ALSM: Chapter 3

Diagnostics and Remedial Measures

Darshan Patel

6/12/2020

```
library(tidyverse)
library(latex2exp)
library(gridExtra)
library(wesanderson)
library(broom)
theme_set(theme_minimal())
```

Problem 1:

Distinguish between (1) residual and semistudentized residual, (2) $\mathrm{E}[\varepsilon_i] = 0$ and $\overline{e} = 0$ and (3) error term and residual.

Answer: The semistudentized residual is the residual divided by an approximation of the standard deviation of the residual itself. $\overline{e} = 0$ is the declaration that the mean of the residuals calculated from the simple linear regression model is zero while $\mathrm{E}[\varepsilon_i] = 0$ is the declaration that the true errors have an expected value of 0. The difference between error term and residual is that the residual is meant to be the observed error between the observed value and the fitted value while the error term is is meant to be the true error error in the regression model.

Problem 2:

Prepare a prototype residual plot for each of the following cases: (1) error variance decreases with X; (2) true regression function is $\bigcup$ shaped, but a linear regression function is fitted.

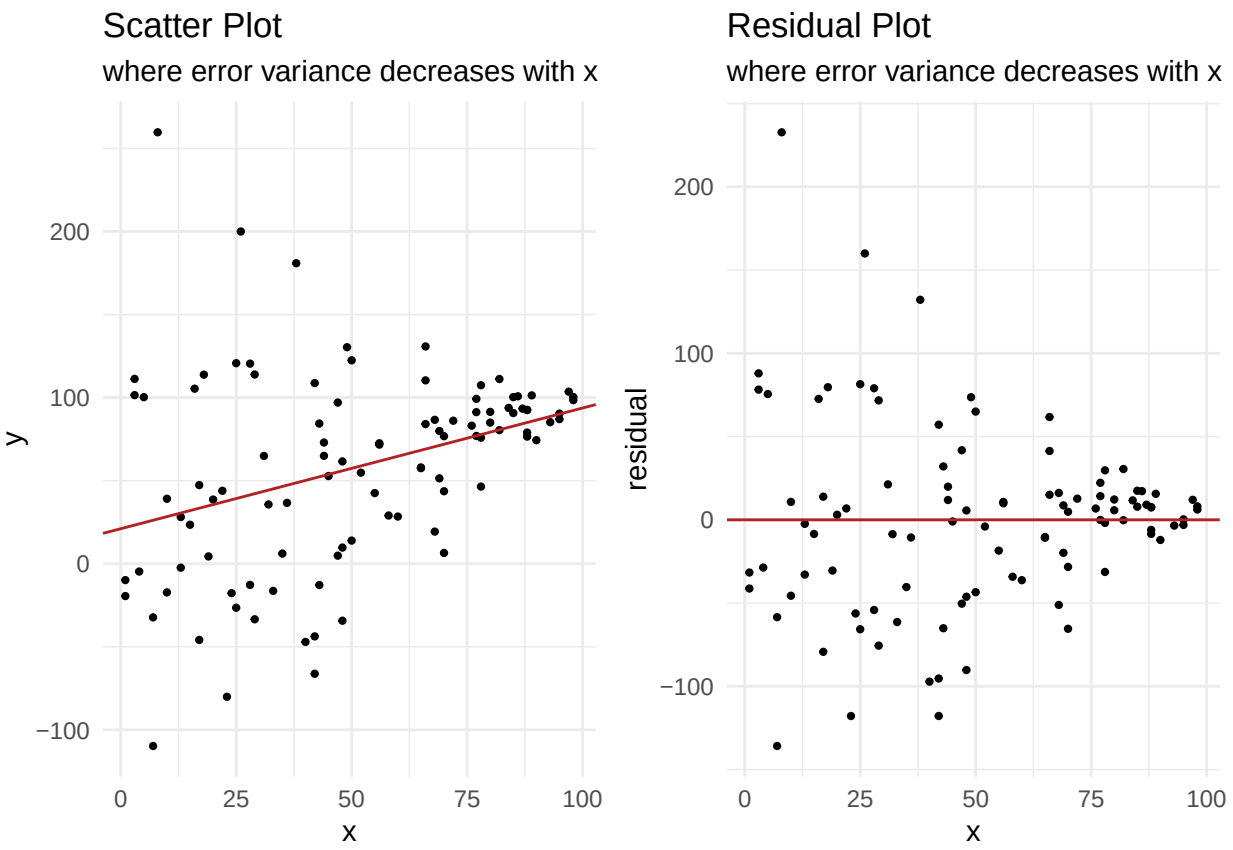
Answer:

```
lm.fit_manual = function(X, Y){
  b1 = sum((X - mean(X))*(Y - mean(Y))) / (sum((X - mean(X))^2))
  b0 = mean(Y) - b1*mean(X)
  return(c(b0, b1))
}
```

In this scenario, the error variance decreases with X:

```
set.seed(2020)
x = sample(1:100, 100, replace = TRUE)
error = rnorm(mean = x, sd = 100-x, n = 100)
y = rexp(x) + error
model = lm.fit_manual(x, y)
pred = model[1] + model[2]*x
resid = y - pred
df = data.frame(x, y, pred, resid)
```

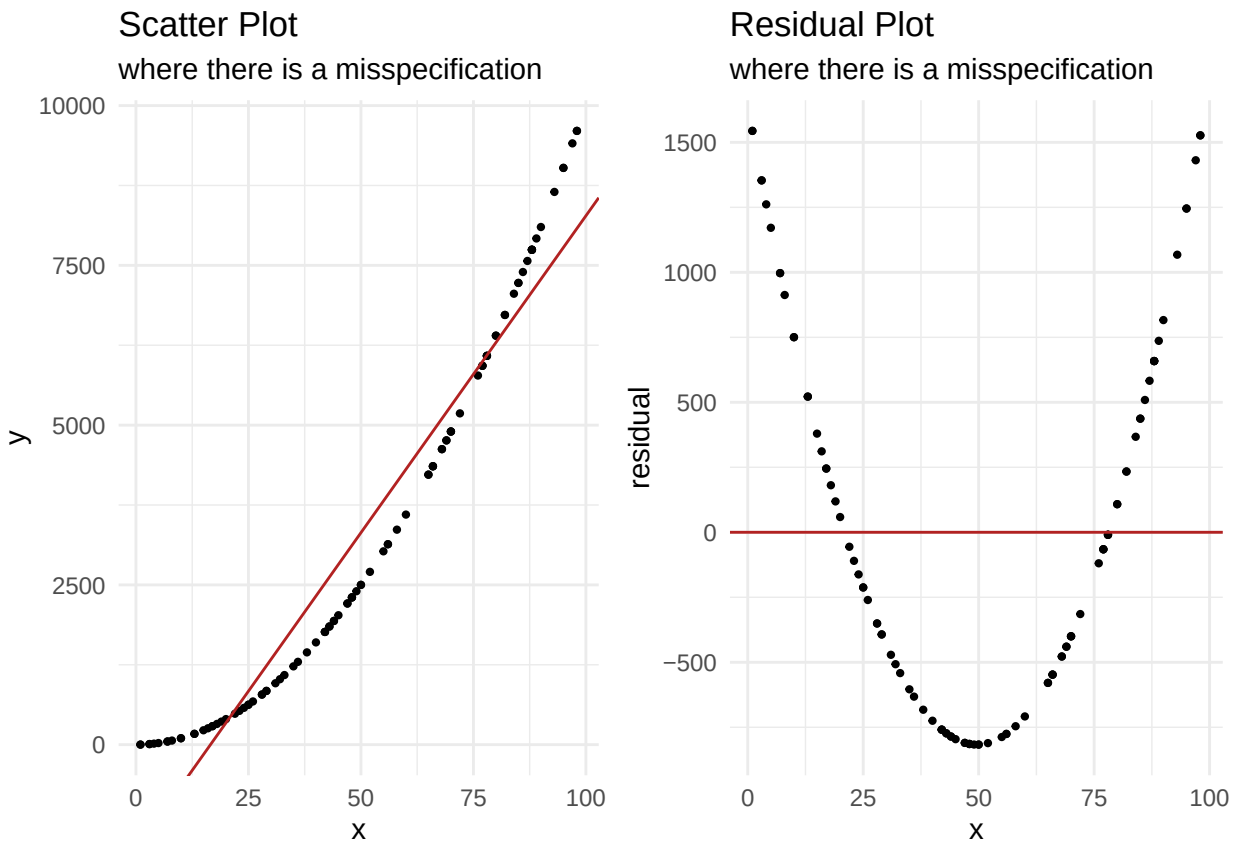
```
plot1 = df %>% ggplot(aes(x=x, y=y)) +
  geom_point(size = .8) +
  geom_abline(intercept = model[1], slope = model[2], color = "firebrick") +
  labs(x = "x", y = "y", title = "Scatter Plot",
       subtitle = "where error variance decreases with x")
plot2 = df %>% ggplot(aes(x=x, y=resid)) +
  geom_point(size = .8) +
  geom_hline(yintercept = 0, color = "firebrick") +
  labs(x = "x", y = "residual",
       title = "Residual Plot",
       subtitle = "where error variance decreases with x")

grid.arrange(plot1, plot2, ncol = 2)
```



In this scenario, the true regression function is ⋃ shaped but a linear regression function is fitted.

```
y = x^2
model = lm.fit_manual(x, y)
pred = model[1] + model[2]*x
resid = y - pred
df = data.frame(x, y, pred, resid)

plot1 = df %>% ggplot(aes(x=x, y=y)) +
  geom_point(size = .8) +
  geom_abline(intercept = model[1], slope = model[2], color = "firebrick") +
  labs(x = "x", y = "y", title = "Scatter Plot",
       subtitle = "where there is a misspecification")
```

```
plot2 = df %>% ggplot(aes(x=x, y=resid)) +
  geom_point(size = .8) +
  geom_hline(yintercept = 0, color = "firebrick") +
  labs(x = "x", y = "residual",
       title = "Residual Plot",
       subtitle = "where there is a misspecification")

grid.arrange(plot1, plot2, ncol = 2)
```



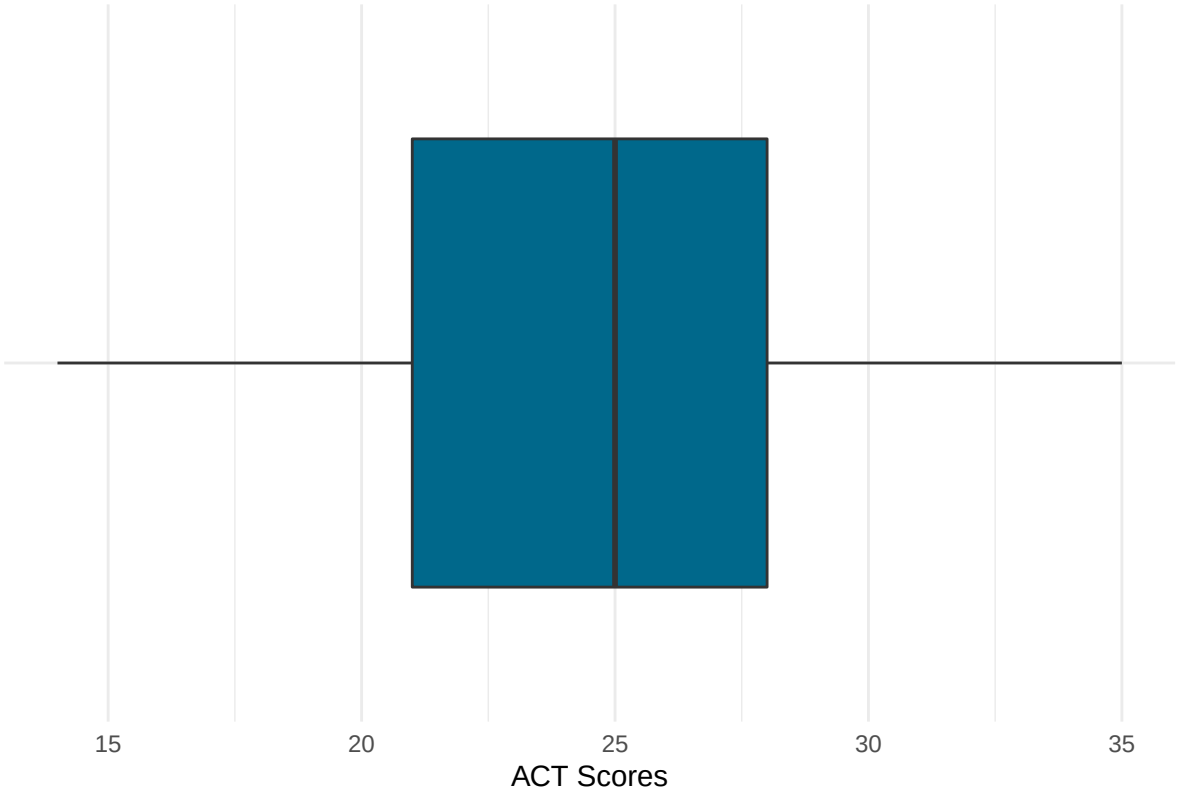
Problem 3:

Refer to Grade point average Problem 1.19.

(a) Prepare a box plot for the ACT scores X_i. Are there any noteworthy features in this plot?

Answer: The box plot is plotted below.

```
gpa = read.csv('CH01PR19.txt', sep = '', header = FALSE,
               col.names = c('y', 'x'),
               colClasses = c('numeric', 'numeric'))

gpa %>%
  ggplot(aes(x = '', y = x)) +
  geom_boxplot(fill = "deepskyblue4") +
  labs(x = '', y = 'ACT Scores', title = "Box Plot of ACT Scores") +
  coord_flip()
```

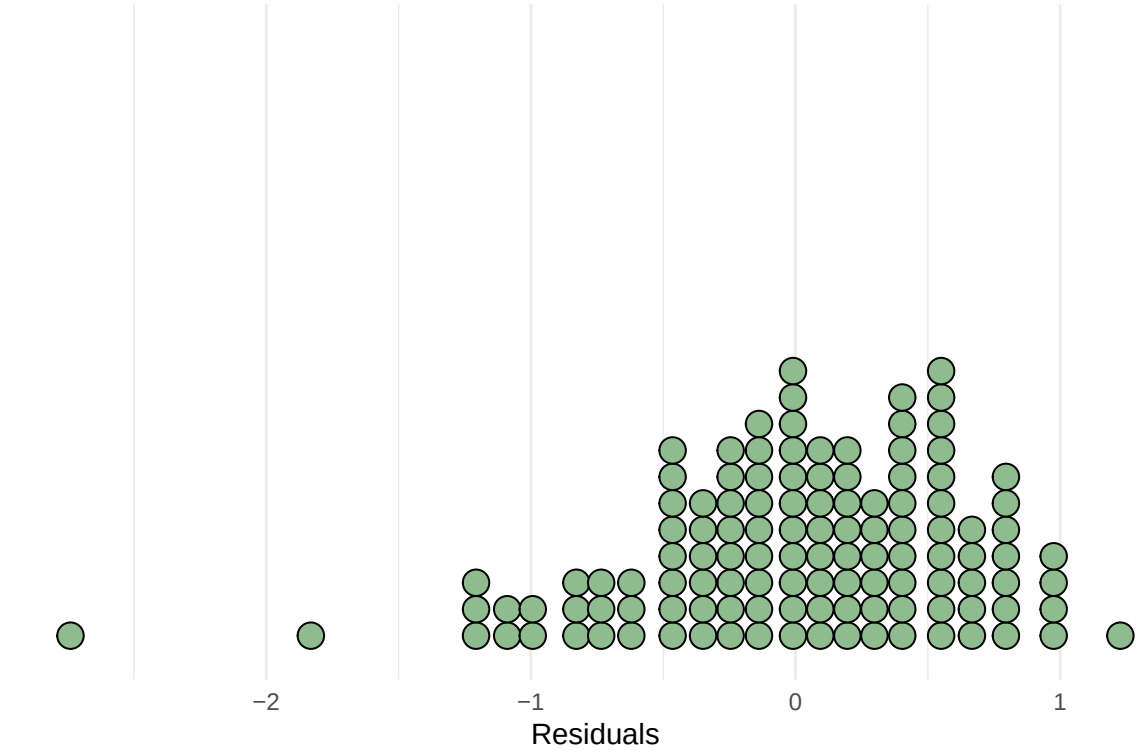


The median ACT score is 25 while scores between 21 and 28 lie between the first and third quartiles. There are no outliers.

(b) Prepare a dot plot of the residuals. What information does this plot provide?

Answer: The dot plot of the residuals is plotted below.

```
model = lm.fit_manual(gpa$x, gpa$y)
pred = model[1] + model[2]*gpa$x
resid = gpa$y - pred
gpa_resids = data.frame(x = gpa$x, y = gpa$y, pred, resid)

gpa_resids %>% ggplot(aes(x = resid)) +
  geom_dotplot(binwidth = .1, fill = "darkseagreen", color = "black") +
  scale_y_continuous(breaks = NULL, name = '') +
  labs(x = "Residuals",
       title = "Dot Plot of the Residuals", subtitle = "of Regressing ACT Scores on GPA")
```

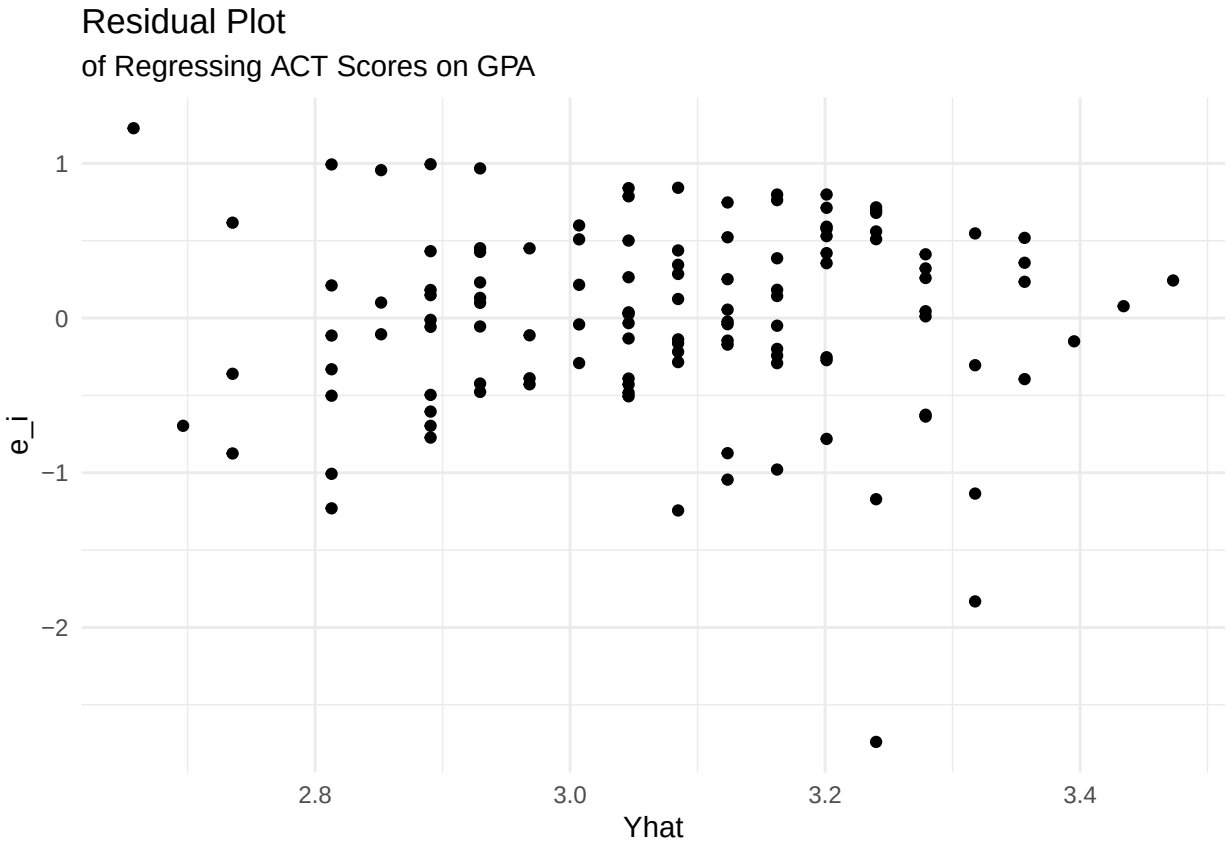


The residuals are skewed to the left and centered around 0. There are two large residuals in the negatives.

(c) Plot the residual e_i against the fitted values $\hat{Y}_i$. What departures from regression model (2.1) can be studied from this plot? What are your findings?

Answer: The residuals are plotted against the fitted values below.

```
gpa_resids %>% ggplot(aes(x = pred, y = resid)) +
  geom_point() +
  labs(x = "Yhat", y = "e_i",
       title = "Residual Plot", subtitle = "of Regressing ACT Scores on GPA")
```



The residuals are distributed randomly. A linear relationship between GPA and ACT scores is probable since there is no pattern in the distribution of the residuals. There appears to be a constant variance.

(d) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumption here using Table B.6 and $\alpha = 0.05$. What do you conclude?

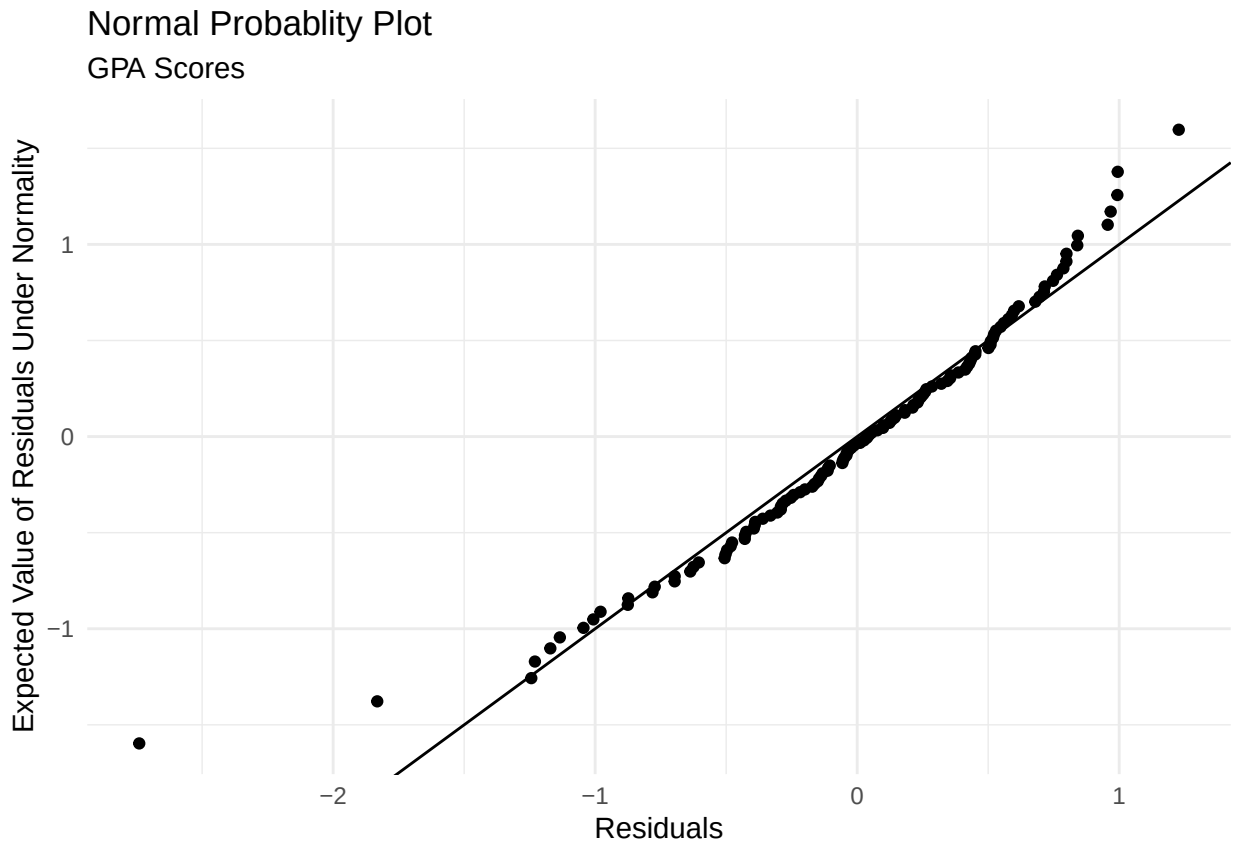
Answer: The normal probability plot is plotted below.

```
prob_plot_table = function(df){
  model = lm.fit_manual(df$x, df$y)
  pred = model[1] + model[2]*df$x
  resid = df$y - pred
  n = nrow(df)
  resids_df = data.frame(x = df$x, y = df$y, pred, resid)

  mse_sq_root = sqrt(sum((resids_df$y - resids_df$pred)^2) / (n - 2))

  resids_df = resids_df %>% mutate("Run" = 1:nrow(resids_df),
                                   "k" = rank(resid)) %>%
    mutate("exp_val" = mse_sq_root * qnorm((k - .375)/(n + .25)))
  return(resids_df)
}

prob_plot = function(df, subtitle){
  prob_plot_table(df) %>%
    ggplot(aes(x = resid, y = exp_val)) +
    geom_point() +
    geom_abline(intercept = 0, slope = 1) +
```

```
    labs(x = "Residuals", y = "Expected Value of Residuals Under Normality",
         title = "Normal Probablity Plot",
         subtitle = subtitle)
}
```

The normal probability plot of the residuals in estimating GPA using ACT scores is shown below.

```
prob_plot(gpa, "GPA Scores")
```



The coefficent of correlation between the ordered residuals and their expected values under normality is

```
prob_plot_table(gpa) %>% select(resid, exp_val) %>% cor()
```

```
##             resid   exp_val
## resid   1.0000000 0.9737275
## exp_val 0.9737275 1.0000000
```

There is a high correlation between the ordered and expected value of the residuals. Using Table B.6, it is found that the critical value for the coefficient of correlation between ordered residuals and expected values under normality, at $\alpha = .05$ and $n = 120$, is .9889. Since the calculated correlation coefficient is less than the critical value, it can be stated that there is no evidence that that the error terms are reasonably normally distributed. This was also evident in the plot above where the points do not line up.

(e) Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into two groups, $X < 26$, $X \geq 26$, and use $\alpha = 0.01$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (c)?

Answer:

```r
brown.forsythe = function(df, split, alpha){
  df = prob_plot_table(df)
  n = nrow(df)

  group1 = df[df$x < split,]
  n1 = nrow(group1)
  e1_median = median(group1$resid)
  d1 = abs(group1$resid - e1_median)
  d1_bar = mean(d1)

  group2 = df[df$x >= split,]
  n2 = nrow(group2)
  e2_median = median(group2$resi)
  d2 = abs(group2$resid - e2_median)
  d2_bar = mean(d2)

  s = sqrt((sum((d1 - d1_bar)^2) + sum((d2 - d2_bar)^2))/(n-2))

  t_ast = abs(round((d1_bar - d2_bar) / (s  * sqrt((1/n1) + (1/n2))), 3))

  if(t_ast <  qt(1 - (alpha/2), n-2)){
    paste("At the alpha level of", alpha, "the test statistic is", t_ast, "and the null hypothesis is fa
  }
  else{
    paste("At the alpha level of", alpha, "the test statistic is", t_ast, "and the null hypothesis is re
  }
}
```

Let the null hypothesis be that the error variance is constant and the alternative hypothesis that the error variance is not constant. Then

```r
brown.forsythe(gpa, 26, 0.01)
```

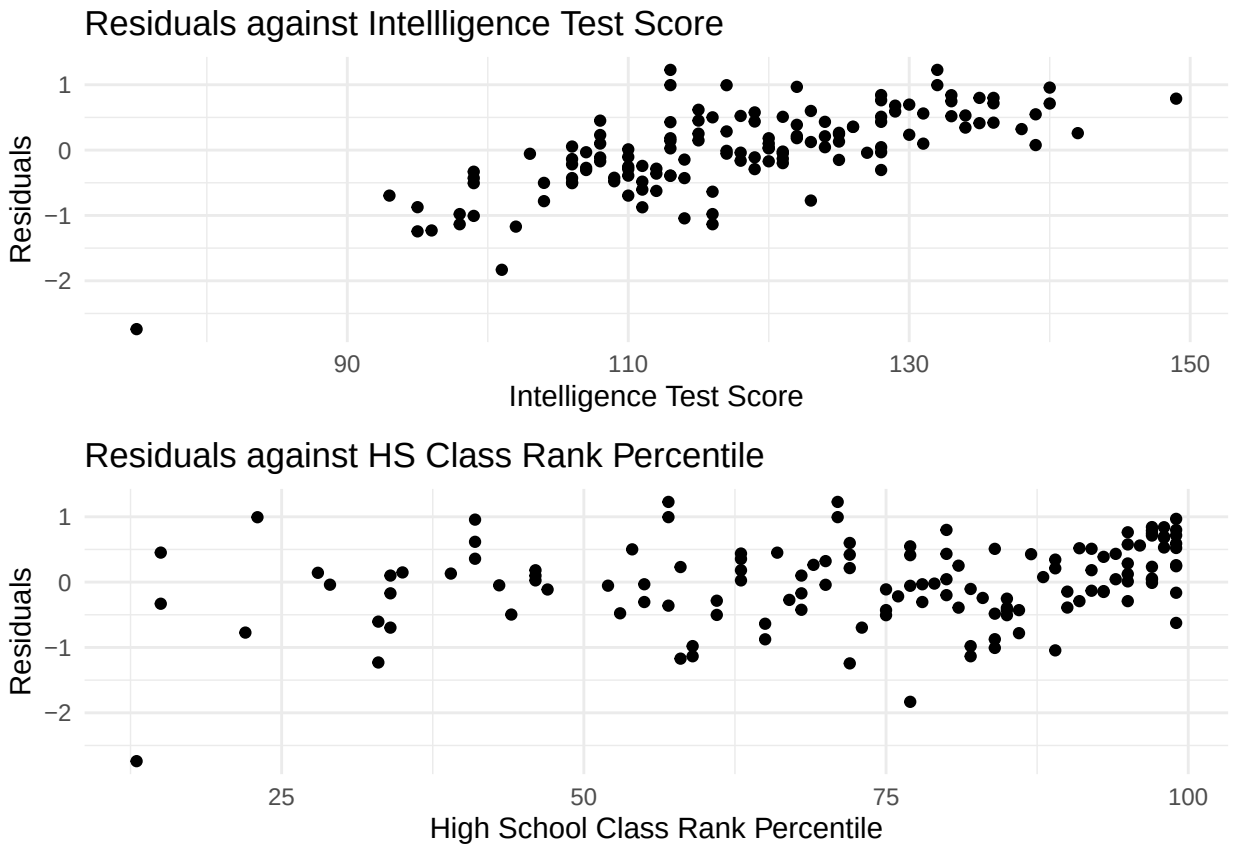
```
## [1] "At the alpha level of 0.01 the test statistic is 0.897 and the null hypothesis is failed to be r
```

(f) Information is given below for each student on two variables not included in the model, namely, intelligence test score (X_2) and high school class rank percentile (X_3). (Note that larger class rank percentiles indicate higher standing in the class, e.g., 1% is near the bottom of the class and 99% is near the top of the class.) Plot the residuals against X_2 and X_3 on separate graphs to ascertain whether the model can be improved by including either of these variables. What do you conclude?

Answer:

```r
gpa_fulldata = read.csv("CH03PR03.txt", sep = '', header = FALSE,
                        col.names = c("y", "x1", "x2", "x3"),
                        colClasses = rep("numeric", 4))
gpa_fulldata_resids = merge(prob_plot_table(gpa), gpa_fulldata)

plot_1 = gpa_fulldata_resids %>% ggplot(aes(x = x2, y = resid)) + geom_point() +
  labs(x = "Intelligence Test Score", y = "Residuals",
       title = "Residuals against Intellligence Test Score")

plot_2 = gpa_fulldata_resids %>% ggplot(aes(x = x3, y = resid)) + geom_point() +
  labs(x = "High School Class Rank Percentile", y = "Residuals",
       title = "Residuals against HS Class Rank Percentile")
```

```r
grid.arrange(plot_1, plot_2, nrow = 2)
```



The residuals increase as intelligence test scores increase and stay relatively constant with increasing high school class rank percentile.

Problem 4:

Refer to Copier maintenance Problem 1.20.

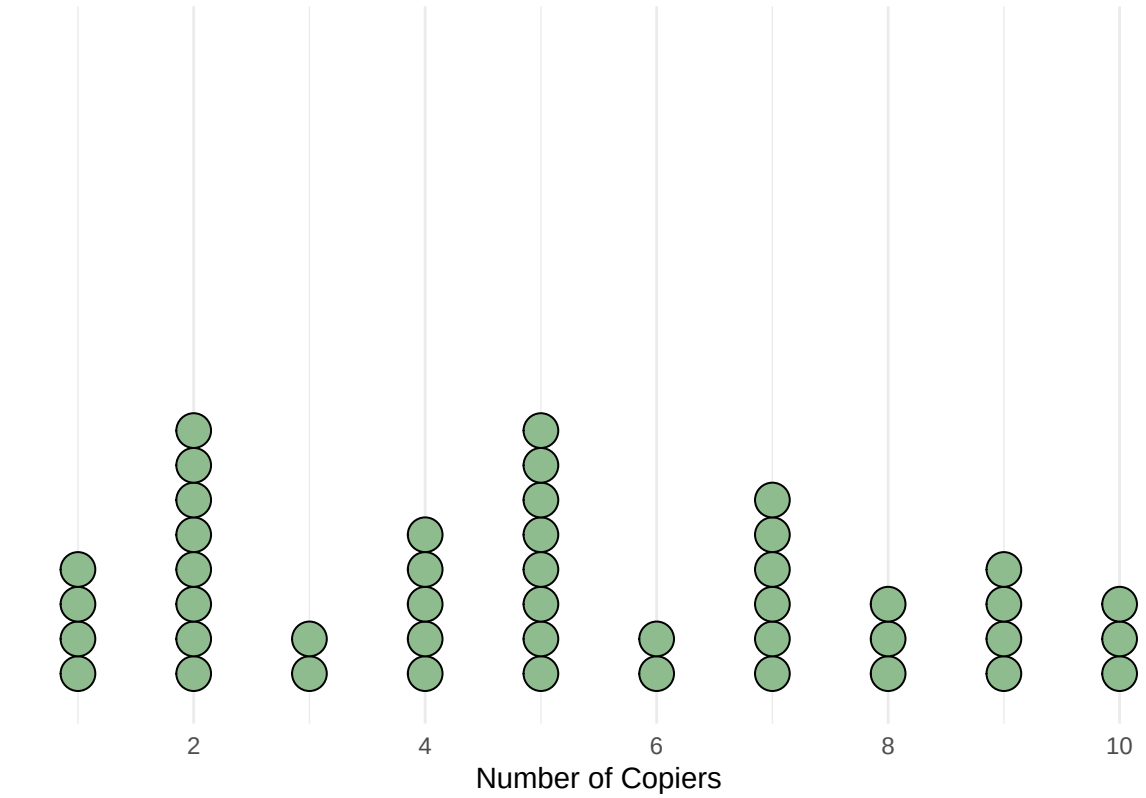
(a) Prepare a dot plot for the number of copiers serviced X_i. What information is provided by this plot? Are there any outlying cases with respect to this variable?

Answer:

The dot plot for the number of copiers serviced is shown below.

```r
copier = read.csv('CH01PR20.txt', sep = '', header = FALSE,
                  col.names = c('y', 'x'),
                  colClasses = c('numeric', 'numeric'))

copier %>%
  ggplot(aes(x = x)) + geom_dotplot(fill = "darkseagreen", color = "black") +
  scale_x_continuous(breaks = seq(0, 10, by = 2)) +
  scale_y_continuous(breaks = NULL, name = '') +
  labs(x = "Number of Copiers", y = '', title = "Dot Plot of the Number of Copiers Serviced")
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



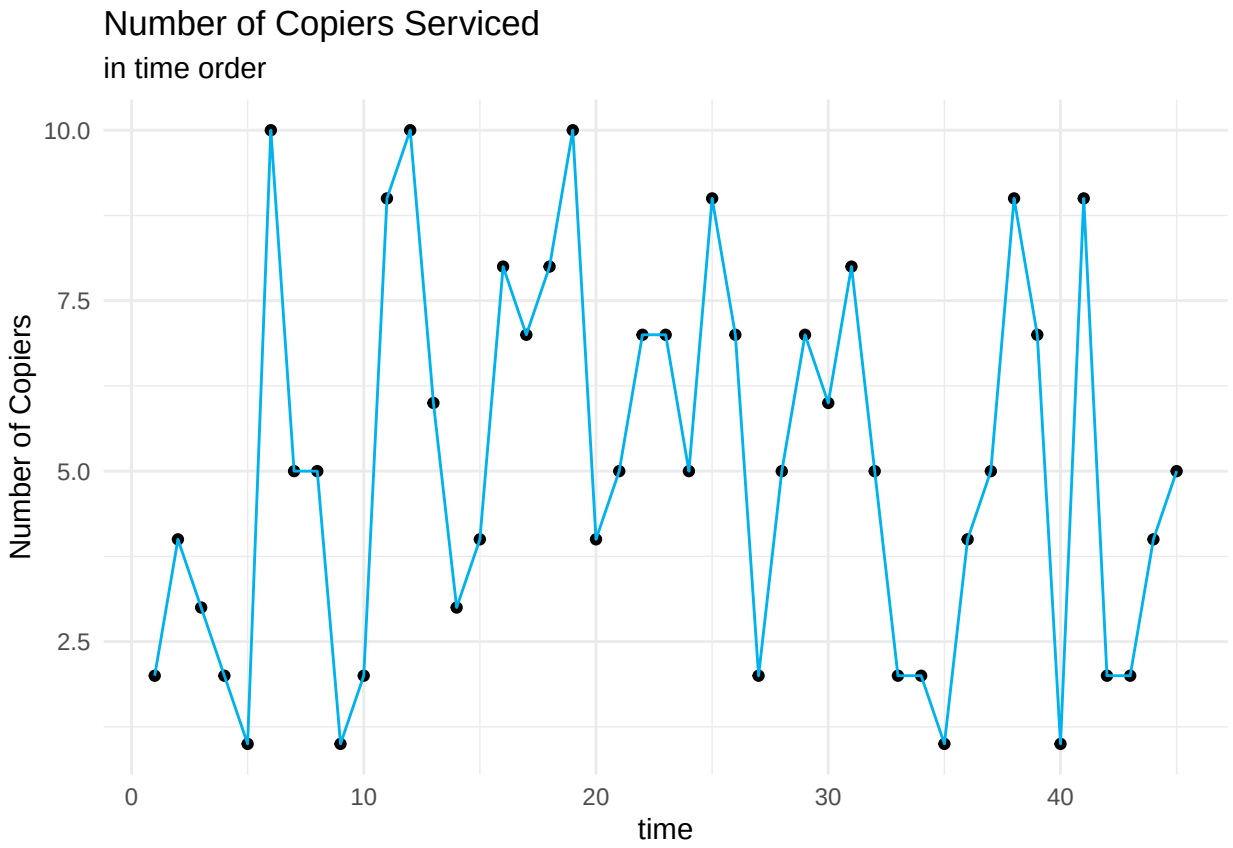
The number of copiers serviced appears to be distributed slightly right skewed. There are no outlying cases with respect to this variable.

(b) The cases are given in time order. Prepare a time plot for the number of copiers serviced. What does your plot show?

Answer:

The time plot for the number of copiers serviced is shown below.

```
copier %>% mutate(time = 1:nrow(.)) %>%
  ggplot(aes(x=time, y = x)) +
  geom_point() +
  geom_line(color = "deepskyblue2") +
  labs(x = "time", y = "Number of Copiers",
       title = "Number of Copiers Serviced", subtitle = "in time order")
```



As time increases, there does not appear to be any overall trend in the number of copiers being serviced. It is noted however that number of copiers serviced increases and decreases roughly every 3/4 measurement of time.

(c) Prepare a stem-and-leaf plot of the residuals. Are there any noteworthy features in this plot?

Answer:

The stem-and-leaf plot of the residuals is shown below.

```
model = lm.fit_manual(copier$x, copier$y)
pred = model[1] + model[2]*copier$x
resid = copier$y - pred
df = data.frame(copier$x, copier$y, pred, resid)

stem(df$resid)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   -2 | 30
##   -1 |
##   -1 | 3110
##   -0 | 99997
##   -0 | 44333222111
##    0 | 001123334
##    0 | 5666779
##    1 | 112234
##    1 | 5
```

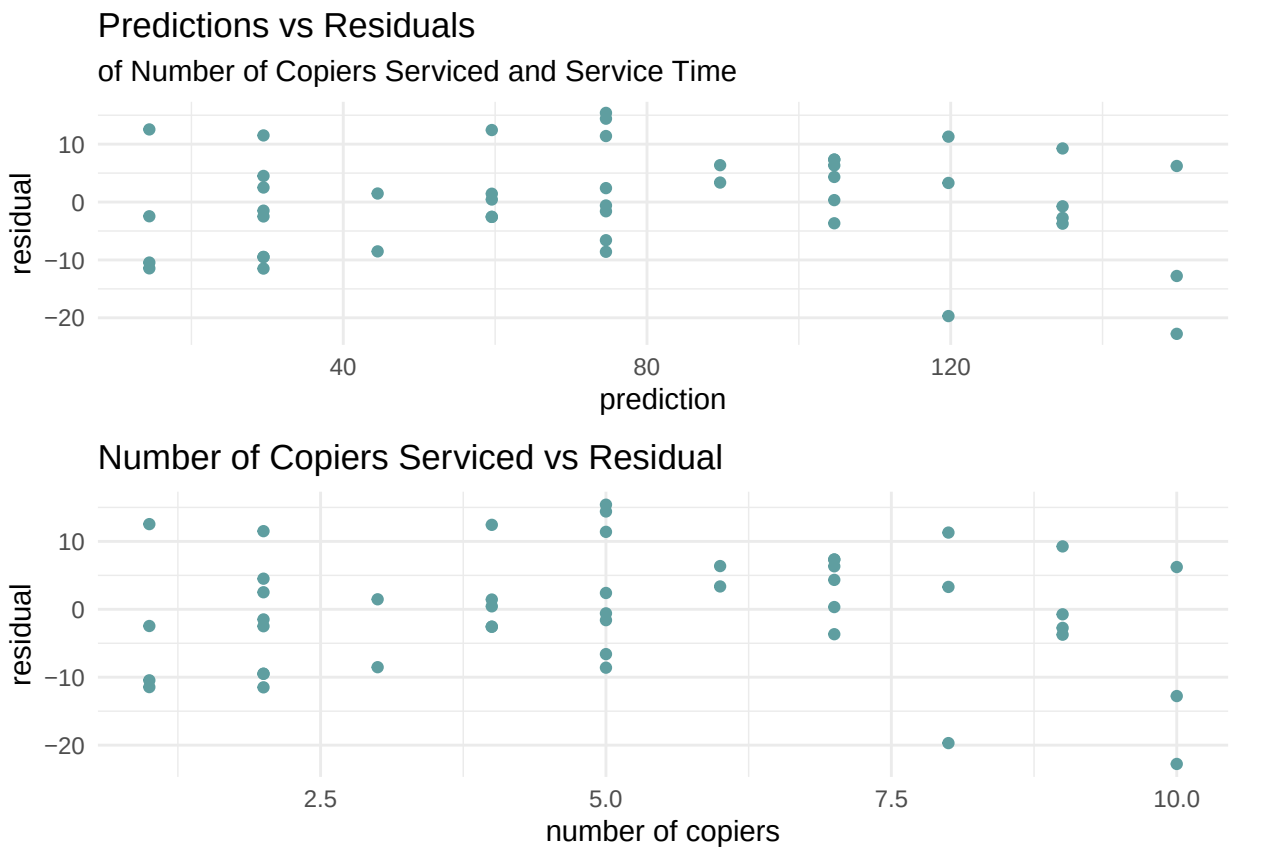
The residuals appear to hover around 0 and are normally distributed.

(d) Prepare residual plots of e_i versus $\hat{Y}_i$ and e_i versus X_i on separate graphs. Do these plots provide the same information? What departures from regression model (2.1) can be studied from these plots? State your findings.

Answer:

The residual plots of e_i versus $\hat{Y}_i$ and e_i versus X_i is shown below.

```
plot_1 = df %>% ggplot(aes(x = pred, y = resid)) +
  geom_point(color = "cadetblue") +
  labs(x = "prediction", y = "residual",
       title = "Predictions vs Residuals",
       subtitle = "of Number of Copiers Serviced and Service Time")

plot_2 = df %>% ggplot(aes(x = copier.x, y = resid)) +
  geom_point(color = "cadetblue") +
  labs(x = "number of copiers", y = "residual",
       title = "Number of Copiers Serviced vs Residual")

grid.arrange(plot_1, plot_2, nrow = 2)
```



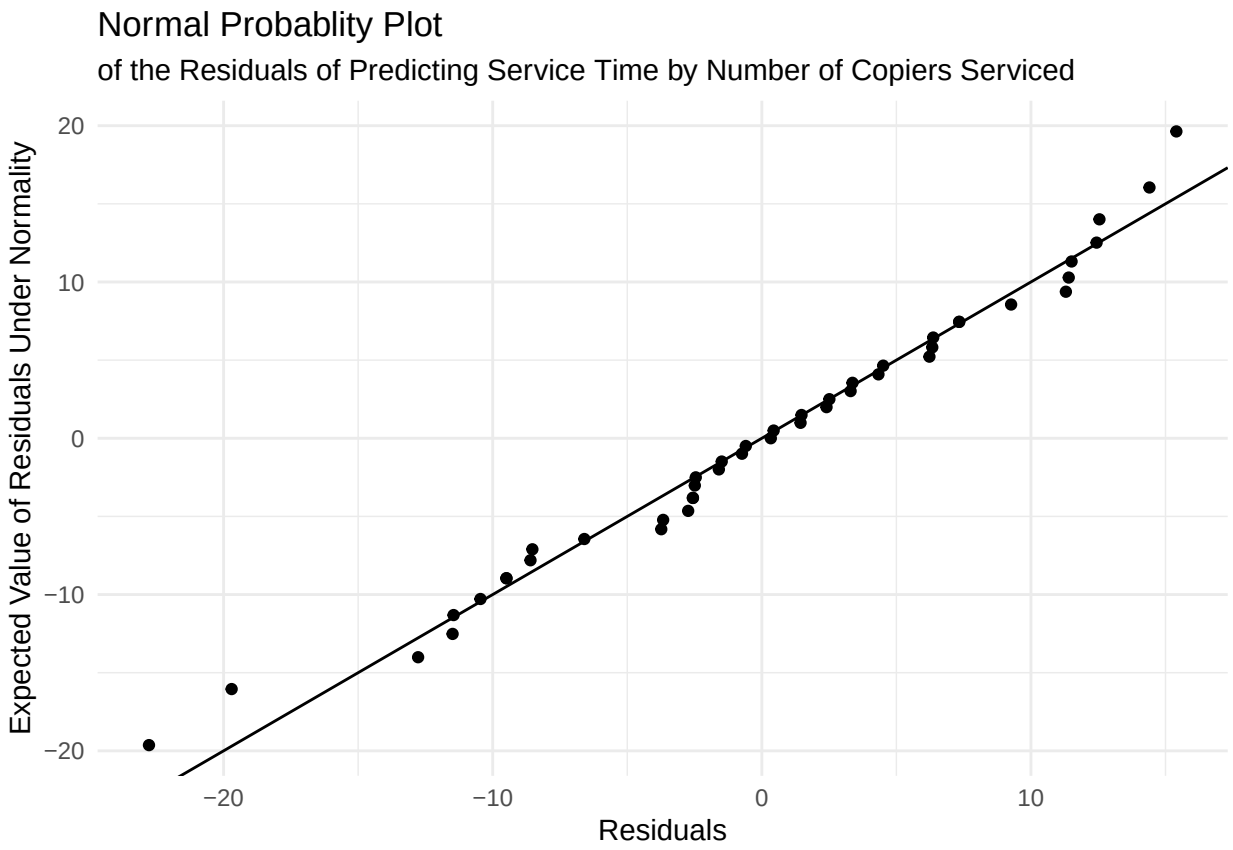
These two plots provide the same information. The nonlinearity of the regression function can be studied from these plots. It is found that there is a reasonably random distribution of points, indicating that there is a linear relationship between the number of copiers serviced and service time.

(e) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Does the normality assumption appear to be tenable here? Use Table B.6 and $\alpha = .10$.

Answer:

The normal probability plot of the residuals is shown below.

```
prob_plot(copier, "of the Residuals of Predicting Service Time by Number of Copiers Serviced")
```



The coefficent of correlation between the ordered residuals and their expected values under normality is

```
prob_plot_table(copier) %>% select(resid, exp_val) %>% cor()
```

```
##             resid   exp_val
## resid   1.0000000 0.9892079
## exp_val 0.9892079 1.0000000
```

There is a high correlation between the ordered and expected value of the residuals. Using Table B.6, it is found that the critical value for the coefficient of correlation between ordered residuals and expected values under normality, at $\alpha = .10$ and $n = 45$, is .981. Since the calculated correlation coefficient is greater than the critical value, it can be stated that there is evidence that that the error terms are reasonably normally distributed. This was also evident in the plot above where the points appeared to line up.

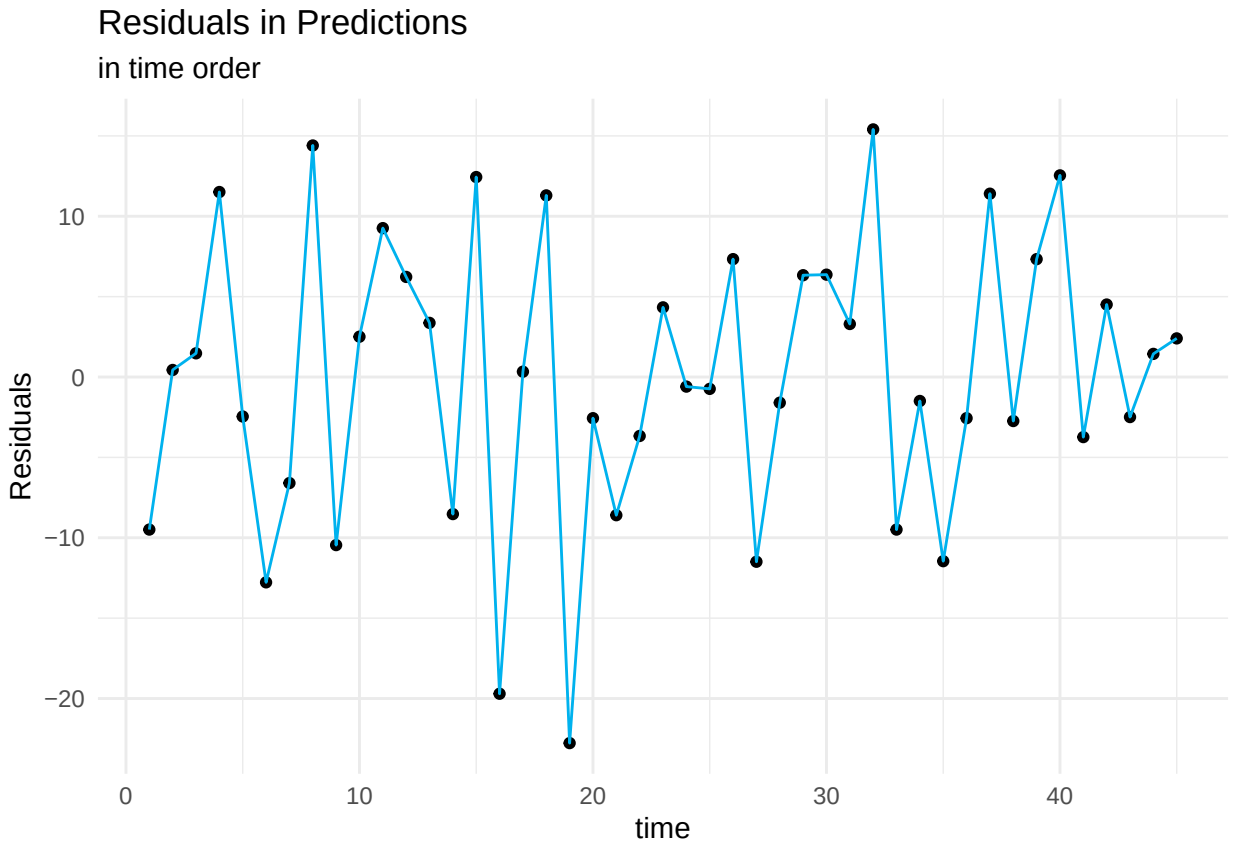
(f) Prepare a time plot of the residuals to ascertain whether the error terms are correlated over time. What is your conclusion?

Answer:

The time plot of the residuals is shown below.

```
df %>% mutate(time = 1:nrow(.)) %>%
  ggplot(aes(x = time, y = resid)) +
  geom_point() +
  geom_line(color = "deepskyblue2") +
```

```
  labs(x = "time", y = "Residuals",
       title = "Residuals in Predictions", subtitle = "in time order")
```



There does not appear to be any correlation in error terms over time.

(g) Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $\alpha = .05$. State the alternatives, decision rule and conclusion.

Answer:

```
breusch.pagan = function(df, alpha){

  model = lm.fit_manual(df$x, df$y)
  pred = model[1] + model[2]*df$x
  resid = df$y - pred
  df_new = data.frame(x = df$x, y = df$y, pred = pred, resid = resid, resid_sq = resid^2)
  sse = sum(resid^2)

  model_sq_error = lm.fit_manual(df_new$x, df_new$resid_sq)
  pred_sq_error = model_sq_error[1] + model_sq_error[2]*df_new$x
  ssr_ast = sum((pred_sq_error - mean(df_new$resid_sq))^2)

  chi_sq_BP = round((ssr_ast/2) / ((sse/nrow(df))^2), 3)

  if(chi_sq_BP < qchisq(1 - alpha, 1)){
    paste("At the alpha level of", alpha, "the test statistic is", chi_sq_BP, "and the null hypothesis
  }
```

```
  else{
    paste("At the alpha level of", alpha, "the test statistic is", chi_sq_BP, "and the null hypothesis i
  }
}
```

Let the null hypothesis be that the error variance is constant and the alternative hypothesis that the error variance is not constant. Then

```
breusch.pagan(copier, alpha = .05)
```

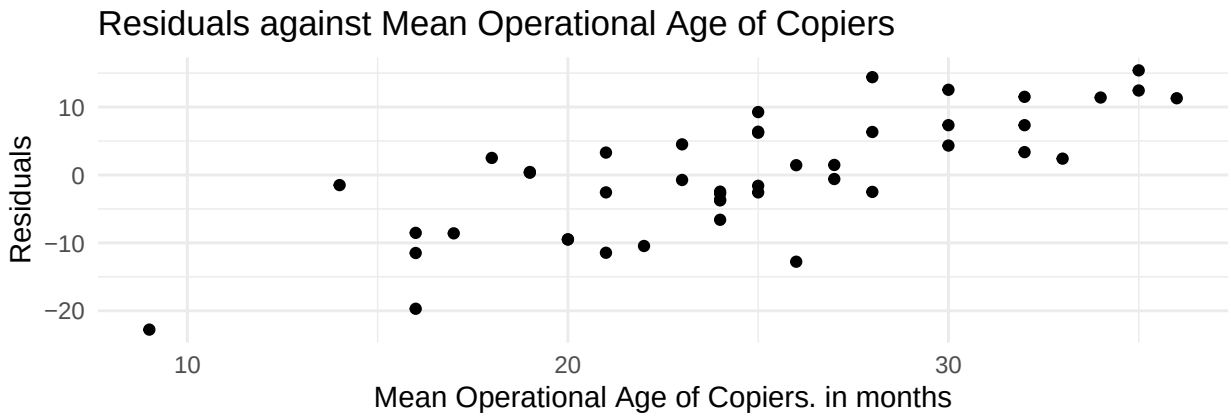
```
## [1] "At the alpha level of 0.05 the test statistic is 1.315 and the null hypothesis is failed to be r
```

(h) Information is given below on two variables not included in the regression model, namely, mean operational age of copiers serviced on the call (X_2, in months) and years of experience of the service person making the call (X_3). Plot the residuals against X_2 and X_3 on separate graphs to ascertain whether the model can be improved by including either or both of these variables. What do you conclude?

Answer:

The residuals against X_2 and X_3 are shown below.

```
copier_fulldata = read.csv('CH03PR04.txt', sep = '', header = FALSE,
                  col.names = c('y', 'x1', 'x2', 'x3'),
                  colClasses = rep('numeric', 4))
copier_fulldata_resids = merge(prob_plot_table(copier), copier_fulldata)

plot_1 = copier_fulldata %>% ggplot(aes(x = x2, y = resid)) + geom_point() +
  labs(x = "Mean Operational Age of Copiers. in months", y = "Residuals",
       title = "Residuals against Mean Operational Age of Copiers")

plot_2 = copier_fulldata %>% ggplot(aes(x = x3, y = resid)) + geom_point() +
  labs(x = "Years of Experience of the Service Person", y = "Residuals",
       title = "Residuals against Years of Experience of the Service Person")

grid.arrange(plot_1, plot_2, nrow = 2)
```



Residuals increase as the mean operational age of copiers, in months, increases, indicating that the model does not predict well for older copiers. There does not seem to be any pattern in model residuals and years of experience of the service person making the call.

Problem 5:

Refer to Airfreight breakage Problem 1.21.

(a) Prepare a dot plot for the number of transfers X_i. Does the distribution of number of transfers appear to be asymmetrical?

Answer:

The dot plot for the number of transfers is shown below.

```
airfreight = read.csv('CH01PR21.txt', sep = '', header = FALSE,
                  col.names = c('y', 'x'),
                  colClasses = c('numeric', 'numeric'))

airfreight %>%
  ggplot(aes(x = x)) + geom_dotplot(fill = "darkseagreen", color = "black") +
  scale_y_continuous(breaks = NULL, name = '') +
  labs(x = "Number of Transfers", y = '', title = "Dot Plot of the Number of Transfers")
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```

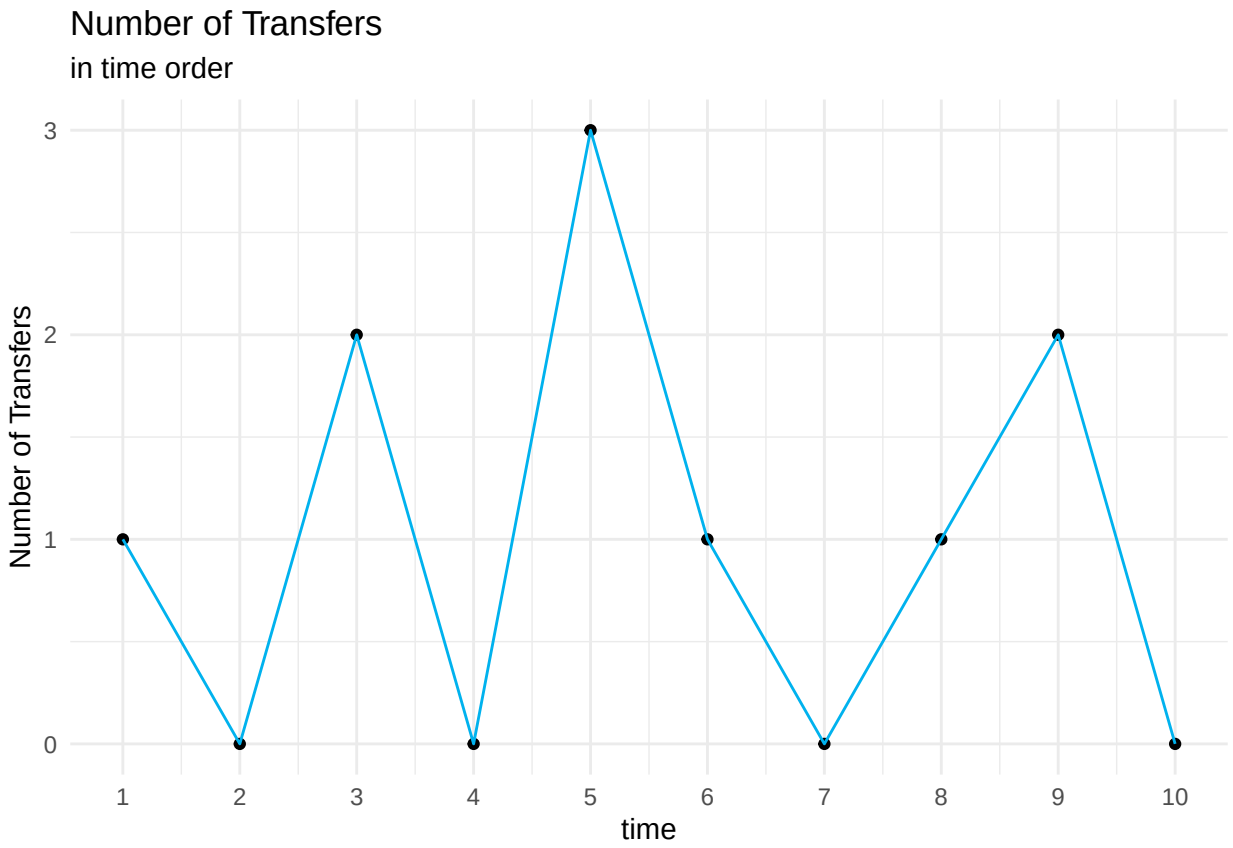
The distribution of the number of transfers appears to be assymmetrical, where as the number of transfers increases, the number of such cases decreases.

(b) The cases are given in time order. Prepare a time plot for the number of transfers. Is any systematic pattern evident in your plot? Discuss.

Answer:

The time plot of the number of transfers is shown below.

```
airfreight %>% mutate(time = 1:nrow(.)) %>%
  ggplot(aes(x = time, y = x)) +
  geom_point() +
  geom_line(color = "deepskyblue2") +
  scale_x_continuous(breaks = 1:10) +
  labs(x = "time", y = "Number of Transfers",
       title = "Number of Transfers", subtitle = "in time order")
```



There does not appear to be any systematic evidence of any trend in the number of transfers over time.

(c) Obtain the residuals e_i and prepare a stem-and-leaf plot of the residuals. What information is provided by your plot?

Answer:

The stem-and-leaf plot of the residuals is shown below.

```
model = lm.fit_manual(airfreight$x, airfreight$y)
pred = model[1] + model[2]*airfreight$x
resid = airfreight$y - pred
df = data.frame(x = airfreight$x, y = airfreight$y, pred, resid)

stem(df$resid)
```

```
##
##   The decimal point is at the |
##
##   -2 | 2
##   -1 | 222
##   -0 | 2
##    0 | 888
##    1 | 88
```

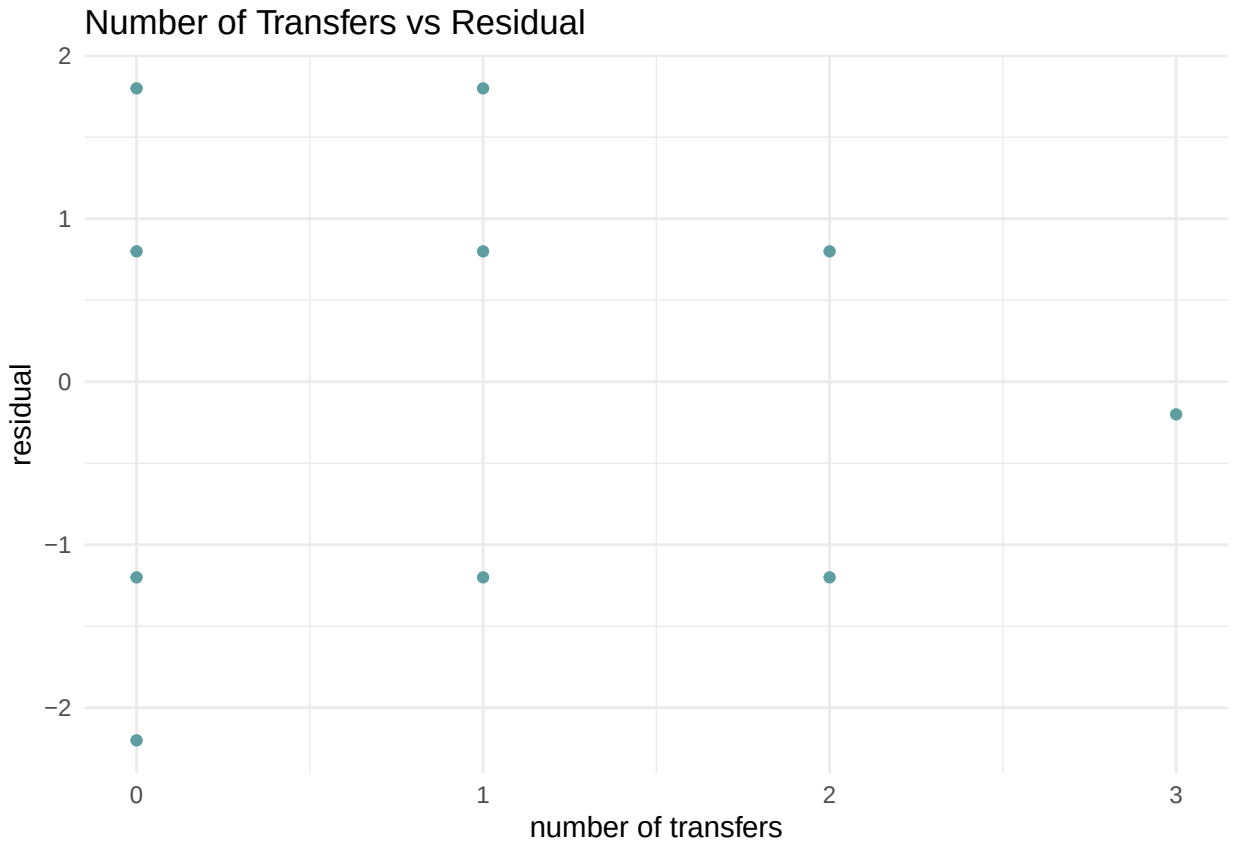
There are two high peaks in the residuals.

(d) Plot the residuals e_i against X_i to ascertain whether any departures from regression model (2.1) are evident. What is your conclusion?

Answer:

The plots of the residuals e_i against X_i is shown below.

```
df %>% ggplot(aes(x = x, y = resid)) +
  geom_point(color = "cadetblue") +
  labs(x = "number of transfers", y = "residual",
       title = "Number of Transfers vs Residual")
```



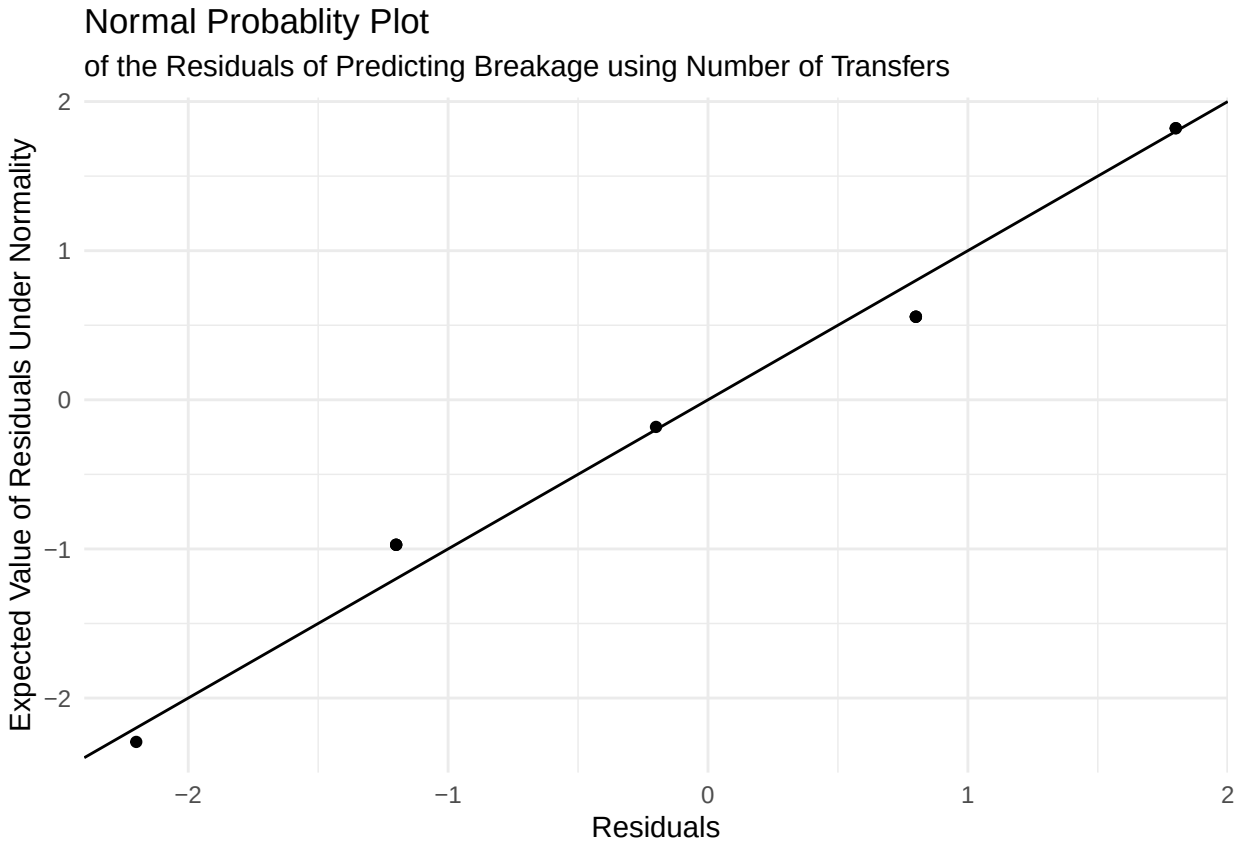
As the number of transfers increases, the residuals become smaller and smaller. This is a sign of nonconstant variance of the error terms.

(e) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality to ascertain whether the normality assumption is reasonable here. Use Table B.6 and $\alpha = .01$. What do you conclude?

Answer:

The normal probability plot of the residuals is shown below.

```
prob_plot(df, "of the Residuals of Predicting Breakage using Number of Transfers")
```



The coefficent of correlation between the ordered residuals and their expected values under normality is

```
prob_plot_table(airfreight) %>% select(resid, exp_val) %>% cor()
```

```
##             resid   exp_val
## resid   1.0000000 0.9913394
## exp_val 0.9913394 1.0000000
```

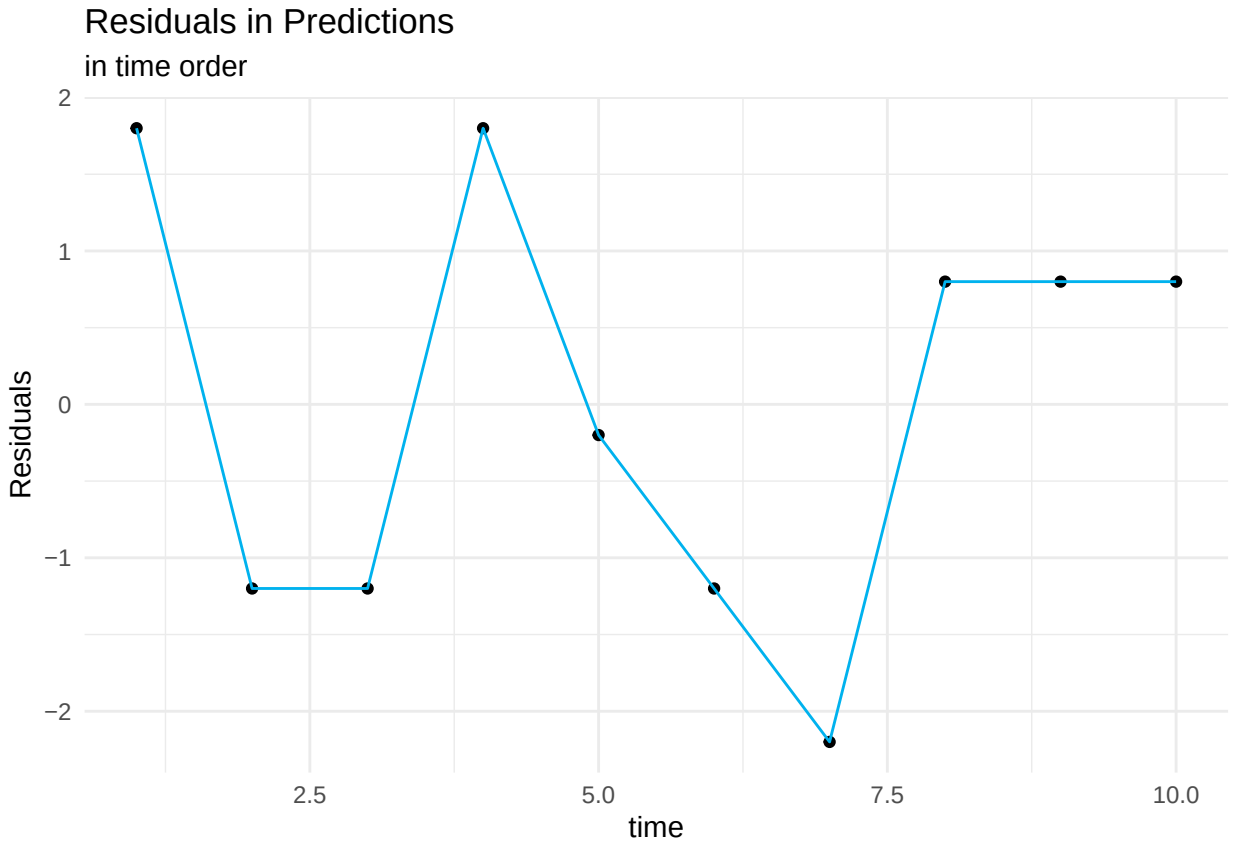
There is a high correlation between the ordered and expected value of the residuals. Using Table B.6, it is found that the critical value for the coefficient of correlation between ordered residuals and expected values under normality, at $\alpha = .01$ and $n = 10$, is .879. Since the calculated correlation coefficient is greater than the critical value, it can be stated that there is evidence that that the error terms are reasonably normally distributed. This was also evident in the plot above where the points appeared to line up.

(f) Prepare a time plot of the residuals. What information is provided by your plot?

Answer:

The time plot of the residuals is shown below.

```
df %>% mutate(time = 1:nrow(.)) %>%
  ggplot(aes(x = time, y = resid)) +
  geom_point() +
  geom_line(color = "deepskyblue2") +
  labs(x = "time", y = "Residuals",
       title = "Residuals in Predictions", subtitle = "in time order")
```



There is no pattern in the residuals as time passes, meaning there is no correlation in the error terms over time.

(g) Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $\alpha = .10$. State the alternatives, decision rule and conclusion. Does your conclusion support your preliminary findings in part (d)?

Answer:

Let the null hypothesis be that the error variance is constant and the alternative hypothesis that the error variance is not constant. Then

```
breusch.pagan(airfreight, alpha = .1)
```

```
## [1] "At the alpha level of 0.1 the test statistic is 1.033 and the null hypothesis is failed to be r
```

This conclusion is supported by the preliminary findings in part (d).

Problem 6:

Refer to Plastic hardness Problem 1.22.

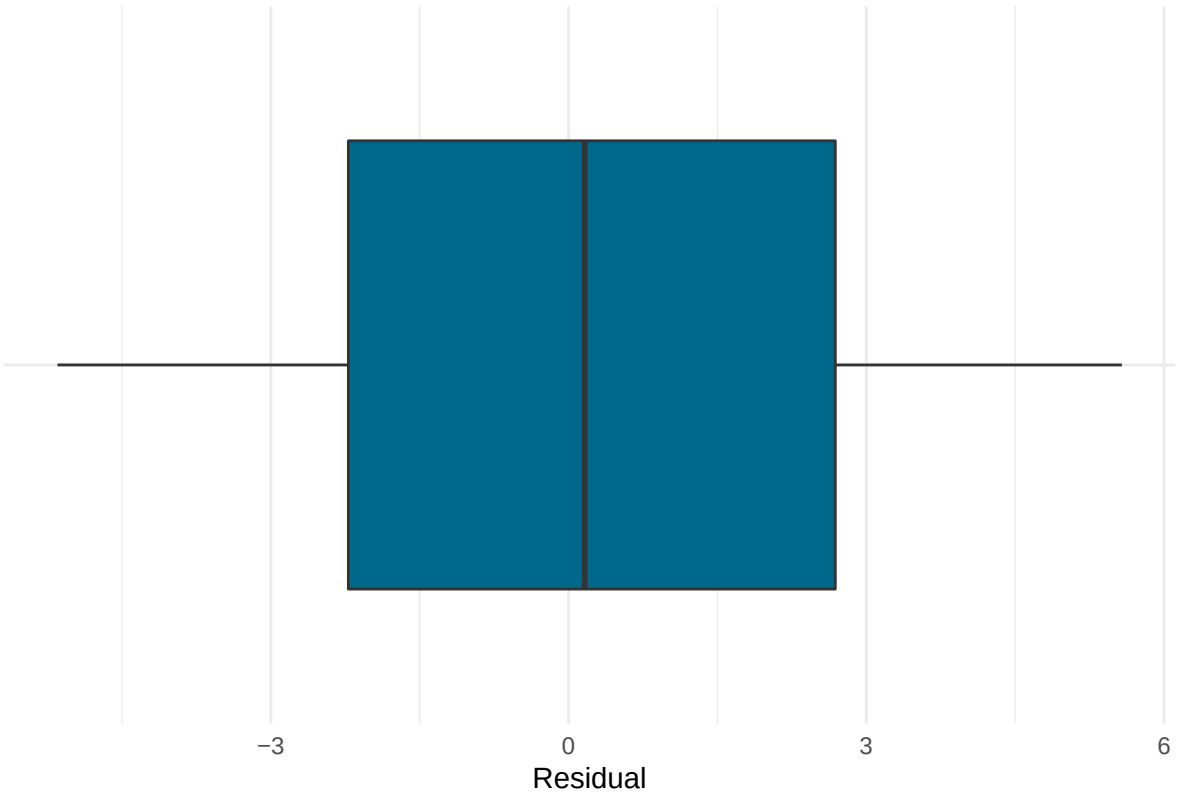
(a) Obtain the residuals e_i and prepare a box plot of the residuals. What information is provided by your plot?

Answer:

The boxplot of the residuals is shown below.

```
plastic = read.csv('CH01PR22.txt', sep = '', header = FALSE,
                   col.names = c('y', 'x'),
                   colClasses = c('numeric', 'numeric'))
```

```
plastic_model = prob_plot_table(plastic)

plastic_model %>%
  ggplot(aes(x = '', y = resid)) +
  geom_boxplot(fill = "deepskyblue4") +
  labs(x = '', y = 'Residual',
       title = "Box Plot of the Residuals of Predicting Hardness of Plastic using Time Elapsed") +
  coord_flip()
```



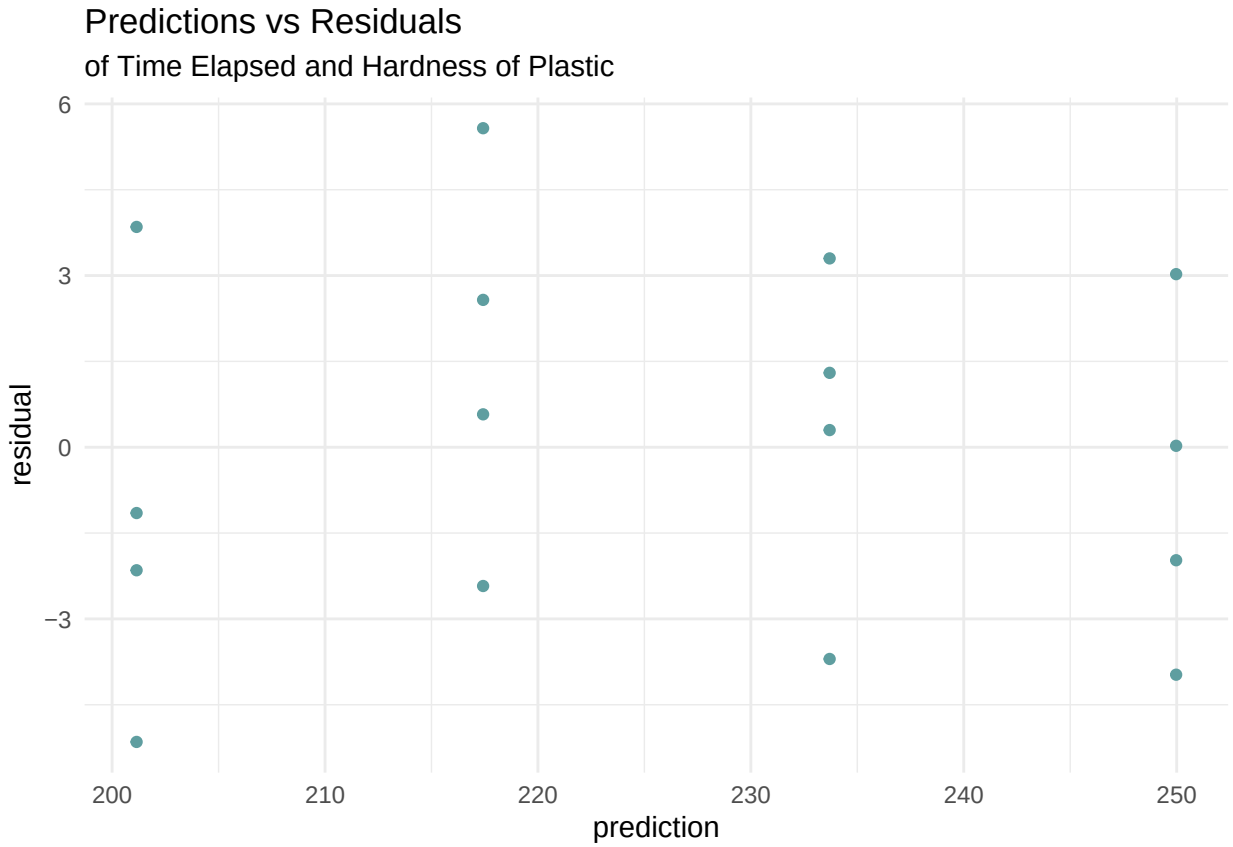
The residuals appear to be symmetrically distributed, with a median slightly above 0

(b) Plot the residuals e_i against the fitted values $\hat{Y}_i$ to ascertain whether any departures from regression model (2.1) are evident. State your findings.

Answer:

The residuals e_i are plotted against the fitted values $\hat{Y}_i$ below.

```
plastic_model %>%
  ggplot(aes(x = pred, y = resid)) +
  geom_point(color = "cadetblue") +
  labs(x = "prediction", y = "residual",
       title = "Predictions vs Residuals",
       subtitle = "of Time Elapsed and Hardness of Plastic")
```



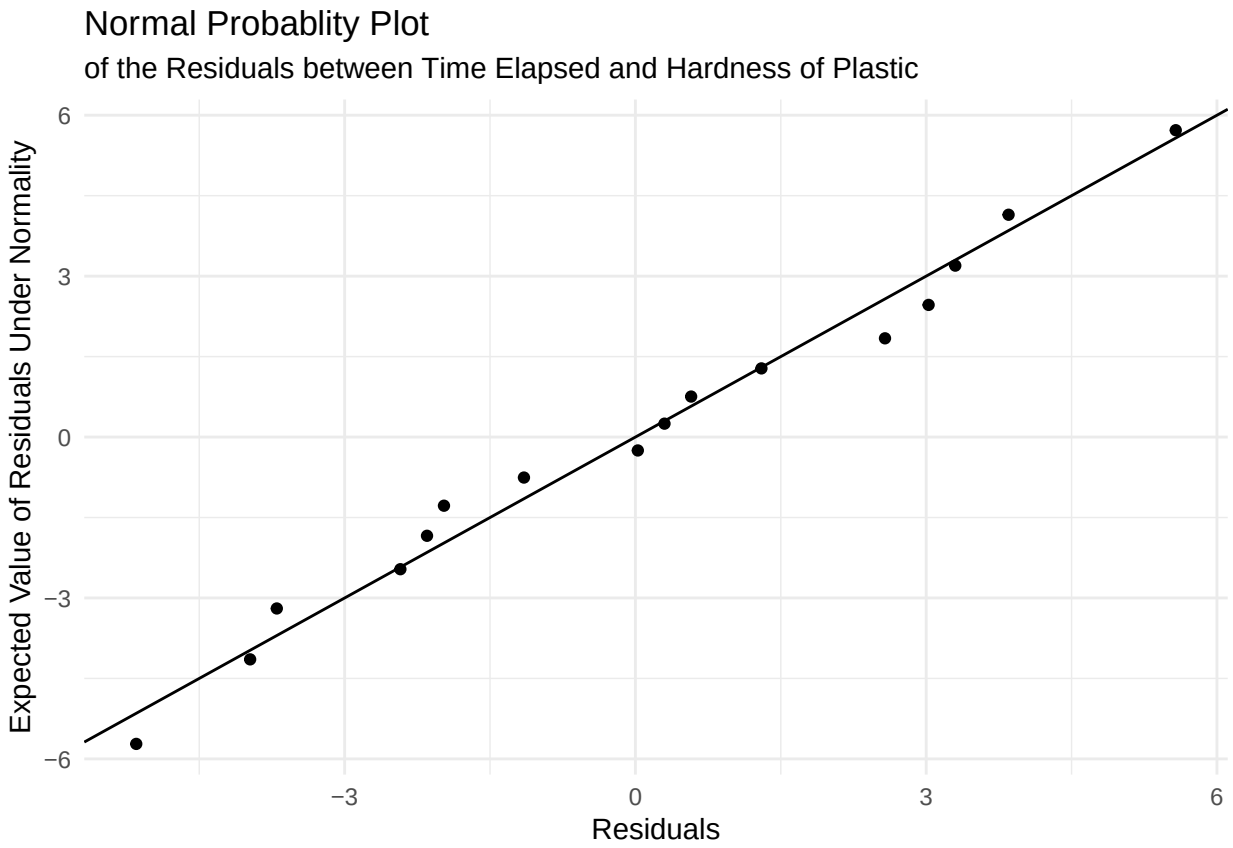
The residuals appear to be random yet have a downward trending slope. This could mean the variance of the error term is nonconstant.

(c) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation betwewn the ordered residuals and their expected values under normality. Does the normality assumption appear to be reasonable here? Use Table B.6 and $\alpha = .05$.

Answer:

The normal probability plot of the residuals is shown below.

```
prob_plot(plastic, "of the Residuals between Time Elapsed and Hardness of Plastic")
```



The coefficent of correlation between the ordered residuals and their expected values under normality is

```
prob_plot_table(plastic) %>% select(resid, exp_val) %>% cor()
```

```
##             resid   exp_val
## resid   1.0000000 0.9916733
## exp_val 0.9916733 1.0000000
```

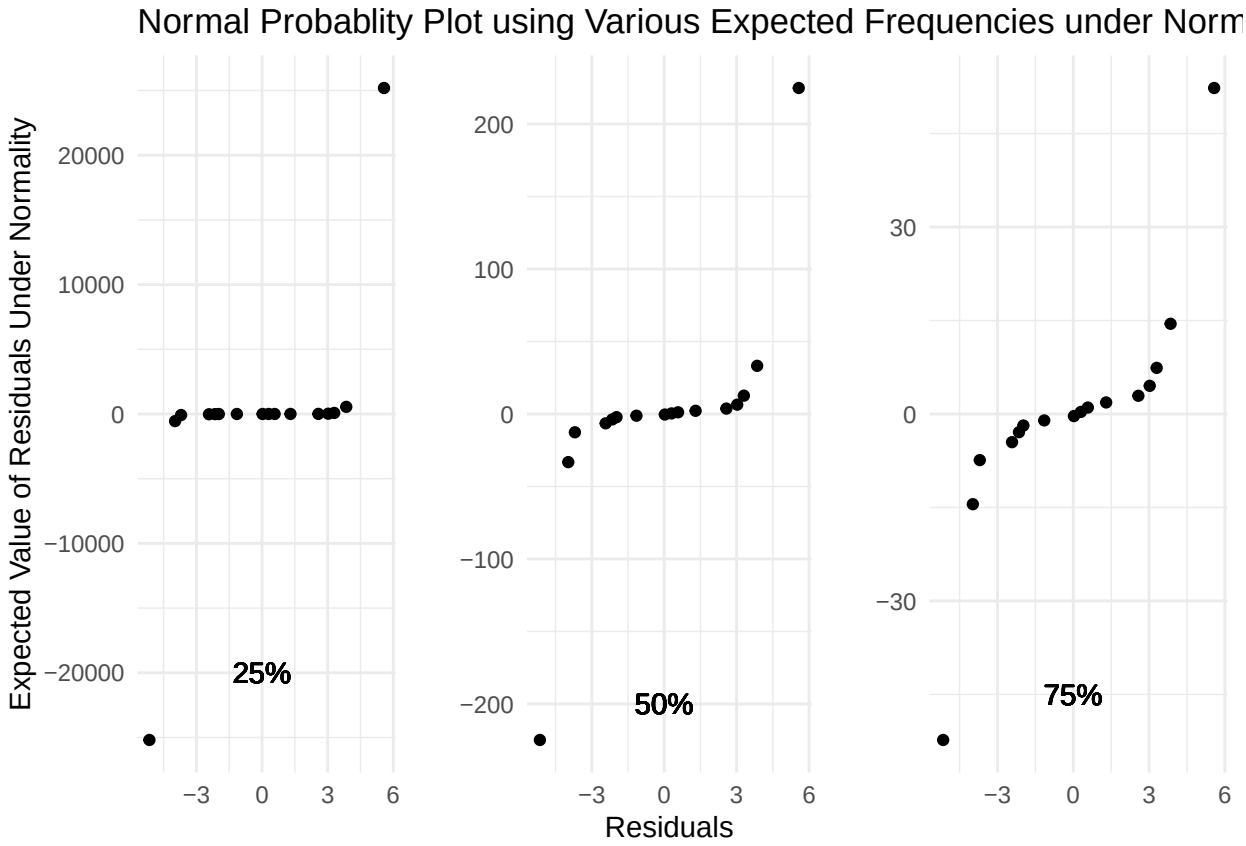
There is a high correlation between the ordered and expected value of the residuals. Using Table B.6, it is found that the critical value for the coefficient of correlation between ordered residuals and expected values under normality, at $\alpha = .05$ and $n = 16$, is .941. Since the calculated correlation coefficient is greater than the critical value, it can be stated that there is evidence that that the error terms are reasonably normally distributed. This was also evident in the plot above where the points appeared to line up.

(d) Compare the frequencies of the residuals against the expected frequencies under normality, using the 25th, 50th and 75th percentiles of the relevant t distribution. Is the information provided by these comparisons consistent with the findings from the normal probability plot in part (c)?

Answer:

```
expected_freq = function(df){
  model = lm.fit_manual(df$x, df$y)
  pred = model[1] + model[2]*df$x
  resid = df$y - pred
  n = nrow(df)
  resids_df = data.frame(x = df$x, y = df$y, pred, resid)

  mse_sq_root = sqrt(sum((resids_df$y - resids_df$pred)^2) / (n - 2))

  resids_df = resids_df %>% mutate("Run" = 1:nrow(resids_df),
```

```
                                    "k" = rank(resid)) %>%
    mutate("exp_val_25" = mse_sq_root * qt((k - .375)/(n + .25), .25),
           "exp_val_50" = mse_sq_root * qt((k - .375)/(n + .25), .5),
           "exp_val_75" = mse_sq_root * qt((k - .375)/(n + .25), .75))
  return(resids_df)
}

expected_freq_table = expected_freq(plastic)

plot_1 = expected_freq_table %>%
  ggplot(aes(x = resid, y = exp_val_25)) +
  geom_point() +
  geom_text(aes(x = 0, y = -20000, label = "25%")) +
  labs(x = '',
       y = "Expected Value of Residuals Under Normality",
       title = "Normal Probablity Plot using Various Expected Frequencies under Normality")

plot_2 = expected_freq_table %>%
  ggplot(aes(x = resid, y = exp_val_50)) +
  geom_point() +
  geom_text(aes(x = 0, y = -200, label = "50%")) +
  labs(x = "Residuals", y = '', title = '')

plot_3 = expected_freq_table %>%
  ggplot(aes(x = resid, y = exp_val_75)) +
  geom_point() +
  geom_text(aes(x = 0, y = -45, label = "75%")) +
  labs(x = '', y = '', title = '')

grid.arrange(plot_1, plot_2, plot_3, nrow = 1)
```



The information provided by these comparisons is consistent with the findings from the normal probability plot in part (c). As the percentile of the releveant t distribution increases, the points start to align themselves on a $y = x$ line and the scale of the expected value of the residual becomes smaller.

(e) Use the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into the two groups, $X \leq 24$ and $X > 24$ and use $\alpha = .05$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (b)?

Answer:

Let the null hypothesis be that the error variance is constant and the alternative hypothesis that the error variance is not constant. Then

```
brown.forsythe(plastic, 25, 0.05)
```

```
## [1] "At the alpha level of 0.05 the test statistic is 0.856 and the null hypothesis is failed to be r
```

Problem 7:

Refer to Muscle mass Problem 1.27.

(a) Prepare a stem-and-leaf plot for the ages X_i. Is this plot consistent with the random selection of women from each 10-year age group? Explain.

Answer:

A stem-and-leaf plot of the ages is shown below.

```
muscle = read.csv('CH01PR27.txt', sep = '', header = FALSE,
                  col.names = c('y', 'x'),
                  colClasses = c('numeric', 'numeric'))
```

```r
stem(muscle$x)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   4 | 11122334
##   4 | 5677788
##   5 | 123344
##   5 | 56777999
##   6 | 00013334
##   6 | 5556889
##   7 | 001223
##   7 | 5666788888
```

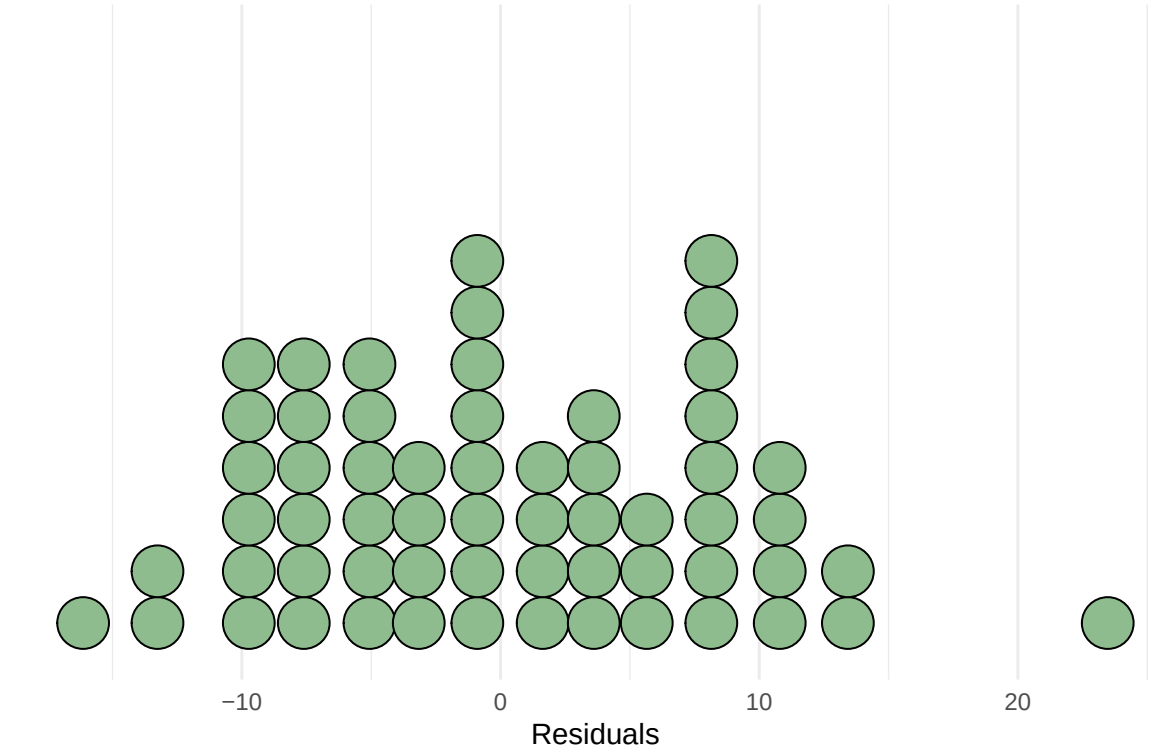
It appears to be that there is randomness in the ages of each age group of women. The distribution appears to be uniform with no skewness in either ends, nor is there any sense of normality.

(b) Obtain the residuals e_i and prepare a dot plot of the residuals. What does your plot show?

Answer:

The dot plot of the residuals is shown below.

```r
muscle_model = prob_plot_table(muscle)

muscle_model %>%
  ggplot(aes(x = resid)) +
  geom_dotplot(binwidth = 2, fill = "darkseagreen", color = "black") +
  scale_y_continuous(breaks = NULL, name = '') +
  labs(x = "Residuals",
       title = "Dot Plot of the Residuals", subtitle = "of Regressing Muscle Mass on Age")
```



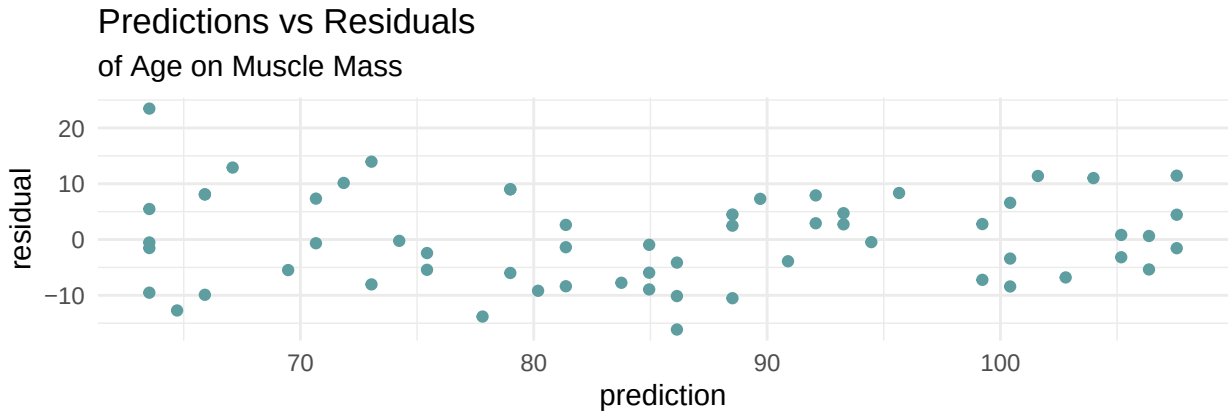
The residuals are skewed to the right due to the presence of an outlier.

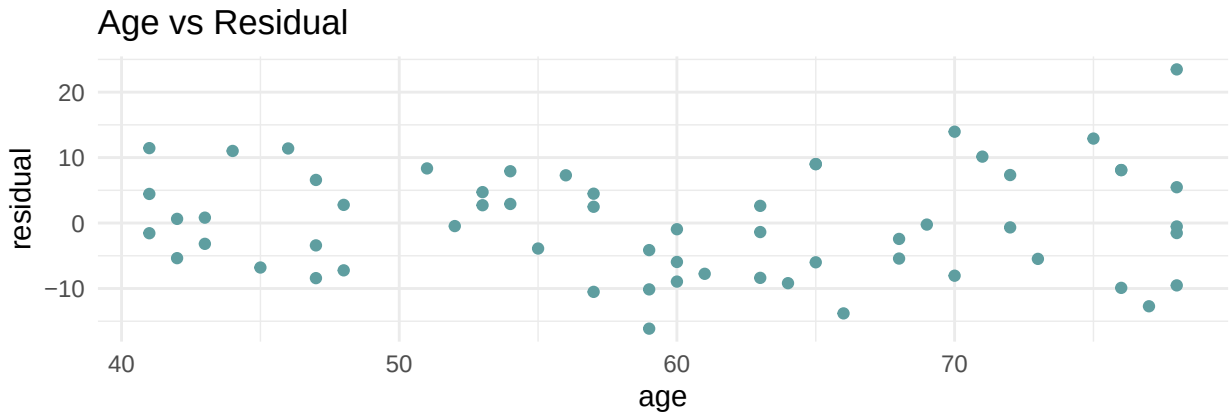
(c) Plot the residuals e_i against $\hat{Y}_i$ and also against X_i on separate graphs to ascertain whether any departures from regression model (2.1) are evident. Do the two plots provide the same information? State your conclusions.

Answer:

The plot of the residual e_i against $\hat{Y}_i$ and also against X_i is shown below.

```
plot_1 = muscle_model %>% ggplot(aes(x = pred, y = resid)) +
  geom_point(color = "cadetblue") +
  labs(x = "prediction", y = "residual",
       title = "Predictions vs Residuals",
       subtitle = "of Age on Muscle Mass ")

plot_2 = muscle_model %>% ggplot(aes(x = x, y = resid)) +
  geom_point(color = "cadetblue") +
  labs(x = "age", y = "residual",
       title = "Age vs Residual")

grid.arrange(plot_1, plot_2, nrow = 2)
```





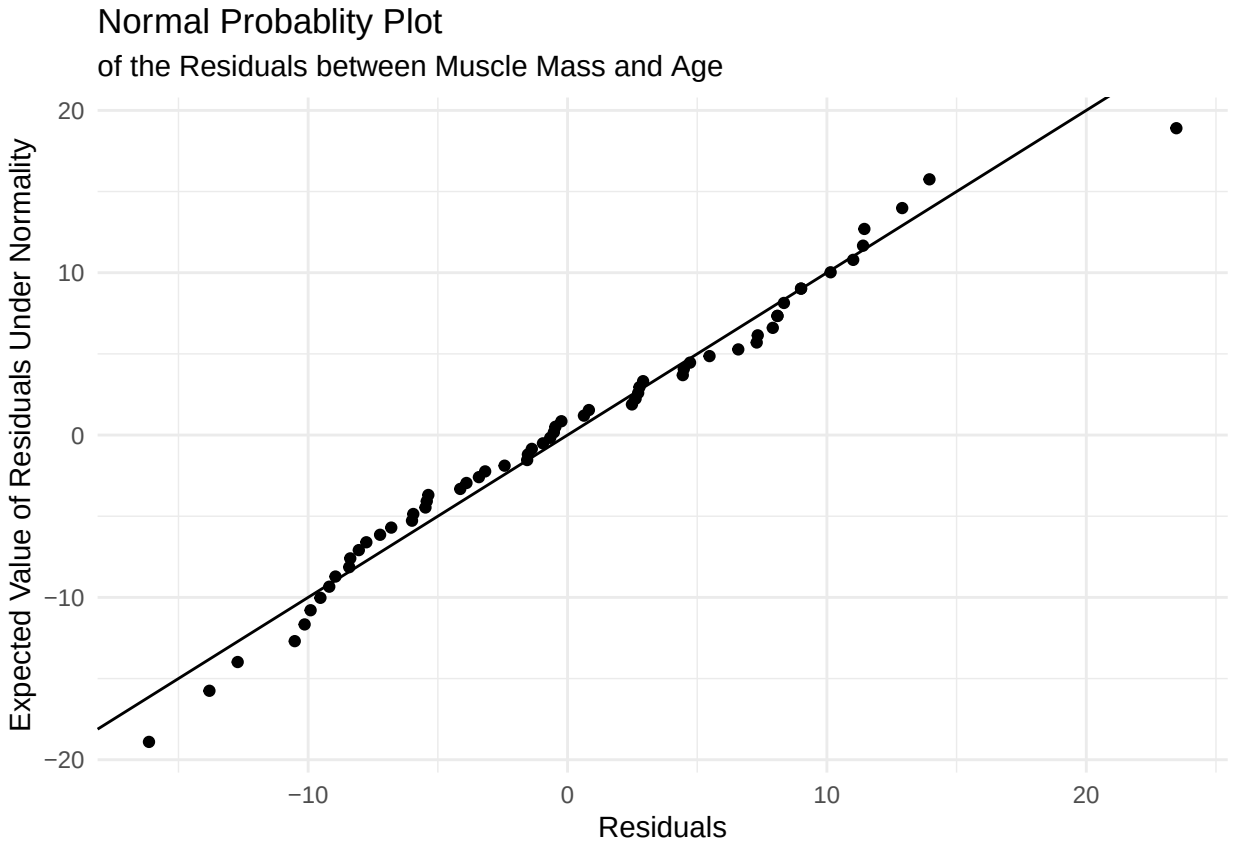
These two plots are showing the same information yet reversed. This is because of the negative trend between muscle mass and age. Nonetheless, since the residuals appear to be scattered and not follow a particular pattern, the error terms should have a constant variance. Residuals are slightly larger at the end of 1 side than the other.

(d) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality to ascertain whether the normality assumption is tenable here. Use Table B.6 and $\alpha = .10$. What do you conclude?

Answer:

The normal probability plot of the residuals is shown below.

```
prob_plot(muscle, "of the Residuals between Muscle Mass and Age")
```



The coefficent of correlation between the ordered residuals and their expected values under normality is

```
prob_plot_table(muscle) %>% select(resid, exp_val) %>% cor()
```

```
##             resid   exp_val
## resid   1.0000000 0.9897499
## exp_val 0.9897499 1.0000000
```

There is a high correlation between the ordered and expected value of the residuals. Using Table B.6, it is found that the critical value for the coefficient of correlation between ordered residuals and expected values under normality, at $\alpha = .10$ and $n = 60$, is .984. Since the calculated correlation coefficient is greater than the critical value, it can be stated that there is evidence that that the error terms are reasonably normally distributed. This was also evident in the plot above where the points appeared to line up.

(e) Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $\alpha = .01$. State the alternatives, decision rule and conclusion. Is your conclusion consistent with your preliminary findings in part (c)?

Answer:

Let the null hypothesis be that the error variance is constant and the alternative hypothesis that the error variance is not constant. Then

```
breusch.pagan(muscle, alpha = .01)
```

```
## [1] "At the alpha level of 0.01 the test statistic is 3.817 and the null hypothesis is failed to be r
```

Problem 8:

Refer to Crime rate Problem 1.28.

(a) Prepare a stem-and-leaf plot for the percentage of individuals in the county having at least a high school diplima X_i. What information does your plot provide?

Answer:

The stem-and-leaf plot of the percentage of individuals in the county having at least a high school diploma is shown below.

```
crime = read.csv('CH01PR28.txt', sep = '', header = FALSE,
                 col.names = c('y', 'x'),
                 colClasses = c('numeric', 'numeric'))

stem(crime$x)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   6 | 1444
##   6 | 5678
##   7 | 00334444
##   7 | 55556666777777788888888999999
##   8 | 000011111112222222233333344444
##   8 | 55578889
##   9 | 11
```

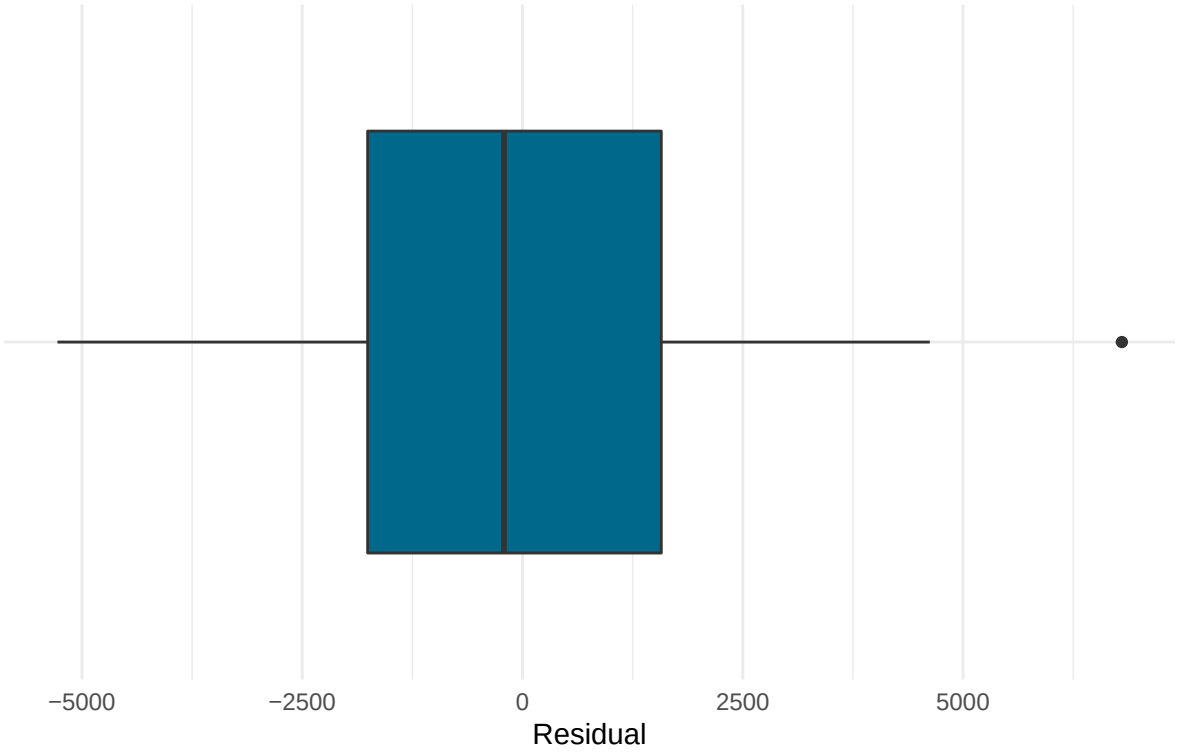
There is much higher number of percentages in the 70s and 80s range than either ends. The percentages are distributed normally.

(b) Obtain the residuals e_i and prepare a box plot of the residuals. Does the distribution of the residuals appear to be symmetrical?

Answer:

The box plot of the residuals is shown below.

```
crime_model = prob_plot_table(crime)

crime_model %>%
  ggplot(aes(x = '', y = resid)) +
  geom_boxplot(fill = "deepskyblue4") +
  labs(x = '', y = 'Residual',
       title = "Box Plot of the Residuals of Predicting Crime Rate",
       subtitle = "using % of People with High School Diplomas") +
  coord_flip()
```



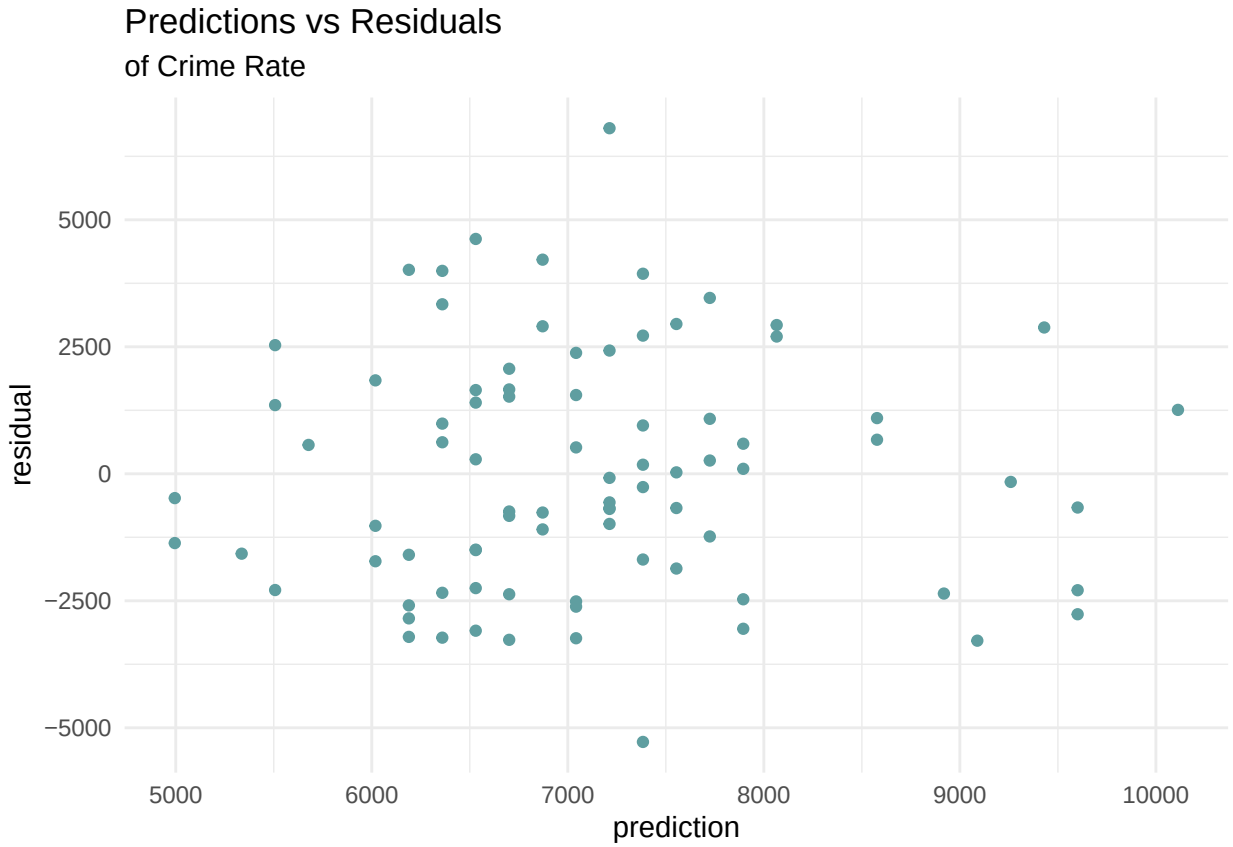
The distribution of the residuals does not appear to be symmetrical.

(c) Make a residual plot of e_i against $\hat{Y}_i$. What does the plot show?

Answer:

The residual plot of e_i against $\hat{Y}_i$ is shown below.

```
crime_model %>%
  ggplot(aes(x = pred, y = resid)) +
  geom_point(color = "cadetblue") +
  labs(x = "prediction", y = "residual",
       title = "Predictions vs Residuals",
       subtitle = "of Crime Rate")
```



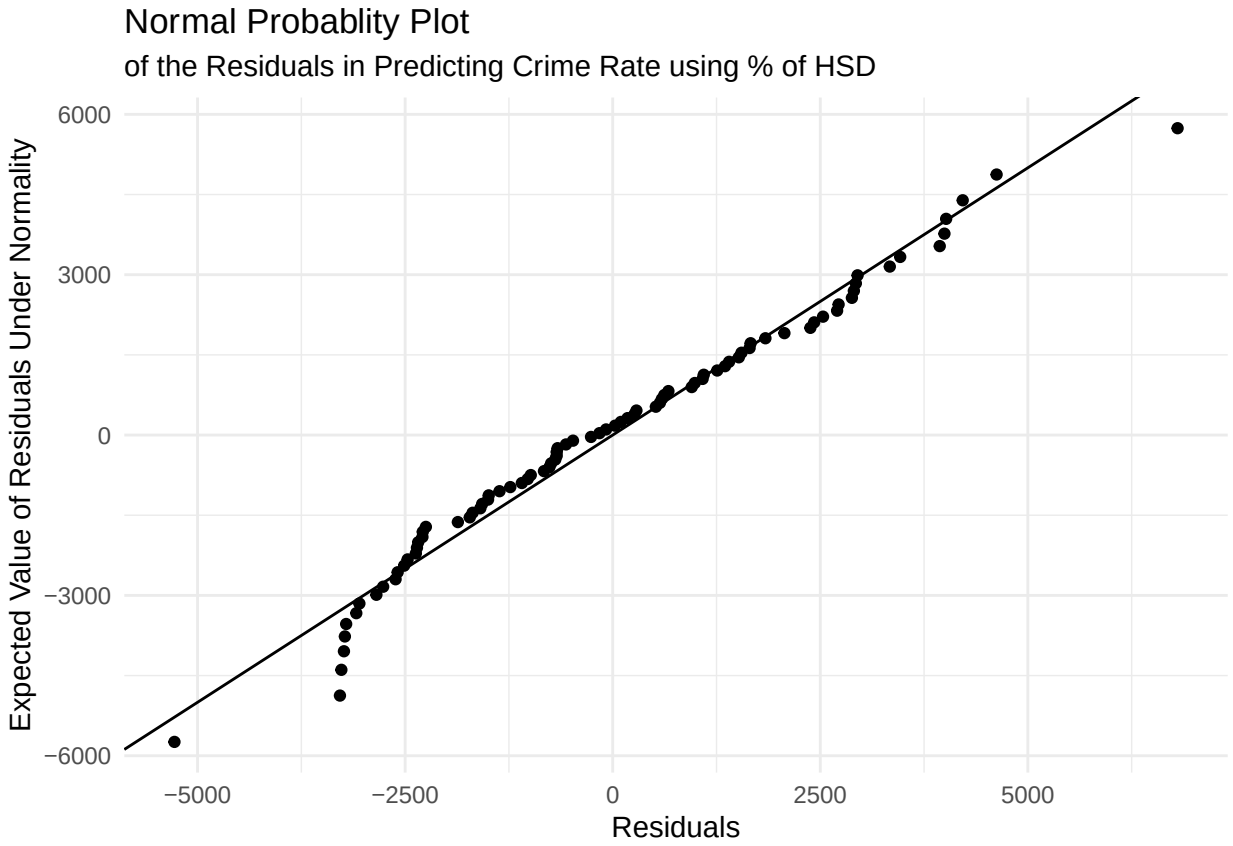
The residuals appear to be more varied in the middle range of predictions whereas at either ends it congregates towards 0.

(d) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality using Table B.6 and $\alpha = .05$. What do you conclude?

Answer:

The normal probability plot of the residuals is shown below.

```
prob_plot(crime, "of the Residuals in Predicting Crime Rate using % of HSD")
```



The coefficent of correlation between the ordered residuals and their expected values under normality is

```
prob_plot_table(crime) %>% select(resid, exp_val) %>% cor()
```

```
##             resid   exp_val
## resid   1.0000000 0.9887589
## exp_val 0.9887589 1.0000000
```

There is a high correlation between the ordered and expected value of the residuals. Using Table B.6, it is found that the critical value for the coefficient of correlation between ordered residuals and expected values under normality, at $\alpha = .05$ and $n = 84$, is .985. Since the calculated correlation coefficient is greater than the critical value, it can be stated that there is evidence that that the error terms are reasonably normally distributed. This was also evident in the plot above where the points appeared to line up.

(e) Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into two groups, $X \leq 69$ and $X > 69$ and use $\alpha = .05$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (c)?

Answer:

Let the null hypothesis be that the error variance is constant and the alternative hypothesis that the error variance is not constant. Then

```
brown.forsythe(crime, 69, 0.05)
```

```
## [1] "At the alpha level of 0.05 the test statistic is 0.355 and the null hypothesis is failed to be r
```

Problem 9:

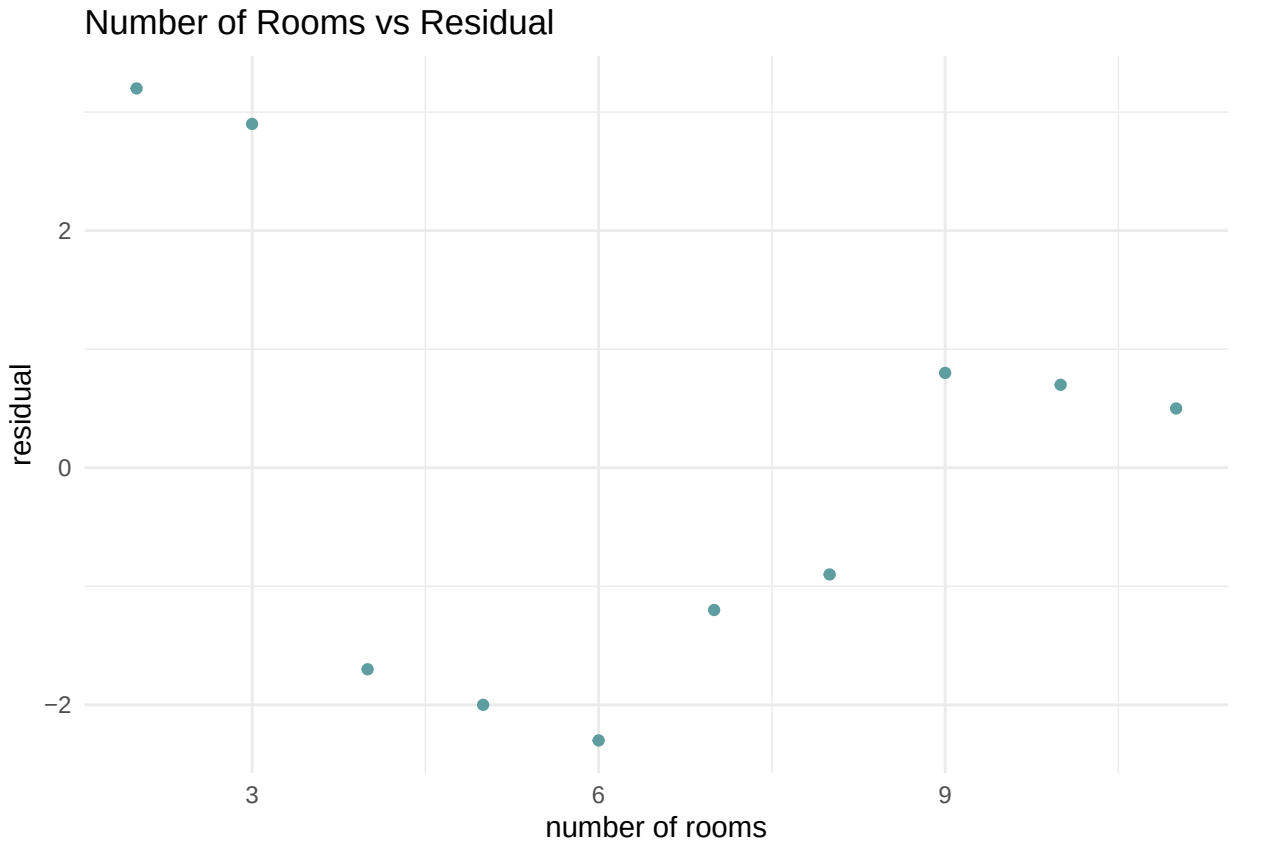
Electricity consumption. An economist studying the relation between household electricity (Y) and number of rooms in the home (X) employed linear regression model (2.1) and obtained the residuals. Plot the residuals

against X_i. What problem appears to be present here? Might a transformation alleviate the problem?

Answer:

A plot of the residuals against X_i is shown below.

```
electricity = read.csv('CH03PR09.txt', sep = '', header = FALSE,
                  col.names = c('x', 'e'),
                  colClasses = c('numeric', 'numeric'))

electricity %>%
  ggplot(aes(x = x, y = e)) +
  geom_point(color = "cadetblue") +
  labs(x = "number of rooms", y = "residual",
       title = "Number of Rooms vs Residual")
```



Residuals appear to become large and small in no detectable pattern. A transformation might alleviate this problem and find a viable way to correct the nonlinearity of the regression function.

Problem 10:

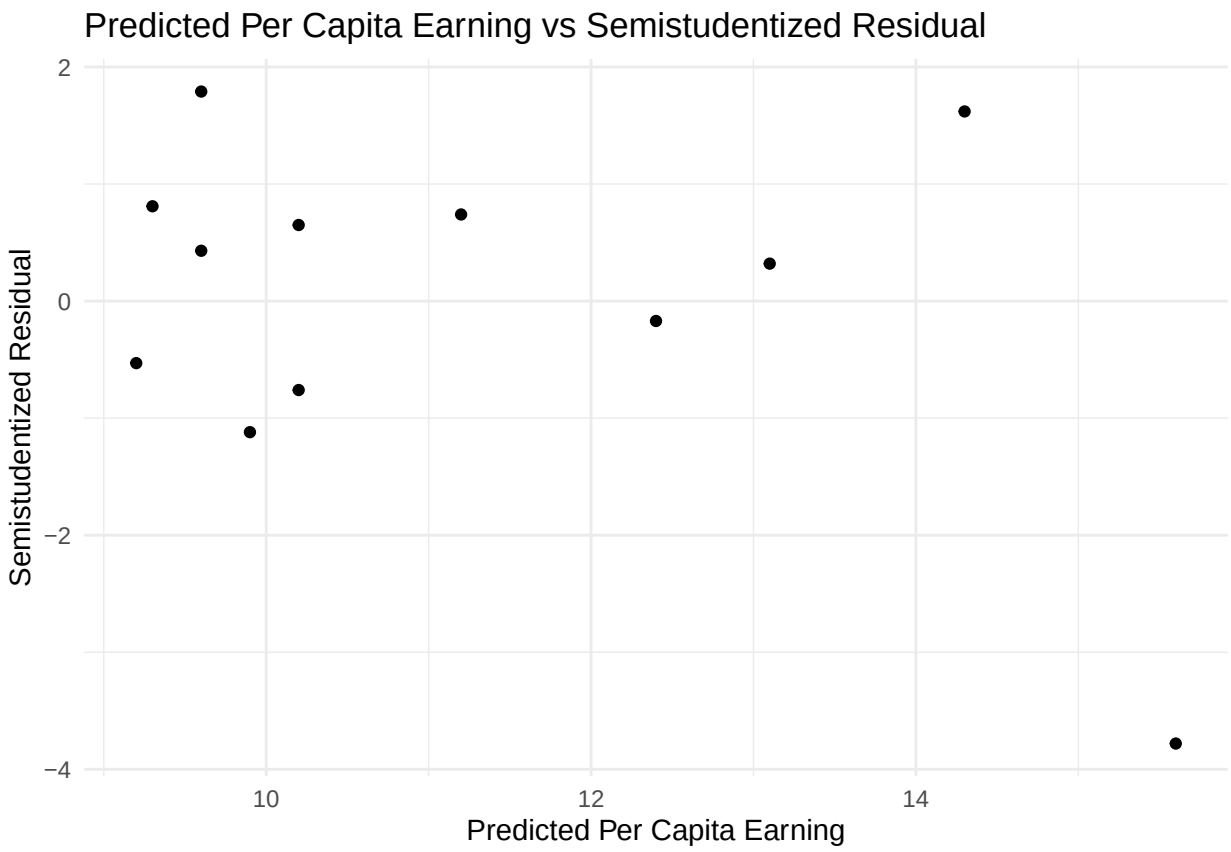
Per capita earnings. A sociologist employed linear regression model (2.1) to relate per capita earnings (Y) to average number of years of schooling (X) for 12 cities. The fitted values $\hat{Y}_i$ and the semistudentized residuals e_i^* are given.

(a) Plot the semistudentized residuals against the fitted values. What does the plot suggest?

Answer:

A plot of the semistudentized residuals against the fitted values is shown below.

```
earnings = read.csv('CH03PR10.txt', sep = '', header = FALSE,
                    col.names = c('y', 'e'),
                    colClasses = c('numeric', 'numeric'))

earnings %>%
  ggplot(aes(x = y, y = e)) +
  geom_point() +
  labs(x = "Predicted Per Capita Earning", y = "Semistudentized Residual",
       title = "Predicted Per Capita Earning vs Semistudentized Residual")
```



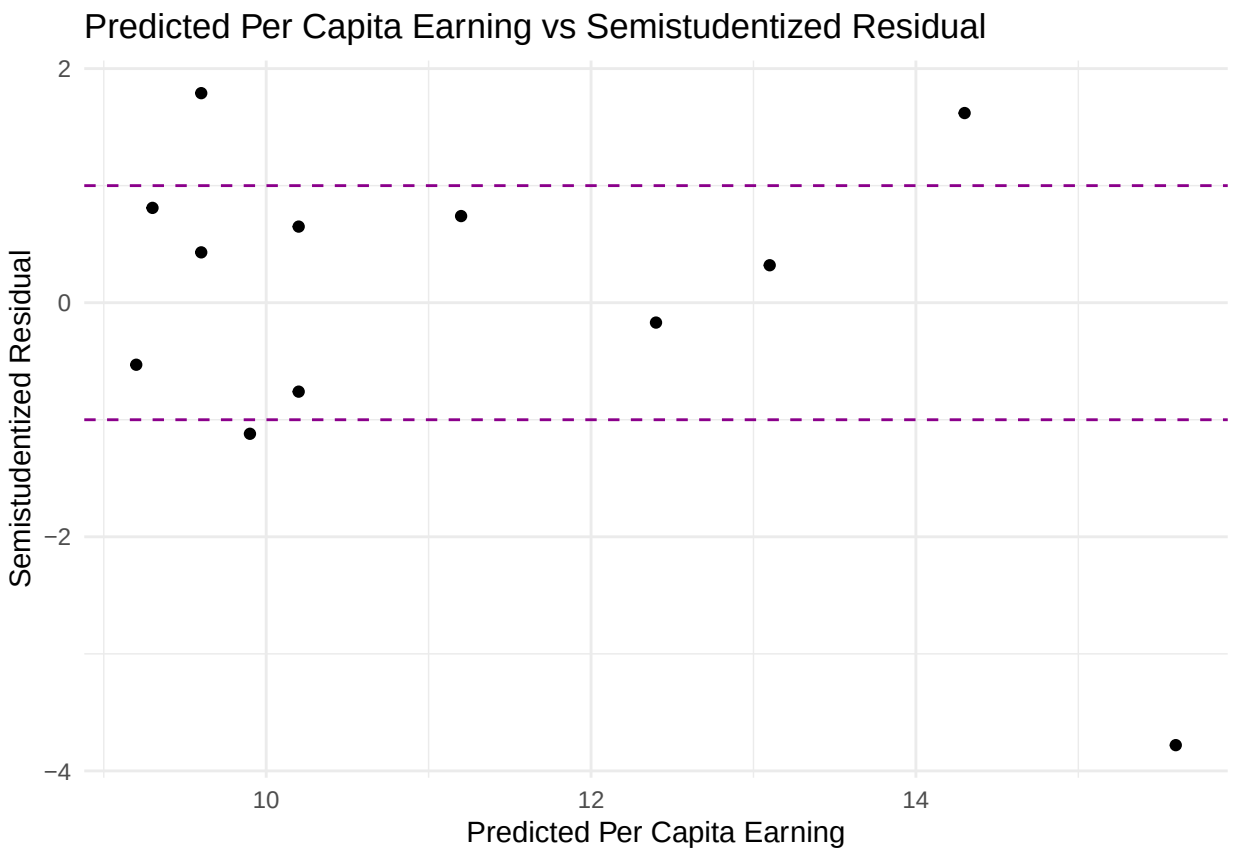
There is no pattern between the predicted per capita earning and the semistudentized residuals. However there appears to be one outlier that has a residual much lower than the other predicted per capita earnings.

(b) How many semistudentized residuals are outside ± 1 standard deviation? Approximately how many would you expect to see if the normal error model is appropriate?

Answer:

The above plot is redrawn with two lines added indicating the area above and below 1 standard deviation.

```
earnings %>%
  ggplot(aes(x = y, y = e)) +
  geom_point() +
  geom_hline(yintercept = -1,
             linetype = "dashed", color = "magenta4") +
  geom_hline(yintercept = 1,
             linetype = "dashed", color = "magenta4") +
  labs(x = "Predicted Per Capita Earning", y = "Semistudentized Residual",
       title = "Predicted Per Capita Earning vs Semistudentized Residual")
```



There appears to be 4 semistudentized residuals outside ± 1 standard deviation. If the normal error model was appropriate, then there should be 4 semistudented residuals outside ± 1 standard deviation, since 68% of the values shall fall in within 1 standard deviation of the mean. In this case, since the size of the data is 12, 68% of it is 8.16, which is within 1 standard deviation, and the remaining 3.84, or 4 falls outside.

Problem 11:

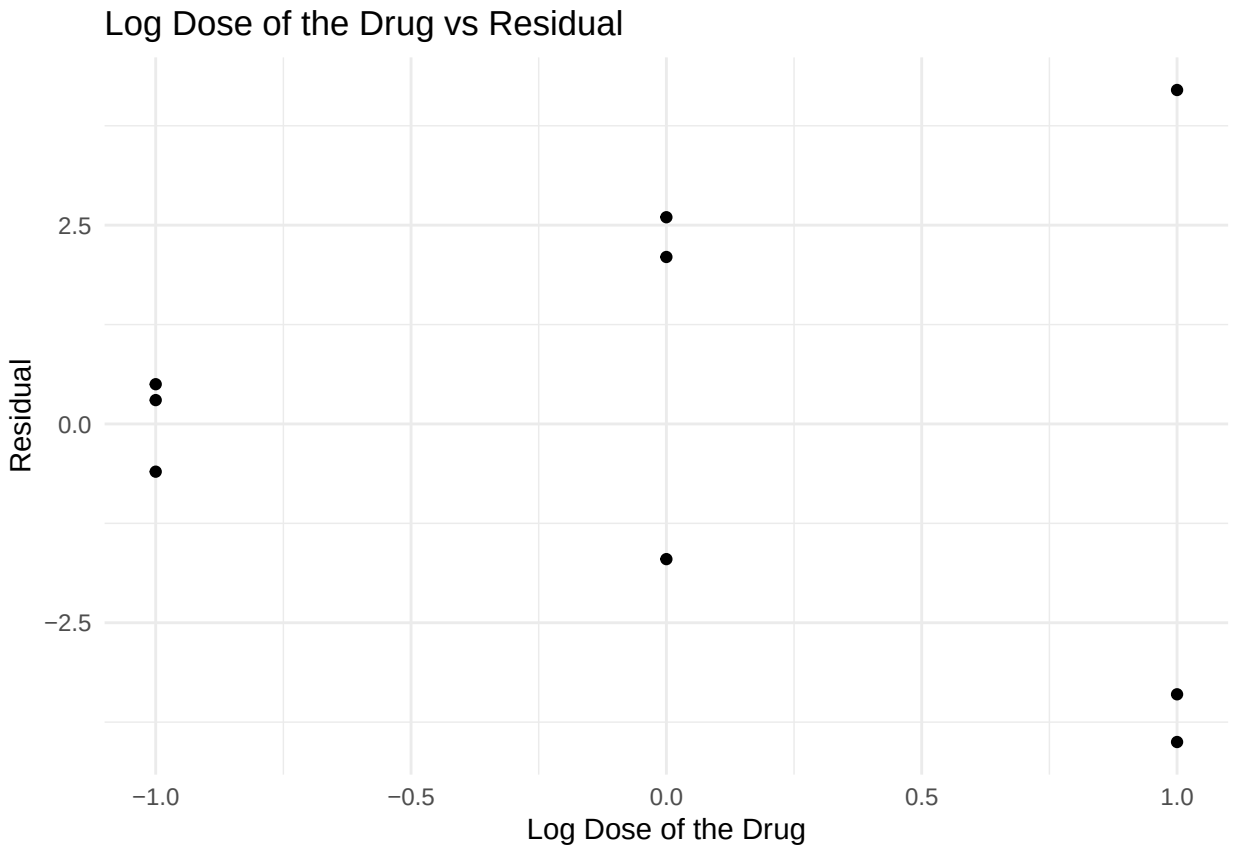
Drug concentration. A pharmacologist employed linear regression model (2.1) to study the relation between the concentration of a drug in plasma (Y) and the log-dose of the drug (X). The residuals and log-dose levels are given.

(a) Plot the residuals e_i against X_i. What conclusions do you draw from the plot?

Answer:

A plot of the residuals against the X_i values is shown below.

```
drug = read.csv('CH03PR11.txt', sep = '', header = FALSE,
                  col.names = c('x', 'e'),
                  colClasses = c('numeric', 'numeric'))

drug %>%
  ggplot(aes(x = x, y = e)) +
  geom_point() +
  labs(x = "Log Dose of the Drug", y = "Residual",
       title = "Log Dose of the Drug vs Residual")
```



As the log dose of the drug increases, the residual grows in magnitude. This shows that the error variance is not constant.

(b) Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with log-dose of the drug (X). Use $\alpha = .05$. State the alternatives, decision rule and conclusion. Does your conclusion support your preliminary findings in part (a)?

Answer:

Let the null hypothesis be that the error variance is constant and the alternative hypothesis that the error variance is not constant. Then

```
breusch.pagan.error = function(df, alpha){

  df_new = data.frame(x = df$x, resid = df$e, resid_sq = (df$e)^2)
  sse = sum(df$e^2)

  model_sq_error = lm.fit_manual(df_new$x, df_new$resid_sq)
  pred_sq_error = model_sq_error[1] + model_sq_error[2]*df_new$x
  ssr_ast = sum((pred_sq_error - mean(df_new$resid_sq))^2)

  chi_sq_BP = round((ssr_ast/2) / ((sse/nrow(df))^2), 3)

  if(chi_sq_BP < qchisq(1 - alpha, 1)){
    paste("At the alpha level of", alpha, "the test statistic is", chi_sq_BP, "and the null hypothesis i
  }
  else{
    paste("At the alpha level of", alpha, "the test statistic is", chi_sq_BP, "and the null hypothesis i
  }
```

```
}

breusch.pagan.error(drug, .05)
```

```
## [1] "At the alpha level of 0.05 the test statistic is 3.718 and the null hypothesis is failed to be r
```

Problem 12:

A student does not understand why the sum of squares defined in (3.16) is called a pure error sum of squares "since the formula looks like one for an ordinary sum of squares" Explain.

Answer:

The equation given in (3.16) is

$$\text{SSE}(F) = \sum_j \sum_i (Y_{ij} - \overline{Y}_j)^2 = \text{SSPE}$$

This is called the pure error sum of squares because it quantifies how much of a lack of fit in Y_{ij} there is from the estimated value for Y_{ij}. It is different from the formula for the ordinary sum of squares because it is derived from the notion that the full model for a simple linear regression model is

$$Y_{ij} = \mu_j + \varepsilon_{ij}$$

where here there is no restrictions on the means μ_j ($E[Y_{ij}] = \mu_j$). In the regression model, the mean responses are linearly related to X, or $E[Y] = \beta_0 + \beta_1 X$.

Problem 13:

Refer to Copier maintenance Problem 1.20.

(a) What are the alternative conclusions when testing for lack of fit of a linear regression function?

Answer: When testing for lack of fit of a linear regression function, the alternative conclusions are such that: the expected value of Y is $\beta_0 + \beta_1 X$ (or $H_0 : E[Y] = \beta_0 + \beta_1 X$) and the expected value of Y is not $\beta_0 + \beta_1 X$ (or $H_A : E[Y] \neq \beta_0 + \beta_1 X$).

(b) Perform the test indicated in part (a). Control the risk of Type I error at .05. State the decision rule and conclusion.

Answer:

```
lack.of.fit.test = function(data, alpha = 0.05){

  X = data$x
  Y = data$y
  Y_bars = data.frame(x = sort(unique(X)))
  mean_vec = c()
  for(i in Y_bars$x){
    temp_mean = mean(data[data$x == i, "y"])
    mean_vec = c(mean_vec, temp_mean)
  }
  Y_bars$ybars = mean_vec
  SSPE= 0
  for(i in Y_bars$x){
    temp_Y = data[data$x == i, "y"]
    SSPE = SSPE + sum((temp_Y - Y_bars[Y_bars$x == i, "ybars"])^2)
  }
  df_F = nrow(data) - length(unique(X))
```

```
  model = lm.fit_manual(X, Y)
  Ypreds = model[1] + model[2]*X
  SSE_R = sum((Ypreds - Y)^2)
  df_R = nrow(data) - 2

  F_ast = round(((SSE_R - SSPE) / (df_R - df_F))/(SSPE / df_F), 3)

  if(F_ast < qf(1 - alpha, length(unique(X)) - 2, nrow(data) - length(unique(X)))){
    paste("At the alpha level of", alpha, "the test statistic is", F_ast, "and the null hypothesis is fa
  }
  else{
    paste("At the alpha level of", alpha, "the test statistic is", F_ast, "and the null hypothesis is re
  }
}

lack.of.fit.test(copier, 0.05)
```

```
## [1] "At the alpha level of 0.05 the test statistic is 0.968 and the null hypothesis is failed to be r
```

(c) Does your test in part (b) detect other departures from regression model (2.1), such as lack of constant variance or lack of normality in the error terms? Could the results of the test of lack of fit be affected by such departures? Discuss.

Answer: The lack of fit test decomposes the error sum of squares SSE as follows

$$SSE = SSPE + SSLF$$

where SSE is the error deviation, SSPE is the pure error deviation and SSLF is the lack of fit deviation. This reveals that the error deviations in SSE depend on the lack of fit deviation. Hence if there is non-constant variance or lack of normality in the error terms, it will be shown in the lack of fit component of the SSE.

Problem 14:

Refer to Plastic hardness Problem 1.22.

(a) Perform the F test to determine whether or not there is lack of fit of a linear regression function; use $\alpha = .01$. State the alternatives, decision rule and conclusion.

Answer: Let the null hypothesis be that there exists a linear regression function associating hardness of plastic and the time elapsed. Let the alternative hypothesis be that there does not exist such a linear regression function. Then

```
lack.of.fit.test(plastic, 0.01)
```

```
## [1] "At the alpha level of 0.01 the test statistic is 0.824 and the null hypothesis is failed to be r
```

(b) Is there any advantage of having an equal number of replications at each of the X levels? Is there any disadvantage?

Answer: There is no advantage of having an equal number of replications at each of the X levels.

(c) Does the test in part (a) indicate what regression function is appropriate when it leads to the conclusion that the regression function is not linear? How would you proceed?

Answer: When the lack of fit test concludes that the regression function is not linear, it does not indicate which family of regression function is the true representation. To do so, studying the residuals can help to identify an appropriate family of functions.

Problem 15:

Solution concentration. A chemist studied the concentration of a solution (Y) over time (X). Fifteen identical solutions were prepared. The 15 solutions were randomly divided into five sets of three, and the five sets were measured, respectively, after 1, 3, 5, 7 and 9 hours.

(a) Fit a linear regression function.

Answer:

```
solutions = read.csv('CH03PR15.txt', sep = '', header = FALSE,
                     col.names = c('y', 'x'),
                     colClasses = c('numeric', 'numeric'))
model = round(lm.fit_manual(solutions$x, solutions$y), 3)

paste("The coefficients of the linear regression function are b0 = ", model[1],
      " and b1 = ", model[2], ".", sep = '')
```

```
## [1] "The coefficients of the linear regression function are b0 = 2.575 and b1 = -0.324."
```

(b) Perform the F test to determine whether or not there is lack of fit of a linear regression function; use $\alpha = .025$. State the alternatives, decision rule and conclusion.

Answer: The null hypothesis is that there exists a linear regression function relating the concentration of a solution to time. The alternative hypothesis is that there is no such linear function. Then

```
lack.of.fit.test(solutions, 0.025)
```

```
## [1] "At the alpha level of 0.025 the test statistic is 58.603 and the null hypothesis is rejected. Th
```

(c) Does the test in part (b) indicate what regression function is appropriate when it leads to the conclusion that lack of fit of a linear regression function exists? Explain.

Answer: The lack of best fit test does not indicate what regression function is appropriate when it leads to the conclusion that lack of fit of a linear regression function exists. It only tells that a linear relationship cannot be established.

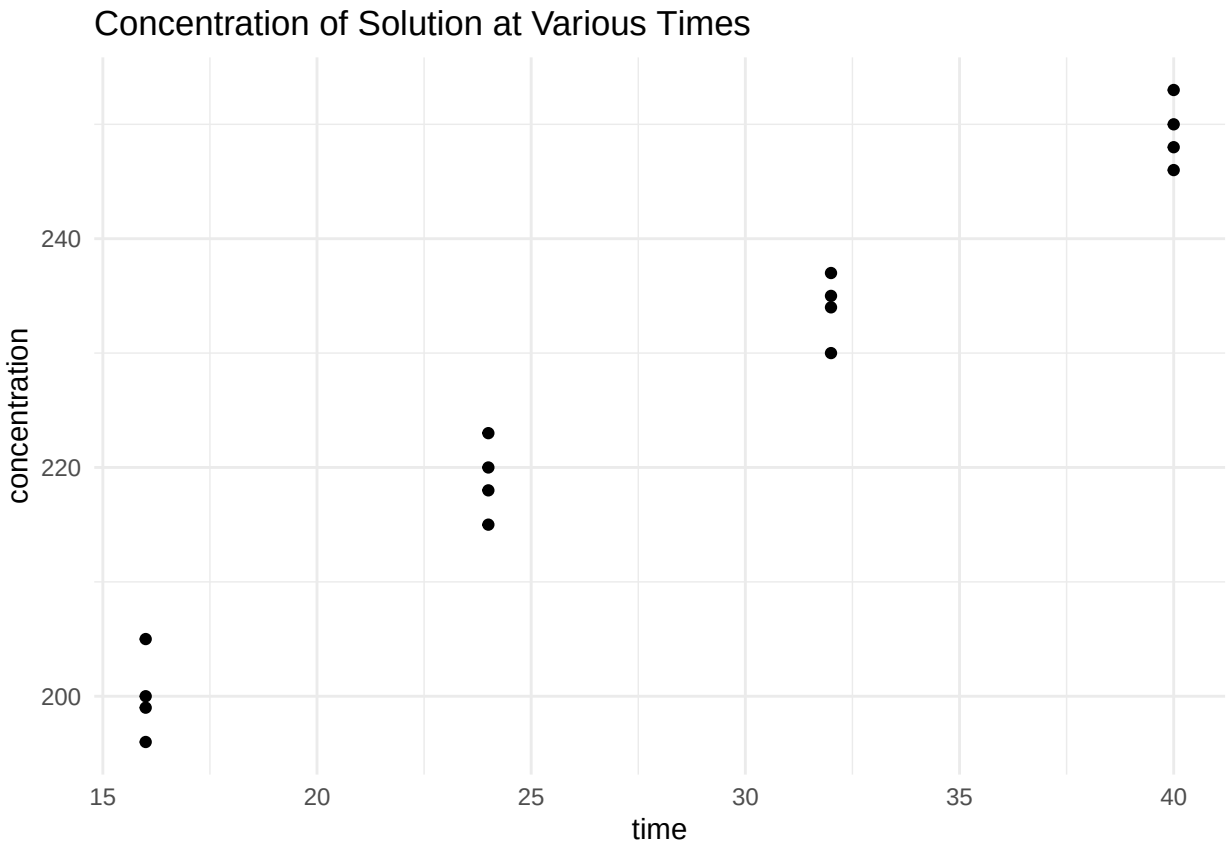
Problem 16:

Refer to Solution concentration Problem 3.15.

(a) Prepare a scatter plot of the data. What transformation of Y might you try, using the prototype patterns in Figure 3.15 to achieve constant variance and linearity?

Answer:

```
plastic %>%
  ggplot(aes(x,y)) +
  geom_point() +
  labs(x = "time", y = "concentration",
       title = "Concentration of Solution at Various Times")
```



To achieve constant variance and linearity, a $\log_{10}$ transformation on Y might be useful.

(b) Use the Box-Cox procedure and standardization (3.36) to find an appropriate power transformation. Evaluate SSE for $\lambda = -.2, -.1, 0, .1, .2$. What transformation of Y is suggested?

Answer:

```
box.cox = function(x, y, lambda){

  sse = c()
  for(i in lambda){
    if(i != 0){
      y_prime = y^i
    }
    else{
      y_prime = log(y)
    }

    k2 = (prod(y))^(1/length(y))
    k1 = 1/(i*k2^(i-1))

    if(i != 0){
      W = k1*(y_prime - 1)
    }
    else{
      W = k2*y_prime
    }

    model = lm.fit_manual(x, W)
```

```
    preds = model[1] + model[2]*x

    sse_temp = sum((W - preds)^2)
    sse = c(sse, sse_temp)
  }
  df = data.frame(lambda = lambda, sse = sse)
  lambda_min_sse = df[df$sse == min(df$sse), "lambda"]
  if(lambda_min_sse == 0){
    print("Using the Box-Cox procedure and strandardization, the most appropriate power transformation
  }
  else{
    paste("Using the Box-Cox procedure and standardizarion, the most appropriate power transformation o
  }

}

box.cox(solutions$x, solutions$y, seq(-.2, .2, length = 5))
```

```
## [1] "Using the Box-Cox procedure and strandardization, the most appropriate power transformation on
```

(c) Use the transformation $Y' = \log_{10} Y$ and obtain the estimated linear regression function for the transformed data.

Answer:

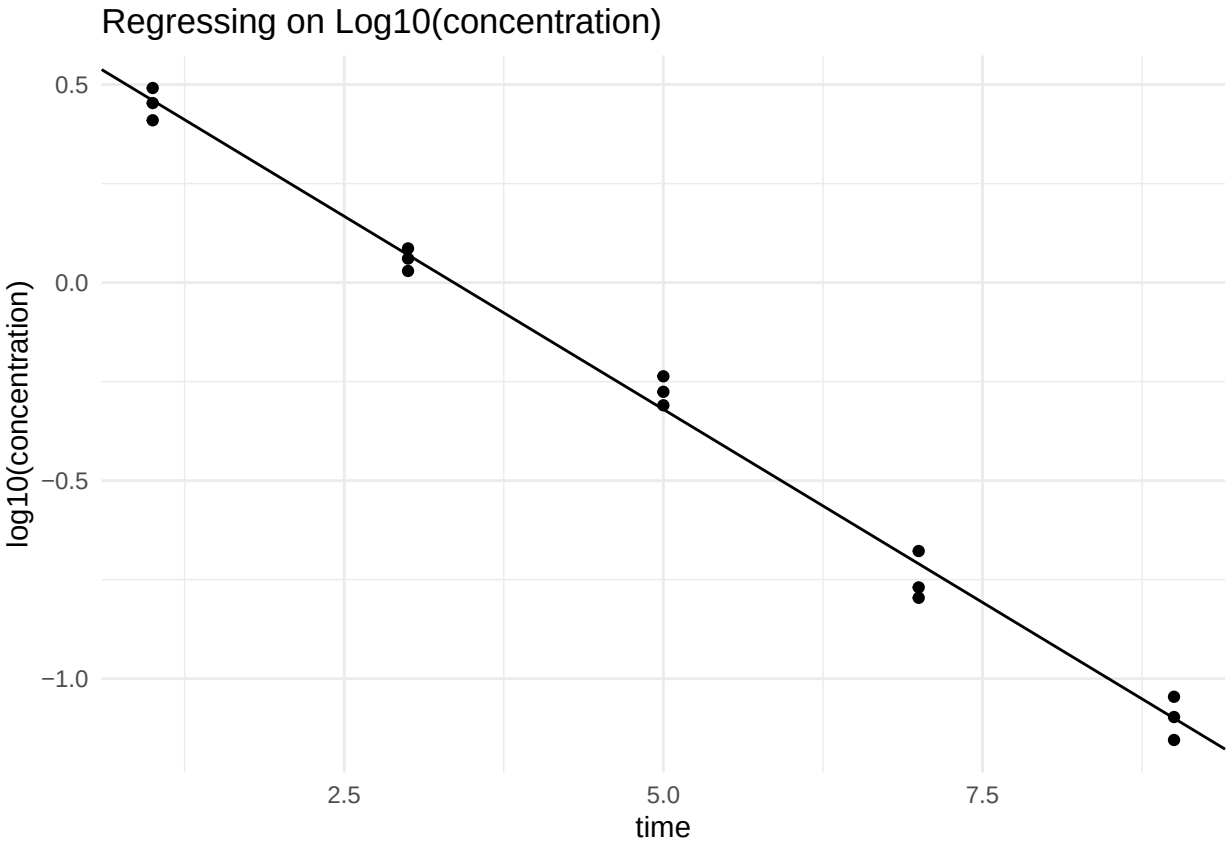
```
Y_prime = log10(solutions$y)
model = round(lm.fit_manual(solutions$x, Y_prime), 3)
paste("b0:", model[1], ", b1:", model[2], sep = ' ')
```

```
## [1] "b0: 0.655 , b1: -0.195"
```

(d) Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

Answer:

```
solutions_transformed = data.frame(x = solutions$x,
                                   y = log10(solutions$y),
                                   y_original = solutions$y,
                                   y_pred = model[1] + model[2]*solutions$x,
                                   resids = model[1] + (model[2]*solutions$x) - log10(solutions$y))

solutions_transformed %>%
  ggplot(aes(x, y)) +
  geom_point() +
  geom_abline(intercept = model[1], slope = model[2]) +
  labs(x = "time",
       y = "log10(concentration)",
       title = "Regressing on Log10(concentration)")
```

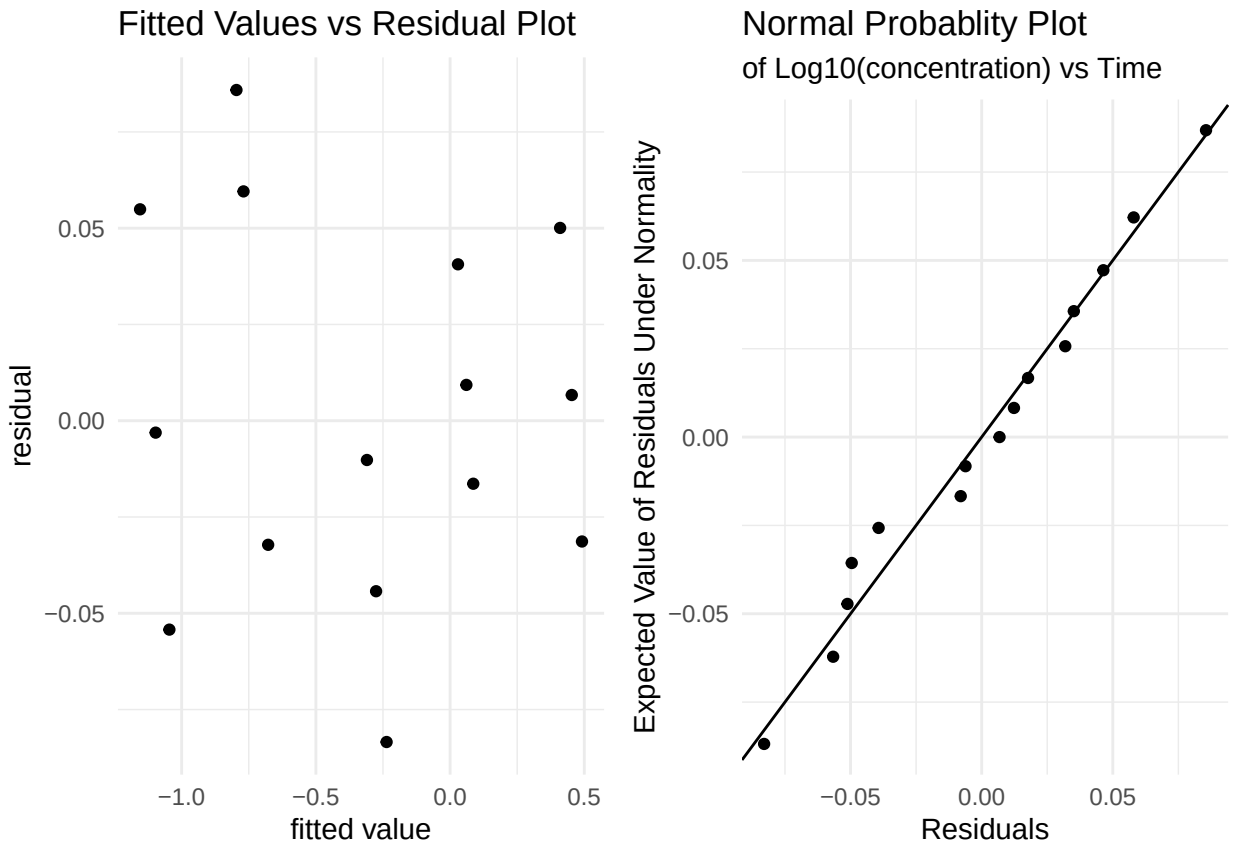


The regression line appears to be a good fit to the transformed data.

(e) Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

Answer:

```
plot_1 = solutions_transformed %>%
  ggplot(aes(x = y, y = resids)) +
  geom_point() +
  labs(x = "fitted value",
       y = "residual",
       title = "Fitted Values vs Residual Plot")
plot_2 = prob_plot(solutions_transformed, "of Log10(concentration) vs Time")

grid.arrange(plot_1, plot_2, nrow = 1)
```



The transformation causes error terms to be distributed normally. The error terms have nonconstant variance.

(f) Express the estimated regression function in the original units.

Answer:

```
paste("Y = concentration = 10^(", model[1], " + ", model[2], "* x)", sep = '')
```

```
## [1] "Y = concentration = 10^(0.655 + -0.195* x)"
```

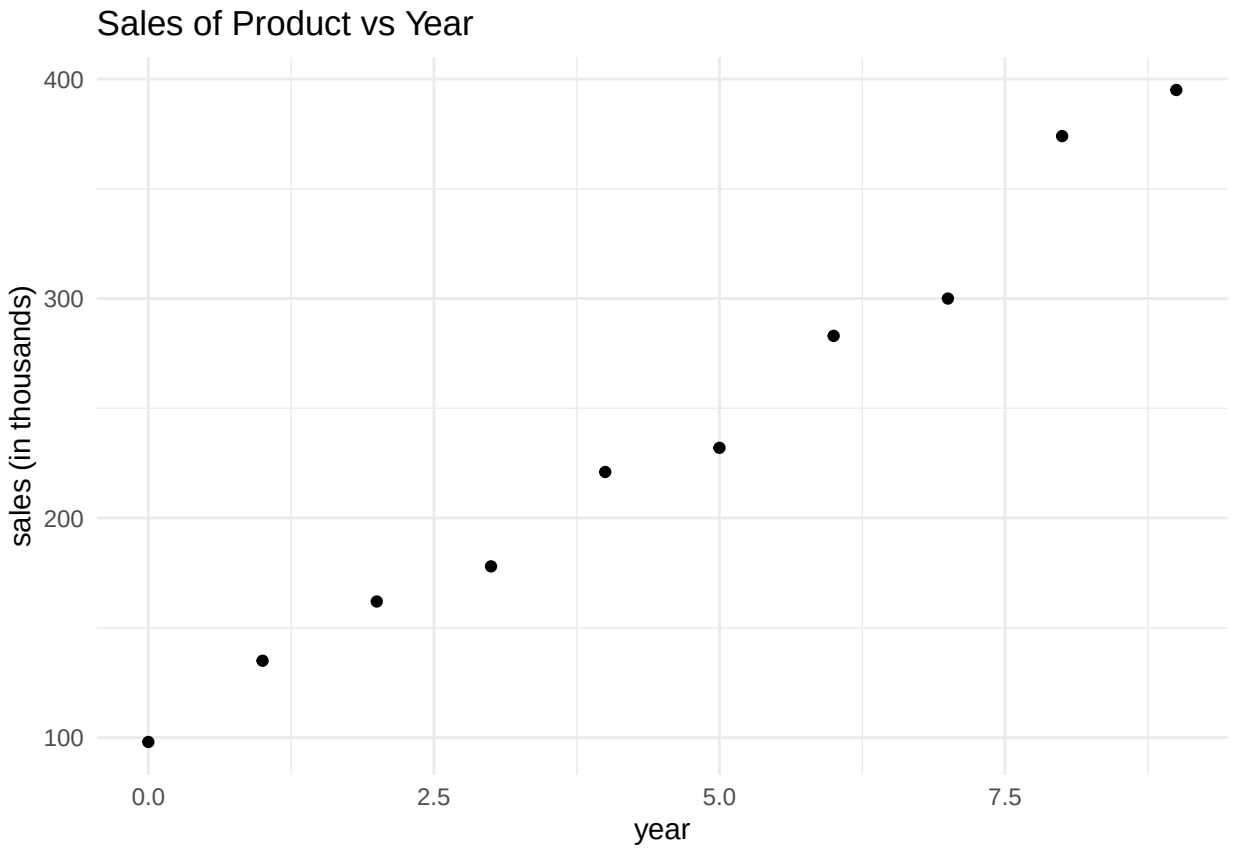
Problem 17:

Sales growth. A marketing researcher studied annual sales of a product that had been introduced 10 years ago. The data is given as follows, where X is the year (coded) and Y is sales in thousands of units.

(a) Prepare a scatter plot of the data. Does a linear relation appear adequate here?

Answer:

```
sales = read.csv('CH03PR17.txt', sep = '', header = FALSE,
                 col.names = c('y', 'x'),
                 colClasses = c('numeric', 'numeric'))

sales %>%
  ggplot(aes(x,y)) +
  geom_point() +
  labs(x = "year",
       y = "sales (in thousands)",
       title = "Sales of Product vs Year")
```



There appears to be a linear relationship between the two variables.

(b) Use the Box-Cox procedure and standardization (3.36) to find an appropriate power transformation of Y. Evaluate SSE for $\lambda = .3, .4, .5, .6, .7$. What transformation of Y is suggested?

Answer:

```
box.cox(sales$x, sales$y, seq(.3, .7, by = .1))
```

```
## [1] "Using the Box-Cox procedure and standardizarion, the most appropriate power transformation on Y
```

(c) Use the transformation $Y' = \sqrt{Y}$ and obtain the estimated linear regression function for the transformed data.

Answer:

```
Y_prime = sqrt(sales$y)
model = round(lm.fit_manual(sales$x, Y_prime), 3)
paste("b0:", model[1], ", b1:", model[2], sep = ' ')
```

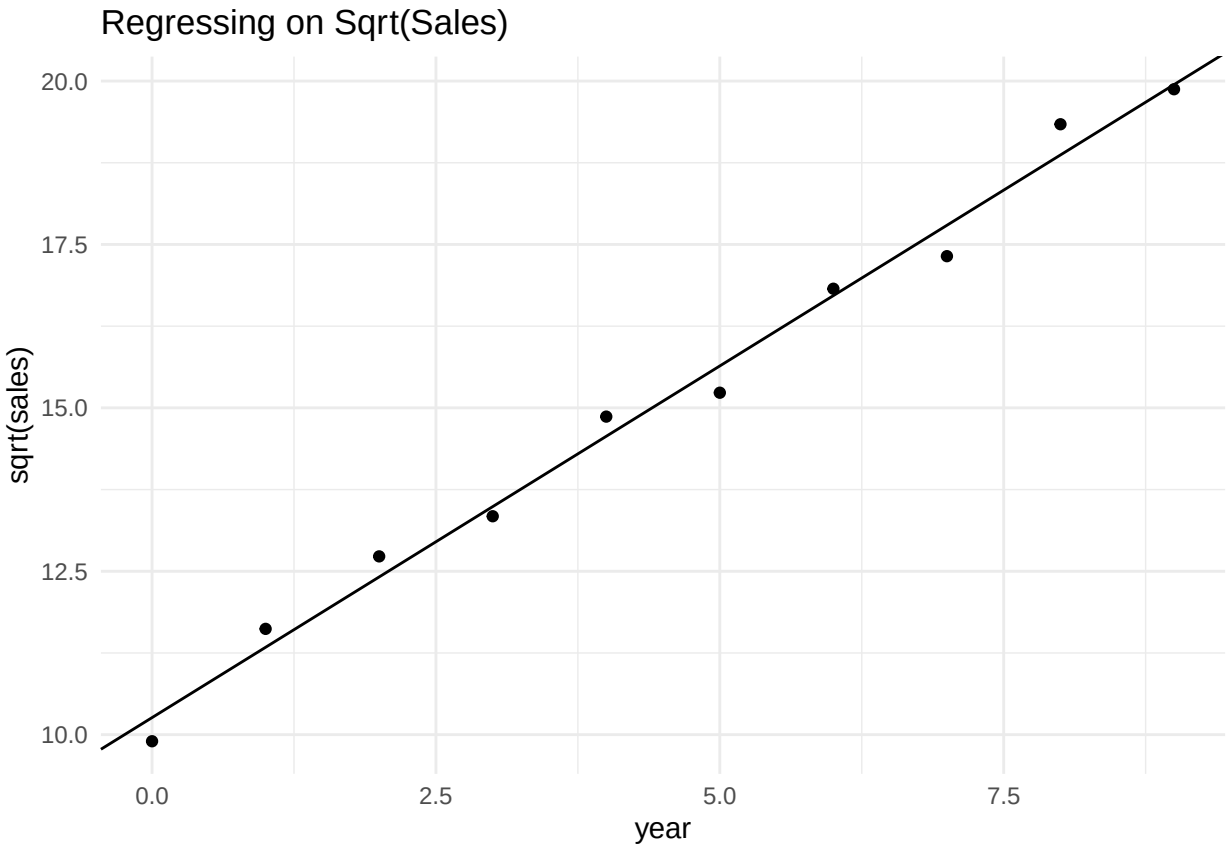
```
## [1] "b0: 10.261 , b1: 1.076"
```

(d) Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

Answer:

```
sales_transformed = data.frame(x = sales$x,
                               y = sqrt(sales$y),
                               y_original = sales$y,
                               y_pred = model[1] + model[2]*sales$x,
                               resids = model[1] + (model[2]*sales$x) - sqrt(sales$y))
```

```r
sales_transformed %>%
  ggplot(aes(x, y)) +
  geom_point() +
  geom_abline(intercept = model[1], slope = model[2]) +
  labs(x = "year",
       y = "sqrt(sales)",
       title = "Regressing on Sqrt(Sales)")
```

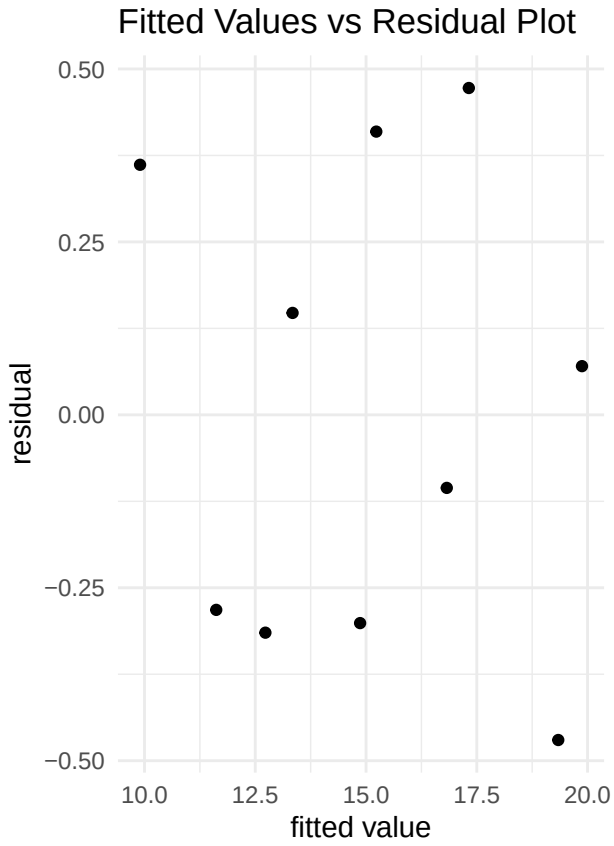


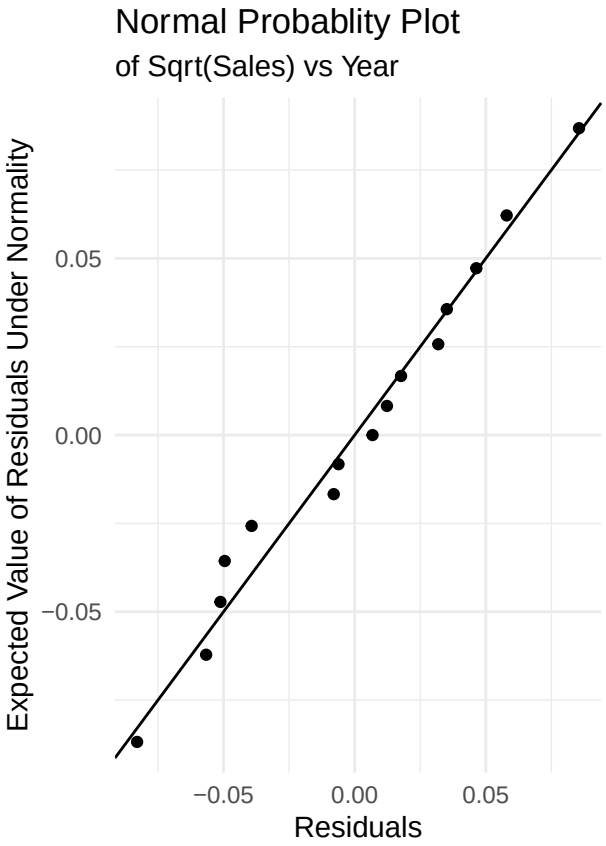
The regression line appears to be a good fit to the transformed data.

(e) Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

Answer:

```r
plot_1 = sales_transformed %>%
  ggplot(aes(x = y, y = resids)) +
  geom_point() +
  labs(x = "fitted value",
       y = "residual",
       title = "Fitted Values vs Residual Plot")
plot_2 = prob_plot(solutions_transformed, "of Sqrt(Sales) vs Year")

grid.arrange(plot_1, plot_2, nrow = 1)
```





The residuals are distributed normally and have a nonconstant variance.

(f) Express the estimated regression function in the original units.

Answer:

```
paste("Y = (", model[1], " + ", model[2], "* x)^2", sep = '')
```

```
## [1] "Y = (10.261 + 1.076* x)^2"
```

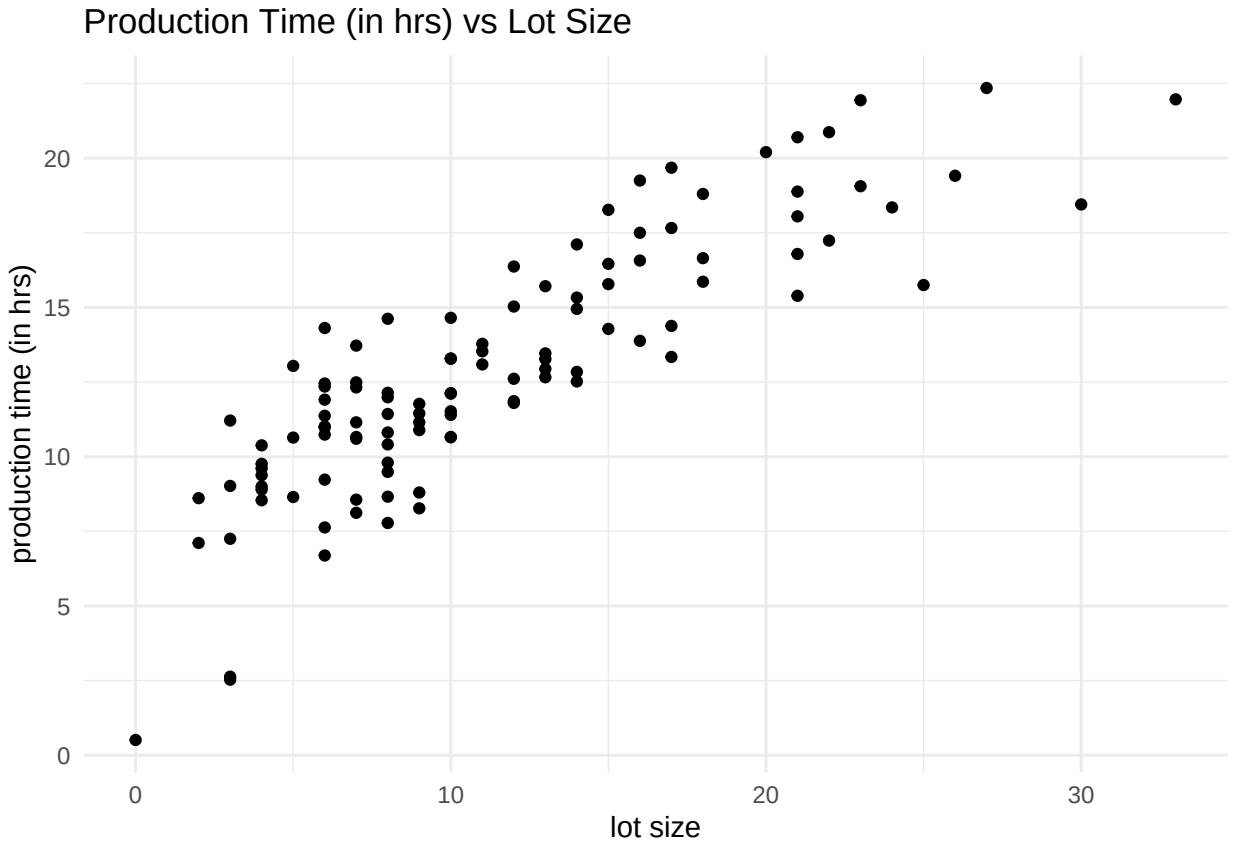
Problem 18:

Production time. In a manufacturing study, the production times for 111 recent production runs were obtained. The table lists for each run the production time in hours (Y) and the production lot size (X).

(a) Prepare a scatter plot of the data. Does a linear relation appear adequate here? Would a transformation on X or Y be more appropriate here? Why?

Answer:

```
prod_time = read.csv('CH03PR18.txt', sep = '', header = FALSE,
                     col.names = c('y', 'x'),
                     colClasses = c('numeric', 'numeric'))

prod_time %>%
  ggplot(aes(x,y)) +
  geom_point() +
  labs(x = "lot size",
       y = "production time (in hrs)",
       title = "Production Time (in hrs) vs Lot Size")
```



There does not appear to exist a linear relation between the two variable. A transformation on the X would be appropriate here since there exists replicates of X in this data set.

(b) Use the transformation $X' = \sqrt{X}$ and obtain the estimated linear regression function for the transformed data.

Answer:

```
prod_time_transformed = data.frame(x = sqrt(prod_time$x), x_original = prod_time$x, y = prod_time$y)

prod_time_transformed = data.frame(x_original = prod_time$x,
                                   x = sqrt(prod_time$x),
                                   y = prod_time$y,
                                   y_pred = lm.fit_manual(sqrt(prod_time$x), prod_time$y)[1] + lm.fit_ma
                                   resids = lm.fit_manual(sqrt(prod_time$x), prod_time$y)[1] + (lm.fit_r

model = round(lm.fit_manual(prod_time_transformed$x, prod_time_transformed$y), 3)
paste("b0:", model[1], ", b1:", model[2], sep = ' ')
```

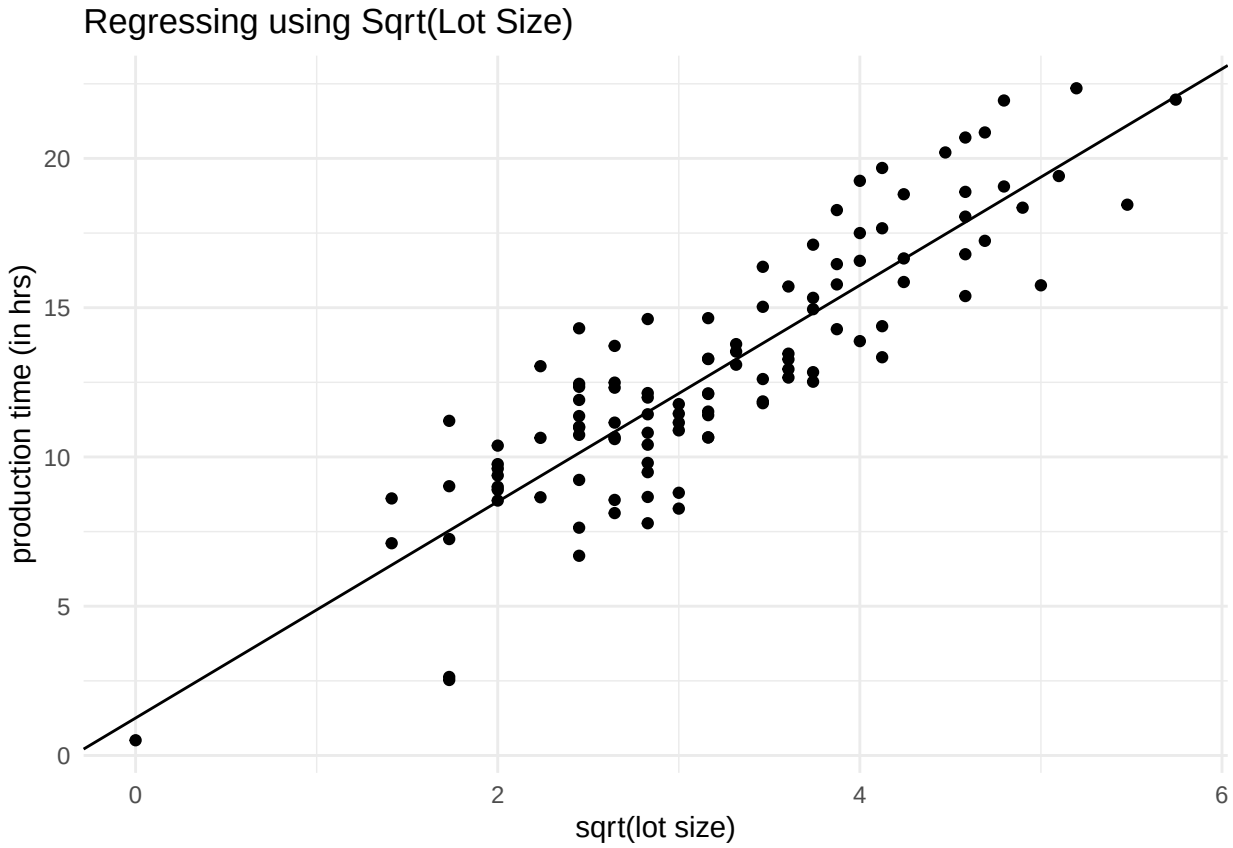
```
## [1] "b0: 1.255 , b1: 3.624"
```

(c) Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

Answer:

```
prod_time_transformed %>%
  ggplot(aes(x, y)) +
  geom_point() +
  geom_abline(intercept = model[1], slope = model[2]) +
```

```
  labs(x = "sqrt(lot size)",
       y = "production time (in hrs)",
       title = "Regressing using Sqrt(Lot Size)")
```

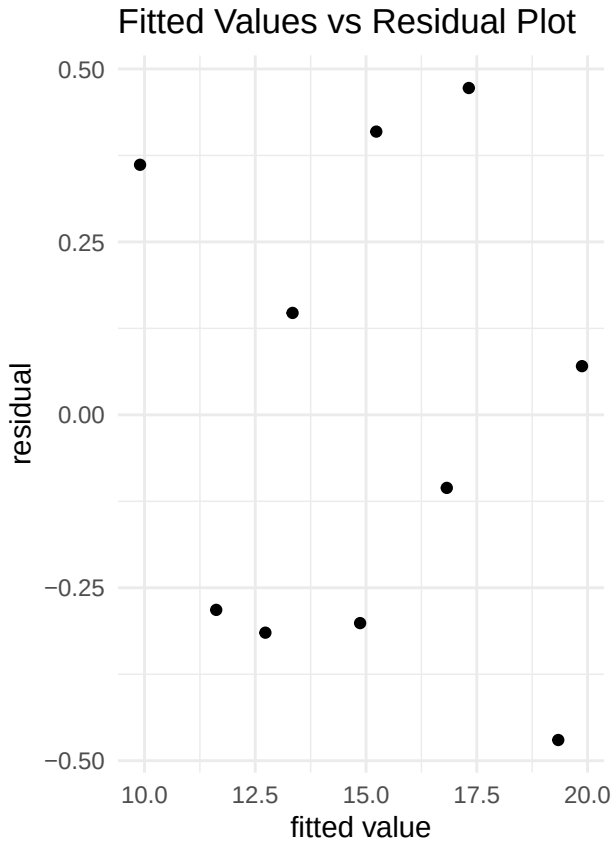


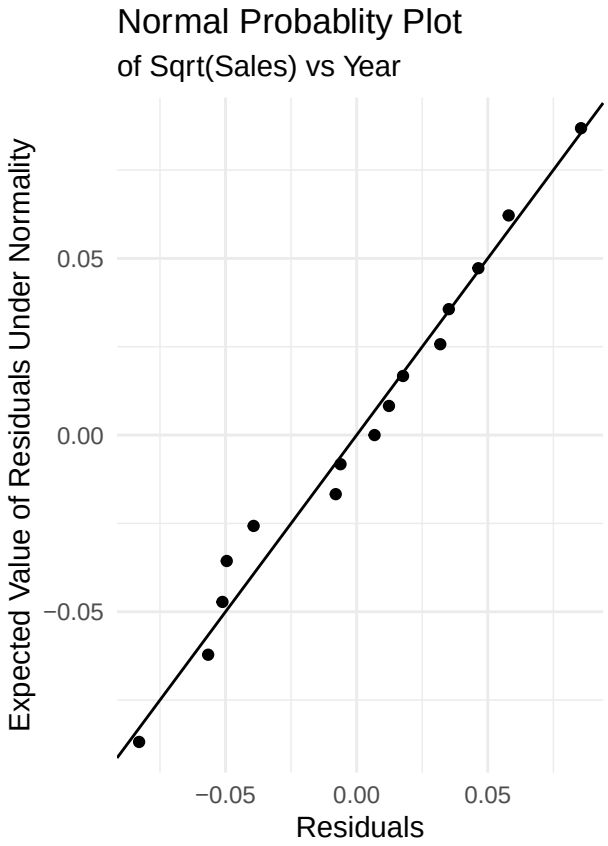
The regression line seems to be an okay fit for the transformed data.

(d) Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

Answer:

```
plot_1 = sales_transformed %>%
  ggplot(aes(x = y, y = resids)) +
  geom_point() +
  labs(x = "fitted value",
       y = "residual",
       title = "Fitted Values vs Residual Plot")
plot_2 = prob_plot(solutions_transformed, "of Sqrt(Sales) vs Year")

grid.arrange(plot_1, plot_2, nrow = 1)
```





The residuals have nonconstant variance and distributed normally.

(e) Express the estimated regression function in the original units.

Answer:

```
paste("Y = ", model[1], " + ", model[2], "sqrt(X)", sep = '')
```

```
## [1] "Y = 1.255 + 3.624sqrt(X)"
```

Problem 19:

A student fitted a linear regression function for a class assignment. The student plotted the residuals e_i against Y_i and found a positive relation. When the residuals were plotted against the fitted values $\hat{Y}_i$, the student found no relation. How could this difference arise? Which is the more meaningful plot?

Answer: Plotting residuals against the original Y values would indicate nothing valuable about the created model. There is always a positive relationship between Y_i and e_i since $e_i = Y_i - \hat{Y}_i$ and because e_i and $\hat{Y}_i$ are independent, their covariance is always positive. The more meaningful plot is e_i against $\hat{Y}_i$.

Problem 20:

If the error terms in a regression model are independent $N(0, \sigma^2)$, what can be said about the error terms after transformation $X' = \frac{1}{X}$ is used? Is the situation the same after transformation $Y' = \frac{1}{Y}$ is used?

Answer: If the error terms in a regression model are independent $N(0, \sigma^2)$, then the transformation $X' = \frac{1}{X}$ will not change the shape of the normal distribution of the error terms, except with a shift in where the mean lies. However if the transformation $Y' = \frac{1}{Y}$ is used, then that can change the distribution of the error terms, since the inverse of a normally distributed random variable is not normally distributed.

Problem 21:

Derive the result in (3.29).

Answer:

$$\underbrace{\sum\sum(Y_{ij}-\hat{Y}_{ij})^2}_{\text{SSE}}=\underbrace{\sum\sum(Y_{ij}-\overline{Y}_j)^2}_{\text{SSPE}}+\underbrace{\sum\sum(\overline{Y}_j-\hat{Y}_{ij})^2}_{\text{SSLF}}$$

The sum of squares can be decomposed into two sums: the full model sum of squares (or pure error sum of squares), and the lack of fit sum of squares. Start from the left and note that $\hat{Y}_{ij}=b_0+b1X_j$ is independent of j.

$$\begin{aligned}
\sum\sum(Y_{ij}-\hat{Y}_{ij})^2 &= \sum\sum[(Y_{ij}-\overline{Y}_j)+(\overline{Y}_j-\hat{Y}_{ij})]^2 \\
&= \sum\sum(Y_{ij}-\overline{Y})^2+\sum\sum(\overline{Y}_j-\hat{Y}_{ij})^2+2\left(\sum\sum(Y_{ij}-\overline{Y}_j)(\overline{Y}_j-\hat{Y}_{ij})\right) \\
&= \sum\sum(Y_{ij}-\overline{Y})^2+\sum\sum(\overline{Y}_j-\hat{Y}_{ij})^2+2\left[\sum\sum Y_{ij}\overline{Y}_j-\sum\sum\overline{Y}_j^2-\sum\sum Y_{ij}\overline{Y}_{ij}+\sum\sum\overline{Y}_j\hat{Y}_{ij}\right. \\
&= \sum\sum(Y_{ij}-\overline{Y})^2+\sum\sum(\overline{Y}_j-\hat{Y}_{ij})^2+2\left[\sum_j n_j\overline{Y}_j^2-\sum_j n_j\overline{Y}_j^2-\sum_j\hat{Y}_{ij}n_j\overline{Y}_j+\sum_j n_j\overline{Y}_j\hat{Y}_{ij}\right] \\
&= \sum\sum(Y_{ij}-\overline{Y})^2+\sum\sum(\overline{Y}_j-\hat{Y}_{ij})^2+0 \\
&= \sum\sum(Y_{ij}-\overline{Y})^2+\sum\sum(\overline{Y}_j-\hat{Y}_{ij})^2
\end{aligned}$$

Problem 22:

Using (A.70), (A.41) and (A.42), show that E[MSPE] $=\sigma^2$ for normal error regression model (2.1).

Answer: Note that, from (A.70), for the sample variance s^2, $\frac{(n-1)s^2}{\sigma^2}$ is distributed as χ^2 with $n-1$ degrees of freedom when the random sample is from a normal population.

MSPE, or pure error mean square, is defined as $\frac{\text{SSPE}}{n-c}$, or pure error sum of squares divided by the sum of the component degrees of freedom. Now, SSPE is made up of the sums od squared deviations at each X level, or $\sum_i(Y_{ij}-\overline{Y}_j)^2$. Therefore, using the above property,

$$\begin{aligned}
\text{E[MSPE]} &= \text{E}\left[\frac{\sum_j\sum_i(Y_{ij}-\overline{Y}_j)^2}{n-c}\right] \\
&= \frac{1}{n-c}\sum_j\text{E}\left[(n_j-1)s_j^2\right] \\
&= \frac{1}{n-c}\sum_j\text{E}[\sigma^2\chi^2(n_j-1)] \\
&= \frac{\sigma^2}{n-c}\sum_j(n_j-1) \\
&= \sigma^2
\end{aligned}$$

Problem 23:

A linear regression model with intercept $\beta_0=0$ is under consideration. Data have been obtained that contain replications. State the full and reduced models for testing the appropriateness of the regression function under consideration. What are the degrees of freedom associated with the full and reduced models if $n=20$ and $c=10$?

Answer: The full model is

$$Y_{ij} = \mu_j + \varepsilon_{ij}$$

and the reduced model is

$$Y_{ij} = \beta_1 X_j + \varepsilon_{ij}$$

When $n = 20$ and $c = 10$, the degrees of freedom are as follows:

$$df_F = n - c = 20 - 10 = 10$$
$$df_R = n - 1 = 20 - 1 = 19$$

Problem 24:

Blood pressure. Data was obtained in a study of the relation between diastolic blood pressure (Y) and age (X) for boys 5 to 13 years old.

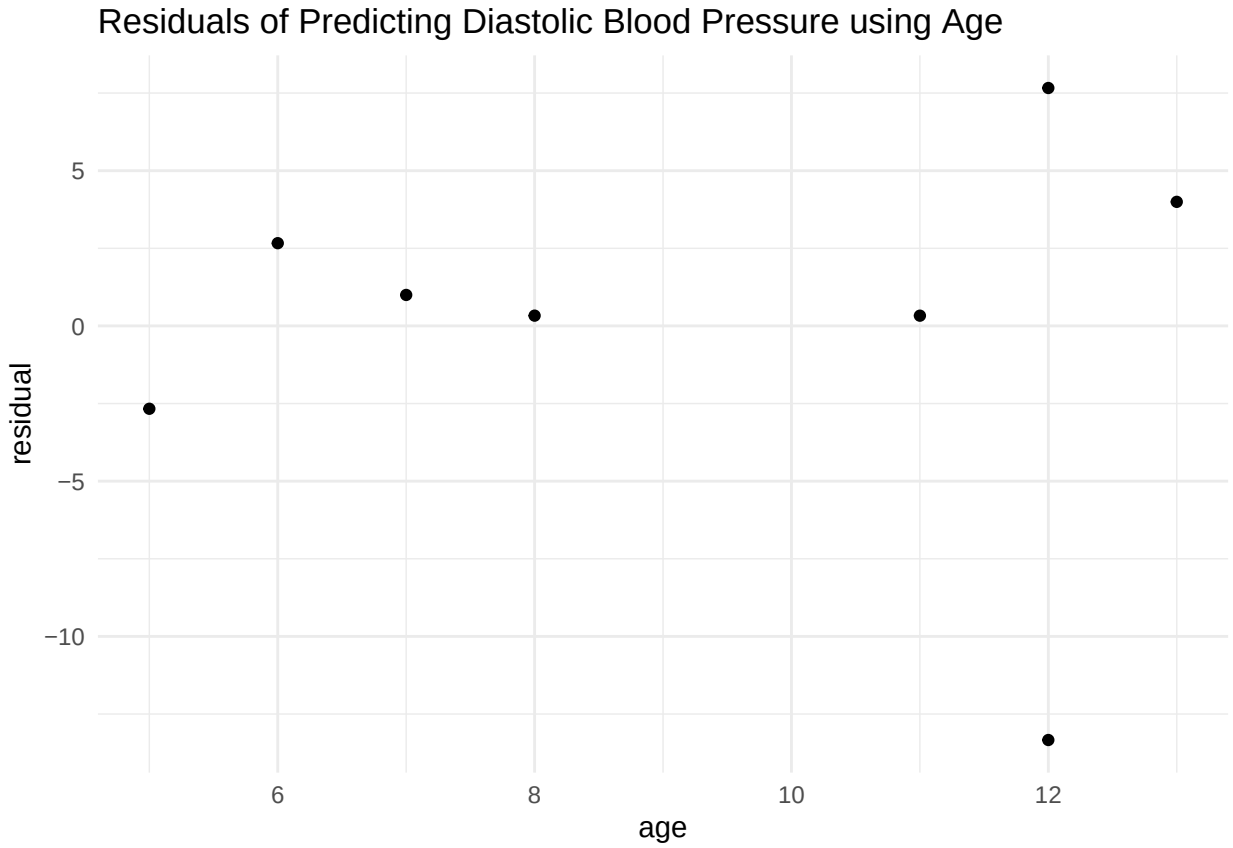
(a) Assuming normal error regression model (2.1) is appropriate, obtain the estimated regression function and plot the residuals e_i against X_i. What does your residual plot show?

Answer:

```
blood_pressure = read.csv('CH03PR24.txt', sep = '', header = FALSE,
                  col.names = c('y', 'x'),
                  colClasses = c('numeric', 'numeric'))

model = round(lm.fit_manual(blood_pressure$x, blood_pressure$y), 3)
paste("b0:", model[1], "b1:", model[2], sep = ' ')
```

```
## [1] "b0: 48.667 b1: 2.333"
```

```
Y_pred = model[1] + model[2]*blood_pressure$x
resid = Y_pred - blood_pressure$y

data.frame(resid, x = blood_pressure$x) %>%
  ggplot(aes(x, resid)) +
  geom_point() +
  labs(x = "age",
       y = "residual",
       title = "Residuals of Predicting Diastolic Blood Pressure using Age")
```



The residual plot shows that there is an outlier at age 12, which makes the distribution of error terms to have nonconstant variance.

(b) Omit case 7 from the data and obtain the estimated regression function based on the remaining seven cases. Compare this estimated regression function to that obtained in part (a). What can you conclude about the effect of case 7?

Answer:

```
blood_pressure_new = blood_pressure[c(1:6, 8:nrow(blood_pressure)),]
model = round(lm.fit_manual(blood_pressure_new$x, blood_pressure_new$y), 3)
paste("b0:", model[1], "b1:", model[2], sep = ' ')
```

```
## [1] "b0: 53.068 b1: 1.621"
```

By omitting case 7, the y intercept of the regression function is higher and the slope is less steep. The effect of case 7 was bringing the regression function down.

(c) Using your fitted regression function in part (b), obtain a 99 percent prediction interval for a new Y observation at $X = 12$. Does observation Y_7 fall outside this prediction interval? What is the significance of this?

Answer:

```
Yhat_pred_interval = function(data, Xhat, alpha = 0.01){

  X = data[['x']]
  Y = data[['y']]
  df = nrow(data) - 2
  b1 = sum((X - mean(X))*(Y - mean(Y))) / (sum((X - mean(X))^2))
  b0 = mean(Y) - b1*mean(X)
```

```
  Ypred = b0 + b1*X
  Yhat = b0 + b1*Xhat
  mse = sum((Y - Ypred)^2) / df
  var_adjust = (1/nrow(data)) + ((Xhat - mean(X))^2)/sum((X - mean(X))^2) + 1
  t = qt(1 - (alpha/2), df = df)
  se = sqrt(mse * var_adjust)
  error = t*se
  lower = round(Yhat - error, 5)
  upper = round(Yhat + error, 5)
  return(c(lower, upper))
}

Yhat_pred_interval(blood_pressure_new, Xhat = 12, alpha = 0.01)
```

```
## [1] 60.31266 84.73588
```

```
blood_pressure[7,]
```

```
##    y  x
## 7 90 12
```

Observation Y_7 falls outside this prediction interval. This signifies that the observation is an outlier from the collection of the data.

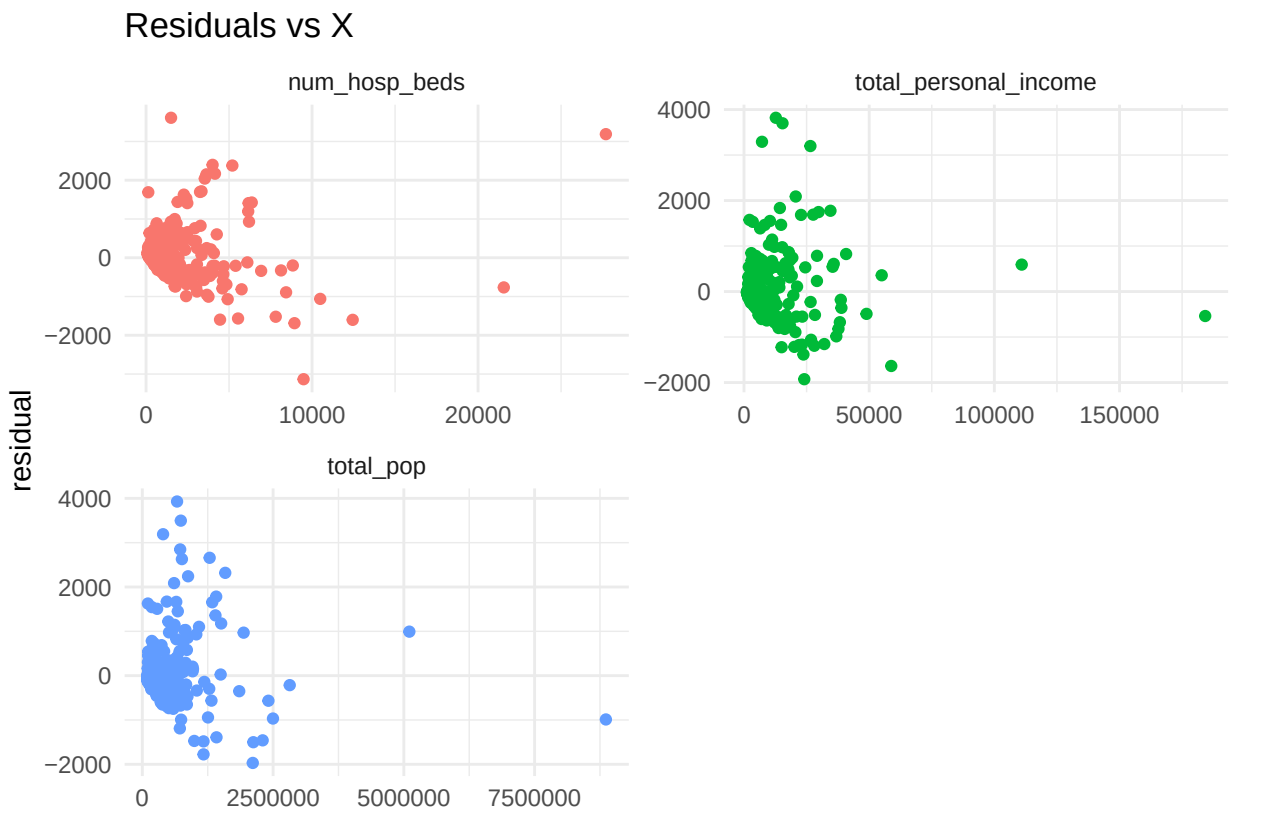
Problem 25:

Refer to the CDI data set in Appendix C.2 and Project 1.43. For each of the three fitted regression models, obtain the residuals and prepare a residual plot against X and a normal probability plot. Summarize your conclusions. Is linear regression model (2.1) more appropriate in one case than in the others?

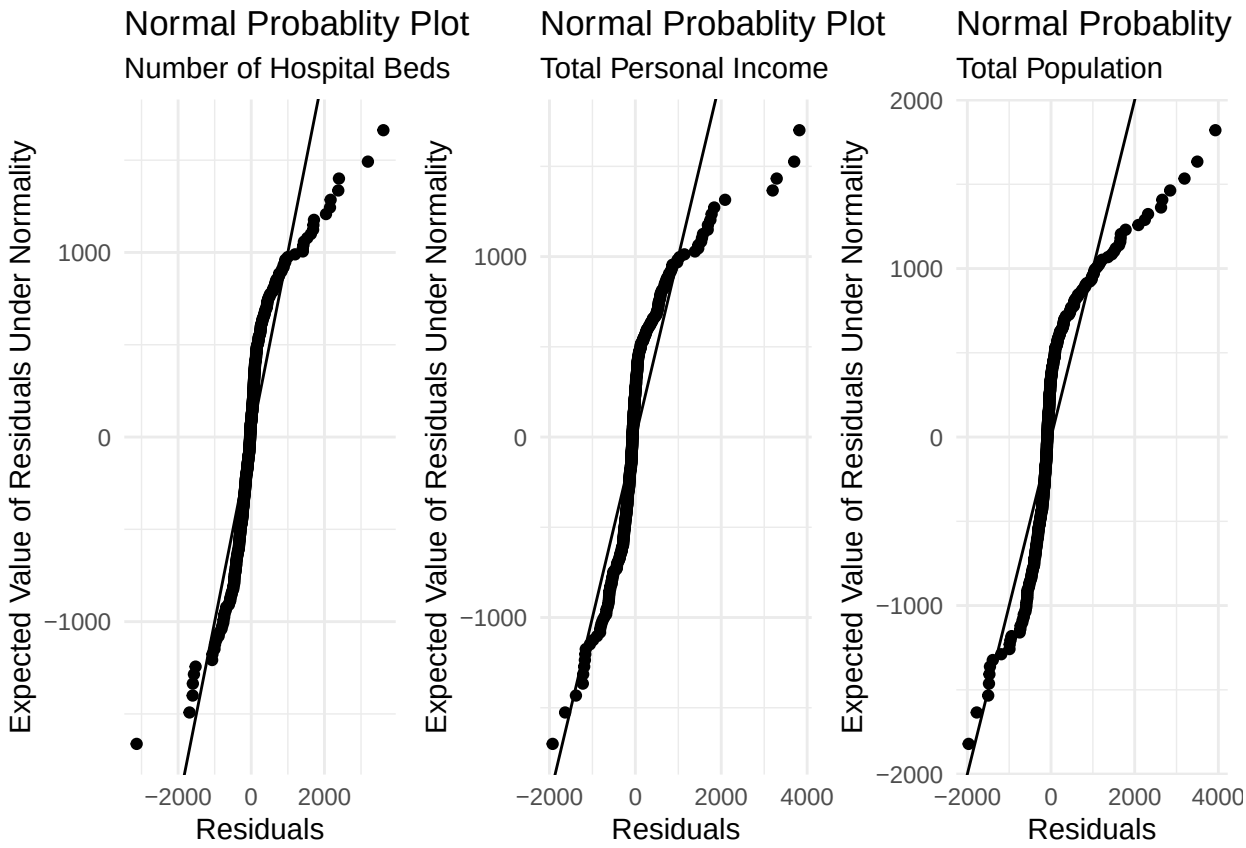
Answer:

```
cdi_cols = c('ID', 'county', 'state', 'area', 'total_pop', 'perc_pop_18to34',
             'perc_pop_65plus', 'num_physicians', 'num_hosp_beds', 'total_crimes',
             'perc_hs_grads', 'perc_bach', 'perc_below_poverty', 'perc_unemploy',
             'per_capita_income', 'total_personal_income', 'geographic_region')
cdi_colclasses = c('integer', 'character', 'character', rep('numeric', 13), 'factor')
cdi = read.csv('APPENC02.txt', sep = '', header = FALSE,
                  col.names = cdi_cols,
                  colClasses = cdi_colclasses)
cdi_43 = cdi %>%
  select(num_physicians, total_pop,
         num_hosp_beds, total_personal_income)
plot_1 = cdi_43 %>%
  gather(type, x, total_pop, num_hosp_beds, total_personal_income) %>%
  group_by(type) %>%
  do(., augment(lm(num_physicians ~ x, data = .))) %>%
  ggplot(aes(x = x, y = .resid, color = type)) +
  geom_point() +
  facet_wrap(type ~ ., ncol = 2, scales = "free") +
  theme(legend.position = '') +
  labs(x = '',
       y = "residual",
       title = "Residuals vs X")
plot_2 = prob_plot(select(cdi_43, x = num_hosp_beds, y = num_physicians), "Number of Hospital Beds")
plot_3 = prob_plot(select(cdi_43, x = total_personal_income, y = num_physicians), "Total Personal Income
```

```
plot_4 = prob_plot(select(cdi_43, x = total_pop, y = num_physicians), "Total Population")
plot_1
```




```
grid.arrange(plot_2, plot_3, plot_4, ncol = 3)
```



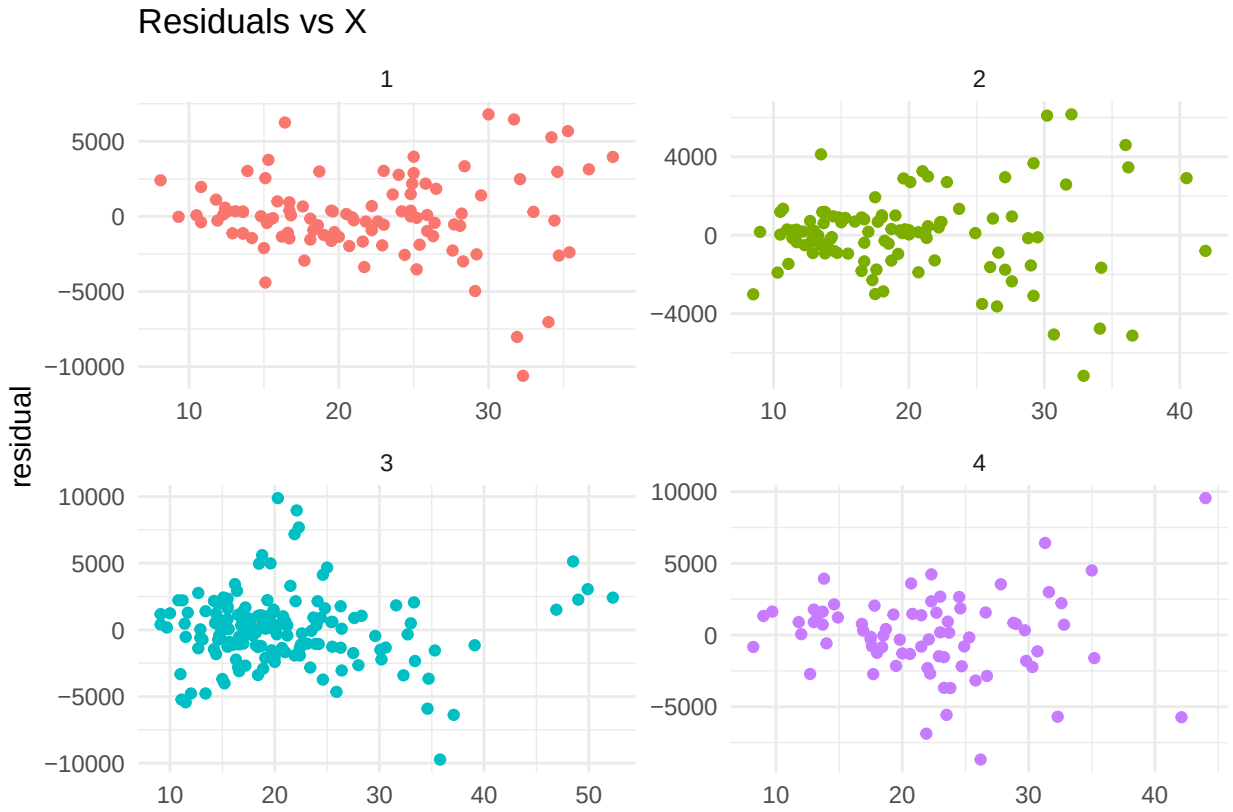
Error terms from all 3 regression models have a constant variance and are distributed normally, with heavy tails at the higher end. The linear regression model is appropriate when using number of hospital beds, as the model diagnostics are most normal in use.

Problem 26:

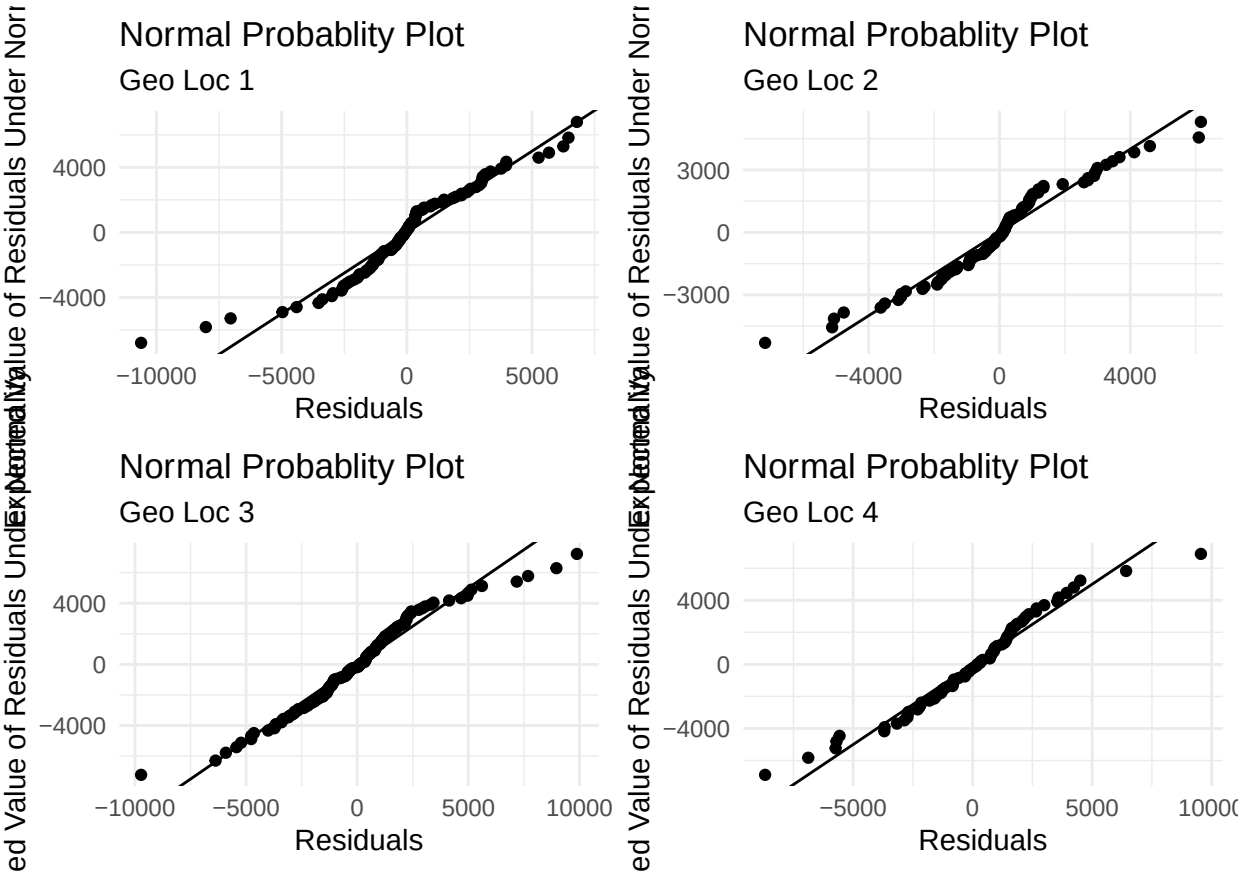
Refer to the CDI data set in Appendix C.2 and Project 1.44. For each geographic region, obtain the residuals and prepare a residual plot against X and a normal probability plot. Do the four regions appear to have similar error variances? What other conclusions do you draw from your plots?

Answer:

```
cdi %>% select(per_capita_income, perc_bach, geographic_region) %>%
  group_by(geographic_region) %>%
  do(., augment(lm(per_capita_income ~ perc_bach, data = .))) %>%
  ggplot(aes(x = perc_bach, y = .resid, color = factor(geographic_region))) +
  geom_point() +
  facet_wrap(geographic_region ~ ., ncol = 2, scales = "free") +
  theme(legend.position = '') +
  labs(x = '',
       y = "residual",
       title = "Residuals vs X")
```




```
plot_1 = prob_plot(select(filter(cdi, geographic_region == 1), x = perc_bach, y = per_capita_income), "
plot_2 = prob_plot(select(filter(cdi, geographic_region == 2), x = perc_bach, y = per_capita_income), "
plot_3 = prob_plot(select(filter(cdi, geographic_region == 3), x = perc_bach, y = per_capita_income), "
plot_4 = prob_plot(select(filter(cdi, geographic_region == 4), x = perc_bach, y = per_capita_income), "

grid.arrange(plot_1, plot_2, plot_3, plot_4, ncol = 2)
```



There is some occurrence of higher variances for high values of X in most of the regression plots. Thus the four regions do not appear to have similar error variances. However residuals are distributed normally in all 4 regression models.

Problem 27:

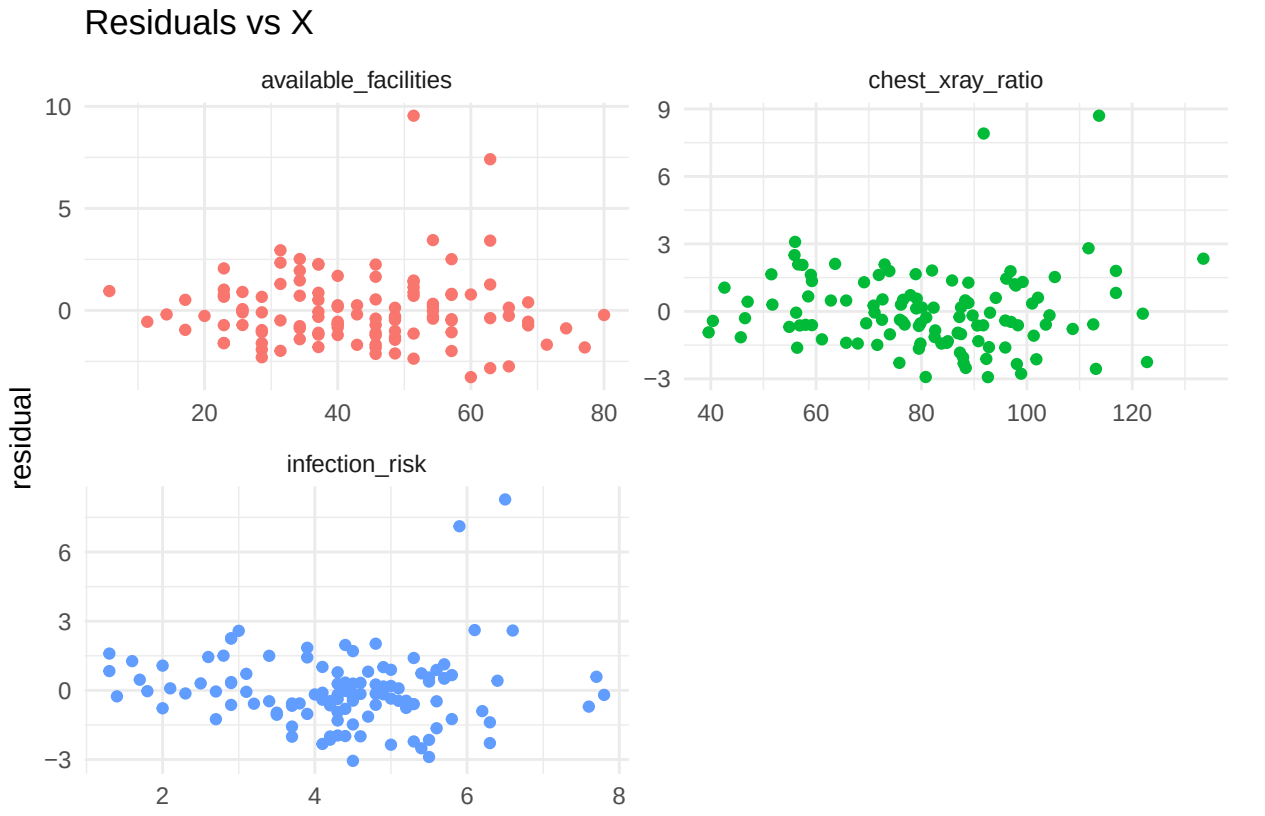
Refer to the SENIC data set in Appendix C.1 and Project 1.45.

(a) For each of the three fitted regression models, obtain the residuals and prepare a residual plot against X and a normal probability plot. Summarize your conclusions. Is linear regression model (2.1) more apt in one case than in the others?

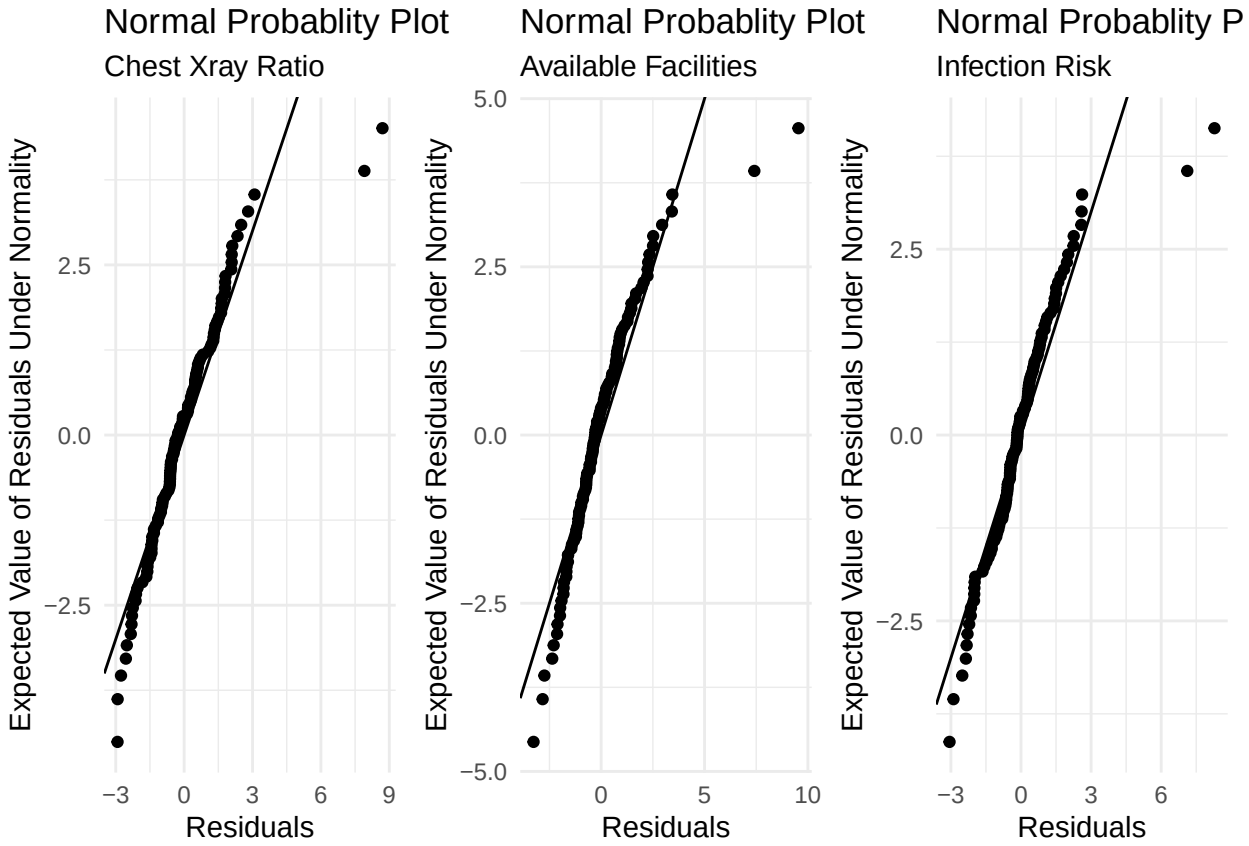
Answer:

```
senic_cols = c('ID', 'length_stay', 'age', 'infection_risk', 'culturing_ratio',
               'chest_xray_ratio', 'num_beds', 'med_school_aff', 'region',
               'avg_daily_census', 'num_nurses', 'available_facilities')
senic_colclasses = c(rep('numeric', 6), rep('factor', 2), rep('numeric', 3))
senic = read.csv('APPENC01.txt', sep = '', header = FALSE,
                 col.names = senic_cols,
                 colClasses = senic_colclasses)
senic_model = senic %>%
  select(length_stay, infection_risk, available_facilities, chest_xray_ratio)

plot_1 = senic_model %>%
  gather(type, x, infection_risk, available_facilities, chest_xray_ratio) %>%
  group_by(type) %>%
  do(., augment(lm(length_stay ~ x, data = .))) %>%
```

```
  ggplot(aes(x = x, y = .resid, color = type)) +
  geom_point() +
  facet_wrap(type ~ ., ncol = 2, scales = "free") +
  theme(legend.position = '') +
  labs(x = '',
       y = "residual",
       title = "Residuals vs X")
plot_2 = prob_plot(select(senic_model, x = chest_xray_ratio, y = length_stay), "Chest Xray Ratio")
plot_3 = prob_plot(select(senic_model, x = available_facilities, y = length_stay), "Available Facilitie
plot_4 = prob_plot(select(senic_model, x = infection_risk, y = length_stay), "Infection Risk")
plot_1
```




```
grid.arrange(plot_2, plot_3, plot_4, ncol = 3)
```



In all 3 regression models, error terms show a constant variance and are distributed normally, with slightly large tails at the high end. The linear regression model is apt in all 3 cases.

(b) Obtain the fitted regression function for the relation between length of stay and infection risk after deleting cases 47 ($X_{47} = 6.5$, $Y_{47} = 19.56$) and 112 ($X_{112} = 5.9$ and $Y_{112} = 17.94$). From this fitted regression function obtain separate 95 percent prediction intervals for new Y observations at $X = 6.5$ and $X = 5.9$, respectively. Do observations Y_{47} and Y_{112} fall outside these prediction intervals? Discuss the significance of this.

Answer:

```
senic_model_2 = senic %>%
  filter(!ID %in% c(47, 112)) %>%
  select(y = length_stay, x = infection_risk)

Yhat_pred_interval(senic_model_2, 6.5, 0.05)
```

```
## [1]  8.31863 13.30654
```

```
senic[47, c("ID", "length_stay", "infection_risk")]
```

```
##    ID length_stay infection_risk
## 47 47       19.56            6.5
```

```
Yhat_pred_interval(senic_model_2, 5.8, 0.05)
```

```
## [1]  7.90776 12.86377
```

```
senic[112, c("ID", "length_stay", "infection_risk")]
```

```
##      ID length_stay infection_risk
```

```
## 112 112       17.94             5.9
```

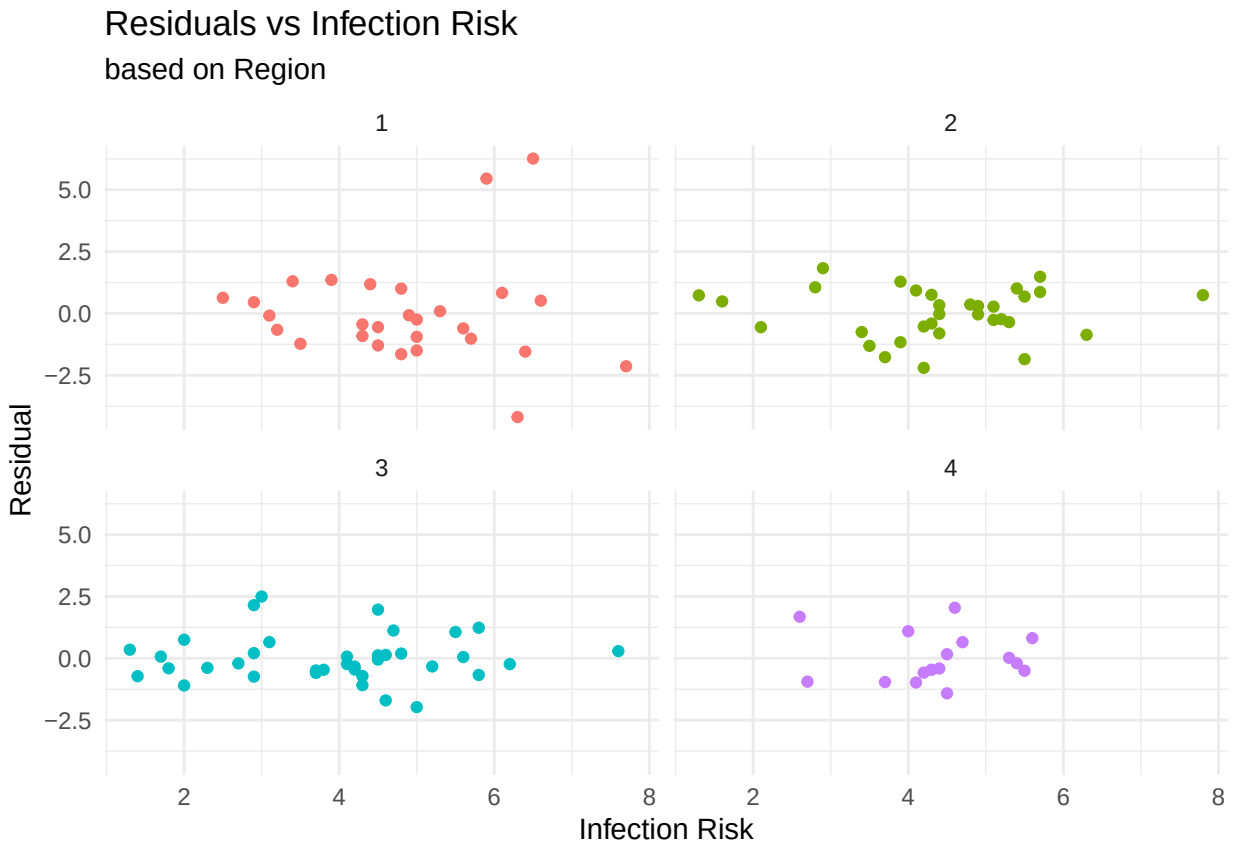
Observations Y_{47} and Y_{112} both fall outside these prediction intervals, indicating that they are outliers in the data collection.

Problem 28:

Refer to the SENIC data set in Appendix C.1 and Project 1.46. For each geographic region, obtain the residuals and prepare a residual plot against X and a normal probability plot. Do the four regions appear to have similar error variance? What other conclusions do you draw from your plots?

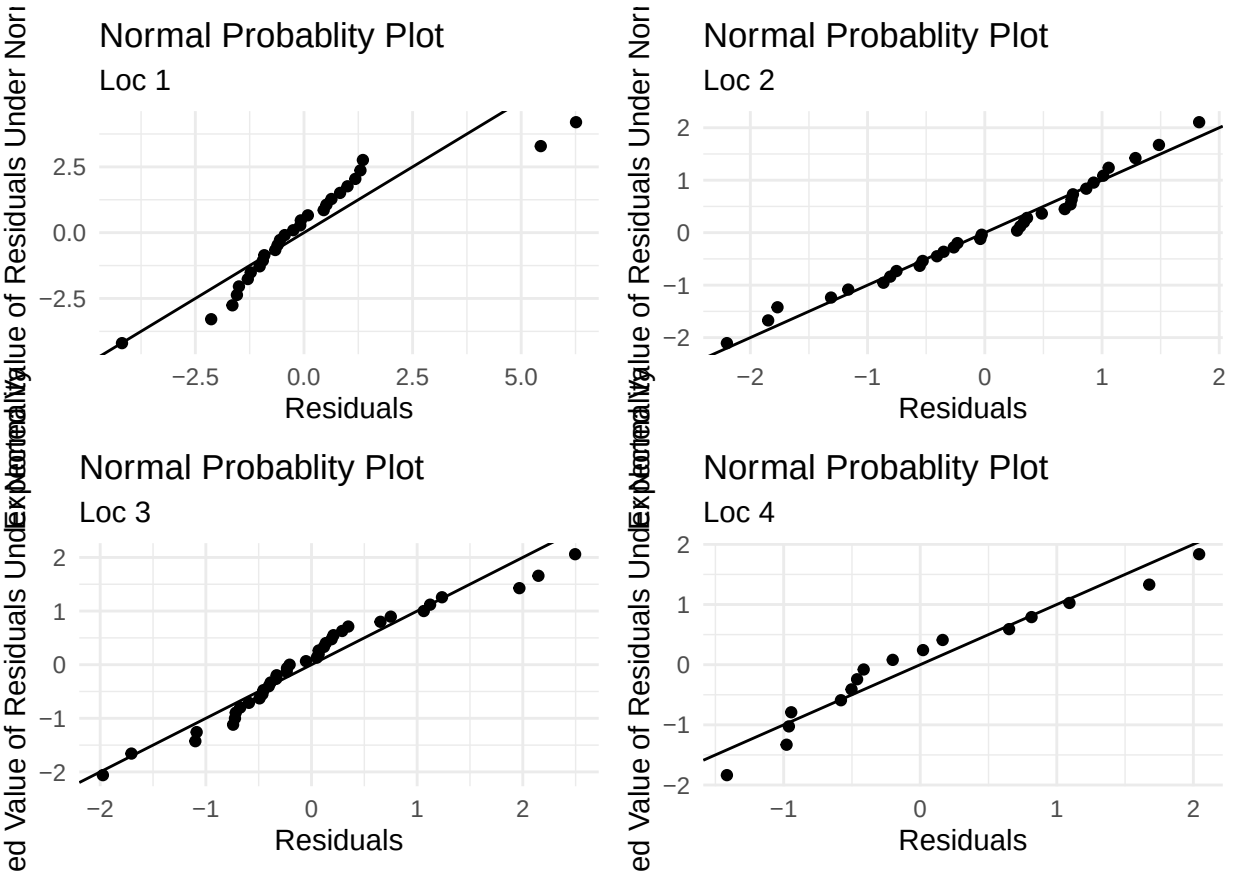
Answer:

```
senic %>%
  select(length_stay, infection_risk, region) %>%
  group_by(region) %>%
  do(., augment(lm(length_stay ~ infection_risk, data = .))) %>%
  ggplot(aes(x = infection_risk, y = .resid, color = factor(region))) +
  geom_point() +
  facet_wrap(region ~ ., ncol = 2) +
  labs(x = "Infection Risk",
       y = "Residual",
       title = "Residuals vs Infection Risk",
       subtitle = "based on Region") +
  theme(legend.position = '')
```




```
plot_1 = prob_plot(select(filter(senic, region == 1), x = infection_risk, y = length_stay), "Loc 1")
plot_2 = prob_plot(select(filter(senic, region == 2), x = infection_risk, y = length_stay), "Loc 2")
plot_3 = prob_plot(select(filter(senic, region == 3), x = infection_risk, y = length_stay), "Loc 3")
```

```
plot_4 = prob_plot(select(filter(senic, region == 4), x = infection_risk, y = length_stay), "Loc 4")

grid.arrange(plot_1, plot_2, plot_3, plot_4, ncol = 2)
```



The error variances appear to be constant and similar for regression models for regions 2, 3 and 4. Error variances may be higher for region 1. Error terms are distributed normally for region 2, 3 and 4. Region 1 error terms are distributed with high tails at both ends.

Problem 29:

Refer to Copier maintenance Problem 1.20.

(a) Divide the data into four bands according to the number of copiers serviced (X). Band 1 ranges from $X = .5$ to $X = 2.5$; band 2 ranges from $X = 2.5$ to $X = 4.5$; and so forth. Determine the median value of X and the median value of Y in each of the bands and develop the band smooth by connecting the four pairs of medians by straight lines on a scatter plot of the data. Does the band smooth suggest the regression relation is linear? Discuss.

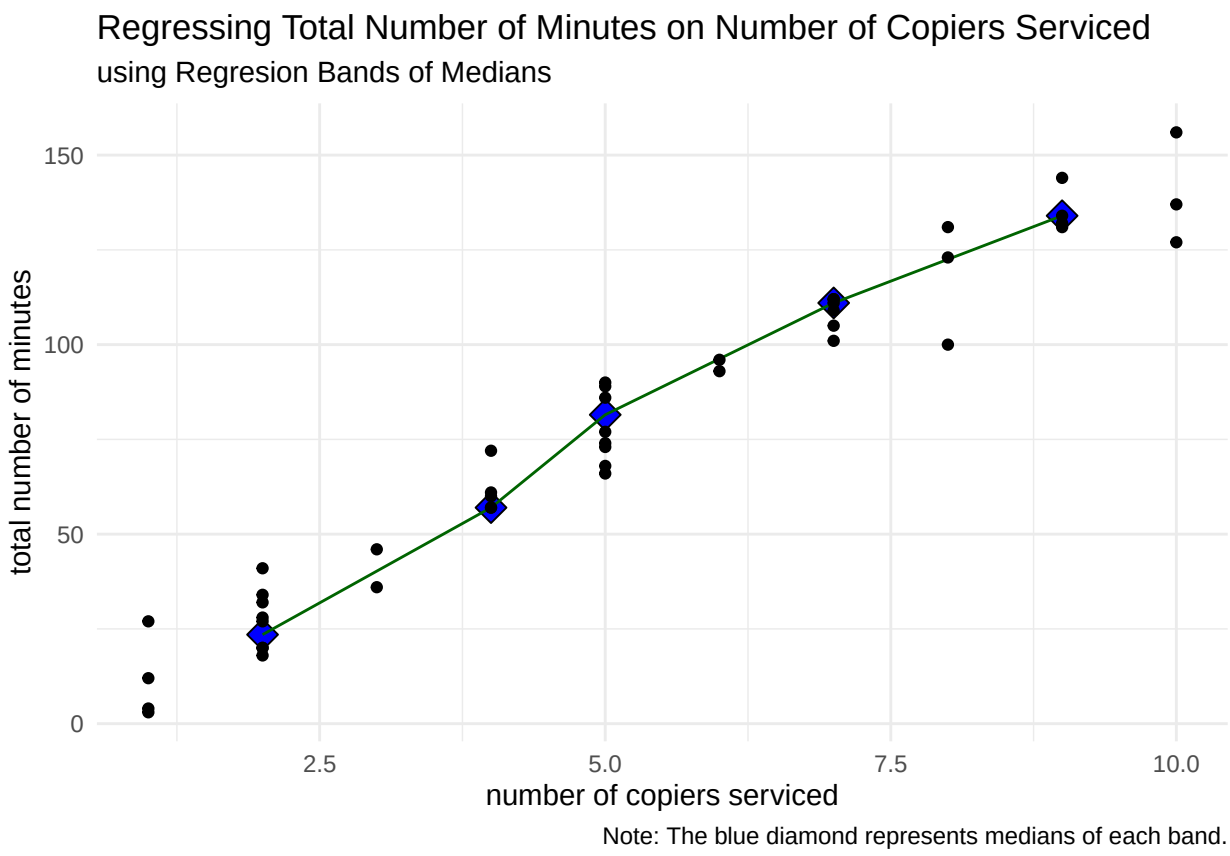
Answer:

```
plot_1 = copier %>%
  mutate(band = case_when(
    x <= 2.5 ~ 1,
    x > 2.5 & x <= 4.5 ~ 2,
    x > 4.5 & x <= 6.5 ~ 3,
    x > 6.5 & x <= 8.5 ~ 4,
    x > 8.5 ~ 5)) %>%
  group_by(band) %>%
```

```
  summarize(x_med = median(x, na.rm = TRUE),
            y_med = median(y, na.rm = TRUE)) %>%
  ggplot(aes(x_med, y_med)) +
  geom_point(shape=23, fill="blue", size=4) +
  geom_line(color = "darkgreen") +
  geom_point(data = copier, aes(x,y)) +
  labs(x = "number of copiers serviced",
       y = "total number of minutes",
       title = "Regressing Total Number of Minutes on Number of Copiers Serviced",
       subtitle = "using Regresion Bands of Medians",
       caption = "Note: The blue diamond represents medians of each band.")
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
plot_1
```



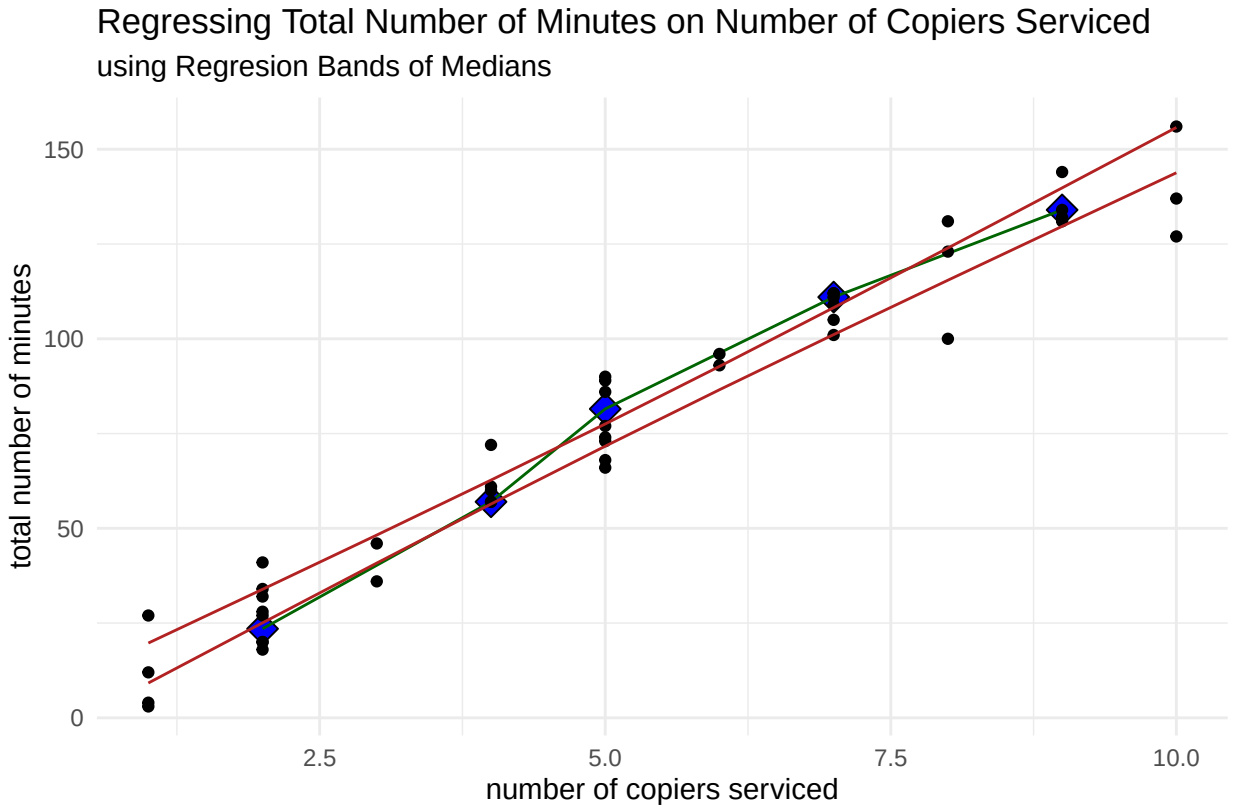
The smooth band suggests that the regression relation is not linear. This is because the medians in each of the bands do not appear to match up on a straight line.

(b) Obtain the 90 percent confidence band for the true regression line and plot it on the scatter plot prepared in part (a). Does the band smooth fall entirely inside the confidence band? What does this tell you about the appropriateness of the linear regression function?

Answer:

```
conf_band_reg = function(data, alpha = 0.05){

  model = lm.fit_manual(data$x, data$y)
  Y_hat = model[1] + model[2]*data$x
```

```
  mse = sum((Y_hat - data$y)^2)/(nrow(data)-2)
  X_bar = mean(data$x)
  X_iX_bar_diff_sum_sq = sum((data$x - X_bar)^2)

  X_h = sort(unique(data$x))
  Y_hat_sd = c()
  for(i in X_h){
    temp_sd = sqrt(mse * ((1/nrow(data)) + ((i - X_bar)^2/X_iX_bar_diff_sum_sq)))
    Y_hat_sd = c(Y_hat_sd, temp_sd)
  }

  W_squared = 2*qf(1 - alpha, 2, nrow(data)-2)
  W = sqrt(W_squared)

  conf_band_df = data.frame(x = X_h,
                            y = model[1] + model[2]*X_h,
                            W = rep(W, length(X_h)),
                            s_Yhat = Y_hat_sd)

  conf_band_df$lower_bound = conf_band_df$y - (conf_band_df$W*conf_band_df$s_Yhat)
  conf_band_df$upper_bound = conf_band_df$y + (conf_band_df$W*conf_band_df$s_Yhat)
  return(conf_band_df)
}

plot_1 +
  geom_path(data = conf_band_reg(copier, alpha = 0.1),
            aes(x = x, y = lower_bound), color = "firebrick") +
  geom_path(data = conf_band_reg(copier, alpha = 0.1),
            aes(x = x, y = upper_bound), color = "firebrick") +
  labs(caption = "Note: The red lines indicates the upper and lower bounds of the regression line")
```



Note: The red lines indicates the upper and lower bounds of the regression line

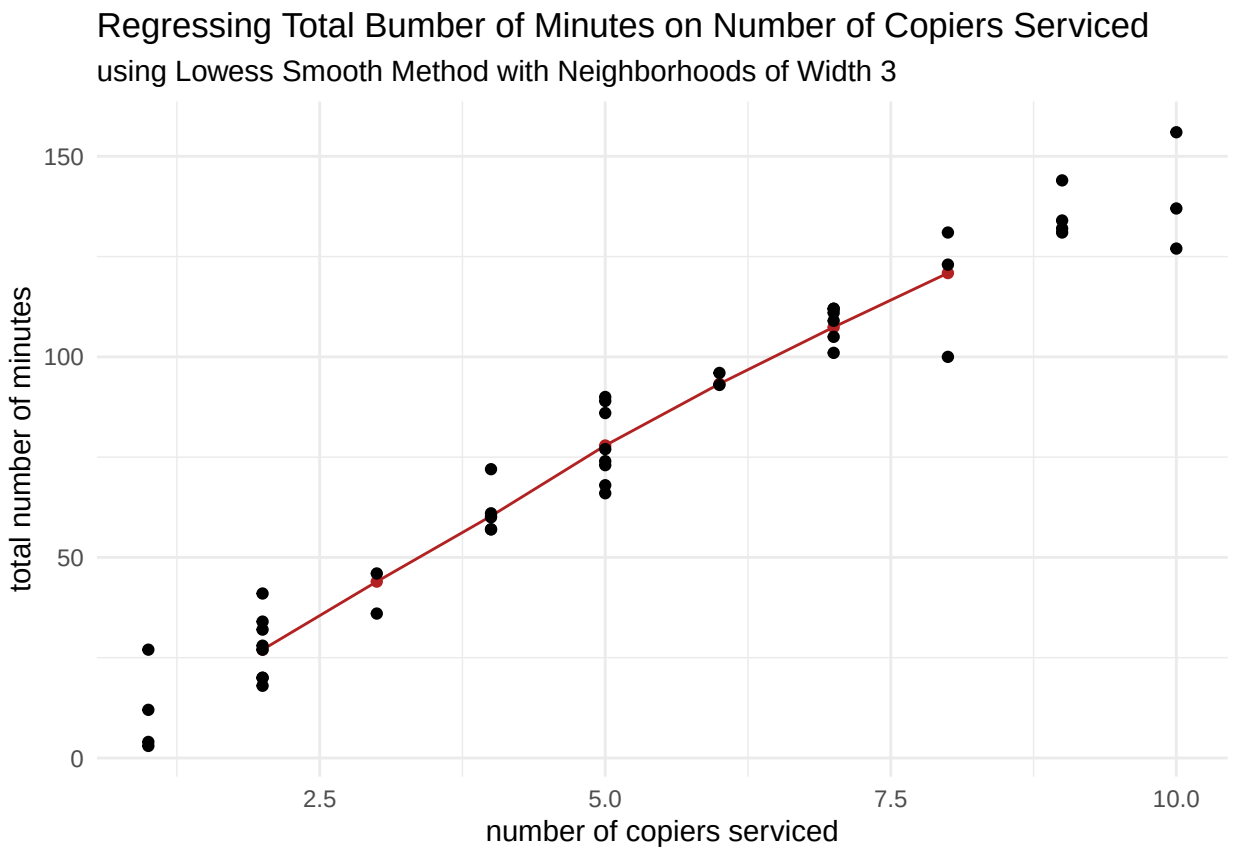
The band smooth does not fall entirely inside the confidence band. Hence the linear regression function may not be the best fit for this data.

(c) Create a series of six overlapping neighborhoods of width 3.0 beginning at $X = .5$. The first neighborhood will range from $X = .5$ to $X = 3.5$; the second neighborhood will range from $X = 1.5$ to $X = 4.5$; and so on. For each of the six overlapping neighborhoods, fit a linear regression function and obtain the fitted value $\hat{Y}_c$ at the center X_c of the neighborhood. Develop a simplified version of the lowess smooth by connecting the six $(X_c, \hat{Y}_c)$ pairs by straight lines on a scatter plot of the data. In what ways does your simplified lowess smooth differ from the band smooth obtained in part (a)?

Answer:

```
X_c_list = c()
pred_list = c()
for(i in seq(0.5, 6.5, by = 1)){
  filtered_data = filter(copier, x >= i & x <= i + 3)
  model = lm.fit_manual(filtered_data$x, filtered_data$y)
  X_c = (i + i + 3)/2
  pred = model[1] + model[2]*X_c

  X_c_list = c(X_c_list, X_c)
  pred_list = c(pred_list, pred)
}

data.frame(X_c = X_c_list, pred = pred_list) %>%
  ggplot(aes(X_c, pred)) +
  geom_point(color = "firebrick") +
  geom_line(color = "firebrick") +
```

```
  geom_point(data = copier, aes(x,y)) +
  labs(x = "number of copiers serviced",
       y = "total number of minutes",
       title = "Regressing Total Bumber of Minutes on Number of Copiers Serviced",
       subtitle = "using Lowess Smooth Method with Neighborhoods of Width 3")
```



The simplified lowess smooth does not fit the data fully from the minimum and maximum value of X. It only covers values that are centers of discrete neighborhoods.

Problem 30:

Refer to Sales growth Problem 3.17.

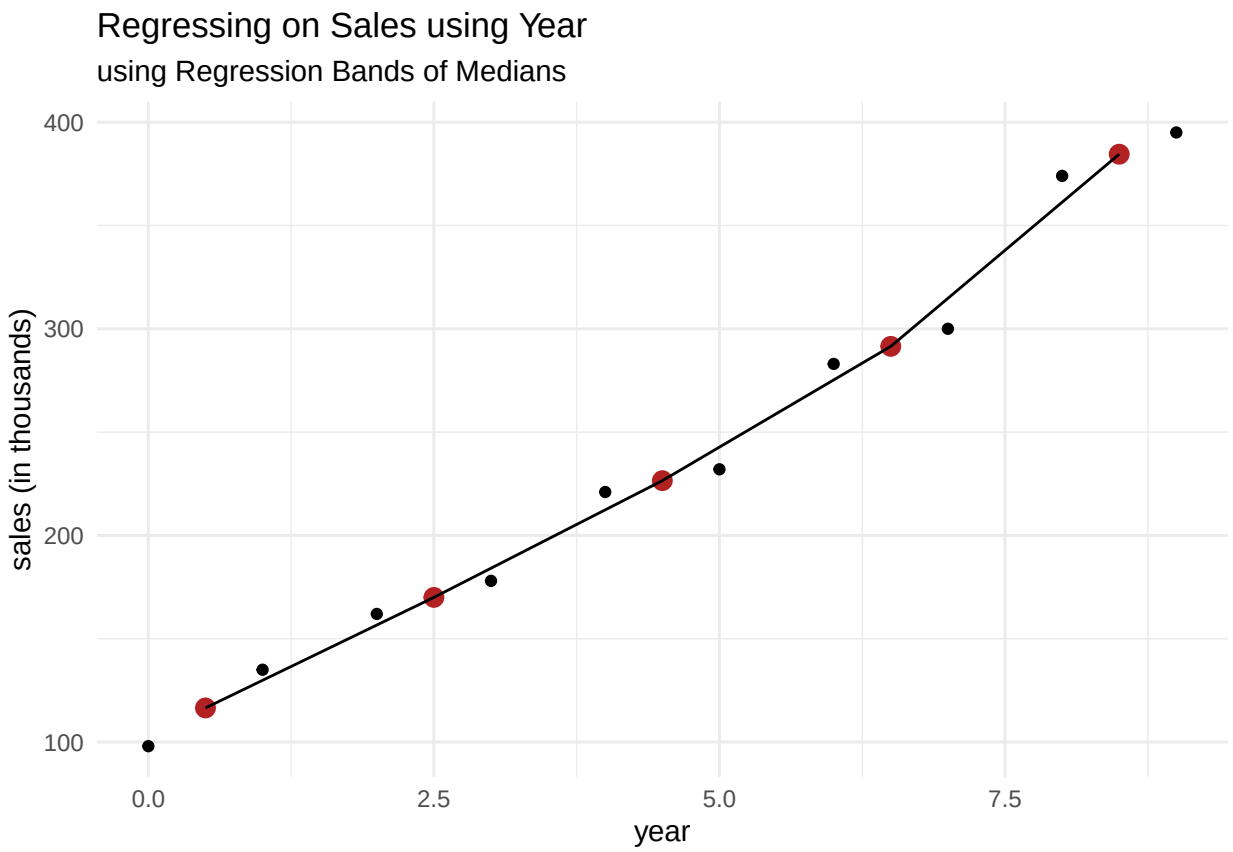
(a) Divide the range of the predictor variable (coded years) into five bands of width 2.0, as follows: Band 1 ranges from $X = -.5$ to $X = 1.5$; band 2 ranges from $X = 1.5$ to $X = 3.5$; and so on. Determine the median value of X and the median value of Y in each band and develop the band smooth by connecting the five pairs of medians by straight lines on a scatter plot of the data. Does the band smooth suggest that the regression relation is linear? Discuss.

Answer:

```
sales %>%
  mutate(band = case_when(
    x > -.5 & x <= 1.5 ~ 1,
    x > 1.5 & x <= 3.5 ~ 2,
    x > 3.5 & x <= 5.5 ~ 3,
    x > 5.5 & x <= 7.5 ~ 4,
    x > 7.5 & x <= 9.5 ~ 5)) %>%
```

```
  group_by(band) %>%
  summarize(x_med = median(x), y_med = median(y)) %>%
  ggplot(aes(x_med, y_med)) +
  geom_point(color = "firebrick", size = 3) +
  geom_path() +
  geom_point(data = sales, aes(x,y)) +
  labs(x = "year", y = "sales (in thousands)",
       title = "Regressing on Sales using Year",
       subtitle = "using Regression Bands of Medians")
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



The band smooth suggests that the regression relation is not linear. The medians do not appear to line up on a straight line.

(b) Create a series of seven overlapping neighborhoods of width 3.0 beginning at $X = -.5$. The first neighborhood will range from $X = -.5$ to $X = 2.5$; the second neighborhood will range from $X = .5$ to $X = 3.5$; and so on. For each of the seven overlapping neighborhoods, fit a linear regression function and obtain the fitted value $\hat{Y}_c$ at the center X_c of the neighborhood. Develop a simplified version of the lowess smooth by connecting the seven $(X_c, \hat{Y}_c)$ pairs by straight lines on a scatter plot of the data.

Answer:

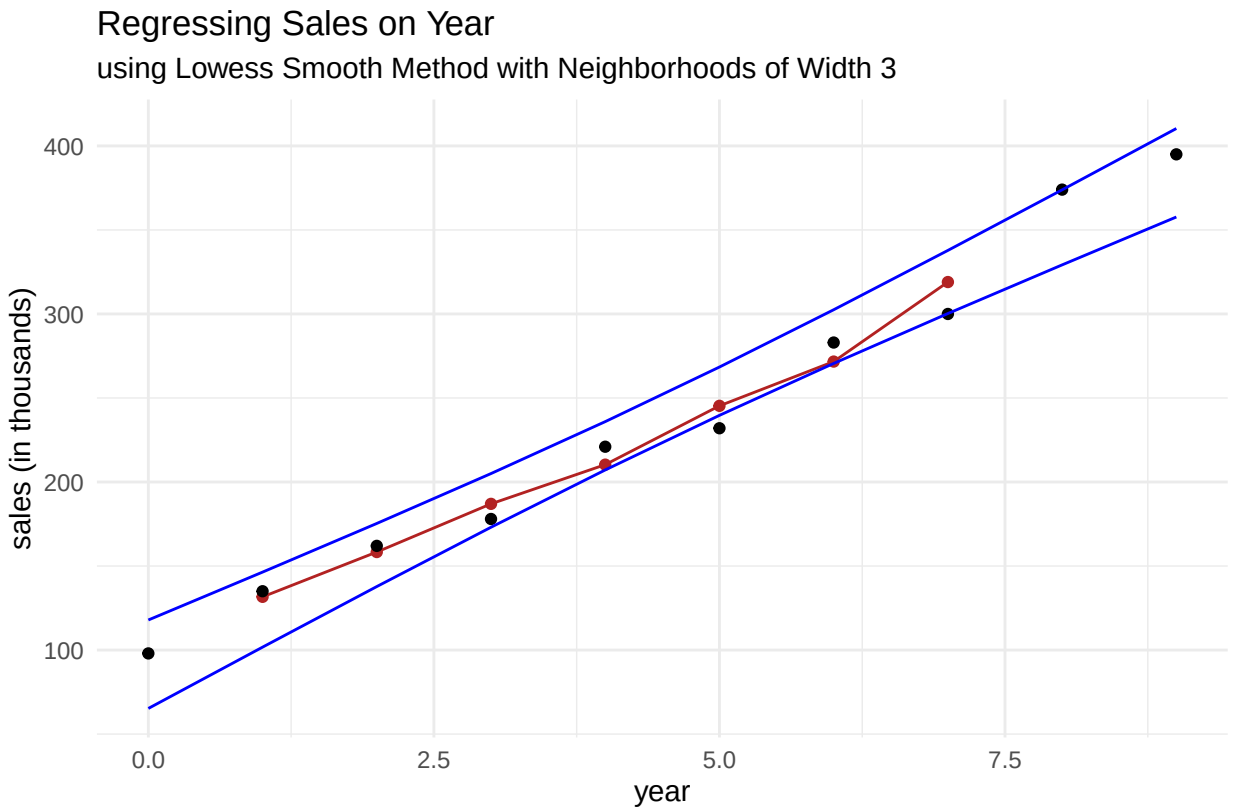
```
X_c_list = c()
pred_list = c()
for(i in seq(-0.5, 5.5, by = 1)){
  filtered_data = filter(sales, x >= i & x <= i + 3)
  model = lm.fit_manual(filtered_data$x, filtered_data$y)
```

```
  X_c = (i + i + 3)/2
  pred = model[1] + model[2]*X_c

  X_c_list = c(X_c_list, X_c)
  pred_list = c(pred_list, pred)
}

plot_1 = data.frame(X_c = X_c_list, pred = pred_list) %>%
  ggplot(aes(X_c, pred)) +
  geom_point(color = "firebrick") +
  geom_line(color = "firebrick") +
  geom_point(data = sales, aes(x,y)) +
  labs(x = "year",
       y = "sales (in thousands)",
       title = "Regressing Sales on Year",
       subtitle = "using Lowess Smooth Method with Neighborhoods of Width 3")
```

(c) Obtain the 95 percent confidence band for the true regression line and plot it on the plot prepared in part (b). Does the simplified lowess smooth fall entirely within the confidence band for the regression line? What does this tell you about the appropriateness of the linear regression function?

Answer:

```
plot_1 +
  geom_path(data = conf_band_reg(sales, alpha = 0.05),
            aes(x = x, y = lower_bound), color = "blue") +
  geom_path(data = conf_band_reg(sales, alpha = 0.05),
            aes(x = x, y = upper_bound), color = "blue") +
  labs(caption = "Note: The blue lines indicates the upper and lower bounds of the regression line")
```



Note: The blue lines indicates the upper and lower bounds of the regression line

The simplified lowess smooth appears to fall entirely within the confidence band for the regression line. This suggests that the linear regression function is appropriate.

Problem 31:

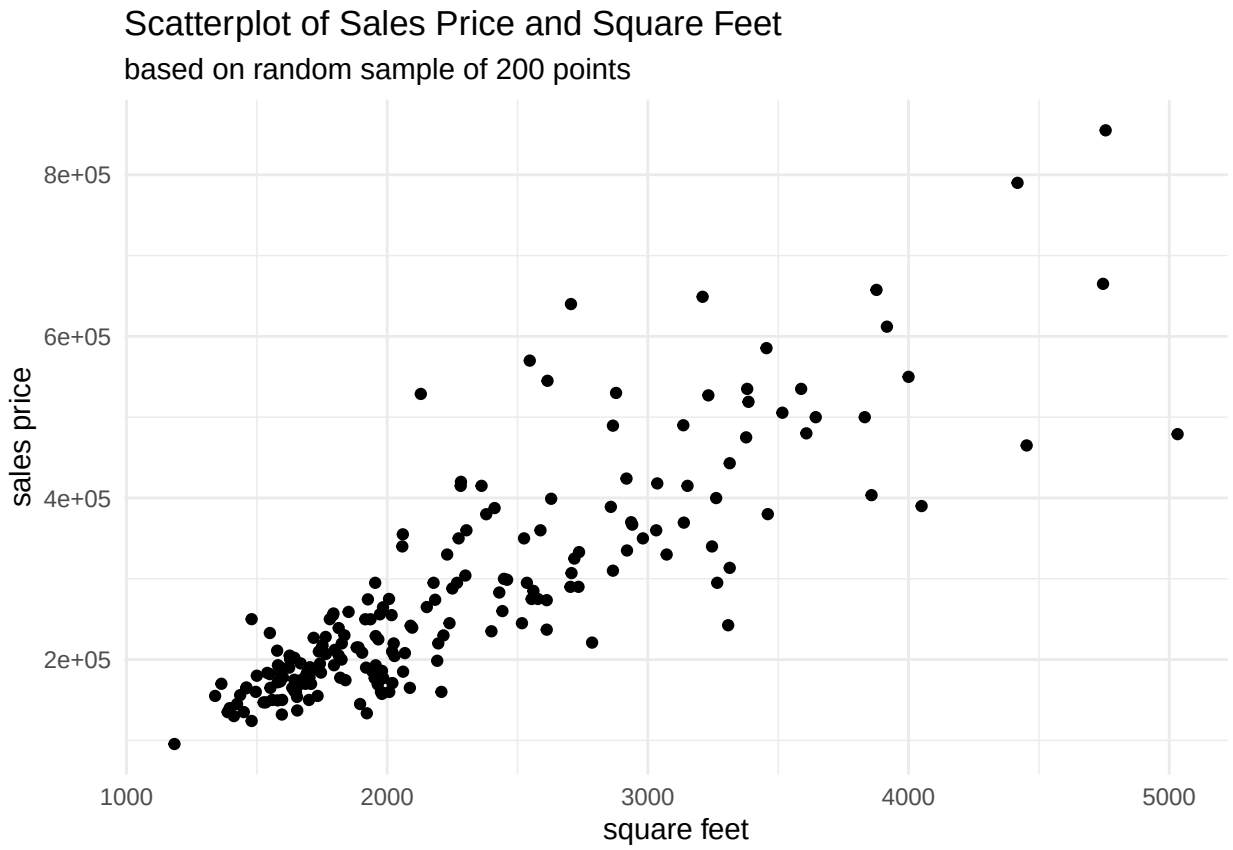
Refer to the Real estate sales data set in Appendix C.7. Obtain a random sample of 200 cases from the 522 cases in this data set. Using the random sample, build a regression model to predict sales price (Y) as a function of finished square feet (X). The analysis should include an assessment of the degree to which the key regression assumptions are satisfied. If the regression assumptions are not met, include and justify appropriate remedial measures. Use the final model to predict sales price for two houses that are about to come on the market: the first has $X = 1100$ finished square feet and the second has $X = 4900$ finished square feet. Assess the strengths and weaknesses of the final model.

Answer:

```
real_estate_cols = c('ID', 'sales_price', 'sq_ft', 'num_bed',
                     'num_bath', 'air_cond', 'garage_size', 'pool', 'year_built',
                     'quality', 'style', 'lot_size', 'adj_to_highway')
real_estate_colclasses = c(rep('numeric', 5), 'factor',
                           'numeric', 'factor', 'numeric',
                           rep('factor', 2), 'numeric', 'factor')
real_estate = read.csv('APPENC07.txt', sep = '', header = FALSE,
                       col.names = real_estate_cols,
                       colClasses = real_estate_colclasses)
```

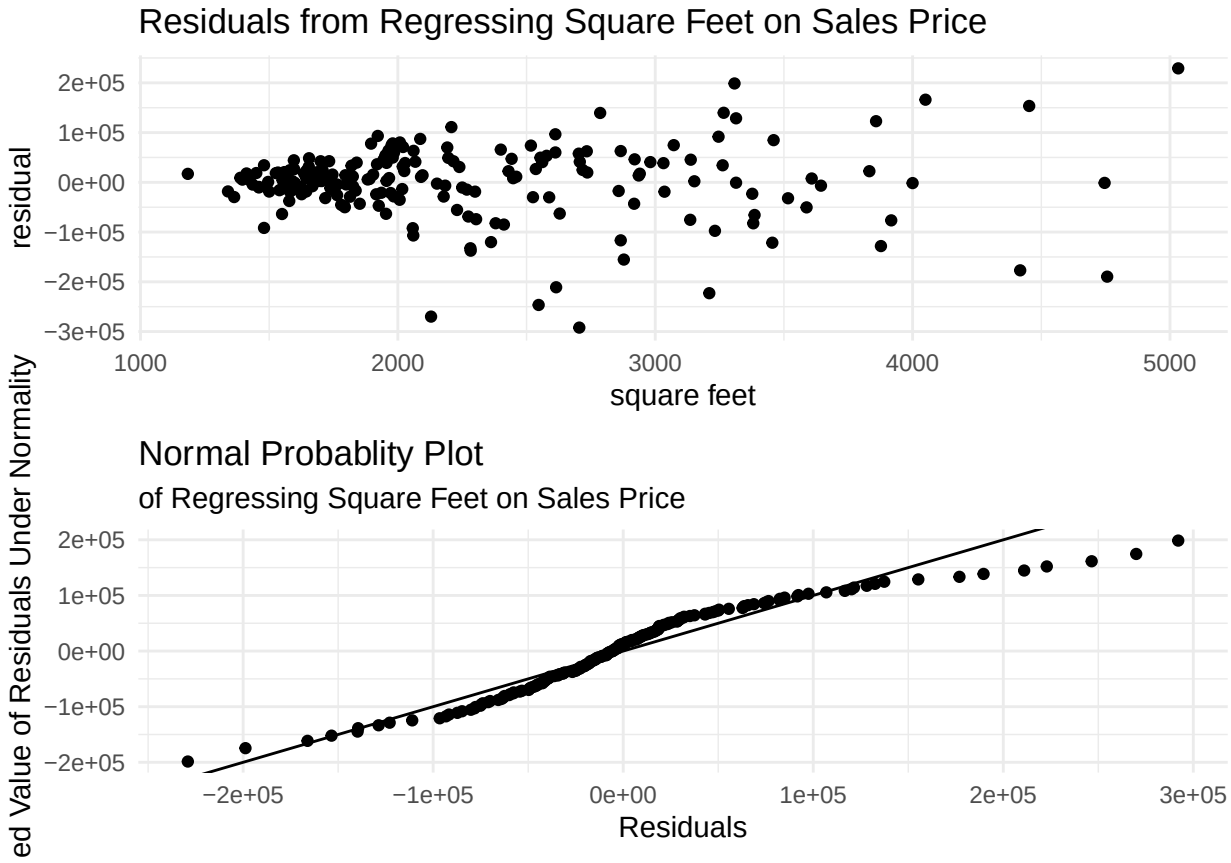
```
set.seed(2020)
real_estate_sampled = real_estate[sample(1:nrow(real_estate), 200, replace = FALSE), c("sales_price", "
model = lm.fit_manual(real_estate_sampled$sq_ft, real_estate_sampled$sales_price)
```

```
real_estate_sampled %>%
  ggplot(aes(sq_ft, sales_price)) +
  geom_point() +
  labs(x = "square feet",
       y = "sales price",
       title = "Scatterplot of Sales Price and Square Feet",
       subtitle = "based on random sample of 200 points")
```



Using the randomly selected points, a model was created. Regression diagnostics plots are shown below.

```
real_estate_model = data.frame(x = real_estate_sampled$sq_ft,
                               y = real_estate_sampled$sales_price,
                               y_pred = model[1] + model[2]*real_estate_sampled$sq_ft,
                               resids = model[1] + (model[2]*real_estate_sampled$sq_ft) - real_estate_sa
plot_1 = real_estate_model %>%
  ggplot(aes(x, resids)) +
  geom_point() +
  labs(x = "square feet",
       y = "residual",
       title = "Residuals from Regressing Square Feet on Sales Price")
plot_2 = prob_plot(data.frame(x=real_estate_sampled$sq_ft, y=real_estate_sampled$sales_price), "of Regr
grid.arrange(plot_1, plot_2, ncol = 1)
```



Error terms appear to have constant variance and are distributed normally.

The sales price for a house that has $X = 1100$ finished square feet is

```
model[1] + model[2]*1100
```

```
## [1] 99535.26
```

and the sales price for a house that has $X = 4900$ finished square feet is

```
model[1] + model[2]*4900
```

```
## [1] 687669.1
```

The final model will prove to be ineffective for houses that are large in square feet, as demonstrated by the giant spread of variance towards the high end of square feet in the diagnostic plot above. It also does not take into account the various other factors that go into pricing a house in real estate.

Problem 32:

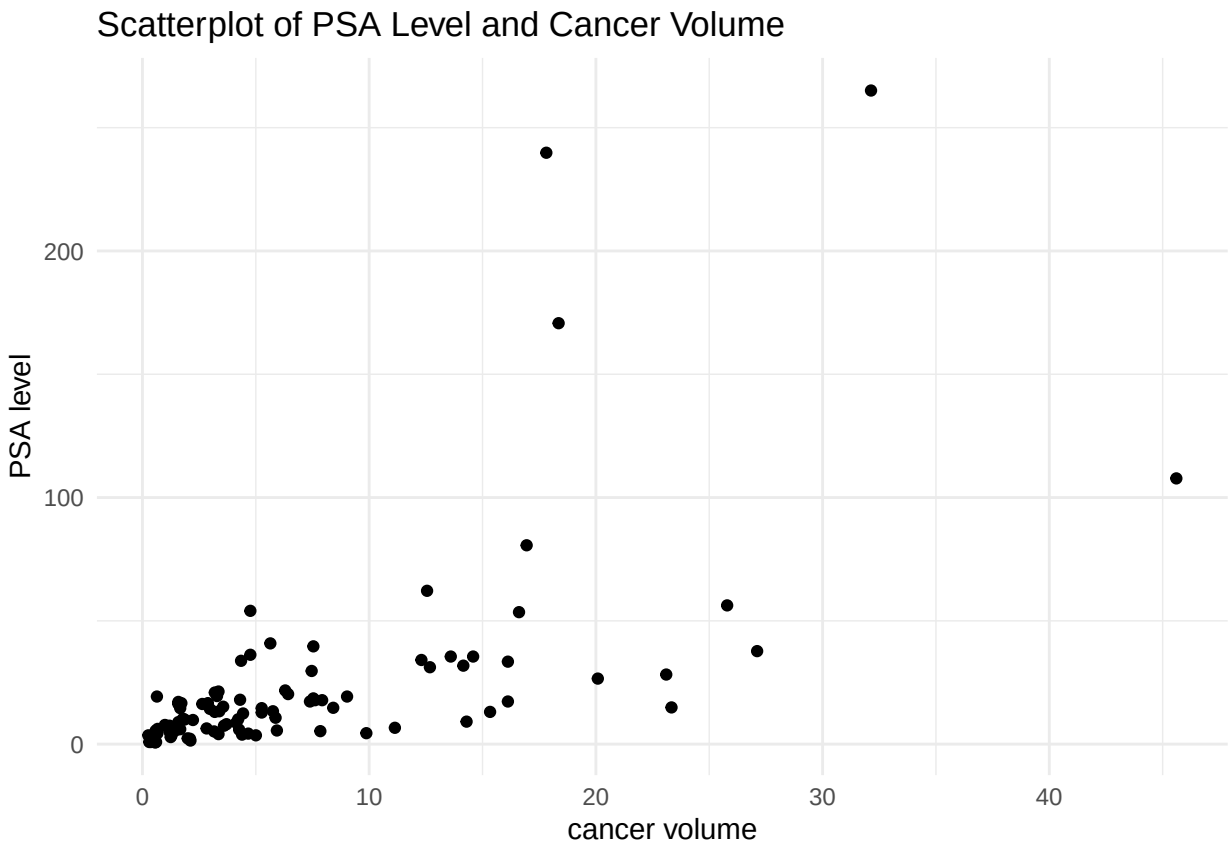
Refer to the Prostate cancer data set in Appendix C.5. Build a regression model to predict PSA level (Y) as a function of cancer volume (X). The analysis should include an assessment of the degree to which the key regression assumptions are satisfied. If the regression assumptions are not met, include and justify appropriate remedial measures. Use the final model to estimate mean PSA level for a patient whose cancer volume is 20 cc. Assess the strengths and weaknesses of the final model.

Answer:

```
prostate_cols = c('ID', 'PSA_level', 'cancer_volume', 'weight', 'age',
                  'benign_prostatic_hyperplasia', 'seminal_vesical_invation',
                  'capsular_penetration', 'gleason_score')
```

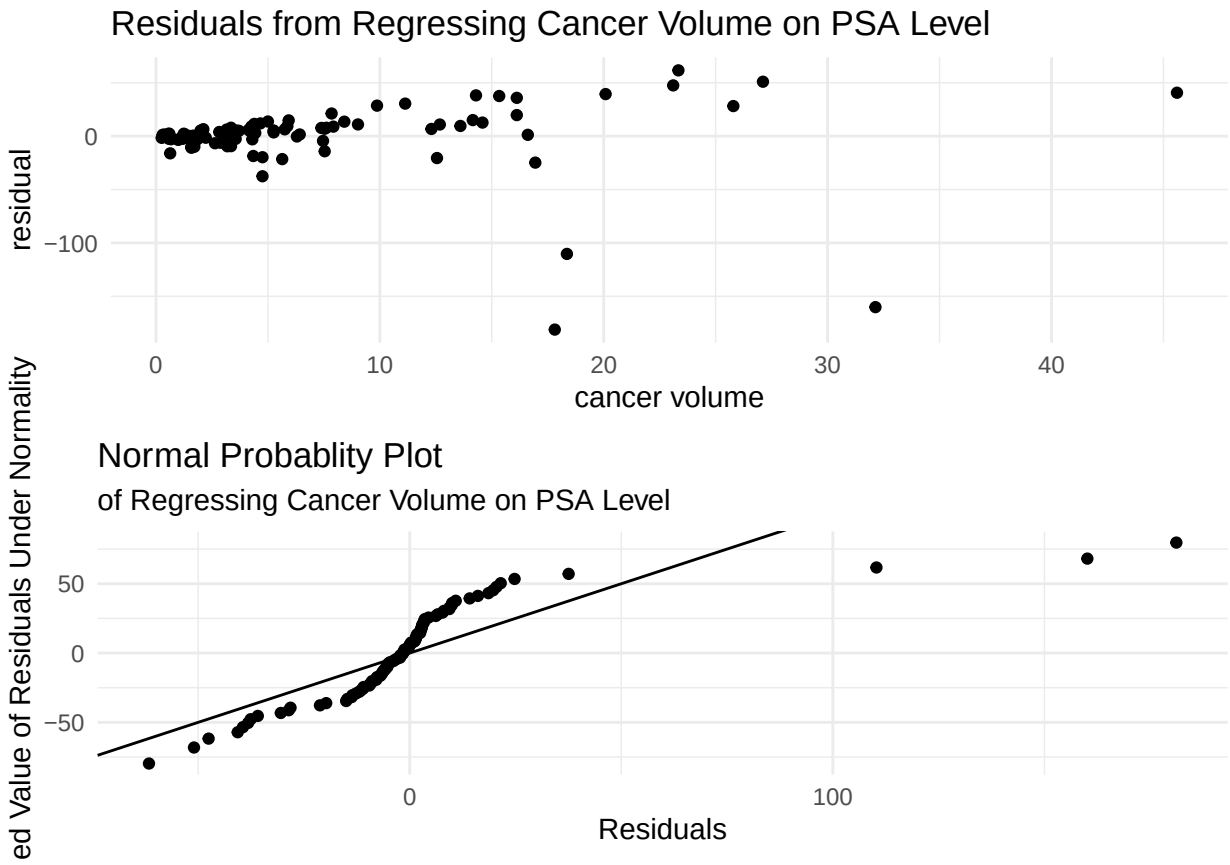
```
prostate_colclasses = c(rep('numeric', 6), 'factor', 'numeric', 'factor')
prostate = read.csv('APPENC05.txt', sep = '', header = FALSE,
                        col.names = prostate_cols,
                        colClasses = prostate_colclasses)
```

```
prostate %>%
  ggplot(aes(cancer_volume, PSA_level)) +
  geom_point() +
  labs(x = "cancer volume",
       y = "PSA level",
       title = "Scatterplot of PSA Level and Cancer Volume")
```



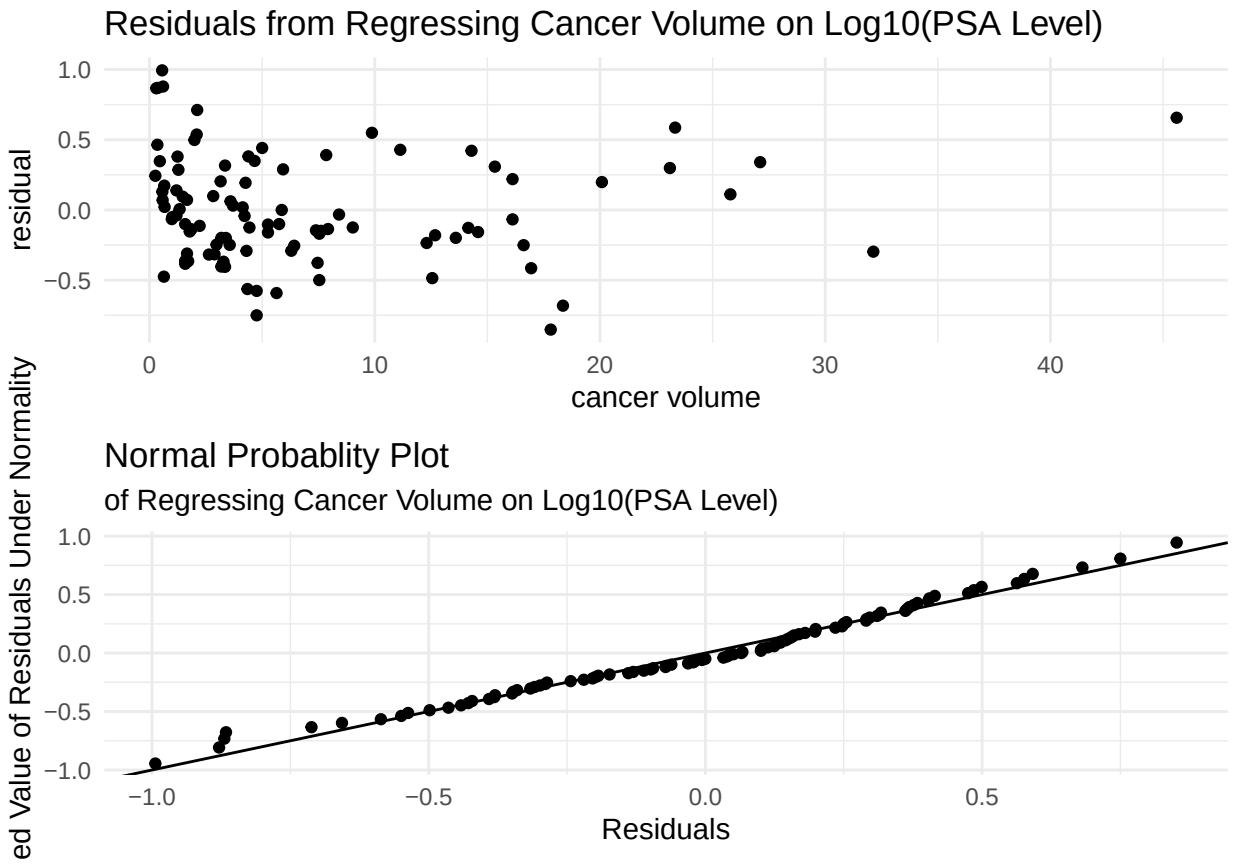
Using all the points, a model was created. Regression diagnostics plots are shown below.

```
model = lm.fit_manual(prostate$cancer_volume, prostate$PSA_level)
prostate_model = data.frame(x = prostate$cancer_volume,
                            y = prostate$PSA_level,
                            y_pred = model[1] + model[2]*prostate$cancer_volume,
                            resids = model[1] + (model[2]*prostate$cancer_volume) - prostate$PSA_level)
plot_1 = prostate_model %>%
  ggplot(aes(x, resids)) +
  geom_point() +
  labs(x = "cancer volume",
       y = "residual",
       title = "Residuals from Regressing Cancer Volume on PSA Level")
plot_2 = prob_plot(data.frame(x=prostate$cancer_volume, y=prostate$PSA_level), "of Regressing Cancer Vo
grid.arrange(plot_1, plot_2, ncol = 1)
```



The model does not appear to be a good fit for the data. Error terms deviate heavily from normality and the error variance is sparse at the high end. To remedy this, a $\log_{10}$ tranformation will be used on the PSA level (Y) variable. By doing so and creating the regression model again, the following diagnostic plots are made.

```
model = lm.fit_manual(prostate$cancer_volume, log10(prostate$PSA_level))
prostate_model = data.frame(x = prostate$cancer_volume,
                            y = log10(prostate$PSA_level),
                            y_pred = model[1] + model[2]*prostate$cancer_volume,
                            resids = model[1] + (model[2]*prostate$cancer_volume) - log10(prostate$PSA_
plot_1 = prostate_model %>%
  ggplot(aes(x, resids)) +
  geom_point() +
  labs(x = "cancer volume",
       y = "residual",
       title = "Residuals from Regressing Cancer Volume on Log10(PSA Level)")
plot_2 = prob_plot(data.frame(x=prostate$cancer_volume, y=log10(prostate$PSA_level)), "of Regressing Ca
grid.arrange(plot_1, plot_2, ncol = 1)
```



By using the $\log_{10}$ tranformation on Y, the error terms are now distributed normally and have a constant variance. The regression function in its original units is:

```
paste("Y = 10^(", round(model[1], 3), "+", round(model[2], 3), "x)", sep = '')
```

```
## [1] "Y = 10^(0.784+0.042x)"
```