STAT 400 - Quiz 3

Colin Gibbons-Fly

Question 1

```r
# Data
x <- c(26.8, 25.4, 28.9, 23.6, 27.7, 23.9, 24.7, 28.1, 26.9, 27.4, 22.6, 25.6)
y <- c(26.5, 27.3, 24.2, 27.1, 23.6, 25.9, 26.3, 22.5, 21.7, 21.4, 25.8, 24.9)

# Part (a): Regression Line
x_mean <- mean(x)
y_mean <- mean(y)
cov_xy <- cov(x, y)
var_x <- var(x)
beta_1 <- cov_xy / var_x
beta_0 <- y_mean - beta_1 * x_mean
cat("Regression Line: y =", beta_0, "+", beta_1, "* x\n")
```

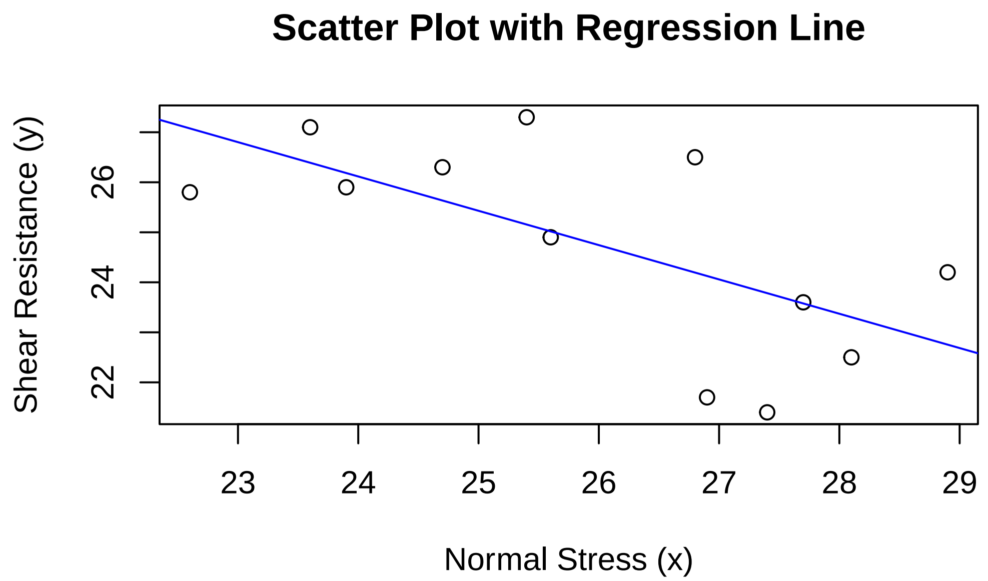
```
Regression Line: y = 42.5818 + -0.6860771 * x
```

```r
# Part (b): Prediction
x_new <- 24.5
y_pred <- beta_0 + beta_1 * x_new
cat("Predicted Shear Resistance:", y_pred, "\n")
```

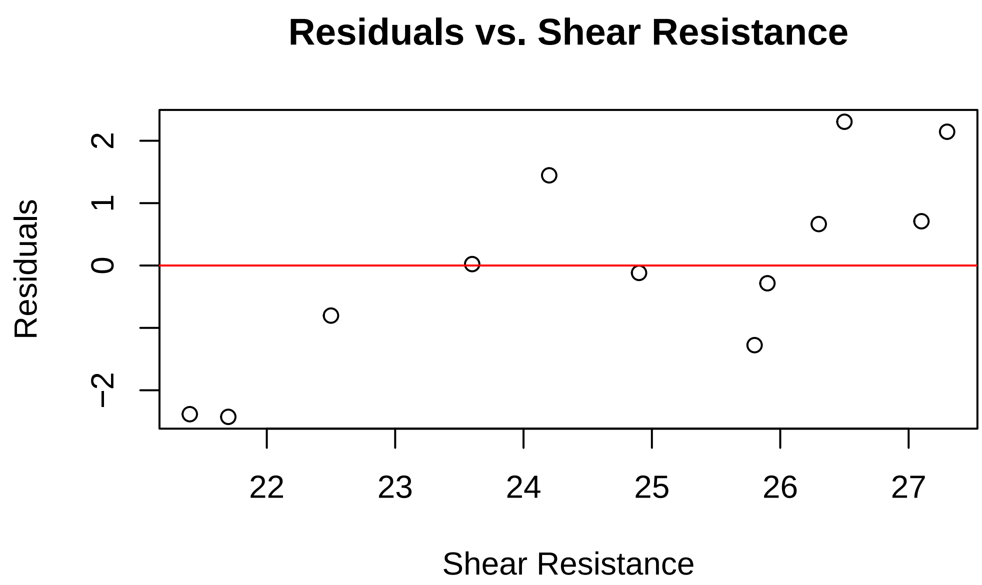
```
Predicted Shear Resistance: 25.77291
```

```r
# Part (c): Scatter Plot
plot(x, y, main="Scatter Plot with Regression Line", xlab="Normal Stress (x)", ylab="Shear Re

# Part (d): Graph Regression Line
abline(beta_0, beta_1, col="blue")
```




```
# Part (e): Residuals Plot
y_hat <- beta_0 + beta_1 * x
residuals <- y - y_hat
plot(y, residuals, main="Residuals vs. Shear Resistance", xlab="Shear Resistance", ylab="Resi
abline(h=0, col='red')
```



Question 2

```
# Part (a): Residual Variance (s^2)
SS_residual <- sum(residuals^2)
n <- length(x)
s2 <- SS_residual / (n - 2)
cat("Residual Variance (s^2):", s2, "\n")
```

```
Residual Variance (s^2): 2.688452
```

```
# Part (b): Hypothesis Test for Beta_1
SE_beta1 <- sqrt(s2 / sum((x - x_mean)^2))
t_value <- beta_1 / SE_beta1
cat("t-statistic for beta_1:", t_value, "\n")
```

```
t-statistic for beta_1: -2.745311
```

```
p_value <- 2 * (1 - pt(abs(t_value), df = n - 2))
cat("p-value for beta_1:", p_value, "\n")
```

```
p-value for beta_1: 0.02064371
```

```r
# Part (c): Confidence Interval for Beta_0
alpha <- 0.05
t_critical <- qt(1 - alpha / 2, df = n - 2)
SE_beta0 <- sqrt(s2 * (1 / n + x_mean^2 / sum((x - x_mean)^2)))
CI_beta0 <- c(beta_0 - t_critical * SE_beta0, beta_0 + t_critical * SE_beta0)
cat("95% CI for beta_0:", CI_beta0, "\n")
```

```
95% CI for beta_0: 28.08434 57.07927
```

```r
# Part (d): Confidence Interval for Beta_1
CI_beta1 <- c(beta_1 - t_critical * SE_beta1, beta_1 + t_critical * SE_beta1)
cat("95% CI for beta_1:", CI_beta1, "\n")
```

```
95% CI for beta_1: -1.242908 -0.1292458
```

Question 3

```r
# Part (a): Coefficient of Determination (R^2)
SS_total <- sum((y - y_mean)^2)
R_squared <- 1 - (SS_residual / SS_total)
cat("R-squared:", R_squared, "\n")
```

```
R-squared: 0.4297683
```

```r
# Part (b): Lack-of-Fit Test
SS_pure_error <- 0  # Replace with actual value if replicated x-values exist
SS_lack_of_fit <- SS_residual - SS_pure_error
df_pure_error <- 0  # Replace if replicated
df_lack_of_fit <- n - 2 - df_pure_error
MS_lack_of_fit <- SS_lack_of_fit / df_lack_of_fit
MS_pure_error <- ifelse(df_pure_error > 0, SS_pure_error / df_pure_error, NA)
F_lack_of_fit <- ifelse(!is.na(MS_pure_error), MS_lack_of_fit / MS_pure_error, NA)
cat("F-statistic for Lack-of-Fit:", F_lack_of_fit, "\n")
```

```
F-statistic for Lack-of-Fit: NA
```

```r
# Part (c): Hypothesis Test Using F-statistic
SS_regression <- SS_total - SS_residual
df_regression <- 1
df_residual <- n - 2
MS_regression <- SS_regression / df_regression
MS_residual <- SS_residual / df_residual
F_statistic <- MS_regression / MS_residual
cat("F-statistic for regression:", F_statistic, "\n")
```

```
F-statistic for regression: 7.536732
```

Question 4

```r
# Part (a): Correlation Coefficient
correlation <- cor(x, y)
cat("Correlation coefficient (r):", correlation, "\n")
```

```
Correlation coefficient (r): -0.6555672
```

```r
# Part (b): Hypothesis Test for Rho
rho_0 <- -0.5
SE_r <- sqrt((1 - correlation^2) / (n - 2))
t_value_r <- (correlation - rho_0) / SE_r
cat("t-statistic for testing rho = -0.5:", t_value_r, "\n")
```

```
t-statistic for testing rho = -0.5: -0.6514669
```

```r
p_value_r <- pt(t_value_r, df = n - 2)
cat("p-value for rho = -0.5:", p_value_r, "\n")
```

```
p-value for rho = -0.5: 0.2647157
```

```r
# Part (c): Percentage of Variation Explained
variation_explained <- correlation^2 * 100
cat("Percentage of variation explained by X:", variation_explained, "%\n")
```

```
Percentage of variation explained by X: 42.97683 %
```

Question 5

```r
# Given Value of X
x_new <- 24.5
SE_CI <- sqrt(s2 * (1 / n + (x_new - x_mean)^2 / sum((x - x_mean)^2)))
SE_PI <- sqrt(s2 * (1 + 1 / n + (x_new - x_mean)^2 / sum((x - x_mean)^2)))
CI <- c(beta_0 + beta_1 * x_new - t_critical * SE_CI, beta_0 + beta_1 * x_new + t_critical *
PI <- c(beta_0 + beta_1 * x_new - t_critical * SE_PI, beta_0 + beta_1 * x_new + t_critical *
cat("95% Confidence Interval for mean response:", CI, "\n")
```

```
95% Confidence Interval for mean response: 24.43903 27.10679
```

```r
cat("95% Prediction Interval for individual observation:", PI, "\n")
```

```
95% Prediction Interval for individual observation: 21.88365 29.66217
```

```r
# Find Lowest Standard Error
SE_values <- sqrt(s2 * (1 + 1 / n + (x - x_mean)^2 / sum((x - x_mean)^2)))
min_SE <- min(SE_values)
min_SE_index <- which.min(SE_values)
cat("Observation with lowest SE (index):", min_SE_index, "\n")
```

```
Observation with lowest SE (index): 12
```

```r
cat("Standard Error for this observation:", min_SE, "\n")
```

```
Standard Error for this observation: 1.70906
```

Question 6

```r
# Data
x1 <- c(14.62, 15.63, 14.62, 15.00, 14.50, 15.25, 16.12, 15.13, 15.50, 15.13, 15.50, 16.12,
x2 <- c(226.0, 220.0, 217.4, 220.0, 226.5, 224.1, 220.5, 223.5, 217.6, 228.5, 230.2, 226.5,
x3 <- c(7.000, 3.375, 6.375, 6.000, 7.625, 6.000, 3.375, 6.125, 5.000, 6.625, 5.750, 3.750,
y <- c(128.40, 52.62, 113.90, 98.01, 139.90, 102.60, 48.14, 109.60, 82.68, 112.60, 97.52, 59

# Fit multiple linear regression model
model <- lm(y ~ x1 + x2 + x3)
summary(model)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9517 -2.3992  0.3168  1.8602  7.8955 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -21.4696    51.6756  -0.415    0.684    
x1           -3.3243     3.5888  -0.926    0.369    
x2            0.2465     0.2747   0.897    0.384    
x3           20.3448     1.2576  16.178 6.65e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.151 on 15 degrees of freedom
Multiple R-squared:  0.9916,	Adjusted R-squared:  0.9899 
F-statistic: 588.1 on 3 and 15 DF,  p-value: 8.976e-16
```

```r
# Part (b): Prediction for x1 = 14, x2 = 220, x3 = 5
new_data <- data.frame(x1 = 14, x2 = 220, x3 = 5)
y_pred <- predict(model, newdata = new_data)
cat("Predicted y:", y_pred, "\n")
```

```
Predicted y: 87.94123
```

Question 7

```r
# Full model
full_model <- lm(y ~ x1 + x2 + x3)

# Reduced model (excluding x1)
reduced_model <- lm(y ~ x2 + x3)

# Part (a): Compare Adjusted R^2
R2_adj_full <- summary(full_model)$adj.r.squared
R2_adj_reduced <- summary(reduced_model)$adj.r.squared
cat("Adjusted R^2 (Full Model):", R2_adj_full, "\n")
```

```
Adjusted R^2 (Full Model): 0.9898834
```

```r
cat("Adjusted R^2 (Reduced Model):", R2_adj_reduced, "\n")
```

```
Adjusted R^2 (Reduced Model): 0.9899731
```

```r
# Part (b): Compare Prediction Interval Widths
new_data <- data.frame(x2 = 220, x3 = 5)

# Full model prediction interval
PI_full <- predict(full_model, new_data = new_data, interval = "prediction")
```

```
Warning in predict.lm(full_model, new_data = new_data, interval = "prediction"): predictions
```

```r
# Reduced model prediction interval
PI_reduced <- predict(reduced_model, new_data = new_data, interval = "prediction")
```

```
Warning in predict.lm(reduced_model, new_data = new_data, interval = "prediction"): predicti
```

```r
# Calculate and print widths
PI_width_full <- PI_full[3] - PI_full[2]
PI_width_reduced <- PI_reduced[3] - PI_reduced[2]
cat("Prediction Interval Width (Full Model):", PI_width_full, "\n")
```

```
Prediction Interval Width (Full Model): 63.75113
```

```r
cat("Prediction Interval Width (Reduced Model):", PI_width_reduced, "\n")
```

```
Prediction Interval Width (Reduced Model): 63.81373
```

Question 8

```r
# ANOVA for the full model
anova_full <- anova(full_model)
print(anova_full)
```

```
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)    
x1         1 18252.3 18252.3 1059.42 2.490e-15 ***
x2         1  7634.1  7634.1  443.11 1.501e-12 ***
x3         1  4509.2  4509.2  261.73 6.646e-11 ***
Residuals 15   258.4    17.2                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Part (a): Sum of Squares
SS_regression <- sum(anova_full$`Sum Sq`[-length(anova_full$`Sum Sq`)])
SS_residual <- anova_full$`Sum Sq`[length(anova_full$`Sum Sq`)]
SS_total <- SS_regression + SS_residual

# Degrees of Freedom
df_regression <- sum(anova_full$Df[-length(anova_full$Df)])
df_residual <- anova_full$Df[length(anova_full$Df)]
df_total <- df_regression + df_residual

# Mean Squares
MS_regression <- SS_regression / df_regression
MS_residual <- SS_residual / df_residual

# F-statistic
F_statistic <- MS_regression / MS_residual
cat("F-statistic:", F_statistic, "\n")
```

```
F-statistic: 588.0842
```

```r
# p-value
p_value <- pf(F_statistic, df1 = df_regression, df2 = df_residual, lower.tail = FALSE)
cat("p-value:", p_value, "\n")
```

```
p-value: 8.975576e-16
```

Question 9

```r
# Data
profit <- c(157, -181, -253, 158, 75, 202, -451, 146, 89, -357, 522, 78, 5, -177, 123, 251, -
income <- c(45000, 55000, 45800, 38000, 75000, 99750, 28000, 39000, 54350, 32500, 36750, 425
gender <- c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) # 1 = Male, 0 = Femal
family_members <- c(1, 2, 4, 3, 4, 4, 1, 2, 1, 1, 1, 3, 2, 3, 2, 1, 1, 1, 1, 2)

# Fit model
credit_model <- lm(profit ~ income + gender + family_members)
summary(credit_model)
```

```
Call:
lm(formula = profit ~ income + gender + family_members)

Residuals:
    Min      1Q  Median      3Q     Max 
-347.24 -150.85    7.16  132.66  341.49 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)  
(Intercept)     3.008e+01  1.267e+02   0.237   0.8153  
income          5.433e-03  2.741e-03   1.982   0.0649 .
gender         -2.367e+02  1.106e+02  -2.141   0.0480 *
family_members -4.924e+01  5.196e+01  -0.948   0.3574  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 227.5 on 16 degrees of freedom
Multiple R-squared:  0.3075,	Adjusted R-squared:  0.1777 
F-statistic: 2.368 on 3 and 16 DF,  p-value: 0.1091
```

Question 10

```r
# -----------------------------------------------------------------------------------------------
# Question 10: Regression Model Selection

# Data
data <- data.frame(
  Y  = c(11.2, 14.5, 17.2, 17.8, 19.3, 24.5, 21.2, 16.9, 14.8, 20.0, 13.2, 22.5),
```

```r
  X1 = c(56.5, 59.5, 69.2, 74.5, 81.2, 88.0, 78.2, 69.0, 58.1, 80.5, 58.3, 84.0),
  X2 = c(71.0, 72.5, 76.0, 79.5, 84.0, 86.2, 80.5, 72.0, 68.0, 85.0, 71.0, 87.2),
  X3 = c(38.5, 38.2, 42.5, 43.4, 47.5, 47.4, 44.5, 41.8, 42.1, 48.1, 37.5, 51.0),
  X4 = c(43.0, 44.8, 49.0, 56.3, 60.2, 62.0, 58.1, 48.1, 46.0, 60.3, 47.1, 65.2)
)

# Part (a): Forward Selection

# Null model (no predictors)
null_model <- lm(Y ~ 1, data = data)

# Full model (all predictors)
full_model <- lm(Y ~ ., data = data)

# Perform forward selection
forward_model <- step(null_model, scope = list(lower = null_model, upper = full_model), dire
```

```
Start:  AIC=33.91
Y ~ 1

       Df Sum of Sq     RSS    AIC
+ X1    1    158.41  12.978  4.940
+ X4    1    145.29  26.100 13.324
+ X2    1    136.01  35.380 16.975
+ X3    1    133.65  37.741 17.750
<none>              171.389 33.908

Step:  AIC=4.94
Y ~ X1

       Df Sum of Sq    RSS    AIC
<none>              12.978 4.9404
+ X2    1   1.92969 11.049 5.0088
+ X4    1   0.02886 12.949 6.9137
+ X3    1   0.01684 12.961 6.9249
```

```r
# Output summary of the selected model
cat("Forward Selection Model Summary:\n")
```

```
Forward Selection Model Summary:
```

```r
summary(forward_model)
```

```
Call:
lm(formula = Y ~ X1, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.75899 -0.86677  0.07325  0.85826  1.53439 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.33592    2.20553  -2.873   0.0166 *  
X1           0.33738    0.03054  11.048 6.33e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.139 on 10 degrees of freedom
Multiple R-squared:  0.9243,    Adjusted R-squared:  0.9167 
F-statistic: 122.1 on 1 and 10 DF,  p-value: 6.331e-07
```

```r
# Part (b): Backward Elimination

# Perform backward elimination
backward_model <- step(full_model, direction = "backward")
```

```
Start:  AIC=8.34
Y ~ X1 + X2 + X3 + X4

       Df Sum of Sq    RSS     AIC
- X3    1    0.0065 10.460  6.3515
- X4    1    0.3963 10.850  6.7905
<none>              10.453  8.3440
- X2    1    2.4315 12.885  8.8536
- X1    1   15.0455 25.499 17.0446

Step:  AIC=6.35
Y ~ X1 + X2 + X4

       Df Sum of Sq    RSS     AIC
- X4    1    0.5889 11.049  5.0088
```

```
<none>              10.460  6.3515
- X2    1    2.4897 12.949  6.9137
- X1    1   15.6378 26.098 15.3232

Step:  AIC=5.01
Y ~ X1 + X2

       Df Sum of Sq    RSS     AIC
- X2    1    1.9297 12.978  4.9404
<none>              11.049  5.0088
- X1    1   24.3318 35.380 16.9750

Step:  AIC=4.94
Y ~ X1

       Df Sum of Sq     RSS    AIC
<none>               12.978  4.940
- X1    1    158.41 171.389 33.908
```

```r
# Output summary of the selected model
cat("Backward Elimination Model Summary:\n")
```

```
Backward Elimination Model Summary:
```

```r
summary(backward_model)
```

```
Call:
lm(formula = Y ~ X1, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.75899 -0.86677  0.07325  0.85826  1.53439 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.33592    2.20553  -2.873   0.0166 *  
X1           0.33738    0.03054  11.048 6.33e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.139 on 10 degrees of freedom
Multiple R-squared:  0.9243,	Adjusted R-squared:  0.9167
F-statistic: 122.1 on 1 and 10 DF,  p-value: 6.331e-07
```

```r
# Part (c): Stepwise Regression

# Perform stepwise regression
stepwise_model <- step(null_model, scope = list(lower = null_model, upper = full_model), dire
```

```
Start:  AIC=33.91
Y ~ 1

       Df Sum of Sq     RSS    AIC
+ X1    1    158.41  12.978  4.940
+ X4    1    145.29  26.100 13.324
+ X2    1    136.01  35.380 16.975
+ X3    1    133.65  37.741 17.750
<none>              171.389 33.908

Step:  AIC=4.94
Y ~ X1

       Df Sum of Sq     RSS    AIC
<none>               12.978  4.940
+ X2    1     1.930  11.049  5.009
+ X4    1     0.029  12.949  6.914
+ X3    1     0.017  12.961  6.925
- X1    1   158.411 171.389 33.908
```

```r
# Output summary of the selected model
cat("Stepwise Regression Model Summary:\n")
```

```
Stepwise Regression Model Summary:
```

```r
summary(stepwise_model)
```

```
Call:
lm(formula = Y ~ X1, data = data)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.75899 -0.86677  0.07325  0.85826  1.53439 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.33592    2.20553  -2.873   0.0166 *  
X1           0.33738    0.03054  11.048 6.33e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.139 on 10 degrees of freedom
Multiple R-squared:  0.9243,	Adjusted R-squared:  0.9167 
F-statistic: 122.1 on 1 and 10 DF,  p-value: 6.331e-07
```