STAT 504 - Homework 1

Due date: Thursday, January 24. Submit your homework solutions to the course Canvas page. Please submit the output and plots, but not your R code unless the quesiton specifically asks for it. Total possible points: 24.

1. (5 points) In this question we want to examine distributions with the help of the software R. In particular, we want to examine the (i) skewness, (ii) symmetry of a distribution, (iii) outliers and (iv) non-normality of the data sets `Anscombe` und `Leinhardt`.

   Both data sets can be found in the package `car`, so first install and load the package `car`. To load the data sets `Anscombe` and `Leinhardt` use the commands `data(Anscombe)` and `data(Leinhardt)`. They are then available as variables `Anscombe` and `Leinhardt`.

   (a) (2 points) You should first get an overview of the data sets. How many variables are there? How many missing values?

   ```
   ## R Hints
   summary(Anscombe)                  # To get an overview

   Anscombe$income                    # To access the variable
                                      # income in the data set Anscombe
   ```

   Solution: (1 Point for each data set.)

   ```
   > library(car)
   > data(Anscombe)
   > data(Leinhardt)
   > summary(Anscombe)

     education         income         young
   Min.   :112.0   Min.   :2081   Min.   :326.2
   1st Qu.:165.0   1st Qu.:2786   1st Qu.:342.1
   Median :192.0   Median :3257   Median :354.1
   Mean   :196.3   Mean   :3225   Mean   :358.9
   3rd Qu.:228.5   3rd Qu.:3612   3rd Qu.:369.1
   Max.   :372.0   Max.   :4425   Max.   :439.7
       urban
   Min.   : 322.0
   1st Qu.: 552.5
   Median : 664.0
   Mean   : 664.5
   3rd Qu.: 790.5
   Max.   :1000.0

   > summary(Leinhardt)

       income           infant            region    oil
   Min.   :  50.0   Min.   :  9.60   Africa  :34   no :96
   1st Qu.: 123.0   1st Qu.: 26.20   Americas:23   yes: 9
   ```

```
Median : 334.0   Median : 60.60   Asia    :30
Mean   : 998.1   Mean   : 89.05   Europe  :18
3rd Qu.:1191.0   3rd Qu.:129.40
Max.   :5596.0   Max.   :650.00
                 NA's   :4
```

The commands `summary(Anscombe)` and `Anscombe` provide a good first overview. We see that there are 4 variables and no missing values. The data set Leinhard consists of 4 variables as well, but the variable `infant` has 4 missing values.

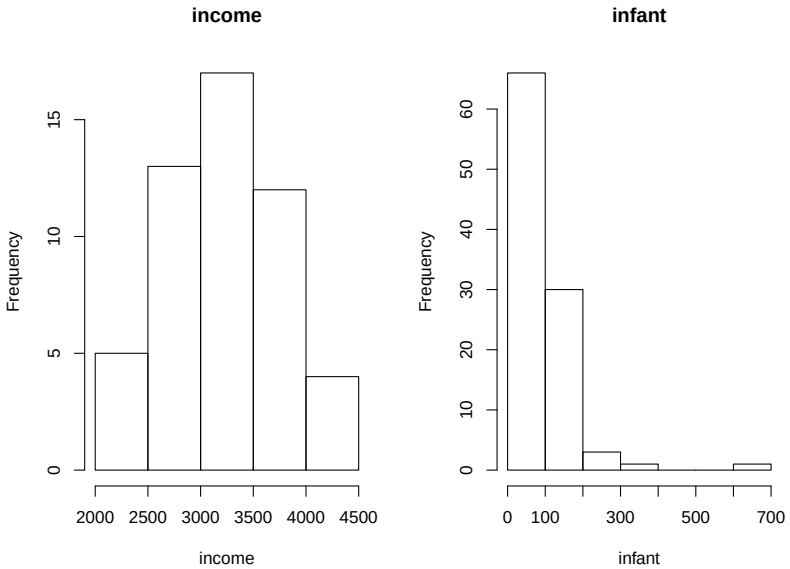
(b) (2 points) Examine the variable `income` in the data set `Anscombe` and the variable `infant` in the data set `Leinhardt` with respect to (i)-(iv) above (in the description).

```
## R Hints
AnsInc <- Anscombe$income          # Saves the variable income
                                   # with another name
summary(AnsInc)                    # For an overview
hist(AnsInc)                       # Draws a histogram
boxplot(AnsInc)                    # Boxplot of the data
```

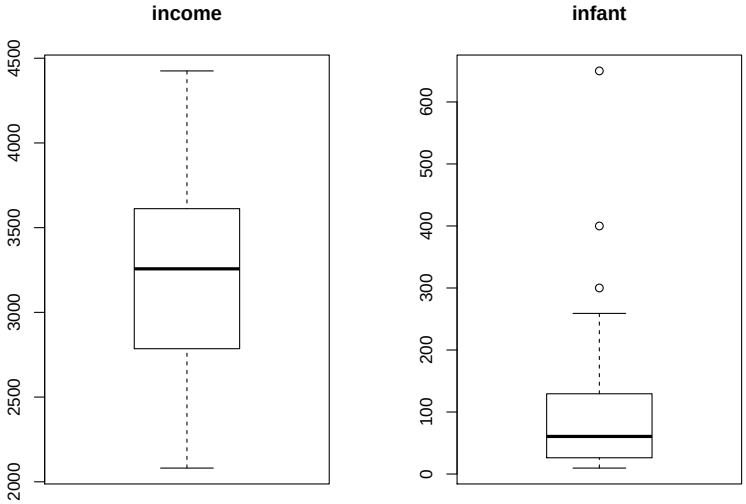
As always, you can access the help for any command with `?command`.

Solution: (0.5 points for each of the correctly analyzed parts (i)-(iv).)

```
> income <- Anscombe$income
> infant <- Leinhardt$infant

> par(mfrow=c(1,2))
> hist(income, main="income")
> hist(infant, main="infant")
```




```
> par(mfrow=c(1,2))
> boxplot(income, main="income")
> boxplot(infant, main="infant")
```



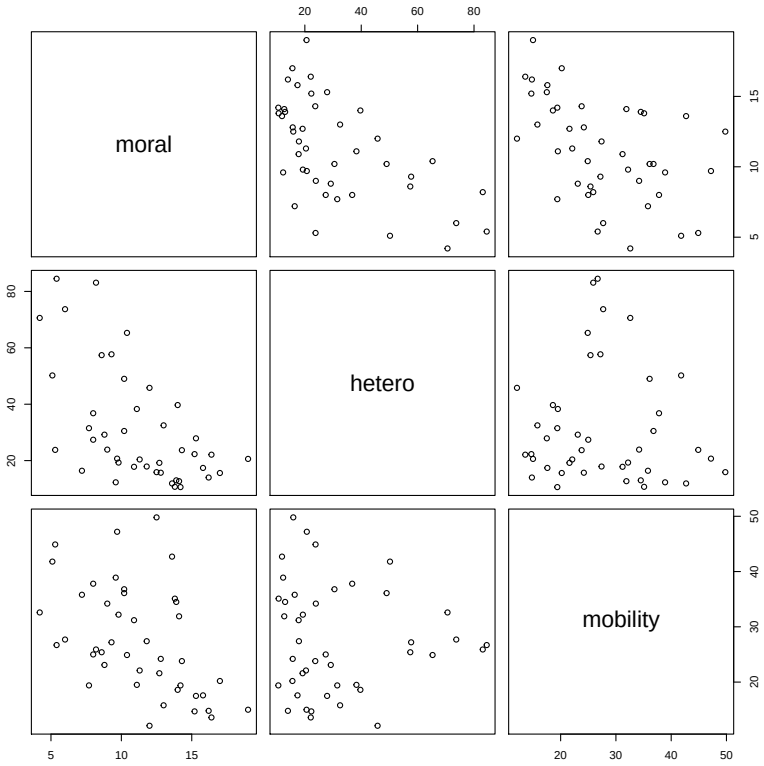
The variable `infant` is nonsymmetric and seems to have outliers. The variable `income` seems to be more or less normally distributed. The nonsymmetry and the outliers can be determined more easily with a boxplot (see the figure). We can see there that the variable `income` is symmetric and has no outliers while these observations are not true for `infant`.

(c) (1 point) For this task, we need to load the data set `Angell`. You may obtain a description of the data set with the command `?Angell`.

Generate a matrix of scatterplots for the variables *Moral integration*, *Ethnic heterogeneity* and *Geographic mobility*. Describe the dependences among the variables. Are there outliers?

Solution:

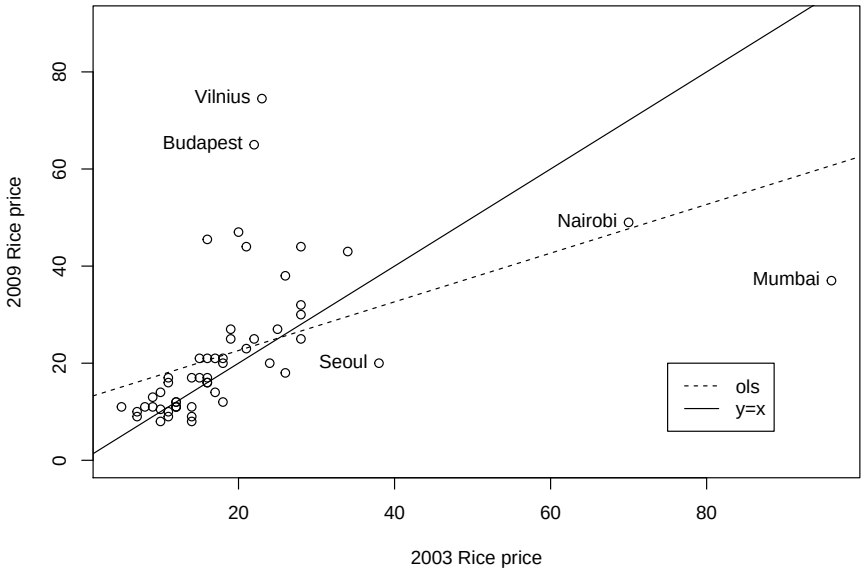
```
> # Generate scatterplot without the fourth column
> plot(Angell[,-4])
```



There seems to be a slight linear trend in all the scatterplots. There seem to be no clear outliers.

2. (12 points) (Problem 2.2 in Sanford Weisberg's "Applied linear regression", 4th edition.) Data: `UBSprices` in `alr4`. The international bank UBS regularly produces a report on prices and earnings in major cities throughout the world. Three of the measures they include are prices of basic commodities, namely 1kg of rice, a 1kg loaf of bread, and the price of a Big Mac hamburger at McDonalds. An interesting feature of the prices they report is that prices are measured in the minutes of labor required for a "typical" worker in that location to earn enough money to purchase the commodity. Using minutes of labor corrects at least in part for currency fluctuations, prevailing wage rates, and local prices. The data file includes measurements for rice, bread, and Big Mac prices from the 2003 and the 2009 reports. The year 2003 was before the major recession hit much of the world around 2006, and the year 2009 may reflect changes in prices due to the recession.

   The figure below is the plot of `y = rice2009` versus `x = rice2003`, the price of rice in 2009 and 2003, respectively, with the cities corresponding to a few of the points marked.



(a) (1 point) The line with equation $y = x$ is shown on this plot as the solid line. What is the key difference between points above this line and points below the line?

Solution: The points above the line $y = x$ had a rice price increase in 2009 compared to 2003.

(b) (1 point) Which city had the largest increase in rice price? Which had the largest decrease in rice price?

Solution:

```
> biggest.increase <- which.max(UBSprices$rice2009-UBSprices$rice2003)
> row.names(UBSprices[biggest.increase,])

[1] "Vilnius"
```

(c) (1 point) The OLS line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is shown on the figure as a dashed line, and evidently $\hat{\beta}_1 < 1$. Does this suggest that prices are lower in 2009 than in 2003? Explain your answer.

Solution: No, in fact we can see that most of the data points are above the $y = x$ line, so the price of rice seems to have increased in general.

The estimate of the slope $\hat{\beta}_1$ only tells us how much we expect the average price of rice in 2009 to increase when comparing $rice2003 = x$ with $rice2003 = x + 1$ for some x.

(d) (3 points) Calculate $\overline{x}, \overline{y}, SXX, SXY, \hat{\beta}_0$, and $\hat{\beta}_1$ from data.

Solution: (0.5 points for each correctly calculated value.)

```
> xbar <- mean(UBSprices$rice2003)
> xbar

[1] 19.46296

> ybar <- mean(UBSprices$rice2009)
> ybar

[1] 22.34259

> sxx <- sum((UBSprices$rice2003 - xbar)^2)
> sxx
```

```
[1] 11367.43

> sxy <- sum((UBSprices$rice2003 - xbar)*(UBSprices$rice2009 - ybar))
> sxy

[1] 5699.435

> beta1.hat <- sxy/sxx
> beta0.hat <- ybar-beta1.hat*xbar
> beta0.hat

[1] 12.58419

> beta1.hat

[1] 0.5013831

> ## compare with
> summary(lm(rice2009~rice2003, data=UBSprices))

Call:
lm(formula = rice2009 ~ rice2003, data = UBSprices)

Residuals:
    Min      1Q  Median      3Q     Max
-23.717  -7.105  -3.602   1.369  50.384

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.5842     2.9442   4.274 8.21e-05 ***
rice2003      0.5014     0.1213   4.134  0.00013 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.93 on 52 degrees of freedom
Multiple R-squared:  0.2474,        Adjusted R-squared:  0.2329
F-statistic: 17.09 on 1 and 52 DF,  p-value: 0.0001302
```

(e) (3 points) Calculate the fitted value $\hat{y}_*$ for $x_* = 50$. Give a 95% confidence interval for $\hat{y}_*$ and a 95% prediction interval for y_*?

Solution: (1 Point for the fitted value and 1 point for each of the intervals.)

You can calculate these values using the formulas from the lectures or using the following R functions:

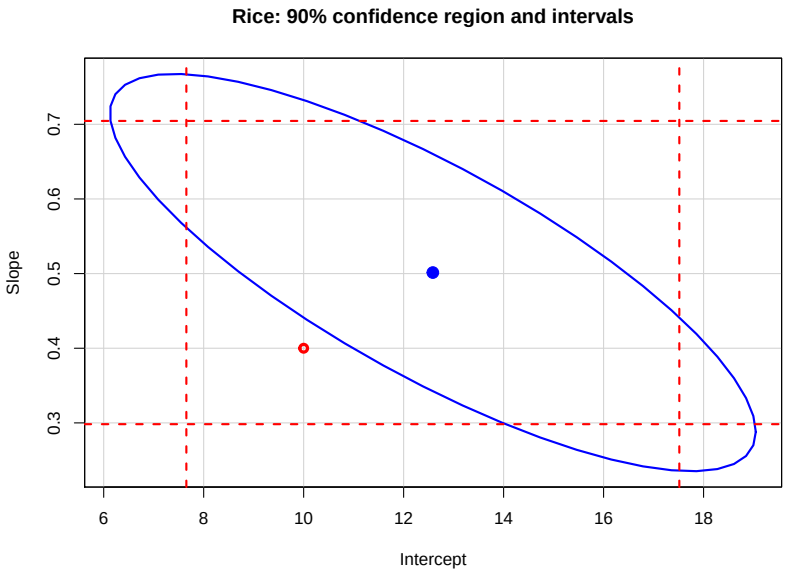
```
> new <- data.frame(rice2003 = 50)
> fit <- lm(rice2009 ~ rice2003,data=UBSprices)
> ## fitted value
> predict(fit, new, se.fit = TRUE)$fit

       1
37.65335

> ## confidence interval
> pred.w.clim <- predict(fit, new, interval = "confidence")
> pred.w.clim
```

```
       fit      lwr      upr
1 37.65335 29.42533 45.88136

> ## prediction interval
> pred.w.plim <- predict(fit, new, interval = "prediction")
> pred.w.plim

       fit      lwr      upr
1 37.65335 10.43231 64.87438
```

(f) (2 points) Would you reject the null hypothesis $H_0 : (\beta_0, \beta_1) = (10, 0.4)$ at a 90% level?

Solution: (1 Point for the plot of the ellipse and 1 point for a correct test decision.)

```
> confidenceEllipse(fit,grid=TRUE,xlab="Intercept",ylab="Slope",
+                   main="Rice: 90% confidence region and intervals",levels=.9)
> abline(h=confint(fit,level=.9)[2,1],lty=2,lwd=2,col="red")
> abline(h=confint(fit,level=.9)[2,2],lty=2,lwd=2,col="red")
> abline(v=confint(fit,level=.9)[1,1],lty=2,lwd=2,col="red")
> abline(v=confint(fit,level=.9)[1,2],lty=2,lwd=2,col="red")
> points(10,.4,col="red",lwd=3)
```

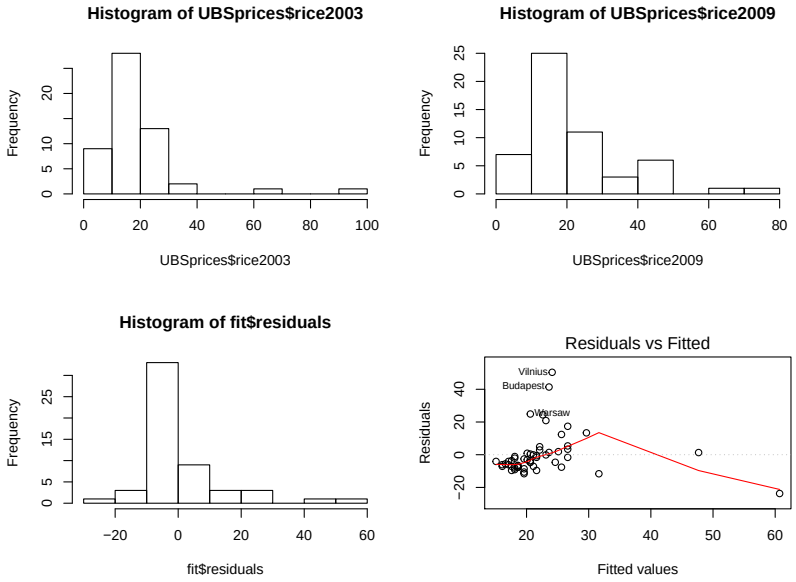


The point $(\beta_0, \beta_1) = (10, 0.4)$ is outside the confidence ellipse and hence, we reject the null hypothesis.

(g) (1 point) Give one reason why fitting simple linear regression to the figure in this problem is not likely to be appropriate.

Solution: Let's look at some diagnostic plots:

```
> par(mfrow=c(2,2))
> hist(UBSprices$rice2003)
> hist(UBSprices$rice2009)
> hist(fit$residuals)
> plot(fit, which =c(1))
```



The distribution of the predictor, response and the residuals appear to be right skewed. So the assumption of normality of the errors does not seem to be satisfied. Furthermore by looking at the residual versus fitted values plot, a curvature of the mean is present, which indicates that $\mathrm{E}[\epsilon|X = x] = 0$ is not satisfied. Similarly, the assumption of constant variance of the errors appears to be violated. Because of the skeweness of the response and the predictor there are few points with extreme values that appear as outliers on the plot.

3. (7 points) We consider a simple linear regression with dependent variable Y and independent variable X. The estimated coefficients using the least squares method are denoted by $\hat{\alpha}$ (intercept) and $\hat{\beta}$ (slope). Moreover, $\hat{\sigma}$ and the correlation r have been calculated already.

Hints: The formula for $\hat{\beta}$ as well as for $\hat{\alpha}$ can be found in the lecture notes. Moreover, $R^2 = \frac{\|\hat{Y}-\bar{Y}\|_2^2}{\|Y-\bar{Y}\|_2^2}$.

(a) (3 points) We now define $X' = X - 10$ and compute the regression of Y on X'. Give estimates of $\hat{\alpha}', \hat{\sigma}'$ and R'^2 (in terms of α, β, $\hat{\sigma}$, or R^2).
`Hint: For the transformation` $X' = X - 10$, $\hat{\beta}' = \hat{\beta}$.
Solution: (1 point for each correct estimate.)

With the formula for the coefficients, we obtain

$$\begin{aligned}
\hat{\beta}' &= \frac{\sum(X_i' - \bar{X}')(Y_i - \bar{Y})}{\sum(X_i' - \bar{X}')^2} = \frac{\sum((X_i - 10) - (\bar{X} - 10))(Y_i - \bar{Y})}{\sum((X_i - 10) - (\bar{X} - 10))^2} \\
&= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \hat{\beta} \\
\hat{\alpha}' &= \bar{Y} - \hat{\beta}'\bar{X}' = \bar{Y} - \hat{\beta}\bar{X}' = \bar{Y} - \hat{\beta}(\bar{X} - 10) = (\bar{Y} - \hat{\beta}\bar{X}) + 10\hat{\beta} = \hat{\alpha} + 10\hat{\beta}
\end{aligned}$$

and

$$\hat{Y}_i' = \hat{\alpha}' + \hat{\beta}'X_i' = (\hat{\alpha} + 10\hat{\beta}) + \hat{\beta}X_i' = (\hat{\alpha} + 10\hat{\beta}) + \hat{\beta}(X_i - 10) = \hat{\alpha} + \hat{\beta}X_i = \hat{Y}_i.$$

This implies that the residuals stay unchanged:

$$\varepsilon_i' = Y_i - \hat{Y}_i' = Y_i - \hat{Y}_i = \varepsilon_i.$$

In particular, it holds that $RSS' = \sum \varepsilon_i'^2 = \sum \varepsilon_i^2 = RSS$ and

$$\hat{\sigma}' = \sqrt{RSS'/(n-2)} = \sqrt{RSS/(n-2)} = \hat{\sigma}.$$

R^2 does not change, $R'^2 = R^2$ (as Y does not change, i.e. $Y_i' = Y_i$, neither do the corresponding estimates, i.e. $\hat{Y}_i' = \hat{Y}_i$).

(b) (4 points) We now consider the transformation $Y' = 5Y$ and the regression of Y' on X. As above, give estimates of $\hat{\alpha}'$, $\hat{\beta}', \hat{\sigma}'$ and R'^2.

Solution: (1 point for each correct estimate.)

Let $Y' = 5Y$. Then,

$$\begin{aligned}\hat{\beta}' &= \tfrac{\sum(X_i-\bar{X})(Y_i'-\bar{Y}')}{\sum(X_i-\bar{X})^2} = \tfrac{5\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = 5\hat{\beta},\\ \hat{\alpha}' &= \bar{Y}' - \hat{\beta}'\bar{X} = 5\bar{Y} - 5\hat{\beta}\bar{X} = 5(\bar{Y} - \hat{\beta}\bar{X}) = 5\hat{\alpha}.\end{aligned}$$

So, we have

$$\hat{Y}_i' = \hat{\alpha}' + \hat{\beta}' X_i' = (5\hat{\alpha}) + (5\hat{\beta})X_i = 5(\hat{\alpha} + \hat{\beta}X_i) = 5\hat{Y}_i.$$

Thus, for the residuals,

$$\varepsilon_i' = Y_i' - \hat{Y}_i' = (5Y_i) - (5\hat{Y}_i) = 5(Y_i - \hat{Y}_i) = 5\varepsilon_i.$$

We obtain

$$RSS' = \sum \varepsilon_i'^2 = 25 \cdot RSS,$$

and

$$\hat{\sigma}' = \sqrt{RSS'/(n-2)} = 5\hat{\sigma}.$$

Finally,

$$R'^2 = \frac{\|\hat{Y}' - \bar{Y}'\|_2^2}{\|Y' - \bar{Y}'\|_2^2} = \frac{25\|\hat{Y} - \bar{Y}\|_2^2}{25\|Y - \bar{Y}\|_2^2} = R^2.$$

The case $Y' = Y + 10$ can be treated the same way; only the intercept changes.