Homework 5 - P8130 Biostatistics Method I

Yongyan Liu (yl6107)

2025-12-16

Problem 1 (20 points)

Background: Accurately predicting clinical outcomes is a central challenge in public health and medical research. For biomarkers like HbA1c—an indicator of long-term blood glucose regulation—building a reliable predictive model can support early identification of individuals at risk for diabetes or metabolic disease. Because health data often include multiple, correlated predictors, selecting an appropriate model is as important as the modeling technique itself.

The data used in this assignment comes from a community health screening initiative conducted through several primary-care clinics in the southeastern United States. As part of routine preventive visits, adults were assessed on demographic, lifestyle, and physiological measures, including age, BMI, diet, and physical activity. The data for this exercise are in the data file "HbA1c.csv".

```
hba1c <- read.csv("data/HbA1c.csv")
```

Part (a)

Question: Fit the full linear regression model to the data and examine standard diagnostics. Are there any suggestions of violation of regression assumptions?

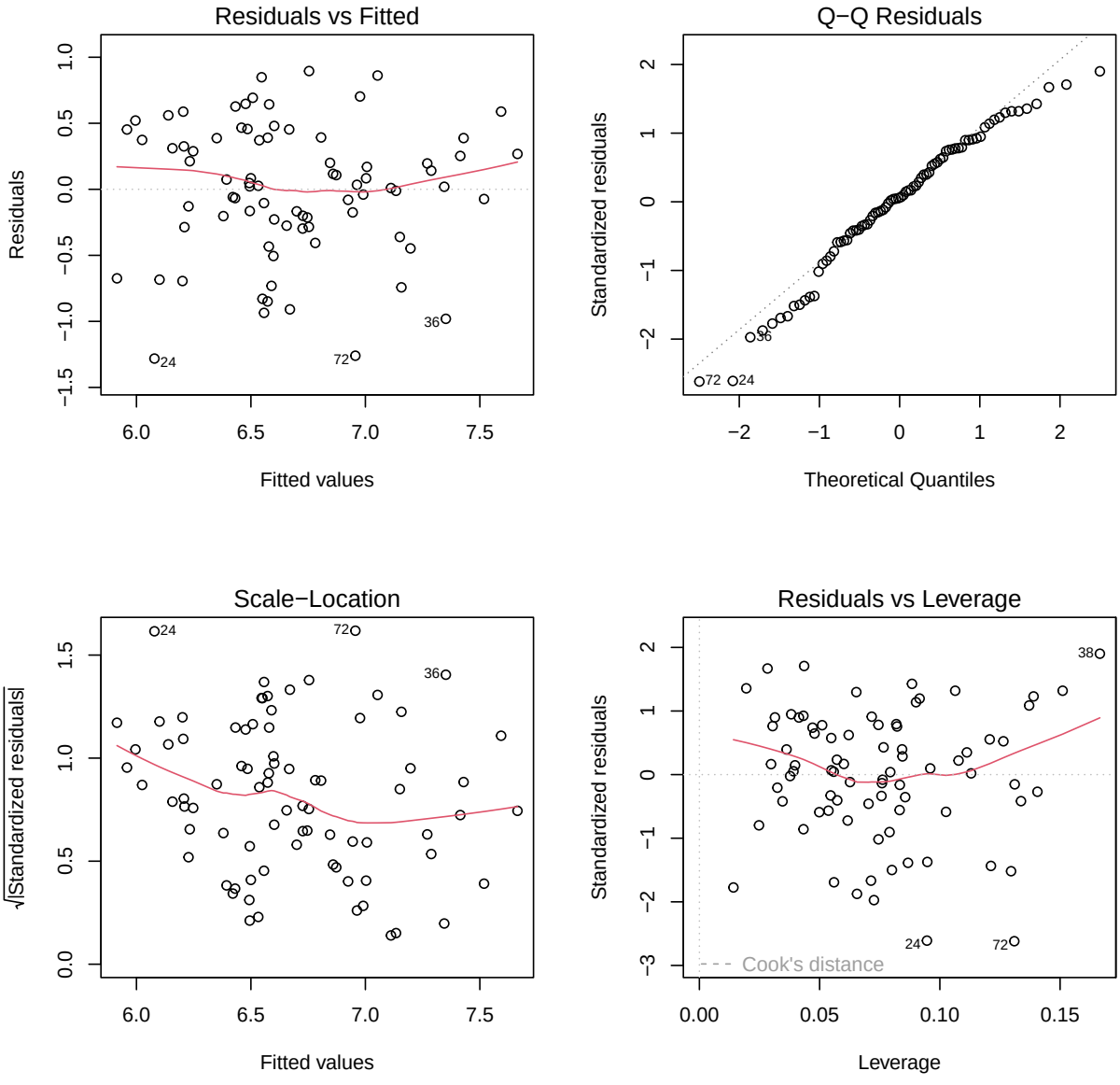
Solution:

```
# Fit full model
full_model <- lm(hba1c ~ age + bmi + physical_activity + diet_score + digit_ratio,
                 data = hba1c)
summary(full_model)
```

```
##
## Call:
## lm(formula = hba1c ~ age + bmi + physical_activity + diet_score +
##     digit_ratio, data = hba1c)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28150 -0.27771  0.03032  0.38759  0.89550
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7.917839   1.972321   4.014 0.000141 ***
## age               0.021879   0.006284   3.482 0.000839 ***
```

```
## bmi                0.043460   0.018159   2.393 0.019232 *
## physical_activity  0.056189   0.031603   1.778 0.079520 .
## diet_score        -0.014306   0.006353  -2.252 0.027306 *
## digit_ratio       -2.940292   1.934180  -1.520 0.132729
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5161 on 74 degrees of freedom
## Multiple R-squared:  0.3973, Adjusted R-squared:  0.3566
## F-statistic: 9.756 on 5 and 74 DF,  p-value: 3.457e-07
```

```r
# Diagnostic plots
par(mfrow = c(2, 2))
plot(full_model)
```



Interpretation:

- **Residuals vs Fitted:** The residuals appear randomly scattered around zero with no clear pattern, suggesting linearity is satisfied.
- **Normal Q-Q:** Points generally follow the diagonal line, indicating approximate normality of residuals.
- **Scale-Location:** The spread of residuals appears relatively constant across fitted values, suggesting homoscedasticity.
- **Residuals vs Leverage:** No points have both high leverage and high residuals (no influential outliers beyond Cook's distance threshold).

Overall, there are no major violations of regression assumptions. The model assumptions appear reasonably satisfied.

Part (b)

Question: Apply the LASSO regression procedure to the data using 10-fold cross-validation for choosing the tuning parameter. Fit LASSO models to the data with two different selections of lambda – one using the minimum of the CV function and one using the "1se". Write an expression for the fitted regression model in each case.

Solution:

```
# Prepare data for glmnet
X <- as.matrix(hba1c[, c("age", "bmi", "physical_activity", "diet_score", "digit_ratio")])
y <- hba1c$hba1c

# 10-fold CV LASSO
set.seed(123)
cv_lasso <- cv.glmnet(X, y, alpha = 1, nfolds = 10)
cv_lasso
```

```
##
## Call:  cv.glmnet(x = X, y = y, nfolds = 10, alpha = 1)
##
## Measure: Mean-Squared Error
##
##      Lambda Index Measure      SE Nonzero
## min 0.00116    62  0.2999 0.05797       5
## 1se 0.16086     9  0.3565 0.07364       2
```

```
# Coefficients for lambda.min
coef(cv_lasso, s = "lambda.min")
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                     lambda.min
## (Intercept)        7.87858660
## age                0.02180192
## bmi                0.04323741
## physical_activity  0.05533419
## diet_score        -0.01415508
## digit_ratio       -2.89476515
```

```r
# Coefficients for lambda.1se
coef(cv_lasso, s = "lambda.1se")
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                    lambda.1se
## (Intercept)       5.76410107
## age               0.01256734
## bmi               0.01335787
## physical_activity .
## diet_score        .
## digit_ratio       .
```

Fitted Models:

Model 1 (lambda.min = 0.0012, log(lambda.min) = -6.76): Uses all predictors with minimal shrinkage:

$$\widehat{\text{HbA1c}} = 7.879 + 0.022 \cdot \text{age} + 0.043 \cdot \text{bmi} + 0.055 \cdot \text{physical\_activity} - 0.014 \cdot \text{diet\_score} - 2.895 \cdot \text{digit\_ratio}$$

Model 2 (lambda.1se = 0.161, log(lambda.1se) = -1.83): More parsimonious model with stronger regularization. Physical activity, diet score, and digit ratio are shrunk to zero:

$$\widehat{\text{HbA1c}} = 5.764 + 0.013 \cdot \text{age} + 0.013 \cdot \text{bmi}$$

Part (c)

Question: Use each model to predict HbA1c in the data. Plot fits vs. residuals for each model. Based on the plots and the MSPE, which model would be preferred?

Solution:

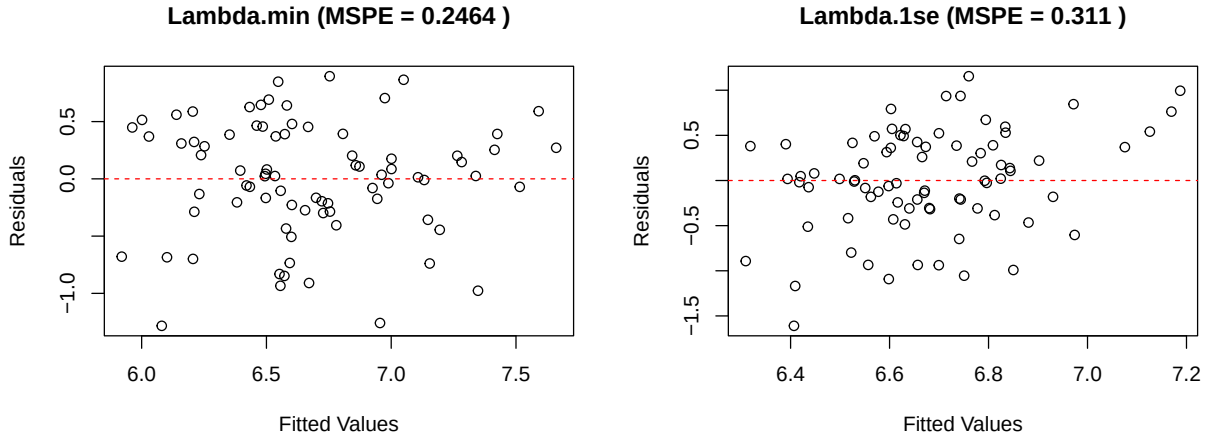
```r
# Predictions
pred_min <- predict(cv_lasso, newx = X, s = "lambda.min")
pred_1se <- predict(cv_lasso, newx = X, s = "lambda.1se")

# Residuals
resid_min <- y - pred_min
resid_1se <- y - pred_1se

# MSPE (in-sample)
mspe_min <- mean(resid_min^2)
mspe_1se <- mean(resid_1se^2)
cat("MSPE (lambda.min):", mspe_min, "\n")
```

```
## MSPE (lambda.min): 0.2463549
```

```r
cat("MSPE (lambda.1se):", mspe_1se, "\n")
```

```
## MSPE (lambda.1se): 0.3109598
```

```r
# Plots
par(mfrow = c(1, 2))
plot(pred_min, resid_min, xlab = "Fitted Values", ylab = "Residuals",
     main = paste("Lambda.min (MSPE =", round(mspe_min, 4), ")"))
abline(h = 0, col = "red", lty = 2)

plot(pred_1se, resid_1se, xlab = "Fitted Values", ylab = "Residuals",
     main = paste("Lambda.1se (MSPE =", round(mspe_1se, 4), ")"))
abline(h = 0, col = "red", lty = 2)
```



Interpretation:

Both residual plots show random scatter around zero with no obvious patterns, indicating both models fit reasonably well.

- **In-sample MSPE (lambda.min):** 0.2464
- **In-sample MSPE (lambda.1se):** 0.3110

Preferred model: Based on in-sample MSPE and the residual plots, the **lambda.min model is preferred**. It has a lower MSPE (0.2464 vs 0.3110), indicating better fit to the training data. Both models show similar residual patterns with no violations of assumptions, but the lambda.min model explains more variance in the outcome.

However, it's important to note that in-sample MSPE tends to favor more complex models and may reflect overfitting. Out-of-sample validation would provide a more reliable assessment of predictive performance.

Part (d)

Question: A new dataset was gathered – it is in the data file "HbA1c_val.csv". Use each model from (c) (fitted only to the original data) to predict HbA1c in the new data. Based on these out-of-sample predictions, which model would be preferred?

Solution:

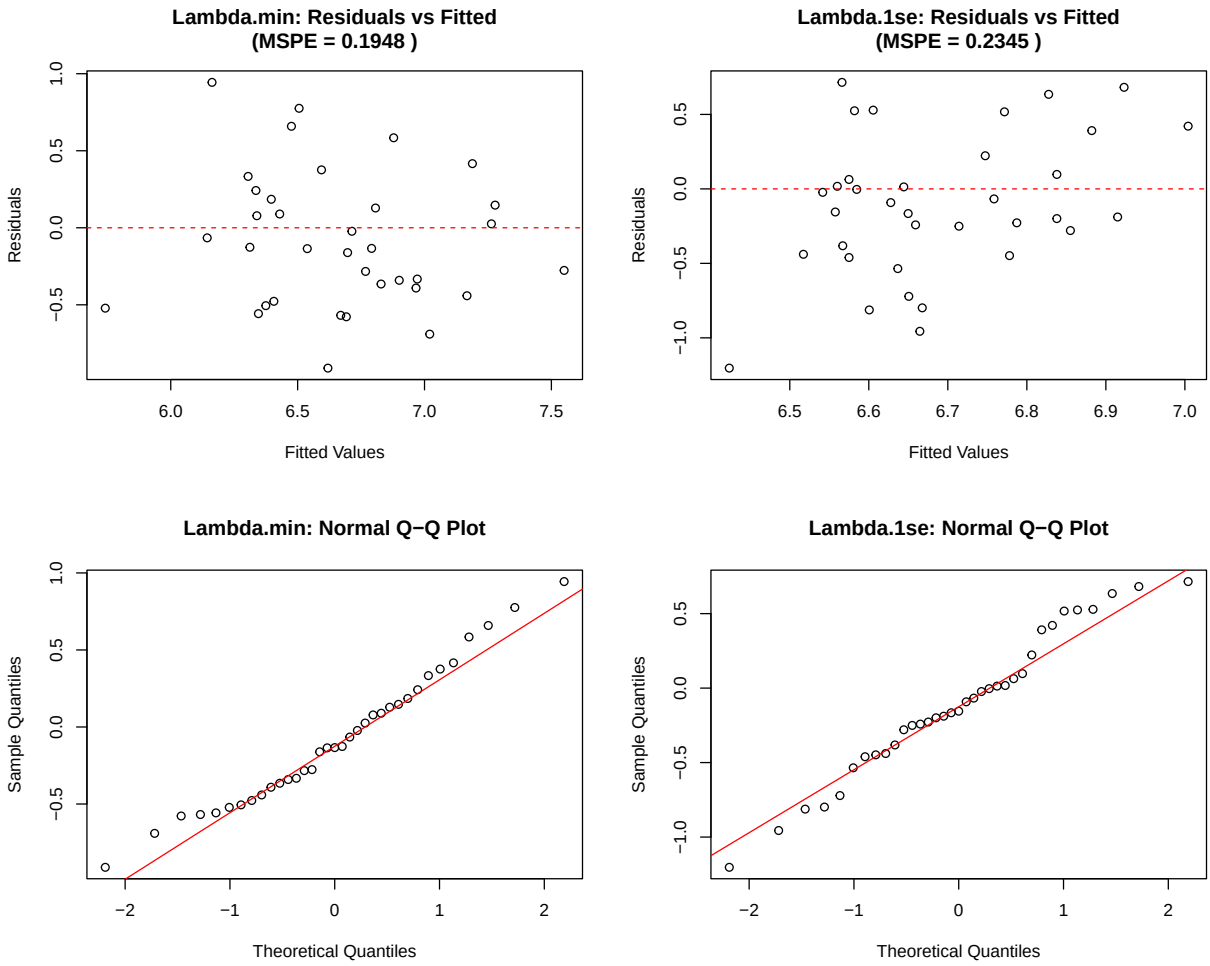
```r
hba1c_val <- read.csv("data/HbA1c_val.csv")

# Prepare validation data
X_val <- as.matrix(hba1c_val[, c("age", "bmi", "physical_activity", "diet_score", "digit_ratio")])
y_val <- hba1c_val$hba1c

# Predictions on validation set
pred_val_min <- predict(cv_lasso, newx = X_val, s = "lambda.min")
pred_val_1se <- predict(cv_lasso, newx = X_val, s = "lambda.1se")

# Residuals on validation set
resid_val_min <- y_val - pred_val_min
resid_val_1se <- y_val - pred_val_1se

# Out-of-sample MSPE
mspe_val_min <- mean(resid_val_min^2)
mspe_val_1se <- mean(resid_val_1se^2)

cat("Out-of-sample MSPE (lambda.min):", mspe_val_min, "\n")
```

```
## Out-of-sample MSPE (lambda.min): 0.1947502
```

```r
cat("Out-of-sample MSPE (lambda.1se):", mspe_val_1se, "\n")
```

```
## Out-of-sample MSPE (lambda.1se): 0.2344715
```

```r
# Diagnostic plots for validation data
par(mfrow = c(2, 2))

# Residuals vs Fitted for lambda.min
plot(pred_val_min, resid_val_min, xlab = "Fitted Values", ylab = "Residuals",
     main = paste("Lambda.min: Residuals vs Fitted\n(MSPE =", round(mspe_val_min, 4), ")"))
abline(h = 0, col = "red", lty = 2)

# Residuals vs Fitted for lambda.1se
plot(pred_val_1se, resid_val_1se, xlab = "Fitted Values", ylab = "Residuals",
     main = paste("Lambda.1se: Residuals vs Fitted\n(MSPE =", round(mspe_val_1se, 4), ")"))
abline(h = 0, col = "red", lty = 2)

# Q-Q plot for lambda.min
qqnorm(resid_val_min, main = "Lambda.min: Normal Q-Q Plot")
qqline(resid_val_min, col = "red")

# Q-Q plot for lambda.1se
qqnorm(resid_val_1se, main = "Lambda.1se: Normal Q-Q Plot")
qqline(resid_val_1se, col = "red")
```



Interpretation:

The out-of-sample MSPE provides a more honest assessment of model performance:

- **Out-of-sample MSPE (lambda.min):** 0.1948
- **Out-of-sample MSPE (lambda.1se):** 0.2345

Diagnostic plots on validation data:

- **Residuals vs Fitted:** Both models show random scatter around zero with no obvious patterns, indicating the linear relationship holds on new data.
- **Normal Q-Q plots:** Both models show residuals that follow the diagonal line reasonably well, suggesting approximate normality of prediction errors.

Preferred model: The **lambda.min model** is preferred based on:

1. Lower out-of-sample MSPE (0.1948 vs 0.2345), indicating approximately 17% better predictive accuracy
2. Similar diagnostic plot patterns between the two models, with no violations of assumptions
3. The lower out-of-sample MSPE is unusual but suggests the validation set might represent a slightly 'easier' subset to predict or has less inherent variability than the training set.

Part (e)

Question: Considering prediction accuracy and the Principle of Parsimony, comment on the relative merits of the two fitted models.

Solution:

Principle of Parsimony (Occam's Razor): Among models with similar predictive performance, prefer the simpler model.

Comparison:

Criterion	Lambda.min	Lambda.1se
Number of predictors	5 (all)	2 (age, bmi)
In-sample MSPE	0.2464	0.3110
Out-of-sample MSPE	0.1948	0.2345

Discussion:

- **Lambda.min model:** Includes all 5 predictors with minimal shrinkage. Lower MSPE on both training and validation data.
- **Lambda.1se model:** More parsimonious with only 2 non-zero coefficients (age and bmi). Physical activity, diet score, and digit ratio are excluded.

In this case, the **lambda.min model is preferred** because:

1. It has substantially lower out-of-sample MSPE (0.1948 vs 0.2345), indicating better generalization
2. The additional predictors (physical activity, diet score, digit ratio) appear to provide meaningful predictive information
3. The model does not appear to be overfitting, as out-of-sample MSPE is actually lower than in-sample MSPE

However, if interpretability is a primary concern, the lambda.1se model offers a simpler explanation with only age and BMI as predictors, at the cost of approximately 20% higher prediction error.

Problem 2 (20 points)

Background: Asthma is a common chronic respiratory condition and accurately understanding the factors that influence symptom severity is critical for guiding public health interventions and improving patient care. Environmental exposures, such as air pollution, as well as lifestyle and household factors, can affect both the level and variability of symptoms.

Adult participants were surveyed for demographic and lifestyle factors, including physical activity and exposure to household smoking, and nearby PM2.5 measurements were collected. The outcome variable represents a continuous measure of daily asthma symptom severity, with higher values indicating more severe symptoms. The data for this exercise are in the data file "asthma.csv".

```
asthma <- read.csv("data/asthma.csv")
```

Part (a)

Question: Fit the full regression model to the data. For each of the predictor variables examine a plot:

- Of the residuals vs. the predictor
- Of the absolute values of the residuals vs. the predictor

Do any of these plots show apparent patterns of non-constant variance?

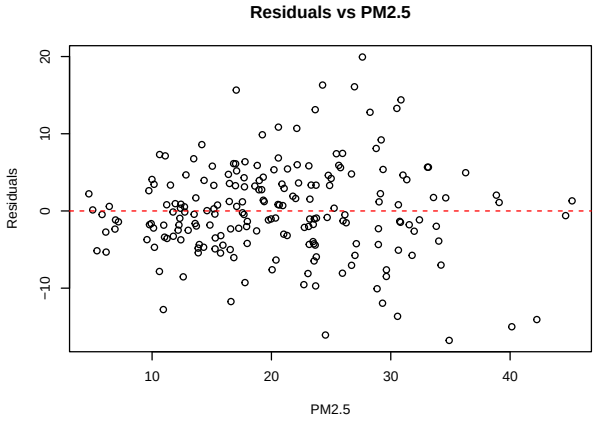
Solution:

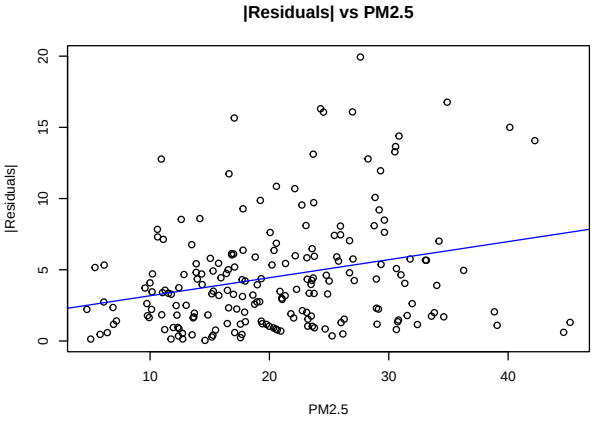
```
# Fit full model
asthma_model <- lm(symptom_sev ~ pm25 + activity + smoke_exposure, data = asthma)
summary(asthma_model)
```

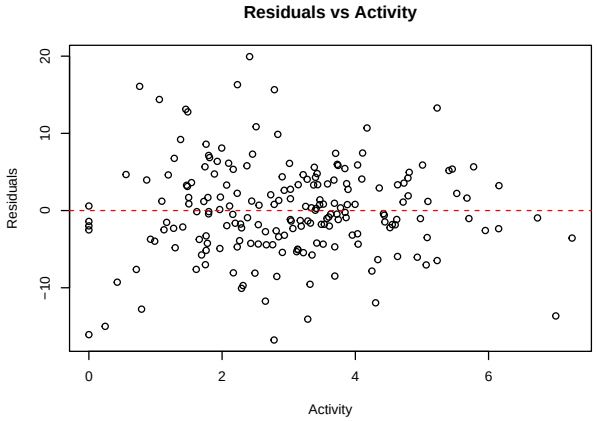
```
##
## Call:
## lm(formula = symptom_sev ~ pm25 + activity + smoke_exposure,
##     data = asthma)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.7681  -3.5137  -0.4183   3.5019  19.9389
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.97883    1.47784   2.692  0.00771 **
## pm25            0.05795    0.05215   1.111  0.26779
## activity       -0.53261    0.29893  -1.782  0.07634 .
## smoke_exposure  0.45167    0.91525   0.493  0.62222
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.029 on 196 degrees of freedom
## Multiple R-squared:  0.02336,    Adjusted R-squared:  0.00841
## F-statistic: 1.563 on 3 and 196 DF,  p-value: 0.1998
```

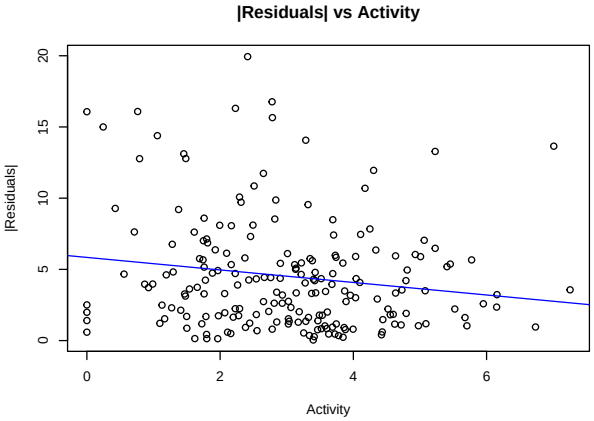
```
resid <- residuals(asthma_model)
abs_resid <- abs(resid)
predictors <- c("pm25", "activity")
labels <- c("PM2.5", "Activity")

par(mfrow = c(3, 2))
for (i in seq_along(predictors)) {
  x <- asthma[[predictors[i]]]
  # Residuals vs predictor
  plot(x, resid, xlab = labels[i], ylab = "Residuals", main = paste("Residuals vs", labels[i]))
  abline(h = 0, col = "red", lty = 2)
  # |Residuals| vs predictor
  plot(x, abs_resid, xlab = labels[i], ylab = "|Residuals|", main = paste("|Residuals| vs", labels[i]))
  abline(lm(abs_resid ~ x), col = "blue")
}
```

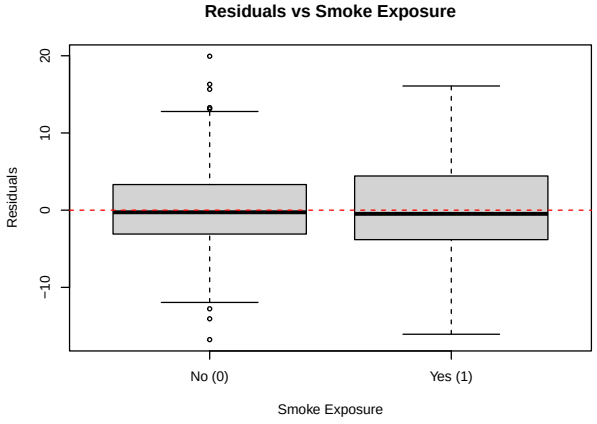
```r
# Smoke exposure (binary) - use boxplots
boxplot(resid ~ asthma$smoke_exposure, xlab = "Smoke Exposure", ylab = "Residuals",
        main = "Residuals vs Smoke Exposure", names = c("No (0)", "Yes (1)"))
abline(h = 0, col = "red", lty = 2)
boxplot(abs_resid ~ asthma$smoke_exposure, xlab = "Smoke Exposure", ylab = "|Residuals|",
        main = "|Residuals| vs Smoke Exposure", names = c("No (0)", "Yes (1)"))
```

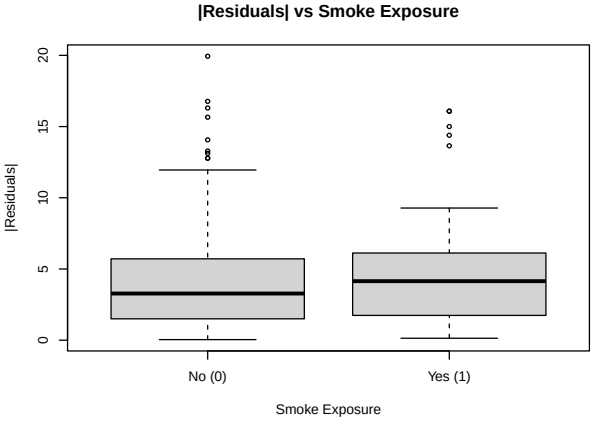












Interpretation:

Examining the plots:

- **PM2.5:** The residuals vs PM2.5 plot shows a **funnel/fan pattern** - variance appears to increase with PM2.5. The absolute residuals plot shows a positive trend, confirming heteroscedasticity related to PM2.5.
- **Activity:** The absolute residuals plot shows a **negative trend** - variance appears to decrease as activity increases. This suggests possible heteroscedasticity related to activity level.
- **Smoke exposure:** The boxplots show relatively similar spread in both groups, indicating no clear pattern of non-constant variance for this binary predictor.

Conclusion: There is evidence of **non-constant variance (heteroscedasticity)** primarily related to PM2.5 levels (variance increasing with PM2.5). There is also a weaker pattern suggesting variance may decrease with higher activity levels.

Part (b)

Question: Regress the squared residuals from the model in (a) on PM2.5. Is there any apparent relationship between these two variables?

Solution:

```r
# Squared residuals
resid_sq <- resid^2

# Regress squared residuals on PM2.5
var_model <- lm(resid_sq ~ pm25, data = asthma)
summary(var_model)
```

```
##
## Call:
## lm(formula = resid_sq ~ pm25, data = asthma)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -83.95 -29.11 -15.10   2.98 347.68
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.9869    11.3388  -0.528 0.598093
## pm25          2.0223     0.5121   3.949 0.000109 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.29 on 198 degrees of freedom
## Multiple R-squared:  0.07302,    Adjusted R-squared:  0.06834
## F-statistic:  15.6 on 1 and 198 DF,  p-value: 0.0001089
```

Interpretation:

The regression output shows: - $\hat{\gamma}_0 = -5.99$ (intercept, p = 0.598, not significant) - $\hat{\gamma}_1 = 2.02$ (PM2.5 coefficient, p = 0.0001, highly significant)

The coefficient for PM2.5 is statistically significant ($p < 0.001$), providing strong evidence of a relationship between the variance of residuals and PM2.5, confirming heteroscedasticity.

The model: $\hat{\sigma}_i^2 = -5.99 + 2.02 \cdot \text{PM2.5}_i$

This relationship indicates that variance increases linearly with PM2.5 concentration—for each unit increase in PM2.5, the residual variance increases by approximately 2.02 units. While the intercept is negative, all observed PM2.5 values in the dataset are sufficiently large (>3) to ensure positive predicted variance estimates.

Part (c)

Question: Using the results from (b), calculate appropriate weights and, using these weights, refit the model in (a). Compare the parameter estimates between the unweighted and weighted version. Which do you think provides a more accurate model for these data? Why?

Solution:

Theory: For weighted least squares (WLS), the appropriate weights are $w_i = 1/\hat{\sigma}^2$.

```
coef(var_model)
```

```
## (Intercept)        pm25
##   -5.986857    2.022334
```

```
# Calculate weights
pred_var <- predict(var_model)
weights <- 1 / pred_var

# Refit model with weights
weighted_model <- lm(symptom_sev ~ pm25 + activity + smoke_exposure,
                     data = asthma, weights = weights)
summary(weighted_model)
```

```
##
## Call:
## lm(formula = symptom_sev ~ pm25 + activity + smoke_exposure,
##     data = asthma, weights = weights)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -2.95483 -0.66441 -0.02522  0.60942  2.99048
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.34095    1.08302   2.162   0.0319 *
## pm25            0.11691    0.04626   2.527   0.0123 *
## activity       -0.41608    0.24318  -1.711   0.0887 .
## smoke_exposure  0.68185    0.84367   0.808   0.4200
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.011 on 196 degrees of freedom
## Multiple R-squared:  0.05489,    Adjusted R-squared:  0.04043
## F-statistic: 3.795 on 3 and 196 DF,  p-value: 0.01123
```

```r
# Compare coefficients
comparison <- data.frame(
  Unweighted = coef(asthma_model),
  Weighted = coef(weighted_model)
)
print(round(comparison, 4))
```

```
##                Unweighted Weighted
## (Intercept)        3.9788   2.3410
## pm25               0.0580   0.1169
## activity          -0.5326  -0.4161
## smoke_exposure     0.4517   0.6819
```

```r
# Compare standard errors
se_comparison <- data.frame(
  Unweighted = summary(asthma_model)$coefficients[, 2],
  Weighted = summary(weighted_model)$coefficients[, 2]
)
print(round(se_comparison, 4))
```

```
##                Unweighted Weighted
## (Intercept)        1.4778   1.0830
## pm25               0.0521   0.0463
## activity           0.2989   0.2432
## smoke_exposure     0.9153   0.8437
```

Interpretation:

Key observations:

- The PM2.5 coefficient increases from 0.058 (not significant, p=0.27) to 0.117 (significant, p=0.01) after weighting
- Standard errors are generally smaller in the weighted model
- WLS gives more weight to observations with lower variance (lower PM2.5 levels)

The **weighted model is preferred**.

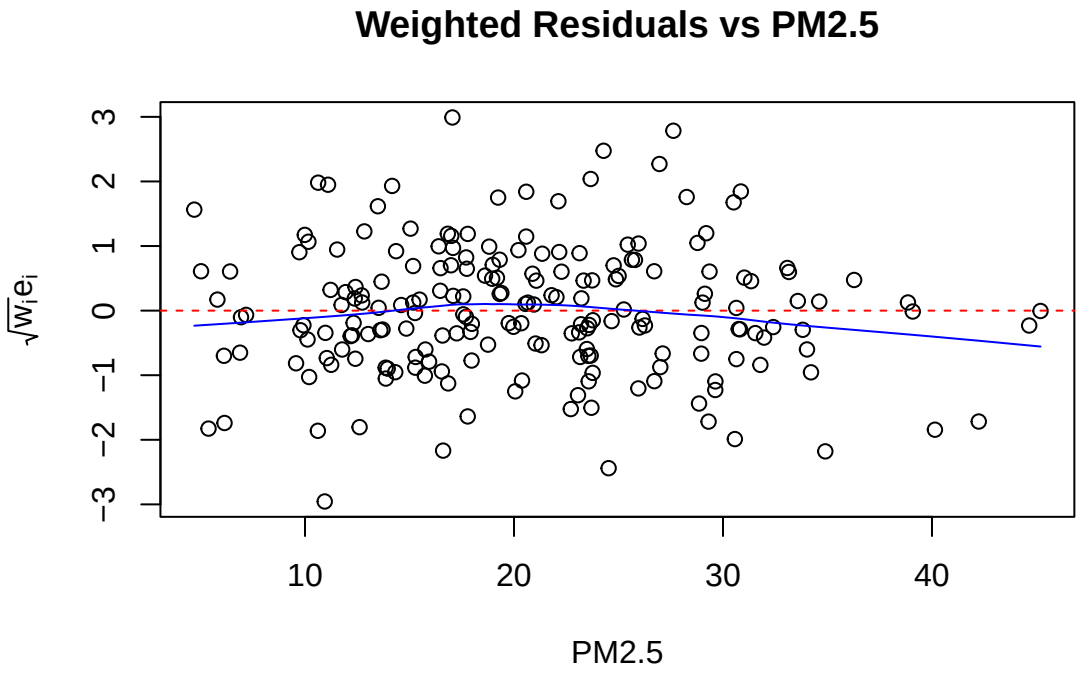
Part (d)

Question: If the weights calculated in (b) are appropriate, then a plot of weighted residuals from the model in (c) against PM2.5 should not show any pattern. Plot $\sqrt{w_i}e_i$ vs. PM2.5 and comment on the appearance of the plot.

Solution:

```r
# Get residuals from weighted model
weighted_resid <- residuals(weighted_model)

# Calculate sqrt(w_i) * e_i
```

```r
sqrt_w_resid <- sqrt(weights) * weighted_resid

# Plot
plot(asthma$pm25, sqrt_w_resid,
     xlab = "PM2.5", ylab = expression(sqrt(w[i]) * e[i]),
     main = "Weighted Residuals vs PM2.5")
abline(h = 0, col = "red", lty = 2)

# Add smoothed line to check for patterns
lines(lowess(asthma$pm25, sqrt_w_resid), col = "blue")
```



The weights are appropriate, since the weighted residuals $\sqrt{w_i}e_i$ show random scatter around zero. This means the weighted residuals have constant variance across PM2.5 values.