MATH20811 - Practical Statistics

Coursework 3 Submission

Firstly, import the data using the following R code:

```
> sweden_ins_data<-read.table("sweden_ins_data.txt",header = TRUE)
```

1. Here is a scatterplot of the data and the R code used to produce the plot. This shows a strong linear relationship between Payment and Claims. The above regression model is suitable for the data.

   ```
   > plot(sweden_ins_data$claims,sweden_ins_data$payment,main =
   "The Sweden Insurance Data",xlab = "Claims",ylab = "Payment")
   ```

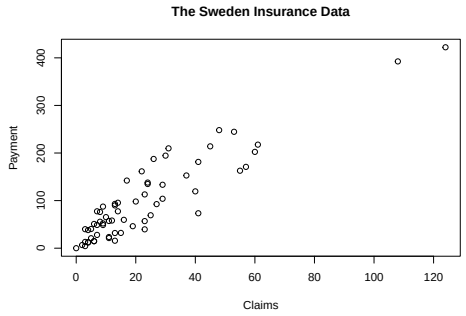

Figure 1: Scatterplot of the Sweden Insurance Data

2. The principle of least squares is to estimate the parameters by values that make the sum of squares as small as possible.

   Let S be the sum of squares,

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

   Its minimum may be found by solving the system

$$\frac{\partial S}{\partial \alpha} = 0, \frac{\partial S}{\partial \beta} = 0$$

Note that

$$\frac{\partial \epsilon_i}{\partial \alpha} = -1, \frac{\partial \epsilon_i}{\partial \beta} = -x_i$$

The following notation will be used

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

and

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

$$\frac{\partial S}{\partial \alpha} = -2n(\bar{y} - \alpha - \beta\bar{x}), \frac{\partial S}{\partial \beta} = -2\sum_{i=1}^{n} x_i(y_i - \alpha - \beta x_i)$$

Equating to zero, this system can be rewritten as

$$\bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = 0, \sum_{i=1}^{n} x_i(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0$$

From the first equation of the system we get

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Putting this in the second equation we get

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

The following R code is my own function to fit a simple linear regression model:

```
> sweden_lm<-function(x,y){
+   n=length(x)
+   df=n-2
+   x_bar=sum(x)/n
+   y_bar=sum(y)/n
```

```
+   Sxx=sum(x^2)-(sum(x))^2/n
+   Sxy=sum(x*y)-sum(x)*sum(y)/n
+   beta_hat=Sxy/Sxx
+   alpha_hat=y_bar-beta_hat*x_bar
+   ei_hat=y-alpha_hat-beta_hat*x
+   my_list<-list(c(alpha_hat,beta_hat),c(x,y),ei_hat,df)
+   return(my_list)
+ }
>
> sweden_lm(sweden_ins_data$claims,sweden_ins_data$payment)
[[1]]
[1] 19.994486  3.413824

[[2]]
  [1] 108.0  19.0  13.0 124.0  40.0  57.0  23.0  14.0  45.0
  10.0   5.0  48.0  11.0  23.0
 [15]   7.0   2.0  24.0   6.0   3.0  23.0   6.0   9.0   9.0
 3.0  29.0   7.0   4.0  20.0
 [29]   7.0   4.0   0.0  25.0   6.0   5.0  22.0  11.0  61.0
 12.0   4.0  16.0  13.0  60.0
 [43]  41.0  37.0  55.0  41.0  11.0  27.0   8.0   3.0  17.0
 13.0  13.0  15.0   8.0  29.0
 [57]  30.0  24.0   9.0  31.0  14.0  53.0  26.0 392.5  46.2
 15.7 422.2 119.4 170.9  56.9
 [71]  77.5 214.0  65.3  20.9 248.1  23.5  39.6  48.8   6.6
 134.9  50.9   4.4 113.0  14.8
 [85]  48.7  52.1  13.2 103.9  77.5  11.8  98.1  27.9  38.1
 0.0  69.2  14.6  40.3 161.5
 [99]  57.2 217.6  58.1  12.6  59.6  89.9 202.4 181.3 152.8
 162.8  73.4  21.3  92.6  76.1
[113]  39.9 142.1  93.0  31.9  32.1  55.6 133.3 194.5 137.9
87.4 209.8  95.5 244.6 187.5

[[3]]
 [1]   3.8125698 -38.6571334 -48.6741920 -21.1086072 -37.1474282
 -43.6824287 -41.6124276
 [8]   9.7119844  40.3834540  11.1672786 -16.1636036  64.2419834
 -34.0465449 -58.9124276
[15]   4.9087493 -20.2221329  32.9737488  10.4225729 -25.8359564
14.4875724 -25.6774271
```

```
[22]  -2.0188978   1.3811022 -17.0359564 -15.0953690  33.6087493
-21.8497800   9.8290430
[29] -15.9912507   4.4502200 -19.9944858 -36.1400748 -25.8774271
3.2363964  66.4013959
[36]  -0.3465449 -10.6377229  -2.8603685 -21.0497800 -15.0156627
25.5258080 -22.4238994
[43]  21.3387483   6.4940425 -44.9547816 -86.5612517 -36.2465449
-19.5677219  28.7949258
[50]   9.6640436  64.0705137  28.6258080 -32.4741920 -39.1018392
8.2949258  14.3046310
[57]  72.0908074  35.9737488  36.6811022  83.9769839  27.7119844
43.6728656  78.7461017

[[4]]
[1] 61
```

3. (i) $\hat{\alpha} = 19.994486$ and $\hat{\beta} = 3.413824$.

   The following R code is used to superimpose the fitted regression line onto the scatterplot of the data:

```
> mod1<-lm(sweden_ins_data$payment~sweden_ins_data$claims)
> summary(mod1)

Call:
lm(formula = sweden_ins_data$payment ~ sweden_ins_data
$claims)

Residuals:
    Min      1Q  Median      3Q     Max
-86.561 -24.051  -0.347  23.432  83.977

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            19.9945     6.3678    3.14   0.0026 **
sweden_ins_data$claims  3.4138     0.1955   17.46
<2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1
```

```
Residual standard error: 35.94 on 61 degrees of freedom
Multiple R-squared:  0.8333,Adjusted R-squared:  0.8306
F-statistic:   305 on 1 and 61 DF,  p-value: < 2.2e-16
> abline(mod1)
```

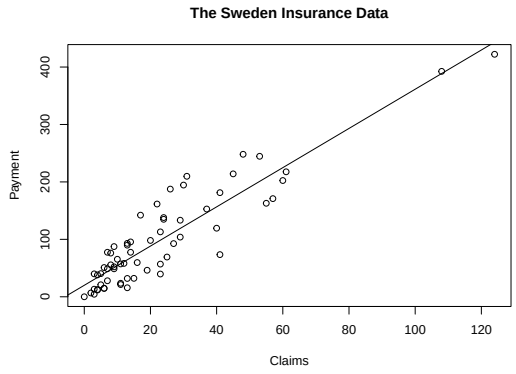

Figure 2: Scatterplot of the Data with the Regression Line

The intercept of the regression line is 19.994486, which means when the number of claims is 0, the total payment for all the claims in thousands of Swedish Kronor for geographical zones in Sweden is 19.9945. The slope is 3.413824, which means that the mean increase of the total payment for all the claims in thousands of Swedish Kronor for geographical zones in Sweden is 3.413824 units for every additional one unit in claims. The regression line is not flat.

The p-value is smaller than 2.2×10^{-16}. Since it is smaller than the significance level, we can conclude that there is significant evidence to reject the null hypotheses that the independent variable (claims) has no correlation with the dependent variable (payment), so there is association between the changes in the independent variable and the shifts in the dependent variable.

(ii) The OLS estimator is known to be unbiased.
The value of the sum of squares evaluated at $(\hat{\alpha}, \hat{\beta})$ is called error sum of squares, SSE, or residual sum of squares, RSS. So,

$$SSE = S(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

We also estimate σ^2 by $\hat{\sigma^2} = SSE/(n-2)$.

```
> output=sweden_lm(sweden_ins_data$claims,sweden_ins_data
$payment)
> ei_hat_list=output[3]
> ei_hat=unlist(ei_hat_list)
> sse=sum((ei_hat)^2)
> n=63
> error_variance=sse/(n-2)
> error_variance
[1] 1291.75
```

From the code above, $\hat{\sigma}^2 = 1291.75$.

(iii) Here is ANOVA table for the fitted model:

```
> anova(mod1)
Analysis of Variance Table

Response: sweden_ins_data$payment
                       Df Sum Sq Mean Sq F value    Pr(>F)
sweden_ins_data$claims  1 394022  394022  305.03 < 2.2e-16 ***
Residuals              61  78797    1292
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We use the F-statistic to test the null hypothesis that our model is not better than a model containing just the intercept. Since the p-value corresponding to the F-statistic is 2.2×10^{-16} , which is smaller than the significance level, we confidently reject the null hypothesis and conclude that our model is better than the model containing only the intercept.

(iv) A 95% confidence interval for $E[Y|x = 80] = \alpha + \beta \times 80$ is:

$$\left(\hat{\alpha} + \hat{\beta} \times 80 \pm t_{\alpha/2;61} \hat{\sigma} \sqrt{\frac{1}{63} + \frac{(80 - \bar{x})^2}{S_{xx}}} \right)$$

```
> n<-length(sweden_ins_data$payment)
> n
[1] 63
> summ1<-summary(mod1)$coefficients
```

```
> est.80<-summ1[1,1]+summ1[2,1]*80
> est.80
[1] 293.1004
> Sxx<-sum((sweden_ins_data$claims-mean(sweden_ins_data
$claims))^2)
> Sxx
[1] 33809.43
> est.se<-sqrt(error_variance)
> est.se
[1] 35.94092
> c11<-est.80-qt(0.975,df=n-2)*est.se*sqrt((1/n)
+(80-mean(sweden_ins_data$claims))^2/Sxx)
> c12<-est.80+qt(0.975,df=n-2)*est.se*sqrt((1/n)
+(80-mean(sweden_ins_data$claims))^2/Sxx)
> c11;c12
[1] 269.0173
[1] 317.1834
```

The estimated 95% CI is (269.0173, 317.1834).

4. a)

```
> std.res1<-rstandard(mod1)
> plot(mod1$fitted.values,std.res1,main="Standardised
residuals vs fitted values")
> abline(h=0)
```

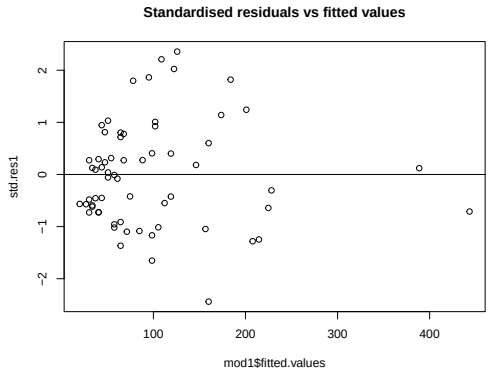

Figure 3: Diagnositic Plot of the Standardised Residuals vs Fitted Values

b)

```
> plot(sweden_ins_data$claims,std.res1,main="Standardised
residuals vs covariate values")
> abline(h=0)
```

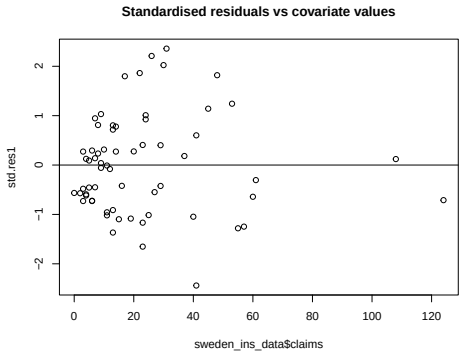

Figure 4: Diagnositic Plot of the Standardised Residuals vs the Predictor

We made assumptions that the model has been fitted to the data. If the assumptions are valid, the variance of the residuals should be a constant and not depend on the values of the predictor. Nevertheless, from the plots we observe that the variance of the residuals seems to depend on the fitted values and the predictor variable in that it first increases and then decreases as both values increase, which is a contradiction to the assumptions of the linear regression model, so the model is not adequate.

5. In this part we manually construct a Normal Q-Q plot of the data. The sample quantiles correspond to the sorted data. The theoretical quantiles here are given by the values in the quantile function $p(k) = \frac{(k-3/8)}{(63+1/4)}$, k=1,...,63.

```
> sample.quantiles<-sort(std.res1)
> pn=0
> for (k in 1:n) {
+   pn[k]=((k-3/8)/(n+1/4))
+ }
> theoretical.quantiles<-qnorm(pn)
```

```
> windows()
> plot(theoretical.quantiles,sample.quantiles,main="Normal
Q-Q plot of Sweden Insurance data (n=63)")
> abline(a=mean(std.res1), b=sd(std.res1), col="green")
```

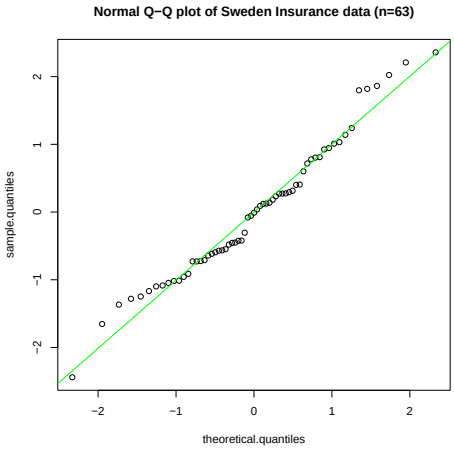

Figure 5: Normal Q-Q Plot of the Data

The Q-Q plot shows that the sample quantiles (ordered data) lie reasonably close to the reference line. Normality of the data does seem to be a reasonable assumption then.

6. We carry out a Kolmogorov-Smirnov test to test the null hypothesis H_0: the data have a N(0, 1) distribution vs H_1: the data are not distributed as specified under H_0.
We firstly use the R function ks.test to conduct a two-sided test at the 5% significance level.

```
> ks.test(x=std.res1, y=pnorm, mean=0, sd=1,alternative
= c("two.sided"))

Exact one-sample Kolmogorov-Smirnov test

data:  std.res1
D = 0.10773, p-value = 0.4277
alternative hypothesis: two-sided
```

The observed value of the test statistic is D = 0.10773 with a p-value of 0.4277. This p-value is greater than 0.05 so we can accept H_0 at the 5% significance level and conclude that the data are a random sample from the N(0, 1) distribution.

We now compute the value of D manually in R.

```
> std.res.ord=sort(std.res1)
> x.ecdf<-(1:n)/n
> y<-pnorm(std.res.ord, mean=0, sd=1)
> diff1=x.ecdf-y
> md1=max(diff1)
> dn.plus=max(md1,0)
> dn.plus
[1] 0.1077294
> x.KS1=std.res.ord[dn.plus==diff1]
> diff2=y-x.ecdf
> md2=max(diff2)
> md2=md2+(1/n)
> dn.minus=max(md2,0)
> dn.minus
[1] 0.05914349
> x.KS2=std.res.ord[dn.minus==diff2+(1/n)]
> KSstat=max(dn.minus, dn.plus)
> KSstat
[1] 0.1077294
> if(dn.minus < dn.plus) x.KSstat=x.KS1
> if(dn.minus > dn.plus) x.KSstat=x.KS2
> x.KSstat
-0.4214452
```

The calculated value of D is 0.1077294 which agrees with value computed using the ks.test function. We also find that the largest difference between the empirical cdf and the N(0, 1) cdf occurs at x = -0.4214452.

7. We firstly calculate the minimum and maximum values to aid plotting the Normal cdf.

```
> min(std.res1)
```

```
[1] -2.439812
> max(std.res1)
[1] 2.357619
```

Here we use R to plot the empirical cdf and superimpose the N(0, 1) cdf.

```
> plot.ecdf(std.res1,main="Ecdf and N(0, 1) cdf for the Sweden
Insurance data")
> x.grid=seq(from=-2.5,to=2.5,length.out=500)
> x.grid.pnorm<-pnorm(x.grid,mean = 0,sd=1)
> lines(x.grid,x.grid.pnorm,col="green",type="l")
> points(x.KSstat,0,pch=15,col="red")
> abline(v=x.KSstat,col="blue")
```

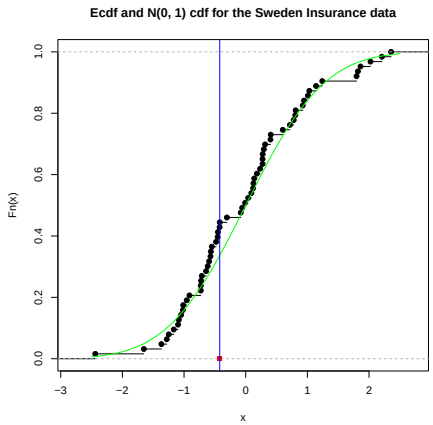

Figure 6: Ecdf and N(0, 1) cdf for the Sweden Insurance data

The red dot and blue vertical line indicate that x=-0.4214452(x.KSstat) is the point at which the maximum difference between the curves occurs.

8. Using the following code we can simulate B = 2000 random vectors from the N(0,1) distribution. For each data set, the statistic is calculated using R function ks.test and the its value stored in a vector.

```
test.sim.fun<-function(N,Nmean,Nsd){
```

```
B=2000
test.sim=0
for (i in 1:B) {
  ysim<-rnorm(N,Nmean,Nsd)
  test.sim[i]<-ks.test(x=ysim, y=pnorm,Nmean,Nsd,alternative
  = c("two.sided"))$statistic
}
return(test.sim)
}
```

The following code then produces a histogram of the estimated sampling distribution and superimposes N(0,1) pdf on to it.

```
> hist(test.sim.fun(63,0,1),freq = F,xlim=c(0,0.25),main =
"Histogram of simulated test statistic values (B=2000) with
N(0,1) pdf",xlab="Simulation size is B=2000;the curve is the
kde of the distribution")
> lines(density(test.sim.fun(63,0,1)), col="red")
```

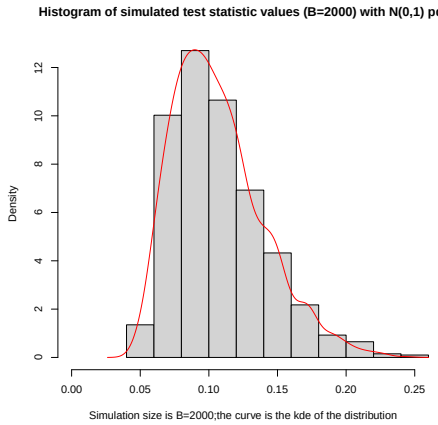

Figure 7: Histogram of simulated test statistic values with N(0,1) pdf

The following code obtains an estimated 5% critical value. The calculated value corresponds to the observed value.

```
> quantile(test.sim.fun(63,0,1),0.95)
      95%
0.1644367
```