Week 3 Assignment

David Russo

1/28/2017

* 3.7 #14
  1. Set up data

set.seed(1)  
x1 <- runif(100)  
x2 <- 0.5\*x1 + rnorm(100)/10  
y <- 2 + 2\*x1 + 0.3\*x2 + rnorm(100)

The form of the linear model is .

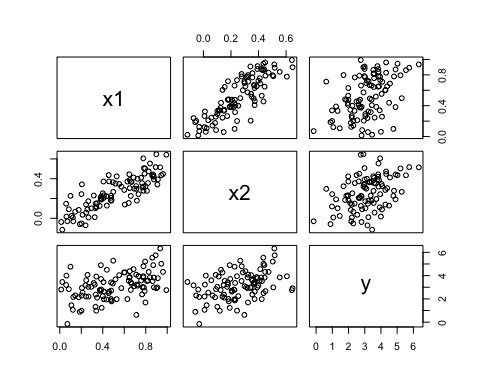
The coefficients are

* 1. cor() and pairs()

cor(x1, x2)

## [1] 0.8351212

pairs(data.frame(x1 = x1, x2 = x2, y = y))



The correlation between x1 and x2 is 0.8351.

* 1. lm() and summary()

lm\_all <- lm(y ~ x1 + x2)  
summary(lm\_all)

##   
## Call:  
## lm(formula = y ~ x1 + x2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8311 -0.7273 -0.0537 0.6338 2.3359   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.1305 0.2319 9.188 7.61e-15 \*\*\*  
## x1 1.4396 0.7212 1.996 0.0487 \*   
## x2 1.0097 1.1337 0.891 0.3754   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.056 on 97 degrees of freedom  
## Multiple R-squared: 0.2088, Adjusted R-squared: 0.1925   
## F-statistic: 12.8 on 2 and 97 DF, p-value: 1.164e-05

of 2.1305 is fairly close to the true of 2; of 1.4396 is about 70% the size of = 2; of 1.0097 is about 3 times the size of = 0.3

We can reject the null hypothesis because the p-value = 0.0487 < 0.05. We cannot reject the null hypothesis because the p-value = 0.3754 > 0.05.

* 1. lm() y onto x1

lm\_x1 <- lm(y ~ x1)  
summary(lm\_x1)

##   
## Call:  
## lm(formula = y ~ x1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.89495 -0.66874 -0.07785 0.59221 2.45560   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.1124 0.2307 9.155 8.27e-15 \*\*\*  
## x1 1.9759 0.3963 4.986 2.66e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.055 on 98 degrees of freedom  
## Multiple R-squared: 0.2024, Adjusted R-squared: 0.1942   
## F-statistic: 24.86 on 1 and 98 DF, p-value: 2.661e-06

We can reject the null hypothesis because the p-value = < 0.05. The for this fit is 0.2024, which is very close to of the regression of onto and , which is 0.2088. This would indicate that does not add much information to the model given the presence of .

* 1. lm() onto x2

lm\_x2 <- lm(y ~ x2)  
summary(lm\_x2)

##   
## Call:  
## lm(formula = y ~ x2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.62687 -0.75156 -0.03598 0.72383 2.44890   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.3899 0.1949 12.26 < 2e-16 \*\*\*  
## x2 2.8996 0.6330 4.58 1.37e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.072 on 98 degrees of freedom  
## Multiple R-squared: 0.1763, Adjusted R-squared: 0.1679   
## F-statistic: 20.98 on 1 and 98 DF, p-value: 1.366e-05

We can reject the null hypothesis because the p-value = 1.3664310^{-5} < 0.05. The for this fit is 0.1763, which is a few percentage points away from the of the regression of onto and , which is 0.2088. This would indicate that does not add as much information to the model as . The variable is significant in the absence of , but is not significant in the presence of .

* 1. Comparison of (c)-(e)

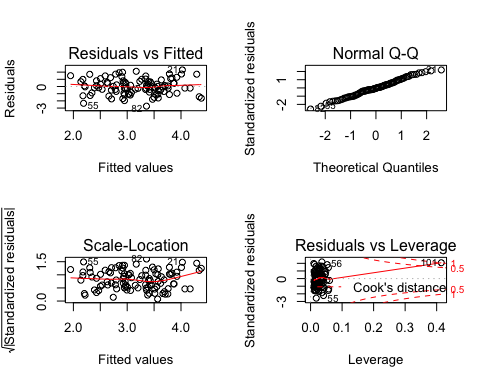
In part (c), we fit a regression model of onto and . In this model, was a significant predictor of , but was not. In part (d), we fit a model of onto only and was significant. In part (e), we fit a model of onto and was significant. These answers contradict eachother on the surface, but upon further investigation it means that is not significant in the presence of , but it is significant in the absence of . This would indicate that and are correlated (i.e., they suffer from collinearity).

+ g. Additional observation

x1 <- c(x1, 0.1)  
x2 <- c(x2, 0.8)  
y <- c(y, 6)  
  
lm\_all\_ao <- lm(y ~ x1 + x2)  
summary(lm\_all\_ao)

##   
## Call:  
## lm(formula = y ~ x1 + x2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.73348 -0.69318 -0.05263 0.66385 2.30619   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.2267 0.2314 9.624 7.91e-16 \*\*\*  
## x1 0.5394 0.5922 0.911 0.36458   
## x2 2.5146 0.8977 2.801 0.00614 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.075 on 98 degrees of freedom  
## Multiple R-squared: 0.2188, Adjusted R-squared: 0.2029   
## F-statistic: 13.72 on 2 and 98 DF, p-value: 5.564e-06

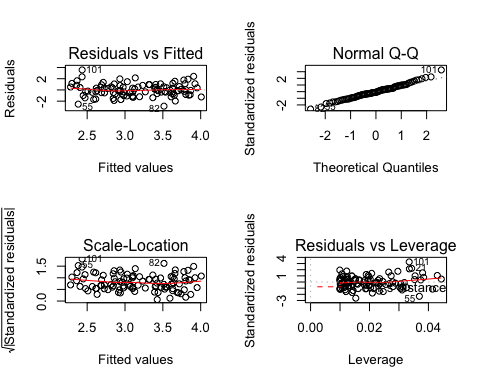
par(mfrow = c(2,2))  
plot(lm\_all\_ao)



lm\_x1\_ao <- lm(y ~ x1)  
summary(lm\_x1\_ao)

##   
## Call:  
## lm(formula = y ~ x1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8897 -0.6556 -0.0909 0.5682 3.5665   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.2569 0.2390 9.445 1.78e-15 \*\*\*  
## x1 1.7657 0.4124 4.282 4.29e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.111 on 99 degrees of freedom  
## Multiple R-squared: 0.1562, Adjusted R-squared: 0.1477   
## F-statistic: 18.33 on 1 and 99 DF, p-value: 4.295e-05

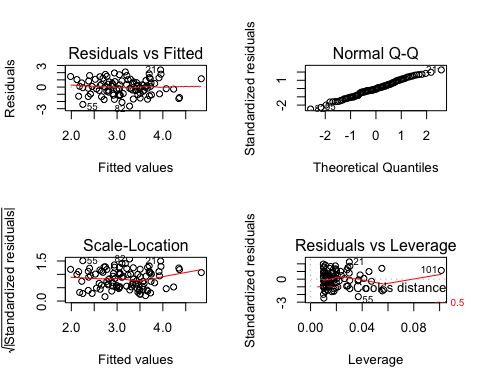
par(mfrow = c(2,2))  
plot(lm\_x1\_ao)



lm\_x2\_ao <- lm(y ~ x2)  
summary(lm\_x2\_ao)

##   
## Call:  
## lm(formula = y ~ x2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.64729 -0.71021 -0.06899 0.72699 2.38074   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.3451 0.1912 12.264 < 2e-16 \*\*\*  
## x2 3.1190 0.6040 5.164 1.25e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.074 on 99 degrees of freedom  
## Multiple R-squared: 0.2122, Adjusted R-squared: 0.2042   
## F-statistic: 26.66 on 1 and 99 DF, p-value: 1.253e-06

par(mfrow = c(2,2))  
plot(lm\_x2\_ao)



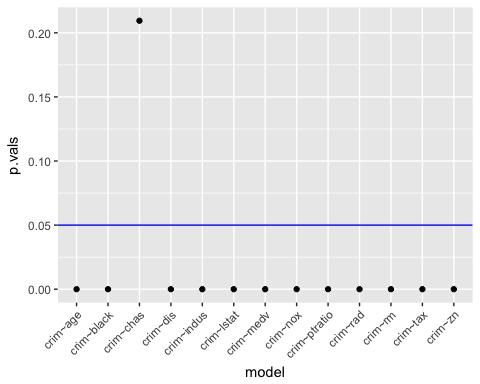
With the new observation, is significant and is not significant in the model featuring both predictors. In the individual models, both and are significant. The only model in which this new observation is an outlier is the model in which is regressed upon only. This can be seen by looking at the plot of the residuals vs. the fitted values. In the plot for the model featuring only, the new observation is about 3 units away from the mean of 0. In the other models, the new observation is not labeled as being far from the mean. In the model featuring both predictors, the new data point is a high leverage point. In the model using only the new point is not a high leverage point. In the model using only the new point is a high leverage point. The assessments of leverage can be made from the residuals vs. leverage plots.

* 3.7 #15
  1. lm() for each predictor

data(Boston)  
models <- lapply(paste("crim", names(Boston)[-1], sep = "~"), formula)  
res\_models <- lapply(models, function(x) {summary(lm(formula = x, data = Boston))})  
names(res\_models) <- paste("crim", names(Boston)[-1], sep = "~")  
res\_models

## $`crim~zn`  
##   
## Call:  
## lm(formula = x, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.429 -4.222 -2.620 1.250 84.523   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.45369 0.41722 10.675 < 2e-16 \*\*\*  
## zn -0.07393 0.01609 -4.594 5.51e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.435 on 504 degrees of freedom  
## Multiple R-squared: 0.04019, Adjusted R-squared: 0.03828   
## F-statistic: 21.1 on 1 and 504 DF, p-value: 5.506e-06  
##   
##   
## $`crim~indus`  
##   
## Call:  
## lm(formula = x, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.972 -2.698 -0.736 0.712 81.813   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.06374 0.66723 -3.093 0.00209 \*\*   
## indus 0.50978 0.05102 9.991 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.866 on 504 degrees of freedom  
## Multiple R-squared: 0.1653, Adjusted R-squared: 0.1637   
## F-statistic: 99.82 on 1 and 504 DF, p-value: < 2.2e-16  
##   
##   
## $`crim~chas`  
##   
## Call:  
## lm(formula = x, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.738 -3.661 -3.435 0.018 85.232   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.7444 0.3961 9.453 <2e-16 \*\*\*  
## chas -1.8928 1.5061 -1.257 0.209   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.597 on 504 degrees of freedom  
## Multiple R-squared: 0.003124, Adjusted R-squared: 0.001146   
## F-statistic: 1.579 on 1 and 504 DF, p-value: 0.2094  
##   
##   
## $`crim~nox`  
##   
## Call:  
## lm(formula = x, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.371 -2.738 -0.974 0.559 81.728   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -13.720 1.699 -8.073 5.08e-15 \*\*\*  
## nox 31.249 2.999 10.419 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.81 on 504 degrees of freedom  
## Multiple R-squared: 0.1772, Adjusted R-squared: 0.1756   
## F-statistic: 108.6 on 1 and 504 DF, p-value: < 2.2e-16  
##   
##   
## $`crim~rm`  
##   
## Call:  
## lm(formula = x, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.604 -3.952 -2.654 0.989 87.197   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 20.482 3.365 6.088 2.27e-09 \*\*\*  
## rm -2.684 0.532 -5.045 6.35e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.401 on 504 degrees of freedom  
## Multiple R-squared: 0.04807, Adjusted R-squared: 0.04618   
## F-statistic: 25.45 on 1 and 504 DF, p-value: 6.347e-07  
##   
##   
## $`crim~age`  
##   
## Call:  
## lm(formula = x, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.789 -4.257 -1.230 1.527 82.849   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.77791 0.94398 -4.002 7.22e-05 \*\*\*  
## age 0.10779 0.01274 8.463 2.85e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.057 on 504 degrees of freedom  
## Multiple R-squared: 0.1244, Adjusted R-squared: 0.1227   
## F-statistic: 71.62 on 1 and 504 DF, p-value: 2.855e-16  
##   
##   
## $`crim~dis`  
##   
## Call:  
## lm(formula = x, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.708 -4.134 -1.527 1.516 81.674   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.4993 0.7304 13.006 <2e-16 \*\*\*  
## dis -1.5509 0.1683 -9.213 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.965 on 504 degrees of freedom  
## Multiple R-squared: 0.1441, Adjusted R-squared: 0.1425   
## F-statistic: 84.89 on 1 and 504 DF, p-value: < 2.2e-16  
##   
##   
## $`crim~rad`  
##   
## Call:  
## lm(formula = x, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.164 -1.381 -0.141 0.660 76.433   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.28716 0.44348 -5.157 3.61e-07 \*\*\*  
## rad 0.61791 0.03433 17.998 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.718 on 504 degrees of freedom  
## Multiple R-squared: 0.3913, Adjusted R-squared: 0.39   
## F-statistic: 323.9 on 1 and 504 DF, p-value: < 2.2e-16  
##   
##   
## $`crim~tax`  
##   
## Call:  
## lm(formula = x, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.513 -2.738 -0.194 1.065 77.696   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8.528369 0.815809 -10.45 <2e-16 \*\*\*  
## tax 0.029742 0.001847 16.10 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.997 on 504 degrees of freedom  
## Multiple R-squared: 0.3396, Adjusted R-squared: 0.3383   
## F-statistic: 259.2 on 1 and 504 DF, p-value: < 2.2e-16  
##   
##   
## $`crim~ptratio`  
##   
## Call:  
## lm(formula = x, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.654 -3.985 -1.912 1.825 83.353   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -17.6469 3.1473 -5.607 3.40e-08 \*\*\*  
## ptratio 1.1520 0.1694 6.801 2.94e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.24 on 504 degrees of freedom  
## Multiple R-squared: 0.08407, Adjusted R-squared: 0.08225   
## F-statistic: 46.26 on 1 and 504 DF, p-value: 2.943e-11  
##   
##   
## $`crim~black`  
##   
## Call:  
## lm(formula = x, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.756 -2.299 -2.095 -1.296 86.822   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.553529 1.425903 11.609 <2e-16 \*\*\*  
## black -0.036280 0.003873 -9.367 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.946 on 504 degrees of freedom  
## Multiple R-squared: 0.1483, Adjusted R-squared: 0.1466   
## F-statistic: 87.74 on 1 and 504 DF, p-value: < 2.2e-16  
##   
##   
## $`crim~lstat`  
##   
## Call:  
## lm(formula = x, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.925 -2.822 -0.664 1.079 82.862   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.33054 0.69376 -4.801 2.09e-06 \*\*\*  
## lstat 0.54880 0.04776 11.491 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.664 on 504 degrees of freedom  
## Multiple R-squared: 0.2076, Adjusted R-squared: 0.206   
## F-statistic: 132 on 1 and 504 DF, p-value: < 2.2e-16  
##   
##   
## $`crim~medv`  
##   
## Call:  
## lm(formula = x, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.071 -4.022 -2.343 1.298 80.957   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.79654 0.93419 12.63 <2e-16 \*\*\*  
## medv -0.36316 0.03839 -9.46 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.934 on 504 degrees of freedom  
## Multiple R-squared: 0.1508, Adjusted R-squared: 0.1491   
## F-statistic: 89.49 on 1 and 504 DF, p-value: < 2.2e-16

With the exception of the variable , each predictor is a significant predictor of crime rate. This can be seen in the chart below which plots p-value for the predictor as a function of model.



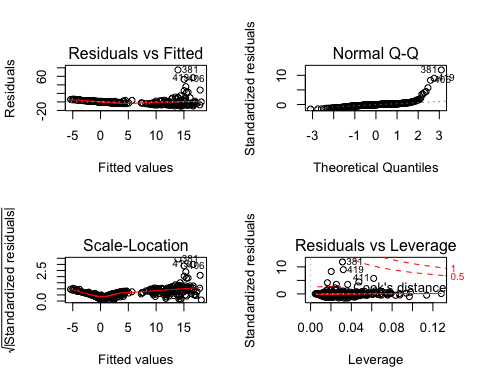
* 1. multiple regression

lm\_all <- lm(crim ~ ., data = Boston)  
summary(lm\_all)

##   
## Call:  
## lm(formula = crim ~ ., data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.924 -2.120 -0.353 1.019 75.051   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.033228 7.234903 2.354 0.018949 \*   
## zn 0.044855 0.018734 2.394 0.017025 \*   
## indus -0.063855 0.083407 -0.766 0.444294   
## chas -0.749134 1.180147 -0.635 0.525867   
## nox -10.313535 5.275536 -1.955 0.051152 .   
## rm 0.430131 0.612830 0.702 0.483089   
## age 0.001452 0.017925 0.081 0.935488   
## dis -0.987176 0.281817 -3.503 0.000502 \*\*\*  
## rad 0.588209 0.088049 6.680 6.46e-11 \*\*\*  
## tax -0.003780 0.005156 -0.733 0.463793   
## ptratio -0.271081 0.186450 -1.454 0.146611   
## black -0.007538 0.003673 -2.052 0.040702 \*   
## lstat 0.126211 0.075725 1.667 0.096208 .   
## medv -0.198887 0.060516 -3.287 0.001087 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.439 on 492 degrees of freedom  
## Multiple R-squared: 0.454, Adjusted R-squared: 0.4396   
## F-statistic: 31.47 on 13 and 492 DF, p-value: < 2.2e-16

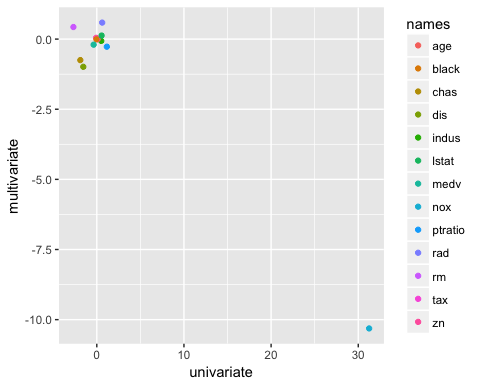
In the multiple regression model, we see that we can reject the null hypothesis that at least one of the predictors is significant as evidenced by the p-value of . We also see that , , , , and are significant predictors of crime at the level. In examining the diagnostic plots, we see that the assumption of constant variance is violated as the residuals form a funnel shape. The residuals get bigger as the fitted values get bigger. The q-q plot indicates a serious deviation from normality.

par(mfrow = c(2,2))  
plot(lm\_all)



* 1. comparisons

univariate <- as.numeric(sapply(res\_models, function(x) x$coefficients[2,1]))  
multivariate <- as.numeric(lm\_all$coefficients[-1])  
names <- names(lm\_all$coefficients)[-1]  
   
cos <- data.frame(  
 names = names,  
 univariate = univariate,  
 multivariate = multivariate  
)  
  
plot2 <- cos %>%  
 ggplot(aes(x = univariate, y = multivariate)) +   
 geom\_point(aes(col = names)) +   
 scale\_fill\_distiller()  
plot2



As can be seen from the plot, the univariate regression estimates are similar to the multivariate regression estimates with the exception of one point. The point with the univariate estimate of 30 and a multivariate estimate of -10 corresponds to the predictor of .