Exercise 4

Tobias Famos

Task 1

Load the data

```
education <- read.table("EducationBis.txt", header=T)
education_men <- education[which(education$Gender== "male"), ]
edcuation_women <- education[which(education$Gender=="female"),]
```

Now build linear model to explain `Wage` by `Edcuation`

```
linear_model_men <- lm(education_men$Wage~education_men$Education)
linear_model_women <- lm(edcuation_women$Wage~edcuation_women$Education)
summary(linear_model_men)
```

```
##
## Call:
## lm(formula = education_men$Wage ~ education_men$Education)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -243.654  -74.152    6.096   70.923  279.797
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               24.199     27.937   0.866    0.387
## education_men$Education  398.250      1.919 207.559   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97.41 on 297 degrees of freedom
## Multiple R-squared:  0.9932, Adjusted R-squared:  0.9931
## F-statistic: 4.308e+04 on 1 and 297 DF,  p-value: < 2.2e-16
```

```
summary(linear_model_women)
```

```
##
## Call:
## lm(formula = edcuation_women$Wage ~ edcuation_women$Education)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -220.461  -78.869   -4.028   73.437  271.939
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -563.61      37.50  -15.03   <2e-16 ***
```

```
## edcuation_women$Education   397.54       2.58  154.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 104.3 on 196 degrees of freedom
## Multiple R-squared:  0.9918, Adjusted R-squared:  0.9918
## F-statistic: 2.375e+04 on 1 and 196 DF,  p-value: < 2.2e-16
```

For both models, the multiple R^2 is quite high, thus we can conclude that the linear model explains much of the variance. The model for the women shows a significant p value in both the intercept and the slope, thus we can accpet it as true. The model for the men has only a significant p value in the slope but not in the intercept.

Task 2

The slopes are significantly different from 0 as they both have a p value $< 0.5\%$.

Task 3

First remove the ID column as we don't need it

```r
education_no_index <- subset(education, select=-c(ID))
```

```r
unifiedModel <- lm(education_no_index$Wage~., data=education_no_index)
summary(unifiedModel)
```

```
##
## Call:
## lm(formula = education_no_index$Wage ~ ., data = education_no_index)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -243.473  -76.073    0.354   73.126  280.275
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -569.784     23.100  -24.67   <2e-16 ***
## Education    397.975      1.543  258.00   <2e-16 ***
## Gendermale   597.904      9.173   65.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 100.1 on 494 degrees of freedom
## Multiple R-squared:  0.9931, Adjusted R-squared:  0.9931
## F-statistic: 3.544e+04 on 2 and 494 DF,  p-value: < 2.2e-16
```

The unified model explains the Wage quite good. All variables are significant, and the R^2 is almost 1.

Task 4

We can't use vendor name and model name to explain the performance.

Task 5

```r
computers <- read.table("Computers.txt", header=T)
computers_no_vendor_model <- subset(computers, select=-c(vendor, model))
```

```r
computeds_model <- lm(computers_no_vendor_model$PRP~. , data=computers_no_vendor_model)
summary(computeds_model)
```

```
##
## Call:
## lm(formula = computers_no_vendor_model$PRP ~ ., data = computers_no_vendor_model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -160.607  -15.233   -2.245    7.561  234.568
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.9096358  6.7797646   1.019   0.3094
## MYCT        -0.0134571  0.0125074  -1.076   0.2832
## MMIN         0.0017770  0.0015113   1.176   0.2411
## MMAX        -0.0006542  0.0005910  -1.107   0.2696
## CACH         0.1742877  0.0990474   1.760   0.0800 .
## CGMIN       -0.1075541  0.5786491  -0.186   0.8527
## CHMAX        0.3477529  0.1657726   2.098   0.0372 *
## ERP          0.9446790  0.0608708  15.519   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.56 on 201 degrees of freedom
## Multiple R-squared:  0.9385, Adjusted R-squared:  0.9364
## F-statistic: 438.5 on 7 and 201 DF,  p-value: < 2.2e-16
```

From the preliminary linear regression, the single variable I would use for a linear regression is the `ERP`.

```r
singleModel <- lm(computers$PRP ~ computers$ERP)
summary(singleModel)
```

```
##
## Call:
## lm(formula = computers$PRP ~ computers$ERP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -148.092  -16.189   -5.973   10.040  259.696
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.85461    3.40493   1.719    0.087 .
## computers$ERP  1.00440    0.01855  54.153   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.4 on 207 degrees of freedom
## Multiple R-squared:  0.9341, Adjusted R-squared:  0.9337
## F-statistic:  2933 on 1 and 207 DF,  p-value: < 2.2e-16
```

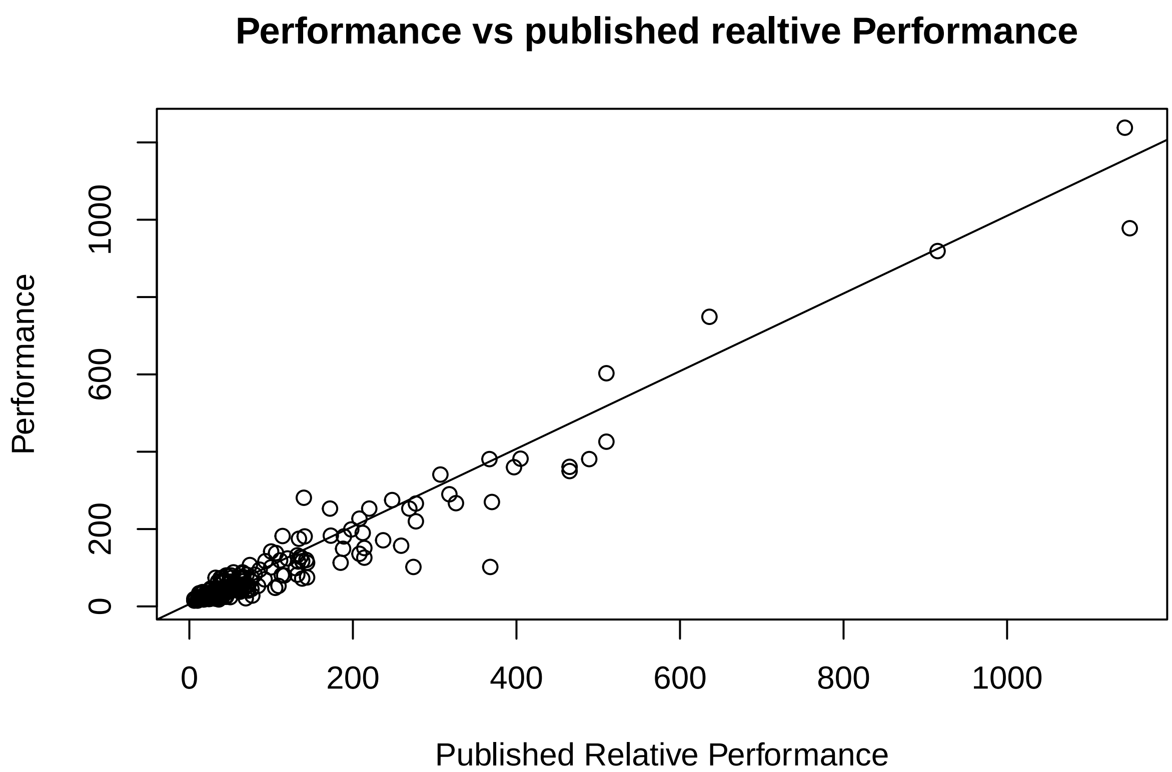
```r
confint(singleModel)
```

```
##                    2.5 %    97.5 %
```

```
## (Intercept)   -0.8581818 12.567408
## computers$ERP  0.9678360  1.040968
```

The model explains the 93% of the variance, and it is quite significant. The confidence interval is 2 times the standard error, thus 2*0.018 = 0.036 ## Task 6

```
plot(computers$PRP, computers$ERP, xlab="Published Relative Performance", ylab="Performance", main="Per
abline(singleModel$coefficients)
```



Task 7

```
cars = read.table("Cars.txt", header=T)
cars = subset(cars, select = -c(name))
all_model <- lm(cars$mpg~.,data=cars)
summary(all_model)
```

```
##
## Call:
## lm(formula = cars$mpg ~ ., data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -17.218435   4.644294  -3.707  0.00024 ***
## cylinders      -0.493376   0.323282  -1.526  0.12780
## displacement    0.019896   0.007515   2.647  0.00844 **
```

```
## horsepower    -0.016951   0.013787  -1.230  0.21963    
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548    
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.328 on 384 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182 
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

The not significant variables are : `cylinders`, `horsepower` and `acceleration`