Chapter 2

Introduction to Multiple Linear Regression

Contents

2.1 Indian systolic blood pressure example 45

2.1.1 Taking Weight Into Consideration 50

2.1.2 Important Points to Notice About the Regression Output . 51

2.1.3 Understanding the Model 54

2.2 GCE exam score example 59

2.2.1 Some Comments on GCE Analysis 67

In **multiple linear regression**, a linear combination of two or more predictor variables (xs) is used to explain the variation in a response. In essence, the additional predictors are used to explain the variation in the response not explained by a simple linear regression fit.

2.1 Indian systolic blood pressure example

Anthropologists conducted a study[1] to determine the long-term effects of an environmental change on systolic blood pressure. They measured the blood pressure and several other characteristics of 39 Indians who migrated from a very primitive environment high in the Andes into the mainstream of Peruvian society at a lower altitude. All of the Indians were males at least 21 years of age, and were born at a high altitude.

```
library(tidyverse)

# load ada functions
source("ada_functions.R")

#### Example: Indian
dat_indian <-
  read_table("http://statacumen.com/teach/ADA2/notes/ADA2_notes_Ch02_indian.dat")
## Parsed with column specification:
## cols(
##  id = col_double(),
##  age = col_double(),
##  yrmig = col_double(),
##  wt = col_double(),
##  ht = col_double(),
##  chin = col_double(),
##  fore = col_double(),
##  calf = col_double(),
##  pulse = col_double(),
##  sysbp = col_double(),
##  diabp = col_double()
## )
# examine the structure of the dataset, is it what you expected?
  # a data.frame containing integers, numbers, and factors
str(dat_indian)
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 39 obs. of  11 variables:
##  $ id   : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ age  : num  21 22 24 24 25 27 28 28 31 32 ...
##  $ yrmig: num  1 6 5 1 1 19 5 25 6 13 ...
##  $ wt   : num  71 56.5 56 61 65 62 53 53 65 57 ...
##  $ ht   : num  1629 1569 1561 1619 1566 ...
##  $ chin : num  8 3.3 3.3 3.7 9 3 7.3 3.7 10.3 5.7 ...
##  $ fore : num  7 5 1.3 3 12.7 3.3 4.7 4.3 9 4 ...
##  $ calf : num  12.7 8 4.3 4.3 20.7 5.7 8 0 10 6 ...
```

[1] This problem is from the Minitab handbook.

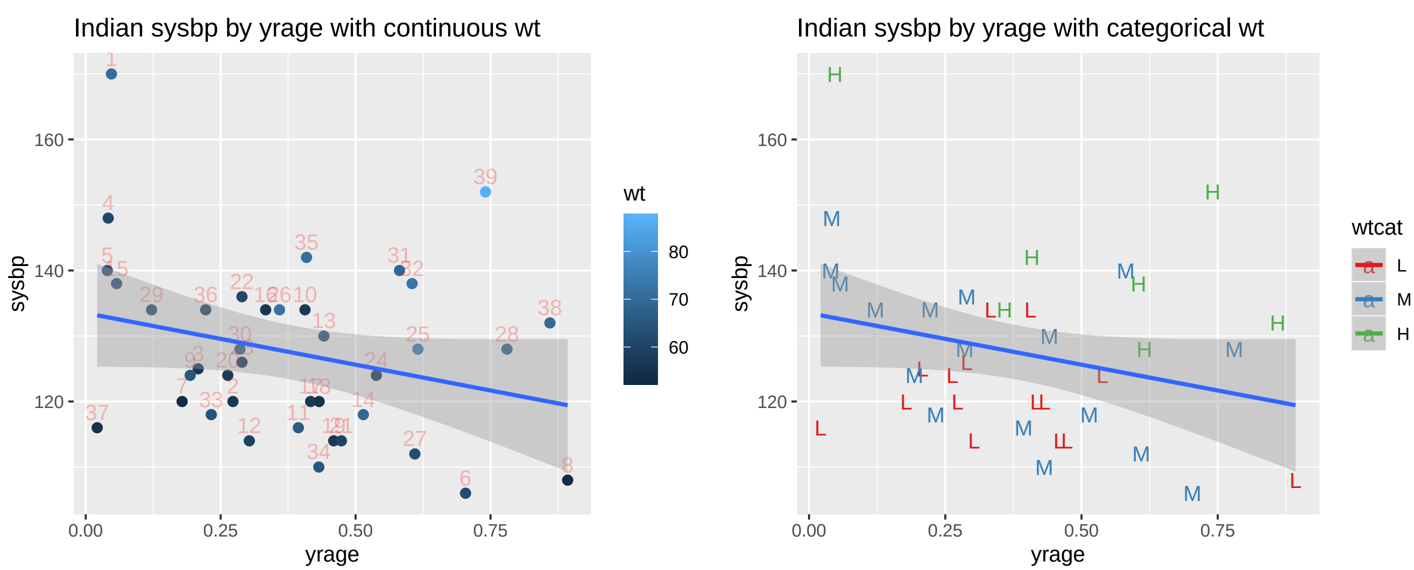
```
##  $ pulse: num  88 64 68 52 72 72 64 80 76 60 ...
##  $ sysbp: num  170 120 125 148 140 106 120 108 124 134 ...
##  $ diabp: num  76 60 75 120 78 72 76 62 70 64 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   id = col_double(),
##   ..   age = col_double(),
##   ..   yrmig = col_double(),
##   ..   wt = col_double(),
##   ..   ht = col_double(),
##   ..   chin = col_double(),
##   ..   fore = col_double(),
##   ..   calf = col_double(),
##   ..   pulse = col_double(),
##   ..   sysbp = col_double(),
##   ..   diabp = col_double()
##   .. )
# Description of variables
#    id = individual id
#   age = age in years           yrmig = years since migration
#    wt = weight in kilos           ht = height in mm
#  chin = chin skin fold in mm    fore = forearm skin fold in mm
#  calf = calf skin fold in mm   pulse = pulse rate-beats/min
# sysbp = systolic bp            diabp = diastolic bp

## print dataset to screen
#dat_indian
```

	id	age	yrmig	wt	ht	chin	fore	calf	pulse	sysbp	diabp
1	1	21	1	71.0	1629	8.0	7.0	12.7	88	170	76
2	2	22	6	56.5	1569	3.3	5.0	8.0	64	120	60
3	3	24	5	56.0	1561	3.3	1.3	4.3	68	125	75
4	4	24	1	61.0	1619	3.7	3.0	4.3	52	148	120
5	5	25	1	65.0	1566	9.0	12.7	20.7	72	140	78
6	6	27	19	62.0	1639	3.0	3.3	5.7	72	106	72
7	7	28	5	53.0	1494	7.3	4.7	8.0	64	120	76
8	8	28	25	53.0	1568	3.7	4.3	0.0	80	108	62
9	9	31	6	65.0	1540	10.3	9.0	10.0	76	124	70
10	10	32	13	57.0	1530	5.7	4.0	6.0	60	134	64
11	11	33	13	66.5	1622	6.0	5.7	8.3	68	116	76
12	12	33	10	59.1	1486	6.7	5.3	10.3	73	114	74
13	13	34	15	64.0	1578	3.3	5.3	7.0	88	130	80
14	14	35	18	69.5	1645	9.3	5.0	7.0	60	118	68
15	15	35	2	64.0	1648	3.0	3.7	6.7	60	138	78
16	16	36	12	56.5	1521	3.3	5.0	11.7	72	134	86
17	17	36	15	57.0	1547	3.0	3.0	6.0	84	120	70
18	18	37	16	55.0	1505	4.3	5.0	7.0	64	120	76
19	19	37	17	57.0	1473	6.0	5.3	11.7	72	114	80
20	20	38	10	58.0	1538	8.7	6.0	13.0	64	124	64
21	21	38	18	59.5	1513	5.3	4.0	7.7	80	114	66
22	22	38	11	61.0	1653	4.0	3.3	4.0	76	136	78
23	23	38	11	57.0	1566	3.0	3.0	3.0	60	126	72
24	24	39	21	57.5	1580	4.0	3.0	5.0	64	124	62
25	25	39	24	74.0	1647	7.3	6.3	15.7	64	128	84
26	26	39	14	72.0	1620	6.3	7.7	13.3	68	134	92
27	27	41	25	62.5	1637	6.0	5.3	8.0	76	112	80
28	28	41	32	68.0	1528	10.0	5.0	11.3	60	128	82
29	29	41	5	63.4	1647	5.3	4.3	13.7	76	134	92
30	30	42	12	68.0	1605	11.0	7.0	10.7	88	128	90
31	31	43	25	69.0	1625	5.0	3.0	6.0	72	140	72
32	32	43	26	73.0	1615	12.0	4.0	5.7	68	138	74
33	33	43	10	64.0	1640	5.7	3.0	7.0	60	118	66
34	34	44	19	65.0	1610	8.0	6.7	7.7	74	110	70
35	35	44	18	71.0	1572	3.0	4.7	4.3	72	142	84
36	36	45	10	60.2	1534	3.0	3.0	3.3	56	134	70
37	37	47	1	55.0	1536	3.0	3.0	4.0	64	116	54
38	38	50	43	70.0	1630	4.0	6.0	11.7	72	132	90
39	39	54	40	87.0	1542	11.3	11.7	11.3	92	152	88

A question we consider concerns the long term effects of an environmental change on the systolic blood pressure. In particular, is there a relationship between the systolic blood pressure and how long the Indians lived in their new environment as measured by the fraction of their life spent in the new environment.

```
dat_indian <-
  dat_indian %>%
  mutate(
    # Create the "fraction of their life" variable
    #   yrage = years since migration divided by age
    yrage = yrmig / age
  )

# continuous color for wt
# ggplot: Plot the data with linear regression fit and confidence bands
```

```
library(ggplot2)
p <- ggplot(dat_indian, aes(x = yrage, y = sysbp, label = id))
p <- p + geom_point(aes(colour=wt), size=2)
# plot labels next to points
p <- p + geom_text(hjust = 0.5, vjust = -0.5, alpha = 0.25, colour = 2)
# plot regression line and confidence band
p <- p + geom_smooth(method = lm)
p <- p + labs(title="Indian sysbp by yrage with continuous wt")
print(p)

# categorical color for wt
dat_indian <-
  dat_indian %>%
  mutate(
    wtcat = case_when(
      wt <  60                    ~ "L"   # low
    , (wt >=  60) & (wt < 70) ~ "M"   # medium
    , wt >= 70                    ~ "H"   # high
    )
  # define as a factor variable with a specific order
  , wtcat = factor(wtcat, levels=c("L", "M", "H"), ordered = TRUE)
  )

library(ggplot2)
p <- ggplot(dat_indian, aes(x = yrage, y = sysbp, colour = wtcat, label = wtcat))
p <- p + geom_text(size = 4)
  p <- p + scale_colour_brewer(palette="Set1")
# plot regression line and confidence band
p <- p + geom_smooth(method = lm, aes(group = 1))
p <- p + labs(title="Indian sysbp by yrage with categorical wt")
print(p)
```



Fit the simple linear regression model reporting the ANOVA table ("Terms") and parameter estimate table ("Coefficients").

```
# fit the simple linear regression model
lm_sysbp_yrage <- lm(sysbp ~ yrage, data = dat_indian)
# use Anova() from library(car) to get ANOVA table (Type 3 SS, df)
library(car)
Anova(lm_sysbp_yrage, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: sysbp
##             Sum Sq Df   F value  Pr(>F)
## (Intercept) 178221  1 1092.9484 < 2e-16 ***
## yrage          498  1    3.0544 0.08881 .
## Residuals     6033 37
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# use summary() to get t-tests of parameters (slope, intercept)
summary(lm_sysbp_yrage)
```

```
##
## Call:
## lm(formula = sysbp ~ yrage, data = dat_indian)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.161 -10.987  -1.014   6.851  37.254
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  133.496      4.038  33.060   <2e-16 ***
## yrage        -15.752      9.013  -1.748   0.0888 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.77 on 37 degrees of freedom
## Multiple R-squared:  0.07626,Adjusted R-squared:  0.05129
## F-statistic: 3.054 on 1 and 37 DF,  p-value: 0.08881
```

A plot above of systolic blood pressure against `yrage` fraction suggests a weak linear relationship. Nonetheless, consider fitting the regression model

$$\text{sysbp} = \beta_0 + \beta_1 \text{ yrage} + \varepsilon.$$

The least squares line (already in the plot) is given by

$$\widehat{\text{sysbp}} = 133.5 + -15.75 \text{ yrage},$$

and suggests that average systolic blood pressure decreases as the fraction of life spent in modern society increases. Note that a unit increase of this variable is an entire lifetime since it is a proportion, thus I'll interpret a 0.1 unit increase: for each 0.1 (10%) additional fraction of life spent at lower altitude, we expect a decrease of (also 10%) 1.575 millimeters of mercury (mmHg) of systolic blood pressure. However, the t-test of $H_0 : \beta_1 = 0$ is not significant at the 5% level (p-value=0.08881). That is, the weak linear relationship observed in the data is not atypical of a population where there is no linear relationship between systolic blood pressure and the fraction of life spent in a modern society.

Even if this test were significant, the small value of $R^2 = 0.07626$ suggests that `yrage` fraction does not explain a *substantial* amount of the variation in the systolic blood pressures. If we omit the individual with the highest blood pressure then the relationship would be weaker.

2.1.1 Taking Weight Into Consideration

At best, there is a weak relationship between systolic blood pressure and the `yrage` fraction. However, it is usually accepted that systolic blood pressure and weight are related. A natural way to take weight into consideration is to include `wt` (weight) and `yrage` fraction as predictors of systolic blood pressure in the multiple regression model:

$$\text{sysbp} = \beta_0 + \beta_1 \text{ yrage} + \beta_2 \text{ wt} + \varepsilon.$$

As in simple linear regression, the model is written in the form:

$$\text{Response} = \text{Mean of Response} + \text{Residual},$$

so the model implies that that average systolic blood pressure is a linear combination of `yrage` fraction and weight. As in simple linear regression, the standard multiple regression analysis assumes that the responses are normally distributed with a constant variance σ^2. The parameters of the regression model β_0, β_1, β_2, and σ^2 are estimated by least squares (LS).

Here is the multiple regression model with `yrage` and `wt` (weight) as predictors. Add `wt` to the right hand side of the previous formula statement.

```r
# fit the multiple linear regression model, (" + wt" added)
lm_sysbp_yrage_wt <- lm(sysbp ~ yrage + wt, data = dat_indian)
library(car)
Anova(lm_sysbp_yrage_wt, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: sysbp
##             Sum Sq Df F value    Pr(>F)
## (Intercept) 1738.2  1  18.183 0.0001385 ***
## yrage       1314.7  1  13.753 0.0006991 ***
## wt          2592.0  1  27.115 7.966e-06 ***
## Residuals   3441.4 36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(lm_sysbp_yrage_wt)
```

```
##
## Call:
## lm(formula = sysbp ~ yrage + wt, data = dat_indian)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4330  -7.3070   0.8963   5.7275  23.9819
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  60.8959    14.2809   4.264 0.000138 ***
## yrage       -26.7672     7.2178  -3.708 0.000699 ***
## wt            1.2169     0.2337   5.207 7.97e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.777 on 36 degrees of freedom
## Multiple R-squared:  0.4731,Adjusted R-squared:  0.4438
## F-statistic: 16.16 on 2 and 36 DF,  p-value: 9.795e-06
```

2.1.2 Important Points to Notice About the Regression Output

1. The LS estimates of the intercept and the regression coefficient for `yrage` fraction, and their standard errors, change from the simple linear model to the multiple regression model. For the simple linear regression:

$$\widehat{\text{sysbp}} = 133.50 - 15.75 \text{ yrage}.$$

For the multiple regression model:

$$\widehat{\text{sysbp}} = 60.89 - 26.76\ \text{yrage} + 1.21\ \text{wt}.$$

2. Looking at the ANOVA tables for the simple linear and the multiple regression models we see that the Regression (model) *df* has increased from 1 to 2 (2=number of predictor variables) and the Residual (error) *df* has decreased from 37 to 36 (=$n - 1-$ number of predictors). Adding a predictor increases the Regression *df* by 1 and decreases the Residual *df* by 1.
3. The Residual SS decreases by $6033.37 - 3441.36 = 2592.01$ upon adding the weight term. The Regression SS increased by 2592.01 upon adding the weight term term to the model. The Total SS does not depend on the number of predictors so it stays the same. The Residual SS, or the part of the variation in the response unexplained by the regression model, never increases when new predictors are added. (You can't add a predictor and explain less variation.)
4. The proportion of variation in the response explained by the regression model:

$$R^2 = \text{Regression SS}/\text{Total SS}$$

never decreases when new predictors are added to a model. The R^2 for the simple linear regression was 0.076, whereas

$$R^2 = 3090.08/6531.44 = 0.473$$

for the multiple regression model. Adding the weight variable to the model increases R^2 by 40%. That is, weight explains 40% of the variation in systolic blood pressure not already explained by fraction.
5. The estimated variability about the regression line

$$\text{Residual MS} = \hat{\sigma}^2$$

decreased dramatically after adding the weight effect. For the simple linear regression model $\hat{\sigma}^2 = 163.06$, whereas $\hat{\sigma}^2 = 95.59$ for the multiple regression model. This suggests that an important predictor has been added to model.

6. The F-statistic for the multiple regression model

$$F_{obs} = \text{Regression MS/Residual MS} = 1545.04/95.59 = 16.163$$

(which is compared to a F-table with 2 and 36 df) tests $H_0 : \beta_1 = \beta_2 = 0$ against H_A : not H_0. This is a test of no relationship between the average systolic blood pressure and fraction and weight, assuming the relationship is linear. If this test is significant, then either fraction or weight, or both, are important for explaining the variation in systolic blood pressure.

7. Given the model

$$\text{sysbp} = \beta_0 + \beta_1 \text{ yrage} + \beta_2 \text{ wt} + \varepsilon,$$

we are interested in testing $H_0 : \beta_2 = 0$ against $H_A : \beta_2 \neq 0$. The t-statistic for this test

$$t_{obs} = \frac{b_2 - 0}{\text{SE}(b_2)} = \frac{1.217}{0.234} = 5.207$$

is compared to a t-critical value with Residual $df = 36$. The test gives a p-value of < 0.0001, which suggests $\beta_2 \neq 0$. The t-test of $H_0 : \beta_2 = 0$ in the multiple regression model tests whether adding weight to the simple linear regression model explains a significant part of the variation in systolic blood pressure not explained by `yrage` fraction. In some sense, the t-test of $H_0 : \beta_1 = 0$ will be significant if the increase in R^2 (or decrease in Residual SS) obtained by adding weight to this simple linear regression model is substantial. We saw a big increase in R^2, which is deemed significant by the t-test. A similar interpretation is given to the t-test for $H_0 : \beta_1 = 0$.

8. The t-tests for $\beta_0 = 0$ and $\beta_1 = 0$ are conducted, assessed, and interpreted in the same manner. The p-value for testing $H_0 : \beta_0 = 0$ is 0.0001, whereas the p-value for testing $H_0 : \beta_1 = 0$ is 0.0007. This implies that

fraction is important in explaining the variation in systolic blood pressure **after** weight is taken into consideration (by including weight in the model as a predictor).

9. We compute CIs for the regression parameters in the usual way: $b_i + t_{\text{crit}} SE(b_i)$, where t_{crit} is the t-critical value for the corresponding CI level with df = Residual df.

2.1.3 Understanding the Model

The t-test for $H_0 : \beta_1 = 0$ is highly significant (p-value=0.0007), which implies that fraction is important in explaining the variation in systolic blood pressure **after weight is taken into consideration** (by including weight in the model as a predictor). Weight is called a **suppressor variable**. Ignoring weight suppresses the relationship between systolic blood pressure and `yrage` fraction.

The implications of this analysis are enormous! Essentially, the correlation between a predictor and a response says very little about the importance of the predictor in a regression model with one or more additional predictors. This conclusion also holds in situations where the correlation is high, in the sense that a predictor that is highly correlated with the response may be unimportant in a multiple regression model once other predictors are included in the model. In multiple regression *"everything depends on everything else."*

I will try to convince you that this was expected, given the plot of systolic blood pressure against fraction. This plot used a weight category variable `wtcat` L, M, or H as a plotting symbol. The relationship between systolic blood pressure and fraction is fairly linear within each weight category, and stronger than when we ignore weight. The slopes in the three groups are negative and roughly constant.

To see why `yrage` fraction is an important predictor after taking weight into consideration, let us return to the multiple regression model. The model implies that the average systolic blood pressure is a linear combination of `yrage` fraction and weight:

$$\widehat{\text{sysbp}} = \beta_0 + \beta_1 \text{ yrage} + \beta_2 \text{ wt}.$$

For each fixed weight, the average systolic blood pressure is linearly related to `yrage` fraction with a constant slope β_1, independent of weight. A similar interpretation holds if we switch the roles of `yrage` fraction and weight. That is, if we fix the value of fraction, then the average systolic blood pressure is linearly related to weight with a constant slope β_2, independent of `yrage` fraction.

To see this point, suppose that the LS estimates of the regression parameters are the true values

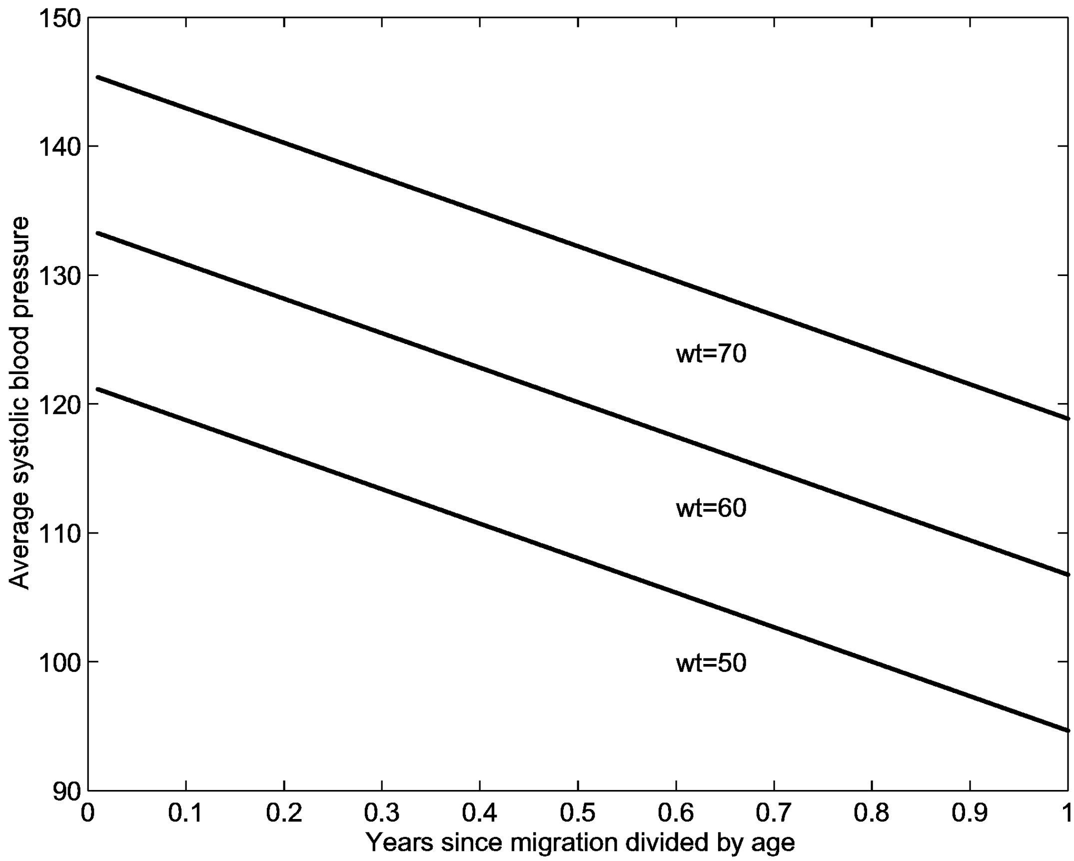
$$\widehat{\text{sysbp}} = 60.89 - 26.76 \text{ yrage} + 1.21 \text{ wt}.$$

If we restrict our attention to 50kg Indians, the average systolic blood pressure as a function of fraction is

$$\widehat{\text{sysbp}} = 60.89 - 26.76 \text{ yrage} + 1.21(50) = 121.39 - 26.76 \text{ yrage}.$$

For 60kg Indians,

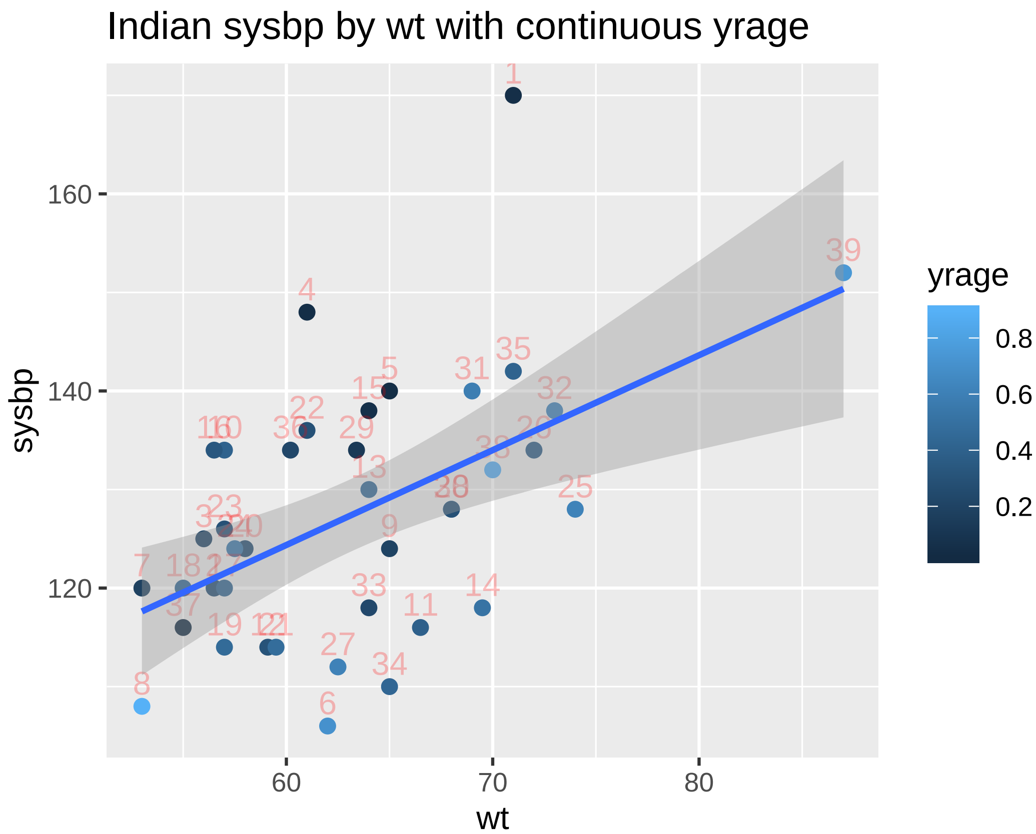
$$\widehat{\text{sysbp}} = 60.89 - 26.76 \text{ yrage} + 1.21(60) = 133.49 - 26.76 \text{ yrage}.$$



Hopefully the pattern is clear: the average systolic blood pressure decreases by 26.76 for each increase of 1 on fraction, regardless of one's weight. If we vary weight over its range of values, we get a set of parallel lines (i.e., equal slopes) when we plot average systolic blood pressure as a function of `yrage` fraction. The intercept increases by 1.21 for each increase of 1kg in weight.

Similarly, if we plot the average systolic blood pressure as a function of weight, for several fixed values of fraction, we see a set of parallel lines with slope 26.76, and intercepts decreasing by 26.76 for each increase of 1 in fraction.

```
# ggplot: Plot the data with linear regression fit and confidence bands
library(ggplot2)
p <- ggplot(dat_indian, aes(x = wt, y = sysbp, label = id))
p <- p + geom_point(aes(colour=yrage), size=2)
# plot labels next to points
p <- p + geom_text(hjust = 0.5, vjust = -0.5, alpha = 0.25, colour = 2)
# plot regression line and confidence band
p <- p + geom_smooth(method = lm)
p <- p + labs(title="Indian sysbp by wt with continuous yrage")
print(p)
```



If we had more data we could check the model by plotting systolic blood pressure against fraction, broken down by individual weights. The plot should show a fairly linear relationship between systolic blood pressure and fraction, with a constant slope across weights. I grouped the weights into categories because of the limited number of observations. The same phenomenon should approximately hold, and it does. If the slopes for the different weight groups changed drastically with weight, but the relationships were linear, we would need to include an **interaction** or product variable wt $\times$ yrage in the model, in addition to weight and `yrage` fraction. This is probably not warranted here.

A final issue that I wish to address concerns the interpretation of the estimates of the regression coefficients in a multiple regression model. For the fitted model

$$\widehat{\text{sysbp}} = 60.89 - 26.76 \text{ yrage} + 1.21 \text{ wt}$$

our interpretation is consistent with the explanation of the regression model given above. For example, focus on the `yrage` fraction coefficient. The negative coefficient indicates that the predicted systolic blood pressure decreases as `yrage` fraction increases **holding weight constant**. In particular, the predicted systolic blood pressure decreases by 26.76 for each unit increase in fraction, holding weight constant at any value. Similarly, the predicted systolic blood pressure increases by 1.21 for each unit increase in weight, holding `yrage` fraction constant at any level.

This example was meant to illustrate multiple regression. A more complete analysis of the data, including diagnostics, will be given later.

2.2 GCE exam score example

The data below are selected from a larger collection of data referring to candidates for the General Certificate of Education (GCE) who were being considered for a special award. Here, Y denotes the candidate's total mark, out of 1000, in the GCE exam, while X_1 is the candidate's score in the compulsory part of the exam, which has a maximum score of 200 of the 1000 points on the exam. X_2 denotes the candidates' score, out of 100, in a School Certificate English Language paper taken on a previous occasion.

```
#### Example: GCE
dat_gce <-
  read_table("http://statacumen.com/teach/ADA2/notes/ADA2_notes_Ch02_gce.dat")

## Parsed with column specification:
## cols(
##   y = col_double(),
##   x1 = col_double(),
##   x2 = col_double()
## )

str(dat_gce)

## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 15 obs. of  3 variables:
##  $ y : num  476 457 540 551 575 698 545 574 645 690 ...
##  $ x1: num  111 92 90 107 98 150 118 110 117 114 ...
##  $ x2: num  68 46 50 59 50 66 54 51 59 80 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   y = col_double(),
##   ..   x1 = col_double(),
##   ..   x2 = col_double()
##   .. )

## print dataset to screen
#dat_gce
```

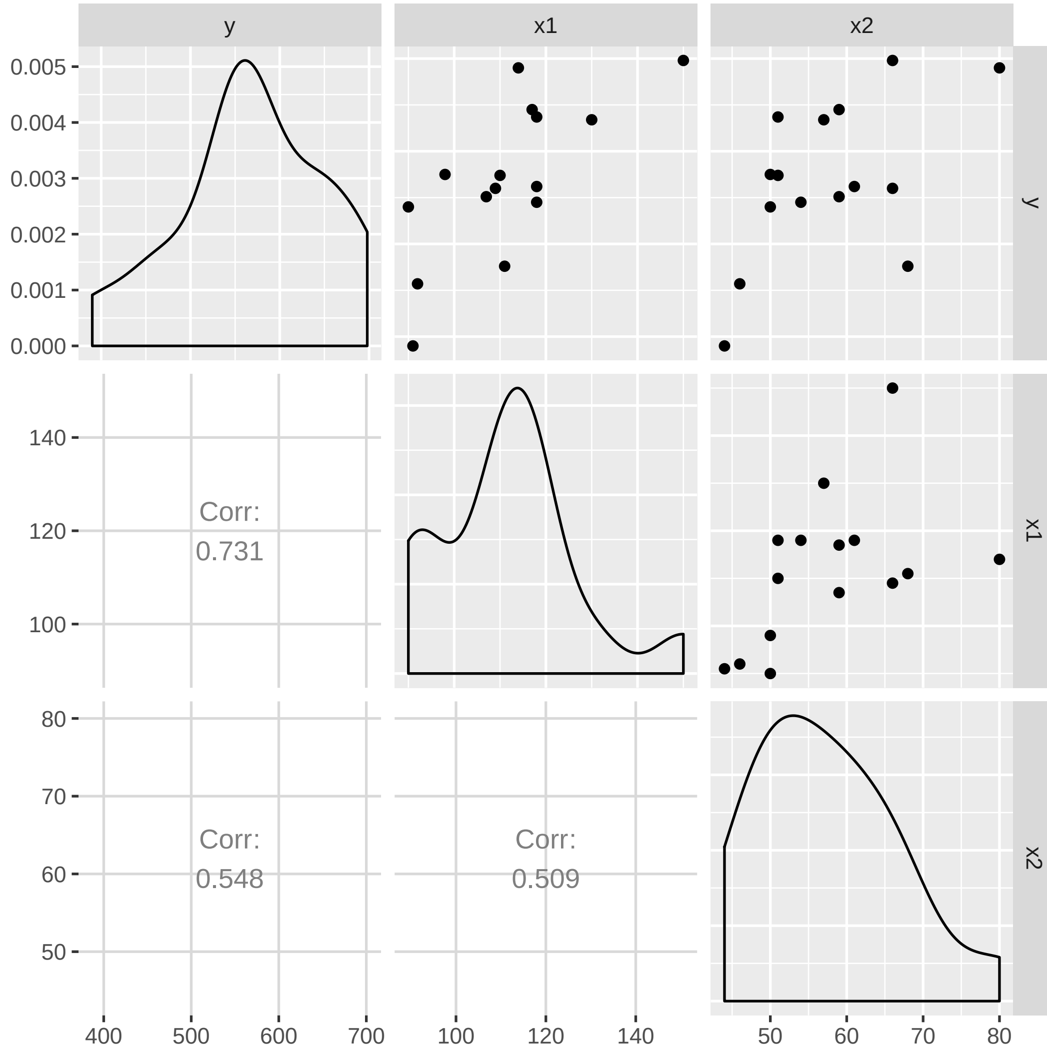
	y	x1	x2
1	476	111	68
2	457	92	46
3	540	90	50
4	551	107	59
5	575	98	50
6	698	150	66
7	545	118	54
8	574	110	51
9	645	117	59
10	690	114	80
11	634	130	57
12	637	118	51
13	390	91	44
14	562	118	61
15	560	109	66

A goal here is to compute a multiple regression of Y on X_1 and X_2, and make the necessary tests to enable you to comment intelligently on the extent to which current performance in the compulsory test (X_1) may be used to predict aggregate performance on the GCE exam (Y), and on whether previous performance in the School Certificate English Language (X_2) has any predictive value independently of what has already emerged from the current performance in the compulsory papers.

I will lead you through a number of steps to help you answer this question. Let us answer the following straightforward questions.

1. Plot Y against X_1 and X_2 individually, and comment on the form (i.e., linear, non-linear, logarithmic, etc.), strength, and direction of the relationships.
2. Plot X_1 against X_2 and comment on the form, strength, and direction of the relationship.
3. Compute the correlation between all pairs of variables. Do the correlations appear sensible, given the plots?

```
library(ggplot2)
library(GGally)
#p <- ggpairs(dat_gce, progress=FALSE)
# put scatterplots on top so y axis is vertical
p <- ggpairs( dat_gce
            , upper = list(continuous = "points")
            , lower = list(continuous = "cor")
            , progress=FALSE
            )
print(p)
```




```
# correlation matrix and associated p-values testing "H0: rho == 0"
#library(Hmisc)
dat_gce %>% as.matrix() %>% Hmisc::rcorr()

##       y   x1   x2
## y  1.00 0.73 0.55
## x1 0.73 1.00 0.51
## x2 0.55 0.51 1.00
##
## n= 15
##
##
## P
##    y      x1     x2
## y         0.0020 0.0346
## x1 0.0020        0.0527
## x2 0.0346 0.0527
```

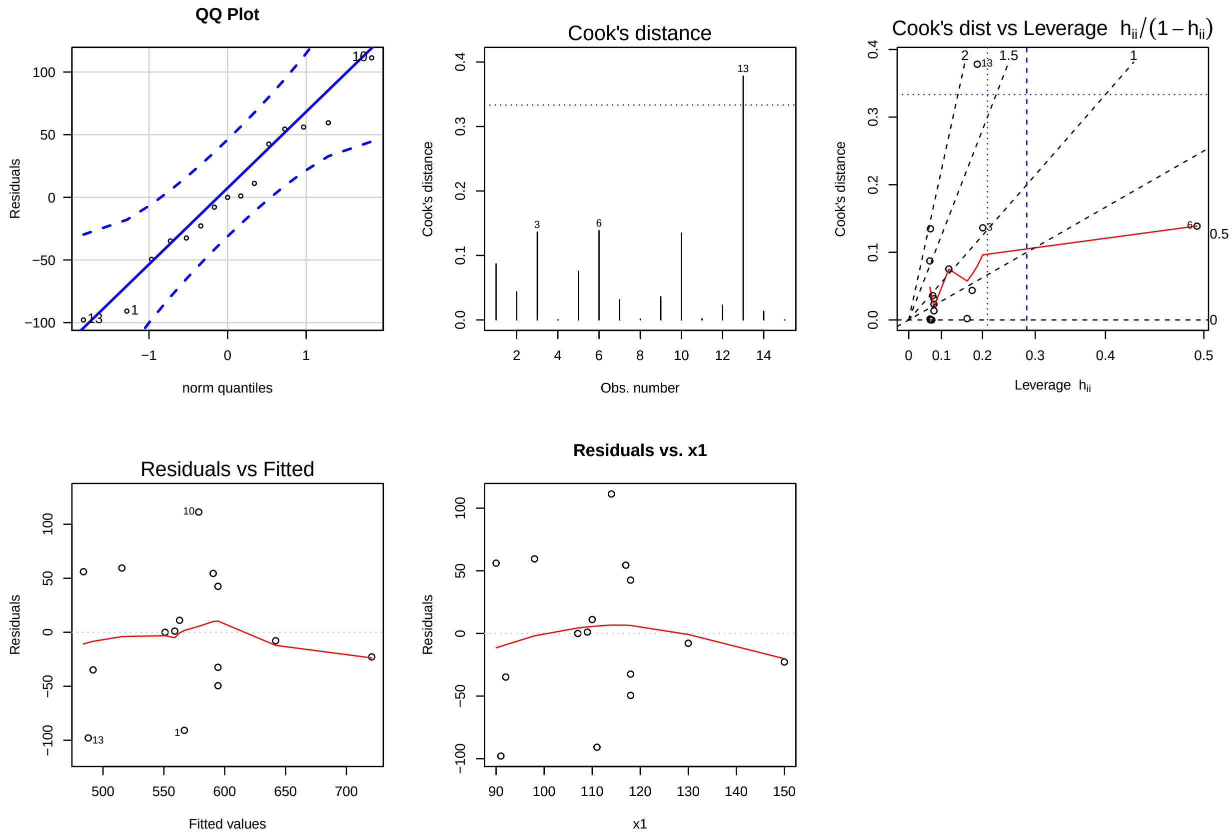
In parts 4 through 9, ignore the possibility that Y, X_1 or X_2 might ideally need to be transformed.

4. Which of X_1 and X_2 explains a larger proportion of the variation in Y? Which would appear to be a better predictor of Y? (Explain).
Model $Y = \beta_0 + \beta_1 X_1 + \varepsilon$:

```
# y ~ x1
lm_y_x1 <- lm(y ~ x1, data = dat_gce)
library(car)
Anova(lm_y_x1, type=3)

## Anova Table (Type III tests)
##
## Response: y
##             Sum Sq Df F value   Pr(>F)
## (Intercept)   4515  1   1.246 0.284523
## x1           53970  1  14.895 0.001972 **
## Residuals    47103 13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(lm_y_x1)

##
## Call:
## lm(formula = y ~ x1, data = dat_gce)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -97.858 -33.637  -0.034  48.507 111.327
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  128.548    115.160   1.116  0.28452
## x1             3.948      1.023   3.859  0.00197 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.19 on 13 degrees of freedom
## Multiple R-squared:  0.534,Adjusted R-squared:  0.4981
## F-statistic:  14.9 on 1 and 13 DF,  p-value: 0.001972
```

Plot diagnostics.

```
# plot diagnostics
lm_diag_plots(lm_y_x1, sw_plot_set = "simple")
```

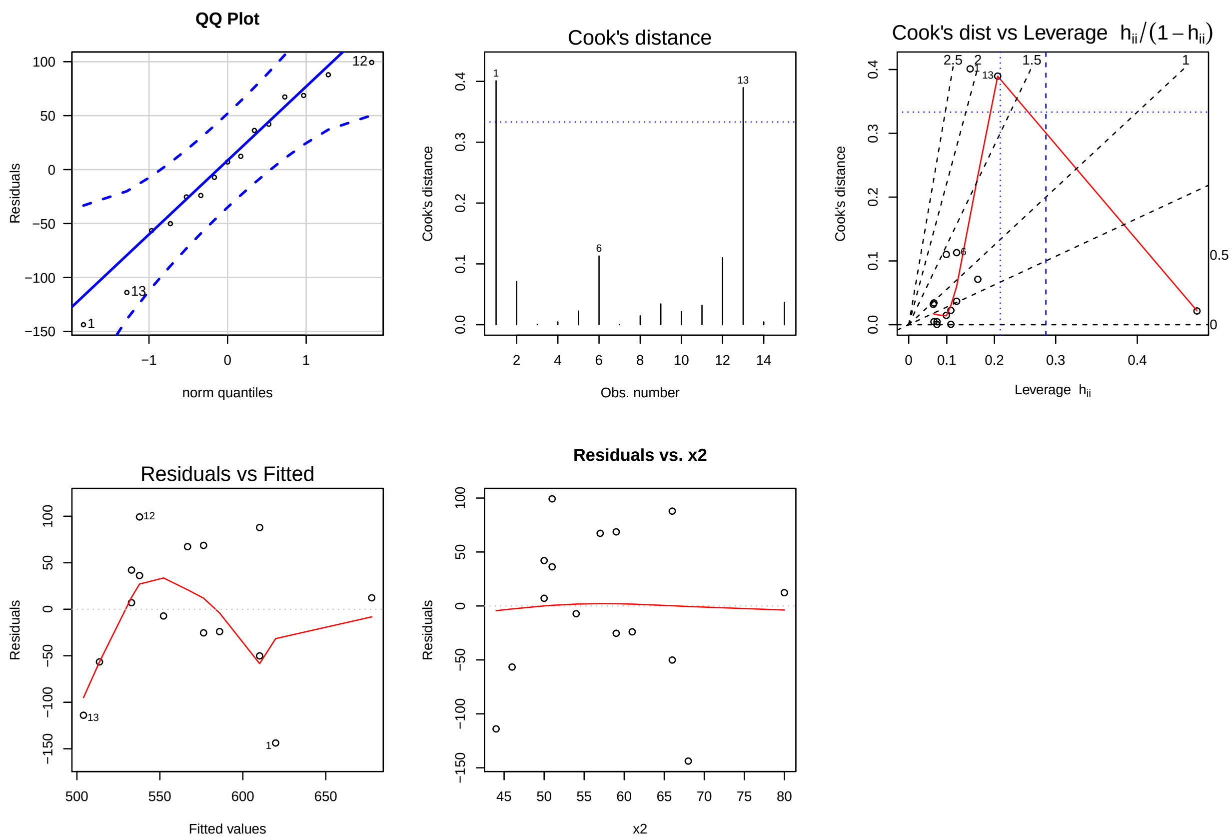


Model $Y = \beta_0 + \beta_1 X_2 + \varepsilon$:

```
# y ~ x2
lm_y_x2 <- lm(y ~ x2, data = dat_gce)
library(car)
Anova(lm_y_x2, type=3)

## Anova Table (Type III tests)
##
## Response: y
##              Sum Sq Df F value  Pr(>F)
## (Intercept)   32656  1  6.0001 0.02924 *
## x2            30321  1  5.5711 0.03455 *
## Residuals     70752 13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(lm_y_x2)

##
## Call:
## lm(formula = y ~ x2, data = dat_gce)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -143.770  -37.725    7.103   54.711   99.276
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  291.586    119.038    2.45   0.0292 *
## x2             4.826      2.045    2.36   0.0346 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 73.77 on 13 degrees of freedom
## Multiple R-squared:    0.3,Adjusted R-squared:  0.2461
## F-statistic: 5.571 on 1 and 13 DF,  p-value: 0.03455
```

```
# plot diagnostics
lm_diag_plots(lm_y_x2, sw_plot_set = "simple")
```



Answer: R^2 is 0.53 for the model with X_1 and 0.30 with X_2. Equivilantly, the Model SS is larger for X_1 (53970) than for X_2 (30321). Thus, X_1 appears to be a better predictor of Y than X_2.

5. Consider 2 simple linear regression models for predicting Y, one with X_1 as a predictor, and the other with X_2 as the predictor. Do X_1 and X_2 individually appear to be important for explaining the variation in Y? (i.e., test that the slopes of the regression lines are zero). Which, if any, of the output, support, or contradicts, your answer to the previous question?

Answer: The model with X_1 has a t-statistic of 3.86 with an associated p-value of 0.0020, while X_2 has a t-statistic of 2.36 with an associated p-value of 0.0346. Both predictors explain a significant amount of variability in Y. This is consistant with part (4).

6. Fit the multiple regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

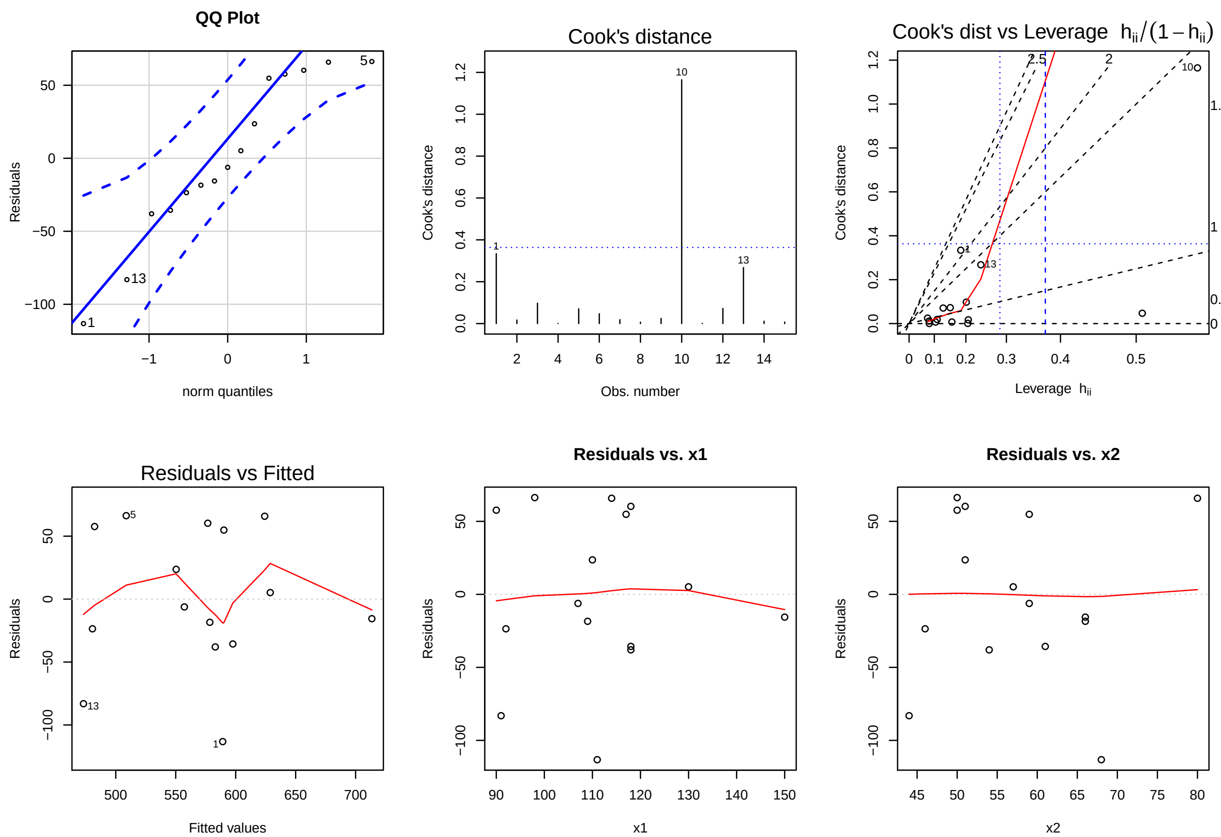
Test $H_0 : \beta_1 = \beta_2 = 0$ at the 5% level. Describe in words what this test is doing, and what the results mean here.
Model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$:

```
# y ~ x1 + x2
lm_y_x1_x2 <- lm(y ~ x1 + x2, data = dat_gce)
library(car)
Anova(lm_y_x1_x2, type=3)

## Anova Table (Type III tests)
##
## Response: y
##             Sum Sq Df F value  Pr(>F)
## (Intercept)   1571  1  0.4396 0.51983
## x1           27867  1  7.7976 0.01627 *
## x2            4218  1  1.1802 0.29866
## Residuals    42885 12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(lm_y_x1_x2)

##
## Call:
## lm(formula = y ~ x1 + x2, data = dat_gce)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -113.201  -29.605   -6.198   56.247   66.285
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   81.161    122.406   0.663   0.5198
## x1             3.296      1.180   2.792   0.0163 *
## x2             2.091      1.925   1.086   0.2987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.78 on 12 degrees of freedom
```

```
## Multiple R-squared:  0.5757,Adjusted R-squared:  0.505
## F-statistic: 8.141 on 2 and 12 DF,  p-value: 0.005835
```

Diagnostic plots suggest the residuals are roughly normal with no substantial outliers, though the Cook's distance is substantially larger for observation 10. We may wish to fit the model without observation 10 to see whether conclusions change.

```
# plot diagnistics
lm_diag_plots(lm_y_x1_x2, sw_plot_set = "simple")
```



Answer: The ANOVA table reports an F-statistic of 8.14 with associated p-value of 0.0058 indicating that the regression model with both X_1 and X_2 explains significantly more variability in Y than a model with the intercept, alone. That is, X_1 and X_2 explain variability in Y together. This does not tell us which of or whether X_1 or X_2 are individually important (recall the results of the Indian systolic blood pressure example).

7. In the multiple regression model, test $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$ individually. Describe in words what these tests are doing, and what the results mean here.

   Answer: Each hypothesis is testing, conditional on all other predictors

being in the model, whether the addition of the predictor being tested explains significantly more variability in Y than without it.
For $H_0 : \beta_1 = 0$, the t-statistic is 2.79 with an associated p-value of 0.0163. Thus, we reject H_0 in favor of the alternative that the slope is statistically significantly different from 0 conditional on X_2 being in the model. That is, X_1 explains significantly more variability in Y given that X_2 is already in the model.
For $H_0 : \beta_2 = 0$, the t-statistic is 1.09 with an associated p-value of 0.2987. Thus, we fail to reject H_0 concluding that there is insufficient evidence that the slope is different from 0 conditional on X_1 being in the model. That is, X_2 does not explain significantly more variability in Y given that X_1 is already in the model.

8. How does the R^2 from the multiple regression model compare to the R^2 from the individual simple linear regressions? Is what you are seeing here appear reasonable, given the tests on the individual coefficients?
Answer: The R^2 for the model with only X_1 is 0.5340, only X_2 is 0.3000, and both X_1 and X_2 is 0.5757. There is only a very small increase in R^2 from the model with only X_1 when X_2 is added, which is consistent with X_2 not being important given that X_1 is already in the model.
9. Do your best to answer the question posed above, in the paragraph after the data "A goal . . . ". Provide an equation (LS) for predicting Y.
Answer: Yes, we've seen that X_1 may be used to predict Y, and that X_2 does not explain significantly more variability in the model with X_1. Thus, the preferred model has only X_1:

$$\hat{y} = 128.55 + 3.95X_1.$$

2.2.1 Some Comments on GCE Analysis

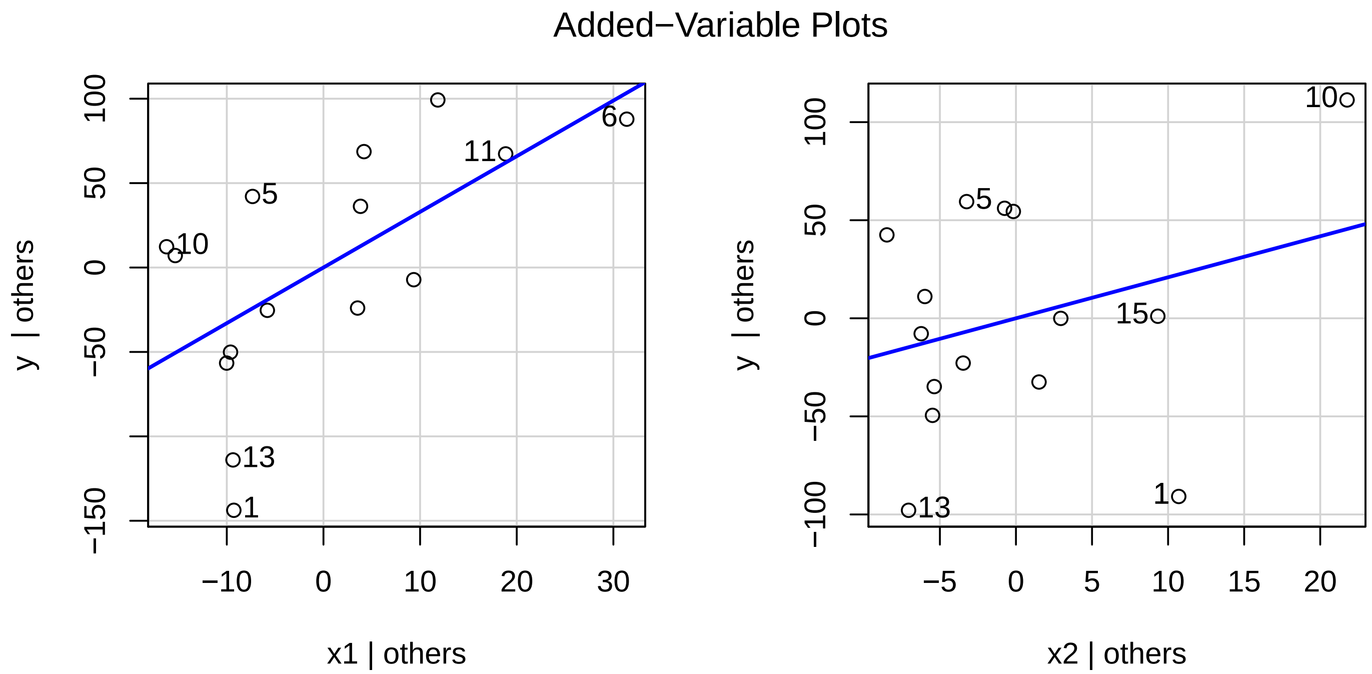
I will give you my thoughts on these data, and how I would attack this problem, keeping the ultimate goal in mind. I will examine whether transformations of the data are appropriate, and whether any important conclusions are dramatically influenced by individual observations. I will use some new tools to attack this problem, and will outline how they are used.

The plot of GCE (Y) against COMP (X_1) is fairly linear, but the trend in the plot of GCE (Y) against SCEL (X_2) is less clear. You might see a non-linear trend here, but the relationship is not very strong. When I assess plots I try to not allow a few observations affect my perception of trend, and with this in mind, I do not see any strong evidence at this point to transform any of the variables.

One difficulty that we must face when building a multiple regression model is that these two-dimensional (2D) plots of a response against individual predictors may have little information about the appropriate scales for a multiple regression analysis. In particular, the 2D plots only tell us whether we need to transform the data in a simple linear regression analysis. If a 2D plot shows a strong non-linear trend, I would do an analysis using the suggested transformations, including any other effects that are important. However, it might be that no variables need to be transformed in the multiple regression model.

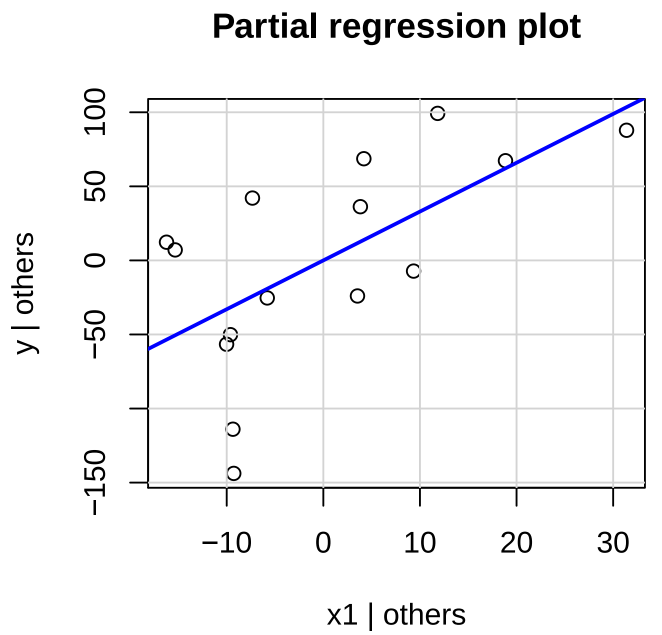
The **partial regression residual plot**, or added variable plot, is a graphical tool that provides information about the need for transformations in a multiple regression model. The following **reg** procedure generates diagnostics and the partial residual plots for each predictor in the multiple regression model that has COMP and SCEL as predictors of GCE.

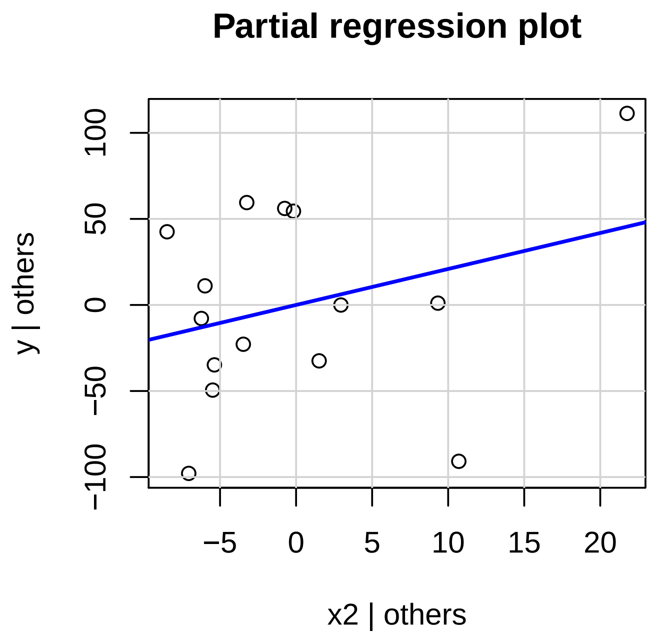
```
library(car)
avPlots(lm_y_x1_x2, id = list(n = 3))
```



The partial regression residual plot (added-variables plot) compares the residuals from two model fits. First, we "adjust" Y for all the other predictors in the model except the selected one. Then, we "adjust" the selected variable X_{sel} for all the other predictors in the model. Lastly, plot the residuals from these two models against each other to see what relationship still exists between Y and X_{sel} after accounting for their relationships with the other predictors.

```
# function to create partial regression plot
partial_regression_plot <- function (y, x, sel, ...) {
  m <- as.matrix(x[, -sel])
  # residuals of y regressed on all x's except "sel"
  y1 <- lm(y ~ m)$res
  # residuals of x regressed on all other x's
  x1 <- lm(x[, sel] ~ m)$res
  # plot residuals of y vs residuals of x
  plot( y1 ~ x1, main="Partial regression plot", ylab="y | others", ...)
  # add grid
  grid(lty = "solid")
  # add red regression line
  abline(lm(y1 ~ x1), col = "blue", lwd = 2)
}

par(mfrow=c(1, 2))
partial_regression_plot(dat_gce$y, cbind(dat_gce$x1, dat_gce$x2), 1, xlab="x1 | others")
partial_regression_plot(dat_gce$y, cbind(dat_gce$x1, dat_gce$x2), 2, xlab="x2 | others")
```





The first partial regression residual plot for COMP, given below, "**adjusts**" GCE (Y) and COMP (X_1) for their common dependence on all the other predictors in the model (only SCEL (X_2) here). This plot tells us whether we need to transform COMP in the multiple regression model, and whether any observations are influencing the significance of COMP in the fitted model. A roughly linear trend suggests that no transformation of COMP is warranted. The positive relationship seen here is consistent with the coefficient of COMP being positive in the multiple regression model. The partial residual plot for COMP shows little evidence of curvilinearity, and much less so than the original 2D plot of GCE against COMP. This indicates that there is no strong evidence for transforming COMP in a multiple regression model that includes SCEL.

Although SCEL appears to somewhat useful as a predictor of GCE on it's own, the multiple regression output indicates that SCEL does not explain a significant amount of the variation in GCE, once the effect of COMP has been taken into account. Put another way, previous performance in the School Certificate English Language (X_2) has little predictive value independently of what has already emerged from the current performance in the compulsory papers (X_1 or COMP). This conclusion is consistent with the fairly weak linear relationship between GCE against SCEL seen in the second partial residual plot.

Do diagnostics suggest any deficiencies associated with this conclusion? The partial residual plot of SCEL highlights observation 10, which has the largest value of Cook's distance in the multiple regression model. If we visually hold observation 10 out from this partial residual plot, it would appear that the relationship observed in this plot would weaken. This suggests that observation 10 is actually **enhancing** the significance of SCEL in the multiple regression model. That is, the p-value for testing the importance of SCEL in the multiple regression model would be inflated by holding out observation 10. The following output confirms this conjecture. The studentized residuals, Cook's distances and partial residual plots show no serious deficiencies.

Model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, excluding observation 10:

```
dat_gce_10 <-
  dat_gce %>%
  slice(
    -10
  )
# y ~ x1 + x2
lm_y10_x1_x2 <- lm(y ~ x1 + x2, data = dat_gce_10)
library(car)
Anova(lm_y10_x1_x2, type=3)
```

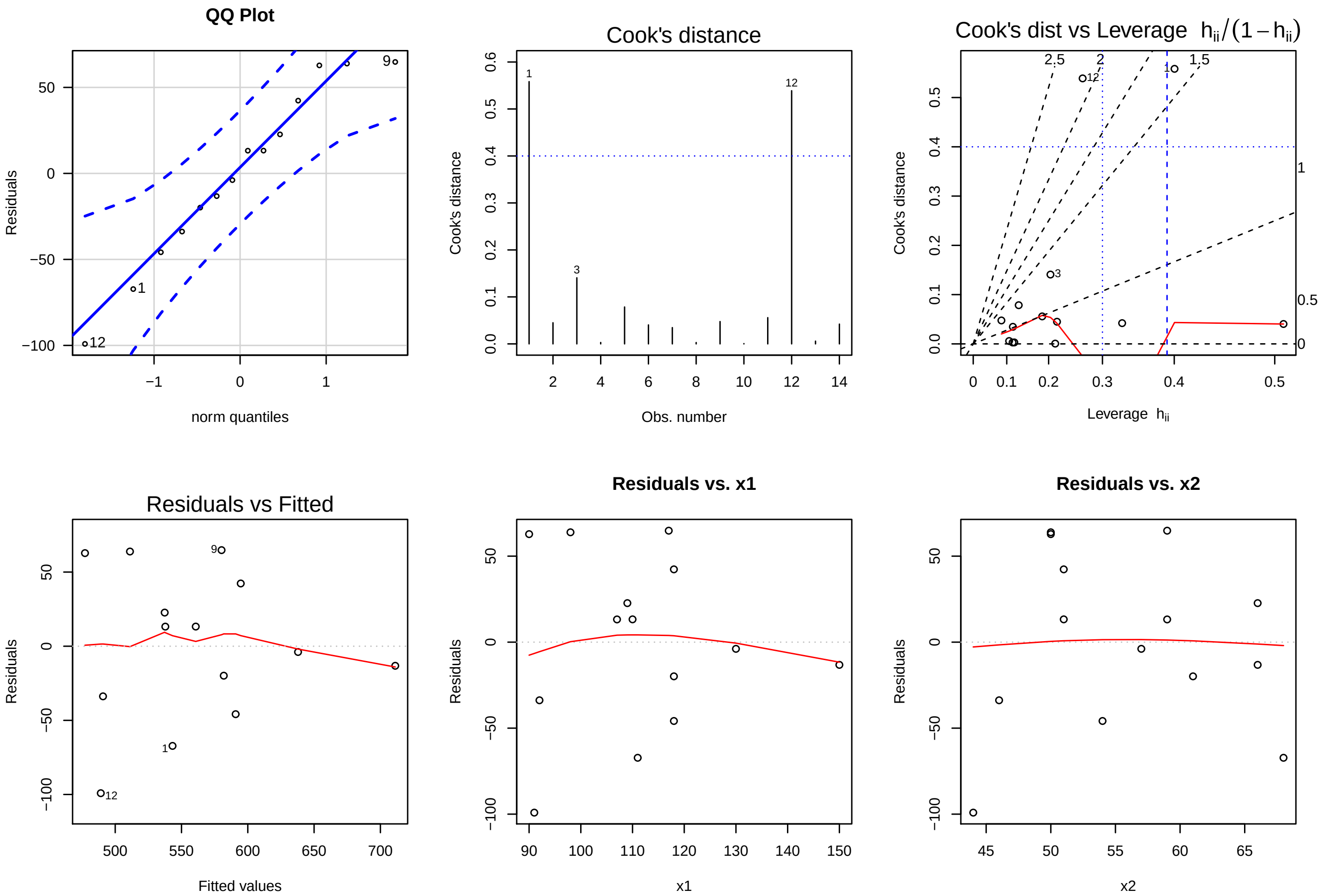
```
## Anova Table (Type III tests)
##
## Response: y
##              Sum Sq Df F value   Pr(>F)
## (Intercept)   5280  1  1.7572 0.211849
## x1           37421  1 12.4540 0.004723 **
## x2             747  1  0.2486 0.627870
## Residuals    33052 11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

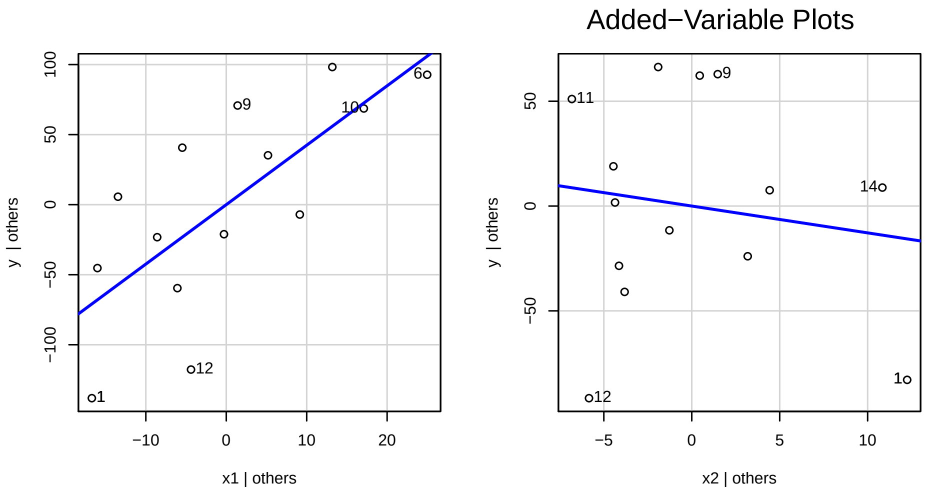
```
summary(lm_y10_x1_x2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = dat_gce_10)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.117 -30.319   4.661  37.416  64.803
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  159.461    120.295   1.326  0.21185
```

```
## x1               4.241      1.202   3.529  0.00472 **
## x2              -1.280      2.566  -0.499  0.62787
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.82 on 11 degrees of freedom
## Multiple R-squared:  0.6128,Adjusted R-squared:  0.5424
## F-statistic: 8.706 on 2 and 11 DF,  p-value: 0.005413
```

```
# plot diagnostics
lm_diag_plots(lm_y10_x1_x2, sw_plot_set = "simpleAV")
```





What are my conclusions? It would appear that SCEL (X_2) is not a useful predictor in the multiple regression model. For simplicity, I would likely use a simple linear regression model to predict GCE (Y) from COMP (X_1) only. The diagnostic analysis of the model showed no serious deficiencies.