Homework 1

Jacob Thielemier

23 January 2024

Question 1

(a) Use the following formula:

$$Obs.X\sqrt{(0-x_1)^2+(0-x_2)^2+(0-x_3)^2}$$

```
Obs <- function(x1, x2, x3) {
  result <- sqrt((0-x1)^2 + (0-x2)^2 + (0-x3)^2)
  return(result)
}

Obs.1 <- Obs(0, 3, 0)
Obs.2 <- Obs(2, 0, 0)
Obs.3 <- Obs(0, 1, 3)
Obs.4 <- Obs(0, 1, 2)
Obs.5 <- Obs(-1, 0, 1)
Obs.6 <- Obs(1, 1, 1)
obs_values <- c(Obs.1, Obs.2, Obs.3, Obs.4, Obs.5, Obs.6)
print(obs_values)
```

```
## [1] 3.000000 2.000000 3.162278 2.236068 1.414214 1.732051
```

(b) **Green**; the nearest single observation is Obs.5.

(c) **Red**; the nearest three observations are green (Obs.5), red (Obs.6) and red (Obs.2). The probability of the test point belonging to red is 2/3 and green is 1/3. Therefore, the prediction is red.

(d) For highly non-linear boundaries, we would expect the best value of K to be small. Small K values yield a model with lots of detailed curves in the boundary, and likely the lowest irreducible error.

Question 2

(a) **ii.** because without considering the interaction term, the base model shows that college graduates earn more. This is indicated by the positive coefficient for Level.

(b) $137,100

Use the following formula:

$$Y = 50 + 20 * GPA + 0.07 * IQ + 35 * Level + 0.01 * GPA : IQ - 10 * GPA : Level$$

```
earn <- function(GPA, IQ, Level) {
  Y = 50 + 20*GPA + 0.07*IQ + 35*Level + 0.01*(GPA*IQ) - 10*(GPA*Level)
  return(Y)
}
questionB <- earn(4.0, 110, 1)
print(questionB)
```

```
## [1] 137.1
```

(c) False the magnitude of the coefficient for the GPA/IQ interaction term being very small does not imply that there is very little evidence of an interaction effect. If the coefficient is small but statistically significant (which can be determined by looking at the p-value), it means the interaction effect is present and meaningful.

Question 3

```
options(repos = c(CRAN = "https://cloud.r-project.org/"))
install.packages("ISLR")
```

(a)

```
## Installing package into 'C:/Users/JThie/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'ISLR' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\JThie\AppData\Local\Temp\RtmpMNQ0AS\downloaded_packages
```

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.3.2
```

```
data(Carseats)
```

```
carseats_lm = lm(Sales~Price+Urban+US,data=Carseats)
summary(carseats_lm)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

(b)

- The `Price` coefficient is negative and so sales will fall by roughly 54 seats (0.054x1000) for every unit ($1) increase in price.
- The `UrbanYes` coefficient is not statistically significant.
- The `USYes` coefficient is 1.2, and this means an average increase in car seat sales of 1200 units when `US=Yes`(this predictor refers to the shop being in the USA).

(c)

$$Sales = 13 + -0.054 \times Price + \begin{cases} -0.022, & \text{if } Urban \text{ is Yes, } US \text{ is No} \\ 1.20, & \text{if } Urban \text{ is No, } US \text{ is Yes} \\ 1.18, & \text{if } Urban \text{ and } US \text{ is Yes} \\ 0, & \text{Otherwise} \end{cases}$$

(d) If we use all variables, the null hypothesis can be rejected for `CompPrice`, `Income`, `Advertising`, `Price`, `ShelvelocGood`, `ShelvelocMedium` and `Age`.

```r
carseats_all_lm = lm(Sales~.,data=Carseats)
summary(carseats_all_lm)
```

```
##
## Call:
## lm(formula = Sales ~ ., data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8692 -0.6908  0.0211  0.6636  3.4115
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.6606231  0.6034487   9.380  < 2e-16 ***
## CompPrice        0.0928153  0.0041477  22.378  < 2e-16 ***
## Income           0.0158028  0.0018451   8.565 2.58e-16 ***
## Advertising      0.1230951  0.0111237  11.066  < 2e-16 ***
```

```
## Population       0.0002079  0.0003705   0.561    0.575    
## Price           -0.0953579  0.0026711 -35.700  < 2e-16 ***
## ShelveLocGood    4.8501827  0.1531100  31.678  < 2e-16 ***
## ShelveLocMedium  1.9567148  0.1261056  15.516  < 2e-16 ***
## Age             -0.0460452  0.0031817 -14.472  < 2e-16 ***
## Education       -0.0211018  0.0197205  -1.070    0.285    
## UrbanYes         0.1228864  0.1129761   1.088    0.277    
## USYes           -0.1840928  0.1498423  -1.229    0.220    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.019 on 388 degrees of freedom
## Multiple R-squared:  0.8734, Adjusted R-squared:  0.8698 
## F-statistic: 243.4 on 11 and 388 DF,  p-value: < 2.2e-16
```

```r
carseats_all_lm2 <- lm(Sales~.-Education-Urban-US-Population,data=Carseats)
summary(carseats_all_lm2)
```

(e)

```
##
## Call:
## lm(formula = Sales ~ . - Education - Urban - US - Population, 
##     data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.7728 -0.6954  0.0282  0.6732  3.3292 
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      5.475226   0.505005   10.84   <2e-16 ***
## CompPrice        0.092571   0.004123   22.45   <2e-16 ***
## Income           0.015785   0.001838    8.59   <2e-16 ***
## Advertising      0.115903   0.007724   15.01   <2e-16 ***
## Price           -0.095319   0.002670  -35.70   <2e-16 ***
## ShelveLocGood    4.835675   0.152499   31.71   <2e-16 ***
## ShelveLocMedium  1.951993   0.125375   15.57   <2e-16 ***
## Age             -0.046128   0.003177  -14.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.019 on 392 degrees of freedom
## Multiple R-squared:  0.872,  Adjusted R-squared:  0.8697 
## F-statistic: 381.4 on 7 and 392 DF,  p-value: < 2.2e-16
```

(f)

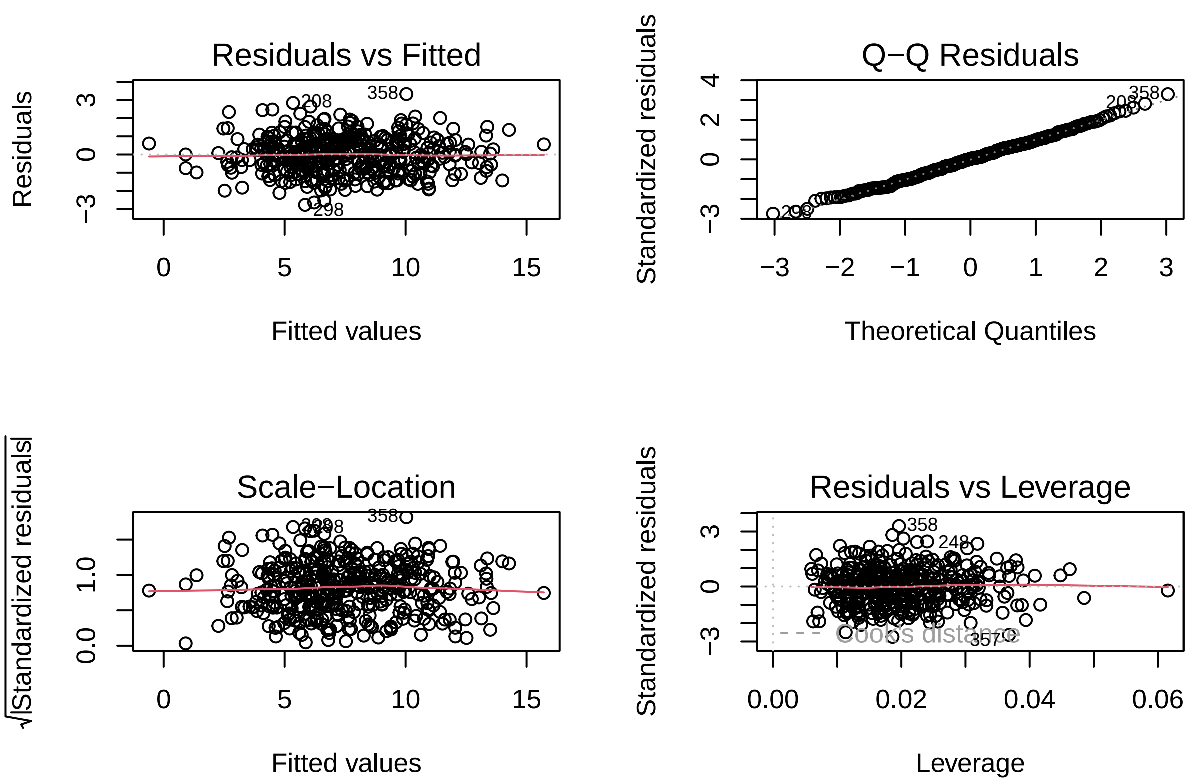
- The RSE goes down from 2.47 **model (a)** to 1.02 **model (e)**. The R2 statistic goes up from 0.24 **(a)** to 0.872 **(e)** and the F-statistic goes up from 41.52 to 381.4.
- The statistical evidence clearly shows that **(e)** is a much better fit.

```r
confint(carseats_all_lm2)
```

(g)

```
##                        2.5 %      97.5 %
## (Intercept)       4.48236820  6.46808427
## CompPrice         0.08446498  0.10067795
## Income            0.01217210  0.01939784
## Advertising       0.10071856  0.13108825
## Price            -0.10056844 -0.09006946
## ShelveLocGood     4.53585700  5.13549250
## ShelveLocMedium   1.70550103  2.19848429
## Age              -0.05237301 -0.03988204
```

```r
par(mfrow=c(2,2))
plot(carseats_all_lm2)
```



(h)

- The residuals vs. fitted values chart doesn't show any distinct shape, so the model appears to be a good fit to the data.

- There appears to be some outliers. We can check by using studentized residuals. Observation 358 appears to an outlier.

```
rstudent(carseats_all_lm2)[which(rstudent(carseats_all_lm2)>3)]
```

```
##     358
## 3.34075
```

- There appears to be one high leverage observation.

```
hatvalues(carseats_all_lm2)[order(hatvalues(carseats_all_lm2), decreasing = T)][1]
```

```
##        311
## 0.06154635
```

Question 4

```
set.seed(1)
x = rnorm(100, mean=0, sd=1)
```

(a)

- Length of y2=100, $\beta_0 = -1$, $\beta_1 = 0.5$

```
eps = rnorm(100, mean=0, sd=0.5)
```

(b)

```
y = -1 + (0.5*x) + eps
```

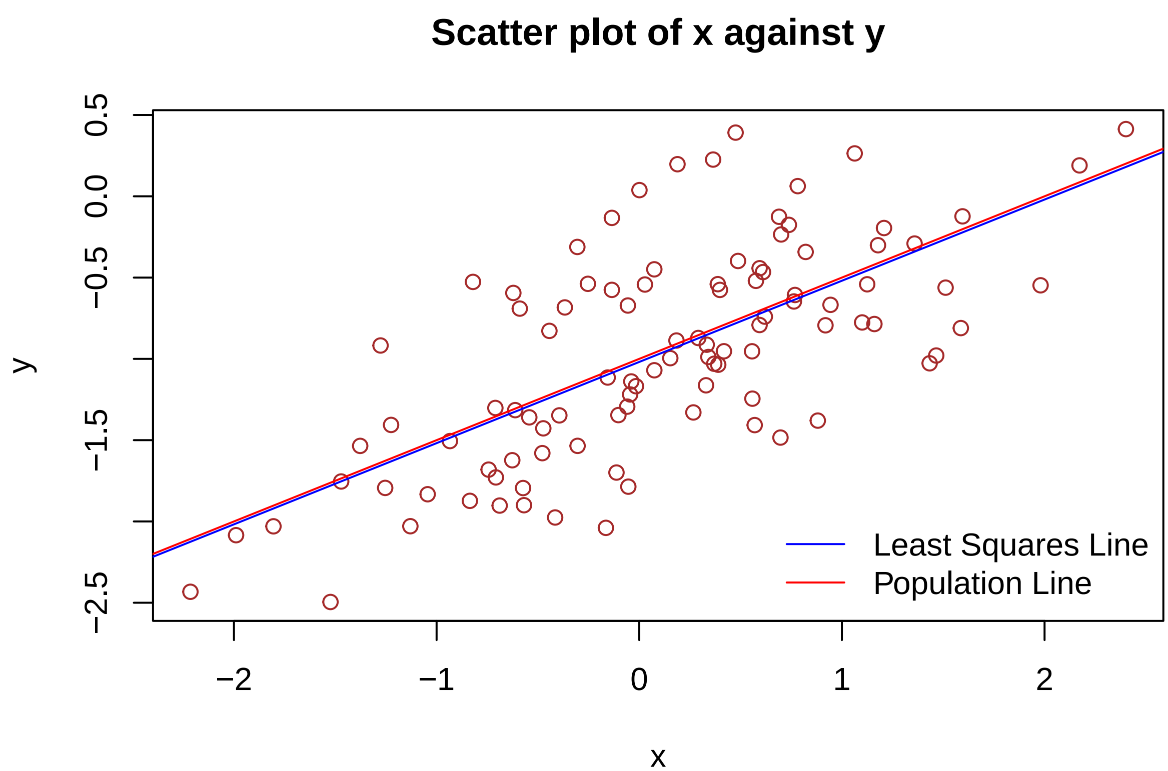
(c) Length of y2=100, $\beta_0 = -1$, $\beta_1 = 0.5$

```
plot(y~x, main= 'Scatter plot of x against y', col='brown')
#Linear regression line for (e)
lm.fit6 = lm(y~x)
summary(lm.fit6)
```

(d),(e),(f)

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93842 -0.30688 -0.06975  0.26970  1.17309
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.01885    0.04849 -21.010  < 2e-16 ***
## x            0.49947    0.05386   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
```

```r
abline(lm.fit6, lwd=1, col ="blue")
#Population regression line for (f)
abline(a=-1, b=0.5, lwd=1, col="red")
legend('bottomright', bty='n', legend=c('Least Squares Line', 'Population Line'), col=c('blue','red'), l
```



- A positive linear relationship exists between x2 and y2, with added variance introduced by the error terms.

- $\hat{\beta}_0 = -1.018$ and $\hat{\beta}_1 = 0.499$.The regression estimates are very close to the true values: $\beta_0 = -1$, $\beta_1 = 0.5$ This is further confirmed by the fact that the regression and population lines are very close to each other. P-values are near zero and F-statistic is large so null hypothesis can be rejected.

```
lm.fit7 = lm(y~x+I(x^2))
summary(lm.fit7)
```

(g)

```
##
## 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
```

The quadratic term does not improve the model fit. The F-statistic is reduced, and the p-value for the squared term is higher than 0.05 and shows that it isn't statistically significant.

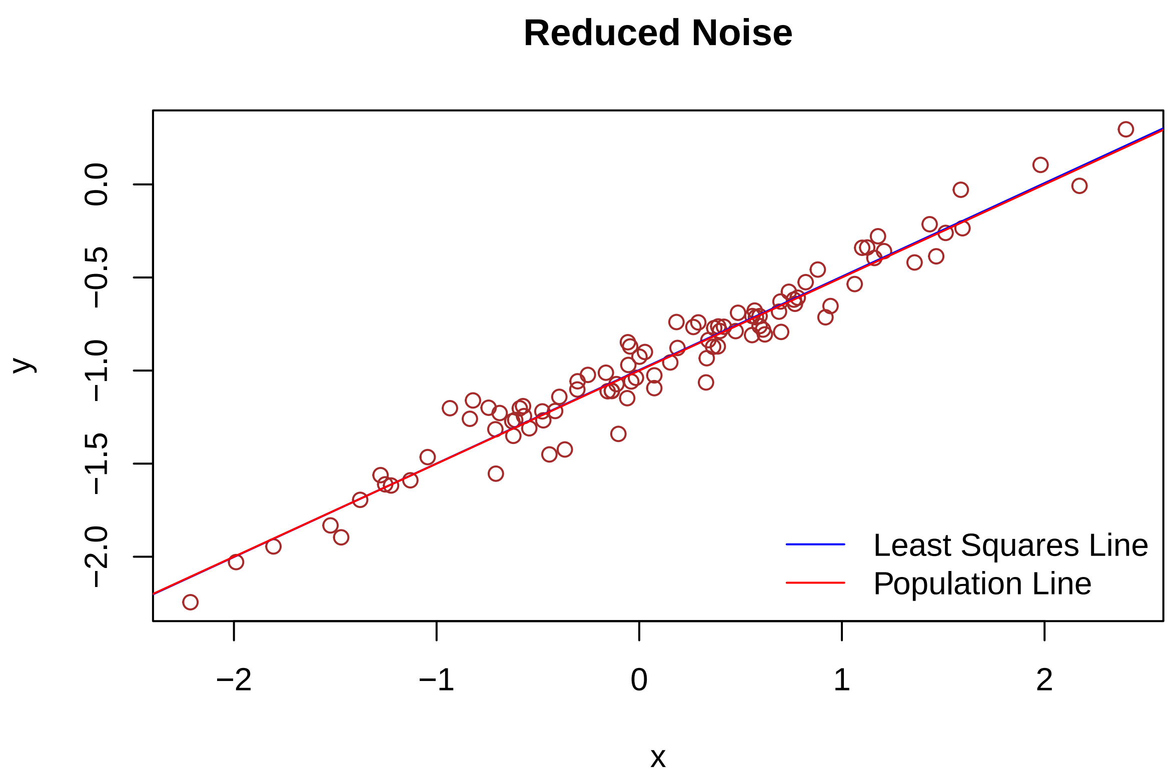
```
eps = rnorm(100, mean=0, sd=sqrt(0.01))
y = -1 +(0.5*x) + eps

plot(y~x, main='Reduced Noise', col='brown')
lm.fit7 = lm(y~x)
summary(lm.fit7)
```

(h)

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -0.291411 -0.048230 -0.004533  0.064924  0.264157
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.99726    0.01047  -95.25   <2e-16 ***
## x            0.50212    0.01163   43.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1039 on 98 degrees of freedom
## Multiple R-squared:  0.9501, Adjusted R-squared:  0.9495
## F-statistic:  1864 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
abline(lm.fit7, lwd=1, col ="blue")
abline(a=-1,b=0.5, lwd=1, col="red")
legend('bottomright', bty='n', legend=c('Least Squares Line', 'Population Line'), col=c('blue','red'), 
```



The points are closer to each other, the RSE is lower, R2 and F-statistic are much higher than with variance of 0.25. The linear regression and population lines are very close to each other as noise is reduced, and the relationship is much more linear.

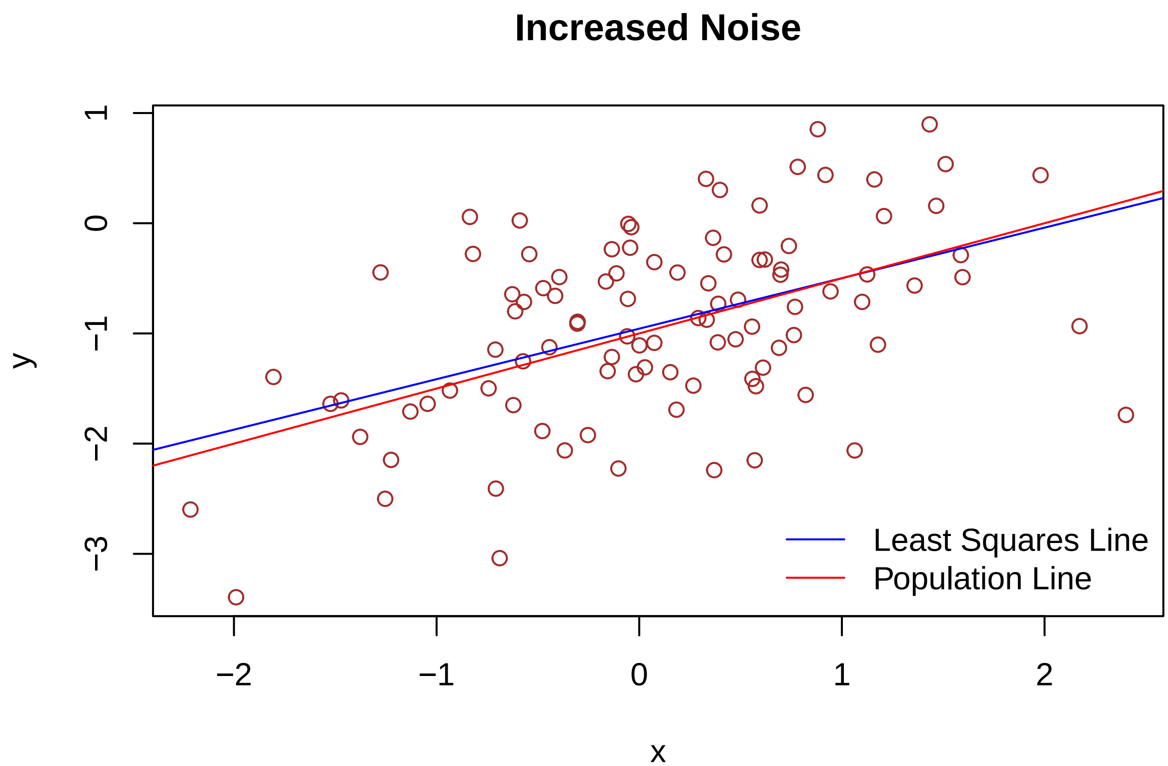
```
eps = rnorm(100, mean=0, sd=sqrt(0.56))
y = -1 +(0.5*x) + eps
```

```r
plot(y~x, main='Increased Noise', col='brown')
lm.fit8 = lm(y~x)
summary(lm.fit8)
```

(i)

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88299 -0.40802 -0.02826  0.50354  1.40602
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.95685    0.07504 -12.751  < 2e-16 ***
## x            0.45833    0.08335   5.499 3.05e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7449 on 98 degrees of freedom
## Multiple R-squared:  0.2358, Adjusted R-squared:  0.228
## F-statistic: 30.23 on 1 and 98 DF,  p-value: 3.046e-07
```

```r
abline(lm.fit8, lwd=1, col ="blue")
abline(a=-1,b=0.5, lwd=1, col="red")
legend('bottomright', bty='n', legend=c('Least Squares Line', 'Population Line'), col=c('blue','red'), l
```



The points are more spread out and so the relationship is less linear. The RSE is higher, the R2 and F-statistic are lower than with variance of 0.25.

```
confint(lm.fit6)
```

(j)

```
##                  2.5 %     97.5 %
## (Intercept) -1.1150804 -0.9226122
## x            0.3925794  0.6063602
```

```
confint(lm.fit7)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0180413 -0.9764850
## x            0.4790377  0.5251957
```

```
confint(lm.fit8)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.1057691 -0.8079252
## x            0.2929158  0.6237410
```

Confidence interval values are narrowest for the lowest variance model, widest for the highest variance model and in-between these two for the original model.