David Bieniakowski, Luke Halecki, Darren Rodoff

Dr. W. Fu

Statistical Modeling

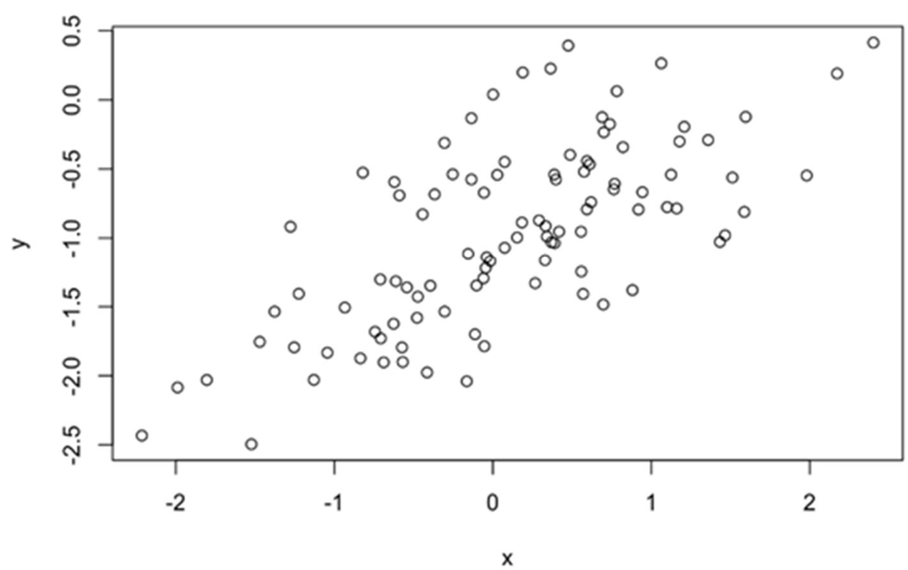
September 21, 2022

Teamwork Formal Presentation and Submission Problems 1

Problem 13:

c) y is of length 100. β_0 is -1 and β_1 is 0.5

d) The plot of x and y is below



There is a positive relationship between x and y. The estimates for both betas seem to be very close to the actual values.

e) The summary of the fit is below:

Call:

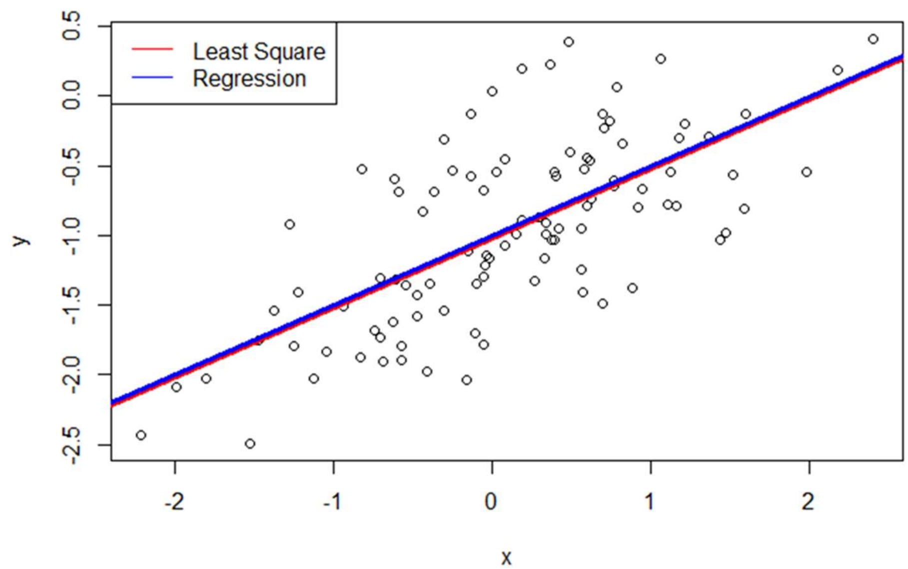
lm(formula = y ~ x)

Residuals:

```
    Min       1Q   Median
-0.93842 -0.30688 -0.06975
     3Q      Max
 0.26970  1.17309

Coefficients:
            Estimate Std. Error
(Intercept) -1.01885    0.04849
x            0.49947    0.05386
            t value Pr(>|t|)
(Intercept) -21.010  < 2e-16 ***
x             9.273 4.58e-15 ***
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’
  0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4814 on 98 degrees of freedom
Multiple R-squared:  0.4674,	Adjusted R-squared:  0.4619
F-statistic: 85.99 on 1 and 98 DF,  p-value: 4.583e-15
```

The linear regression fits a model close to the true value of the coefficients as was constructed. The model has a large F-statistic with a p-value close to 0 so the H_0 can be rejected.

f) The plot with the 2 model lines is shown below:



g) There is evidence that model fit has increased over the training data given the slight increase in R^2 and RSE. However, the p-value of the t-statistic suggests that there isn't a relationship between y and x^2. The summary of fit_sq is shown below:

```
Call:

lm(formula = y ~ x + I(x^2))


Residuals:

    Min       1Q   Median       3Q      Max

-0.98252 -0.31270 -0.06441  0.29014  1.13500


Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.97164    0.05883 -16.517  < 2e-16 ***

x            0.50858    0.05399   9.420  2.4e-15 ***

I(x^2)      -0.05946    0.04238  -1.403    0.164

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.479 on 97 degrees of freedom
```

Multiple R-squared: 0.4779,	Adjusted R-squared: 0.4672

F-statistic: 44.4 on 2 and 97 DF, p-value: 2.038e-14

h) The error seen in R^2 and the RSE both decrease significantly, which is expected. The summary and the plot for lm.fit are shown below:

Call:

lm(formula = y1 ~ x1)

Residuals:

     Min   1Q   Median   3Q   Max

-0.136567 -0.028264 0.001012 0.031550 0.131670

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.998814   0.005173 -193.09   <2e-16 ***

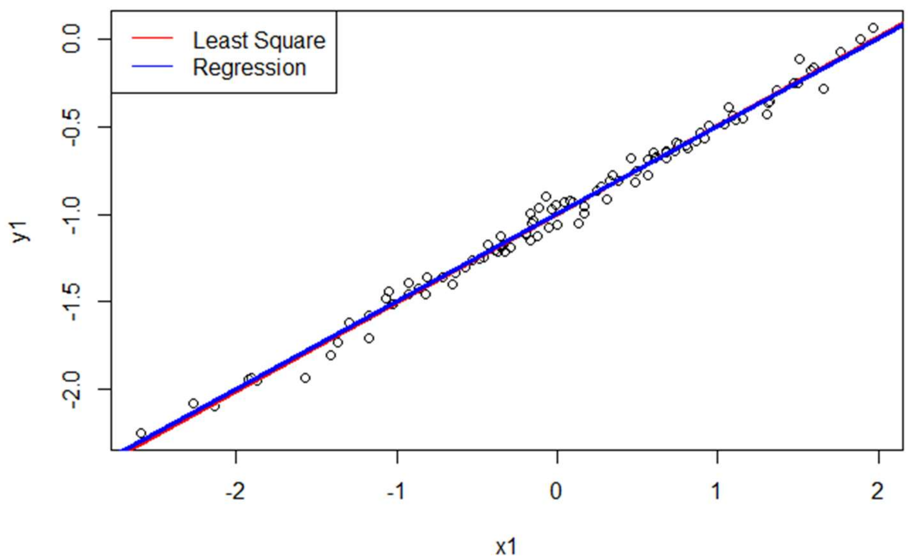
x1   0.505777   0.005235   96.61   <2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05166 on 98 degrees of freedom

Multiple R-squared: 0.9896,	Adjusted R-squared: 0.9895

F-statistic: 9333 on 1 and 98 DF, p-value: < 2.2e-16



i) The error seen in R^2 and the RSE both increase significantly from part h), which is expected. The summary and the plot for lm.fit2 are shown below:

Call:

lm(formula = y2 ~ x2)

Residuals:

```
    Min       1Q   Median       3Q      Max 
-1.16208 -0.30181  0.00268  0.29152  1.14658 
```

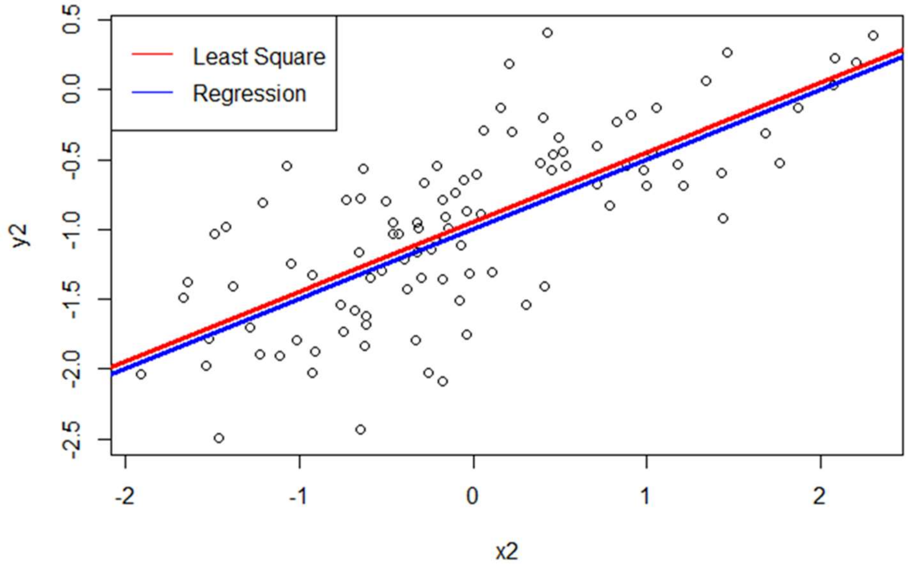
Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.94557    0.04517  -20.93   <2e-16 ***
x2           0.49953    0.04736   10.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 0.4514 on 98 degrees of freedom

Multiple R-squared: 0.5317,   Adjusted R-squared: 0.5269

F-statistic: 111.2 on 1 and 98 DF, p-value: 2.2e-16



j) All 3 intervals seem to be centered on about 0.5, with the second fit's interval being narrowest and the last fit's interval being widest. All three intervals are printed below in order.

```
            2.5 %     97.5 %
(Intercept) -1.1150804 -0.9226122
x           0.3925794  0.6063602
> confint(lm.fit)
            2.5 %     97.5 %
(Intercept) -1.0090795 -0.9885493
x1          0.4953877  0.5161661
> confint(lm.fit2)
            2.5 %     97.5 %
(Intercept) -1.0352203 -0.8559276
x2          0.4055479  0.5935197
```

Conclusion Paragraph: The above data showcases that all three of the models have similar performances, and we can see that in the confidence intervals above that are all centered on approximately 0.5. The narrowest interval is the second model's and the widest interval is the first model's, with the third model only slightly more narrow than the first one. This leads us to the conclusion that as the noise increases the interval widens and the model becomes less predictable, and as the noise decreases the interval becomes narrower and the model becomes more predictable.

Supplemental:

To analyze the data, we used a linear model with all variables involved and a correlation of all variables. Some predictors that look important are:

- zn (proportion of residential land zoned for lots over 25,000 sq.ft.),
- Dis (weighted mean of distances to five Boston employment centres.),
- Rad (index of accessibility to radial highways.),
- and medv (median value of owner-occupied homes in $1000s.).

	crim	zn	indus	chas	nox	rm	age	dis
crim	1.00000000	-0.20046922	0.40658341	-0.055891582	0.42097171	-0.21924670	0.35273425	-0.37967009
zn	-0.20046922	1.00000000	-0.53382819	-0.042696719	-0.51660371	0.31199059	-0.56953734	0.66440822
indus	0.40658341	-0.53382819	1.00000000	0.062938027	0.76365145	-0.39167585	0.64477851	-0.70802699
chas	-0.05589158	-0.04269672	0.06293803	1.000000000	0.09120281	0.09125123	0.08651777	-0.09917578
nox	0.42097171	-0.51660371	0.76365145	0.091202807	1.00000000	-0.30218819	0.73147010	-0.76923011
rm	-0.21924670	0.31199059	-0.39167585	0.091251225	-0.30218819	1.00000000	-0.24026493	0.20524621
age	0.35273425	-0.56953734	0.64477851	0.086517774	0.73147010	-0.24026493	1.00000000	-0.74788054
dis	-0.37967009	0.66440822	-0.70802699	-0.099175780	-0.76923011	0.20524621	-0.74788054	1.00000000
rad	0.62550515	-0.31194783	0.59512927	-0.007368241	0.61144056	-0.20984667	0.45602245	-0.49458793
tax	0.58276431	-0.31456332	0.72076018	-0.035586518	0.66802320	-0.29204783	0.50645559	-0.53443158
ptratio	0.28994558	-0.39167855	0.38324756	-0.121515174	0.18893268	-0.35550149	0.26151501	-0.23247054
lstat	0.45562148	-0.41299457	0.60379972	-0.053929298	0.59087892	-0.61380827	0.60233853	-0.49699583
medv	-0.38830461	0.36044534	-0.48372516	0.175260177	-0.42732077	0.69535995	-0.37695457	0.24992873

	rad	tax	ptratio	lstat	medv
crim	0.625505145	0.58276431	0.2899456	0.4556215	-0.3883046
zn	-0.311947826	-0.31456332	-0.3916785	-0.4129946	0.3604453
indus	0.595129275	0.72076018	0.3832476	0.6037997	-0.4837252
chas	-0.007368241	-0.03558652	-0.1215152	-0.0539293	0.1752602
nox	0.611440563	0.66802320	0.1889327	0.5908789	-0.4273208
rm	-0.209846668	-0.29204783	-0.3555015	-0.6138083	0.6953599
age	0.456022452	0.50645559	0.2615150	0.6023385	-0.3769546
dis	-0.494587930	-0.53443158	-0.2324705	-0.4969958	0.2499287
rad	1.000000000	0.91022819	0.4647412	0.4886763	-0.3816262
tax	0.910228189	1.00000000	0.4608530	0.5439934	-0.4685359
ptratio	0.464741179	0.46085304	1.0000000	0.3740443	-0.5077867
lstat	0.488676335	0.54399341	0.3740443	1.0000000	-0.7376627
medv	-0.381626231	-0.46853593	-0.5077867	-0.7376627	1.0000000

The following summary is a result of all the predictors being used.

```
Call:
lm(formula = crim ~ ., data = Boston)

Residuals:
   Min     1Q Median     3Q    Max 
-8.534 -2.248 -0.348  1.087 73.923 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.7783938  7.0818258   1.946 0.052271 .  
zn           0.0457100  0.0187903   2.433 0.015344 *  
indus       -0.0583501  0.0836351  -0.698 0.485709    
chas        -0.8253776  1.1833963  -0.697 0.485841    
nox         -9.9575865  5.2898242  -1.882 0.060370 .  
rm           0.6289107  0.6070924   1.036 0.300738    
age         -0.0008483  0.0179482  -0.047 0.962323    
dis         -1.0122467  0.2824676  -3.584 0.000373 ***
rad          0.6124653  0.0875358   6.997 8.59e-12 ***
tax         -0.0037756  0.0051723  -0.730 0.465757    
ptratio     -0.3040728  0.1863598  -1.632 0.103393    
lstat        0.1388006  0.0757213   1.833 0.067398 .  
medv        -0.2200564  0.0598240  -3.678 0.000261 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.46 on 493 degrees of freedom
Multiple R-squared:  0.4493,	Adjusted R-squared:  0.4359 
F-statistic: 33.52 on 12 and 493 DF,  p-value: < 2.2e-16
```

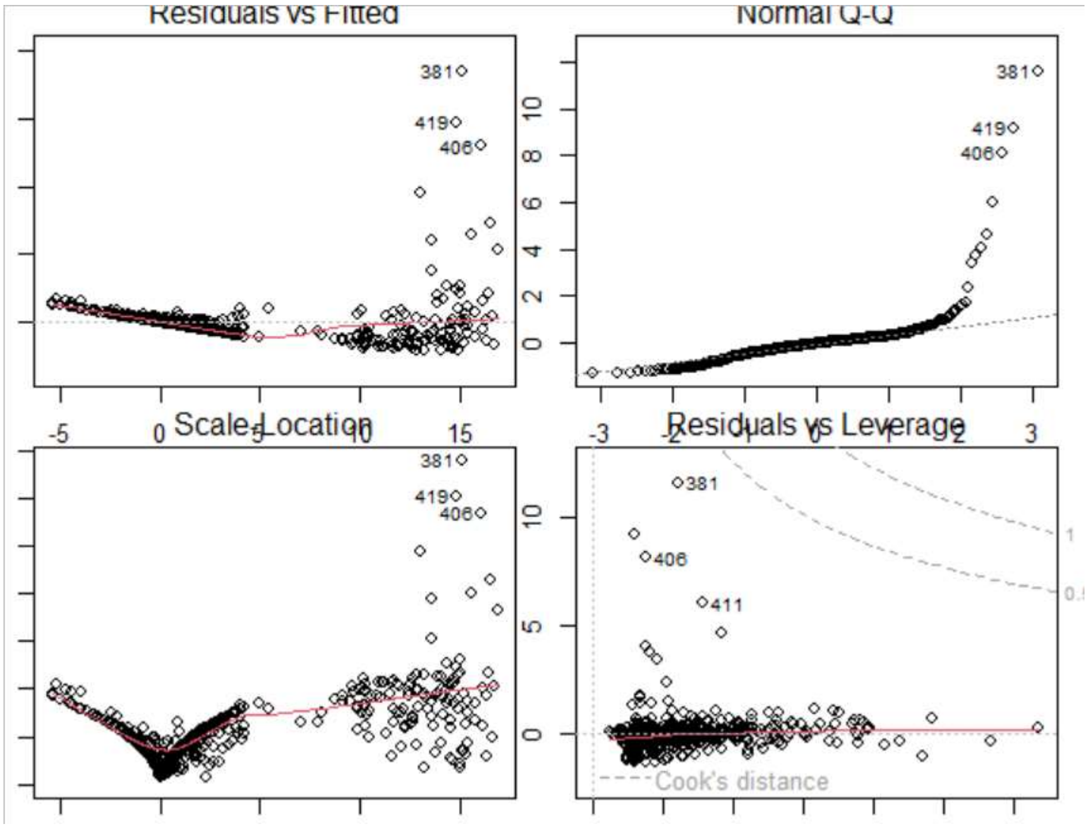
With just the 4 mentioned predictors, the F-statistic increases from 33.52 to 95.84.

```
Call:
lm(formula = crim ~ zn + dis + rad + medv, data = Boston)

Residuals:
   Min     1Q Median     3Q    Max 
-8.459 -1.960 -0.331  0.857 74.718 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.26548    1.34674   3.910 0.000105 ***
zn           0.05487    0.01735   3.163 0.001658 ** 
dis         -0.72291    0.20254  -3.569 0.000393 ***
rad          0.50021    0.04044  12.370  < 2e-16 ***
medv        -0.19122    0.03566  -5.362 1.26e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.5 on 501 degrees of freedom
Multiple R-squared:  0.4335,	Adjusted R-squared:  0.429 
F-statistic: 95.84 on 4 and 501 DF,  p-value: < 2.2e-16
```

We ran many different models on the data:

Name	RSE	Adjusted R^2	F-Statistic
All predictors	6.46	.4359	33.52
Medv,dis,rad,zn	6.5	.429	95.84
Medv,dis^0.5,rad,zn	6.465	.4351	98.24
Medv,dis,rad,zn^0.5	6.505	.4281	95.52
Crim^0.5~all predictors	.6936	.7716	143.2
Ln(Crim)~medv,dis,zn,rad	.878	.8349	639.4
Ln(crim)~all predictors	.781	.8694	281
Ln(crim)~all predictors on Boston1	.768	.8704	282

We analyzed the diagnostic plots for linear model with all predictors and found that the residual values increased significantly at higher fitted values. We tried to square the response but found that the natural log was better. The 4 predictors we picked out we a worse overall model than using all predictors. Finally, we removed 3 outliers we believed were skewing out data and renamed the dataset "Boston1" and see the model is only slightly better.



Residuals vs Fitted
Normal Q-Q
Scale-Location
Residuals vs Leverage
381
419
406
411
Cook's distance

The best model we found was the take the natural log of the response and to include all the predictors.

```
Call:
lm(formula = I(log(crim, base = 2.72)) ~ ., data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.58529 -0.56856 -0.04957  0.47295  2.66877 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.3784836  0.8561394  -5.114 4.52e-07 ***
zn          -0.0115074  0.0022716  -5.066 5.76e-07 ***
indus        0.0208393  0.0101109   2.061  0.03982 *  
chas        -0.0632406  0.1430637  -0.442  0.65865    
nox          3.9152451  0.6394999   6.122 1.88e-09 ***
rm          -0.0093813  0.0733929  -0.128  0.89834    
age          0.0055267  0.0021698   2.547  0.01117 *  
dis         -0.0104253  0.0341482  -0.305  0.76027    
rad          0.1475944  0.0105824  13.947  < 2e-16 ***
tax         -0.0001312  0.0006253  -0.210  0.83394    
ptratio     -0.0476792  0.0225295  -2.116  0.03482 *  
lstat        0.0341910  0.0091541   3.735  0.00021 ***
medv         0.0062483  0.0072323   0.864  0.38803    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.781 on 493 degrees of freedom
Multiple R-squared:  0.8725,	Adjusted R-squared:  0.8694 
F-statistic:   281 on 12 and 493 DF,  p-value: < 2.2e-16
```

In this model, we used all other 12 predictors to help predict the Boston suburbs' crime rate. In the model created, we transformed the response variable by taking the natural log of each which resulted in the adjusted R^2 being 86.94%. After removing 3 outliers and renaming the dataset “Boston1” we got our R^2 up to 87.04%. We chose to use this model because out of all the ones we tested included other transforming functions of the response variable, using only the predictor variable with the lowest p-scores, and transforming the predictor variables as well, the model with just taking the natural log of the response variable explained the most variability of the response and had a high F-Statistic.