to the sum of squares in the model

Q: Why do we add the variable being studied last?

A: Because the analysis of variance is order-dependent so adding the variable to be studied last accounts for every other variable beforehand

In a sense, we're playing devil's advocate for the beauty variable by saying that everything else might be more significant

Q: Why not add every variable given into a model from the get-go?

A: Because sometimes the relationship between highly correlated variables can skew the model

Permutation Test

Main question to answer is: "How significant is this value? Could this have happened due to chance?"

To test, shuffle the cards!

Values far from the median and outside of a few standard deviations are either absurd miracles or prove that the H_0 is wrong.

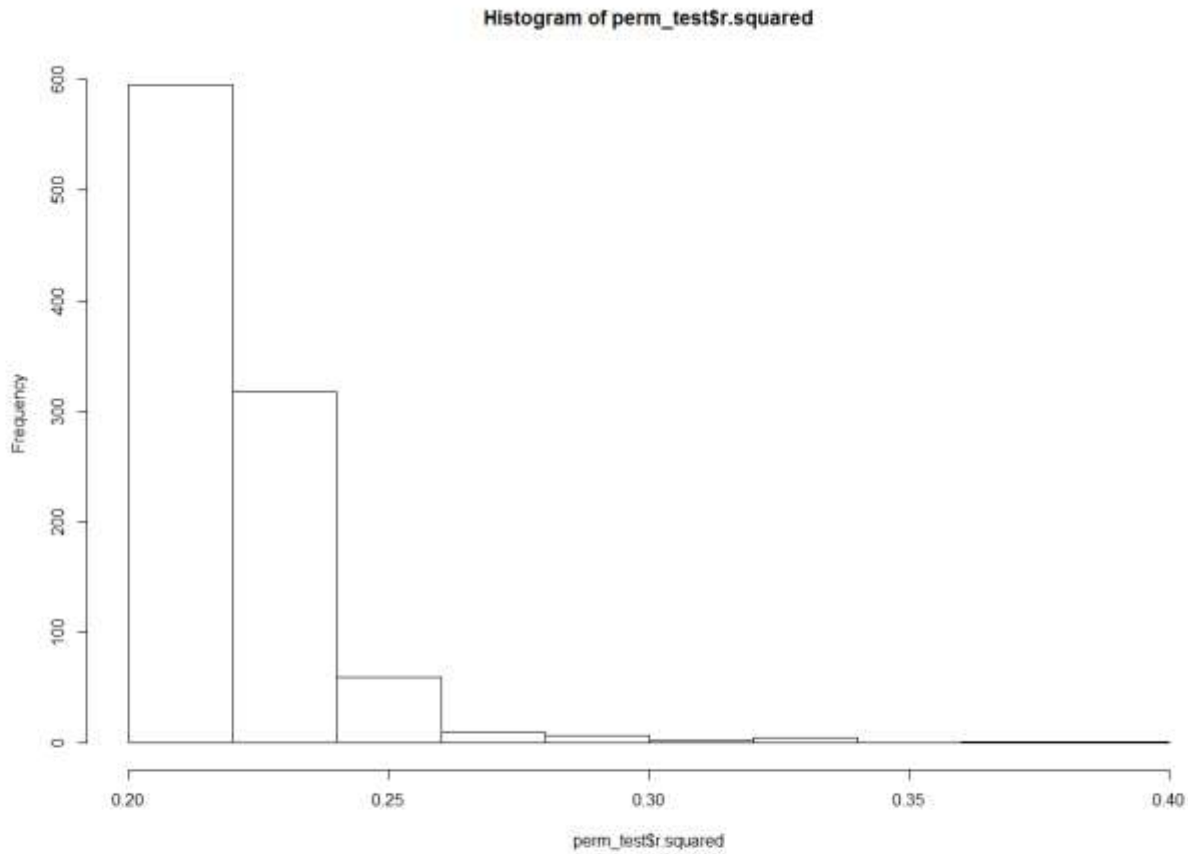
Steps for a Permutation Test:

1) Identify H_0 – nothing special is happening, everything is due to chance
2) Choose a test statistic
   a. E.g. relative risk, odds ratio, anything that is sensitive to departures from the mean
3) Simulate (calculate the test statistic) – $P(T|H_0)$
   a. Get a sampling distribution of the test statistic under the assumption that the null hypothesis is true.
4) Check whether your test statistic is consistent with $P(T|H_0)$

R^2 is a great test statistic to use because a small deviation often means the null could be true while a large deviation gives credibility to the idea of the null being false.

Back to Exercise 3

```
perm_test = do(1000)*{
  lm_perm = lm(undercount ~ poor + urban + atlanta + repshare + shuffle(equip), data=georgia2000)
  lm_perm
}
hist(perm_test$r.squared)
```



This histogram of the shuffled model for undercount shows that the mean R^2 is near 21-22%. A value of 27%, as seen in the model with equip not being shuffled, is far from the norm, meaning that equipment is a significant variable in the model.