STA5167 Homework 1

Rey Govea

1/13/2021

```
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
data()
```

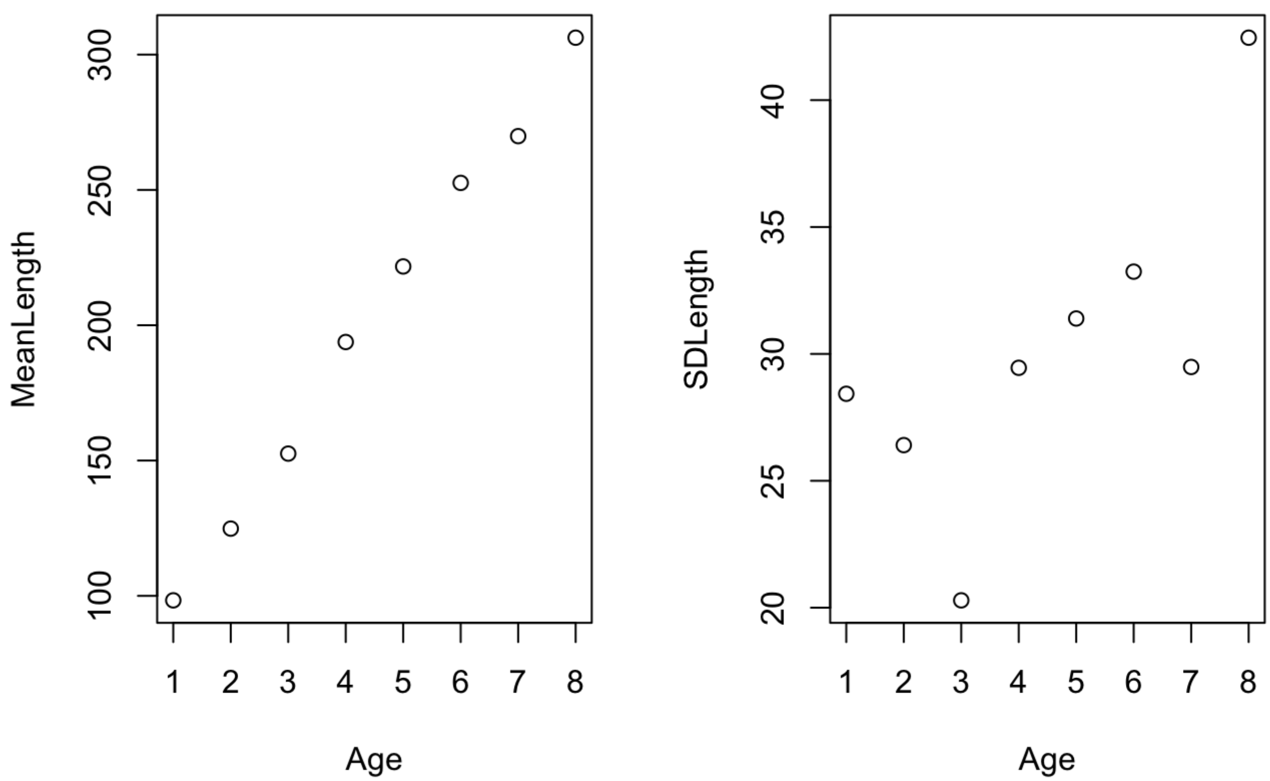
1.2 Smallmouth bass data

Compute the means and the variances for each of the eight subpopulations in the smallmouth bass data. Draw a graph of average length versus Age and compare to Figure 1.5. Draw a graph of the standard deviations versus age. If the variance function is constant, then the plot of standard deviation versus Age should be a null plot. Summarize the information.

```
#computing the mean and Var. for each of the 8 years of bass age
# tapply function applies a statistics to different levels of a factor
MeanLength <- with(wblake, tapply(Length, Age, mean))
SDLength <- with(wblake, tapply(Length, Age, sd))
Age <- c(1, 2, 3, 4, 5, 6, 7 ,8)
Q1.1 <- data.frame(Age, MeanLength, SDLength) #merging the three vectors into a matrix
print(Q1.1)
```

```
##   Age MeanLength SDLength
## 1   1   98.34211 28.42941
## 2   2  124.84722 26.40618
## 3   3  152.56383 20.28960
## 4   4  193.80000 29.45263
## 5   5  221.72059 31.39581
## 6   6  252.59770 33.24275
## 7   7  269.86885 29.48529
## 8   8  306.25000 42.46077
```

```
par(mfrow= c(1,2))
# Plotting the MeanLength to the age to see how it corresponds to Figure 1.5 and OLS lin
e
Q1.2a <- plot(MeanLength ~ Age, data = Q1.1 )
#Plotting the SDLength to age for constant varianace analysis
Q1.2b <- plot(SDLength ~ Age, data = Q1.1)
```



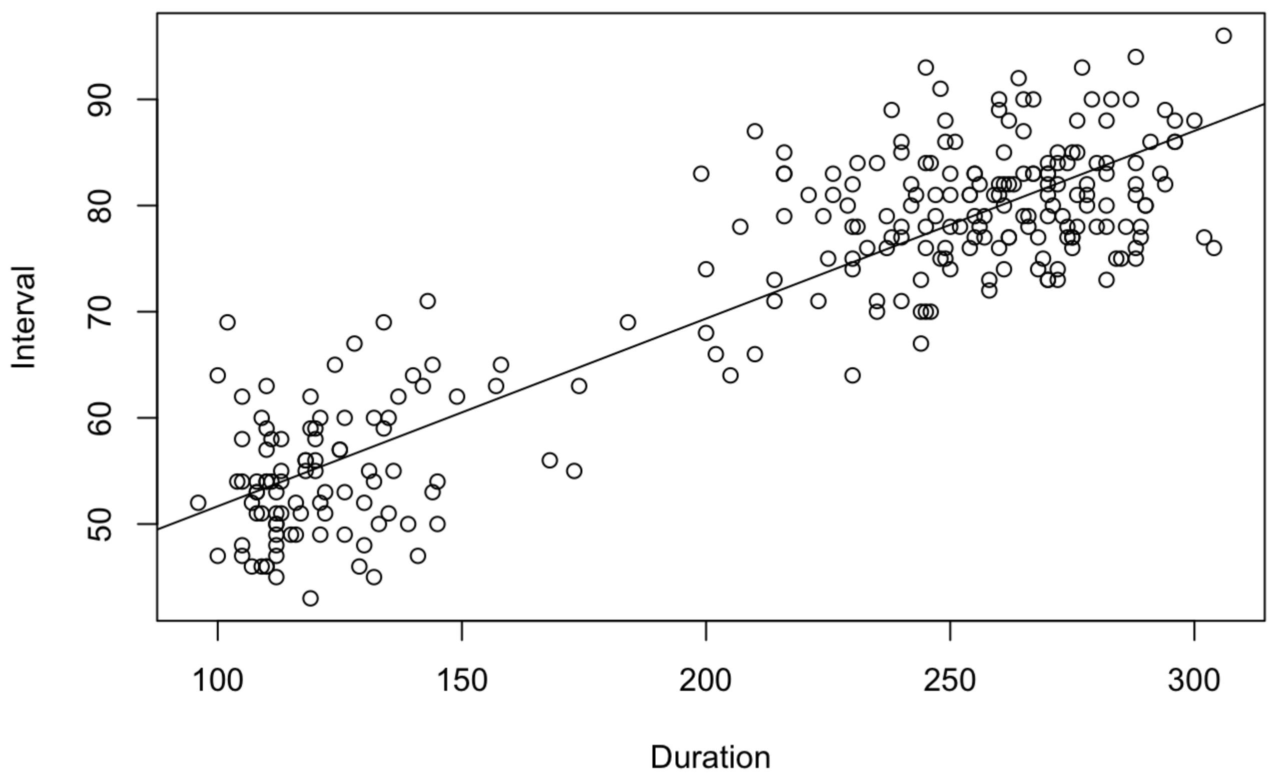
We see that the means plotted have a positive linear relationship similar to the sample data in figure 1.5. the OLS regression line in the textbook is a very close fit to the means of the data. As for the Standard Deviation to age, Age 8 and age 3 show the most variation from the 30 SDLength value. There is a slight increase in the slope of the standard deviations (SD). This could show us both that the SD of bass length increases with age and that the SD length is mostly constant through the first 8 years of growth; This indicates that there is most likely constant variance.

#1.4 Old Faithful *The data in the data file oldfaith.txt gives information about eruptions of Old Faithful Geyser during October 1980. Variables are the Duration in seconds of the current eruption, and the Interval, the time in minutes to the next eruption. The data were collected by volunteers and were provided by R. Hutchinson. Apart from missing data for the period from midnight to 6 AM, this is a complete record of eruptions for that month… [Insert Graph (pg. 19)]

…Old Faithful Geyser is an important tourist attraction, with up to several thousand people watching it erupt on pleasant summer days. The park service uses data like these to obtain a prediction equation for the time to the next eruption.

Draw the relevant summary graph for predicting interval from duration, and summarize your results.*

```
par(mfrow=c(1,1)) #redefining the dimensions of the graph
plot(Interval~Duration, data = oldfaith)
LinearReg <- lm(Interval~Duration, data = oldfaith) #using lm to establis the linear ana
lysis for graphing
LinearPlot <- abline(LinearReg) #graph using lm
```



This graph highlights how the data exists in two clusters in the bottom left quadrant and the top right quadrant. Without many observations in between the two groups we can observe that there is a relationship between short duration's having short intervals and large duration's having large intervals. If we were to splice the two clusters, the variance in the lower left quadrant would seem to be slightly bigger as well compared to the other cluster.

2.2 UBSprices

1. *The line with equation $y = x$ is shown on this plot as the solid line. What is the Key difference between points above this line and points below the line?*

The Key difference between the points above and below the $y = x$ point is the difference in the ratio of labor hours needed to buy 1kg of rice in 2003 vs. 2009. The $y = x$ line represents the slope of the line where there is a 1:1 relationship between the prices between the two years for each country. For example, a country like Budapest which is above the $y = x$ line has a roughly 1:3 ratio of 2003 labor hours to 2009 labor hours, meaning in 2009 rice cost 3x as more labor hours to buy; compared to Seoul with a roughly 2:1 ratio, the $x = y$ line is significant in determining what countries have to work more or less labor hours for the same amount of rice.

2. *Which city has the Largest increase in rice price? Which has the largest decrease in rice price?*

- Largest Increase: Vilnius
- Largest Decrease: Mumbaio

3. *The OLS line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is shown on the figure as a dashed line and evidently $\hat{\beta}_1 < 1$. Does this suggest that prices are lower in 2009 than in 2003?*

The OLS line suggests that the prices are lower in 2009 than in 2003. With ${*1} = estimate()* {1} < 1 x>y $. This means that with a slope < 1, the 2009 rice labor hours needed are estimated to be lower than the 2003 labor hours needed.

4. *Give two reasons why fitting a simple linear regression to the figure in this problem is not likely to be appropriate.*

- non-constant variance: The cluster of observations is extremely dense in the bottom left portion of the scatter plot and is more spread out otherwise.
- Non-Normality: After the 20:20 point on the graph, the observations seem to not follow the OLS line and shift upward with outliers potentially pulling the OLS line down. The points seem to curve and not follow the linear pattern the OLS line suggests.

2.3 UBSprices Cont.

1. *Explain why this graph and the graph in problem 2.2 suggest that using log-scale is preferable if fitting SLR is desired?*

When comparing both graphs, the Log transformation better accounted for both the variance and the non-normality. The distribution no longer appears to be as skewed and the outliers do not seem as influential in the new log graph. Overall, the log transformation has the data better fitting the OLS line.

2.
$$E(y|x) = \gamma_0 x^{\beta_1}$$
$$log(E(y|x)) = log(\gamma_0) + \beta_1 log(x)$$
$$E(log(y)|x) = \beta_0 + \beta_1 log(x)$$

… Give an interpretation of β_0 and β_1 in this setting, assuming $\beta_1 > 0$

- Interpretation of β_0: γ_0 is considered a constant in the first equation that is used to scale x^{β_1}. After applying the log to both sides, $log(\gamma_0)$ becomes the intercept of the log expectation of y|x. This means that for $\gamma_0 > 1$ the $E(log(y)|x)$ will have an upward shift in its graph. Conversely, a $\gamma_0 < 1$ will have a downward shift effect on the $E(log(y)|x)$ graph.
- Interpretation of β_1: The first equation has the β_1 as an exponent to x. This means that for $\beta_1 > 0$ there is exponential growth with exception to $\beta_1 = 1$ which is linear growth.As the log is taken, $E(log(y)|x)$ uses

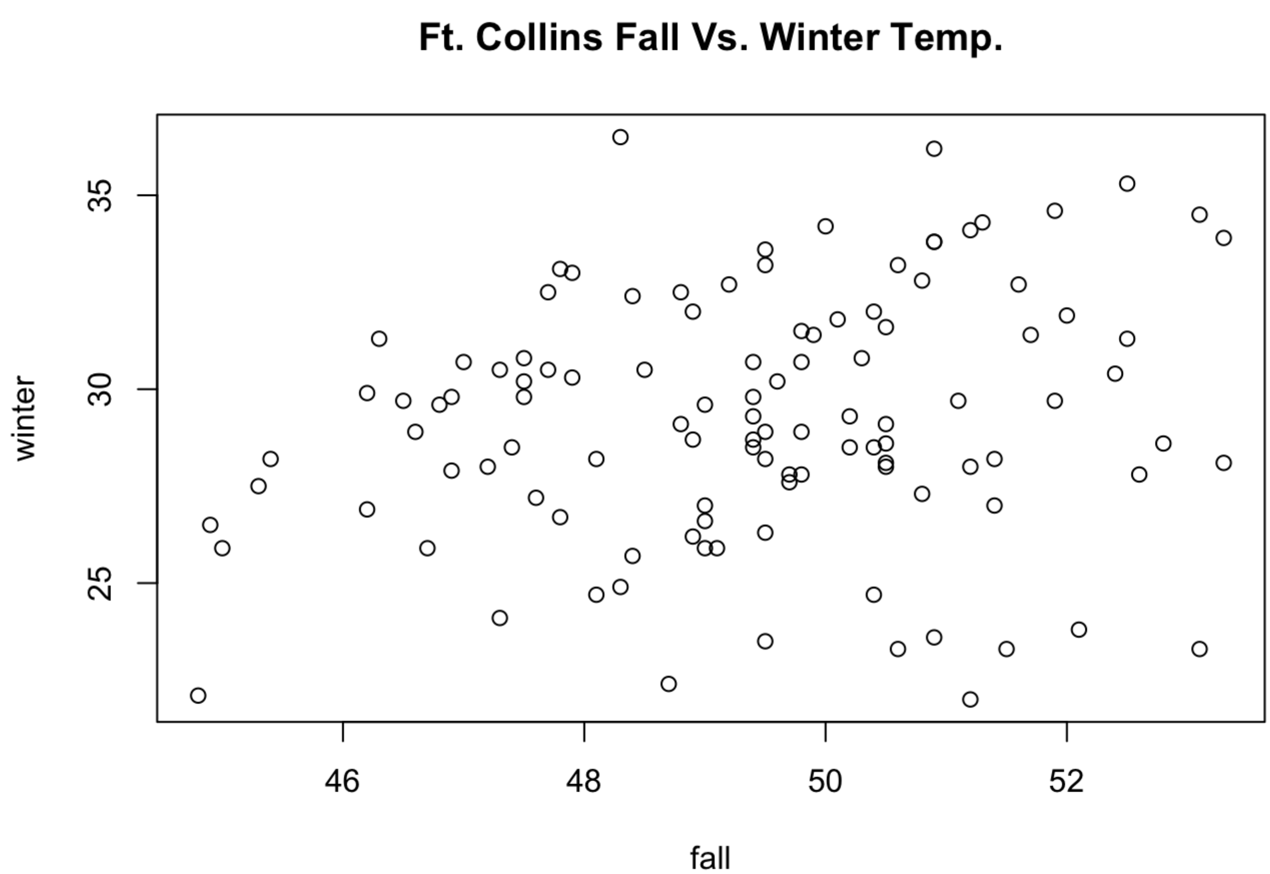
β_1 as a constant to scale the log(x) and provide a slope for the equations graph.

2.6: Ft. Collins Temperature Data

(data file: ftcollinstemp) The data file gives the mean temperature in the fall of each year, defined as September 1 - November 30, and the mean temp. in the following winter, def. as Dec 1 - end of Feb. in the following calendar year, in degrees Fahrenheit, for Ft. Collins, CO. These data cover the time period from 1900 - 2010. DOES THE AVERAGE FALL TEMP. PREDICT THE AVG. WINTER TEMP?

1. *Draw a scatter plot of the response vs predictor, describe any pattern you see in the plot*

```
Collins_fit <- plot(winter~fall, data = ftcollinstemp, main= "Ft. Collins Fall Vs. Winte
r Temp.")
```

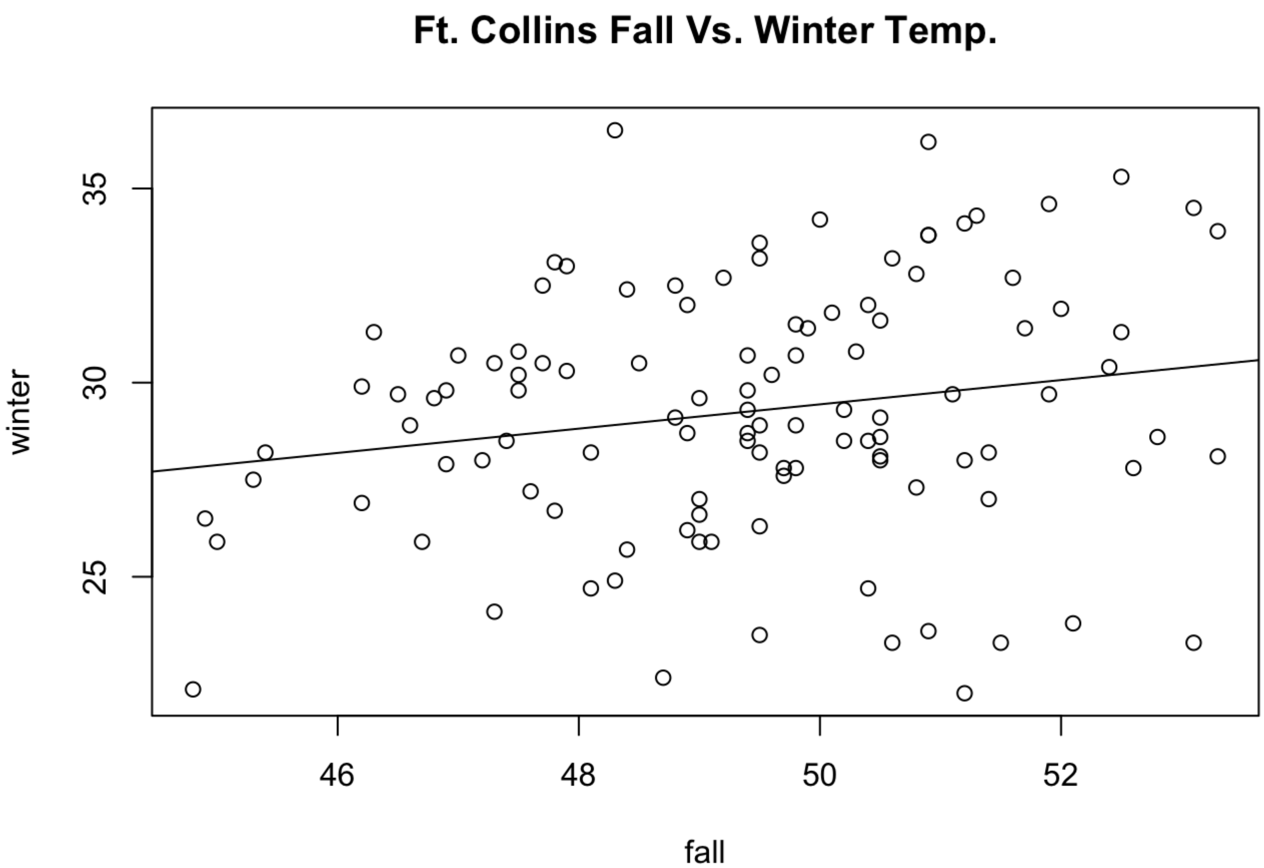


The scatter plot seems to show a weak, positive, linear relationship between the two variables. This weak positive correlation is not very apparent and the variance seems to increase in a coned shape as fall temperatures increase.

2. *Use statistical software to fit the regression of the response on the predictor. Add the fitted line to your graph. Test the slope to be 0 against a two-sided alternative, and summarize your results.*

```
Collins_fit <- plot(winter~fall, data = ftcollinstemp, main= "Ft. Collins Fall Vs. Winte
r Temp.")

# Applying a fitted line to the distribution
Collins_lm <- lm(winter~fall, data= ftcollinstemp)
abline(Collins_lm)
```




```
# Creating a two tailed hypothesis test with alpha = .05 to test if slope is equal to ze
ro
summary(Collins_lm)
```

```
##
## Call:
## lm(formula = winter ~ fall, data = ftcollinstemp)
##
## 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
```

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$
$$\alpha = 0.05$$

With a p-value of 0.0428 being less than our $\alpha = 0.05$ we can reject the null hypothesis and conclude that there is some relationship between fall and winter temperatures under a two-tailed t-test.

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

```
summary(Collins_lm)$r.squared
```

```
## [1] 0.03709854
```

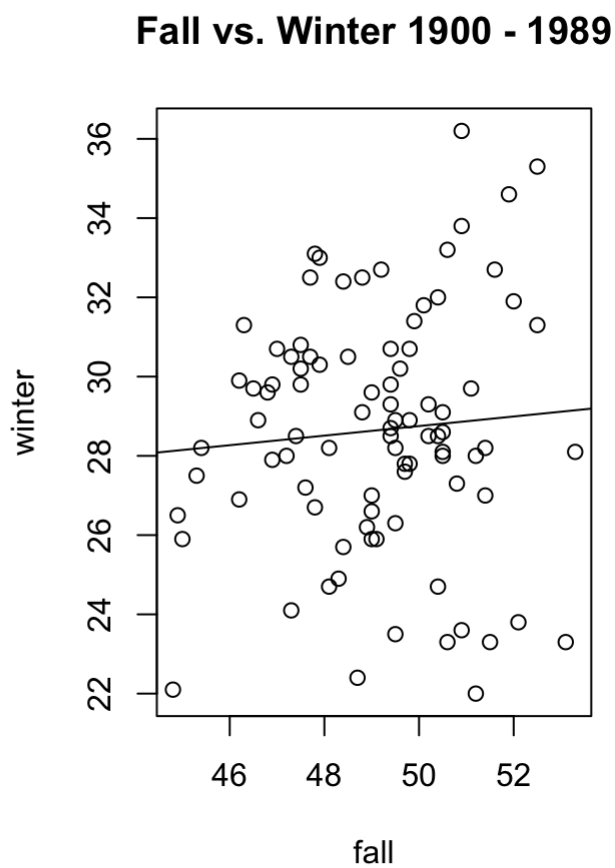
Analysis: About 3.7% of the variation in winter temperatures at Ft. Collins can be explained by the fall temperatures.

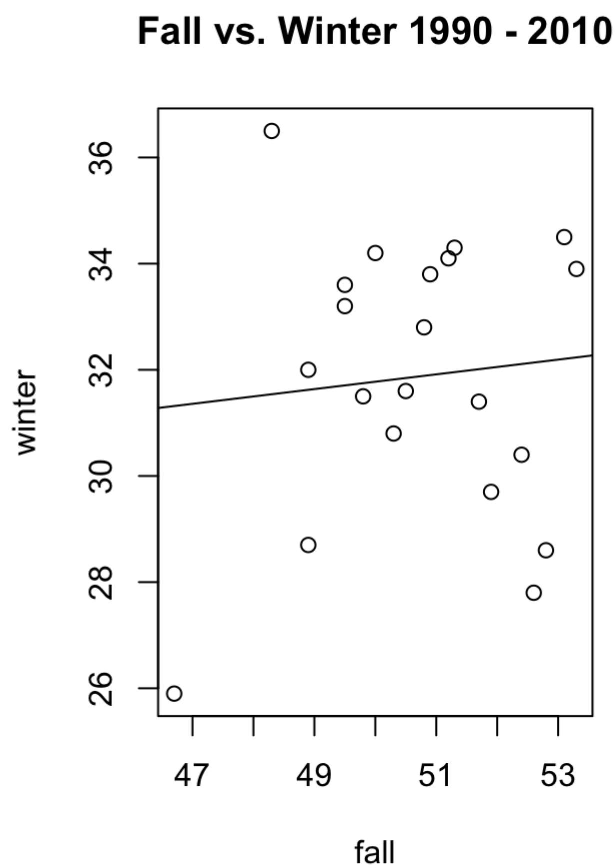
4. *Dicide the data into 2 time periods, an early period form 1900 to 1989, and a later period from 1990 - 2010. You can do this using the variable year in the data file. are the results different in the two time periods?*

```
# Subsetting the data to have the two time periods
Early_pd <- subset(ftcollinstemp, year < 1990)
Late_pd <- subset(ftcollinstemp, year > 1989)

par(mfrow= c(1,2))
# graphing the early period
Early_plot <- plot(winter~fall, data = Early_pd, main = "Fall vs. Winter 1900 - 1989")
Early_lm <- lm(winter~fall, data = Early_pd) #using lm to establis the linear analysis f
or graphing
Early_fit <- abline(Early_lm) #graph using lm

# Graphing the Late period
Latey_plot <- plot(winter~fall, data = Late_pd, main = "Fall vs. Winter 1990 - 2010")
Late_lm <- lm(winter~fall, data = Late_pd) #using lm to establis the linear analysis for
graphing
Late_fit <- abline(Late_lm) #graph using lm
```





Noticeable differences in the data show that the early period had more observations and thus had slightly greater variability. ALso, with fewer observations the estimate line is shifted upward with a relatively similar slope. These two graphs are similar in that their slopes are the same weakly positive relationship between fall and winter. However, they are different in that the intercept for the later period starts at roughly 31 degrees while the early period begins at roughly 28. The fewer observations probably account for this difference.

2.8: Deviations From the Mean

Sometimes it is convienient to write the simple linear regression model in a diff. form that is a little easier to manipulate. Taking Eq. (2.1), and adding $\beta_1\bar{x} - \beta_1\bar{x}$, which equals 0, to the right-hand side, and combining terms. we write…

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

… Where we have defined $\alpha = \beta_0 + \beta_1\bar{x}$. This is called the deviations from the sample mean form for simple regression.

1. what is the meaning of parameter α?

The meaning of α is the $E(Y|X = \bar{x})$. This parameter is an unknown quantity measuring the rate of change in $E(Y|X = \bar{x})$ for a unit change in $\bar{x}$. Essentially, the parameter is the true population representation of Least Squares Estimates, measuring the true population response mean.

2. Show the Least Squares Estimates are $\hat{\alpha} = \bar{y}, \hat{\beta}_1$ as given by (2.5).

$\alpha = \beta_0 + \beta_1\bar{x} \Rightarrow \hat{\alpha} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}$ and by pg. 27, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ s.t. $\hat{\alpha} = \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1\bar{x} = \bar{y}$ the

$\hat{E}(Y|X = \bar{x}) = \bar{y} - \hat{\beta}_1 + \hat{\beta}_1 = \bar{y} = \alpha$ making it the least squares estimate.

3. Find the expressions for the variances of the estimates and the covariances between them

$$Var(\hat{\beta}_1|X) = \sigma^2\frac{1}{SXX}$$

$$Var(\hat{\beta}_0|X) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{SXX})$$

$$\begin{aligned} Cov(\hat{\beta}_0, \hat{\beta}_1|X) &= Cov(\bar{y} - \hat{\beta}_1\bar{x}, \hat{\beta}_1|X) \\ &= Cov(\bar{y}, \hat{\beta}_1|X) - \bar{x}Var(\hat{\beta}_1|X) \\ &= -\sigma^2\frac{\bar{x}}{SXX} \end{aligned}$$

2.9: Invariance

1. In the SLR model (2.1), suppose the value of the predictor X is replaced by $Z = aX + b$, where $a \neq 0$ and b are constants. Thus, we are considering 2 Simple Regression Models,

$$\begin{aligned} I &: E(Y|X = x) = \beta_0 + \beta_1 x \\ II &: E(Y|Z = z) = \gamma_0 + \gamma_1 z \\ &= \gamma_0 + \gamma 1(ax + b) \end{aligned}$$

Find the relationships between β_0 and γ_0; β_1 and γ_1; between the estimates of variance in the two regressions, and between the t-tests of $\beta_1 = 0$ and of $\gamma_1 = 0$

ANSWER:

$$\beta_0 = \gamma_0 + \gamma_1 b; \beta_1 = \gamma_1 a$$
$$\Rightarrow \gamma_1 = \frac{\beta_1}{a}; \gamma_0 = \beta_0 - \frac{\beta_1 b}{a}$$
$$\Rightarrow E(Y|Z = z) = \gamma_0 + \gamma_1 z$$
$$= \gamma_0 + \gamma_1(ax + b)$$
$$= \gamma_0 + \gamma_1 ax + \gamma_1 b$$
$$= (\gamma_0 + \gamma_1 b) + (\gamma_1 a)x$$
$$= \beta_0 + \beta_1 x$$
$$= E(Y|X = x)$$

Because the variable Y was not changed between the two models, statistics such as the Variance and R^2 will not be affected; For the t-tests, $\gamma_1 = 0$ has the realationship of $\gamma_1 b$ relying on the constant b. This means that the intercept with will changing and become rejected.

2. Suppose each value of the response Y is replaced by V=dY, for some $d \neq 0$, so we consider the two reg models

$$I : E(Y|X = x) = \beta_0 + \beta_1 x$$
$$II : E(V|X = x) = \lambda_0 + \lambda_1 x$$

Find the relationship between $\beta_0 \cap \lambda_0$; between $\beta_1 \cap \lambda_1$; between the estimates of variances in the 2 regressions; and between the t test of $\beta_1 = 0 \cap \lambda_1 = 0$

ANSWER:

$$E(Y|X = x) = \beta_0 + \beta_1 x$$
$$= dE(Y|X = x) = d\beta_0 + \beta_1 x$$
$$= E(dY|X = x) = d\beta_0 + d\beta_1 x$$
$$= E(V|X = x)$$

Because the intercept, Y, and the slope are multiplied by the constant d, their parameters are changed. This means that the variance changes with the slope and intercept. However, these changes do not appply to R^2 or the t-tests due to the fact that scaling the function does not manipulate the outcome of the parameter.

2.13: Heights of Mothers and Daughters (data file: Heights)

1. Compute the regression of dheight and mheight, and report the estimates, their standard errors, the value of the coefficient of determination, and the estimate of variance. Write a sentence or two that summarizes the result of these computations.

ANSWER:

```
# Finding the Estimates, standard error, R^2
Heights_lm <- lm(dheight~mheight, data = Heights)
summary(Heights_lm)
```

```
##
## Call:
## lm(formula = dheight ~ mheight, data = Heights)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.397 -1.529  0.036  1.492  9.053
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.91744    1.62247   18.44   <2e-16 ***
## mheight      0.54175    0.02596   20.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.266 on 1373 degrees of freedom
## Multiple R-squared:  0.2408, Adjusted R-squared:  0.2402
## F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
```

```
# finding the mean and var of both columns
summary(Heights)
```

```
##     mheight         dheight
##  Min.   :55.40   Min.   :55.10
##  1st Qu.:60.80   1st Qu.:62.00
##  Median :62.40   Median :63.60
##  Mean   :62.45   Mean   :63.75
##  3rd Qu.:63.90   3rd Qu.:65.60
##  Max.   :70.80   Max.   :73.10
```

```
var(Heights)
```

```
##          mheight  dheight
## mheight 5.546511 3.004806
## dheight 3.004806 6.760274
```

With a p-value of <2e-16 we can reject the null hypothesis with an α of 0.05 that $\beta_1 = 0$. We can also see that 24% of the variation in dheight is explained by the model.

2. Obtain a 99% confidence interval for β_1 from the data using "Confint", "predict", and then other method.

```
attach(Heights) #sets the Heights data as a default table s.t. i can use variable names
 in functions

#99% Confidence w/ confint
M1 <- confint(Heights_lm, level = 0.99)
M1
```

```
##                  0.5 %     99.5 %
## (Intercept) 25.7324151 34.1024585
## mheight      0.4747836  0.6087104
```

```
# 99% confidence w/ predict -- need to create a new data frame for the interval to be ap
plied
M2 <- predict(Heights_lm, newdata = data.frame(mheight=c(1)), interval = "prediction", l
evel = 0.99)
M2
```

```
##        fit      lwr      upr
## 1 30.45918 23.30854 37.60983
```

```
#other method for confidence interval, manual by hand. ref. eq. on pg. 30
M3 <-cbind(CIlower = 0.542- 2.329*0.026, CIupper= 0.542+2.329*0.026 )
M3
```

```
##       CIlower  CIupper
## [1,] 0.481446 0.602554
```

3. Obtain a prediction and 99% confidence interval for a daughter whose mother is 69 inches tall using “Confint”, “predict”, and then other method.

```
# Applying condition of confidence interval for predicting daughters height for specific
mother height
M1a <- confint(Heights_lm, mheight= 64, level = 0.99)
M1a
```

```
##                  0.5 %     99.5 %
## (Intercept) 25.7324151 34.1024585
## mheight      0.4747836  0.6087104
```

```
M2a <- predict(Heights_lm, data.frame(mheight=64), interval = "prediction", level = .99)
M2a
```

```
##        fit      lwr      upr
## 1 64.58925 58.74045 70.43805
```

```
#prediction interval by hand (ref. pg. 32)
n  = length(Heights)
SampleSigmaSq = var(mheight)
SXX = var(mheight)*(n-1)
pred <- 29.917+0.542*64
pred
```

```
## [1] 64.605
```

```
SePred <- SampleSigmaSq*(1+(1/n)+(64-62.45)/SXX)
SePred
```

```
## [1] 9.869767
```

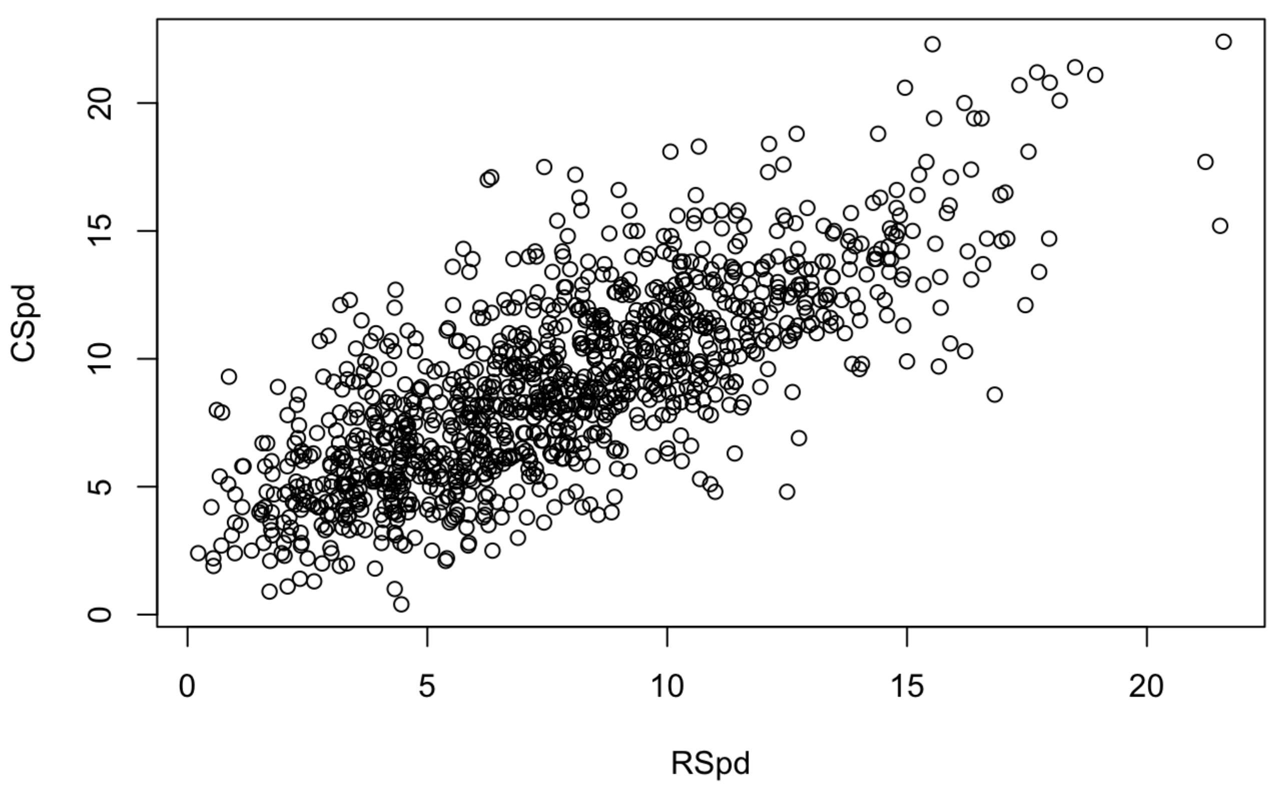
```
Pred_Int <- cbind(CIlower = pred-2.329*9.87, CIupper= pred+2.329*9.87)
Pred_Int
```

```
##       CIlower  CIupper
## [1,] 41.61777 87.59223
```

2.21: Windmills

1. Draw a scatterplot of the response CSpd vs. the predictor RSpd, and present the appropriate regression summaries.

```
WindSc <- plot(CSpd~RSpd, data = wm1)
```



A simple regression model looks reasonable for this data as the variance seems to be constant, the points seem to have a stong linear relationship, and there doesn't seem to be outliers that have undue influence on the data.

2. Fit the simple regression of the response on the predictor, and present the approptriate regression summaries.

```
Wind_lm <- lm(CSpd~RSpd, data = wm1)
summary(Wind_lm)
```

```
##
## Call:
## lm(formula = CSpd ~ RSpd, data = wm1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7877 -1.5864 -0.1994  1.4403  9.1738
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.14123    0.16958   18.52   <2e-16 ***
## RSpd         0.75573    0.01963   38.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.466 on 1114 degrees of freedom
## Multiple R-squared:  0.5709, Adjusted R-squared:  0.5705
## F-statistic:  1482 on 1 and 1114 DF,  p-value: < 2.2e-16
```

We can see that 57% of the variation in CSpd is explained by its relationship with RSpd. We can also reject the null hypothesis that $\beta_0 = 0$ at a $\alpha = 0.05$ which suggests that there is a relationship between both of the variables.

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

```
Wind_predint <- predict(Wind_lm, data.frame(RSpd=7.4285), interval = "prediction", level
= .95)
Wind_predint
```

```
##        fit      lwr      upr
## 1 8.755197 3.914023 13.59637
```

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, and (2) using the first result, the standard error the avg. of *m* predictions is… (ref. pg. 50)

- Let $\tilde{y}_{*i}$ be the variable representing all individual predictions of CSpd where $\tilde{y}_{*i} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_{*i}$ is the prediction taken at $\bar{x}_*$
  Let the average of the m predictions be $\frac{1}{m} \sum_{i=1}^{m} \tilde{y}_{*i}$
  Also, let the prediction taken at $\bar{x}_*$ be $\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_*$

$$\Rightarrow \frac{1}{m}\sum_{i=1}^{m}\tilde{y}_{*i} = \frac{1}{m}\sum_{i=1}^{m}(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_{*i})$$

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

$$= \hat{\beta}_0 + \hat{\beta}_1\bar{x}_*$$

s.t. the average of the *m* predictions is equal to the prediction taken at the average value $\bar{x}_*$

- if $Var(\hat{y}|X = x) = Var(\hat{\beta}_0 + \hat{\beta}_1 x|X = x) = \sigma^2(\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX})$

$$\Rightarrow Var(\tilde{y}_*|X = x_*) = \sigma^2(\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}) + \sigma^2$$

  this is because according to the appendix (A.4, pg. 296-7) “the variance of a prediction consists of the variance of the fitted value at x_* … and the variance of the error that will be attacked to the future value”

now, this shows that the standard error is…

$$\sqrt{\sigma^2(\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}) + \frac{\sigma^2}{m}}$$

… This is because for the se of the average prediction, we need to take the square root of the variance and the added error needs to be devided by *m* to provide the mean value for the future error prediction.