STAT 425 Assignment 2

Due Monday, February 22, 11:59pm. Submit through Moodle.

Name: (insert your name here)

Netid: (insert)

Submit your computational work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown.

Problem 1

In this problem we have data of the form $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. We construct a new variable z_i from the x_i values by subtracting their sample mean, so $z_i = x_i - \bar{x}$ for $i = 1, 2, \ldots, n$, where $\bar{x}_i = n^{-1} \sum_{i=1}^{n} x_i$. We consider two models:

$$\begin{aligned} &\text{Model 1:} \quad y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, \ldots, n \\ &\text{Model 2:} \quad y_i = \alpha_0 + \alpha_1 z_i + e_i, \quad i = 1, 2, \ldots, n \end{aligned}$$

Here the x_i values are considered a fixed set of numbers, and the errors e_i are considered to be uncorrelated random variables with $E(e_i) = 0$ and $\text{var}(e_i) = \sigma^2$.

a) By subtracting one model from the other and averaging, show that $\alpha_0 = \beta_0 + \beta_1 \bar{x}$.

Answer: Since $z_i = x_i - \bar{x}$, by subtracting one model from the other yields

$$\beta_0 + \beta_1 x_i = \alpha_0 + \alpha_1 (x_i - \bar{x}), \quad i = 1, 2, \ldots, n.$$

Summing the above equation for $i = 1, \ldots, n$ and taking average yields

$$\beta_0 + \beta_1 \bar{x} = \alpha_0,$$

which follows from the fact that

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) = \bar{x} - \bar{x} = 0 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}.$$

It finishes the proof.

(here or indicate where it is in the attached pdf file.)

b) Substituting $\alpha_0 = \beta_0 + \beta_1\bar{x}$ into Model 2, show that $(\alpha_1 - \beta_1)(x_i - \bar{x}) = 0$, for $i = 1, 2, \ldots, n$. Assume $\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$, i.e., the x_i values are not all the same. Show why this implies $\alpha_1 = \beta_1$.

Answer: Substituting $\alpha_0 = \beta_0 + \beta_1\bar{x}$ into Model 2 yields

$$\begin{aligned} y_i &= \beta_0 + \beta_1\bar{x} + \alpha_1 z_i + e_i \\ &= \beta_0 + \beta_1\bar{x} + \alpha_1(x_i - \bar{x}) + e_i, \quad i = 1, \ldots, n. \qquad (z_i = x_i - \bar{x}) \end{aligned}$$

Subtituting y_i in the right hand side of the above equation with $y_i = \beta_0 + \beta_1 x_i + e_i$ as in Model 1 shows that

$$\beta_0 + \beta_1 x_i + e_i = \beta_0 + \beta_1\bar{x} + \alpha_1(x_i - \bar{x}) + e_i, \quad i = 1, \ldots, n. \qquad (z_i = x_i - \bar{x})$$

Cancelling terms results in

$$\beta_1 x_i = \beta_1\bar{x} + \alpha_1(x_i - \bar{x}) \Rightarrow (\alpha_1 - \beta_1)(x_i - \bar{x}) = 0, \quad i = 1, \ldots, n.$$

It further implies

$$(\alpha_1 - \beta_1)^2(x_i - \bar{x})^2 = 0, \quad i = 1, \ldots, n.$$

Summing the above equation over $i = 1, \ldots, n$ yields

$$(\alpha_1 - \beta_1)^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = 0.$$

Because by assumption $\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$, the above equation implies $\alpha_1 = \beta_1$.

Now we use matrix notation and rewrite Model 2 as $\mathbf{y} = \mathbf{Z}\alpha + \mathbf{e}$ with:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2, \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{Z} = \begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_n \end{pmatrix} \qquad \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \qquad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2, \\ \vdots \\ e_n \end{pmatrix}$$

c) Show

$$\mathbf{Z}^T\mathbf{Z} = \begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Answer: We can write

$$\mathbf{Z}^T\mathbf{Z} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_n \end{pmatrix} \begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} z_i \\ \sum_{i=1}^{n} z_i & \sum_{i=1}^{n} z_i^2 \end{pmatrix}.$$

Because

$$\sum_{i=1}^{n} z_i = \sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$$
$$\sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = S_{xx},$$

we can obtain

$$\mathbf{Z}^T\mathbf{Z} = \begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix}.$$

d) The LS estimate of the coefficient vector α solves the matrix equation:

$$\mathbf{Z}^T\mathbf{Z}\,\hat{\alpha} = \mathbf{Z}^T\mathbf{y}$$

Solve the equation algebraically to get simplified expressions for $\hat{\alpha}_1$ and $\hat{\alpha}_2$ expressed in terms of the original x_i, y_i and n values.

Answer: From c), for the equation $\mathbf{Z}^T\mathbf{Z}\hat{\boldsymbol{\alpha}} = \mathbf{Z}^T\mathbf{y}$, we can expand it to

$$\begin{aligned} \begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix} \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_n \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n}(x_i - \bar{x})y_i \end{pmatrix}. \qquad (z_i = x_i - \bar{x}) \end{aligned}$$

The above equation further implies

$$\begin{aligned} \hat{\alpha}_1 &= \frac{1}{n}\sum_{i=1}^{n} y_i, \\ \hat{\alpha}_2 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \end{aligned}$$

The above two identities are the simplified expressions for $\hat{\alpha}_1$ and $\hat{\alpha}_2$.

e) Derive a simplified expressions for the entries of $\mathrm{cov}(\hat{\alpha})$, the 2×2 covariance matrix of $\hat{\alpha} = \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix}$. (Note that this still depends on the unknown σ^2.)

Answer: It suffices to compute $\text{var}(\hat{\alpha}_1)$, $\text{var}(\hat{\alpha}_2)$, and $\text{cov}(\hat{\alpha}_1, \hat{\alpha}_2)$. Note that by the assumption that e_i and e_j are uncorrelated for $i \neq j$,

$$\text{var}(\hat{\alpha}_1) = \text{var}(\bar{y}) = \frac{\sigma^2}{n},$$
$$\text{var}(\hat{\alpha}_2) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \text{var}(y_i)}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} = \sigma^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

By linearity property of covariance,

$$\begin{aligned}\text{cov}(\hat{\alpha}_1, \hat{\alpha}_2) &= \text{cov}\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \\ &= \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \text{cov}\left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n}(x_i - \bar{x})y_i\right).\end{aligned}$$

Now, we further simplifies the last covariance term as below.

$$\begin{aligned}\text{cov}\left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n}(x_i - \bar{x})y_i\right) &= \sum_{i=1}^{n}(x_i - \bar{x})\text{var}(y_i) + \sum_{i\neq j}(x_j - \bar{x}) \cdot \text{cov}(y_i, y_j) \\ &= \sigma^2 \sum_{i=1}^{n}(x_i - \bar{x}) + \sum_{i\neq j}(x_j - \bar{x}) \cdot 0 \quad (\text{cov}(y_i, y_j) = \text{cov}(e_i, e_j) = 0) \\ &= 0. \qquad (\sum_{i=1}^{n}(x_i - \bar{x}) = 0)\end{aligned}$$

Thus, it shows that $\text{cov}(\hat{\alpha}_1, \hat{\alpha}_2) = 0$. As a result,

$$\text{cov}(\hat{\boldsymbol{\alpha}}) = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2 \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2 \end{pmatrix}.$$

Alternative solution:

From general results about LS regression,

$$\text{cov}(\hat{\boldsymbol{\alpha}}) = \sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1}.$$

Using Part c we obtain

$$\sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1} = \sigma^2 \begin{pmatrix} n^{-1} & 0 \\ 0 & S_{xx}^{-1} \end{pmatrix}$$

Problem 2

This problem considers the `prostate` data in the library `faraway`. See `help(prostate)` for more information about the data set.

a) Fit a linear model with `lpsa` as the response and all the other variables as predictor variables. Display the model summary and identify all variables whose coefficient t values are statistically significant at the $\alpha = 0.05$ level. In each case, state what the null hypothesis is for the t test.

Answer:

```
library(faraway)
fit_2a = lm(lpsa ~ ., data = prostate)
summary(fit_2a)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

The lcavol, lweight, and svi predictors are statistically significant at the $\alpha = 0.05$ significance level. The null hypothesis in each case is that the coefficient of a single predictor is 0 and that there is no linear association between this predictor and the lpsa response variable after adjusting for all other predictors.

b) Fit a reduced model with `lpsa` as the response but removing `lcp`, `gleason`, and `pgg45`

as predictors in the model. Display the model summary, and provide both 90% and 95% confidence intervals for the coefficient of `lbph`. Do either or both of them include 0?

Answer:

```
fit_2b = lm(lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
summary(fit_2b)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83505 -0.39396  0.00414  0.46336  1.57888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.95100    0.83175   1.143 0.255882
## lcavol       0.56561    0.07459   7.583 2.77e-11 ***
## lweight      0.42369    0.16687   2.539 0.012814 *
## age         -0.01489    0.01075  -1.385 0.169528
## lbph         0.11184    0.05805   1.927 0.057160 .
## svi          0.72095    0.20902   3.449 0.000854 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7073 on 91 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6245
## F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

```
confint(fit_2b, 'lbph', level = .95)
```

```
##              2.5 %    97.5 %
## lbph -0.003474551 0.2271544
```

```
confint(fit_2b, 'lbph', level = .90)
```

```
##              5 %      95 %
## lbph 0.01536969 0.2083102
```

Only the 95% confidence interval includes 0. This makes sense because the p-value of the t-test for the lbph predictor outputted in the summary is low enough to reject at the 90% confidence level, but not at the 95% confidence level.

c) Perform an F test comparing the model from Part b) as the null model, and the model from Part a) as the alternative model. Based on your calculations, does the test reject or accept the null hypothesis at the level $\alpha = 0.05$?

Answer:

```
anova(fit_2b, fit_2a)
```

```
## Analysis of Variance Table
##
## Model 1: lpsa ~ lcavol + lweight + age + lbph + svi
## Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     91 45.526
## 2     88 44.163  3    1.3625 0.905 0.4421
```

With a p-value of 0.4421, this test fails to reject the null hypothesis at the $\alpha = 0.05$ level. We accept the reduced model over the full model.

d) Here are the data for observation #32:

```
> prostate[32,]
      lcavol lweight age     lbph svi      lcp gleason pgg45    lpsa
32 0.1823216  6.1076  65 1.704748   0 -1.38629       6     0 2.00821
```

Pretending you did not know the value for `lpsa` for this observation, use your model from Part b) to compute a 95% prediction interval for observation 32 based on the predictors in the model. Does the interval include the value observed in the data?

Answer:

```
predict(fit_2b, newdata = prostate[32,], interval = "prediction", level = .95)
```

```
##         fit      lwr      upr
## 32 2.864524 1.244189 4.484858
```

```
prostate[32,]$lpsa
```

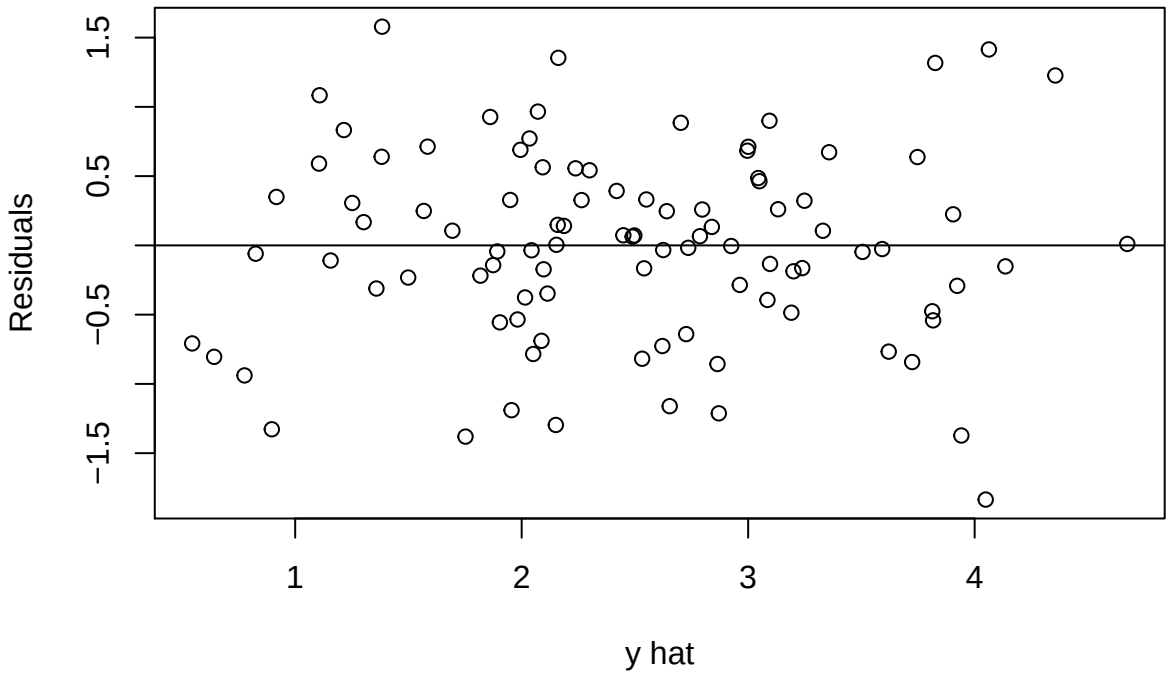
```
## [1] 2.00821
```

This interval does include the observed value.

e) It is often useful to make a scatter plot of the residuals, $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$, on the vertical axis versus the fitted values, $\hat{\mathbf{y}}$ on the horizontal axis. Make such a scatter plot for the model in Part b), and add a horizontal line at a vertical height 0. Hint: `abline(h=0)`. Is there any trend or pattern, or does the point cloud appear random without any systematic trend or curvature? Explain briefly.

Answer:

```
graph = plot(fit_2b$fitted.values, fit_2b$residuals,
             xlab = "y hat",
             ylab = "Residuals") +
  abline(h = 0)
```



There seems to be no clear trend or pattern between residuals and fitted values. The residuals seem to be fairly clustered around the residuals = 0 line regardless of the y hat value and there appears to be no extreme outliers.

Problem 2

This problem considers the `prostate` data in the library `faraway`. See `help(prostate)` for more information about the data set.

a) Fit a linear model with `lpsa` as the response and all the other variables as predictor variables. Display the model summary and identify all variables whose coefficient t values are statistically significant at the $\alpha = 0.05$ level. In each case, state what the null hypothesis is for the t test.

Answer:

```
library(faraway)
fit_2a = lm(lpsa ~ ., data = prostate)
summary(fit_2a)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

The lcavol, lweight, and svi predictors are statistically significant at the $\alpha = 0.05$ significance level. The null hypothesis in each case is that the coefficient of a single predictor is 0 and that there is no linear association between this predictor and the lpsa response variable after adjusting for all other predictors.

b) Fit a reduced model with `lpsa` as the response but removing `lcp`, `gleason`, and `pgg45` as predictors in the model. Display the model summary, and provide both 90% and 95% confidence intervals for the coefficient of `lbph`. Do either or both of them include 0?

Answer:

```
fit_2b = lm(lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
summary(fit_2b)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83505 -0.39396  0.00414  0.46336  1.57888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.95100    0.83175   1.143 0.255882
## lcavol       0.56561    0.07459   7.583 2.77e-11 ***
## lweight      0.42369    0.16687   2.539 0.012814 *
## age         -0.01489    0.01075  -1.385 0.169528
## lbph         0.11184    0.05805   1.927 0.057160 .
```

```
## svi          0.72095    0.20902   3.449 0.000854 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7073 on 91 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6245
## F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

```
confint(fit_2b, 'lbph', level = .95)
```

```
##              2.5 %    97.5 %
## lbph -0.003474551 0.2271544
```

```
confint(fit_2b, 'lbph', level = .90)
```

```
##              5 %      95 %
## lbph 0.01536969 0.2083102
```

Only the 95% confidence interval includes 0. This makes sense because the p-value of the t-test for the lbph predictor outputted in the summary is low enough to reject at the 90% confidence level, but not at the 95% confidence level.

c) Perform an F test comparing the model from Part b) as the null model, and the model from Part a) as the alternative model. Based on your calculations, does the test reject or accept the null hypothesis at the level $\alpha = 0.05$?

Answer:

```
anova(fit_2b, fit_2a)
```

```
## Analysis of Variance Table
##
## Model 1: lpsa ~ lcavol + lweight + age + lbph + svi
## Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     91 45.526
## 2     88 44.163  3    1.3625 0.905 0.4421
```

With a p-value of 0.4421, this test fails to reject the null hypothesis at the $\alpha = 0.05$ level. We accept the reduced model over the full model.

d) Here are the data for observation #32:

```
> prostate[32,]
      lcavol lweight age     lbph svi      lcp gleason pgg45    lpsa
32 0.1823216  6.1076  65 1.704748   0 -1.38629       6     0 2.00821
```

Pretending you did not know the value for `lpsa` for this observation, use your model from Part b) to compute a 95% prediction interval for observation 32 based on the predictors in

the model. Does the interval include the value observed in the data?

Answer:

```
predict(fit_2b, newdata = prostate[32,], interval = "prediction", level = .95)
```

```
##         fit      lwr      upr
## 32 2.864524 1.244189 4.484858
```

```
prostate[32,]$lpsa
```

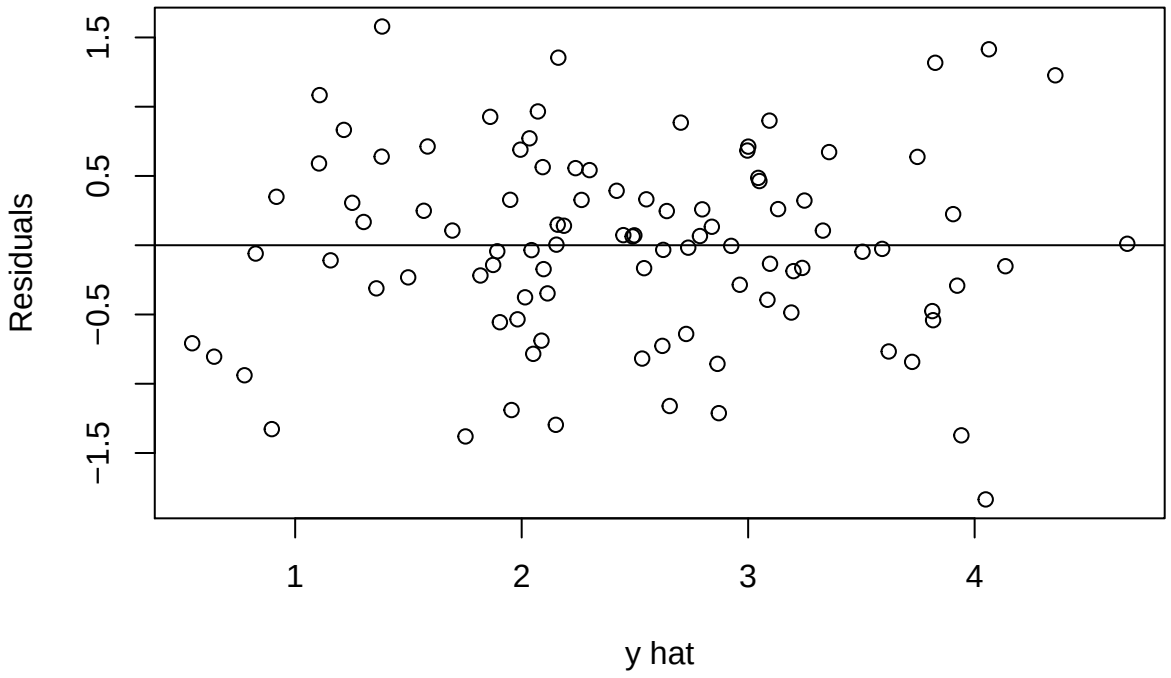
```
## [1] 2.00821
```

This interval does include the observed value.

e) It is often useful to make a scatter plot of the residuals, $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$, on the vertical axis versus the fitted values, $\hat{\mathbf{y}}$ on the horizontal axis. Make such a scatter plot for the model in Part b), and add a horizontal line at a vertical height 0. Hint: `abline(h=0)`. Is there any trend or pattern, or does the point cloud appear random without any systematic trend or curvature? Explain briefly.

Answer:

```
graph = plot(fit_2b$fitted.values, fit_2b$residuals,
             xlab = "y hat",
             ylab = "Residuals") +
  abline(h = 0)
```



There seems to be no clear trend or pattern between residuals and fitted values. The residuals seem to be fairly clustered around the residuals = 0 line regardless of the y hat value and there appears to be no extreme outliers.

Problem 3:

This problem refers to the `punting` data in the `faraway` library. The average distance punted and hang times of 10 punts of a football were measured for 13 volunteers. The left and right leg strength and flexibility were also recorded for each volunteer.

a) Fit a regression model with `Distance` as the response, and `RStr`, `LStr`, `RFlex` and `LFlex` as predictors (left and right leg strength, and left and right leg flexibility). Present a summary of the fitted model. Which if any predictors are significant at the 5% level?

Answer:

```
library("faraway")
model = lm(Distance~RFlex + LFlex + LStr + RStr, data = punting)
summary(model)
```

```
##
## Call:
## lm(formula = Distance ~ RFlex + LFlex + LStr + RStr, data = punting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.941  -8.958  -4.441  13.523  17.016
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -79.6236    65.5935  -1.214    0.259
## RFlex         2.3745     1.4374   1.652    0.137
## LFlex        -0.5277     0.8255  -0.639    0.541
## LStr         -0.1862     0.5130  -0.363    0.726
## RStr          0.5116     0.4856   1.054    0.323
##
## Residual standard error: 16.33 on 8 degrees of freedom
## Multiple R-squared:  0.7365, Adjusted R-squared:  0.6047
## F-statistic:  5.59 on 4 and 8 DF,  p-value: 0.01902
```

As we can see from p value column in the summary of the model, none of the variables are significant at this level since they are all greater than 0.05.

b) Use an F-test to determine whether collectively these four predictors have any relationship with the response, i.e., test the (null) hypothesis that $\beta_{\texttt{RStr}} = \beta_{\texttt{LStr}} = \beta_{\texttt{RFlex}} = \beta_{\texttt{LFlex}} = 0$. (Here we are referring to the coefficient for predictor X_j in the model as β_{X_j}.) What do you conclude?

Answer:

```
modelreduce = lm(Distance~1 , data  = punting)
anova(modelreduce, model)
```

```
## Analysis of Variance Table
##
## Model 1: Distance ~ 1
## Model 2: Distance ~ RFlex + LFlex + LStr + RStr
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     12 8093.3
## 2      8 2132.6  4    5960.7 5.5899 0.01902 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is smaller than 0.05, which means we reject the null hypothesis that all the coefficients are 0.

c) Now we wish to test whether $\beta_{\texttt{RStr}} = \beta_{\texttt{LStr}}$ but not necessarily 0. Under the hypothesis that these two coefficients are equal write out the regression model formula and show that it is equivalent to replacing `RStr` and `LStr` in the model by the single variable `Str` = `RStr` + `LStr`.

Answer: If these two coefficients are equal ($\beta_{RStr} = \beta_{LStr} = \beta_{Str}$), the regression model could be written as:

$$\begin{aligned} y &= \beta_{RStr}x_{RStr} + \beta_{LStr}x_{LStr} + \beta_{RFlex}x_{RFlex} + \beta_{LFlex}x_{LFlex} + e \\ &= \beta_{Str}x_{RStr} + \beta_{Str}x_{LStr} + \beta_{RFlex}x_{RFlex} + \beta_{LFlex}x_{LFlex} + e \\ &= \beta_{Str}(x_{RStr} + x_{LStr}) + \beta_{RFlex}x_{RFlex} + \beta_{LFlex}x_{LFlex} + e \\ &= \beta_{Str}x_{Str} + \beta_{RFlex}x_{RFlex} + \beta_{LFlex}x_{LFlex} + e \end{aligned}$$

We can see it is equivalent to substitute these two variables with a single one by defining the new one "Str" as sum of "RStr" and "LStr".

d) Use an F-test to test whether $\beta_{\texttt{RStr}} = \beta_{\texttt{LStr}}$. Note that the reduced model implied by this hypothesis entails replacing `RStr` and `LStr` in the `lm` model formula by `I(RStr+LStr)` (using the syntax of R).

Answer:

```
modelred = lm(Distance ~ I(RStr + LStr) + RFlex + LFlex, data = punting)
anova(modelred, model)
```

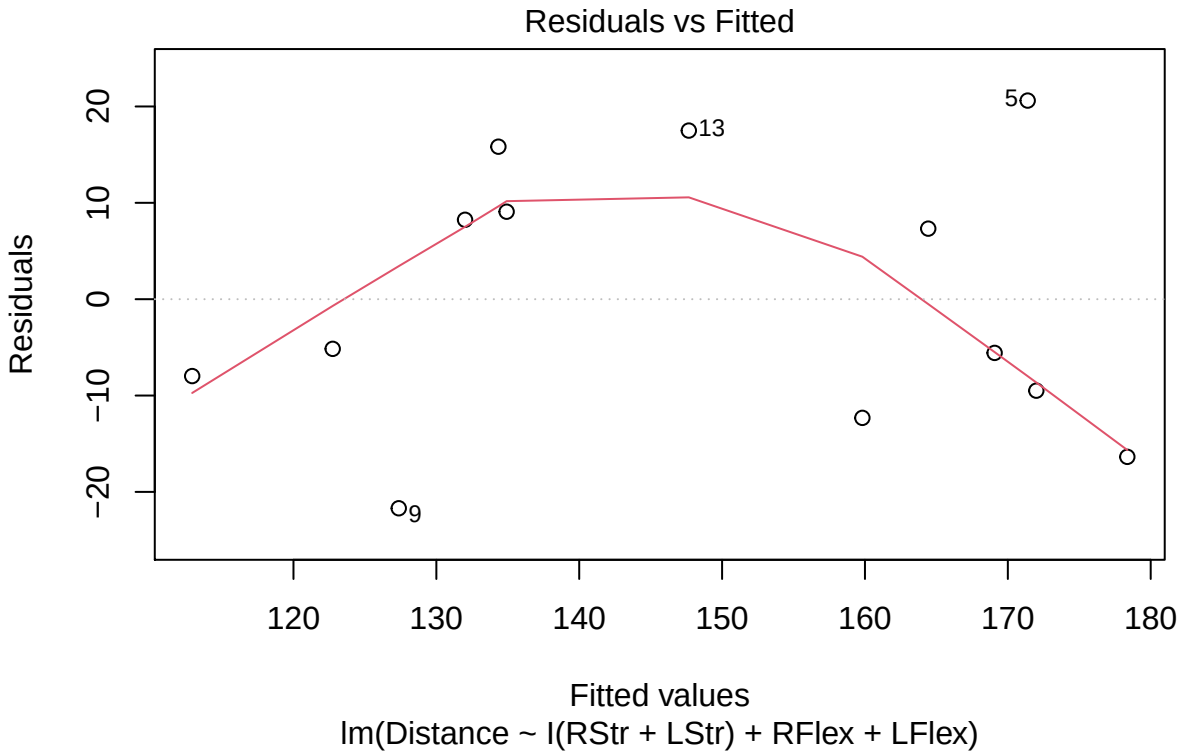
```
## Analysis of Variance Table
##
## Model 1: Distance ~ I(RStr + LStr) + RFlex + LFlex
## Model 2: Distance ~ RFlex + LFlex + LStr + RStr
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      9 2287.4
## 2      8 2132.6  1    154.72 0.5804  0.468
```

From the test result, p value is much larger than 0.05. The result shows the the null hypothesis is acceptable. Therefore, we can accept these two coefficients are equal.

e) Make a plot of residuals versus fitted values for the reduced model considered in Part d). Does the plot show any trend or pattern, or does it appear to be random noise? Explain briefly.

Answer:

```
plot(modelred, which = 1)
```



From the plot we can see there is an obvious trend. But the points seem to be evenly spread around the trend so the noise appears to be random.