STK310 PRACTICAL ASSIGNMENT A6 – SUGGESTED SOLUTION

Where applicable, the given answers are from the SAS output. The answers from the R output will be equivalent, but might differ slightly with respect to the number of decimal places given.

Question 1

SAS Code & Output

```
goptions reset=all;
proc format;
value $city  'RdJ'='Rio de Janeiro'
             'Bra'='Brasilia'
             'SPa'='São Paulo'
             'For'='Fortaleza'
             'BHo'='Belo Horizonte'
             'PAl'='Porto Alegre'
             'Sal'='Salvador'
             'Rec'='Recife'
             'Cui'='Cuiabá'
             'Man'='Manaus'
             'Nat'='Natal'
             'Cur'='Curitiba';
value $constr 'R'='Renovated'
             'N'='New';
data fifa2014;
input city$ constr$ capacity cost @@;
datalines;
RdJ R 78800 320
Bra N 70064 460
SPa N 65807 230
For R 64846 171
BHo R 62547 220
PAl R 48849  95
Sal N 48747 192
Rec N 44248 181
Cui N 42968 195
Man N 42374 174
Nat N 42086 315
Cur R 41456  56
;
run;
data q1_dummy;
set fifa2014;
y=cost;
x=capacity;
d=0;
if constr='N' then d=1;
keep city y x d;
run;
goptions reset=all;
title1 'Regression model using capacity and construction type to explain construction cost';
proc reg data=q1_dummy plot=none;
      model y=x d / cli;
      id city;
      format city $city.;
run;
```

```
data q1_model;
set q1_dummy;
yhat_r=-201.78472+0.00631*x;
yhat_n=-71.60581+0.00631*x;
run;
goptions reset=all;
axis1 label=('Capacity');
axis2 label=(angle=90 'Construction cost (million pounds)');
axis3 label=none;
legend1 label=('Construction type:') value=('Renovated' 'New');
legend2 label=('Regression line:')
             value=('Renovated: y=-201.78472+0.00631*x' 'New: y=-71.60581+0.00631*x');
symbol1 color=blue value=dot;
symbol2 color=red value=trianglefilled;
symbol3 color=blue i=join line=1;
symbol4 color=red i=join line=3;
title1 'Scatter diagram of construction cost against capacity of stadiums';
title2 'Different regression lines based on construction type (renovated vs new)';
proc gplot data=q1_model;
      plot y*x=d / haxis=axis1 vaxis=axis2 legend=legend1;
      plot2 (yhat_r yhat_n)*x / overlay haxis=axis1 vaxis=axis3 legend=legend2;
run;
```

Regression model using capacity and construction type to explain construction cost

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read 12

Number of Observations Used 12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	85092	42546	9.22	0.0066
Error	9	41521	4613.39227		
Corrected Total	11	126613			

Root MSE	67.92196	**R-Square**	0.6721
Dependent Mean	217.41667	**Adj R-Sq**	0.5992
Coeff Var	31.24046		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-201.78472	102.27777	-1.97	0.0800
x	1	0.00631	0.00165	3.83	0.0040
d	1	130.17891	42.10871	3.09	0.0129

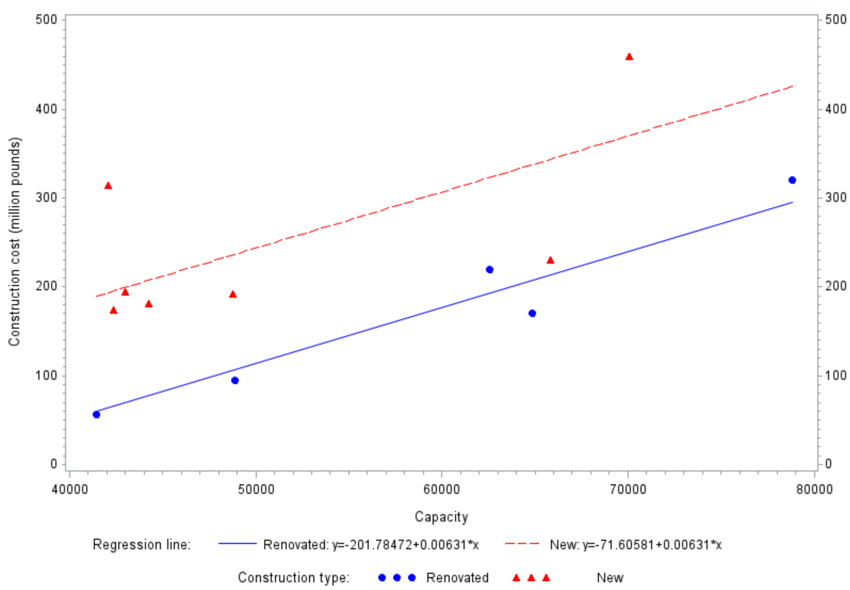
Output Statistics

Obs	city	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	Rio de J	320.0000	295.4489	44.2054	112.1233	478.7745	24.5511
2	Brasilia	460.0000	370.5030	40.6854	191.3965	549.6096	89.4970
3	São Paul	230.0000	343.6411	35.5229	170.2461	517.0361	-113.6411
4	Fortalez	171.0000	207.3982	31.7194	37.8192	376.9772	-36.3982
5	Belo Hor	220.0000	192.8913	30.8429	24.1418	361.6409	27.1087
6	Porto Al	95.0000	106.4560	34.9130	-66.3040	279.2159	-11.4560
7	Salvador	192.0000	235.9913	25.9156	71.5368	400.4457	-43.9913
8	Recife	181.0000	207.6022	27.9114	41.4847	373.7198	-26.6022
9	Cuiabá	195.0000	199.5253	28.8041	32.6298	366.4209	-4.5253
10	Manaus	174.0000	195.7772	29.2607	28.4757	363.0787	-21.7772
11	Natal	315.0000	193.9599	29.4912	26.4514	361.4684	121.0401
12	Curitiba	56.0000	59.8056	42.2646	-121.1626	240.7738	-3.8056

Sum of Residuals	0
Sum of Squared Residuals	41521
Predicted Residual SS (PRESS)	76050

Scatter diagram of construction cost against capacity of stadiums

Different regression lines based on construction type (renovated vs new)

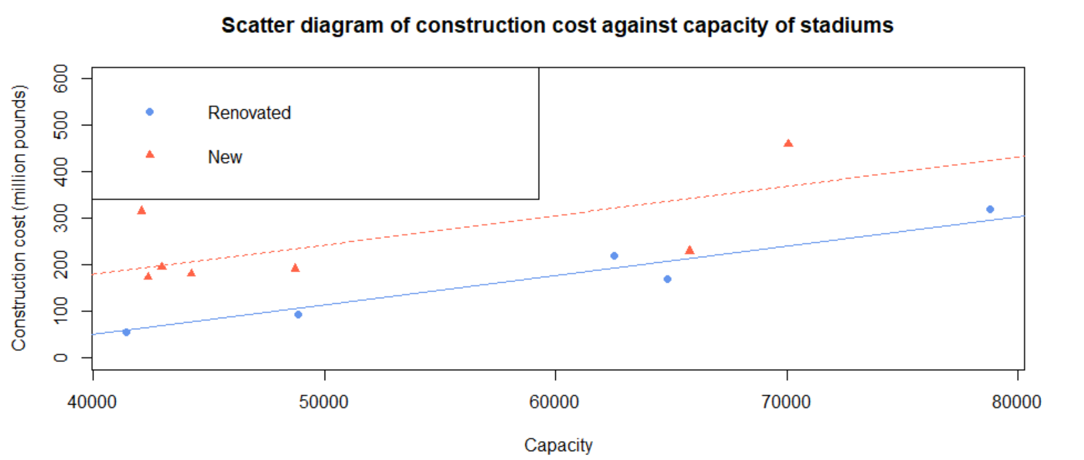


R Code & Output

```
> fifa2014 <- read.csv("c:\\2014-fifa-world-cup.csv", header = T)
> y <- fifa2014$Construction.cost
> x <- fifa2014$Capacity
> d <- ifelse(fifa2014$Construction.type == "R", 0, 1)
> (lrm_q1 <- lm(y ~ x + d, data = fifa2014))

Call:
lm(formula = y ~ x + d, data = fifa2014)

Coefficients:
(Intercept)            x            d
 -201.78472      0.00631    130.17891

> summary(lrm_q1)

Call:
lm(formula = y ~ x + d, data = fifa2014)

Residuals:
     Min       1Q   Median       3Q      Max
-113.641  -29.051   -7.991   25.190  121.040

Coefficients:
               Estimate  Std. Error t value Pr(>|t|)
(Intercept) -201.784720  102.277772  -1.973  0.07997 .
x              0.006310    0.001647   3.831  0.00402 **
d            130.178914   42.108711   3.091  0.01290 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67.92 on 9 degrees of freedom
Multiple R-squared:  0.6721,     Adjusted R-squared:  0.5992
F-statistic: 9.222 on 2 and 9 DF,  p-value: 0.006623

> predict(lrm_q1, interval="predict")
         fit        lwr      upr
1  295.44892  112.12330 478.7745
2  370.50305  191.39652 549.6096
3  343.64107  170.24606 517.0361
4  207.39818   37.81918 376.9772
5  192.89133   24.14176 361.6409
6  106.45597  -66.30395 279.2159
7  235.99125   71.53676 400.4457
8  207.60224   41.48473 373.7198
9  199.52535   32.62984 366.4209
10 195.77717   28.47567 363.0787
11 193.95987   26.45135 361.4684
12  59.80561 -121.16256 240.7738
```



```
> beta1hat <- summary(lrm_q1)$coef[1,1]
> beta2hat <- summary(lrm_q1)$coef[2,1]
> beta3hat <- summary(lrm_q1)$coef[3,1]
> cols <- c("cornflowerblue", "tomato")
> plot(x, y, main = "Scatter diagram of construction cost against capacity of stadiums",
+      pch = d + 16, col = cols[factor(d)], xlab = 'Capacity',
+      ylab = 'Construction cost (million pounds)', ylim = c(0, 600))
> legend("topleft", legend = c("Renovated", "New"), pch = d + 16, col = cols[factor(d)])
> abline(a = beta1hat, b = beta2hat, col = "cornflowerblue")
> abline(a = beta1hat + beta3hat, b = beta2hat, col = "tomato", lty = 2)
```



Fitted regression model:

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\beta}_3 D_i \\ &= -201.78472 + 0.00631 X_i + 130.17891 D_i\end{aligned}$$

Regression line for renovated stadium:

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i \text{ since } D_i = 0 \\ &= -201.78472 + 0.00631 X_i\end{aligned}$$

Regression line for new stadium:

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\beta}_3 \text{ since } D_i = 1 \\ &= -201.78472 + 0.00631 X_i + 130.17891 \\ &= -71.60581 + 0.00631 X_i\end{aligned}$$

Question 2

SAS Code & Output

```
data q2_dummy;
set sasuser.videos;
y=file_size;
x=song_length;
d=0;
if video_quality='High' then d=1;
dx=d*x;
keep y x d dx;
run;
goptions reset=all;
title1 'Regression model using song lengths and video quality to explain file sizes';
proc reg data=q2_dummy plot=none;
      model y=x dx;
run;
data q2_model;
set q2_dummy;
yhat_l=-12.12687+0.10436*x;
yhat_h=-12.12687+0.14191*x;
run;
goptions reset=all;
axis1 label=('Song length (seconds)') order = 50 to 450 by 50;
axis2 label=(angle=90 'File size (MB)') order = -10 to 70 by 10;
axis3 label=none order = -10 to 70 by 10;
legend1 label=('Video quality:') value=('Low' 'High');
legend2 label=('Regression line:')
             value=('Low: y=-12.12687+0.10436*x' 'High: y=-12.12687+0.14191*x');
symbol1 color=green value=dot;
symbol2 color=brown value=trianglefilled;
symbol3 color=green i=join line=1;
symbol4 color=brown i=join line=1;
title1 'Scatter diagram for the file sizes against the song lengths';
title2 'Different regression lines based on video quality (low vs high)';
proc gplot data=q2_model;
      plot y*x=d / haxis=axis1 vaxis=axis2 legend=legend1;
      plot2 (yhat_l yhat_h)*x / overlay haxis=axis1 vaxis=axis3 legend=legend2;
run;
```

Regression model using song lengths and video quality to explain file sizes

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	26
Number of Observations Used	26

Analysis of Variance

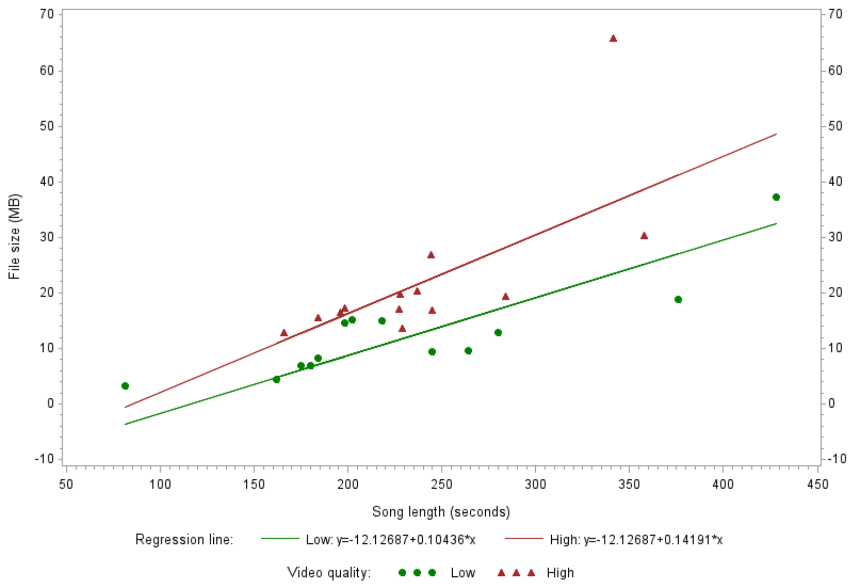
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2463.93544	1231.96772	19.94	<.0001
Error	23	1421.37071	61.79873		
Corrected Total	25	3885.30615			

Root MSE	7.86122	**R-Square**	0.6342
Dependent Mean	17.47692	**Adj R-Sq**	0.6024
Coeff Var	44.98055		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-10.27856	5.19073	-1.98	0.0598
x	1	0.09671	0.02157	4.48	0.0002
dx	1	0.04107	0.01251	3.28	0.0033

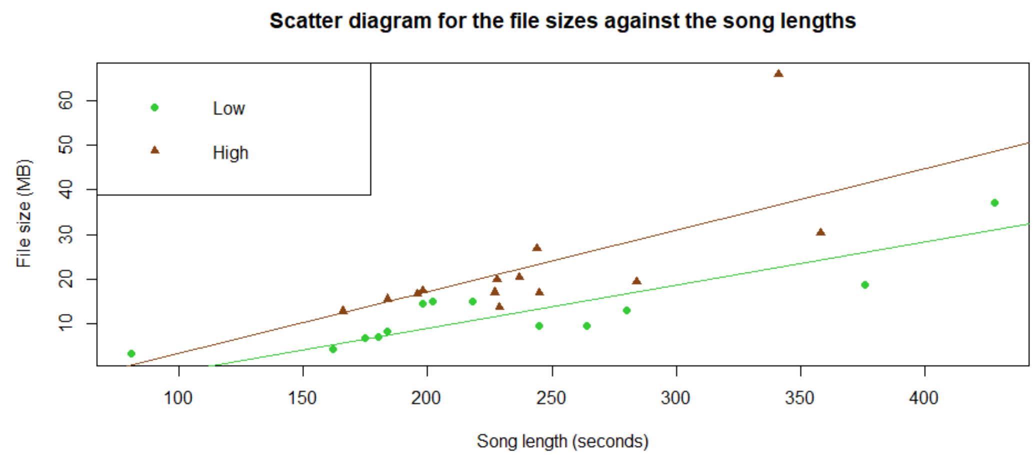
Scatter diagram for the file sizes against the song lengths
Different regression lines based on video quality (low vs high)



R Code & Output

```
> videos <- read.csv("c:\\videos.csv", header = T)
> y <- videos$File.Size
> x <- videos$Song.Length
> d <- ifelse(videos$Video.Quality == "Low", 0, 1)
> dx <- d * x
> (lrm_q2 <- lm(y ~ x + dx, data = videos))

Call:
lm(formula = y ~ x + dx, data = videos)

Coefficients:
(Intercept)            x           dx
  -10.27856      0.09671      0.04107

> summary(lrm_q2)

Call:
lm(formula = y ~ x + dx, data = videos)

Residuals:
   Min     1Q Median     3Q    Max
-9.550 -3.976 -0.127  2.767 29.097
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.27856    5.19073  -1.980 0.059769 .
x             0.09671    0.02157   4.484 0.000168 ***
dx            0.04107    0.01251   3.283 0.003262 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.861 on 23 degrees of freedom
Multiple R-squared:  0.6342,     Adjusted R-squared:  0.6024
F-statistic: 19.94 on 2 and 23 DF,  p-value: 0.0000095

> beta1hat <- summary(lrm_q2)$coef[1,1]
> beta2hat <- summary(lrm_q2)$coef[2,1]
> beta3hat <- summary(lrm_q2)$coef[3,1]
> cols <- c("limegreen", "saddlebrown")
> plot(x, y, main = "Scatter diagram for the file sizes against the song lengths",
+      pch = d + 16, col = cols[factor(d)], xlab = 'Song length (seconds)',
+      ylab = 'File size (MB)')
> legend("topleft", legend = c("Low", "High"), pch = d + 16, col = cols[factor(d)])
> abline(a = beta1hat, b = beta2hat, col = "limegreen")
> abline(a = beta1hat, b = beta2hat + beta3hat, col = "saddlebrown")
```



Fitted regression model:

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\beta}_3 D_i X_i \\ &= -10.27856 + 0.09671 X_i + 0.04107 D_i X_i\end{aligned}$$

Regression line for videos with low quality:

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i \text{ since } D_i = 0 \\ &= -10.27856 + 0.09671 X_i\end{aligned}$$

Regression line for videos with high quality:

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\beta}_3 X_i \text{ since } D_i = 1 \\ &= -10.27856 + 0.09671 X_i + 0.04107 X_i \\ &= -10.27856 + 0.13778 X_i\end{aligned}$$

Question 2 EXTRA

SAS Code & Output

```
data q2_dummy;
set sasuser.videos;
y=file_size;
x=song_length;
d=0;
if video_quality='High' then d=1;
dx=d*x;
keep y x d dx;
run;
goptions reset=all;
title1 'Regression model using song lengths and video quality to explain file sizes';
proc reg data=q2_dummy plot=none;
      model y=x d dx;
run;
data q2_xtra;
set q2_dummy;
yhat_l=-6.31131+0.08167*x;
yhat_h=-20.65506+0.17868*x;
run;
goptions reset=all;
axis1 label=('Song length (seconds)') order = 50 to 450 by 50;
axis2 label=(angle=90 'File size (MB)') order = -10 to 70 by 10;
axis3 label=none order = -10 to 70 by 10;
legend1 label=('Video quality:') value=('Low' 'High');
legend2 label=('Regression line:')
              value=('Low: y=-6.31131+0.08167*x' 'High: y=-20.65506+0.17868*x');
symbol1 color=green height=2 value=dot;
symbol2 color=brown height=2 value=trianglefilled;
symbol3 color=green i=join line=1;
symbol4 color=brown i=join line=1;
title1 'Scatter diagram for the file sizes against the song lengths';
title2 'Different regression lines based on video quality (low vs high)';
proc gplot data=q2_xtra;
      plot y*x=d / haxis=axis1 vaxis=axis2 legend=legend1;
      plot2 (yhat_l yhat_h)*x / overlay haxis=axis1 vaxis=axis3 legend=legend2;
run;
```

Regression model using song lengths and video quality to explain file sizes

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read 26

Number of Observations Used 26

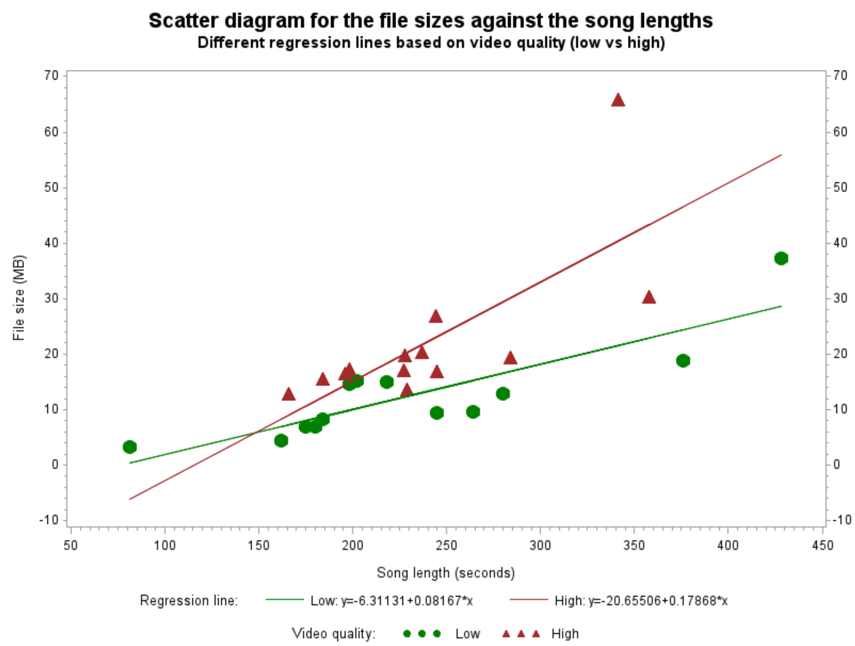
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2558.35546	852.78515	14.14	<.0001
Error	22	1326.95069	60.31594		
Corrected Total	25	3885.30615			

Root MSE	7.76633	**R-Square**	0.6585
Dependent Mean	17.47692	**Adj R-Sq**	0.6119
Coeff Var	44.43765		

Parameter Estimates

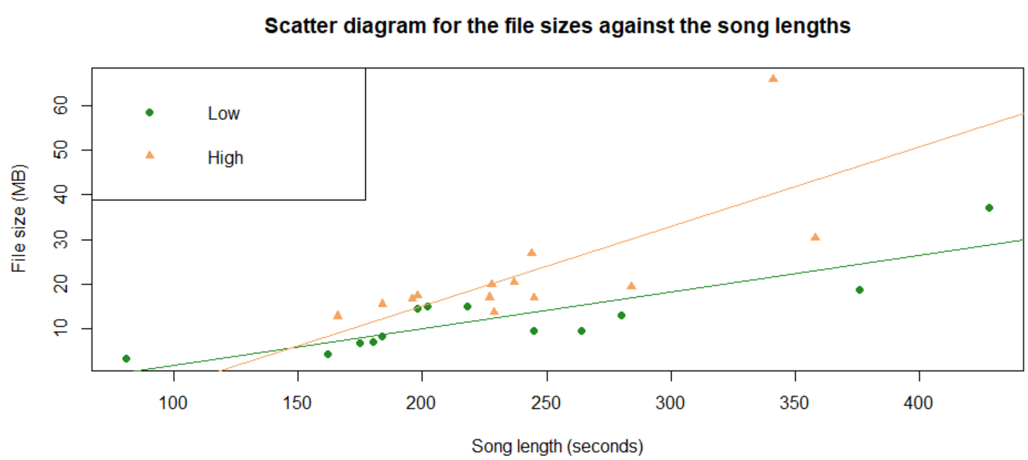
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-6.31131	6.02921	-1.05	0.3066
x	1	0.08167	0.02446	3.34	0.0030
d	1	-14.34375	11.46428	-1.25	0.2240
dx	1	0.09701	0.04638	2.09	0.0483



R Code & Output

```
> videos <- read.csv("c:\\videos.csv", header = T)
> y <- videos$File.Size
> x <- videos$Song.Length
> d <- ifelse(videos$Video.Quality == "Low", 0, 1)
> dx <- d * x
> (lrm_q2_xtra <- lm(y ~ x + d + dx, data = videos))

Call:
lm(formula = y ~ x + d + dx, data = videos)

Coefficients:
(Intercept)            x            d           dx
   -6.31131      0.08167    -14.34375      0.09701

> summary(lrm_q2_xtra)

Call:
lm(formula = y ~ x + d + dx, data = videos)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-13.012  -4.063  -0.799   3.450  25.526 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -6.31131    6.02921  -1.047  0.30656   
x             0.08167    0.02446   3.339  0.00297 **
d           -14.34375   11.46428  -1.251  0.22401   
dx            0.09701    0.04638   2.091  0.04825 * 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.766 on 22 degrees of freedom
Multiple R-squared:  0.6585,	Adjusted R-squared:  0.6119 
F-statistic: 14.14 on 3 and 22 DF,  p-value: 0.00002364

> beta1hat <- summary(lrm_q2_xtra)$coef[1,1]
> beta2hat <- summary(lrm_q2_xtra)$coef[2,1]
> beta3hat <- summary(lrm_q2_xtra)$coef[3,1]
> beta4hat <- summary(lrm_q2_xtra)$coef[4,1]
> cols <- c("forestgreen", "sandybrown")
> plot(x, y, main = "Scatter diagram for the file sizes against the song lengths",
+      pch = d + 16, col = cols[factor(d)], xlab = 'Song length (seconds)',
+      ylab = 'File size (MB)')
> legend("topleft", legend = c("Low", "High"), pch = d + 16, col = cols[factor(d)])
> abline(a = beta1hat, b = beta2hat, col = "forestgreen")
> abline(a = beta1hat + beta3hat, b = beta2hat + beta4hat, col = "sandybrown")
```



Fitted regression model:

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\beta}_3 D_i + \hat{\beta}_4 D_i X_i \\ &= -6.31131 + 0.08167 X_i - 14.34375 D_i + 0.09701 D_i X_i\end{aligned}$$

Regression line for videos with low quality:

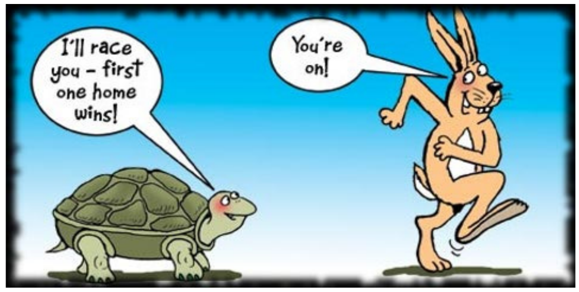
$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i \text{ since } D_i = 0 \\ &= -6.31131 + 0.08167 X_i\end{aligned}$$

Regression line for videos with high quality:

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\beta}_3 + \hat{\beta}_4 X_i \text{ since } D_i = 1 \\ &= -6.31131 + 0.08167 X_i - 14.34375 + 0.09701 X_i \\ &= -20.65506 + 0.17868 X_i\end{aligned}$$

Question 3

SAS Code & Output




```
goptions reset=all;
data tortoise;
infile 'c:\tortoises.txt';
input length clutch;
run;
data q3_poly;
set tortoise;
y=clutch;
x=length;
xsq=x**2;
run;
goptions reset=all;
title1 'Polynomial regression model using carapace length to explain clutch size';
proc reg data=q3_poly plot=none;
      model y=x xsq;
      output out=polyout p=yhat;
run;
goptions reset=all;
axis1 label=('Carapace length (mm)');
axis2 label=(angle=90 'Clutch size (number of eggs)') minor=(number=2) order = 0 to 15 by 3;
legend1 label=('Values:') value=('Observed' 'Predicted');
symbol1 color=grey value=dot;
symbol2 color=black i=spline value=trianglefilled;
title1 'Scatter diagram of clutch size against carapace length';
title2 'Second-order polynomial regression model';
proc gplot data=polyout;
      plot (y yhat)*x / overlay haxis=axis1 vaxis=axis2 legend=legend1;
run;
data q3_dummy;
set tortoise;
y=clutch;
x=length;
xstar=311;
d=0;
if x>xstar then d=1;
xmxstard=(x-xstar)*d;
run;
goptions reset=all;
title1 'Piecewise linear regression model using carapace length to explain clutch size';
proc reg data=q3_dummy plot=none;
      model y=x xmxstard;
      output out=dummyout p=yhat;
run;
goptions reset=all;
axis1 label=('Carapace length (mm)');
axis2 label=(angle=90 'Clutch size (number of eggs)') minor=(number=2) order = 0 to 15 by 3;
legend1 label=('Values:') value=('Observed' 'Predicted');
symbol1 color=grey value=dot;
symbol2 color=black i=join value=trianglefilled;
title1 'Scatter diagram of clutch size against carapace length';
title2 'Piecewise linear regression model';
proc gplot data=dummyout;
      plot (y yhat)*x / overlay haxis=axis1 vaxis=axis2 legend=legend1;
run;
```


Polynomial regression model using carapace length to explain clutch size

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read 18

Number of Observations Used 18

Analysis of Variance

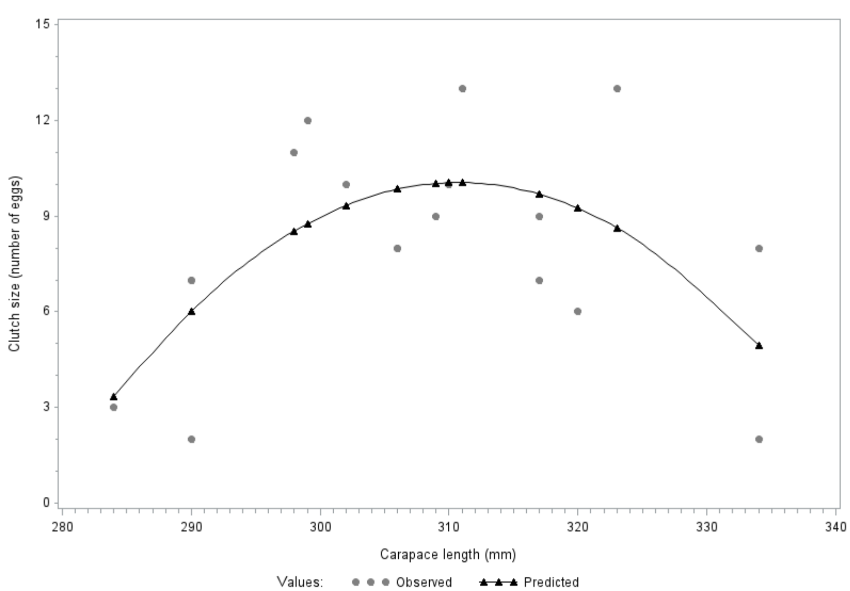
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	81.97087	40.98544	5.75	0.0140
Error	15	106.97357	7.13157		
Corrected Total	17	188.94444			

Root MSE	2.67050	**R-Square**	0.4338
Dependent Mean	8.05556	**Adj R-Sq**	0.3583
Coeff Var	33.15104		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	-899.93459	270.29576	-3.33	0.0046
x	**1**	5.85716	1.75010	3.35	0.0044
xsq	**1**	-0.00942	0.00283	-3.33	0.0045

Scatter diagram of clutch size against carapace length
Second-order polynomial regression model



Piecewise linear regression model using carapace length to explain clutch size

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read 18

Number of Observations Used 18

Analysis of Variance

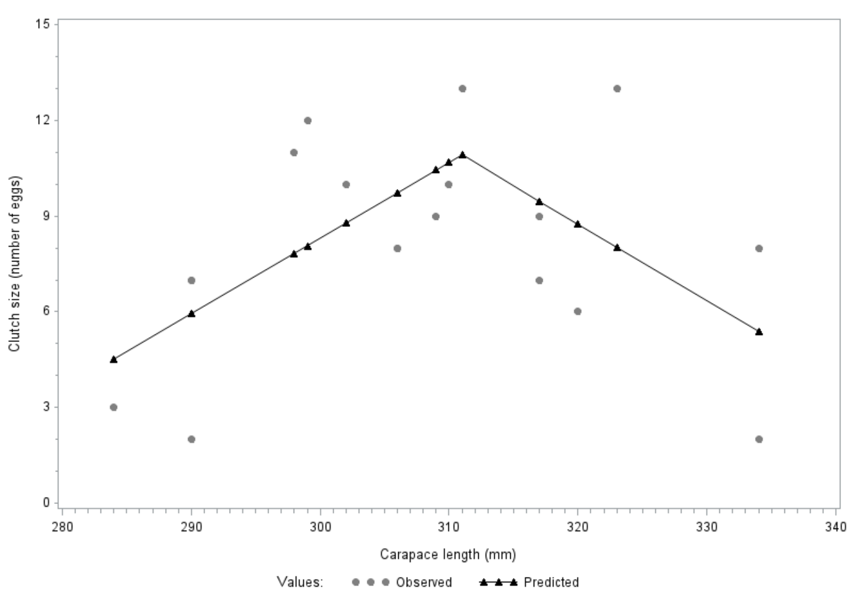
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	72.16635	36.08318	4.63	0.0271
Error	15	116.77809	7.78521		
Corrected Total	17	188.94444			

Root MSE	2.79020	**R-Square**	0.3819
Dependent Mean	8.05556	**Adj R-Sq**	0.2995
Coeff Var	34.63694		

Parameter Estimates

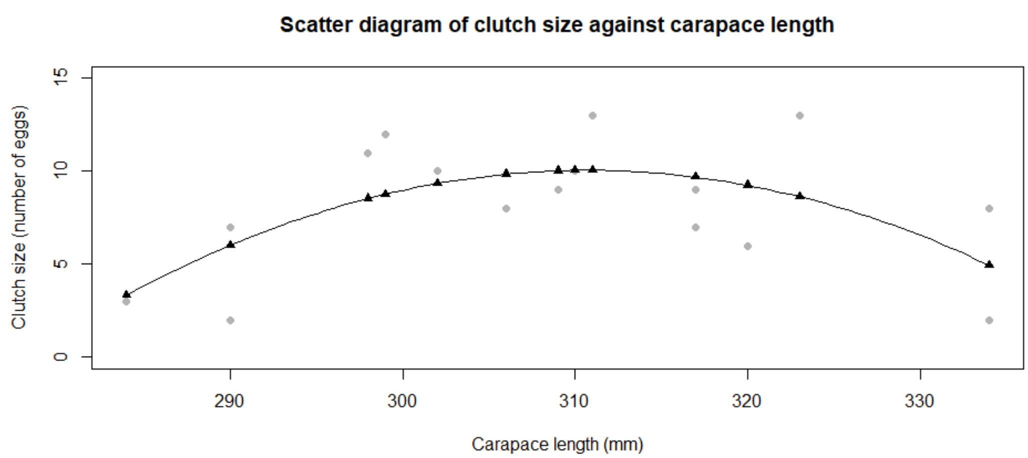
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	-63.02163	25.31801	-2.49	0.0250
x	**1**	0.23777	0.08414	2.83	0.0128
xmxstard	**1**	-0.47913	0.16050	-2.99	0.0092

Scatter diagram of clutch size against carapace length
Piecewise linear regression model

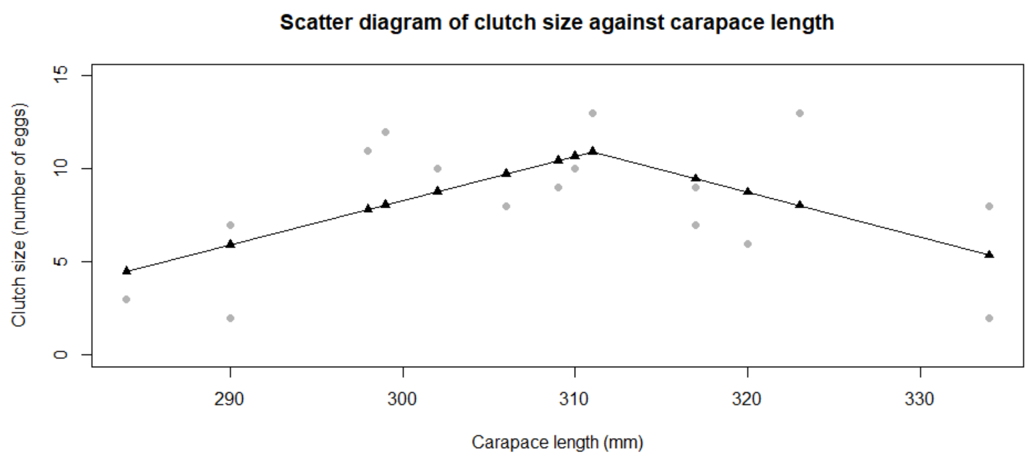


R Code & Output

```
> tortoise <- read.table("c:\\tortoises.txt")
> x <- tortoise$V1
> y <- tortoise$V2
> xsq <- x ^ 2
> (lrm_q3a <- lm(y ~ x + xsq, data = tortoise))

Call:
lm(formula = y ~ x + xsq, data = tortoise)

Coefficients:
(Intercept)            x          xsq
-899.934594     5.857158    -0.009425

> summary(lrm_q3a)

Call:
lm(formula = y ~ x + xsq, data = tortoise)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0091 -1.8480 -0.1896  2.0989  4.3605

Coefficients:
               Estimate  Std. Error t value Pr(>|t|)
(Intercept) -899.934594  270.295756  -3.329  0.00457 **
x              5.857158    1.750103   3.347  0.00441 **
xsq           -0.009425    0.002829  -3.332  0.00455 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.671 on 15 degrees of freedom
Multiple R-squared:  0.4338,     Adjusted R-squared:  0.3583
F-statistic: 5.747 on 2 and 15 DF,  p-value: 0.01403

> yhat <- lrm_q3a$fitted.values
> alpha0hat <- summary(lrm_q3a)$coef[1,1]
> alpha1hat <- summary(lrm_q3a)$coef[2,1]
> alpha2hat <- summary(lrm_q3a)$coef[3,1]
> xmin <- min(x)
> xmax <- max(x)
> plot(x, y, main = "Scatter diagram of clutch size against carapace length", pch = 16,
+      col= "gray70", xlab = 'Carapace length (mm)', ylab = 'Clutch size (number of eggs)',
+      ylim = c(0, 15))
> points(x, yhat, pch = 17, col= "black")
> curve(alpha0hat + alpha1hat * x + alpha2hat * x ^ 2, xmin, xmax, add = TRUE)
```



```
> xstar <- 311
> d <- ifelse(x > xstar, 1, 0)
> xmxstard <- (x - xstar) * d
> (lrm_q3b <- lm(y ~ x + xmxstard, data = tortoise))

Call:
lm(formula = y ~ x + xmxstard, data = tortoise)

Coefficients:
(Intercept)            x     xmxstard
   -63.0216       0.2378      -0.4791

> summary(lrm_q3b)

Call:
lm(formula = y ~ x + xmxstard, data = tortoise)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9314 -1.7357 -0.5816  1.8605  4.9717

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -63.02163   25.31801  -2.489  0.02503 *
x             0.23777    0.08414   2.826  0.01277 *
xmxstard     -0.47913    0.16050  -2.985  0.00925 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.79 on 15 degrees of freedom
Multiple R-squared:  0.3819,     Adjusted R-squared:  0.2995
F-statistic: 4.635 on 2 and 15 DF,  p-value: 0.02708

> yhat <- lrm_q3b$fitted.values
> beta1hat <- summary(lrm_q3b)$coef[1,1]
> beta2hat <- summary(lrm_q3b)$coef[2,1]
> beta3hat <- summary(lrm_q3b)$coef[3,1]
> plot(x, y, main = "Scatter diagram of clutch size against carapace length", pch = 16,
+      col= "gray70", xlab = 'Carapace length (mm)', ylab = 'Clutch size (number of eggs)',
+      ylim = c(0, 15))
> points(x, yhat, pch = 17, col= "black")
> segments(xmin, beta1hat + beta2hat * xmin, xstar, beta1hat + beta2hat * xstar, col= "black")
> segments(xstar, beta1hat + beta2hat * xstar,
+          xmax, beta1hat + beta2hat * xmax + beta3hat * (xmax - xstar), col= "black")
```



(a) Fitted polynomial regression model:

$$\begin{aligned}\hat{Y}_i &= \hat{\alpha}_0 + \hat{\alpha}_1 X_i + \hat{\alpha}_2 X_i^2 \\ &= -899.93459 + 5.85716 X_i - 0.00942 X_i^2\end{aligned}$$

(b) Fitted piecewise linear regression model:

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\beta}_3 D_i (X_i - X^*) \\ &= -63.02163 + 0.23777 X_i - 0.47913 D_i (X_i - 311)\end{aligned}$$

Regression model for length ≤ 311:

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i \text{ since } D_i = 0 \\ &= -63.02163 + 0.23777 X_i\end{aligned}$$

Regression model for length > 311:

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\beta}_3 (X_i - X^*) \text{ since } D_i = 1 \\ &= -63.02163 + 0.23777 X_i - 0.47913 (X_i - 311) \\ &= 85.9878 - 0.24136 X_i\end{aligned}$$