HW4

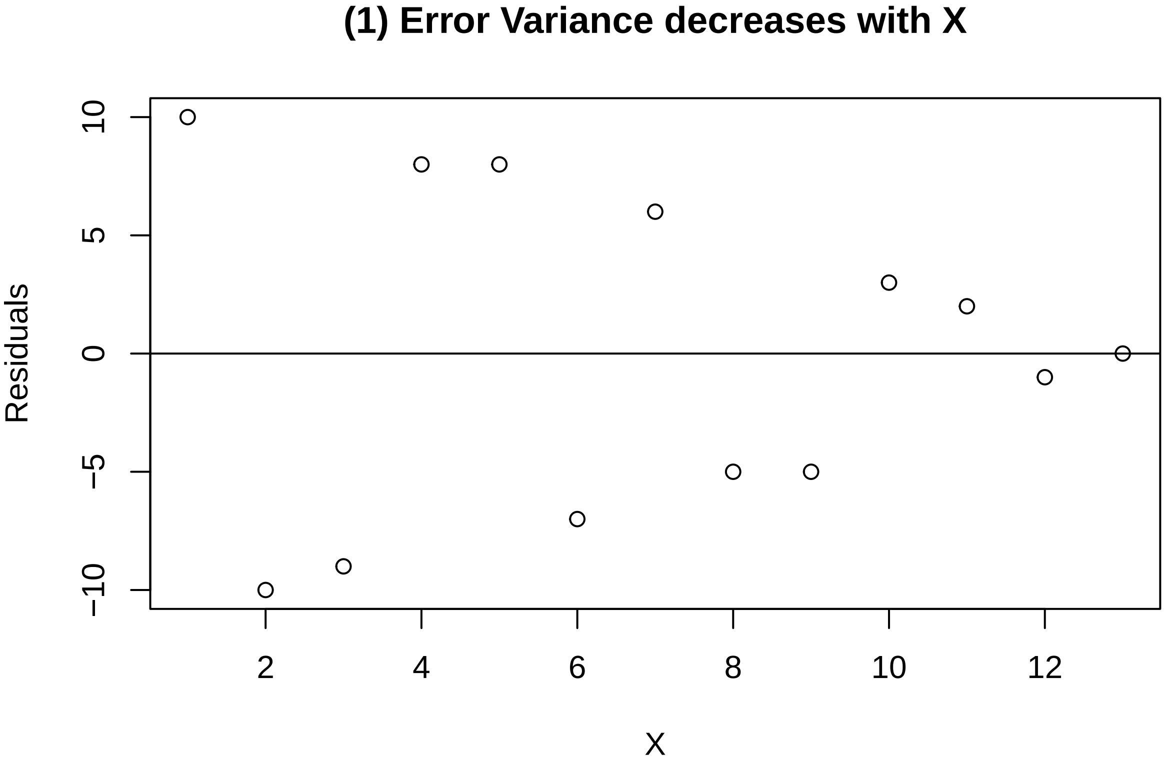
Dylan M Ang

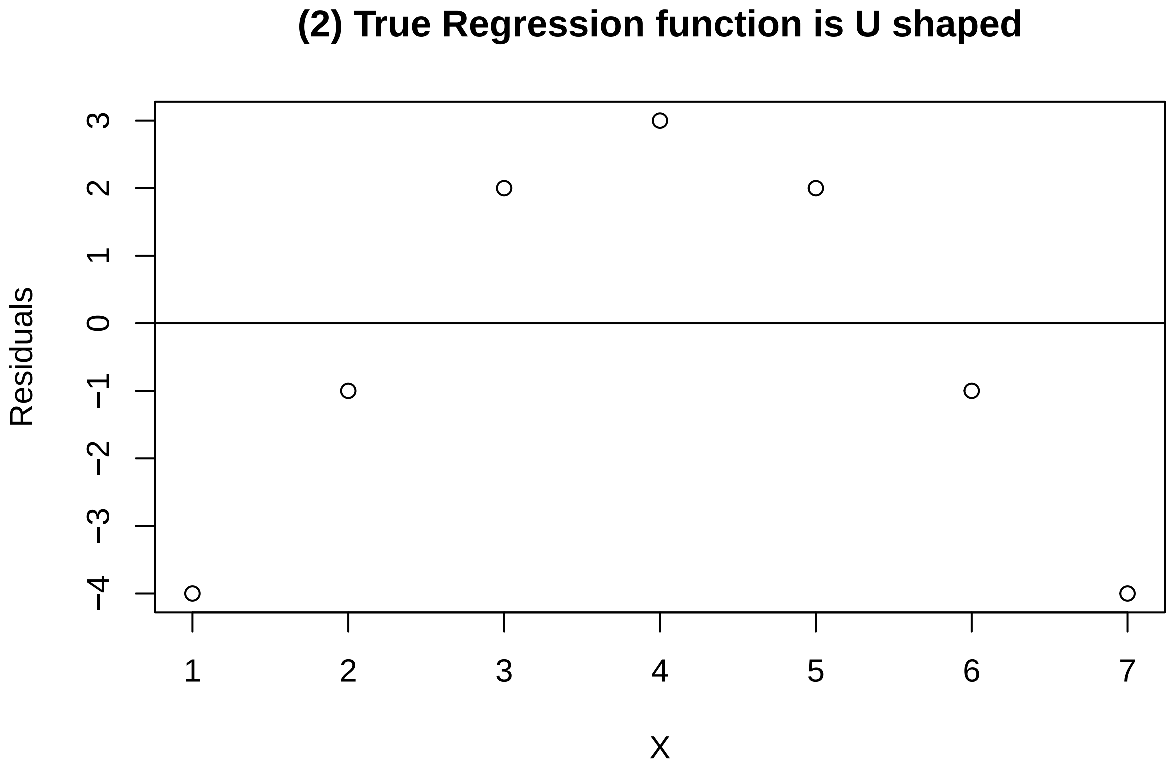
2/8/2022

`problem set: 3.2, 3.3, 3.10, 3.13, 3.15, 3.16.`

3.2

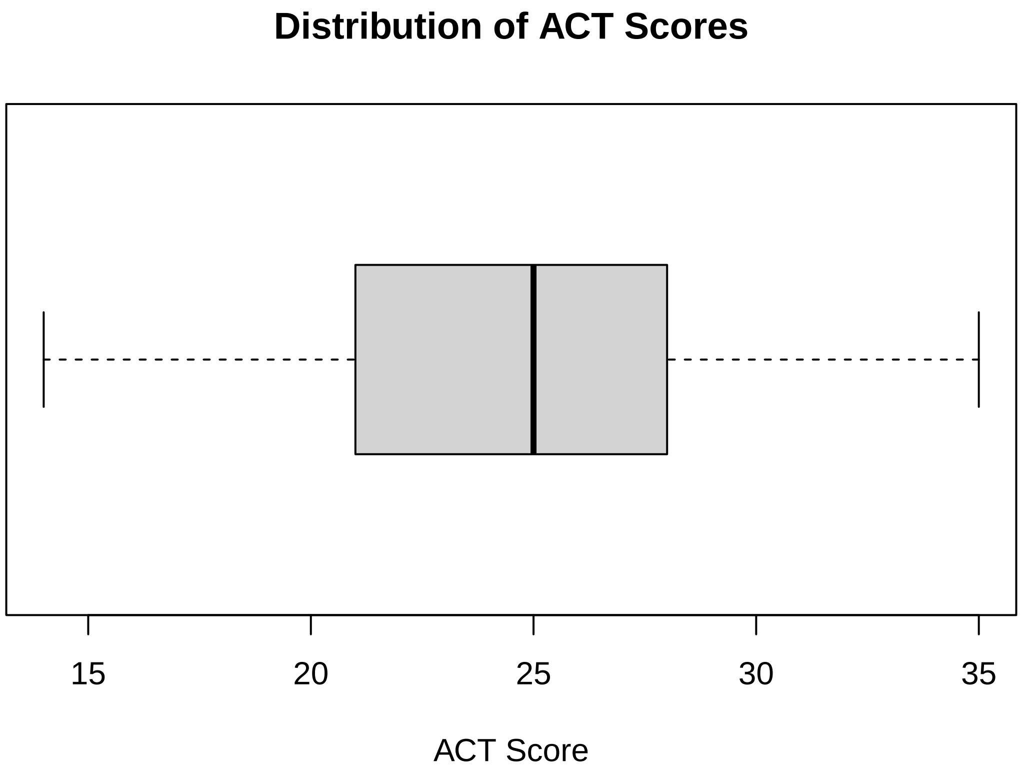
> ***Q*** Prepare a prototype residual plot for each of the following cases: (1) error variances decreases with X; (2) true regression function is $\cup$ shaped, but a linear regression function is fitted.




(2) True Regression function is U shaped
Residuals
3
2
1
0
−1
−2
−3
−4
1
2
3
4
5
6
7
X

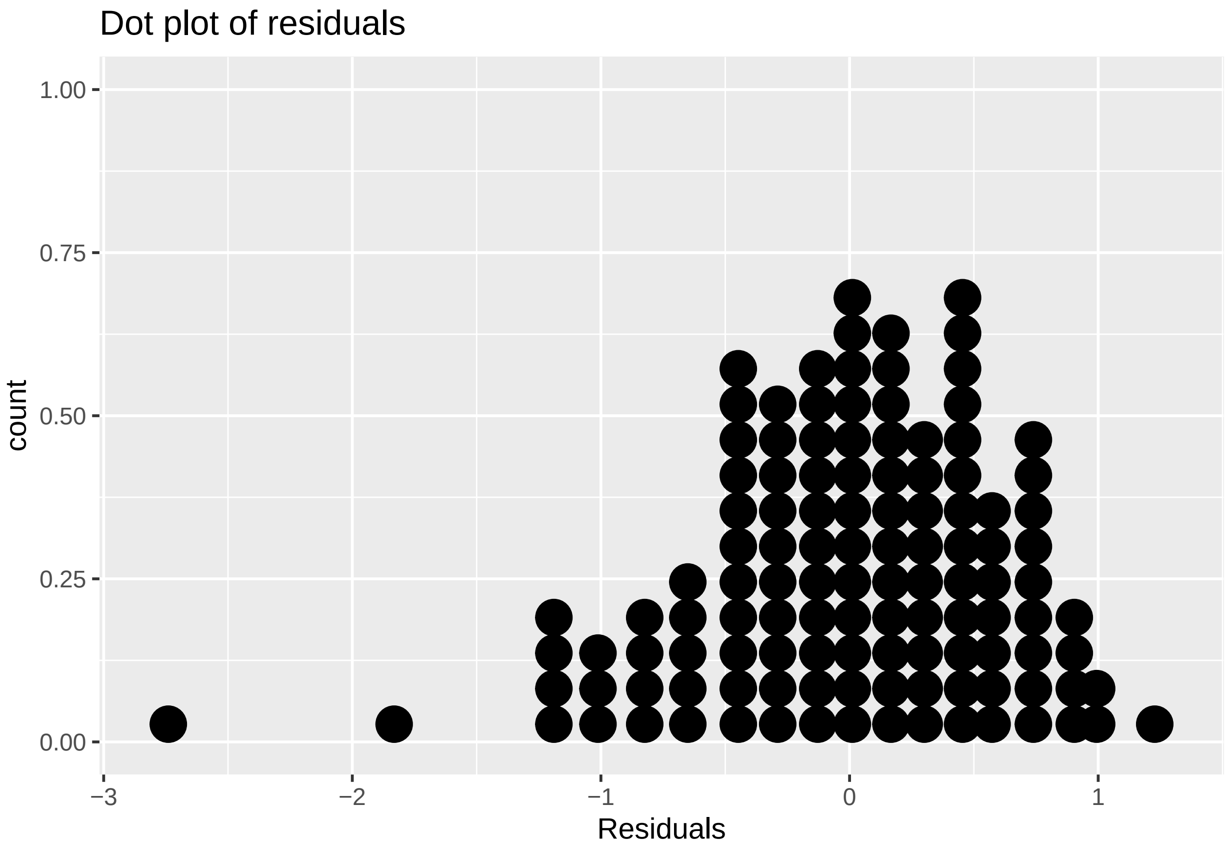
3.3

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



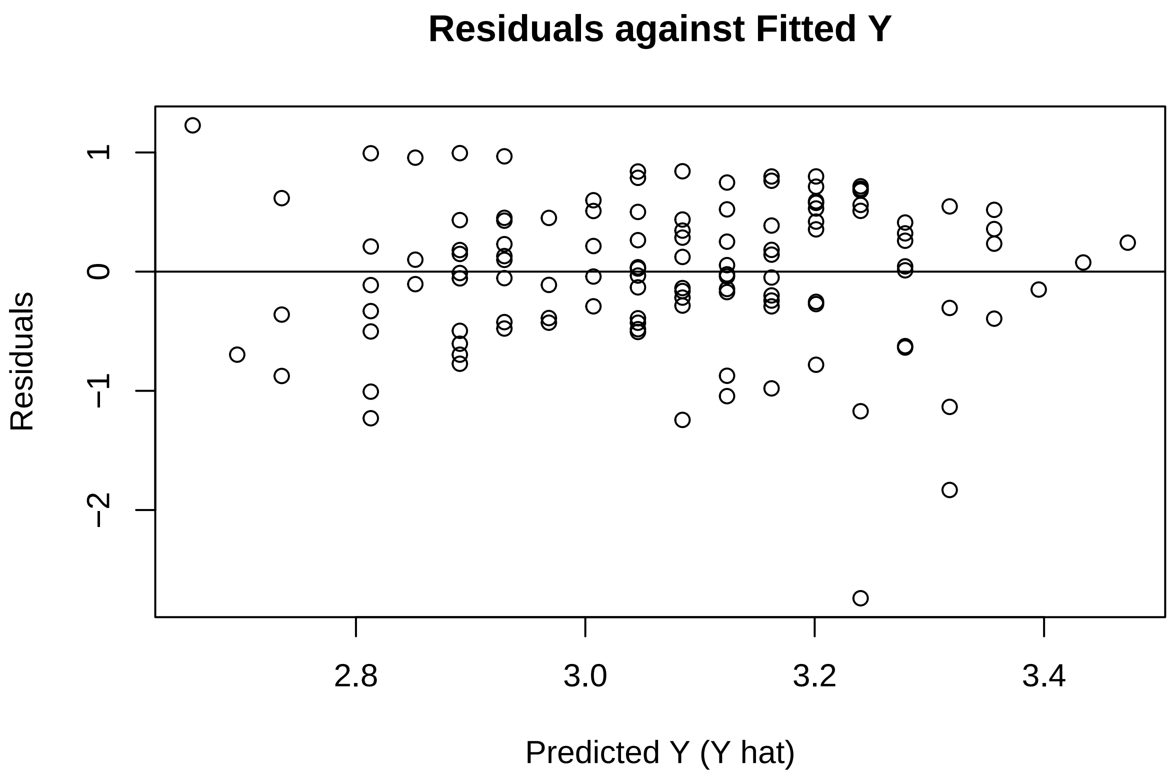
Based on the box plot, the median score is 25, the min and max are 14 and 35 respectively, and there are no outliers.

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



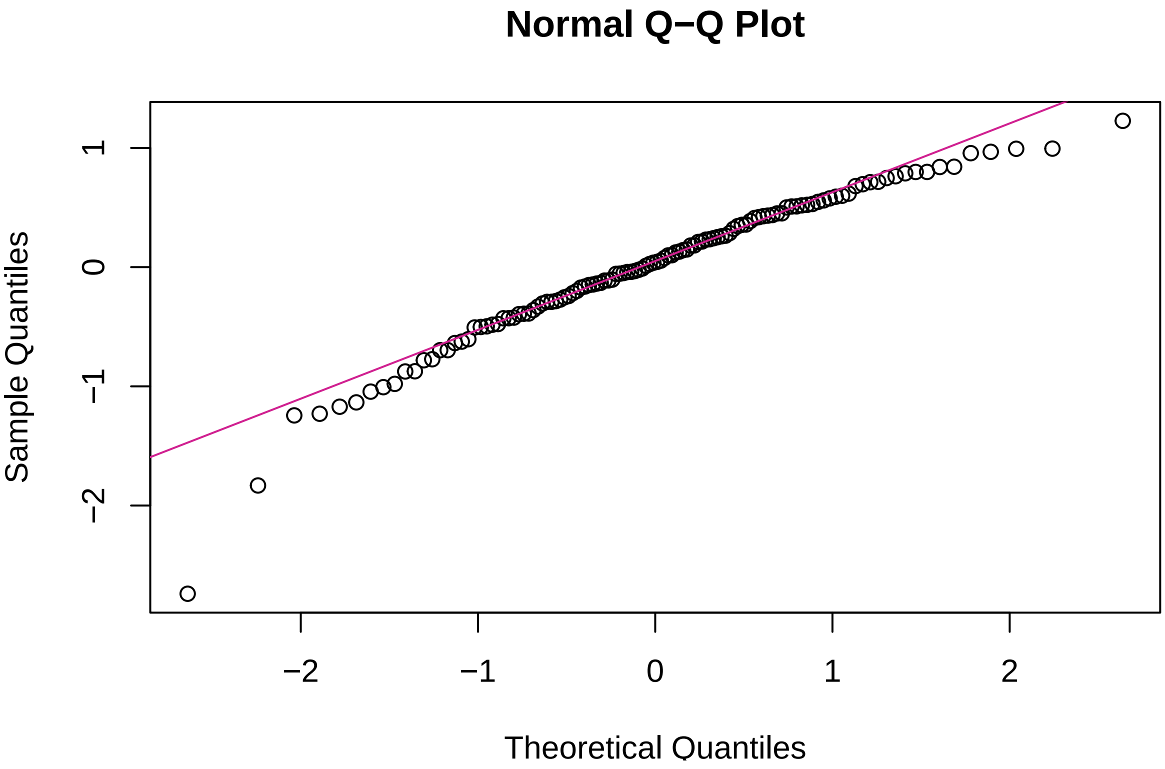
Based on the dot plot, the residuals are roughly normally distributed and centered around 0. This means that our data fits a linear regression function.

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?



The largest departure from the model is the lowest x point that is -2.74. Aside from that, most points are still settled around 0. Since the distribution of the points appears random, we can say that there is likely 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?



H_0: The data is normally distributed ($\rho = 1$).

H_a: The distribution is non-normal ($\rho < 1$).

The critical value of r for $n = 120$ is 0.987

$$r : 0.974 < r^* : 0.987$$

Since our correlation coefficient is less than the critical r value, we fail to reject the null hypothesis. There is sufficient evidence to conclude that there is a significant linear relationship between ACT scores (x) and GPA(y) because the correlation coefficient is sufficiently different from zero.

> ***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 preiminary findings in part (c).

$H_0 : \sigma_1^2 = \sigma_2^2$

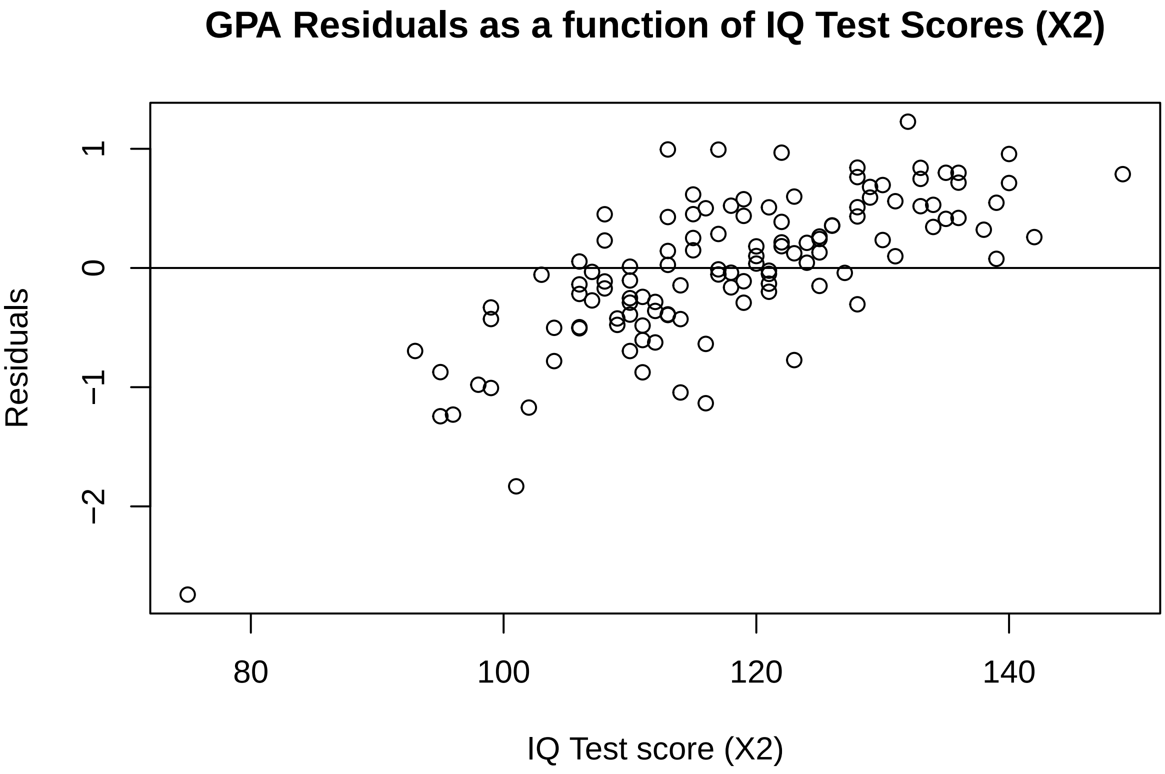
$H_a : \sigma_1^2 \neq \sigma_2^2$

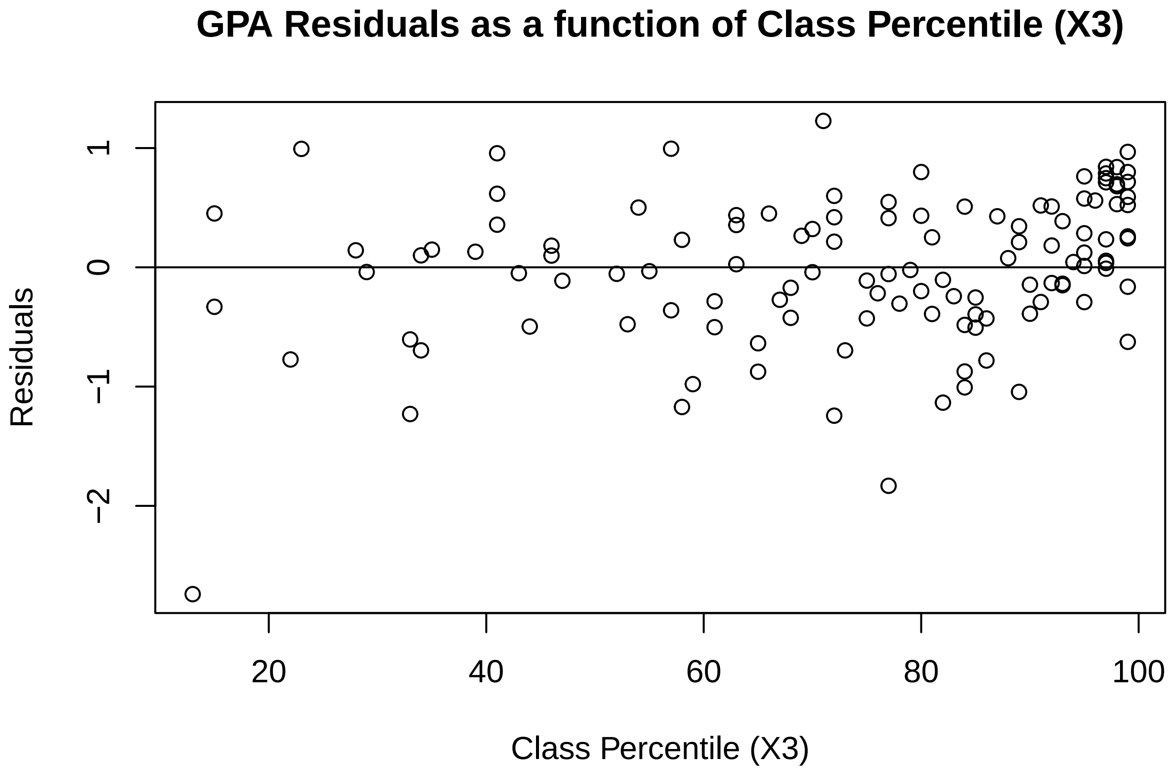
Since $t_{BF} \sim t_{n-2}$ under H_0, reject H_0 when $|t_{BF} > t_{n-2;1-\alpha/2}|$.

$$
\begin{aligned}
s^2 &= \frac{\Sigma_{i=1}^{n_1}(d_{i1}-\bar{d}_1)^2 + \Sigma_{i=1}^{n_2}(d_{i2}-\bar{d}_2)^2}{n-2} \\
&= 0.1741184 \\
s &= \sqrt{s^2} \\
&= 0.417275 \\
t_{BF} &= \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \\
&= -0.8967448 \\
t_{n-2;1-\alpha/2} &= 2.6181369
\end{aligned}
$$

Since $|t_{BF}| = 0.8967448 < 2.6181369$, we fail to reject H_0, so the residuals have constant variance. This supports the conclusion in part (c), which found that the data appeared to have constant variance.

> ***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?



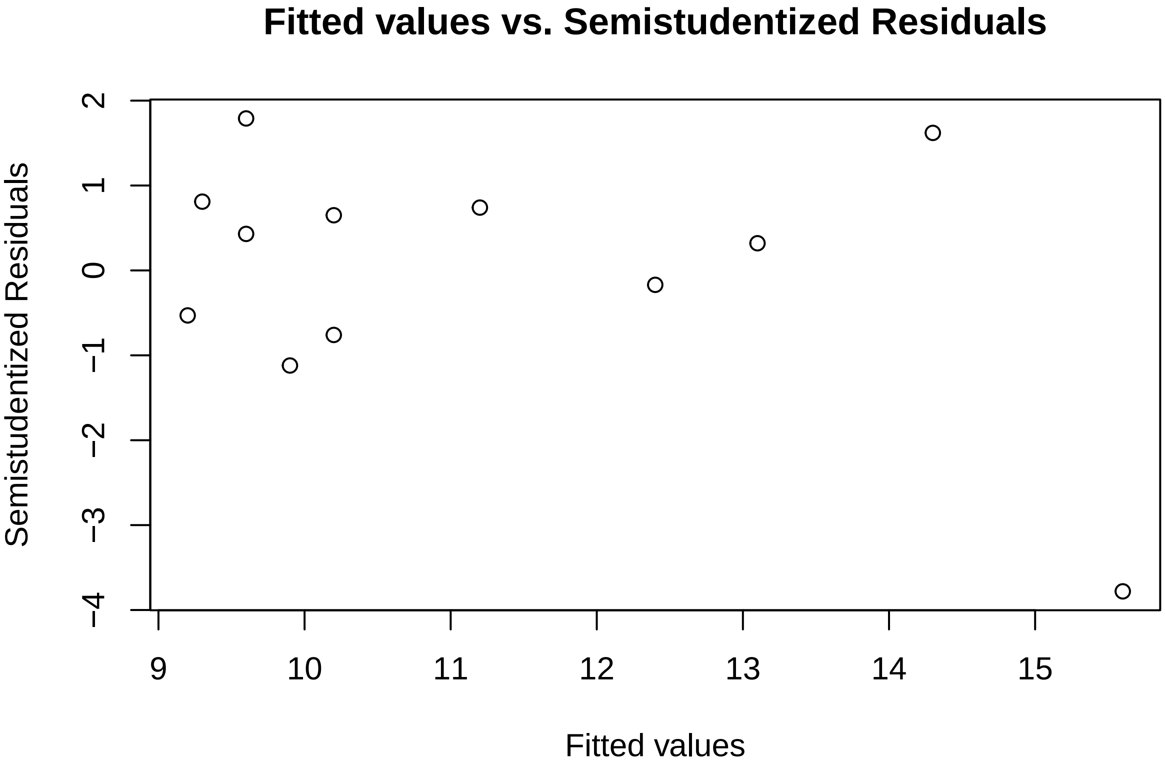


The Class Percentile (X3) graph appears random so we can probably add that to the model, but IQ test score (X2) seems to have a pattern to the data.

3.10

Q 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^* follow.

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



The semistudentized residual is just the residual divided by the square root of the MSE. The residual plot suggests that the error terms are independent.

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?

There are 3 elements more than 1 standard deviation away from the mean. That means that $\frac{12-3}{12} * 100 = 75\%$ of the total data is within 1 standard deviation of the mean. For a standard normal distribution, 68% of the data should fall within 1 standard deviation of the mean, so out result seems about right.

3.13

Q Refer to Copier Maintenance Problem 1.20

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

$H_0 : E(Y) = \beta_0 + \beta_1 X$

$H_a : E(Y) \neq \beta_0 + \beta_1 X$

b Perform the test indicated in part (a). Control the risk of Type I error at 0.05. State the decision rule and conclusion.

$p = P(F_{df_R - df_F, df_F} > F^*)$

$p > \alpha \implies H_0$

$p < \alpha \implies H_a$

$F = 1013.711019, p = 0.3351708$. In this case we see that $p > \alpha \implies$ fail to reject H_0.

In conclusion, the data fits a linear regression mode.

c Does the 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.

No, the lack of constant variance is tested with the Brown-Forsythe statistic. A lack of constant variance would affect the results of the test because lack of variance is a good reason that data would not fit a linear model, even if the error terms are quite small for our sample.

3.15

Q 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, respectivey, after 1, 3, 5, 7, and 9 hours. The results follow.

a Fit a linear regression function.

$$Y_i = 2.5753333 + -0.324X_i$$

b Perform the F test to determine whether or not there is a lack of fit of a linear regression function; use $\alpha = 0.025$. State the alternatives, decision rule, and conclusion.

$H_0 : E(Y) = \beta_0 + \beta_1 X_i$

$H_a : E(Y) \neq \beta_0 + \beta_1 X_i$

Decision rule, if $p < \alpha$, reject H_0. Else, fail to reject the null.

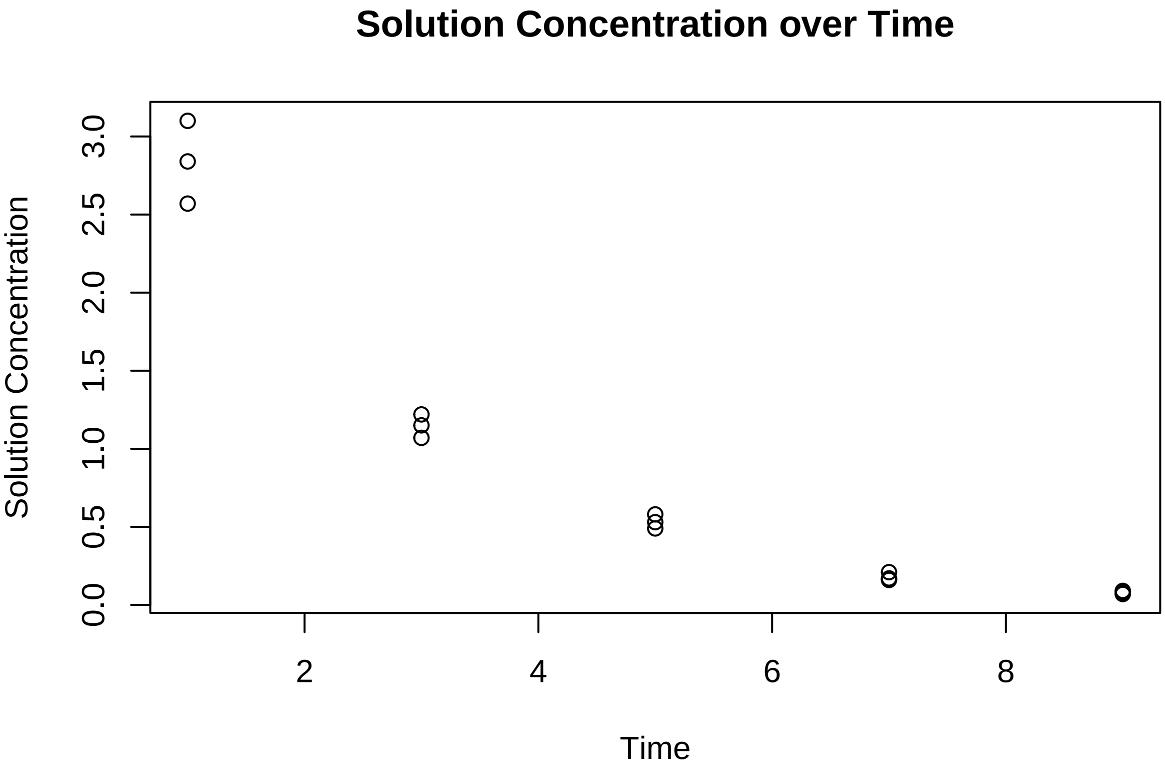
Our p value is $0.3475079 > \alpha$, therefore we fail to reject the null hypothesis. Therefore the data does fit a linear model, and time isa good predictor of solution concentration.

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?

The F test determines if the relationship is statistically significant, but in the even that it determines that a linear relationship is innapropriate, it does not tell you what type of regression function is appropriate.

3.16

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?



A box cox transformation would transform Y to achieve non-normality of the errors, which in turn achieves linearity and consistent variance.

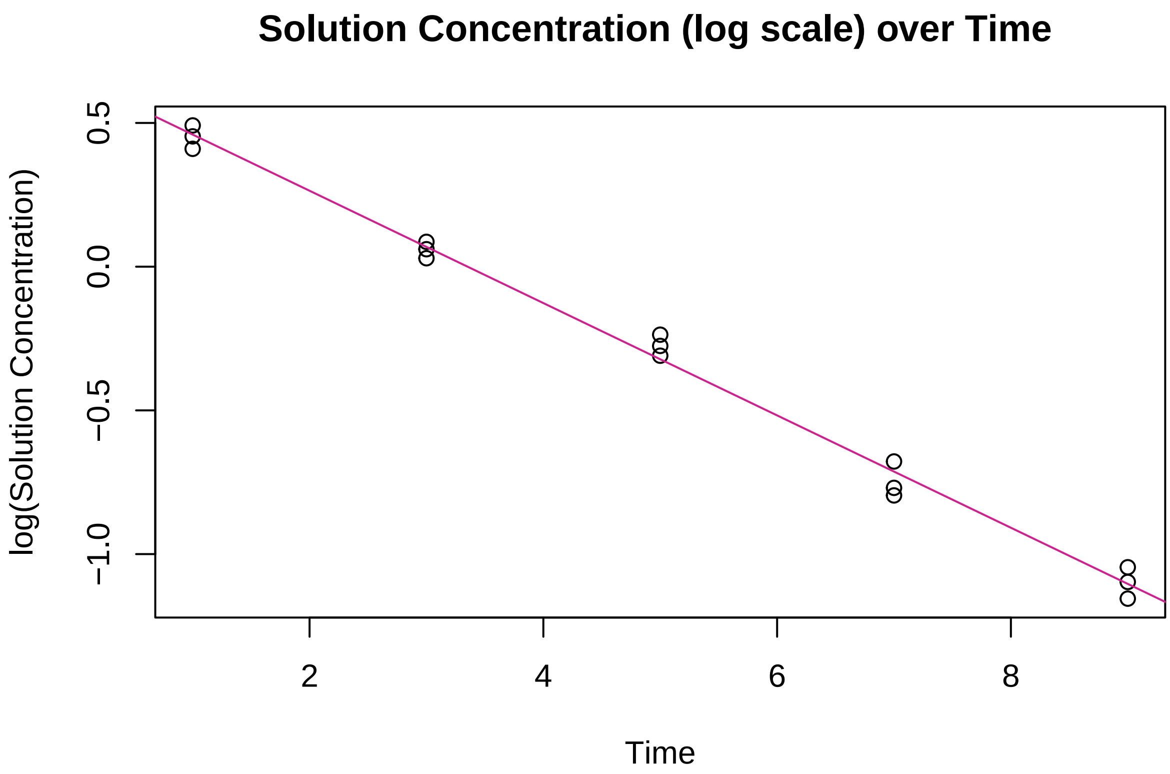
b Use the Box-Cox procedure and standardization (3.36) to find an appropriate power transformation. Evaluate SSE for $\lambda = -0.2, -0.1, 0, 0.1, 0.2$. What transformation of Y is suggested?

SSE is minimized to 19.7343698 when $\lambda = 0$

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

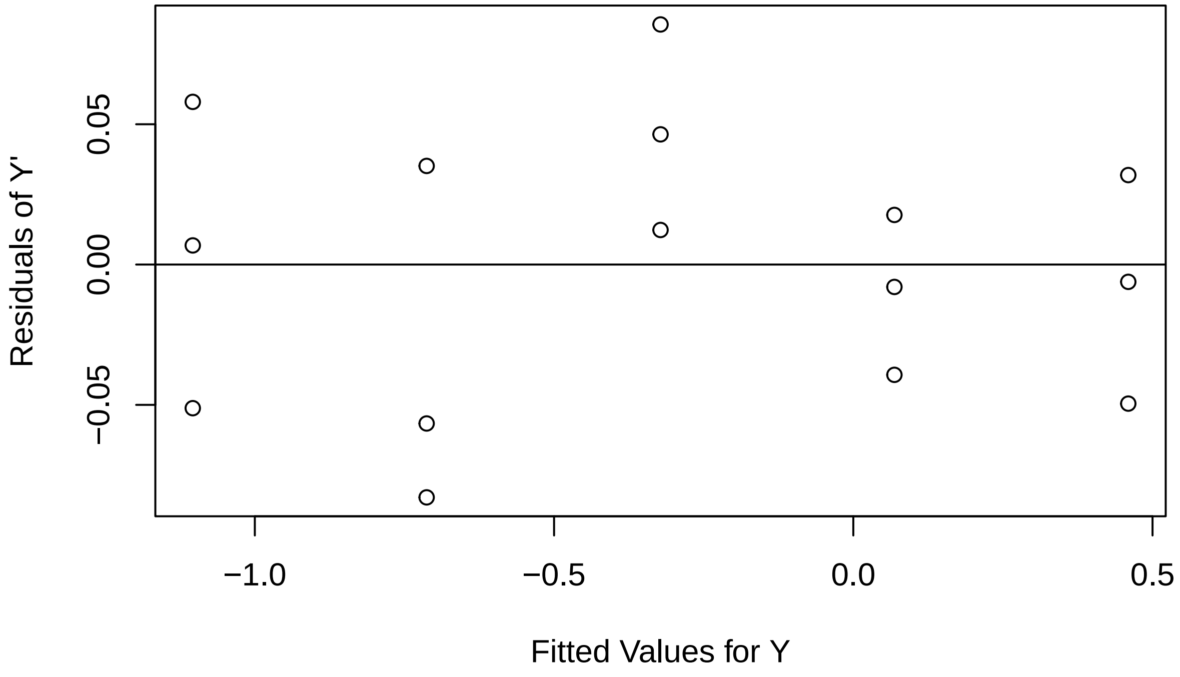
$$Y_i = 0.6548798 + -0.1954003X_i$$

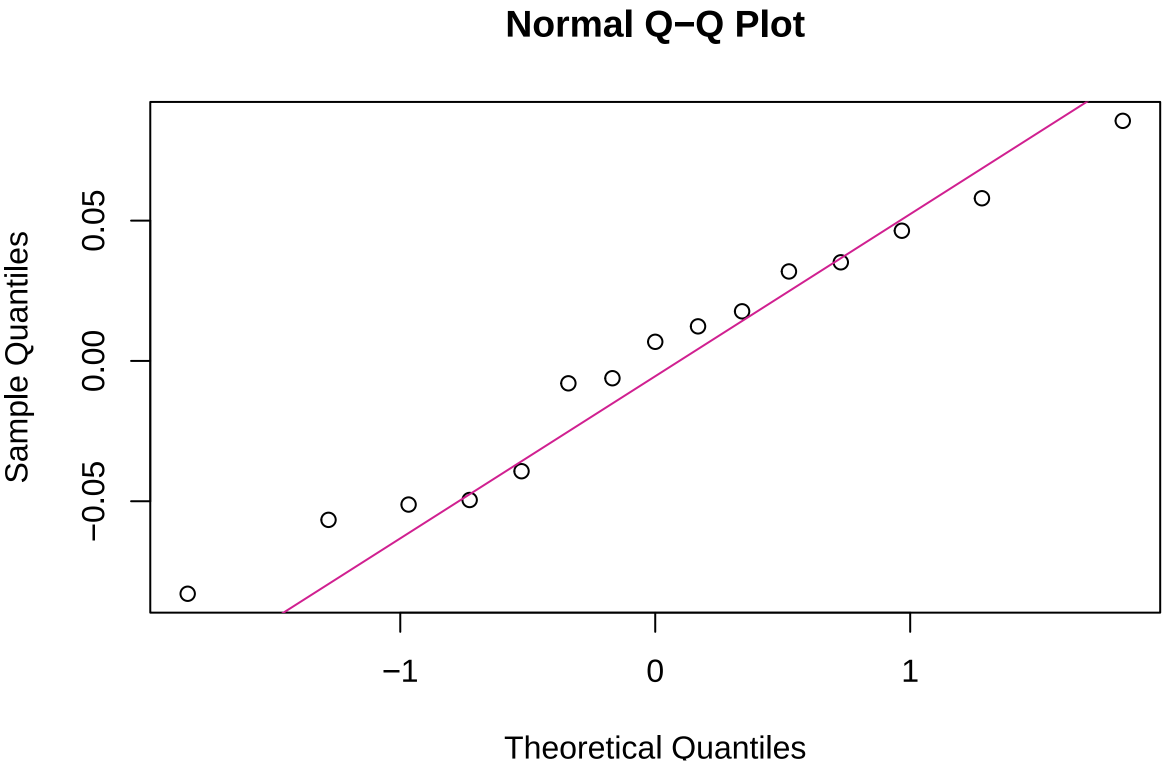
d Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?



The new regression line for Y' is a good fit for the data, much better than the non-transformed data was.

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





The plots show that the model is linear and has constant variance. We can see the linearity because the normal probability plot is roughly linear, and the residual graph is seemingly random. There is no pattern to the residual graph, so we can say that it has constant variance.

f Express the estimated regression functin in the original units.

The logarithmic regression function is $Y_i = 0.6548798 + -0.1954003X_i$.

Expressed in the original units,

$$Y_i = 10^{2.5753333+-0.324X_i}$$

Appendix

```r
knitr::opts_chunk$set(echo = FALSE)
# 3.2 Residual Plots
res = c(10, -10, -9, 8, 8, -7, 6, -5, -5, 3, 2, -1, 0)
plot(res,
     ylab = "Residuals", xlab = "X",
     main = "(1) Error Variance decreases with X")
abline(0,0, )

res = c(-4, -1, 2, 3, 2, -1, -4)
plot(res,
     ylab = "Residuals", xlab = "X",
     main = "(2) True Regression function is U shaped")
abline(0,0, )
# 3.3 GPAS
gpas_set = read.table("../datasets/grade+point+average.txt")
gpas = gpas_set$V1
acts = gpas_set$V2

# 3.3a Box plot
boxplot(acts, horizontal = TRUE,
        xlab = "ACT Score",
        main = "Distribution of ACT Scores")
library(ggplot2)
# 3.3b Dot plot
fit_gpas = lm(gpas ~ acts)
ggplot() + aes(fit_gpas$residuals) + geom_dotplot(binwidth = 1/7) + xlab("Residuals") + ggtitle("Dot pl
# 3.3c Res against fitted Y
beta_0 = fit_gpas$coefficients[1]
beta_1 = fit_gpas$coefficients[2]
y_hat = beta_0 + beta_1 * acts
plot(y = fit_gpas$residuals, x = y_hat,
     xlab = "Predicted Y (Y hat)", ylab = "Residuals",
     main = "Residuals against Fitted Y")

abline(0,0)
# 3.3d Normal probability plot
res <- qqnorm(fit_gpas$residuals)
qqline(fit_gpas$residuals, col = "violetred")
# correlation coefficient
r = cor(res$x, res$y)
alpha = 0.05
n = length(fit_gpas$residuals)
t_star = (r * sqrt(n - 2)) / (sqrt(1 - r^2))
# 3.3e Brown-Forsythe test
# split data
rule1 = acts < 26
rule2 = acts >= 26

res = fit_gpas$residuals

d1 = abs( res[rule1] - median(res[rule1]) )
d2 = abs( res[rule2] - median(res[rule2]) )
```

```r
s_sq = ( sum((d1 - mean(d1))^2) + sum((d2 - mean(d2))^2) ) / (n - 2)
s = sqrt(abs(s_sq))
tbf = (mean(d1) - mean(d2)) / (s * sqrt( 1/length(d1) + 1/length(d2)))

alpha = 0.01
crit = qt(1 - alpha/2, df = n - 2)
# 3.3f
  # V3 - X2 - IQ test score
  # V4 - X3 - high school rank percentile
iqs <- gpas_set$V3
per <- gpas_set$V4

# residuals plotted against other predictors
res <- fit_gpas$residuals
fit_iq <- lm(res ~ iqs)
plot(y = res, x = iqs,
     ylab = "Residuals", xlab = "IQ Test score (X2)",
     main = "GPA Residuals as a function of IQ Test Scores (X2)")
abline(h = 0)

fit_per <- lm(res ~ per)
plot(y = res, x = per,
     ylab = "Residuals", xlab = "Class Percentile (X3)",
     main = "GPA Residuals as a function of Class Percentile (X3)")
abline(h = 0)
# clear environment to prevent using old values
rm(list = ls())

earnings <- read.table("../datasets/per+capita+earnings.txt")

Y_hat <- earnings$V1
res_sst <- earnings$V2

n <- length(Y_hat)
plot(y = res_sst, x = Y_hat,
     xlab = "Fitted values", ylab = "Semistudentized Residuals",
     main = "Fitted values vs. Semistudentized Residuals")
res_sd = sd(res_sst)
m = mean(res_sst)
lower = m - res_sd
upper = m + res_sd
out_rule =  (res_sst < lower) | (res_sst > upper)
outer = res_sst[out_rule]
rm(list = ls()) # clear environment variables to prevent accidentally using old

copier_frame = read.table("../datasets/copier+maintenance.txt")
coppier_frame2 = read.table("../datasets/copier+maintenance+X2.txt")

X = copier_frame$V2
Y = copier_frame$V1
library(knitr)
fit <- lm(Y ~ X)
```

```r
b0 = fit$coefficients[1]
b1 = fit$coefficients[2]

# F statistic = MSR/MSE
SSR = sum( ( (b0 + b1 * X) - mean(Y) )^2 )
SSE = sum( ( Y - (b0 + b1 * X) )^2 )

MSE = SSE / length(X)
MSR = SSR # 1 df

F_stat = MSR / MSE

# p val
alpha = 0.05
p = 1 - pf(1 - alpha, 1, length(X) - 2)
rm(list = ls())

# 3.15 Solution concentration F test
solution_data = read.table("../datasets/solution+concentration.txt")
Y = solution_data$V1
X = solution_data$V2

fit = lm(Y ~ X)

beta_0 = fit$coefficients[1]
beta_1 = fit$coefficients[2]
SSR = sum( ( (beta_0 + beta_1 * X) - mean(Y) )^2 )
SSE = sum( ( Y - (beta_0 + beta_1 * X) )^2 )

MSE = SSE / length(X)
MSR = SSR # 1 df

F_stat = MSR / MSE

# p val
alpha = 0.05
p = 1 - pf(1 - alpha, 1, length(X) - 2)
plot(Y ~ X,
     xlab = "Time", ylab = "Solution Concentration",
     main = "Solution Concentration over Time")
box_cox_transform <- function(lambda, x, y) {
  n = length(y)
  K2 = prod(y)^(1/n)
  K1 = 1/ (lambda * K2^(lambda - 1))
  if (lambda != 0) {
    W = K1 * (y^lambda - 1)
  }
  else {
    W = K2 * log(y)
  }
  return(W)
}
```

```r
# Picking Lambda
lambdas = c(-0.2, -0.1, 0, 0.1, 0.2)
Min_SSE = 100
Min_lambda = 100

# Calculate SSE and set smallest
for (lambda in lambdas) {
  Y_prime = box_cox_transform(lambda, X, Y)
  fit_cox = lm(Y_prime ~ X)
  b0_cox = fit_cox$coefficients[1]
  b1_cox = fit_cox$coefficients[2]
  SSE = sum( ( Y - (b0_cox + b1_cox * Y_prime) )^2 )
  if (SSE < Min_SSE) {
    Min_SSE = SSE
    Min_lambda = lambda
  }
}
Y_prime = log(Y, base = 10)
fit_log = lm(Y_prime ~ X)
b0_log = fit_log$coefficients[1]
b1_log = fit_log$coefficients[2]
plot(Y_prime ~ X,
     xlab = "Time", ylab = "log(Solution Concentration)",
     main = "Solution Concentration (log scale) over Time ")
abline(b0_log, b1_log, col = "violetred")
res = fit_log$residuals
fitted = fit_log$fitted.values
plot(res ~ fitted,
     xlab = "Fitted Values for Y", ylab = "Residuals of Y'")
abline(h = 0)
qqnorm(res)
qqline(res, col = "violetred")
```