STAT 5385: Homework 5

Willliam Ofosu Agyapong

29/04/2022

Problem 7.3

Once again we make use of the **brand preference** data set with the following variables.

- X_1: Moisture content
- X_2: Sweetness
- Y: Degree of brand liking (brand preference)

```
brand_pref=read.table("../Data Sets/Chapter  6 Data Sets/CH06PR05.txt")
colnames(brand_pref)=c("Y","X1","X2")
# kable(head(brand), caption = "First 6 observations")
```

Part (a): ANOVA Table for Extra Sums of Squares

```
brand_mdl1 <- lm(Y ~ X1 + X2, data = brand_pref)
# summary(brand_mdl1)
anova_tbl <- anova(brand_mdl1)
rownames(anova_tbl) <- c("X1", "X2|X1", "Residuals")
# Obtain Type I sum of squares
kable(anova_tbl, caption = "ANOVA Table with extra sums of squares")
```

Table 1: ANOVA Table with extra sums of squares

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1566.5	1566.450	215.95	0
X2|X1	1	306.2	306.250	42.22	0
Residuals	13	94.3	7.254	NA	NA

Part (b):

```
# Compute F critical value
alpha <- 0.01
(F_crit <- qf(1-alpha, 1, 13))
```

```
## [1] 9.074
```

- Desired significance level, $\alpha = 0.01$.
- The alternative hypotheses for testing whether X_2 can be dropped from the regression model given that X_1 is retained are:

$$H_0 : \beta_2 = 0 \quad \text{versus} \quad H_a : \beta_2 \neq 0,$$

where β_2 denotes the true coefficient of X_2 given that X_1 is in the model.

- From the ANOVA table in part (a) the F statistic value associated with the test is $F_{calc} = 42.22$. Given that $\alpha = 0.01$, the F critical value is $F_{crit} = F(1-0.01; 1, 13) = 9.074$.
- The decision rule is to reject H_0 if $F_{calc} > F_{crit}$, otherwise fail to reject H_0.
- Since $42.22 > 9.074$, we reject H_0 and conclude H_a that X_2 is significant and that it cannot be dropped from the regression model given that X_1 is retained.
- Again, from the same ANOVA table, the corresponding p-value is **0**. This leads us to the same conclusion since $p - value = 0 < \alpha = 0.01$.

Problem 7.12

Using the **brand preference** data set, we wish to calculate $R^2_{Y1}$, $R^2_{Y2}$, $R^2_{12}$, $R^2_{Y1|2}$, $R^2_{Y2|1}$, and R^2.

```r
library(rsq)
# Some values require the new models fitted below:
mod_YX1 <- lm(Y ~ X1, data = brand_pref)
mod_YX2 <- lm(Y ~ X2, data = brand_pref)
mod_X1X2 <- lm(X1 ~ X2, data = brand_pref)

# Extract the appropriate measures
rsqY1 <- mod_YX1 %>% glance() %>% pull(r.squared)
rsqY2 <- mod_YX2 %>% glance() %>% pull(r.squared)
rsq12 <- mod_X1X2 %>% glance() %>% pull(r.squared)
overall_rsq <- brand_mdl1 %>% glance() %>% pull(r.squared)
prsq <- rsq.partial(brand_mdl1)$partial.rsq # get the partial R squared

# Display table of values
values <- data.frame(rsqY1, rsqY2, rsq12, prsq[1], prsq[2], overall_rsq)
kable(values, col.names = c("$R^2_{Y1}$", "$R^2_{Y2}$", "$R^2_{12}$", "$R^2_{Y1|2}$", "$R^2_{Y2|1}$", "$R^
  kable_paper() %>%
  kable_styling(position = "center", latex_options = "hold_position")
```

Table 2: Various Coefficients of Determination

$R^2_{Y1}$	$R^2_{Y2}$	$R^2_{12}$	$R^2_{Y1\|2}$	$R^2_{Y2\|1}$	R^2
0.7964	0.1557	0	0.9432	0.7646	0.9521

- $R^2_{Y1} = 0.7964$ measures the coefficient of simple determination between Y and X_1, which means that the variation in brand preference is reduced by 79.64% when moisture content of the product is considered as the only predictor in the linear regression model.
- $R^2_{Y2} = 0.1557$ measures the coefficient of simple determination between Y and X_2. Thus, the variation in brand preference is reduced by 15.57% when the sweetness of the product is considered as the only predictor in the linear regression model.
- $R^2_{12} = 0$ is a measure of the coefficient of simple determination between X_1 and X_2, which signifies that using sweetness to predict moisture content in a linear regression model leads to no reduction in the variation in the moisture content. In other words, there appears to be no linear relationship between sweetness and the moisture content of the product.
- $R^2_{Y1|2} = 0.9432$ measures the coefficient of partial determination between Y and X_1, given that X_2 is in the model. That is, the unique contribution of the moisture content in reducing the variation in the brand preference is approximately 94.3%, when the sweetness of the product is already included in the model.
- Similarly, $R^2_{Y2|1} = 0.7646$ measures the coefficient of partial determination between Y and X_2, given that X_1 is in the model. By this measure we can say that the sweetness of the product is able to explain additional

76.46% of the variation in the brand preference that remains given that the moisture content variable in already included in the model.

- Finally, $R^2 = 0.9521$ measures the coefficient of multiple determination between Y and the two predictors X_1 and X_2. This signifies that approximately 95.2% of the variation in brand preference can be explained by including both the moisture content and sweetness of the product in the model. This is actually a large reduction in heterogeneity.

Problem 7.16

Part (a)

```
# Transform the variables using the correlation coefficient transformation in equation 7.44 (p.273)
brand_pref <- brand_pref %>%
  mutate(Y_std = scale(Y)/(n()-1),
         X1_std = scale(X1)/(n()-1),
         X2_std = scale(X2)/(n()-1)
         )

# Fit the standardized regression model in equation 7.45 (p.273)
std_mdl <- lm(Y_std ~ 0 + X1_std + X2_std, data = brand_pref)
# summary(std_mdl)

# obtain model outputs for the estimates
std_mdl %>%
  tidy() %>%
  kable(caption = "Parameter estimates for the standardized regression model")
```

Table 3: Parameter estimates for the standardized regression model

term	estimate	std.error	statistic	p.value
X1_std	0.8924	0.0585	15.250	0
X2_std	0.3946	0.0585	6.743	0

```
# obtain model performance metrics
std_mdl %>%
  glance() %>%
  select(r.squared, adj.r.squared, sigma, F.statistic = statistic, df, df.residual, p.value) %>%
  kable(caption = "Model performance metrics for the standardized regression model")
```

Table 4: Model performance metrics for the standardized regression model

r.squared	adj.r.squared	sigma	F.statistic	df	df.residual	p.value
0.9521	0.9452	0.0151	139	2	14	0

From results in Table 3, the estimated standardized regression function is given as:

$$\hat{Y^*} = 0.8924X_1^* + 0.3946X_2^*,$$

where Y^*, X_1^* and X_2^* are the transformed brand preference, moisture content and sweetness, respectively.

Part (b)

$b_1^* = 0.8924$. This signifies that for every one standard deviation increase in the moisture content (X_1), the expected brand preference (Y) increases by **0.8924** in units of the standard deviations of Y, while the sweetness (X_2) level is held constant.

Part (c)

```r
# Using formulas in equation 7.53
betas <- as.numeric(coef(std_mdl))

brand_pref %>%
  summarise(b1 = betas[1]*(sd(Y)/sd(X1)),
            b2 = betas[2]*(sd(Y)/sd(X2)),
            b0 = (mean(Y) - b1*mean(X1) - b2*mean(X2))
            ) %>%
  kable(caption = "Transformed regression coefficients")
```

Table 5: Transformed regression coefficients

b1	b2	b0
4.425	4.375	37.65

Interestingly, these are the same regression coefficients obtained in Problem 6.5 (b) in Homework 4.

Problem 8.4

Importing the Muscle mass data set

Here, X and Y are used to denote the age and the muscle mass of a woman participant, respectively.

```r
muscle_mass <- read.table("../Data Sets/Chapter  1 Data Sets/CH01PR27.txt",
                          col.names = c("Y", "X"))
# dim(muscle_mass)
kable(head(muscle_mass), caption = "First few observations")
```

Table 6: First few observations

Y	X
106	43
106	41
97	47
113	46
96	45
119	41

Part (a)

This problem requires centering of the X predictor vector, so we create a new vector x (lower case) as follows

$$x = X - \bar{X}, \quad \text{where } \bar{X} \text{ is the sample mean of } X.$$

```r
# Create the centered predictor variable and its squared term
muscle_mass <- muscle_mass %>%
  mutate(x = X - mean(X),
         x_sq = x^2
         )

# Fit the required model
muscle_mass_mdl <- lm(Y ~ x + x_sq, data = muscle_mass)
# summary(muscle_mass_mdl)

# obtain model outputs for the estimates
muscle_mass_mdl %>%
  tidy() %>%
  kable(caption = "Parameter estimates")
```

Table 7: Parameter estimates

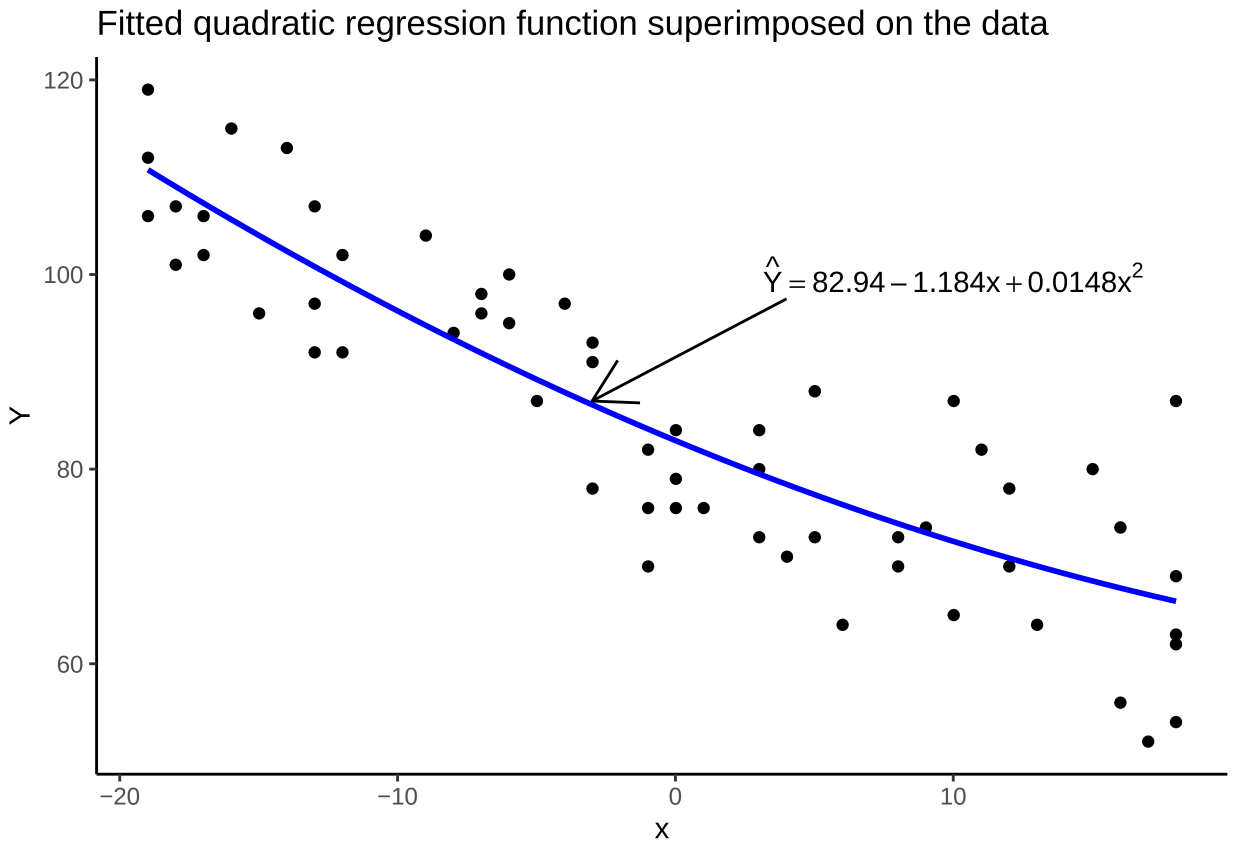
term	estimate	std.error	statistic	p.value
(Intercept)	82.9357	1.5431	53.745	0.0000
x	-1.1840	0.0886	-13.358	0.0000
x_sq	0.0148	0.0084	1.776	0.0811

```r
# obtain model performance metrics
muscle_mass_mdl %>%
  glance() %>%
  select(r.squared, adj.r.squared, sigma, F.statistic = statistic, df, df.residual, p.value) %>%
  kable(caption = "Model performance metrics")
```

Table 8: Model performance metrics

r.squared	adj.r.squared	sigma	F.statistic	df	df.residual	p.value
0.7632	0.7549	8.025	91.84	2	57	0

```r
# Generating plot
coefs <- as.numeric(coef(muscle_mass_mdl))

ggplot(muscle_mass, aes(x, Y)) +
  geom_point() +
  stat_function(fun = function(x){coefs[1] + coefs[2]*x + coefs[3]*(x^2)}, color = "blue", lwd = 1) +
  ggtitle("Fitted quadratic regression function superimposed on the data") +
  annotate("segment", x = 4, y = 97.5, xend = -3, yend = 87, arrow = arrow()) +
  annotate("text", x= 10, y = 100, parse = TRUE, size = 4,
           label = "hat(Y) == 82.9358 - 1.1840*x + 0.0148*x^2")
```



From the plot we can see that the quadratic regression function appears to provide a good fit since it lies rougly at the center of the data points.

And from Table 5, $R^2 = 0.7632$.

Part (b): Testing whether there is a regression relation

- Significance level, $\alpha = 0.05$.
- Given the second-order regression model:

  $$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \epsilon_i, i = 1, 2, \cdots, 60.$$

  We are interested in testing the following two competing hypotheses:

  $$H_0 : \beta_1 = \beta_{11} = 0 \quad \text{against} \quad H_a : \text{at least one of } \; \beta_1 \text{ and } \beta_{11} \neq 0.$$

- From the model output in part (a) Table 8, the overall F test statistic is $F^* = 91.84$ on 2 and 57 degrees of freedom, with associated p-value = 0.
- The decision rule is to reject H_0 if p-value $\leq \alpha = 0.05$.
- Hence, we reject H_0 since the p-value is less than `0.05`, and conclude H_a that there is a regression relation.
- Thus, the test implies that at least one of β_1 and β_{11} is significant. Individual tests may be needed to assess the significance of each term.

Part (c): Estimating the mean response for women aged 48 years with a 95% confidence interval

```
# First center the age of the woman
age_centered <- 48 - mean(muscle_mass$X)
```

```r
kable(predict(muscle_mass_mdl, newdata = data.frame(x= age_centered, x_sq = age_centered^2),
              interval = "confidence", level = 0.95))
```

fit	lwr	upr
99.25	96.28	102.2

The required 95% confidence interval estimate of $E\{Y_h\}$ is: $96.28 \leq E\{Y_h\} \leq 102.2$. Thus, we conclude with 95% confidence that the mean muscle mass for women aged 48 lies somewhere between `96.28` and `102.2`.

Part (d): 95% Prediction interval for a woman aged 48 years

```r
kable(predict(muscle_mass_mdl, newdata = data.frame(x= age_centered, x_sq = age_centered^2),
              interval = "prediction", level = 0.95))
```

fit	lwr	upr
99.25	82.91	115.6

The required 95% prediction interval is: $82.91 \leq Y_{h(new)} \leq 115.6$. Meaning, we predict with 95% confidence that the muscle mass for a woman aged 48 years is somewhere between `82.91` and `115.6`.

Part (e)

```r
# Using the t-test
muscle_mass_mdl %>%
  tidy() %>%
  filter(term == "x_sq") %>%
  kable(caption = "Hypothesis test results")
```

Table 11: Hypothesis test results

term	estimate	std.error	statistic	p.value
x_sq	0.0148	0.0084	1.776	0.0811

- Desired significance level, $\alpha = 0.05$.
- The alternative hypotheses for testing whether the quadratic term x^2 can be dropped from the regression model are:

$$H_0 : \beta_{11} = 0 \quad \text{versus} \quad H_a : \beta_{11} \neq 0,$$

- The decision rule is to reject H_0 if the p-value $< \alpha = 0.05$, otherwise fail to reject H_0.
- Since p-value $= 0.08 > 0.05$, we fail to reject H_0 and conclude that the quadratic term is not statistically significant and thus it can be dropped from the regression model at the 5% significance level.

Part (f): Expressing the regression function in terms of the original variable X

We use Equations 8.12 a through c from the Textbook (Kutner et al., 2004) to compute the appropriate regression coefficients.

```r
# Compute regression coefficients in terms of the original X
muscle_mass %>%
  summarise(b0 = coefs[1] - coefs[2]*mean(X) + coefs[3]*mean(X)^2,
            b1 = coefs[2] - 2*coefs[3]*mean(X),
            b11 = coefs[3]
            ) %>%
  kable(caption = "Derived regression coefficients")
```

Table 12: Derived regression coefficients

b0	b1	b11
207.3	-2.964	0.0148

It follows from the above table that the regression function in terms of the original X variable is

$$\hat{Y} = 207.3 - 2.964X + 0.0148X^2,$$

where $\hat{Y}$ is the point estimate of the mean muscle mass and X is the age of the women in years.

Part (g): Coefficients of simple correlation

```
muscle_mass %>%
  summarise(rsqa = cor(X, X^2),
            rsqb = cor(x, x_sq)) %>%
  kable(col.names = c("$r_{X, X^2}$", "$r_{x, x^2}$"))
```

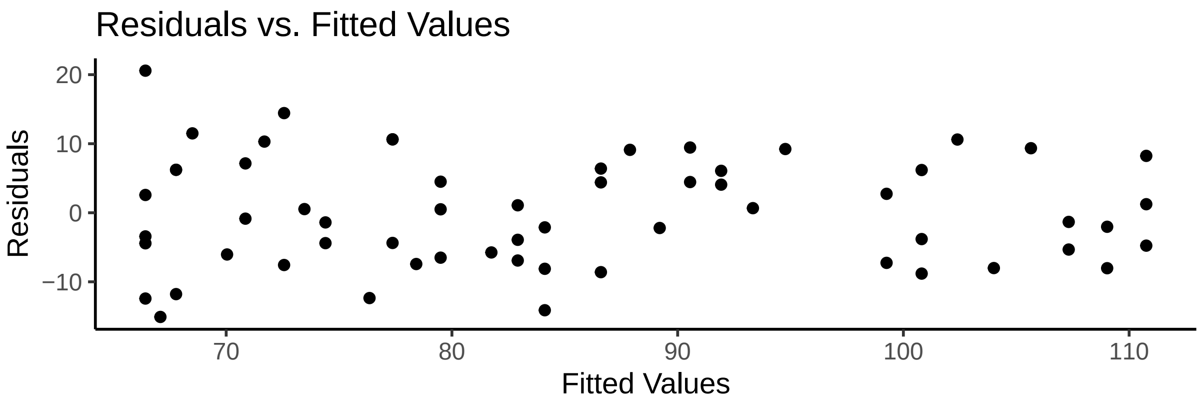
$r_{X,X^2}$	$r_{x,x^2}$
0.9961	-0.0384

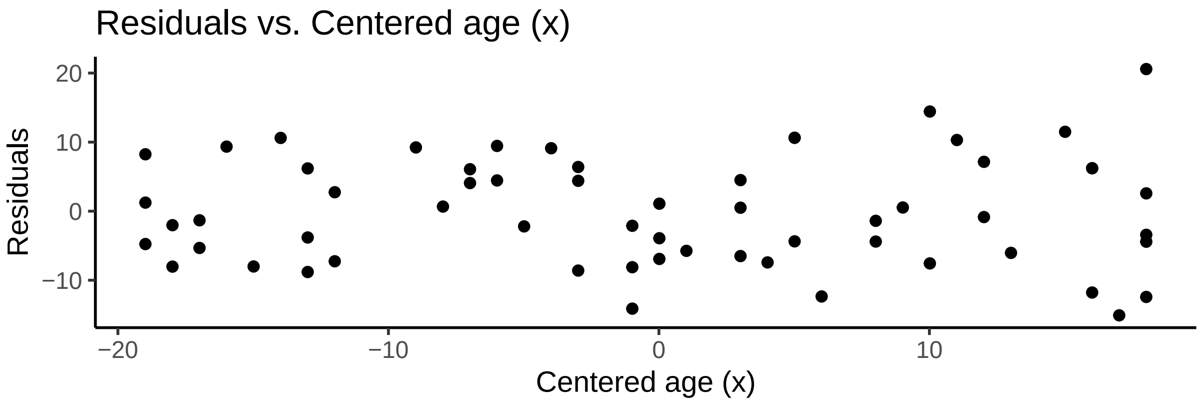
From the output, the coefficient of simple correlation between X and X^2 and between x and x^2 are `0.9961` and `-0.0384`, respectively. These results suggest that the use of a centered variable is very helpful since it leads to a substantial reduction in the multicollinearity between the original variable and its squared term. The end benefit is that the computational burden often associated with high multicollinearity in estimating the regression parameters for the polynomial regression model involving the original variable could be avoided.

Problem 8.5

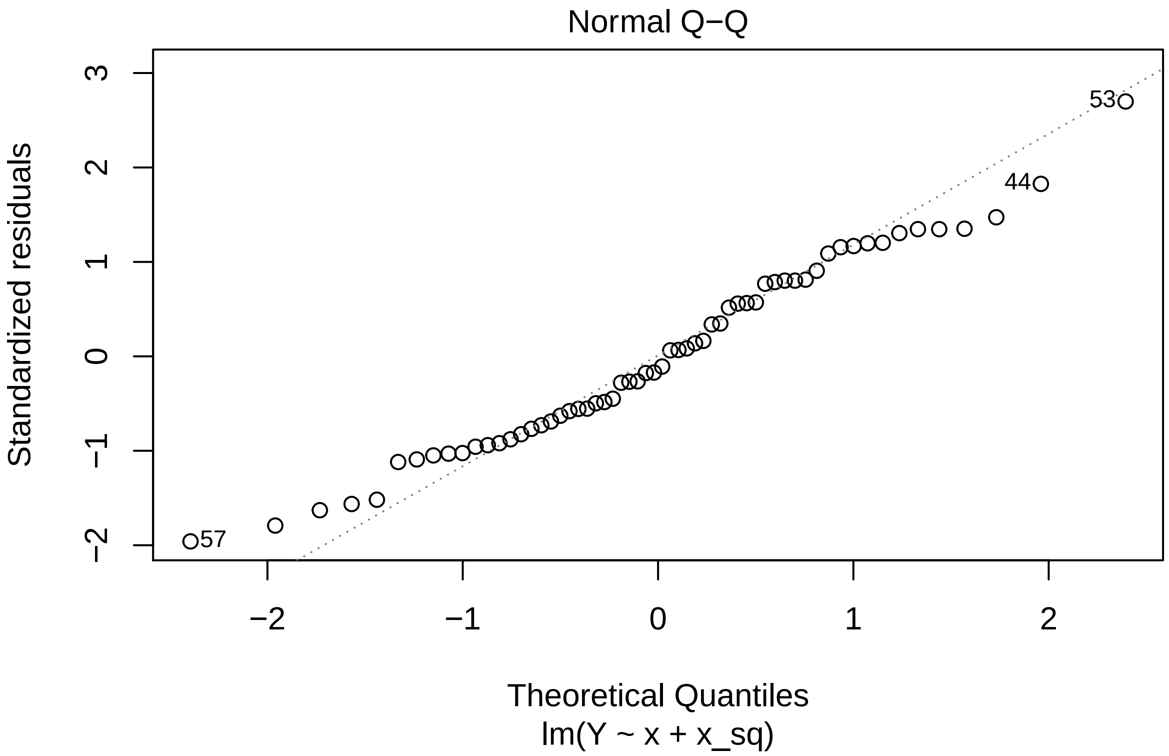
Part (a): Residual plots

```
# Attach fitted values and residuals to the muscle_mass data set for plotting
muscle_mass_aug <- data.frame(muscle_mass, fitted = fitted(muscle_mass_mdl),
                              resid = resid(muscle_mass_mdl))

# par(mfrow = c(3,2))
(ggplot(muscle_mass_aug, aes(fitted, resid)) +
  geom_point()  +
  labs(x="Fitted Values", y="Residuals", title = "Residuals vs. Fitted Values")) /

(ggplot(muscle_mass_aug, aes(x, resid)) +
  geom_point()  +
  labs(x="Centered age (x)", y="Residuals", title = "Residuals vs. Centered age (x)"))
```





```r
# Normal probability plot
plot(muscle_mass_mdl, 2)
```



The residuals versus fitted plot suggests that the nonconstancy of the error variance is not substantial and that there does not appear to be a serious systematic deviations from the response plane. Similarly, the residuals against x plot looks about consistent with the conclusions of good fit by the response plane and somewhat constant variance of the error terms. From the normal probability plot, though the bulk of the points lie on the straight line, a some amount of them depart at the tails, suggesting that violations of the error terms distribution from normality may not be that serious. However, observations related to the points at the extreme tails would probably have to be investigated.

Part (b): Lack of fit test of the quadratic regression function

```r
lack_of_fit_mdl <- lm(Y ~ factor(x) * factor(x_sq), data=muscle_mass)

# Lack of fit test
kable(anova(muscle_mass_mdl, lack_of_fit_mdl), caption = "Lack of fit hypothesis test results")
```

Table 14: Lack of fit hypothesis test results

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
57	3671	NA	NA	NA	NA
28	1850	29	1822	0.9509	0.5539

- Desired significance level, $\alpha = 0.05$.
- The following two alternative hypotheses are of interest:

$$H_0 : \text{the regression function is quadratic (there is no lack of fit)}$$

$$H_a : \text{the regression function is not quadratic (there is lack of fit)}$$

- The decision rule is to reject H_0 if p-value of the test $\leq \alpha = 0.05$.
- The F test-statistic turns out to be `0.9509` with corresponding p-value `0.55`. Since this p-value is way greater than the 5% significance level, we fail to reject H_0 and conclude that the quadratic regression function seems adequate. That is, there is no evidence of a lack of fit.

The implicit assumptions made in this test include the assumptions that the observations Y (muscle mass) for given x (centered age) are independent and normally distributed, and that the distributions of Y have the same variance.

Part (c): A third-order regression model

```r
# Fit the required model
muscle_mass_mdl2 <- lm(Y ~ x + x_sq + I(x^3), data = muscle_mass)
# summary(muscle_mass_mdl)

# obtain model outputs for the estimates
muscle_mass_mdl2 %>%
  tidy() %>%
  kable(caption = "Parameter estimates")
```

Table 15: Parameter estimates

term	estimate	std.error	statistic	p.value
(Intercept)	82.9273	1.5552	53.3217	0.0000
x	-1.2679	0.2489	-5.0935	0.0000
x_sq	0.0150	0.0084	1.7821	0.0802
I(x^3)	0.0003	0.0009	0.3612	0.7193

```r
# obtain model performance metrics
muscle_mass_mdl2 %>%
  glance() %>%
  select(r.squared, adj.r.squared, sigma, F.statistic = statistic, df, df.residual, p.value) %>%
  kable(caption = "Model performance metrics")
```

Table 16: Model performance metrics

r.squared	adj.r.squared	sigma	F.statistic	df	df.residual	p.value
0.7637	0.7511	8.087	60.34	3	56	0

From the above results, the estimated quadratic regression function is given as:

$$E\{Y\} = \hat{Y} = 82.9273 - 1.2679x + 0.0150x^2 + 0.0003x^3,$$

where Y is the muscle mass and x is the centered age.

Next, we are required to test whether or not $\beta_{111} = 0$ (the true coefficient associated with x^3) at $\alpha = 0.05$. This gives rise to the following competing hypotheses:

$$H_0 : \beta_{111} = 0 \quad \text{versus} \quad H_a : \beta_{111} \neq 0.$$

From Table 15 the t-test statistic value is `0.3612` with corresponding p-value `0.7193`. Since p-value $= 0.7193 > \alpha = 0.05$, we fail to reject H_0 and conclude that $\beta_{111} = 0$, which suggests that the cubic term (x^3) is not statistically significant. And yes, I find this conclusion to be consistent with my finding in part (b) since I concluded in part(b) that a quadratic regression function appear to provide a better fit.

Reference

- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Wasserman, W. (2004). Applied linear regression models (Vol. 4). New York: McGraw-Hill/Irwin..