

Weekly Summary Template

Brady Miller

Table of contents

Tuesday, February 7th	2
Tuesday, February 9th	7

```
library(tidyverse)
```

```
-- Attaching packages ----- tidyverse 1.3.2 --
v ggplot2 3.4.0      v purrr   1.0.1
v tibble  3.1.8      v dplyr   1.1.0
v tidyr   1.3.0      v stringr 1.5.0
v readr   2.1.3      v forcats 1.0.0
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()     masks stats::lag()
```

```
library(ISLR2)
library(cowplot)
library(kableExtra)
```

Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

group_rows

Tuesday, February 7th

! TIL

Include a *very brief* summary of what you learnt in this class here.

Today, I learnt the following concepts in class:

1. What the interpretation of β_0 and β_1 is (regression coefficients)
2. Categorical covariates
3. How to reorder factors

Interpretation of β_0 and β_1 (regression coefficients)

- The regression model is...

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- β_0 and β_1 are the regression coefficients with β_0 being the intercept and β_1 being the slope

```
library(ggplot2)
attach(mtcars)
```

The following object is masked from package:ggplot2:

mpg

```
x <- mtcars$hp
y <- mtcars$mpg

model <- lm(y ~ x)
summary(model)
```

Call:

```
lm(formula = y ~ x)
```

Residuals:

Min	1Q	Median	3Q	Max
-5.7121	-2.1122	-0.8854	1.5819	8.2360

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	30.09886	1.63392	18.421	< 2e-16 ***
x	-0.06823	0.01012	-6.742	1.79e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.863 on 30 degrees of freedom

Multiple R-squared: 0.6024, Adjusted R-squared: 0.5892

F-statistic: 45.46 on 1 and 30 DF, p-value: 1.788e-07

Based on the summary model...

- The intercept means that a car with 'hp=0' would have 'mpg=30.09' ## Thursday, February 9th
- From this, if we have some co-variate x_0 , then the expected value for $y(x_0)$ is given by:

$$y(x_0) = \beta_0 + \beta_1 x_0$$

- Using this we can find the expected value for $x_0 + 1$, which is...

$$y(x_0 + 1) = \beta_0 + \beta_1 \times (x_0 + 1) = \beta_0 + \beta_1 x_0 + \beta_1 = y(x_0) + \beta_1$$

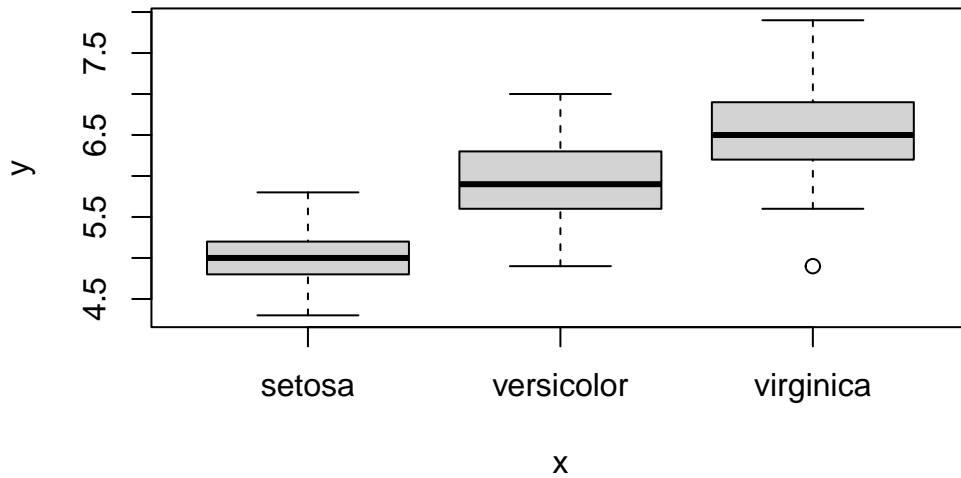
- This implies that $\beta_1 = y(x_0 + 1) - y(x_0)$

Categorical covariates

In class, we looked at categorical covariates in the iris dataset, and created a boxplot and a summary of the model for the categorical covariates

```
x <- iris$Species
y <- iris$Sepal.Length

# boxplot(Sepal.Length ~ Species, df)
boxplot(y ~ x)
```



```
cat_model <- lm(Sepal.Length ~ Species, iris)
cat_model
```

Call:

```
lm(formula = Sepal.Length ~ Species, data = iris)
```

Coefficients:

(Intercept)	Speciesversicolor	Speciesvirginica
5.006	0.930	1.582

Even if x is categorical, we can still write down the regression model as follows:

$$y_i = \beta_0 + \beta_1 x_i$$

where $x_i \in \{setosa, versicolor, virginica\}$. This means that we end up with, three different models with each one having a different intercept.

1. $y_i = \beta_0 + \beta_1(x_i == 'setosa')$
2. $y_i = \beta_0 + \beta_1(x_i == 'versicolor')$
3. $y_i = \beta_0 + \beta_1(x_i == 'virginica')$

- The interpretation of the intercept (β_0) is the expected y value when x belongs to the base category

- The slope (β_1) with the name ‘Species.versicolor’ represents the following:
- ‘(Intercept)’ = $y(x = \text{setosa})$
- ‘Species.versicolor’ = $y(x = \text{versicolor}) - y(x = \text{setosa})$
- ‘Species.virginica’ = $y(x = \text{virginica}) - y(x = \text{setosa})$

Reordering factors

Lets say that we didn’t want ‘setosa’ to be the baseline level, and instead, we wanted ‘virginica’ to be the baseline level. How would we do this?

First we reorder/relevel the categorical covariate

```
# before
iris$Species
```

```
[1] setosa    setosa    setosa    setosa    setosa    setosa
[7] setosa    setosa    setosa    setosa    setosa    setosa
[13] setosa    setosa    setosa    setosa    setosa    setosa
[19] setosa    setosa    setosa    setosa    setosa    setosa
[25] setosa    setosa    setosa    setosa    setosa    setosa
[31] setosa    setosa    setosa    setosa    setosa    setosa
[37] setosa    setosa    setosa    setosa    setosa    setosa
[43] setosa    setosa    setosa    setosa    setosa    setosa
[49] setosa    setosa    versicolor versicolor versicolor versicolor
[55] versicolor versicolor versicolor versicolor versicolor versicolor
[61] versicolor versicolor versicolor versicolor versicolor versicolor
[67] versicolor versicolor versicolor versicolor versicolor versicolor
[73] versicolor versicolor versicolor versicolor versicolor versicolor
[79] versicolor versicolor versicolor versicolor versicolor versicolor
[85] versicolor versicolor versicolor versicolor versicolor versicolor
[91] versicolor versicolor versicolor versicolor versicolor versicolor
[97] versicolor versicolor versicolor versicolor virginica  virginica
[103] virginica  virginica  virginica  virginica  virginica  virginica
[109] virginica  virginica  virginica  virginica  virginica  virginica
[115] virginica  virginica  virginica  virginica  virginica  virginica
[121] virginica  virginica  virginica  virginica  virginica  virginica
[127] virginica  virginica  virginica  virginica  virginica  virginica
[133] virginica  virginica  virginica  virginica  virginica  virginica
[139] virginica  virginica  virginica  virginica  virginica  virginica
[145] virginica  virginica  virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica
```

```
iris$Species <- relevel(iris$Species, "virginica")
```

```
# after  
iris$Species
```

```
[1] setosa    setosa    setosa    setosa    setosa    setosa  
[7] setosa    setosa    setosa    setosa    setosa    setosa  
[13] setosa    setosa    setosa    setosa    setosa    setosa  
[19] setosa    setosa    setosa    setosa    setosa    setosa  
[25] setosa    setosa    setosa    setosa    setosa    setosa  
[31] setosa    setosa    setosa    setosa    setosa    setosa  
[37] setosa    setosa    setosa    setosa    setosa    setosa  
[43] setosa    setosa    setosa    setosa    setosa    setosa  
[49] setosa    setosa    versicolor versicolor versicolor versicolor  
[55] versicolor versicolor versicolor versicolor versicolor versicolor  
[61] versicolor versicolor versicolor versicolor versicolor versicolor  
[67] versicolor versicolor versicolor versicolor versicolor versicolor  
[73] versicolor versicolor versicolor versicolor versicolor versicolor  
[79] versicolor versicolor versicolor versicolor versicolor versicolor  
[85] versicolor versicolor versicolor versicolor versicolor versicolor  
[91] versicolor versicolor versicolor versicolor versicolor versicolor  
[97] versicolor versicolor versicolor versicolor virginica  virginica  
[103] virginica  virginica  virginica  virginica  virginica  virginica  
[109] virginica  virginica  virginica  virginica  virginica  virginica  
[115] virginica  virginica  virginica  virginica  virginica  virginica  
[121] virginica  virginica  virginica  virginica  virginica  virginica  
[127] virginica  virginica  virginica  virginica  virginica  virginica  
[133] virginica  virginica  virginica  virginica  virginica  virginica  
[139] virginica  virginica  virginica  virginica  virginica  virginica  
[145] virginica  virginica  virginica  virginica  virginica  virginica  
Levels: virginica setosa versicolor
```

Once we do the re-leveling, we can now run the regression model:

```
new_cat_model <- lm(Sepal.Length ~ Species, iris)  
new_cat_model
```

Call:

```
lm(formula = Sepal.Length ~ Species, data = iris)
```

Coefficients:

(Intercept)	Speciessetosa	Speciesversicolor
6.588	-1.582	-0.652

Tuesday, February 9th

! TIL

Include a *very brief* summary of what you learnt in this class here.

Today, I learnt the following concepts in class:

1. How to make a plot for a model that incorporates more than 1 quantitative covariate
2. The impact of noise and β values on R^2
3. Multiple regression with categorical covariates

```
library(tibble)
library(ISLR2)
attach(Credit)

df <- Credit %>%
  tibble()
colnames(df) <- tolower(colnames(df))
df
```

A tibble: 400 x 11

	income	limit	rating	cards	age	educat~1	own	student	married	region	balance
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<fct>	<fct>	<fct>	<fct>	<dbl>
1	14.9	3606	283	2	34	11	No	No	Yes	South	333
2	106.	6645	483	3	82	15	Yes	Yes	Yes	West	903
3	105.	7075	514	4	71	11	No	No	No	West	580
4	149.	9504	681	3	36	11	Yes	No	No	West	964
5	55.9	4897	357	2	68	16	No	No	Yes	South	331
6	80.2	8047	569	4	77	10	No	No	No	South	1151
7	21.0	3388	259	2	37	12	Yes	No	No	East	203
8	71.4	7114	512	2	87	9	No	No	No	West	872
9	15.1	3300	266	5	66	13	Yes	No	No	South	279
10	71.1	6819	491	3	41	19	Yes	Yes	Yes	East	1350

... with 390 more rows, and abbreviated variable name 1: education

Plotting a model with more than 1 quantitative covariate

We will look at the following 3 columns: 'income, rating, limit'.

```
df3 <- df %>%  
  select('income', 'rating', 'limit')  
df3
```

```
# A tibble: 400 x 3  
  income rating limit  
  <dbl> <dbl> <dbl>  
1   14.9   283 3606  
2  106.   483 6645  
3  105.   514 7075  
4  149.   681 9504  
5   55.9   357 4897  
6   80.2   569 8047  
7   21.0   259 3388  
8   71.4   512 7114  
9   15.1   266 3300  
10  71.1   491 6819  
# ... with 390 more rows
```

If we want to see how the credit limit is related to income and credit rating, we can visualize the following plot

```
library(plotly)
```

Attaching package: 'plotly'

The following object is masked from 'package:ggplot2':

```
last_plot
```

The following object is masked from 'package:stats':

```
filter
```

The following object is masked from 'package:graphics':

```
layout
```



```
fig <- plot_ly(df3, x=~income, y=~rating, z=~limit)
fig %>% add_markers()
```

This models a linear relationship, which in 3-dimensions is a plane (a hyperplane)

- Was shown in class what the hyperplane looked like but weren't given the code, so couldn't include that in the summary

The regression model is as follows:

```
model <- lm(limit ~ income + rating, df3)
model
```

Call:

```
lm(formula = limit ~ income + rating, data = df3)
```

Coefficients:

(Intercept)	income	rating
-532.4711	0.5573	14.7711

- Have 3 different coefficients
- 2nd/3rd numbers are the slopes associated with the income and rating

What is the interpretation for the coefficients?

1. β_0 is the expected value of y when $income = 0$ and $rating = 0$ (the credit limit with 0 income and 0 rating is -532) \rightarrow this is an extrapolation
2. β_1 is saying that if $rating$ is held constant and $income$ changes by 1 unit, then the corresponding change in the 'limit' is 0.5573.
3. β_2 is saying that if $income$ is held constant and $rating$ changes by 1 unit, then the corresponding change in the 'limit' is 14.7711.

What about the significance?

```
summary(model)
```

Call:

```
lm(formula = limit ~ income + rating, data = df3)
```

Residuals:

Min	1Q	Median	3Q	Max
-420.97	-121.77	14.97	126.72	485.48

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-532.47115	24.17283	-22.028	<2e-16 ***
income	0.55727	0.42349	1.316	0.189
rating	14.77115	0.09647	153.124	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 182.3 on 397 degrees of freedom

Multiple R-squared: 0.9938, Adjusted R-squared: 0.9938

F-statistic: 3.18e+04 on 2 and 397 DF, p-value: < 2.2e-16

The p -value for 'rating' is very significant, while its not significant for 'income'

- Multi-collinearity issue with the 3D model
 1. Not idealistic that if we hold rating constant and income increases, by one unit then limit increase by 0.5573. → You can't change income by 1 without impacting rating

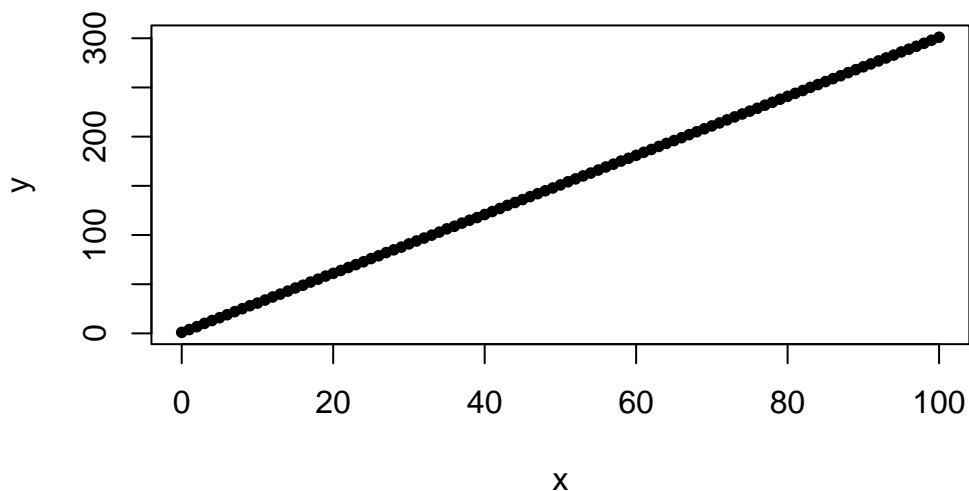
The impact of noise and β values on R^2

This first code chunk shows the model with very little noise and the output model summary to demonstrate the very high R^2 and p -value that occurs with very low noise

```
x <- seq(0,100,1)
b0 <- 1.0
b1 <- 3.0
y <- b0 + b1 * x + rnorm(100) * 0.1
```

Warning in `b0 + b1 * x + rnorm(100) * 0.1`: longer object length is not a multiple of shorter object length

```
plot(x, y, pch = 20)
```



```
model <- lm(y ~ x)
summary(model)
```

Call:

```
lm(formula = y ~ x)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.268338	-0.069057	-0.002244	0.063086	0.261756

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.0050066	0.0200146	50.21	<2e-16 ***
x	2.9997871	0.0003458	8674.94	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1013 on 99 degrees of freedom

Multiple R-squared: 1, Adjusted R-squared: 1

F-statistic: 7.525e+07 on 1 and 99 DF, p-value: < 2.2e-16

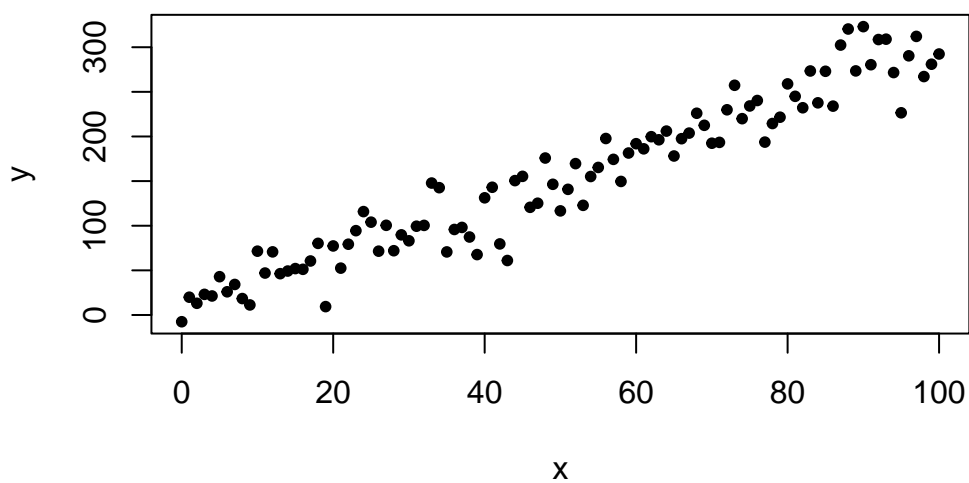
This next code chunk shows what happens with a much greater noise value. The intercept

p -value is much high than the previous model and the R^2 value decreases as well.

```
x <- seq(0,100,1)
b0 <- 1.0
b1 <- 3.0
y <- b0 + b1 * x + rnorm(100) * 20
```

Warning in `b0 + b1 * x + rnorm(100) * 20`: longer object length is not a multiple of shorter object length

```
plot(x, y, pch = 20)
```



```
model <- lm(y ~ x)
summary(model)
```

Call:

```
lm(formula = y ~ x)
```

Residuals:

Min	1Q	Median	3Q	Max
-71.490	-15.396	1.909	14.959	55.693

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.17627	4.75472	1.299	0.197
x	2.93812	0.08215	35.766	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.07 on 99 degrees of freedom

Multiple R-squared: 0.9282, Adjusted R-squared: 0.9274

F-statistic: 1279 on 1 and 99 DF, p-value: < 2.2e-16

To summarize:

- Lower noise increases the R^2 value
- If you make β_0 , β_1 , and noise values super small, then p-value will be super high and lead to a low R^2 value
- Can have high p-value with low R^2 value, but CANT have low p-value with high R^2 value

Multiple regression with categorical covariates

To demonstrate how to incorporate both categorical and quantitative covariates into a regression model, we will use the rating and marital status from the Credit dataset to try and predict the limit.

To create the models, you need to run 2 separate models using income, rating, and marital status to predict limit. 1. once for when student is yes 1. once for when student is no

- By adding the categorical variable, you are adding an additional intercept term

```
model <- lm(limit ~ rating + married, df)
summary(model)
```

Call:

```
lm(formula = limit ~ rating + married, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-----	----	--------	----	-----

-404.83 -126.56 20.86 131.80 456.19

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-528.088	25.204	-20.952	<2e-16 ***
rating	14.875	0.059	252.123	<2e-16 ***
marriedYes	-25.972	18.714	-1.388	0.166

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

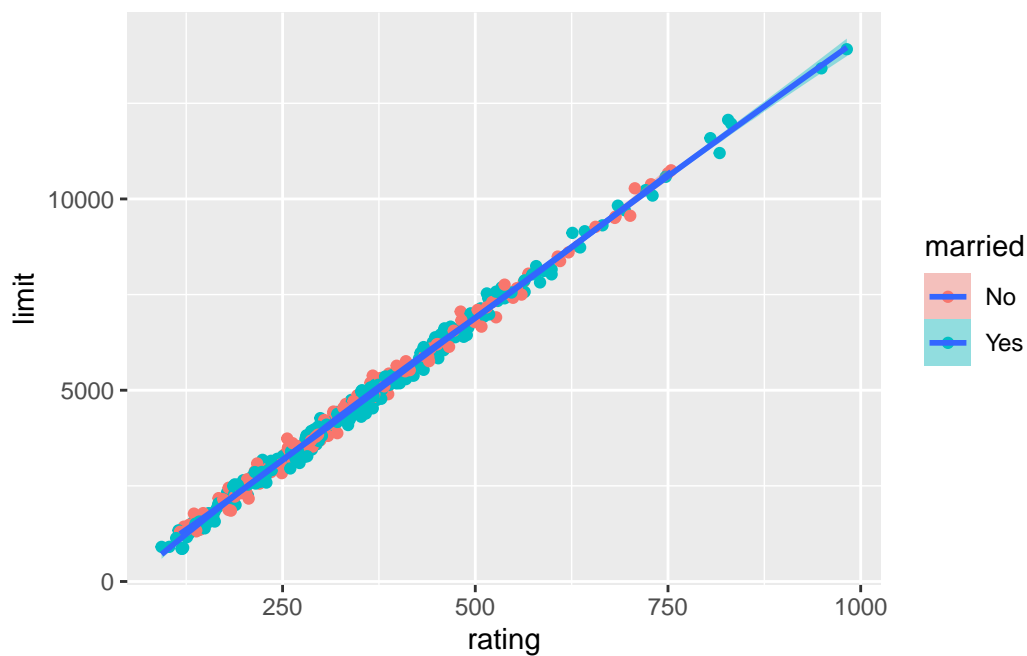
Residual standard error: 182.2 on 397 degrees of freedom

Multiple R-squared: 0.9938, Adjusted R-squared: 0.9938

F-statistic: 3.181e+04 on 2 and 397 DF, p-value: < 2.2e-16

```
ggplot(df) +  
  geom_point(aes(x=rating, y=limit, color=married)) +  
  geom_smooth(aes(x=rating, y=limit, fill=married))
```

`geom_smooth()` using method = 'loess' and formula = 'y ~ x'



The model above does have 2 different lines, they are just so close together, that you can't really see it, which implies that they have the same regression line. From the model summary, you can see that the p -value for marital status is very high, which indicates that it is not a good predictor of someones credit limit.