Homework 10

Stat 135: Concepts of Statistics
UC Berkeley, Fall 2022

Due: **December 2nd, 11:59PM**

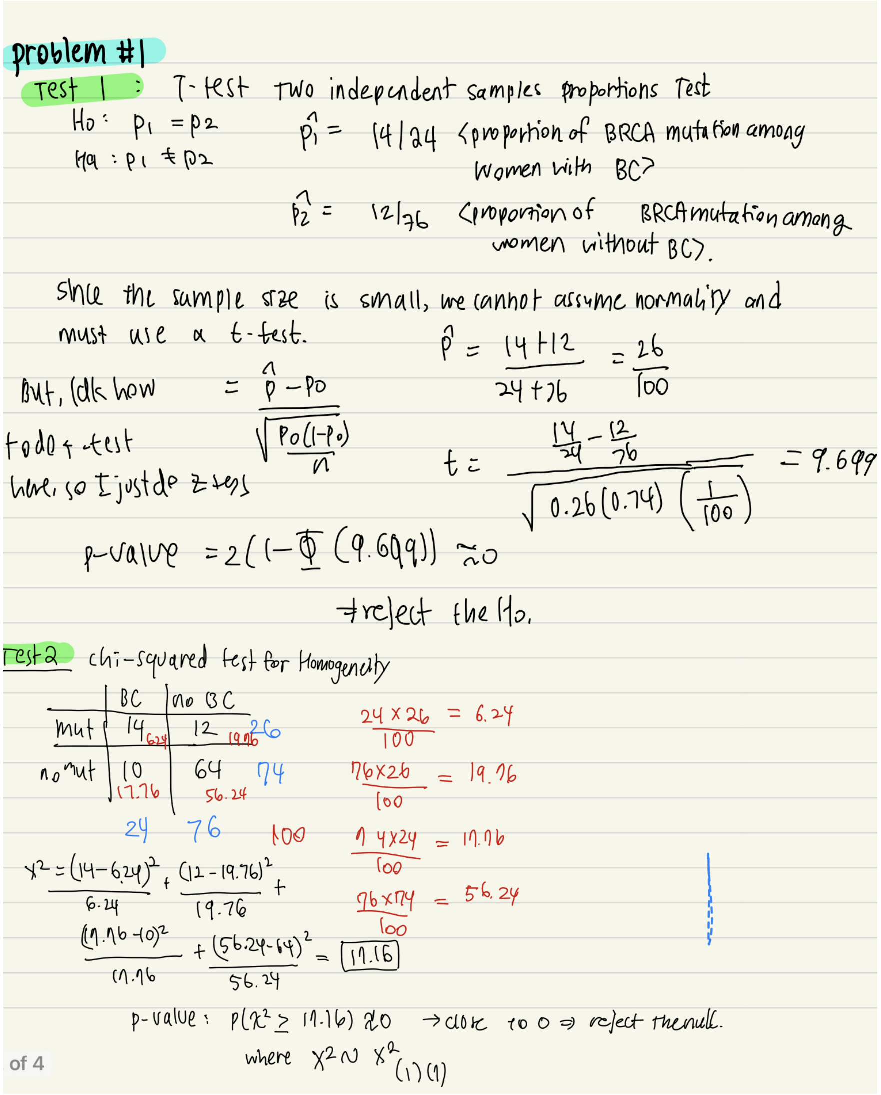
- Read sections 13.3, 13.4 and 14.1 up to and including 14.3. Go over the slides on categorical data and linear regression. We also cover some parts of 14.4 in the textbook.
- Your homework must be submitted to Gradescope **as a single PDF file.**
- To obtain full credit, please **write your answers clearly and show your reasoning.**

Problem 1

Scientists have shown that women with a BRCA genetic mutation are substantially more likely to develop breast cancer over their lifetimes. The following table shows the breast cancer incidence by age 50 in 100 women, some of whom have the BRCA mutation, and some who do not.

	Breast Cancer	**No Breast Cancer**
BRCA Mutation	14	12
No BRCA Mutation	10	64

Test the hypothesis that BRCA mutation status is independent of breast cancer (or that the proportion of breast cancer is the same among those with the BRCA and without the BRCA mutation) *using two different hypothesis test techniques.* For each test, write down the name of the test, the null and alternative hypothesis, the value of the test-statistic and the p-value. You should compute the test-statistic manually, but may directly use the relevant inbuilt tests for the p-values. You need to show your code that computed each test.



problem #1

Test 1 : T-test Two independent samples proportions Test

$H_0: p_1 = p_2$

$H_a: p_1 \neq p_2$

$\hat{p}_1 = 14/24$ <proportion of BRCA mutation among women with BC>

$\hat{p}_2 = 12/76$ <proportion of BRCA mutation among women without BC>.

Since the sample size is small, we cannot assume normality and must use a t-test.

$$\hat{p} = \frac{14+12}{24+76} = \frac{26}{100}$$

But, Idk how to do t-test here, so I just do z-test

$$= \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$$t = \frac{\frac{14}{24} - \frac{12}{76}}{\sqrt{0.26(0.74)\left(\frac{1}{100}\right)}} = 9.699$$

$$\text{p-value} = 2(1 - \Phi(9.699)) \approx 0$$

→ reject the H_0.

Test 2 chi-squared test for Homogeneity

	BC	no BC	
mut	14 6.24	12 19.76	26
no mut	10 17.76	64 56.24	74
	24	76	100

$$\frac{24 \times 26}{100} = 6.24$$

$$\frac{76 \times 26}{100} = 19.76$$

$$\frac{74 \times 24}{100} = 17.76$$

$$\frac{76 \times 74}{100} = 56.24$$

$$X^2 = \frac{(14-6.24)^2}{6.24} + \frac{(12-19.76)^2}{19.76} + \frac{(17.76-10)^2}{17.76} + \frac{(56.24-64)^2}{56.24} = \boxed{17.16}$$

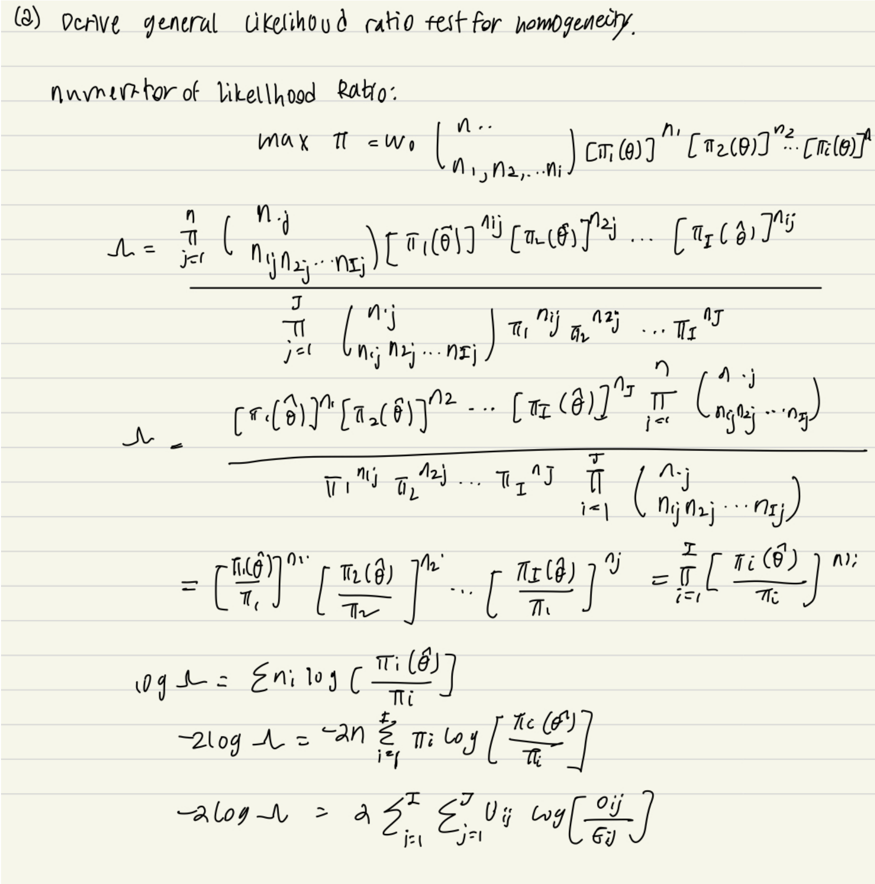
p-value : $P(X^2 \geq 17.16) \approx 0$ → close to 0 ⇒ reject the null.

where $X^2 \sim X^2_{(1)(1)}$

of 4

Problem 2

(a) Derive the (general) likelihood ratio test of homogeneity. The setup is exactly the same as the test of homogeneity for categorical data, the only difference is that we are deriving the LR test statistic instead of the Chi-squared test statistic.



(a) Derive general likelihood ratio test for homogeneity.

numerator of likelihood Ratio:

$$\max \pi = w_0 \binom{n_{\cdot\cdot}}{n_1, n_2, \ldots n_i} [\pi_1(\theta)]^{n_1} [\pi_2(\theta)]^{n_2} \cdots [\pi_i(\theta)]^{n}$$

$$\Lambda = \frac{\prod_{j=1}^{n} \binom{n_{\cdot j}}{n_{1j} n_{2j} \cdots n_{Ij}} [\pi_1(\hat\theta)]^{n_{1j}} [\pi_2(\hat\theta)]^{n_{2j}} \cdots [\pi_I(\hat\theta)]^{n_{ij}}}{\prod_{j=1}^{J} \binom{n_{\cdot j}}{n_{1j} n_{2j} \cdots n_{Ij}} \pi_1^{n_{1j}} \pi_2^{n_{2j}} \cdots \pi_I^{n_J}}$$

$$\Lambda = \frac{[\pi_1(\hat\theta)]^{n_1} [\pi_2(\hat\theta)]^{n_2} \cdots [\pi_I(\hat\theta)]^{n_J} \prod_{j=1}^{n} \binom{n_{\cdot j}}{n_{1j} n_{2j} \cdots n_{Ij}}}{\pi_1^{n_{1j}} \pi_2^{n_{2j}} \cdots \pi_I^{n_J} \prod_{i=1}^{J} \binom{n_{\cdot j}}{n_{1j} n_{2j} \cdots n_{Ij}}}$$

$$= \left[\frac{\pi_1(\hat\theta)}{\pi_1}\right]^{n_1} \left[\frac{\pi_2(\hat\theta)}{\pi_2}\right]^{n_2} \cdots \left[\frac{\pi_I(\hat\theta)}{\pi_1}\right]^{n_j} = \prod_{i=1}^{I} \left[\frac{\pi_i(\hat\theta)}{\pi_i}\right]^{n_i}$$

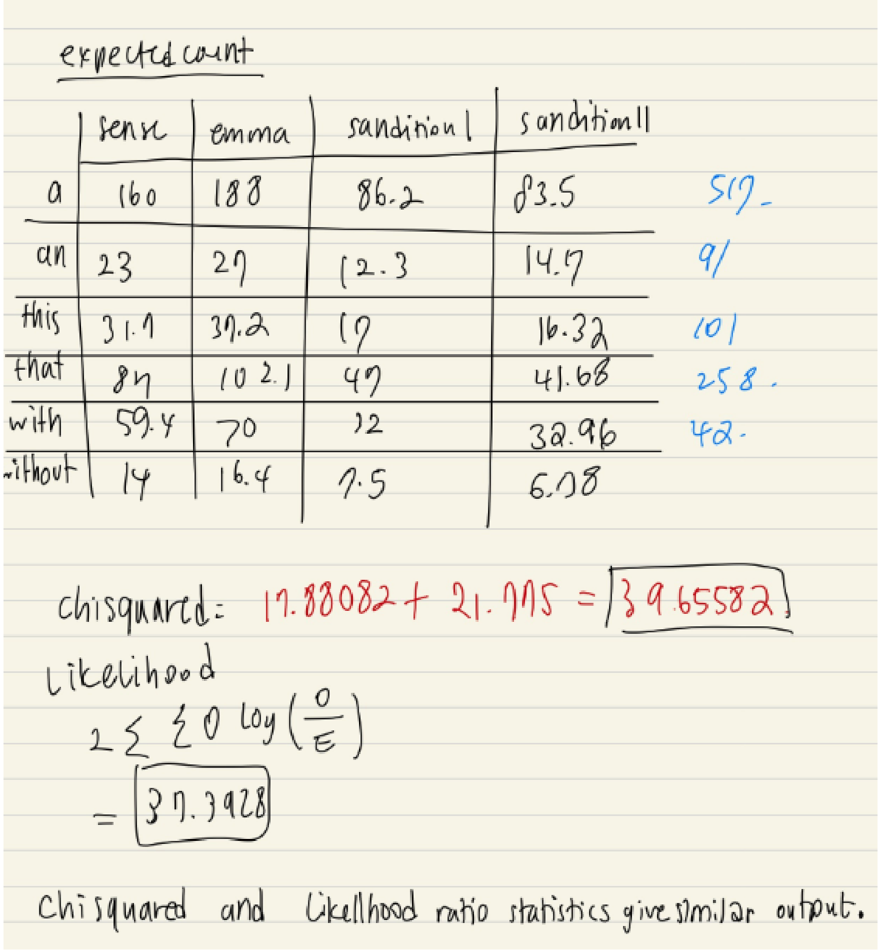
$$\log \Lambda = \sum n_i \log \left[\frac{\pi_i(\hat\theta)}{\pi_i}\right]$$

$$-2\log \Lambda = -2n \sum_{i=1}^{I} \pi_i \log \left[\frac{\pi_i(\hat\theta)}{\pi_i}\right]$$

$$-2\log \Lambda = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} O_{ij} \log \left[\frac{O_{ij}}{E_{ij}}\right]$$

(b) Calculate the likelihood ratio test statistic for the table below from the example we covered in the class. You should use R to do this. How does the test statistic compare to Pearson's chi-squared statistic?

Word	**Sense and Sensibility**	**Emma**	**Sanditon I**	**Sanditon II**
a	147	186	101	83
an	25	26	11	29
this	32	39	15	15
that	94	105	37	22
with	59	74	28	43
without	18	10	10	4
Total	375	440	202	196



expected count

	sense	emma	sandition I	sandition II	
a	160	188	86.2	83.5	517.
an	23	27	12.3	14.7	91
this	31.7	37.2	17	16.32	101
that	87	102.1	47	41.68	258.
with	59.4	70	32	32.96	42.
without	14	16.4	7.5	6.08	

chisquared: $17.88082 + 21.775 = \boxed{39.65582}$

Likelihood

$2\sum\{O\log\left(\frac{O}{E}\right)$

$= \boxed{37.7928}$

Chisquared and Likellhood ratio statistics give similar output.

Problem 3

Plot x versus y for the following pairs:

x	y
0.34	0.27
1.38	1.34
-0.65	-0.53
0.68	0.35
1.40	1.28
-0.88	-0.98
-0.30	-0.72
-1.18	-0.81
0.50	0.64
-1.75	1.59

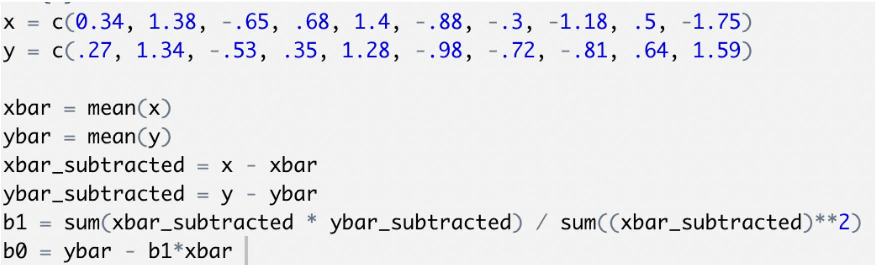
(a) Fit a line $y = \beta_0 + \beta_1 x$ using least squares. Show your working for computing the coefficient values β_0 and β_1. You should use the formula we derived in class to compute them, and you should check your answer using the `lm()` function.

Solution:

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

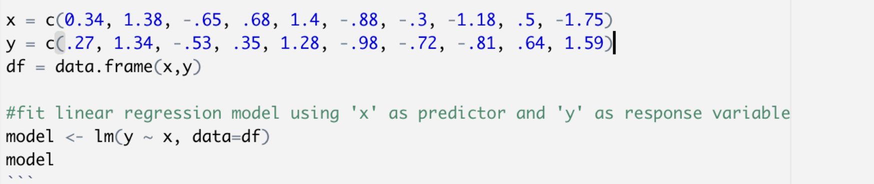
$\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)}$

Using these equations, we get $\beta_1 = 0.3840358$ and $\beta_0 = 0.2606656$.

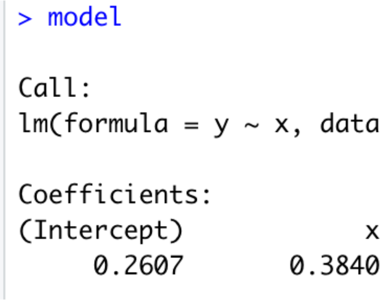


```
x = c(0.34, 1.38, -.65, .68, 1.4, -.88, -.3, -1.18, .5, -1.75)
y = c(.27, 1.34, -.53, .35, 1.28, -.98, -.72, -.81, .64, 1.59)

xbar = mean(x)
ybar = mean(y)
xbar_subtracted = x - xbar
ybar_subtracted = y - ybar
b1 = sum(xbar_subtracted * ybar_subtracted) / sum((xbar_subtracted)**2)
b0 = ybar - b1*xbar
```

I checked with the `lm()` function.



~~~
x = c(0.34, 1.38, -.65, .68, 1.4, -.88, -.3, -1.18, .5, -1.75)
y = c(.27, 1.34, -.53, .35, 1.28, -.98, -.72, -.81, .64, 1.59)
df = data.frame(x,y)

#fit linear regression model using 'x' as predictor and 'y' as response variable
model <- lm(y ~ x, data=df)
model
```
~~~

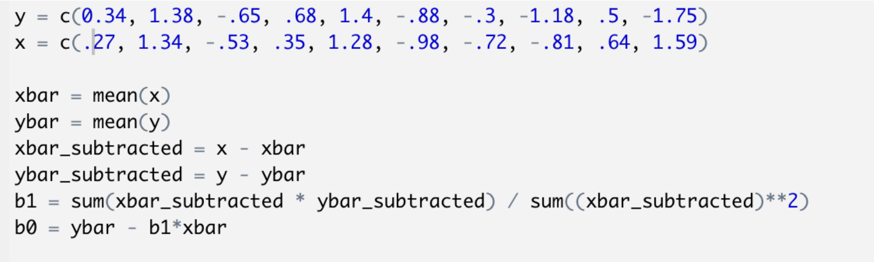


```
> model

Call:
lm(formula = y ~ x, data = df)

Coefficients:
(Intercept)            x
     0.2607       0.3840
```

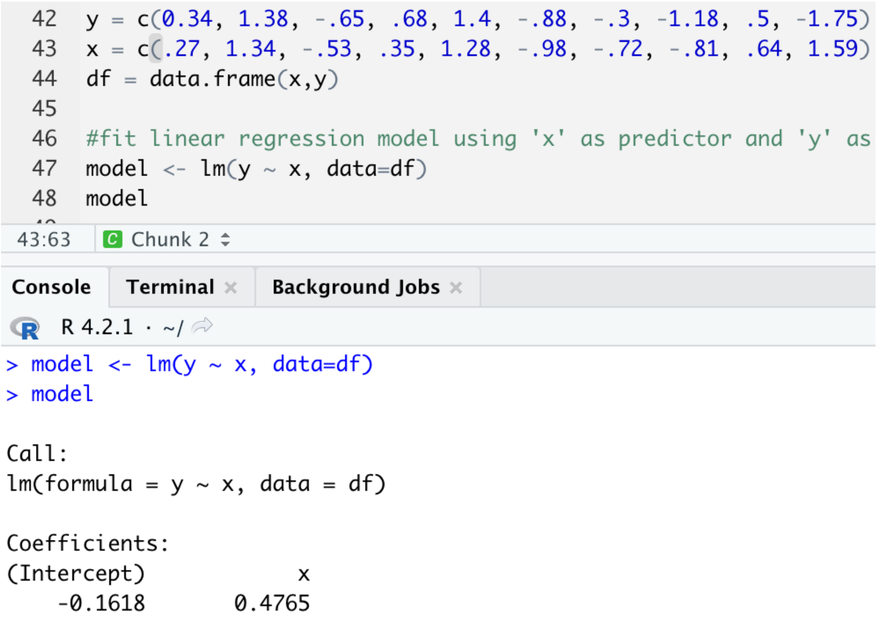
(b) Fit a line $x = \alpha_0 + \alpha_1 y$ using least squares. Show your working for computing the coefficient values α_0 and α_1. You should use the formula we derived in class to compute them, and you should check your answer using the `lm()` function.

Solution:

Same calculation method as (a). For $x = \alpha_0 + \alpha_1 y$, we can think of x being the y and y being the x (in (a)). With the equation derived in class, we find that $\beta_0 = -0.1617894$ and $\beta_1 = 0.4764996$.



```
y = c(0.34, 1.38, -.65, .68, 1.4, -.88, -.3, -1.18, .5, -1.75)
x = c(.27, 1.34, -.53, .35, 1.28, -.98, -.72, -.81, .64, 1.59)

xbar = mean(x)
ybar = mean(y)
xbar_subtracted = x - xbar
ybar_subtracted = y - ybar
b1 = sum(xbar_subtracted * ybar_subtracted) / sum((xbar_subtracted)**2)
b0 = ybar - b1*xbar
```



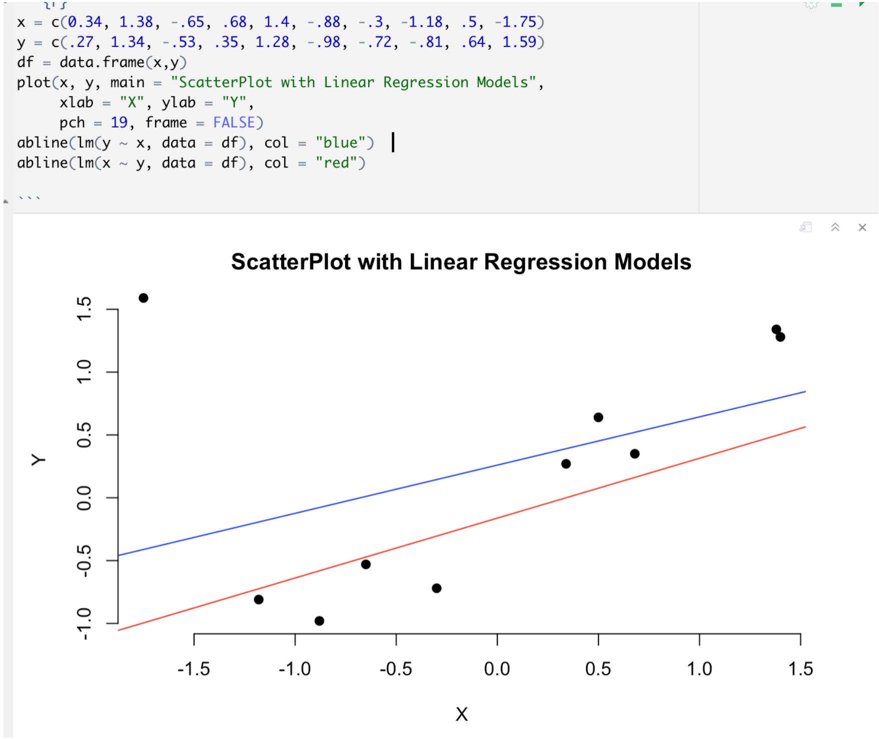
```
y = c(0.34, 1.38, -.65, .68, 1.4, -.88, -.3, -1.18, .5, -1.75)
x = c(.27, 1.34, -.53, .35, 1.28, -.98, -.72, -.81, .64, 1.59)
df = data.frame(x,y)

#fit linear regression model using 'x' as predictor and 'y' as
model <- lm(y ~ x, data=df)
model
```

```
> model <- lm(y ~ x, data=df)
> model

Call:
lm(formula = y ~ x, data = df)

Coefficients:
(Intercept)            x
    -0.1618       0.4765
```

The answers from the `lm()` function confirm my calculation.

(c) On a single scatterplot of y against x, show the two fitted lines (indicate which is which). Are the lines in parts (a) and (b) the same? If not, why not? Provide an intuitive explanation.

Solution:



```
x = c(0.34, 1.38, -.65, .68, 1.4, -.88, -.3, -1.18, .5, -1.75)
y = c(.27, 1.34, -.53, .35, 1.28, -.98, -.72, -.81, .64, 1.59)
df = data.frame(x,y)
plot(x, y, main = "ScatterPlot with Linear Regression Models",
     xlab = "X", ylab = "Y",
     pch = 19, frame = FALSE)
abline(lm(y ~ x, data = df), col = "blue")
abline(lm(x ~ y, data = df), col = "red")

```
```

The lines in part (a) and (b) are different because they have different slope and intercepts. It makes
```

sense because in (a), x is an independent variable and y is a dependent variable. On the other hand, in (b), x is a dependent variable and y is an independent variable. The linear regression line will look different because the slopes and intercept calculations are dependent on the independent and dependent variables.

## Problem 4

Suppose that $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where the $\epsilon_i$ are i.i.d with $E[\epsilon_i] = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$. The Least Squares estimates of $\beta_0$ and $\beta_1$ are

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

Prove the following identities:

(a) $\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x}) y_i}{\sum_{i=1}^n (x_i - \overline{x})^2}$

**Solution:** $\frac{\sum_{i=1}^n (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i y_i - x_i \bar{y}) - n\bar{x}\bar{y} + n\bar{x}\bar{y}}{\sum_{i=1}^n (x_i - \bar{x})^2}$ since $\sum_{i=1}^n \bar{x} y_i = n\bar{x}\bar{y}$.
$= \frac{\sum_{i=1}^n (x_i y_i - x_i \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \overline{x}) y_i}{\sum_{i=1}^n (x_i - \overline{x})^2}$ since $\sum_{i=1}^n \bar{y} x_i = \sum_{i=1}^n y_i \bar{x}$

(b) $\text{Cov}(\overline{y}, \widehat{\beta}_1) = 0$

**Solution:** $Cov(\bar{y}, \hat{\beta}_1) = Cov(\frac{1}{n} \sum_{i=1}^n y_i, \sum_{i=1}^n c_i y_i)$ where $c_i = \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$
$\frac{1}{n} \sum_{i=1}^n c_i Var(y_i) = \frac{2}{n} \sum_{i=1}^n c_i Var(y_i) = \frac{\sigma^2}{n} \sum_{i=1}^n c_i = 0$

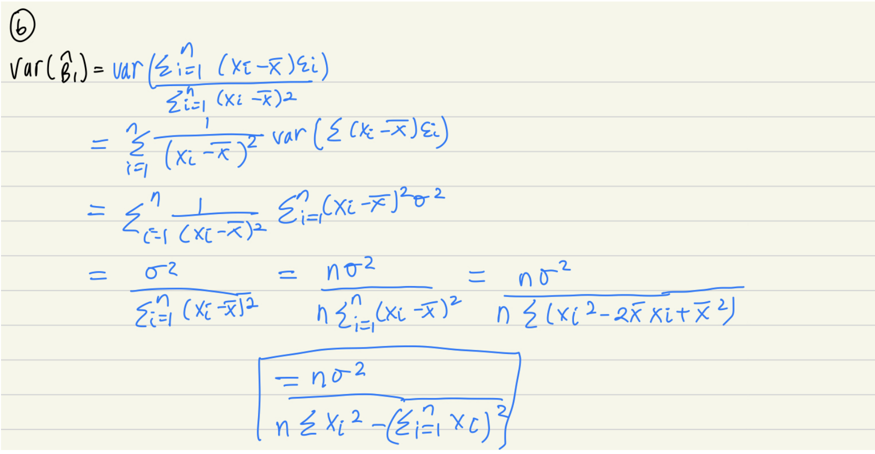
Use these identities to show that the variance and covariance of the Least Squares estimates are given by:

(a) $\text{Var}(\widehat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$

**Solution:**
$Var(\hat{\beta}_0) = Var(\bar{Y} - \hat{\beta}_1 \bar{x}) = Var(\bar{Y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x} cov(\bar{Y}, \hat{\beta}_1)$.
$Var(\bar{Y}) = Var(\frac{1}{n} \sum_{i=1}^n Y_i) = \frac{1}{n^2} \sum_{i=1}^n Var(Y_i) = \frac{\sigma^2}{n}$.
$Var(\beta_1) = \frac{1}{[\sum_{i=1}^n (x_i - \bar{x})^2]^2} \sum_{i=1}^n (x_i - \bar{x})^2 Var(Y_i) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Therefore, $Var(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{n \sum_{i=1}^n}$

(b) $\text{Var}(\widehat{\beta}_1) = \frac{n\sigma^2}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$



(b)

$$Var(\hat{\beta}_1) = var\left(\frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

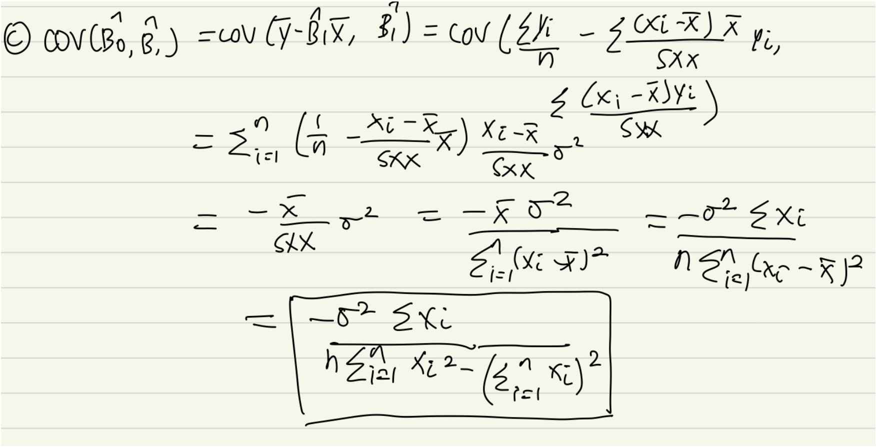
$$= \sum_{i=1}^n \frac{1}{(x_i - \bar{x})^2} var\left(\sum (x_i - \bar{x})\varepsilon_i\right)$$

$$= \sum_{i=1}^n \frac{1}{(x_i - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2$$

$$= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{n\sigma^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{n\sigma^2}{n \sum (x_i^2 - 2\bar{x} x_i + \bar{x}^2)}$$

$$\boxed{= \frac{n\sigma^2}{n \sum x_i^2 - (\sum_{i=1}^n x_i)^2}}$$

(c) $\text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{-\sigma^2 \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{\sigma^2}{n} +$



$$\text{ⓒ } \text{cov}(\hat{\beta}_0, \hat{\beta}_1) = \text{cov}(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1) = \text{cov}\left( \frac{\sum y_i}{n} - \sum \frac{(x_i - \bar{x})\bar{x}}{Sxx} y_i, \sum \frac{(x_i - \bar{x})y_i}{Sxx} \right)$$

$$= \sum_{i=1}^n \left( \frac{1}{n} - \frac{x_i - \bar{x}}{Sxx} \bar{x} \right) \frac{x_i - \bar{x}}{Sxx} \sigma^2$$

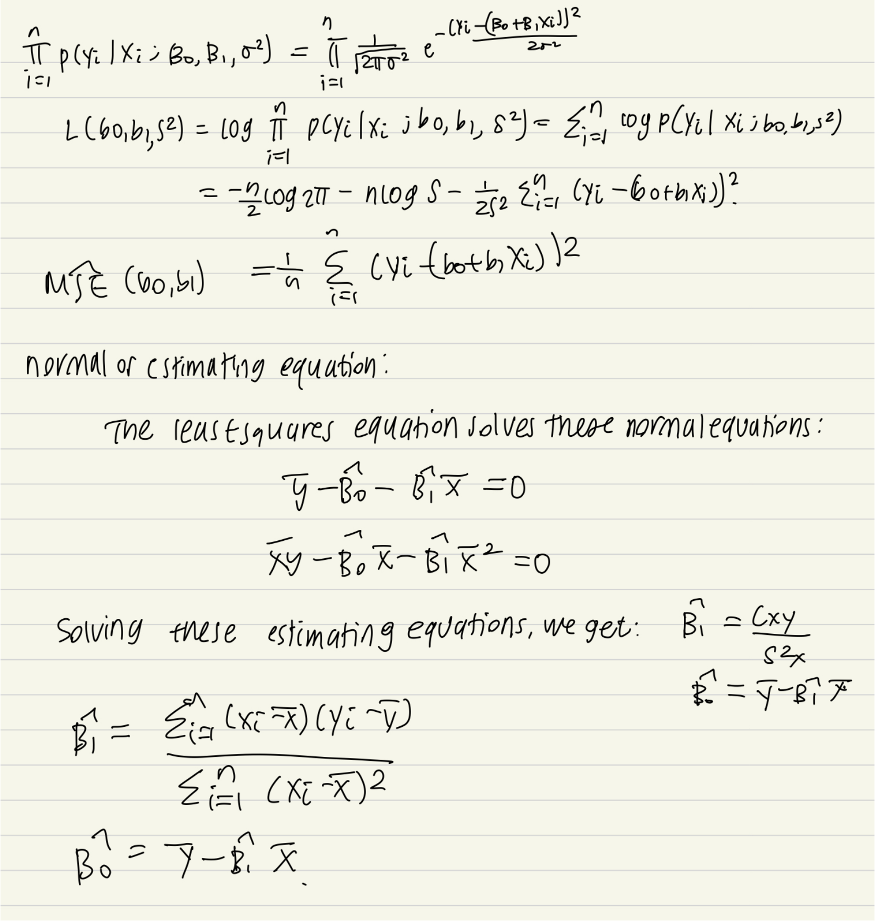
$$= -\frac{\bar{x}}{Sxx} \sigma^2 = \frac{-\bar{x}\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{-\sigma^2 \sum x_i}{n \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= \boxed{\frac{-\sigma^2 \sum x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}}$$

Note that the $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are random variables (because they depend on the $y_i$s, which are random variables due to randomness in $\epsilon_i$s) and you are being asked to compute the population variance and covariance of these estimators.

## Problem 5

If $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$, show that the Maximum Likelihood estimators of $\beta_0$ and $\beta_1$ correspond to the Least Squares estimates of $\beta_0$ and $\beta_1$.
*Hint: The likelihood is the joint density function of the $y_i$s. What is the distribution of a single $y_i$?*



$$\prod_{i=1}^{n} p(y_i | x_i ; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - (\beta_0+\beta_1 x_i))^2}{2\sigma^2}}$$

$$L(b_0, b_1, s^2) = \log \prod_{i=1}^{n} p(y_i | x_i ; b_0, b_1, s^2) = \sum_{i=1}^{n} \log p(y_i | x_i ; b_0, b_1, s^2)$$

$$= -\frac{n}{2}\log 2\pi - n\log s - \frac{1}{2s^2}\sum_{i=1}^{n}(y_i - (b_0+b_1x_i))^2.$$

$$\widehat{MSE}(b_0,b_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i-(b_0+b_1x_i))^2$$

normal or estimating equation:

The least squares equation solves these normal equations:

$$\bar{y} - \hat{\beta}_0 - \hat{\beta}_1\bar{x} = 0$$

$$\overline{xy} - \hat{\beta}_0\bar{x} - \hat{\beta}_1\overline{x^2} = 0$$

Solving these estimating equations, we get: $\hat{\beta}_1 = \frac{c_{xy}}{s^2_x}$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

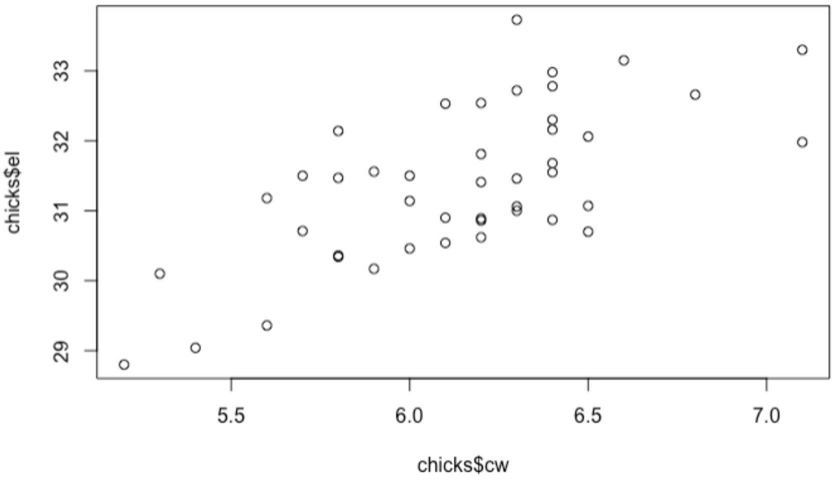
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}.$$

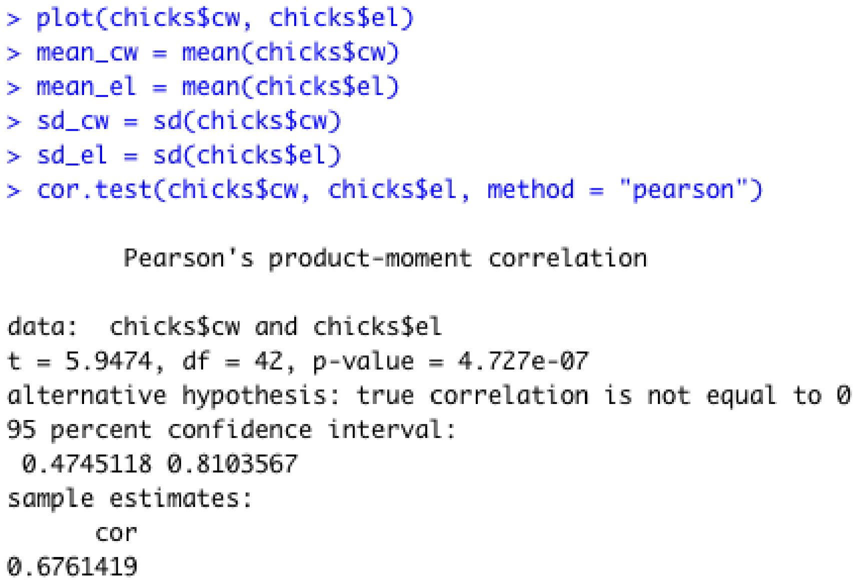
# Problem 6

The dataset chicks was obtained from BLSS: The Berkeley Interactive Statistical System by Abrahams and Rizzardi. Each observation corresponds to an egg (and the resulting chick) of a bird called the Snowy Plover. The data were taken at Point Reyes Bird Observatory. Column 1 contains the egg length in millimeters, Column 2 the egg breadth in millimeters, Column 3 the egg weight in grams, and Column 4 the chick weight in grams. The object is to estimate the size of the chick based on dimensions of the egg.

(a) First you are going to regress chick weight on egg length. Write out the standard regression model ($y_i = \beta_0 + \beta_1 x_i + \epsilon_i$) in terms of these two variables. You may assume that $\epsilon_i$ is normally distributed in this assignment. Plot the data and comment on whether the assumptions of the model are reasonable. Find the means and SDs of both variables, as well as the correlation between them. Manually compute the slope and the intercept of the regression line. Provide units for the slope and the intercept, and write the equation of the regression line. Draw the regression line on your plot. Do not use `geom smooth()` or `lm()` for part a. You can use `geom_point()`.

**Solution:**



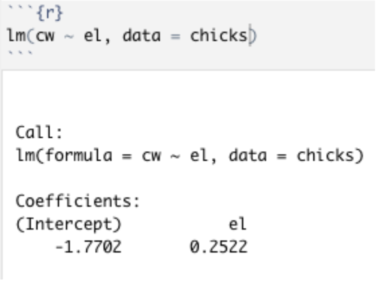


```
> plot(chicks$cw, chicks$el)
> mean_cw = mean(chicks$cw)
> mean_el = mean(chicks$el)
> sd_cw = sd(chicks$cw)
> sd_el = sd(chicks$el)
> cor.test(chicks$cw, chicks$el, method = "pearson")

	Pearson's product-moment correlation

data:  chicks$cw and chicks$el
t = 5.9474, df = 42, p-value = 4.727e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4745118 0.8103567
sample estimates:
      cor
0.6761419
```

Slope: $\frac{cov(x,y)}{var(x)}$= 0.3056258 / 1.211963 = 0.2521742 (gram/milimeters)
Intercept $= \bar{y} - \hat{\beta}_1 \bar{x}$ = -1.770179 gram
Equation of the regression line: (chick weight) = .2521742 * (egg length) - 1.770179

(b) Now use R to regress chick weight on egg length. Check that R produces the same slope and intercept that you got in part a For the t statistic in the R output, state the null and alternative hypotheses that are being tested, and state the conclusion of the test.

Solution:
Using R, we can verify the answer I got on a.



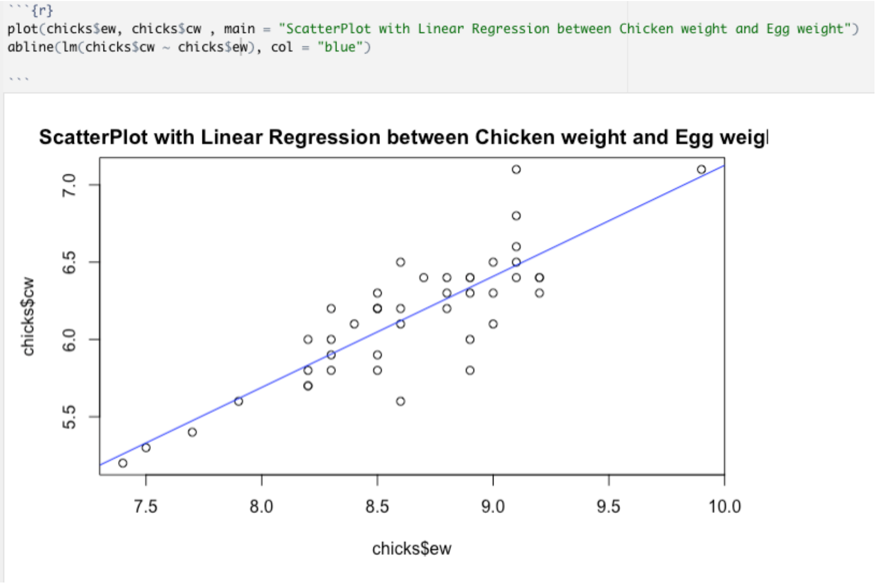
~~~
```{r}
lm(cw ~ el, data = chicks)
```
~~~

```
Call:
lm(formula = cw ~ el, data = chicks)

Coefficients:
(Intercept)           el
    -1.7702       0.2522
```

The t-test in linear regression tests the linear relationship between the independent and dependent variables. Null hypothesis is that the slope of the regression model is zero whereas the alternative hypothesis states that the slope is nonzero. We fail to reject the null.

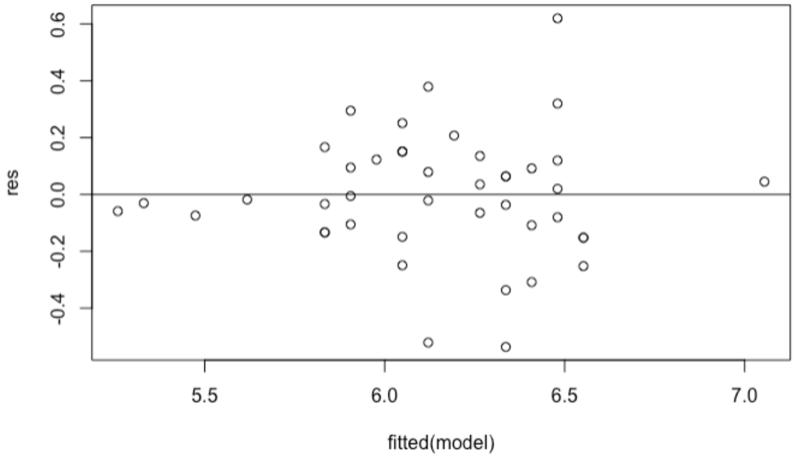
(c) Which of the three variables egg length, egg breadth, and egg weight is most highly correlated with chick weight? Call this one the best predictor for now. Draw the scatter diagram of chick weight versus this best predictor and put the regression line through it. Draw the residual plot. Is there any noticeable heteroscedasticity?

Solution:



~~~
```{r}
plot(chicks$ew, chicks$cw , main = "ScatterPlot with Linear Regression between Chicken weight and Egg weight")
abline(lm(chicks$cw ~ chicks$ew), col = "blue")

```
~~~

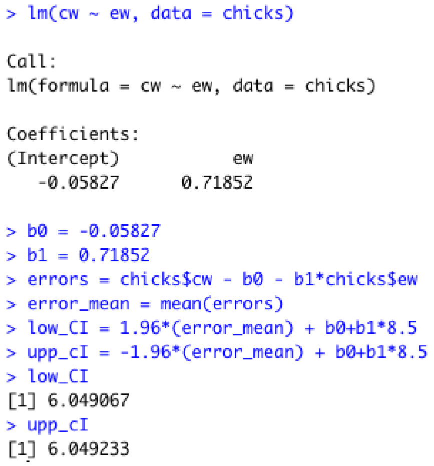
ScatterPlot with Linear Regression between Chicken weight and Egg weigl
chicks$cw
7.0
6.5
6.0
5.5
7.5
8.0
8.5
9.0
9.5
10.0
chicks$ew



Egg weight has the highest correlation with chick weight because its correlation coefficient is .84722, which is the greatest of .73368 (correlation with egg breath) and .6761 (correlation with egg length). I do notice some heteroscedasticity because as the value of the fitted model increases the range of residuals also increase.

(d) If possible, construct a 95%-confidence interval for the mean weight of Snowy Plover chicks that hatch from eggs weighing 8.5 grams.
Hint: You need to estimate the variance of error terms σ^2 *using the errors from the estimated model.*

Solution: The variance of the error term can be estimated by the equation $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i))^2$.

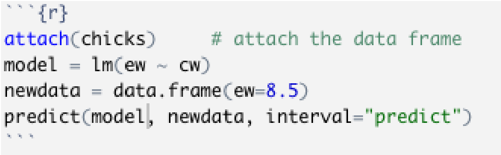


```
> lm(cw ~ ew, data = chicks)

Call:
lm(formula = cw ~ ew, data = chicks)

Coefficients:
(Intercept)           ew
   -0.05827      0.71852

> b0 = -0.05827
> b1 = 0.71852
> errors = chicks$cw - b0 - b1*chicks$ew
> error_mean = mean(errors)
> low_CI = 1.96*(error_mean) + b0+b1*8.5
> upp_cI = -1.96*(error_mean) + b0+b1*8.5
> low_CI
[1] 6.049067
> upp_cI
[1] 6.049233
```

(e) I have a Snowy Plover egg that weighs 8.5 grams. If possible, construct a 95%- prediction interval for the weight of the chick that will hatch from this egg. Note that this is a prediction interval, rather than a confidence interval, because it is trying to predict the value of a random variable instead of estimating a fixed parameter.
Hint: Be careful about the variance you use to construct the interval.

Prediction interval is different from the confidence interval in that it attempts to calculate the random variable whereas the confidence interval is the estimation of the fixed value. SE is 0.2181174. $t_{0.025,42}$

is approximately 2.971. Y hat is 6.04915. Therefore, the prediction interval is (5.401123,6.697177).



~~~
```{r}
attach(chicks)     # attach the data frame
model = lm(ew ~ cw)
newdata = data.frame(ew=8.5)
predict(model, newdata, interval="predict")
```
~~~

(f) If you can, repeat parts d and e when the egg weight is 12 grams instead of 8.5 grams, and if you cannot explain why.

(Hint: beware of extrapolation)

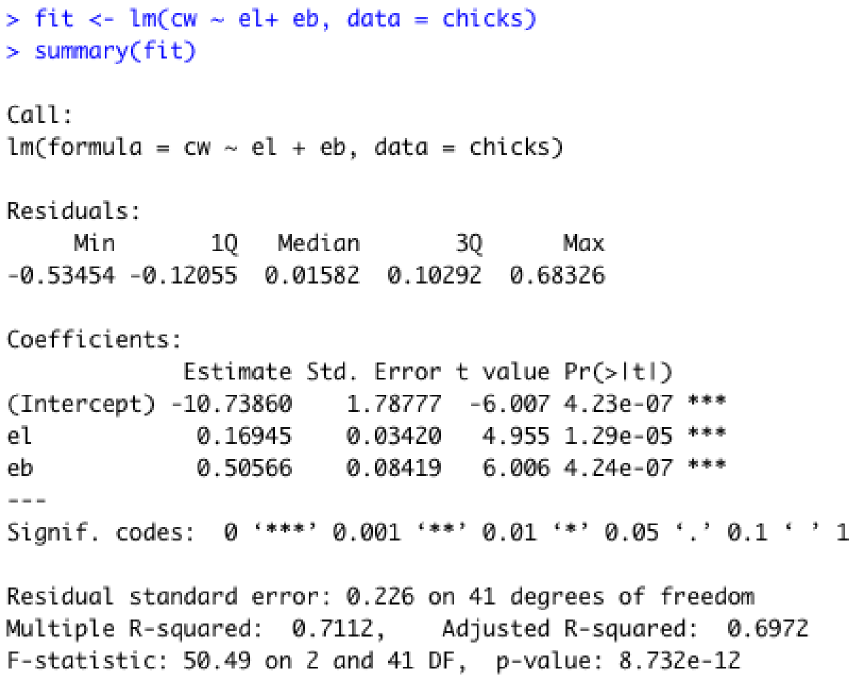
solution: We cannot estimate with the egg weight with 12 grams because none of the data points in the datasets have egg weight above 10. Therefore, it will be inaccurate to estimate the chick weight using the egg weight which we don't have any information with in the sample.

Problem 7

We will continue with problem 6. The object is still to find a good way to predict the weight of a chick given measurements on the egg, using linear regression as the only tool. The difference between this problem and problem 6 is that now you are going to use a combination of variables to estimate the weights of the chicks.

(a) Regress the weights of the chicks on the lengths and breadths of the eggs. Assess the regression. Compare it with the best simple regression you performed Problem 6. Is one noticeably better than the other? Make sure you check the assumptions.

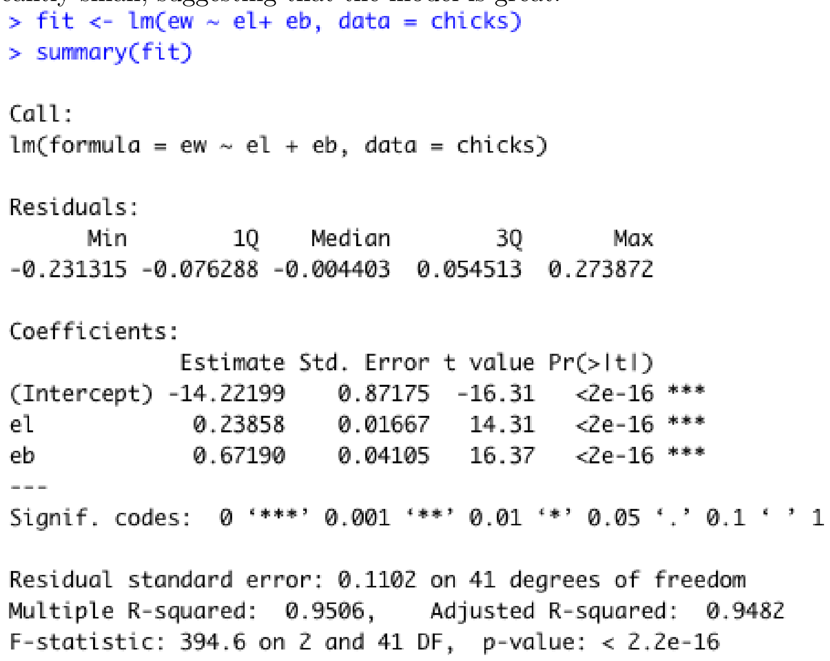
Solution: The model created in the problem 6 is actually pretty good when compared to the model created here. When I regress chick weights on the egg weight, R squared value is .7178 and the p-value is 4.148e-13, wheras when I used lengths of egg and the breadths of the eggs, the p-value is 8.732e-12 and the R squared value is .7112. The models' p-value and the R squared value do not differ much, suggesting having an extra varible does not necessarily mean that the model performs better.ß



```
> fit <- lm(cw ~ el+ eb, data = chicks)
> summary(fit)

Call:
lm(formula = cw ~ el + eb, data = chicks)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.53454 -0.12055  0.01582  0.10292  0.68326 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -10.73860    1.78777  -6.007 4.23e-07 ***
el            0.16945    0.03420   4.955 1.29e-05 ***
eb            0.50566    0.08419   6.006 4.24e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.226 on 41 degrees of freedom
Multiple R-squared:  0.7112,	Adjusted R-squared:  0.6972 
F-statistic: 50.49 on 2 and 41 DF,  p-value: 8.732e-12
```

(b) Regress egg weight on egg length and egg breadth. Assess this regression. Use this regression to explain the similarity (or difference) between the two regressions in that were compared in a.

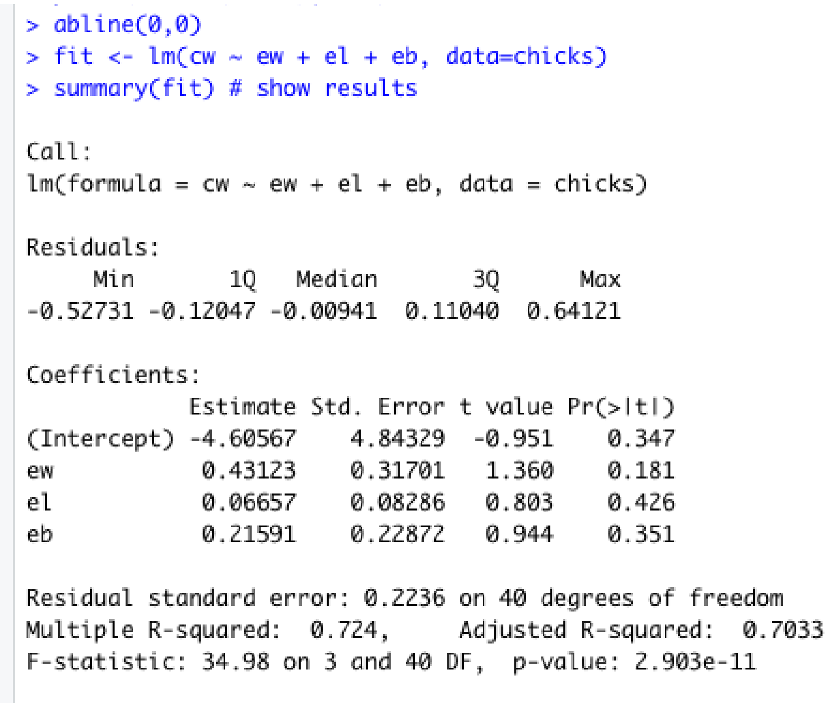
Solution: The R squared value of the regression model is significantly high and the p-value is significantly small, suggesting that the model is great!



```
> fit <- lm(ew ~ el+ eb, data = chicks)
> summary(fit)

Call:
lm(formula = ew ~ el + eb, data = chicks)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.231315 -0.076288 -0.004403  0.054513  0.273872 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -14.22199    0.87175  -16.31   <2e-16 ***
el            0.23858    0.01667   14.31   <2e-16 ***
eb            0.67190    0.04105   16.37   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1102 on 41 degrees of freedom
Multiple R-squared:  0.9506,	Adjusted R-squared:  0.9482 
F-statistic: 394.6 on 2 and 41 DF,  p-value: < 2.2e-16
```

(c) Now regress the weights of the chicks on all three predictor variables: egg length, egg breadth, and egg weight. How do you reconcile the results of the t-tests with the R^2 (or the F-test if you are familiar with it in the context of regression)? Explain why this regression is not as impressive as either of the two you compared in a, even though it has a higher R^2.

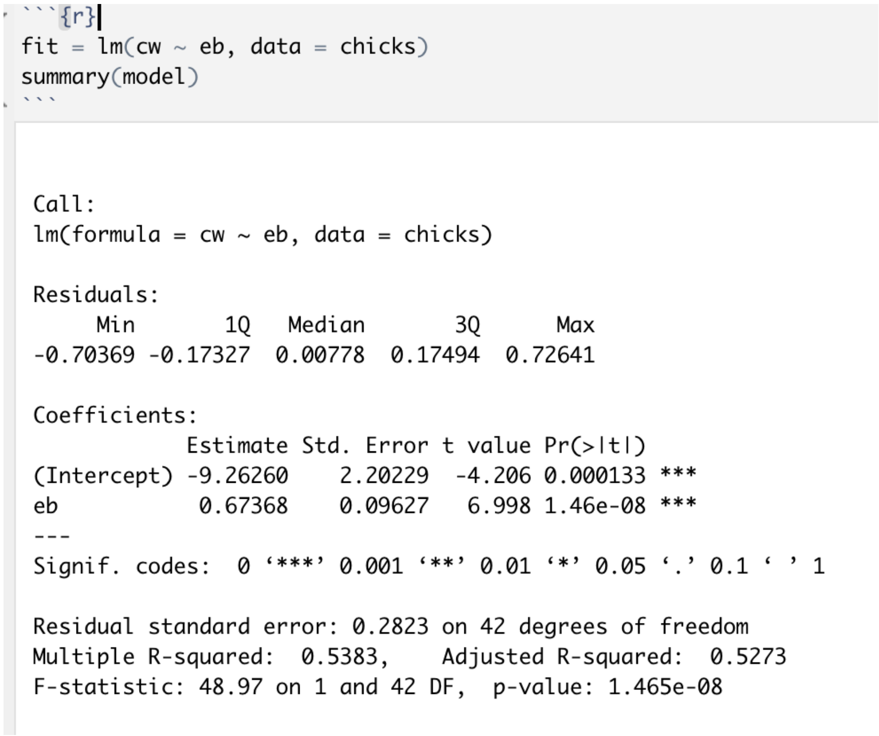
Solution: Using all three predictor variables, I get the multiple Rsquared of .724 and the F-statistic: 34.98 on 3 and 40 DF, p-value: 2.903e-11. Since t-test is about averages and not about individual values, the t-test tells that the predictions are not systematically biased. On the other hand, the high and positive values of the R squared tells you that the prediction model covers and explains a high proportion of the dependent variables. Although this model has a higher R squared value, it is not as significant as the other two because this one uses all the data and falls into the issue of overfitting.



```
> abline(0,0)
> fit <- lm(cw ~ ew + el + eb, data=chicks)
> summary(fit) # show results

Call:
lm(formula = cw ~ ew + el + eb, data = chicks)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.52731 -0.12047 -0.00941  0.11040  0.64121 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.60567    4.84329  -0.951    0.347
ew           0.43123    0.31701   1.360    0.181
el           0.06657    0.08286   0.803    0.426
eb           0.21591    0.22872   0.944    0.351

Residual standard error: 0.2236 on 40 degrees of freedom
Multiple R-squared:  0.724,	Adjusted R-squared:  0.7033 
F-statistic: 34.98 on 3 and 40 DF,  p-value: 2.903e-11
```

(d) Perform all possible regressions of chick weight using combinations of the three predictor variables used in c. You do not turn in all the results. Are there any that are clearly better than the others? Which ones, and why?

Solution: Another combination I could think of is chick weight on egg breadth. R squared value for this regression is 0.5383, which is significantly lower than the regression of chick weight on the egg weight. I think there is a higher correlation between chick weight and egg weight more than egg weight is correlated with egg breadth.



~~~
```{r}
fit = lm(cw ~ eb, data = chicks)
summary(model)
```
~~~

```


Call:
lm(formula = cw ~ eb, data = chicks)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.70369 -0.17327  0.00778  0.17494  0.72641 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -9.26260    2.20229  -4.206 0.000133 ***
eb           0.67368    0.09627   6.998 1.46e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2823 on 42 degrees of freedom
Multiple R-squared:  0.5383,	Adjusted R-squared:  0.5273 
F-statistic: 48.97 on 1 and 42 DF,  p-value: 1.465e-08
```