STA5167 HW2

Woo Min Kim

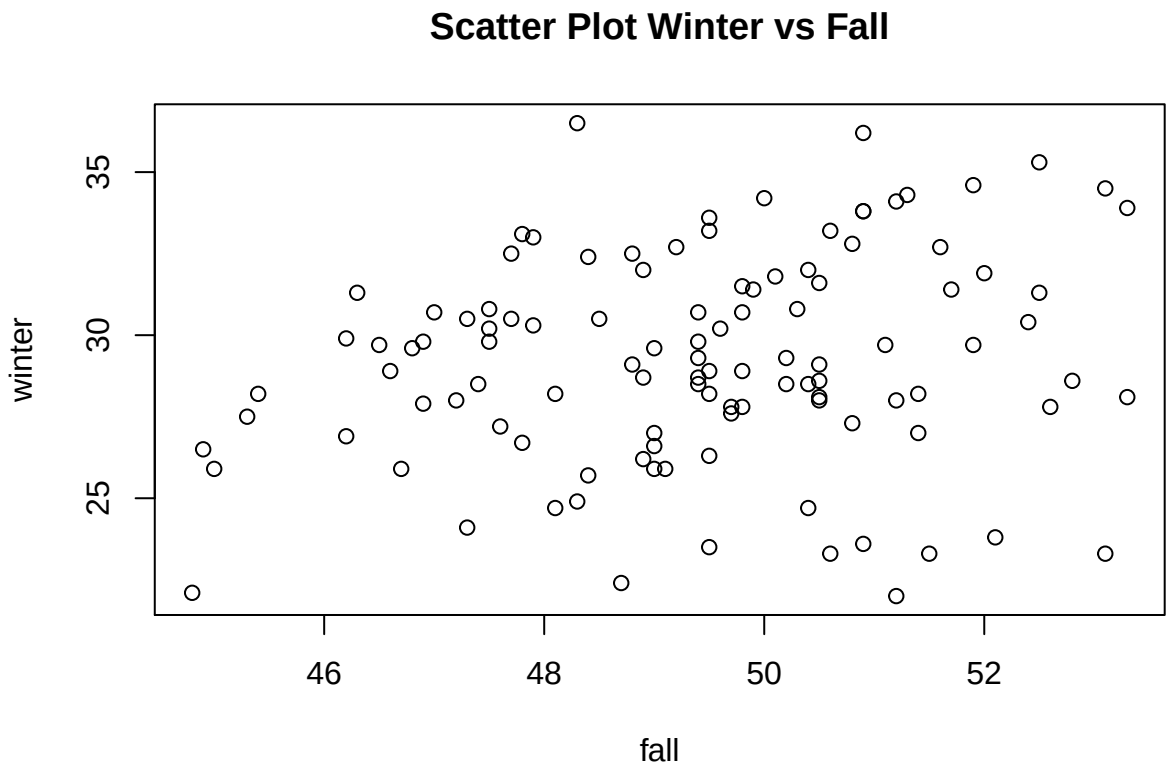
2/4/2019

2.6

```
library(alr4)
attach(ftcollinstemp)
```

2.6.1. Draw a scatterplot of the response versus the predictor, and describe any pattern you might see in the plot.

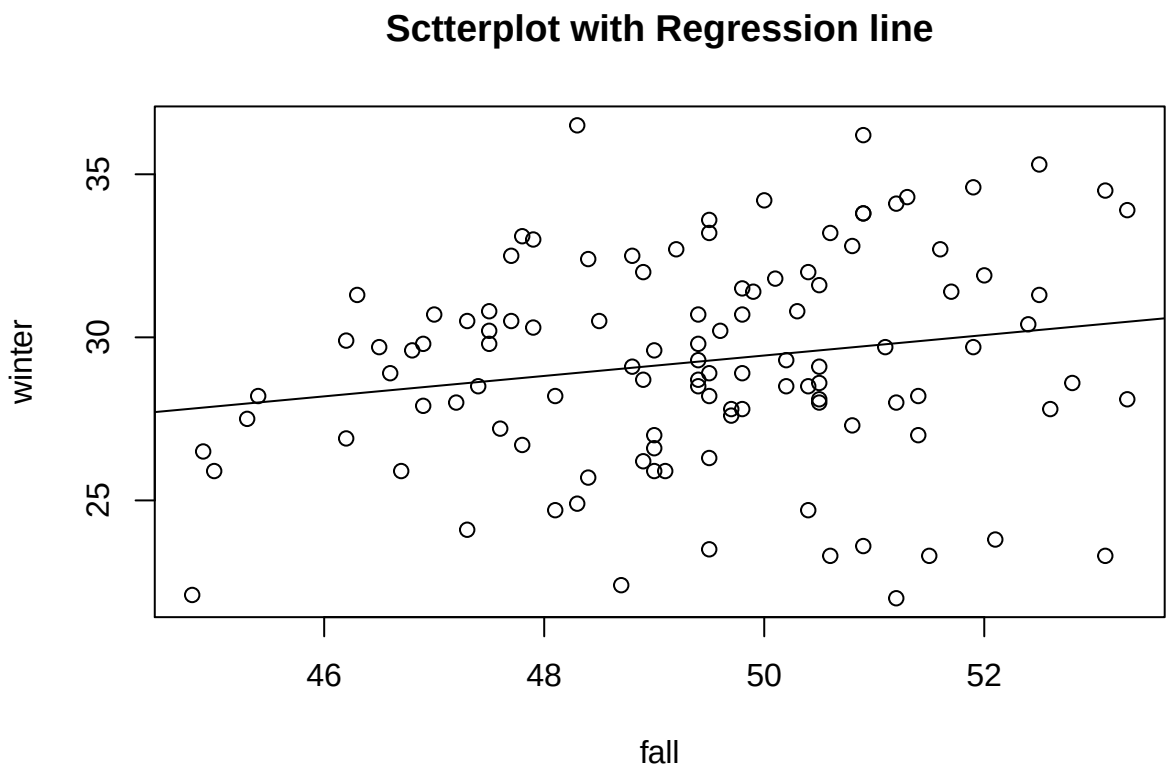
```
plot(fall, winter, main='Scatter Plot Winter vs Fall')
```



As can be seen in the scatter plot, the average temperatures in fall and winter may not have significant relationship because no notable patterns can be found.

2.6.2. Fit the regression model and add the fitted line to your graph. Test the slope with two-sided test and summarize the results.

```
fit = lm(winter ~ fall)
plot(fall, winter, main="Sctterplot with Regression line")
abline(fit)
```




```
summary(fit)
```

```
##
## Call:
## lm(formula = winter ~ fall)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8186 -1.7837 -0.0873  2.1300  7.5896
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.7843     7.5549   1.825   0.0708 .
## fall          0.3132     0.1528   2.049   0.0428 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.179 on 109 degrees of freedom
## Multiple R-squared:  0.0371, Adjusted R-squared:  0.02826
## F-statistic:   4.2 on 1 and 109 DF,  p-value: 0.04284
```

```
# alternatively, we can calculate the estimate of the slope.
# by not using lm function

# Calculate estimates for b0, b1
n = length(winter)
SXY = sum( (fall-mean(fall)) * (winter-mean(winter)))
SXX = sum( (fall-mean(fall))^2 )
beta1 = SXY/SXX
beta0 = mean(winter) - beta1*mean(fall)
```

```
# Calculate Std errs, t-stat, p-vals for b0, b1
s2 = var(winter - beta0 - beta1*fall) * (n-1) / (n-2)
beta1_sd = sqrt(s2 / SXX)
beta1_t = (beta1 - 0) / beta1_sd
beta1_p = (1 - pt(beta1_t,n-2)) * 2

beta0_sd = sqrt(s2 * (1/n + mean(fall)^2/SXX))
beta0_t = (beta0 - 0) / beta0_sd
beta0_p = (1 - pt(beta0_t,n-2)) * 2

b0 = cbind(beta0, beta0_sd, beta0_t, beta0_p)
b1 = cbind(beta1, beta1_sd, beta1_t, beta1_p)
res = rbind(b0,b1)
colnames(res) = c('estimate', 'std.err', 't-stat', 'p-val')
rownames(res) = c('b0', 'b1')
print(res)
```

```
##      estimate   std.err   t-stat      p-val
## b0 13.7843452 7.5548896 1.824559 0.07080657
## b1  0.3131691 0.1528193 2.049277 0.04283611
```

From the result, we obtained the p-value of the slope of 0.0428. This implies 95% confidence interval does not contain 0, and we reject the null hypothesis that slope is 0.

2.6.3. Compute or obtain from your computer output the value of the variability in winter explained by fall and explain what this means.

```
SYY = sum((winter - mean(winter))^2)
r2 = (SXY^2 / SXX) / SYY
sprintf("R-square : %f",r2)
```

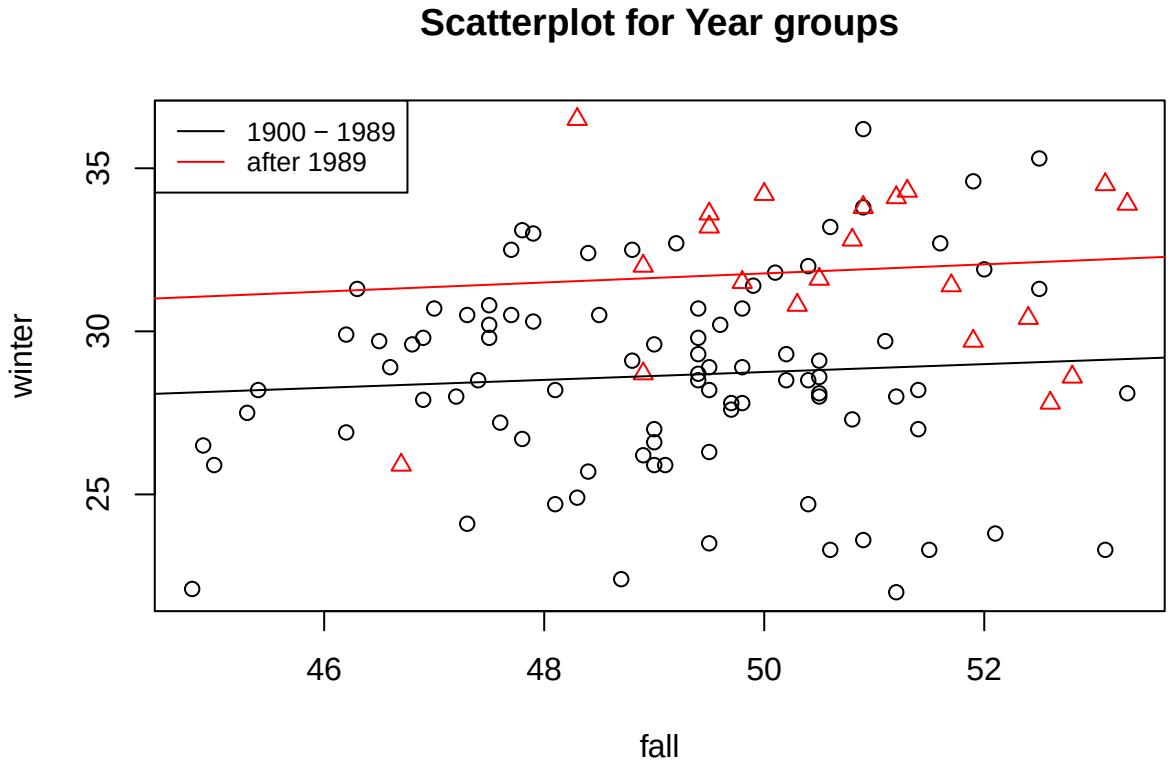
```
## [1] "R-square : 0.037099"
```

R^2 is 0.0371, and this implies the linear model captures 3.71% of the variability of the response variable, winter. Since our predictor variable explains only 3.71% of the response variable, we may say that fall temperature is a good predictor of the winter temperature.

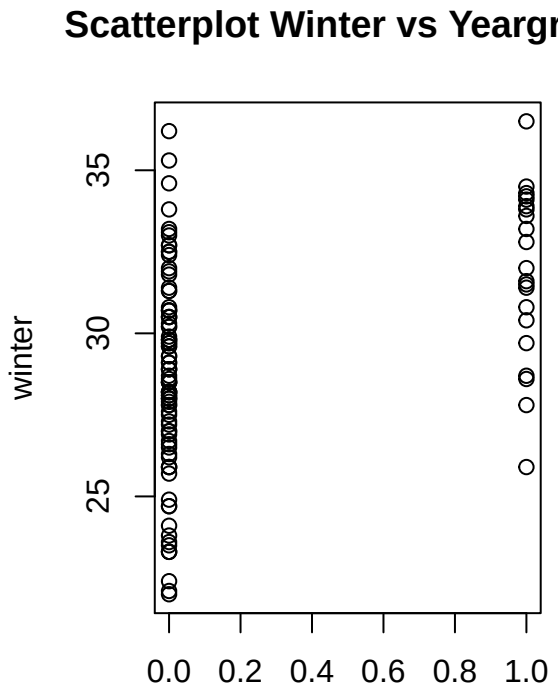
2.6.4. Divide the data into 2 time periods, an early period from 1900 to 1989, and a late period from 1990 to 2010. Are the results different in the two time periods?

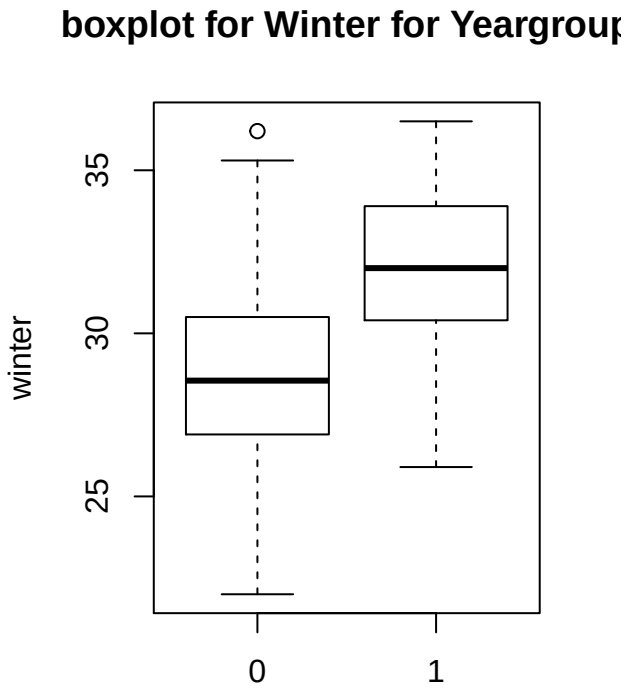
```
# Divide the data
y_group = as.integer(year > 1989)
```

```
# Scatterplot for each year group
idx_y_group = which(y_group == 0)
plot(fall, winter, pch=y_group+1, col=y_group+1,
     main = "Scatterplot for Year groups")
abline(lm(winter[idx_y_group] ~ fall[idx_y_group]))
abline(lm(winter[-idx_y_group] ~ fall[-idx_y_group]), col=2)
legend("topleft", legend=c("1900 - 1989", "after 1989"),
       col=c("black", "red"), lty=c(1,1), cex=0.8)
```




```
par(mfrow=c(1,2))
plot(y_group, winter,
     main="Scatterplot Winter vs Yeargroup",
     xlab="Year Group")
boxplot(winter~ y_group,
        main="boxplot for Winter for Yeargroup",
        xlab="Year Group", ylab="winter")
```






```
fit3 = lm(winter[idx_y_group] ~ fall[idx_y_group])
fit4 = lm(winter[-idx_y_group] ~ fall[-idx_y_group])

summary(fit3)
```

```
##
## Call:
## lm(formula = winter[idx_y_group] ~ fall[idx_y_group])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8976 -1.6349  0.0118  2.0079  7.3387
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         22.7079     8.2600   2.749  0.00725 **
## fall[idx_y_group]    0.1209     0.1681   0.719  0.47397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.057 on 88 degrees of freedom
## Multiple R-squared:  0.005842,   Adjusted R-squared:  -0.005455
## F-statistic: 0.5171 on 1 and 88 DF,  p-value: 0.474
```

```
summary(fit4)
```

```
##
## Call:
## lm(formula = winter[-idx_y_group] ~ fall[-idx_y_group])
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4174 -1.7097  0.3768  1.8988  4.9602
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          24.8260    17.7973   1.395    0.179
## fall[-idx_y_group]    0.1390     0.3509   0.396    0.696
##
## Residual standard error: 2.699 on 19 degrees of freedom
## Multiple R-squared:  0.00819,    Adjusted R-squared:  -0.04401
## F-statistic: 0.1569 on 1 and 19 DF,  p-value: 0.6965
```

```
summary(aov(winter~y_group))
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## y_group       1  177.4  177.43   20.01 1.9e-05 ***
## Residuals   109  966.7    8.87
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First, we can compare the mean of each group by using ANOVA test with assuming the data fairly follows the ANOVA assumptions. From the box-plot and the ANOVA table above, we can conclude that the winter temperature in two groups are different.

Then we can also compare the slopes in two groups. From the linear model summary tables for two groups above, it is quite clear that the slopes are almost the same considering the estimates and their standard errors. Furthurmore, their t-test statistics and p-values indicate that the slope is not significantly effective.

The overall difference can be found in the scatter plot with regression lines above. The regression lines are almost parallel, but they have different intercept.

2.8.

$$y_i = (\beta_0 + \beta_1 \bar{x}) + \beta_1(x_i - \bar{x}) + e_i = \alpha + \beta_1(x_i - \bar{x}) + e_i$$

2.8.1. What is the meaning of the parameter α?

It is an intercept term of centered predictor x. The parameter α also can be interpreted as follows:

$$\begin{aligned} y_i = \beta_0 + \beta_1 x_i + e_i \Rightarrow \frac{1}{n}\sum_{i=1}^{n} y_i &= \beta_0 + \frac{1}{n}\sum_{i=1}^{n}\beta_1 x_i + \frac{1}{n}\sum_{i=1}^{n} e_i \\ \bar{y} &= \beta_0 + \beta_1 \bar{x} \\ \text{This implies : } y_i - \bar{y} &= (\beta_0 + \beta_1 x_i + e_i) - (\beta_0 + \beta_1 \bar{x}) \\ y_i &= \bar{y} + \beta_1(x_i - \bar{x}) + e_i \\ &= \alpha + \beta_1(x_i - \bar{x}) + e_i \end{aligned}$$

As can be seen from the equation above, α can be interpreted as sample mean of response variable, y.

2.8.2. Show that the least squares estimates are $\hat{\alpha} = \bar{y}$, $\hat{\beta}_1 = SXY/SXX$.

Since the linear models (1) $y_i = \alpha + \beta_1(x_i - \bar{x}) + e_i$ and (2) $y_i = \beta_0 + \beta_1 x_i + e_i$ are the same, it is necessary that the estimates of the parameter β_1 in (1) and (2) are the same. Since in model (2), $\hat{\beta}_1 = \frac{SXY}{SXX}$, $\hat{\beta}_1$ in model (1) should have the same value. The estimate, $\hat{\alpha}$, should be $\bar{y}$, and its explanation can be found in the previous section 2.8.1.

2.8.3. Find expressions for the variances of the estimates and the covariance between them.

It is obvious that $\hat{\beta}_1$ has variance of $\frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})^2} = \sigma^2/SXX$ (where σ^2 is the variance of the response variable) as in the original model (2). This can be derived from the fact that $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{SXX} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{SXX} = \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n \bar{x}y_i}{SXX} = \sum_{i=1}^n c_i y_i$, where $c_i = \frac{x_i-\bar{x}}{SXX}$. And then, we have $\text{Var}(\sum_{i=1}^n c_i y_i|X) = \sigma^2 \sum_{i=1}^n c_i^2 = \sigma^2/SXX$.

The variance of $\hat{\alpha}$ can be derived from the fact that $\text{Var}(\hat{\beta}_0|X) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{SXX})$ and $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1|X) = -\sigma^2 \frac{\bar{x}}{SXX}$. Since $\hat{\alpha} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}$, $\text{Var}(\hat{\alpha}|X) = \text{Var}(\hat{\beta}_0|X) + \bar{x}^2\text{Var}(\hat{\beta}_1|X) + 2\bar{x}\text{Cov}(\hat{\beta}_0, \hat{\beta}_1|X)$. This leads to $\text{Var}(\hat{\alpha}|X) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{SXX}) + \bar{x}^2\frac{\hat{\sigma^2}}{SXX} + 2\bar{x}(-\sigma^2\frac{\bar{x}}{SXX}) = \sigma^2/n = \text{Var}(\bar{y}|X)$.

the covariance of $(\hat{\alpha}, \hat{\beta}_1)$ can be derived as follows: $\text{Cov}(\hat{\alpha}, \hat{\beta}_1|X) = \text{Cov}(\hat{\beta}_0 + \hat{\beta}_1\bar{x}, \hat{\beta}_1|X) = \text{Cov}(\hat{\beta}_0, \hat{\beta}_1|X) + \bar{x}\text{Var}(\hat{\beta}_1|X) = -\sigma^2\frac{\bar{x}}{SXX} + \bar{x}\frac{\sigma^2}{SXX} = 0$

This corresponds to the fact $\text{Cov}(\bar{y}, \beta_1|X) = 0$.

2.9 (Invariance)

2.9.1. Find relationship (1) between β_0 and γ_0; (2) between β_1 and γ_1; (3) between the estimates of variance in the 2 regressions, and (4) t-tests of $\beta_1 = 0$ and $gamma_1 = 0$.

$$\text{Model I: } E(Y|X = x) = \beta_0 + \beta_1 x$$
$$\text{Model II: } E(Y|Z = z) = \gamma_0 + \gamma_1 z = \gamma_0 + \gamma_1(ax + b)$$

By closely comparing each term in both models we can infer the relationship between parameters. So we have:

$$\beta_0 = \gamma_0 + b\gamma_1$$
$$\beta_1 = a\gamma_1$$

From the fact that the estimate of variance is $\hat{\sigma}^2 = RSS/(n-2)$ where RSS is residual sum of squares, we have:

$$\text{Model 1 : } \hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{n-2}$$
$$\text{Model 2 : } \hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \gamma_0 - b\gamma_1 - a\gamma_1 x_i)^2}{n-2}$$

But since $y_i = \beta_0 + \beta_1 x_i = \gamma_0 + b\gamma_1 + a\gamma_1 x_i$, the estimate of variances for both models essentially the same.

We can also compare the t-test statistics for both models as follows:

$$\text{Model 1 : } t_{\beta_1} = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}^2/SXX}}$$
$$\text{Model 2 : } t_{\gamma_1} = \frac{\hat{\gamma_1} - 0}{se(\hat{\gamma_1})} = \frac{\hat{\gamma_1}}{\sqrt{\hat{\sigma}^2/(a^2 SXX)}} = \frac{a\hat{\gamma_1}}{\sqrt{\hat{\sigma}^2/SXX}} = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/SXX}}$$

From the fact that $\beta_1 = a\gamma_1$, we can derive $\text{Var}(\gamma_1) = \frac{\text{Var}(\beta_1)}{a^2}$. Then the derivation of t-test statistics above makes sense and we can check they are the same for both models.

2.9.2. Let $V = dY$. Find relationship (1) between β_0 and δ_0; (2) between β_1 and δ_1; (3) between the estimates of variance in the 2 regressions, and (4) t-tests of $\beta_1 = 0$ and $delta_1 = 0$.

$$\text{Model I: } E(Y|X = x) = \beta_0 + \beta_1 x$$
$$\text{Model II: } E(V|X = x) = \delta_0 + \delta_1 x$$

Since $V = dY$, we have $\hat{v}_i = d\hat{y}_i$. Then we can rewrite the Model II: $d\hat{y}_i = \delta_0 + \delta_1 x_i \Rightarrow \hat{y}_i = \frac{\delta_0}{d} + \frac{\delta_1}{d} x_i$. Then we can find the relationship between parameters. So we have:

$$\beta_0 = \frac{\delta_0}{d}$$

$$\beta_1 = \frac{\delta_1}{d}$$

From the fact that the estimate of variance is $\hat{\sigma}^2 = RSS/(n-2)$ where RSS is residual sum of squares, we have:

$$\text{Model 1 : } \hat{\sigma}_y^2 = \frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{n-2}$$
$$\text{Model 2 : } \hat{\sigma}_v^2 = \frac{\sum_{i=1}^n (v_i - \delta_0 - \delta_1 x_i)^2}{n-2} = d^2 \frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{n-2}$$

From the result above, we can conclude that the estimates of variance are different because RSS changes as we transform the response variable. It is quite plausible changing RSS by d^2 times as the variance of the response variable changes by d^2 times.

We can also compare the t-test statistics for both models as follows:

$$\text{Model 1 : } t_{\beta_1} = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}^2/SXX}}$$
$$\text{Model 2 : } t_{\delta_1} = \frac{\hat{\delta}_1 - 0}{se(\hat{\delta}_1)} = \frac{\hat{\delta}_1}{\sqrt{d^2\hat{\sigma}^2/SXX}} = \frac{\hat{\delta}_1/d}{\sqrt{\hat{\sigma}^2/SXX}} = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/SXX}}$$

From the fact that $\beta_1 = \delta_1/d$, we can derive $\text{Var}(\gamma_1) = d^2\text{Var}(\beta_1)$. Then the derivation of t-test statistics above makes sense, and we can check they are the same for both models.

2.13

```
attach(Heights)
y = dheight
x = mheight
```

2.13.1. Compute the regression of dheight on mheight, and report (1) the estimates, (2) their standard errors, (3) the value of the cofficient of determination, and (4) the estimate of variance.

For the computation, I defined a function to compute all estimates and statistics needed. The function is as follows:

```
fit = function(y,x){
  n = length(y)
  SXX = sum((x - mean(x))^2)
  SXY = sum((x-mean(x))*(y-mean(y)))
  SYY = sum((y-mean(y))^2)
```

```
  b1 = SXY / SXX
  b0 = mean(y) - b1 * mean(x)
  RSS = sum((y - b0 - b1*x)^2)
  var_est = RSS / (n-2)
  b1_se = sqrt(var_est / SXX)
  b0_se = sqrt(var_est * (1/n + mean(x)^2 / SXX))
  b1_t = b1 / b1_se; b0_t = b0 / b0_se
  R2 = 1 - RSS/SYY
  xbar = mean(x)

  # returns SXX, SXY, estimates of b0, b1,
  # standard errors and t-statistics of estimates,
  # R-square, variance of estimation, RSS, and
  # sample mean of predictor
  res = list(SXX = SXX, SXY = SXY, SYY = SYY,
             b0 = b0, b0_se = b0_se, b0_t = b0_t,
             b1 = b1, b1_se = b1_se, b1_t = b1_t,
             var_est = var_est, R2 = R2, RSS = RSS,
             xbar = xbar, n = n)
  return(res)
}
```

Using the function above, the results were generated and printed as below.

```
fit13 = fit(y,x)

# (1) The estimates of beta0 and beta1
res1 = cbind(fit13$b0, fit13$b1)
colnames(res1) = c("b0", "b1")
rownames(res1) = "Estimates"
# (2) standard errors of beta0 and beta1
res2 = cbind(fit13$b0_se, fit13$b1_se)
rownames(res2) = "Standard Error"

# (3) the value of the cofficient of determination
res3 = 1 - fit13$RSS / fit13$SYY

# (4) the estimate of variance
res4 = fit13$var_est


# Print the results
print(
  list(Estimate_beta = rbind(res1,res2),
       R_squared = res3,
       Estimate_Variance = res4)
)
```

```
## $Estimate_beta
##                       b0         b1
## Estimates      29.917437 0.54174701
## Standard Error  1.622469 0.02596069
##
## $R_squared
## [1] 0.2407957
```

```
##
## $Estimate_Variance
## [1] 5.136167
```

2.13.2. Obtain a 99% confidence interval for β_1.

```
upper = fit13$b1 + qt(0.005, fit13$n-2, lower.tail = F) * fit13$b1_se
lower = fit13$b1 - qt(0.005, fit13$n-2, lower.tail = F) * fit13$b1_se

res = cbind(lower,upper)
rownames(res) = "99% CI of b1"
print(res)
```

```
##                   lower     upper
## 99% CI of b1 0.4747836 0.6087104
```

2.13.3. Obtain a (1) prediction and (2) 99% prediction interval for a daughter whose mother is 64 inches tall.

In order to avoid using predict function, I defined my own function for prediction interval as follows:

```
# function for standard error for predictions
se_pred = function(y, x, new_data){
  fit = fit(y,x)
  se_pred = sqrt(fit$var_est) * (1 + 1/fit$n + (new_data - fit$xbar)^2/fit$SXX)^(1/2)
  return(se_pred)
}

# function for prediction interval
pred_intv = function(y, x, new_data, alpha){
  fit = fit(y,x)
  se = se_pred(y,x,new_data)
  upper = fit$b0 + fit$b1*new_data + qt(alpha/2, fit$n-2, lower.tail = F) * se
  lower = fit$b0 + fit$b1*new_data - qt(alpha/2, fit$n-2, lower.tail = F) * se
  fit_val = fit$b0 + fit$b1*new_data
  res1 = cbind(new_data, fit_val)
  res2 = cbind(lower, upper)
  rownames(res1) = "Prediction"
  rownames(res2) = sprintf("%i%% Pred_Intv",(1-alpha)*100)
  return(list(Prediction = res1, Pred_Interval = res2, Pred_Std.Err = se))
}
pred_intv(y,x,new_data = 64,0.01)
```

```
## $Prediction
##            new_data  fit_val
## Prediction       64 64.58925
##
## $Pred_Interval
##                   lower    upper
## 99% Pred_Intv 58.74045 70.43805
##
## $Pred_Std.Err
## [1] 2.267491
```

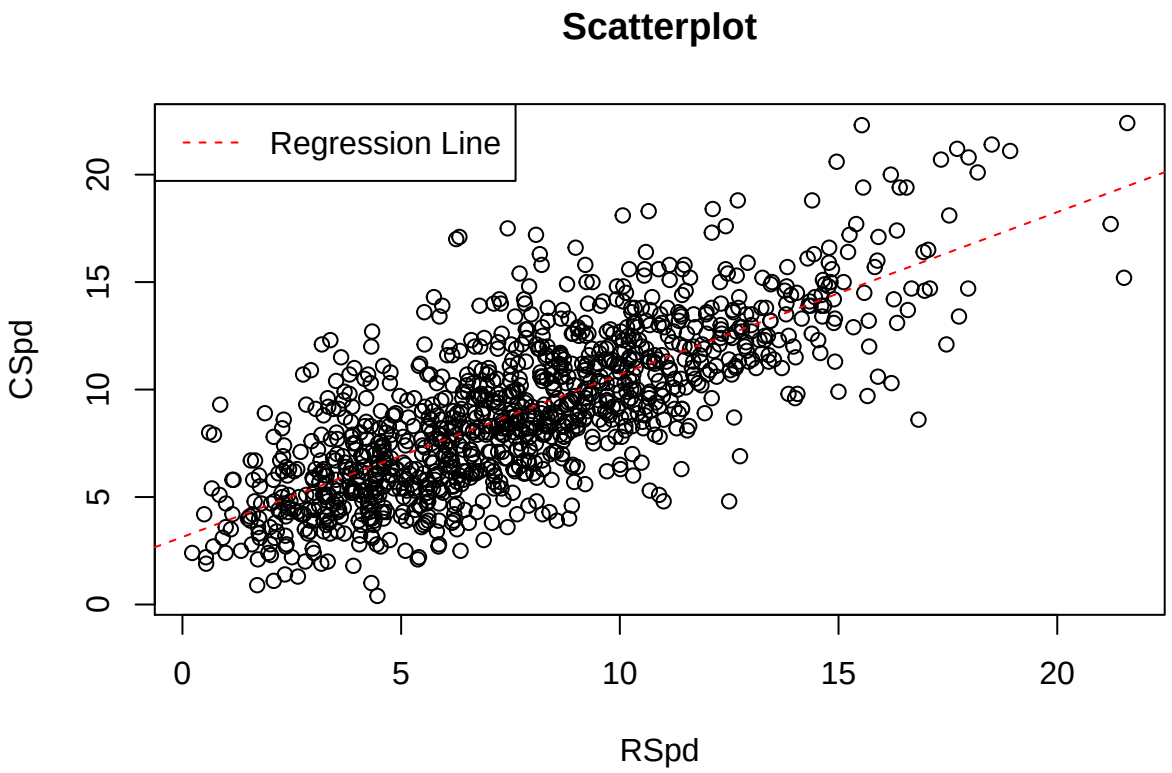
When mheight is 64, the prediction is 64.59. And the 99% prediction interval is [58.74, 70.44].

2.21

```
attach(wm1)
```

2.21.1. Sctterplot of the response CSpd vs the predictor RSpd. Is the simple linear regression model plausible for these data?

```
plot(RSpd, CSpd, main="Scatterplot")
abline(lm(CSpd~RSpd), col='red', lty=2)
legend("topleft", "Regression Line", col="red", lty=2)
```



As can be seen in the scatterplot above, a clear linear trend can be found. This implies a linear model is appropriate. I also added a simple regression line, and it captures the data quite well as expected. Furthurmore, the residuals seem not to have heteroskedasticity problem.

2.21.2.Fit the simple regression of the response on the predictor, and present the appropriate regression summaries.

```
fit21 = fit(CSpd,RSpd)

# (1) The estimates of beta0 and beta1
res1 = cbind(fit21$b0, fit21$b1)
colnames(res1) = c("b0", "b1")
rownames(res1) = "Estimates"
# (2) standard errors of beta0 and beta1
res2 = cbind(fit21$b0_se, fit21$b1_se)
rownames(res2) = "Standard Error"

# (3) t-statistics of beta0 and beta1
```

```r
res3 = cbind(fit21$b0_t, fit21$b1_t)
rownames(res3) = "t-test"

# (4) p-values
b0_p = 2 * pt(fit21$b0_t, fit21$n-2, lower.tail = F)
b1_p = 2 * pt(fit21$b1_t, fit21$n-2, lower.tail = F)

res4 = cbind(b0_p, b1_p)
colnames(res4) = c("b0", "b1")
rownames(res4) = "p-value"

# (5) the value of the cofficient of determination
res5 = 1 - fit21$RSS / fit21$SYY

# (6) the estimate of standard error
res6 = sqrt(fit21$var_est)


# Print the results
print(
  list(Estimate_beta = rbind(res1,res2,res3),
       p_values = res4,
       R_squared = res5,
       Estimate_SE = res6)
)
```

```
## $Estimate_beta
##                          b0          b1
## Estimates       3.1412324  0.75573333
## Standard Error  0.1695765  0.01962928
## t-test         18.5239830 38.50030584
##
## $p_values
##                  b0            b1
## p-value 5.4493e-67 6.632394e-207
##
## $R_squared
## [1] 0.5709235
##
## $Estimate_SE
## [1] 2.466234
```

As can be seen in the summary table above, we can say that both β_0 and β_1 are significant with very small p-values. It corresponds to the observation of the previous scatter plot. R-squared value indicates that the model explains 57% of the variance of the response which also fairly nice.

2.21.3. Obtain a 95% prediction interval for CSpd at a time when RSpd $= 7.4285$.

I used the function defined in the problem 2.13.

```r
pred_intv(CSpd, RSpd, new_data = 7.4285, alpha=0.05)
```

```
## $Prediction
##            new_data  fit_val
## Prediction   7.4285 8.755197
```

```
##
## $Pred_Interval
##                   lower    upper
## 95% Pred_Intv 3.914023 13.59637
##
## $Pred_Std.Err
## [1] 2.467349
```

2.21.4. Show that (1) the average of the m predictions is equal to the prediction taken at the average value $\bar{x}_*$ of the m values of the predictor.

This implies we need to show :

$$\frac{1}{m}\sum_{i=1}^{m}\hat{\beta}_0 + \hat{\beta}_1 x_{*i} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x_*}$$

We can manipulate LHS, then we can get the result above as desired.

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\hat{\beta}_0 + \hat{\beta}_1 x_{*i} &= \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{m}\sum_{i=1}^{m} x_{*i} \\ &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x_*}\end{aligned}$$

2.21.4. Show that (2) Standard Error of the Average of m predictions is

$\sqrt{\frac{\hat{\sigma}^2}{m} + \hat{\sigma}^2\Big(\frac{1}{n} + \frac{(\bar{x}_* - \bar{x})}{SXX}\Big)}$.

First we can derive the variance of a fitted value, $\bar{x}_*$ first. Then we have,

$$\begin{aligned}\mathrm{Var}(\hat{y}|X = \bar{x}_*) &= \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 \bar{x_*}|X = \bar{x}_*) \\ &= \mathrm{Var}(\hat{\beta}_0|X = \bar{x}_*) + \bar{x}_*^2 \mathrm{Var}(\hat{\beta}_1|X = \bar{x}_*) + 2\bar{x}_* \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1|X = \bar{x}_*) \\ &= \sigma^2\Big(\frac{1}{n} + \frac{\bar{x}_*^2}{SXX}\Big) + \sigma^2 \bar{x}_*^2 \frac{1}{SXX} - 2\sigma^2 \frac{\bar{x}_*}{SXX} \\ &= \sigma^2\Big(\frac{1}{n} + \frac{(\bar{x}_* - \bar{x})^2}{SXX}\Big)\end{aligned}$$

Then we can have, variance of a prediction of average of m predictions.

$$\mathrm{Var}(\hat{y_*}|X = \bar{x}_*) = \sigma^2\Big(\frac{1}{n} + \frac{(\bar{x}_* - \bar{x})^2}{SXX}\Big) + S$$

where S is the variance of the error for a future value. Since we are dealing with the future value as $\frac{1}{m}\sum_i^m \hat{y}_*$, its variance is:

$$\mathrm{Var}(\frac{1}{m}\sum_{i}^{m}\hat{y}_*|X = x_*) = \frac{1}{m^2} m\sigma^2 = \frac{\sigma^2}{m}$$

Thus, $S = \sigma^2/m$, and consequently, we have std. err. of prediction $= \sqrt{\frac{\hat{\sigma}^2}{m} + \hat{\sigma}^2\Big(\frac{1}{n} + \frac{(\bar{x}_* - \bar{x})}{SXX}\Big)}$ as desired.

2.21.5. Let $m = 62039$, $\bar{x}_* = 7.4285$. Give a 95% prediction interval on the long-term average wind speed at the candidate site.

To compute the prediction interval, I manipulated my prediction interval function as follows:

```r
m = 62039
x_bar_star = 7.4285

se_avg_pred = function(y, x, new_data, m){
  fit = fit(y,x)
  se_pred = sqrt(fit$var_est)/m +
    sqrt(fit$var_est) * (1/fit$n + (new_data - fit$xbar)^2/fit$SXX)^(1/2)
  return(se_pred)
}

pred_intv_avg = function(y, x, new_data, alpha, m){
  fit = fit(y,x)
  se = se_avg_pred(y,x,new_data,m)
  upper = fit$b0 + fit$b1*new_data + qt(alpha/2, fit$n-2, lower.tail = F) * se
  lower = fit$b0 + fit$b1*new_data - qt(alpha/2, fit$n-2, lower.tail = F) * se
  fit_val = fit$b0 + fit$b1*new_data
  res1 = cbind(new_data, fit_val)
  res2 = cbind(lower, upper)
  rownames(res1) = "Prediction"
  rownames(res2) = sprintf("%i%% Pred_Intv",(1-alpha)*100)
  return(list(Prediction = res1, Pred_Interval = res2, Pred_Std.Err = se))
}

pred_intv_avg(CSpd, RSpd, new_data = x_bar_star, alpha=0.05, m)
```

```
## $Prediction
##            new_data  fit_val
## Prediction   7.4285 8.755197
##
## $Pred_Interval
##                   lower    upper
## 95% Pred_Intv 8.609646 8.900748
##
## $Pred_Std.Err
## [1] 0.0741814
```

With long-term average wind speed in predictor variable, wind speed at reference sites, we can obtain much more conservative prediction interval which is almost the same with the estimated variance of a fitted value.