498818_SDS496_HW2

2024-09-10

Homework Set 2 SDS 496

```
# Read in the Data for Problem 1 and 2
photo = read.csv("/Users/aidanashrafi/Downloads/Reich2018NaturePaperDataAug2018.csv")
data = read.csv("/Users/aidanashrafi/Downloads/horse-colic.csv")
str(photo)
```

```
## 'data.frame':    2052 obs. of  18 variables:
##  $ site                        : chr  "cfc" "cfc" "cfc" "cfc" ...
##  $ block                       : chr  "d" "d" "d" "d" ...
##  $ warming_treatment           : chr  "ambient" "ambient" "ambient" "ambient" ...
##  $ plot_id                     : chr  "d4" "d4" "d5" "d5" ...
##  $ species                     : chr  "betpa" "aceru" "aceru" "betpa" ...
##  $ plant_id                    : int  403 383 503 519 1248 1222 1728 1714 2069 2099 ...
##  $ phylo                       : chr  "angio" "angio" "angio" "angio" ...
##  $ year                        : int  2009 2009 2009 2009 2009 2009 2009 2009 2009 2009 ...
##  $ doy                         : int  162 162 162 162 162 162 162 162 162 162 ...
##  $ Asat                        : num  15.26 7.61 9.49 14.8 16.33 ...
##  $ gs                          : num  0.195 0.103 0.182 0.278 0.25 ...
##  $ ci                          : num  237 259 288 281 259 ...
##  $ soil_water_VWC              : num  0.219 0.219 0.234 0.234 0.223 ...
##  $ tleaf                       : num  27.5 27.1 27.3 25.1 25.4 ...
##  $ VPG                         : num  2.11 1.85 2.15 1.41 1.53 ...
##  $ percent_stomatal_limitation: num  0.278 0.379 0.225 0.3 0.227 ...
##  $ Agmax                       : num  21.1 12.3 12.3 21.1 21.1 ...
##  $ Vcmax25                     : num  74.2 34 38.6 61.4 73.1 ...
```

Problem 1

i. What's the correlation between photosynthesis level and soil water content? Is there a positive or negative relationship between these variables?

```
# Clean Dataset before performing calculations if necessary
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
# Change data to data frame
photo_df<-as.data.frame(photo)
# Drop missing values
clean_photo<- drop_na(photo_df)
# Show identity of new data frame
str(clean_photo)
```

```
## 'data.frame':    1888 obs. of  18 variables:
##  $ site                        : chr  "cfc" "cfc" "cfc" "cfc" ...
##  $ block                       : chr  "d" "d" "d" "d" ...
##  $ warming_treatment           : chr  "ambient" "ambient" "ambient" "ambient" ...
##  $ plot_id                     : chr  "d4" "d4" "d5" "d5" ...
##  $ species                     : chr  "betpa" "aceru" "aceru" "betpa" ...
##  $ plant_id                    : int  403 383 503 519 1248 1222 1728 1714 2069 2099 ...
##  $ phylo                       : chr  "angio" "angio" "angio" "angio" ...
##  $ year                        : int  2009 2009 2009 2009 2009 2009 2009 2009 2009 2009 ...
##  $ doy                         : int  162 162 162 162 162 162 162 162 162 162 ...
##  $ Asat                        : num  15.26 7.61 9.49 14.8 16.33 ...
##  $ gs                          : num  0.195 0.103 0.182 0.278 0.25 ...
##  $ ci                          : num  237 259 288 281 259 ...
##  $ soil_water_VWC              : num  0.219 0.219 0.234 0.234 0.223 ...
##  $ tleaf                       : num  27.5 27.1 27.3 25.1 25.4 ...
##  $ VPG                         : num  2.11 1.85 2.15 1.41 1.53 ...
##  $ percent_stomatal_limitation: num  0.278 0.379 0.225 0.3 0.227 ...
##  $ Agmax                       : num  21.1 12.3 12.3 21.1 21.1 ...
##  $ Vcmax25                     : num  74.2 34 38.6 61.4 73.1 ...
```

```r
# Another way to clean dataset without tidyverse
photo_clean <- photo[!rowSums(is.na(photo)),]
str(photo_clean)
```

```
## 'data.frame':    1888 obs. of  18 variables:
##  $ site                        : chr  "cfc" "cfc" "cfc" "cfc" ...
##  $ block                       : chr  "d" "d" "d" "d" ...
##  $ warming_treatment           : chr  "ambient" "ambient" "ambient" "ambient" ...
##  $ plot_id                     : chr  "d4" "d4" "d5" "d5" ...
##  $ species                     : chr  "betpa" "aceru" "aceru" "betpa" ...
##  $ plant_id                    : int  403 383 503 519 1248 1222 1728 1714 2069 2099 ...
##  $ phylo                       : chr  "angio" "angio" "angio" "angio" ...
##  $ year                        : int  2009 2009 2009 2009 2009 2009 2009 2009 2009 2009 ...
##  $ doy                         : int  162 162 162 162 162 162 162 162 162 162 ...
##  $ Asat                        : num  15.26 7.61 9.49 14.8 16.33 ...
##  $ gs                          : num  0.195 0.103 0.182 0.278 0.25 ...
##  $ ci                          : num  237 259 288 281 259 ...
##  $ soil_water_VWC              : num  0.219 0.219 0.234 0.234 0.223 ...
##  $ tleaf                       : num  27.5 27.1 27.3 25.1 25.4 ...
##  $ VPG                         : num  2.11 1.85 2.15 1.41 1.53 ...
##  $ percent_stomatal_limitation: num  0.278 0.379 0.225 0.3 0.227 ...
##  $ Agmax                       : num  21.1 12.3 12.3 21.1 21.1 ...
##  $ Vcmax25                     : num  74.2 34 38.6 61.4 73.1 ...
```

Now that the data is clean, we can perform correlation calculations.

```r
# Extract all data from the specific columns of interest
photo_level<- photo_clean[c('Asat')]
soil_content<- photo_clean[c('soil_water_VWC')]
```

```r
# Make sure it works
str(photo_level)
```

```
## 'data.frame':    1888 obs. of  1 variable:
##  $ Asat: num  15.26 7.61 9.49 14.8 16.33 ...
```

```r
str(soil_content)
```

```
## 'data.frame':    1888 obs. of  1 variable:
##  $ soil_water_VWC: num  0.219 0.219 0.234 0.234 0.223 ...
```

```r
min(soil_content)
```

```
## [1] 0.05026949
```

```r
max(soil_content)
```

```
## [1] 0.2674575
```

```r
min(photo_level)
```

```
## [1] -0.07440594
```

```r
max(photo_level)
```

```
## [1] 31.986
```

Now that we have ensured everything is as desired we can answer the question.

```r
# This will calculate the correlation between photosynthesis level
# and soil water content
cor(photo_level,soil_content)
```

```
##      soil_water_VWC
## Asat       0.362666
```

The correlation between soil water content and photosynthesis level is 0.362666. There is a positive relationship between these variables.

ii. Fit a suitable simple linear model and obtain a summary of its results. Is there a positive or negative relationship between the predictor and the response variable? Does your answer agree with that of the previous part?

```r
lm <- lm(Asat~soil_water_VWC,data=clean_photo) # Fit simple linear model
summary(lm) # obtain summary of its results
```

```
##
## Call:
## lm(formula = Asat ~ soil_water_VWC, data = clean_photo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5167  -4.2702  -0.3467   3.9434  22.4416
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.0523     0.3542   14.26   <2e-16 ***
## soil_water_VWC  36.9640     2.1872   16.90   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.34 on 1886 degrees of freedom
## Multiple R-squared:  0.1315, Adjusted R-squared:  0.1311
## F-statistic: 285.6 on 1 and 1886 DF,  p-value: < 2.2e-16
```

There is a positive relationship between the predictor and the response variable. Previously, I found a positive correlation of 0.362666 between the predictor and the response variable, which also confirms the positive relationship.

iii. What's the null hypothesis of the t test for the soil water content coefficient shown in the summary of the linear model results? What's the conclusion of the hypothesis test based on the corresponding p-value? Write your answer in terms of the relationship between the predictor and the response variable rather than in terms of the slope coefficient.

Assumption: Testing for 95% confidence level, as tests are typically done at this level. The null hypothesis of the t test for the soil water content coefficient is $H_0 : B_1 = 0$, where B_1 is the beta coefficient for soil_water_content. The conclusion of the hypothesis test based on the corresponding p-value is to reject the null in favor of the alternative, $H_a : B_1 \neq 0$. We arrive at this conclusion, because the p-value, <2e-16 < 0.05. Because we have achieved statistical significance for this variable, we can now claim something about the relationship between photosynthesis level (Asat) and soil water content (soil_water_VWC).

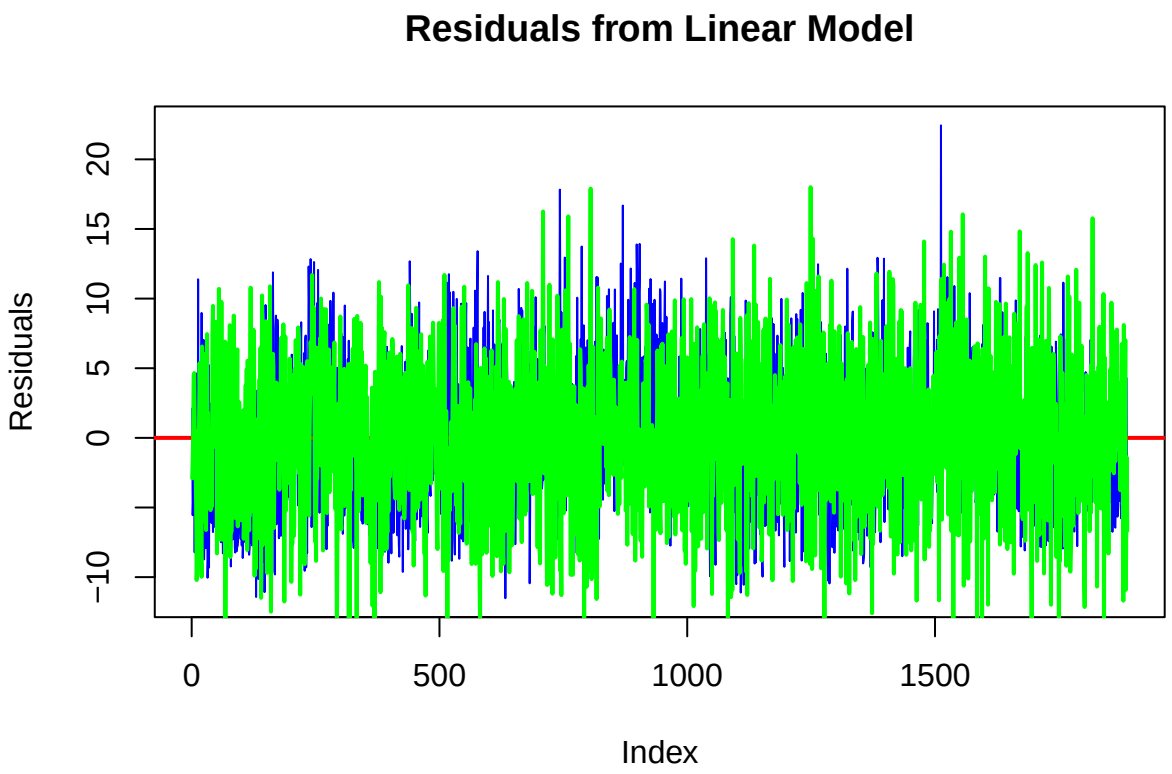
The estimated coefficient of 36.964 means that for each percentage increase in soil water content (soil_water_VWC), the photosynthesis level (Asat) is expected to increase by 36.964 * 0.01 = 0.36964 units based on this linear model, holding all else constant. We are 95 % confident that the true coefficient (true effect of soil water content on photosynthesis level) lies within the confidence interval around 0.36964. (This is how I structured statistical insights from a previous course in Linear Regression)

iv. Does the distribution of the residuals from the linear model appear to be roughly symmetric?

```r
# Grab residuals from lm model
resid <- lm$residuals
# Plot residuals
plot(resid, main = "Residuals from Linear Model",
     xlab = "Index", ylab = "Residuals",
     pch = 19, col = "blue", type='l')

# Add a horizontal line at 0 for analysis
abline(h = 0, col = "red", lwd = 2)

# Fit a normal distribution for the residuals to see if they could be symmetric
x_vals<- seq_along(resid)
normal<- rnorm(x_vals, mean = mean(resid), sd=sd(resid))
lines(normal,col='green',lwd=2)
```



The distribution of the residuals from the linear model appear to be roughly symmetric as they follow a normal distribution pretty well, the overlap between green and blue lines are very evident of this symmetry.

v. What insight does the coefficient of determination give about this linear model?

The coefficient of determination, denoted as $R^2 = 0.1315$. In this linear model, R^2 tells us the proportion of the variance in the response variable (photosynthesis level) that is explained by the predictor variable (soil water content). **An R^2 of 0.1315 means that 13.15 % of the variability in the photosynthesis level can be explained by the soil water content based on this linear model.**

This suggests that while soil water content has some predictive power, it does not explain a large portion of the variance in photosynthesis levels. There may be factors or variables contributing to the variability in photosynthesis levels (Asat) that are not included in this simple linear model!

vi. How many are the residual degrees of freedom for this model? Can you get the same answer by just looking at the number of observations in the sample and the number of predictors in the linear model? What's an unbiased estimate for the variance of the random error terms?

The residual DF in a linear regression model are calculated as: \$DF = n - p - 1 \$. where n is the number of observations, and p is the number of predictors in the model. Yes, you can get the answer by just looking at the number of observatios in the sample and the number of predictors in the linear model. This is a simple linear model, so we have p = 1, we have 1888 observations so n = 1888, so DF = 1888 - 2 = 1886. **Residual DF = 1886**.

An unbiased estimate for the variance of the random error terms is called the Mean Square Error (MSE). It is calculated as SSE/DF. The Residual DF = square root(MSE)

```
# WARNING
# THIS RESIDUAL DF MAY BE DIFFERENT
# AS A RESULT OF NA HANDLING FOR THE DATASET
```

vii. What's your prediction about the photosynthesis level of a plant with a soil water content value of 0.15? Provide the corresponding 95% prediction interval. What about a plant with a soil water content value of 0.25? Which of the 2 prediction intervals is wider?

```
# Make predictions with confidence intervals
new_data <- data.frame(soil_water_VWC=c(0.15, 0.25))
predictions <- predict(lm, new_data, interval = "prediction", level = 0.95)
predictions
```

```
##        fit       lwr      upr
## 1 10.59689 0.1219596 21.07181
## 2 14.29329 3.8099119 24.77666
```

The second prediction interval is wider, which intuitively makes sense because we have increased soil water content by more than 66 %, and 0.27 is the maximum observed value in the dataset. **Since we are farther from the mean of the data, this model has to account for several factors:**

Extrapolation beyond the mean: The model is predicting values further from the central range of the data, where it is less certain about the relationship.

More room for error: The data becomes sparse in the regions, so the models allows for greater variability in its precisions.

Increased uncertainity in the fit: Predictions made far from the mean have more uncertainity because the model's reliability decreases as it moves away from regions densely populated with data points.

viii. Create a scatterplot illustrating the relationship between photosynthesis level and soil water content. Overlay the fitted regression line with 95% point-wise prediction intervals for every value of the predictor.

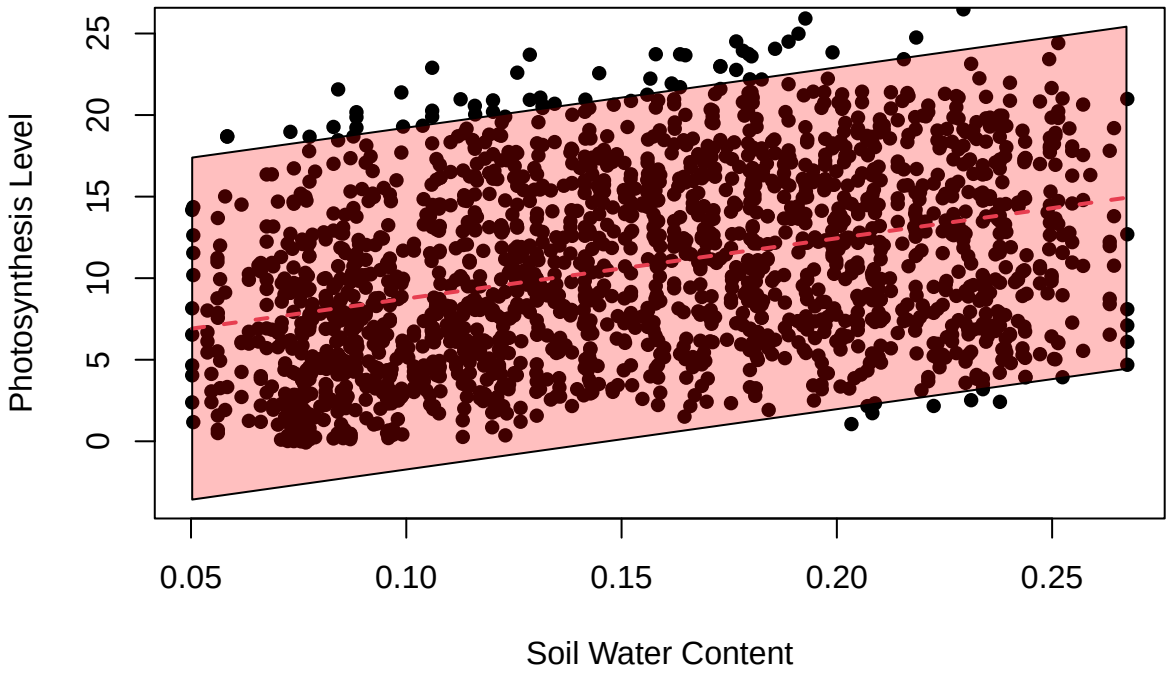
```
# Define the sequence of x values (soil water content in this case)
x <- seq(min(clean_photo$soil_water_VWC), max(clean_photo$soil_water_VWC), 1e-3)

# Get predictions along with prediction intervals
predictions <- predict(lm, newdata = data.frame(soil_water_VWC = x), interval = "prediction")

# Plot the original data
plot(Asat ~ soil_water_VWC, data = clean_photo,
     ylim = range(predictions),
     pch = 16,
     xlab = "Soil Water Content",
     ylab = "Photosynthesis Level")

# Add the fitted line (predicted values)
lines(x, predictions[,1],'l', col = 2, lty = 2, lwd = 2)

# Add the prediction intervals as shaded areas
polygon(c(x, rev(x)),
        c(predictions[,2], rev(predictions[,3])),
        col = rgb(1, 0, 0, 0.25))
```



This code was taken from lecture slides and translated to fit my linear model and data.

Problem 2

```
# Grab the variables of interest
target_data <- data[, c("respiratory.rate", "pulse","packed.cell.volume","rectal.temperature")]
# Make sure we extracted the variables of interest
str(target_data)
```

```
## 'data.frame':    300 obs. of  4 variables:
##  $ respiratory.rate  : int  28 20 24 84 35 NA 16 NA 36 NA ...
##  $ pulse             : int  66 88 40 164 104 NA 48 60 80 90 ...
##  $ packed.cell.volume: num  45 50 33 48 74 NA 37 44 38 40 ...
##  $ rectal.temperature: num  38.5 39.2 38.3 39.1 37.3 NA 37.9 NA NA 38.3 ...
```

```
# Clean data using tidyverse, drop_na() function
clean_data <- drop_na(target_data)
str(clean_data)
```

```
## 'data.frame':    194 obs. of  4 variables:
##  $ respiratory.rate  : int  28 20 24 84 35 16 12 52 28 28 ...
##  $ pulse             : int  66 88 40 164 104 48 66 72 92 76 ...
##  $ packed.cell.volume: num  45 50 33 48 74 37 44 50 37 46 ...
##  $ rectal.temperature: num  38.5 39.2 38.3 39.1 37.3 37.9 38.1 39.1 38 38.2 ...
```

```
# Now that we have cleaned the data, we are able to perform regression as desired
```

i. Fit linear models:

```
model1 <- lm(respiratory.rate ~ pulse ,data=clean_data)
summary(model1)
```

```
##
## Call:
```

```
## lm(formula = respiratory.rate ~ pulse, data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.511  -9.154  -3.640   4.928  71.203
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.00806    2.93798   3.406 0.000802 ***
## pulse        0.28439    0.03859   7.369 4.97e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.82 on 192 degrees of freedom
## Multiple R-squared:  0.2205, Adjusted R-squared:  0.2164
## F-statistic:  54.3 on 1 and 192 DF,  p-value: 4.97e-12
```

```
model2 <- lm(respiratory.rate ~ rectal.temperature, data=clean_data)
summary(model2)
```

```
##
## Call:
## lm(formula = respiratory.rate ~ rectal.temperature, data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.244 -12.695  -4.462   7.242  65.155
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -205.318     65.296  -3.144 0.001929 **
## rectal.temperature    6.166      1.711   3.604 0.000399 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.34 on 192 degrees of freedom
## Multiple R-squared:  0.06337,    Adjusted R-squared:  0.05849
## F-statistic: 12.99 on 1 and 192 DF,  p-value: 0.000399
```

```
model3 <- lm(respiratory.rate ~ packed.cell.volume, data = clean_data)
summary(model3)
```

```
##
## Call:
## lm(formula = respiratory.rate ~ packed.cell.volume, data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.435 -11.877  -4.781   4.565  66.642
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         19.4345     5.9010   3.293  0.00118 **
## packed.cell.volume   0.2308     0.1262   1.829  0.06890 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.77 on 192 degrees of freedom
## Multiple R-squared:  0.01713,    Adjusted R-squared:  0.01201
## F-statistic: 3.347 on 1 and 192 DF,  p-value: 0.0689
```

Can you compare the performance of these 3 candidate linear models on the basis of the coefficient of determination? If so, which of the 3 appears to be performing best?

Yes, you can compare the performance of these 3 candidate linear models on the basis of the coefficient of determination because all the models have the same number of predictors. The best model based on the 3 candidate models created thus far, is the model with the highest R^2, because R^2 tells us the proportion of the variance in the response variable (respiratory rate) that is explained by the predictor variable. Therefore, the model with the highest R^2 is the model that explains more of the variability in the data, which generally indicates a better fit to the data. **The first model with respiratory rate as the response and pulse as the predictor is the best simple linear model based on the 3 candidate linear models thus far.**

ii. Fit the multiple linear models:

```
model4<- lm(respiratory.rate ~ pulse + rectal.temperature, data=clean_data)
summary(model4)
```

```
##
## Call:
## lm(formula = respiratory.rate ~ pulse + rectal.temperature, data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.520  -9.076  -2.698   5.131  70.271
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -143.1526    59.4129  -2.409   0.0169 *
## pulse                 0.2646     0.0388   6.819 1.17e-10 ***
## rectal.temperature    4.0502     1.5693   2.581   0.0106 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.59 on 191 degrees of freedom
## Multiple R-squared:  0.2467, Adjusted R-squared:  0.2389
## F-statistic: 31.28 on 2 and 191 DF,  p-value: 1.771e-12
```

```
model5 <- lm(respiratory.rate ~ packed.cell.volume + pulse + rectal.temperature, data=clean_data)
summary(model5)
```

```
##
## Call:
## lm(formula = respiratory.rate ~ packed.cell.volume + pulse +
##     rectal.temperature, data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.912  -9.348  -2.832   5.034  70.251
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -139.17049   59.86085  -2.325   0.0211 *
## packed.cell.volume   -0.07338    0.11928  -0.615   0.5392
## pulse                 0.27400    0.04177   6.560 4.96e-10 ***
## rectal.temperature    4.01637    1.57278   2.554   0.0114 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.62 on 190 degrees of freedom
## Multiple R-squared:  0.2482, Adjusted R-squared:  0.2364
## F-statistic: 20.91 on 3 and 190 DF,  p-value: 9.435e-12
```

Why can't you compare the performance of these 2 candidate linear models on the basis of the coefficient of determination? On the basis of which other measure can you compare them and which of the 2 appears to be performing best?

You cannot compare the performance of these 2 candidate linear models on the basis of the coefficient of determination because the number of predictor variables in the various models are not the same. **You can compare the performance of these 2 candidate linear models on the basis of the Adjusted R^2.**

The Adjusted R^2 is the coefficient of determinant accounting for the loss of degrees of freedom when adding additional predictor variables. Adjusted R^2 is used because R^2 will always increase when we add more predictor variables, so Adjusted R^2 is used to compare models with different number of predictors, penalizing the additional variables that don't contribute meaningful explanatory power. **This means that if the additional predictors do not improve the model, Adjusted R^2 will decrease.**

iii. Which one of models 4 and 5 has the smaller error sum of squares? Can you directly answer the question without calculating the error sum of squares of these models?

Error Sum of Squares (ESS), also known as Sum of Squares Error (SSE), represents the discrepancy between the observed data and the values predicted by the regression model. R gives us summary statistics, which include Residual Standard Error (RSE). The relationship between RSE and SSE is as follow:

$RSE = \sqrt{\frac{SSE}{\text{DF}}}$

Therefore, we can derive which model has the smaller SSE by using the information provided from RSE in the summary statistics. **Model 4: RSE = 15.59, Model 5: RSE = 15.62.**

Model 4 SSE = $15.59^2 * 191 $ = 46422.19

Model 5 SSE = $15.62^2 * 190 $ = 46357.04

Model 5 has the smaller SSE. This can be answered without calculating the SSE of these models because of the added flexibility additional predictors bring to the model.

As you add more variables to the regression model, the model becomes more complex, allowing it to better fit the relationship between the predictors and response variable. However, this added complexity can also capture unwanted noise, leading to overfitting. **Although R^2 will always increase when new variables are added, this doesn't necessarily improve the model's predictive power for unseen data. Adjusted R^2 helps account for this by penalizing the loss of degrees of freedom, allowing us to determine whether the additional variables truly add value or simply overfit the model.**

iv. What's the null hypothesis for the overall F test of significance for model 5? What's the conclusion of the hypothesis test based on the corresponding p-value? Write your answer in terms of the relationship between the predictors and the response variable rather than in terms of the regression coefficients.

The null hypothesis for the overall F test of significance for model 5 is as follows: H_0 : All predictor coefficients are equal to 0. This means that none of the predictor variables have any significant effect on the response variable, and thus they do not help explain the variability in the response. **In simple terms, the F test is checking whether at least one of the predictor variables contriubtes to explaining the variability in the response.** The alternative hypothesis is: $H_a : B_i \neq 0$ for at least one i. This alternative hypothesis means that at least one predictor variable has a significant impact on the response variable. Based on the p-value from the F test,$9.435e-12$, which is extremely small, we can reject the null in favor of the alternative.

This means that there is strong evidence to suggest that at least one of the predictor variables in Model 5 explains the variability in the response variable.

I was confused by the last question, so I am going to add the relationship between each of the predictors and the response variable as well as the relationship from the overall F test conclusion.

Based on the p-value for packed cell volume, > 0.05, it is not statistically significant, meaning that we cannot claim that packed cell volume is a good predictor variable to explain the variability in respiratory rate.

Based on the p-value for pulse, < 0.05, it is statistically significant, meaning that we can claim that pulse is a good predictor variable to help us explain the variability in respiratory rate.

Based on the p-value for rectal temperature, < 0.05, it is statistically significant, meaning we can claim that rectal temperature is a good predictor variable to help us explain the variability in respiratory rate.

In summary, while the model overall is significant (as indicated by the F-test), not all individual predictors contribute significantly to explaining the response.

v. Which one of models 4 and 5 has the better predictive performance? Try out 2 different ways of estimating the mean squared prediction error of each model to make sure that your result is consistent regardless of the method you used to estimate it.

Model 4 MPSE

```
# Training and Test Split
# Set seed for reproducibility
set.seed(123)
n = nrow(clean_data)
train = sample(n, 0.8*n)
test = setdiff(1:n,train)
fit = lm(respiratory.rate ~  pulse  + rectal.temperature , clean_data, train)
predictions = predict(fit, data.frame(pulse=clean_data$pulse[test],rectal.temperature=clean_data$rectal
MPSE = mean((clean_data$respiratory.rate[test]-predictions)^2)
MPSE
```

```
## [1] 494.4867
```

Another way to calculated MPSE for Model 4

```
# Training and Test Split
# Set seed for reproducibility
set.seed(123)
n = nrow(clean_data)
train = sample(n, 0.8*n)  # 80% of data for training
test = setdiff(1:n, train)  # Remaining 20% for testing

# Fit model using training data
```

```r
fit2 = lm(respiratory.rate ~ pulse + rectal.temperature, data = clean_data[train,])

# Make predictions on the test data
predictions2 = predict(fit2, newdata = clean_data[test,])

# Calculate Mean Squared Prediction Error (MSPE)
MSPE2 = mean((clean_data$respiratory.rate[test] - predictions2)^2)
MSPE2
```

```
## [1] 494.4867
```

The Above code was two ways to calculate the MSPE of Model 4 using Training-Test Split Strategy.

Now we will calculate the MSPE of Model 5 using Training-Test Split.

```r
set.seed(123)
n = nrow(clean_data)
train = sample(n, 0.8*n)
test = setdiff(1:n,train)
fit3 = lm(respiratory.rate ~  pulse  + rectal.temperature + packed.cell.volume, clean_data, train)
predictions3 = predict(fit3, data.frame(pulse=clean_data$pulse[test],rectal.temperature=clean_data$recta
MPSE3 = mean((clean_data$respiratory.rate[test]-predictions3)^2)
MPSE3
```

```
## [1] 494.8867
```

```r
# Training and Test Split
# Set seed for reproducibility
set.seed(123)
n = nrow(clean_data)
train = sample(n, 0.8*n)  # 80% of data for training
test = setdiff(1:n, train)  # Remaining 20% for testing

# Fit model using training data
fit4 = lm(respiratory.rate ~ pulse + rectal.temperature + packed.cell.volume, data = clean_data[train,]

# Make predictions on the test data
predictions4 = predict(fit4, newdata = clean_data[test,])

# Calculate Mean Squared Prediction Error (MSPE)
MSPE4 = mean((clean_data$respiratory.rate[test] - predictions4)^2)
MSPE4
```

```
## [1] 494.8867
```

We have just used Training-Test Split Strategy for MSPE and found that MSPE for Model 4 is lower than Model 5.

We will now try Leave-One-Out Cross Validation for Model 4 and Model 5.

```r
# Model 4
set.seed(123)
MSPE5 = 0
for (i in 1:n)
{
  fit5 = lm(respiratory.rate ~ pulse + rectal.temperature,clean_data,setdiff(1:n,i))
  prediction5 = predict(fit5,data.frame(pulse=clean_data$pulse[i],rectal.temperature=clean_data$rectal.
```

```
  MSPE5 = MSPE5 + (clean_data$respiratory.rate[i]-prediction5)^2
}
MSPE5 = MSPE5/n
MSPE5
```

```
##        1
## 248.4778
```

```
# Model 5
set.seed(123)
MSPE6 = 0
for (i in 1:n)
{
  fit = lm(respiratory.rate ~ pulse + rectal.temperature + packed.cell.volume,clean_data,setdiff(1:n,i))
  prediction6 = predict(fit,data.frame(pulse=clean_data$pulse[i],rectal.temperature=clean_data$rectal.te
  MSPE6 = MSPE6 + (clean_data$respiratory.rate[i]-prediction6)^2
}
MSPE6 = MSPE6/n
MSPE6
```

```
##       1
## 250.752
```

Based on the Leave-One-Out Cross Validation Strategy, Model 4 has a lower MSPE than Model 5.

Therefore, Model 4 has the better predictive performance, because after looking at 2 different ways of estimating the MSPE, Model 4 has a lower MSPE for both strategies, meaning that the results are consistent regardless of the method used.

Additionally, we are happy because this result agrees with the conclusion derived from our Adjusted R^2 comparison as well, since Model 4 has a higher Adjusted R^2 than Model 5, our predetermined inference that Model 4 is a better model than Model 5 for the data based on Adjusted R^2 is also justified by the MSPE results.