# STAT 504 - Homework 3

**Due date:** Thursday, February 14th. Submit your homework solutions to the course Canvas page. Please submit the output and plots, but not your R code unless the quesiton specifically asks for it. Total possible points: 24 points.

1. (4 + 2 Bonus points) Consider the following linear model for the observations $i = 1, \ldots, n$

$$y_i = \alpha + \beta x_i^{(1)} + \sum_{j=2}^{K} \gamma_j x_i^{(j)} + \sum_{j=2}^{K} \delta_j x_i^{(1)} x_i^{(j)} + \epsilon_i.$$

The model contains continuous and categorical independent variables. The continuous variable is denoted by $x^{(1)}$ and the categorical variable represents $K$ categories:

$$x^{(j)} = \begin{cases} 1 & \text{if category} = j \\ 0 & \text{else} \end{cases}$$

for $j = 2, \ldots, K$. The errors $\epsilon_i$ are assumed to be i.i.d. $\sim \mathcal{N}(0, \sigma^2)$.

(a) (1 + 1 Bonus points) Give an interpretation of the model and each of the coefficients. (Bonus 1 Point:) What are the differences to conducting a simple linear regression for each of the $K$ levels of the categorical variable?

**Solution:**

In the full, general linear model, $K$ regression lines are fitted. The parameters $\alpha$ and $\beta$ are intercept and slope for the observations of category 1. The parameters $\gamma_i$ describe for any category $j = 2, \ldots, K$ how much the intercept is changed with respect to category 1. The parameters $\delta_j$ describe how much the slope is changed with respect to category 1. **(1 Point.)**

**Bonus:** The regression slopes of the full, general model would be identical as the ones from the $K$ simple regression models for each of the categories separately. Tests, however, that compare the slopes and intercepts of the categories are only possible in the full, general model. Moreover, the estimation of the error variance is possibly more precise in the full model as we share the information across the $K$ categories. Of course, a crucial assumption for the latter is that the errors have the same variance across the different categories. This gain in precision (or to be more precise, in degrees of freedom) can be also useful for testing the coefficients in the full, general model as opposed to the separate ones. **(2 Points. Errors have the same variance, fewer samples)**

(b) (3 points) Give a mathematical formulation for the null hypothesis "the effect of an increase of $x^{(1)}$ by one unit is the same for all categories". Specify a suitable test statistic and the probability distribution that this test statistic follows under $H_0$.

**Solution:** The null hypothesis can be formulated as: "The slopes of the regression lines are the same for all categories." This means $\delta_2 = \delta_3 = \ldots = \delta_K = 0$. **(1 Point.)** Since the models are nested, the F-test is a suitable test. If the null hypothesis is true, we have the F statistic

$$F = \frac{(RSS_K - RSS_{2K-1})/(K-1)}{RSS_{2K-1}/(n-2K)}.$$

The quantities $RSS_K$ and $RSS_{2K-1}$ are the resudial sums of squares of the reduced model (i.e. $\delta_j = 0$) and of the full model (i.e. $\delta_j \neq 0$) **(1 Point.)**. The F statistic follows an $F_{(K-1),(n-2K)}$ under $H_0$. **(1 Point.)**

2. (12 points) Two runners: (Marcel and Dani) are put under a cardiac stress test (Conconi test) which involves running on a treadmil. The test is conducted as follows:

- The athlete warms up for 10 minutes.

- The assistant sets the treadmill speed to the runners desired start speed.
- The assistant records the heart rate of the runner every 200 metres (.125 miles),
- The assistant increases the treadmill speed every 200 metres by 0.5km/hr. (0.31mph)
- The assistant stops the stopwatch when the athlete is unable to continue.

(a) (1 point) We first need to preprocess the data. Create a data frame that contains all (non NA) observations of the variables `pulse`, `speed`, and `runner`. The `pulse` is the response and `speed` and `runner` are predictors, where `runner` should be a categorical predictor with the levels "Dani" and "Marcel" (0 and 1). Hint: There should be 39 samples.

**Print your processed data frame.**

**Solution:**

```
> ## load data
> conconi <- readRDS("runners.RDS")
> ## preprocess
> speed  <- conconi$Speed[c(1:19,7:26)]
> pulse   <- c(conconi$Marcel.Puls[1:19], conconi$Dani.Puls[7:26])
> runner <- factor(c(rep("Marcel",19), rep("Dani",20)))
> conconi2     <- data.frame(pulse, speed, runner)
> conconi2
```

```
   pulse speed runner
1    145   9.0 Marcel
2    148   9.5 Marcel
3    152  10.0 Marcel
4    156  10.5 Marcel
5    156  11.0 Marcel
6    163  11.5 Marcel
7    159  12.0 Marcel
8    166  12.5 Marcel
9    166  13.0 Marcel
10   170  13.5 Marcel
11   177  14.0 Marcel
12   180  14.5 Marcel
13   184  15.0 Marcel
14   187  15.5 Marcel
15   190  16.0 Marcel
16   196  16.5 Marcel
17   194  17.0 Marcel
18   199  17.5 Marcel
19   201  18.0 Marcel
20   130  12.0   Dani
21   136  12.5   Dani
22   138  13.0   Dani
23   138  13.5   Dani
24   141  14.0   Dani
25   145  14.5   Dani
26   148  15.0   Dani
27   149  15.5   Dani
28   150  16.0   Dani
29   153  16.5   Dani
30   154  17.0   Dani
31   155  17.5   Dani
32   158  18.0   Dani
33   161  18.5   Dani
34   162  19.0   Dani
35   163  19.5   Dani
36   166  20.0   Dani
```

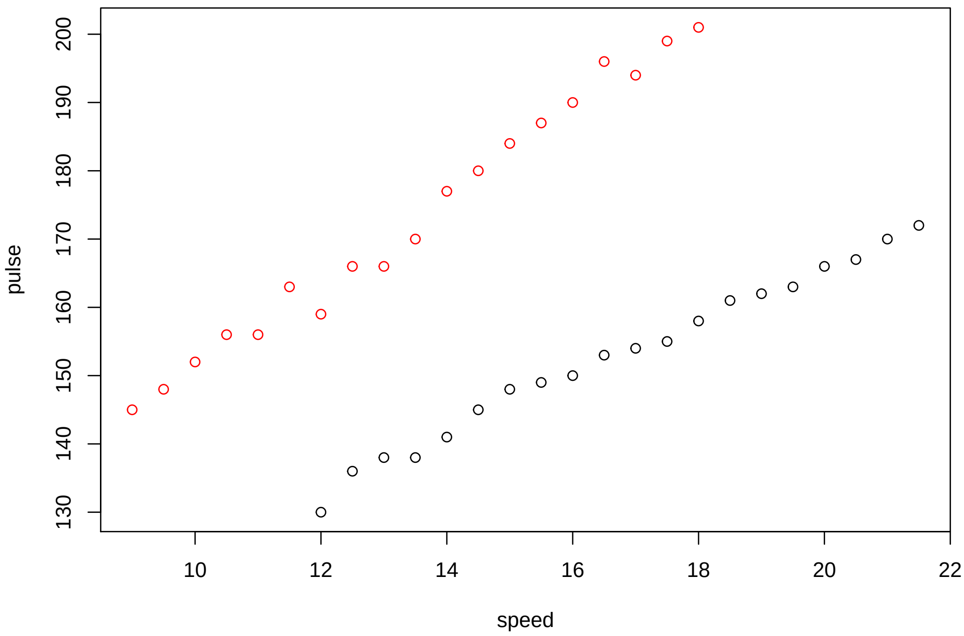
```
37    167   20.5    Dani
38    170   21.0    Dani
39    172   21.5    Dani
```

**(1 Point.)**

(b) (1 point) Print the scatter plot of `pulse` vs. `speed` with different colored points indicating each of the runners. Which model do you think is reasonable in this case?

**Solution:**

```
> plot(pulse~speed, col=runner, data=conconi2)
```




**(0.5 Points.)**

A linear model seems reasonable in this case. There is possibly an interaction between speed and runner, but the strength of the possible interaction cannot be judged from the plot. **(0.5 Points.)**

(c) (2 points) Now fit an OLS regression model: `pulse` $\sim$ `speed + runner`. What does this model assume with respect to the average starting pulse of each runner? What does it assume about the average increase in pulse for a 1 km/hr increase in speed for each of the two runners?

**Solution:**

```
> ## perform regression
> fit1  <- lm(pulse ~ speed + runner, data=conconi2)
> summary(fit1)

Call:
lm(formula = pulse ~ speed + runner, data = conconi2)

Residuals:
   Min     1Q Median     3Q    Max
-6.364 -3.340  0.217  2.992  7.411

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   66.3510     3.7310   17.78   <2e-16 ***
speed          5.1611     0.2169   23.80   <2e-16 ***
```

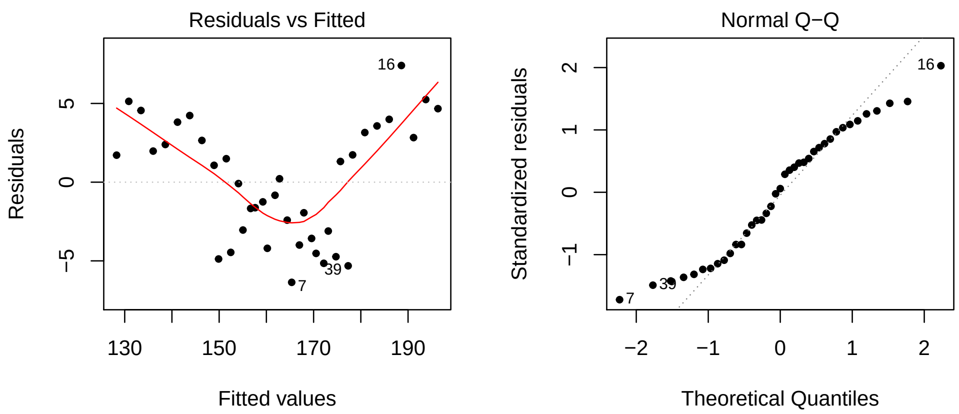
```
runnerMarcel  37.0789     1.4096   26.30   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.811 on 36 degrees of freedom
Multiple R-squared:  0.959,	Adjusted R-squared:  0.9568
F-statistic: 421.5 on 2 and 36 DF,  p-value: < 2.2e-16
```

The main effects model assumes that the average increase in pulse of both runners for a 1 km/hr increase in speed is identical (the slope), while the model allows that the average starting pulse differs between these two runners (the intercept). **(2 Points.)**

(d) (2 points) Perform a residual analysis by plotting the "residuals vs. fitted" plot and the Normal QQ plot. Which model violations can we detect? **State all the assumptions you can check with these plots and whether you think they are satisfied.**

**Solution:**

```
> ## residual analysis
> par(mfrow=c(1,2))
> plot(fit1, which=1:2, pch=20)
```




We observe a large systematic error (curvature of the mean). Possibly some non-constant variance of the residuals. **(1 Point.)** In addition, the distribution of the residuals appears to be short-tailed in comparison to a Normal distribution (so the normality assumption is likely violated). **(1 Point.)** These model violations could be due to the absence of the interaction term from the fitted model.

(e) (2 points) Now, fit a model with an interaction term between `speed` and `runner`. What does this model assume with respect to the average starting pulse of each runner? What does it assume about the average increase in pulse for a 1 km/hr increase in speed for each of the two runners?

**Solution:**

```
> ## new model
> fit2 <- lm(pulse ~ speed + runner + speed:runner, data=conconi2)
> summary(fit2)

Call:
lm(formula = pulse ~ speed + runner + speed:runner, data = conconi2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4947 -0.9034  0.2667  1.0588  3.6737

Coefficients:
```

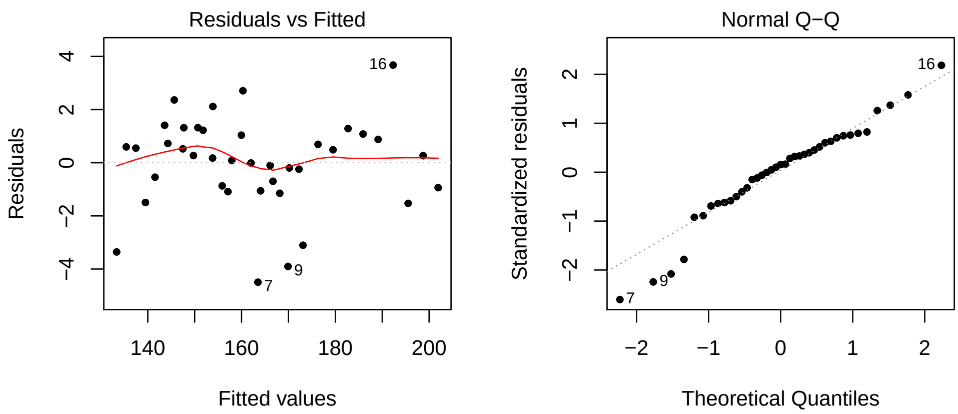
```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          84.2383     2.3574  35.734  < 2e-16 ***
speed                 4.0932     0.1387  29.512  < 2e-16 ***
runnerMarcel          2.3722     3.1330   0.757    0.454
speed:runnerMarcel    2.3138     0.2042  11.333 2.91e-13 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.788 on 35 degrees of freedom
Multiple R-squared:  0.9912,       Adjusted R-squared:  0.9905
F-statistic:  1319 on 3 and 35 DF,  p-value: < 2.2e-16
```

This model allows for both different average starting pulses, as well as different average increase in pulse for each of the runners, i.e. two regression lines with different intercepts and different slopes are fitted. **(2 Points.)**

(f) (2 points) Perform a residual analysis (TA plot and Normal QQ plot) and discuss the model assumptions. **State all the assumptions you can check with these plots and whether you think they are satisfied.**

**Solution:**

```
> ## residual analysis
> par(mfrow=c(1,2))
> plot(fit2, which=1:2, pch=20)
```




The TA-plot does not indicate any model assumption violation (no curvatur of the mean or non-constant variance violation). **(1 Point.)** There are a few outliers in the Normal plot, i.e. observations with large negative residuals. These deviations could be due to some measurement error in the test or the small sample size (it is unclear whether the normality assumption is violated). **(1 Point.)**

(g) (2 points) Using the full model (with interaction), compute the estimates of the average initial pulse (i.e. when `speed=0`) for each runner, as well as the estimates of the average pulse increase with every additional 1 km/hr in speed (for each runner).

**Solution:**

```
> summary(fit2)

Call:
lm(formula = pulse ~ speed + runner + speed:runner, data = conconi2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4947 -0.9034  0.2667  1.0588  3.6737
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          84.2383     2.3574  35.734  < 2e-16 ***
speed                 4.0932     0.1387  29.512  < 2e-16 ***
runnerMarcel          2.3722     3.1330   0.757    0.454
speed:runnerMarcel    2.3138     0.2042  11.333 2.91e-13 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.788 on 35 degrees of freedom
Multiple R-squared:  0.9912,	Adjusted R-squared:  0.9905
F-statistic:  1319 on 3 and 35 DF,  p-value: < 2.2e-16
```

The initial pulse of Dani corresponds to the intercept, i.e. 84.238. For Marcel, the coefficient $\hat{\beta}_2 = 2.372$ needs to be added, so his initial pulse is 86.61. **(1 Point.)**

For Dani, the average pulse increase is 4.093 beats with every additional km/hr in speed (coefficient $\hat{\beta}_1$). For Marcel, with every 1 km/hr increase in speed the average pulse increase is estimated as $\hat{\beta}_1 + \hat{\beta}_3 + \hat{\beta}_4 \approx 4.093 + 2.314 + 2.372 = 8.779$. **(1 Point.)**

3. (8 points) The Australian Bureau of Agricultural and Resource Economics conducts an annual survey of the agroindustry. In 1991, 451 farms in New South Wales took part. The raw data is contained in the file `farm.RDS` available on Canvas. The variables have the following meanings:

`revenue` : target variable, total revenue of the farm
`costs` : predictor, total costs of the farm
`region` : predictor, code for different regions within New South Wales
`industry` : predictor, code for the cultivation (1=(wheat), 2= (wheat, sheep, cattle), 3=(sheep), 4=(cattle), 5=(sheep, cattle)).

The aim is to fit a suitable regression model that explains the revenue of a farm. You will need to perform the following steps:

(a) (1 point) Preprocess the data as needed, i.e. define the necessary factor variables, assess whether transformations are necessary, etc. Check whether there are sufficiently many observations for all levels of the factor variables. The recommendation is that there are at least five observations for each level.

**Solution:** First, we check the structure of the data frame:

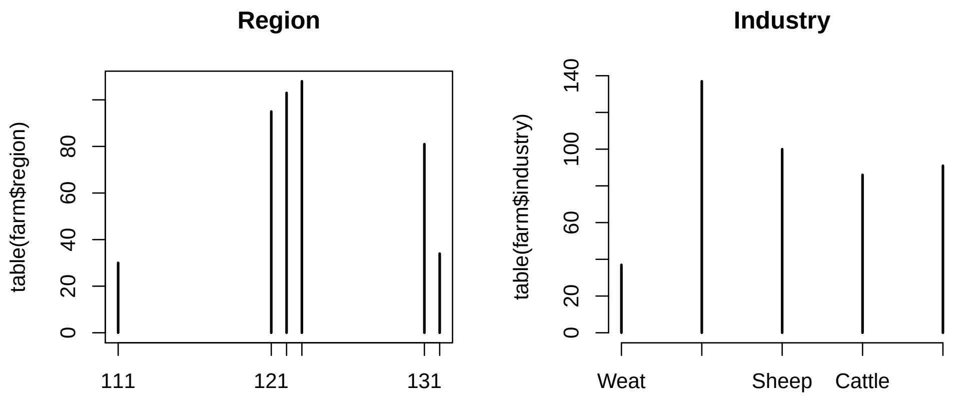
```
> ## load data
> farm <- readRDS("farm.RDS")
> ## check properties of the data
> str(farm)
```

```
'data.frame':	451 obs. of  4 variables:
 $ region  : int  111 111 111 111 111 111 111 111 111 111 ...
 $ industry: int  3 5 2 1 2 5 2 3 3 3 ...
 $ costs   : int  115096 75443 378857 433590 347417 327745 714462 221258 241868 194837 ...
 $ revenue : int  147652 82920 442726 649628 407836 472569 576372 241864 339215 356625 ...
```

All variables are of data type “int”. This is incorrect for the factor variables `region` and `industry` and would lead to incorrect regression results. We define the factor variables as follows:

```
> farm$region   <- factor(farm$region)
> farm$industry <- factor(farm$industry, labels=c("Weat","Weat_Sheep_Cattle",
+                                        "Sheep", "Cattle", "Sheep_Cattle"))
```

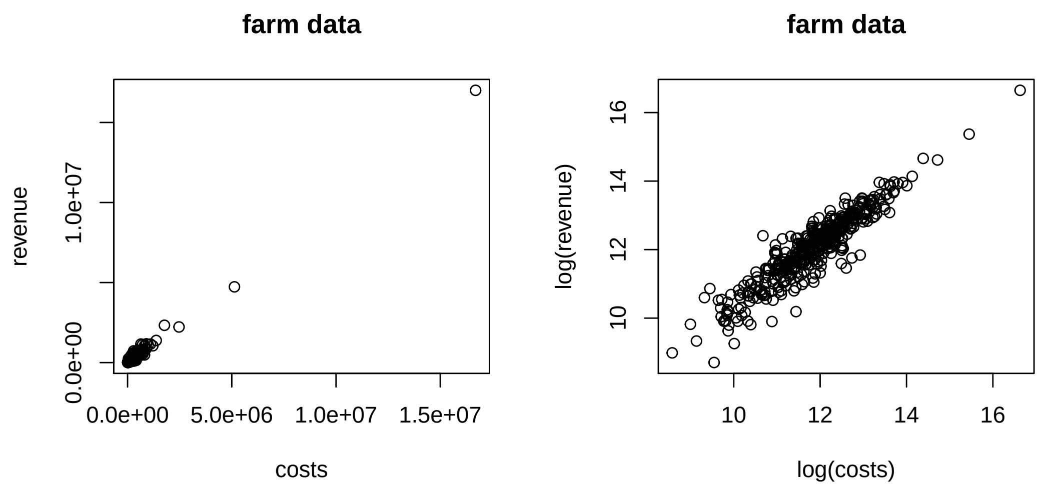
```
> ## visualization
> par(mfrow=c(1,2))
> plot(table(farm$region), main="Region")
> plot(table(farm$industry), main="Industry")
```




The number of observations are sufficient for all levels of the factor variables.
**(1 Point.)**

(b) (2 points) Based on the following two plots:

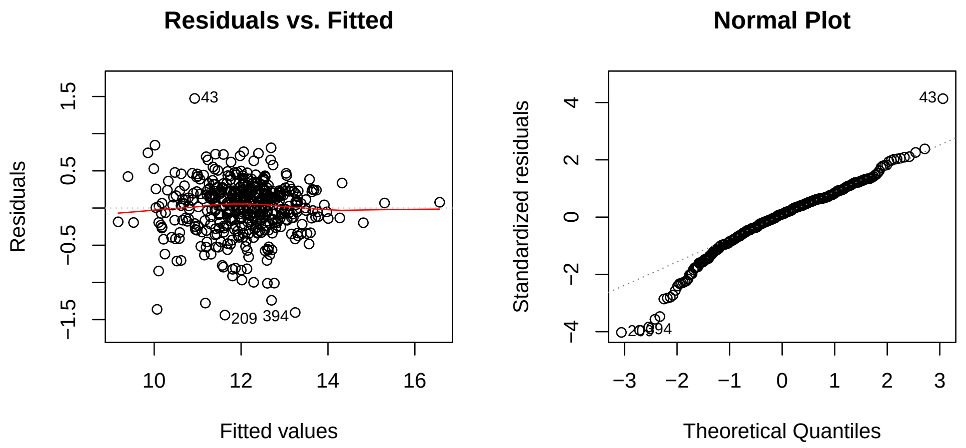
```
> ## visualization
> par(mfrow=c(1,2))
> plot(revenue~costs,data=farm,main="farm data")
> plot(log(revenue)~log(costs),data=farm,main="farm data")
```




we decide to fit the following model:
`fit.farm <- lm(log(revenue) ~ log(costs) + region + industry, data=farm)`.
Fit this model in R and perform a residual analysis (using the TA and QQ plots). Comment on the possible assumption violations. **State all the assumptions you can check with these plots and whether you think they are satisfied.**

**Solution:**

```
> ## fit main effects model
> fit.farm <- lm(log(revenue) ~ log(costs) + region + industry, data=farm)
> ## residual analysis
> par(mfrow=c(1,2))
> plot(fit.farm, which=1, caption="", main="Residuals vs. Fitted")
> plot(fit.farm, which=2, caption="", main="Normal Plot")
```




The Tukey-Anscombe plot does not indicate any model assumption violations (no curvature of the mean or non-constant variance). The Normal plot appears to show that the distribution of the residuals is skewed to the left (the normality assumption is possibly violated). **(2 Points.)**

(c) (1 point) What is the expected revenue of a cattle farm in region 111 with costs of 100'000?

**Solution:**

```
> ## predict
> newdat <- data.frame(costs=10^5, region="111", industry="Cattle")
> predi  <- predict(fit.farm, newdata=newdat)
> exp(predi + 0.5*summary(fit.farm)$sigma^2)

       1
165357.7
```

Using `predict()` we obtain the prediction on the log scale. We thus need to transform the value back to the original scale. So the expected revenue is 165'357.7.

Other acceptable answers:

```
> exp(predi)

       1
154914.2

> ## and
> n <- length(farm$revenue)
> exp(predi)*1/n*sum(exp(resid(fit.farm)))

       1
164455.1
```

**(1 Point.)**

(d) (1 point) Test whether `region` has an influence on `revenue` when the other predictors are given at the 1% level.

**Solution:**

```
> fit.farm2 <- lm(log(revenue) ~ log(costs) +industry, data=farm)
> anova(fit.farm,fit.farm2)

Analysis of Variance Table

Model 1: log(revenue) ~ log(costs) + region + industry
Model 2: log(revenue) ~ log(costs) + industry
```

```
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    440 57.411
2    445 58.775 -5   -1.3639 2.0906 0.06551 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The predictor `region` is not significant at the 1% level, as can be seen from the p-value 0.0655 of the partial F-test. **(1 Point.)**

(e) (3 points) Add an interaction term between `region` and `industry`:
`fit.farm <- lm(log(revenue) ~ log(costs) + region + industry + region:industry, data=farm)`.

i. (1 point) How many parameters are estimated in total?
ii. (1 point) Is the interaction term significant at the 1% level?
iii. (1 point) Based on this whole exercise, which model would you choose to predict the revenue of a farm?

**Solution:**

- 31 parameters are estimated as the model has 420 degrees of freedom and there are 451 observations. **(1 Point.)**
- To test the interaction term we do a partial F-test.

```
> ## option 1
> f.big  <- lm(log(revenue) ~ log(costs) + region + industry + region:industry, data=farm)
> f.small <- lm(log(revenue) ~ log(costs) + region + industry, data=farm)
> anova(f.small, f.big)
```

```
Analysis of Variance Table

Model 1: log(revenue) ~ log(costs) + region + industry
Model 2: log(revenue) ~ log(costs) + region + industry + region:industry
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    440 57.411
2    420 54.540 20    2.8706 1.1053 0.3404
```

The interaction term is not significant at the 1% level. **(1 Point.)**

- The interaction term is not significant at the 1% level (or even the 5% level), as we have seen above. Also, we have seen that `region` is not significant at the 1% level (or 5% level), so we will exclude it as well. Hence, we choose the model where the (logarithmic) revenue is explained with the (logarithmic) costs and the industry. **(1 Point.)**