Linear Regression

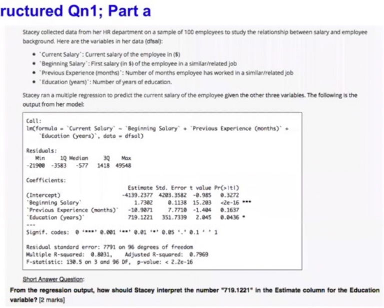
- p-value is small (< 0.05) and F-statistic is large — model has good prediting power — can reject null hypothesis

- R-square is variance

Probability

$$P(X<a) = P\left(Z < \frac{a - mean}{\sqrt{variance}}\right)$$

Structured Qn1; Part a



Stacey collected data from her HR department on a sample of 100 employees to study the relationship between salary and employee background. Here are the variables in her data (dfsal):

- `Current Salary`: Current salary of the employee in ($)
- `Beginning Salary`: First salary (in $) of the employee in a similar/related job
- `Previous Experience (months)`: Number of months employee has worked in a similar/related job
- `Education (years)`: Number of years of education.

Stacey ran a multiple regression to predict the current salary of the employee given the other three variables. The following is the output from her model:

```
Call:
lm(formula = `Current Salary` ~ `Beginning Salary` + `Previous Experience (months)` +
    `Education (years)`, data = dfsal)

Residuals:
   Min     1Q Median     3Q    Max
-21900  -3583   -577   1418  49548

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -4139.2377  4203.3582  -0.985   0.3272
`Beginning Salary`               1.7302     0.1138  15.203   <2e-16 ***
`Previous Experience (months)`  -10.9071     7.7710  -1.404   0.1637
`Education (years)`             719.1221   351.7339   2.045   0.0436 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7791 on 96 degrees of freedom
Multiple R-squared:  0.8031,	Adjusted R-squared:  0.7969
F-statistic: 130.5 on 3 and 96 DF,  p-value: < 2.2e-16
```

Short Answer Question:

From the regression output, how should Stacey interpret the number "719.1221" in the Estimate column for the Education variable? [2 marks]

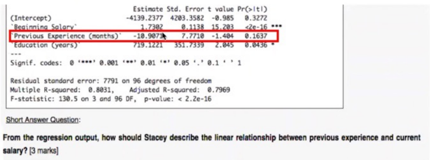
Common Mistakes:

- not including "holding all other IVs constant"
- not including the right units for each variable

Example of a student's answer that got full credits

The number can be interpreted as, holding all the other variables constant, for every 1 year increase in `Education (years)`, `Current Salary` increases by $719.1221.

Structured Qn1; Part b



```
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -4139.2377  4203.3582  -0.985   0.3272
`Beginning Salary`               1.7302     0.1138  15.203   <2e-16 ***
`Previous Experience (months)`  -10.9071     7.7710  -1.404   0.1637
`Education (years)`             719.1221   351.7339   2.045   0.0436 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7791 on 96 degrees of freedom
Multiple R-squared:  0.8031,	Adjusted R-squared:  0.7969
F-statistic: 130.5 on 3 and 96 DF,  p-value: < 2.2e-16
```

Short Answer Question:

From the regression output, how should Stacey describe the linear relationship between previous experience and current salary? [3 marks]

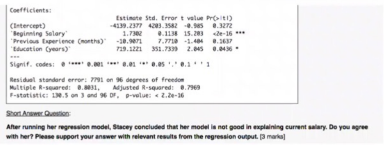
Common Mistakes:

- only interpreting the coefficient and not the p-value
- some stated the linear relationship is negative but due to p>0.05, then it's positive

Example of a student's answer that got full credits

The number "-10.9071" under the "Estimate" column for `Previous Experience (months)` shows that there is a negative linear relationship between previous experience and current salary. Such that, for every 1 month increase in `Previous Experience (months)`, `Current Salary` decreases by "-10.9071". However, as there isn't an "*" next to its line of results, it implies that its p-value is greater than 0.05 which means there isn't sufficient evidence to reject H0 and we cannot accept that `Previous Experience (months)` has a negative linear relationship with `Current Salary`.

Structured Qn1; Part c



```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -4139.2377  4203.3582  -0.985   0.3272
`Beginning Salary`               1.7302     0.1138  15.203   <2e-16 ***
`Previous Experience (months)`  -10.9071     7.7710  -1.404   0.1637
`Education (years)`             719.1221   351.7339   2.045   0.0436 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7791 on 96 degrees of freedom
Multiple R-squared:  0.8031,	Adjusted R-squared:  0.7969
F-statistic: 130.5 on 3 and 96 DF,  p-value: < 2.2e-16
```

Short Answer Question:

After running her regression model, Stacey concluded that her model is not good in explaining current salary. Do you agree with her? Please support your answer with relevant results from the regression output. [3 marks]

Common Mistakes:

- not referring to the Goodness of fit statistics
- interpreted only F statistic or R square and not both

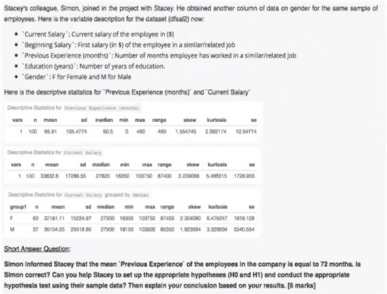
Example of a student's answer that got full credits

H0: All betas are 0
H1: At least one of the betas is nonzero
Since the p-value is 2.2e^-16 < 0.05, which is very small, and the f-statistic is large, we can reject the null hypothesis that the model has no predictive power and we can accept H1 and conclude that the model is useful and has predictive power.

Multiple R-Squared of 0.8031 implies that the model explains 80.31% of the variance of Y(Current salary)
Adjusted R-Squared of 0.7969 explains the variance of Y (Current salary) but it accounts for the adjustments for the number of variables in the model as byadding more variables will always increase R^2 , adjusted R^2 provides a penalty for the number of variables in the model)

Therefore, I do not agree with Stacey that the model is not good in explaining current salary.

Structured Qn2; Part a



Stacey's colleague, Simon, joined in the project with Stacey. He obtained another column of data on gender for the same sample of employees. Here is the variable description for the dataset (dfsal2) now:

- `Current Salary`: Current salary of the employee in ($)
- `Beginning Salary`: First salary (in $) of the employee in a similar/related job
- `Previous Experience (months)`: Number of months employee has worked in a similar/related job
- `Education (years)`: Number of years of education.
- `Gender`: F for Female and M for Male

Here is the descriptive statistics for `Previous Experience (months)` and `Current Salary`

Descriptive Statistics for `Previous Experience (months)`

vars	n	mean	sd	median	min	max	range	skew	kurtosis	se
1	100	95.61	105.4774	60.5	0	480	480	1.554745	2.082174	10.54774

Descriptive Statistics for `Current Salary`

vars	n	mean	sd	median	min	max	range	skew	kurtosis	se
1	100	33632.6	17289.55	27825	16350	103750	87400	2.239068	5.466215	1728.955

Descriptive Statistics for `Current Salary` grouped by `Gender`

group1	n	mean	sd	median	min	max	range	skew	kurtosis	se
F	63	32181.11	15224.67	27300	16350	103750	87400	2.304080	6.475047	1918.128
M	37	36104.05	20318.80	27900	18150	103500	85350	1.923094	3.329604	3340.554

Short Answer Question:

Simon informed Stacey that the mean `Previous Experience` of the employees in the company is equal to 72 months. Is Simon correct? Can you help Stacey to set up the appropriate hypotheses (H0 and H1) and conduct the appropriate hypothesis test using their sample data? Then explain your conclusion based on your results. [6 marks]

Common Mistakes:

- not conducting the hypothesis test
- incorrect computation of test statistic

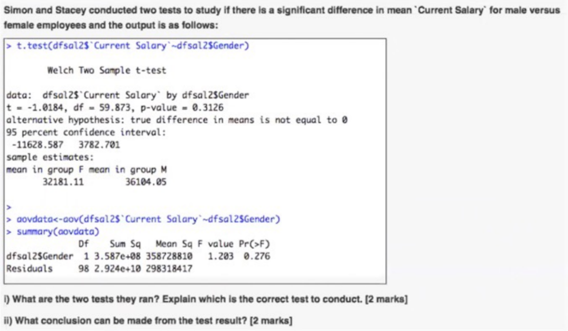
Example of a student's answer that got full credits

H0: the mean previous experience of employees equals to 72
H1: the mean previous experience of employees does not equal to 72
set significance level of 95%

conduct one sample 2-tail T test
t-score: (95.61-72)/(105.4774/sqrt(100)) = 2.238394
qt(0.025, 99, lower.tail = FALSE) = 1.984217
2.238394 > 1.984217

since t-score of hypothesis is greater than 97.5% of the probabilities, it falls into reject region. So we reject H0.
conclusion, the mean previous experience of employees does not equal to 72

Structured Qn2; Part b



Simon and Stacey conducted two tests to study if there is a significant difference in mean `Current Salary` for male versus female employees and the output is as follows:

```
> t.test(dfsal2$`Current Salary`~dfsal2$Gender)

	Welch Two Sample t-test

data:  dfsal2$`Current Salary` by dfsal2$Gender
t = -1.0184, df = 59.873, p-value = 0.3126
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11628.587   3782.701
sample estimates:
mean in group F mean in group M 
       32181.11        36104.05 

>
> aovdata<-aov(dfsal2$`Current Salary`~dfsal2$Gender)
> summary(aovdata)
              Df    Sum Sq   Mean Sq F value Pr(>F)
dfsal2$Gender  1 3.587e+08 358728810   1.203  0.276
Residuals     98 2.924e+10 298318417
```

i) What are the two tests they ran? Explain which is the correct test to conduct. [2 marks]

ii) What conclusion can be made from the test result? [2 marks]

Common Mistakes:

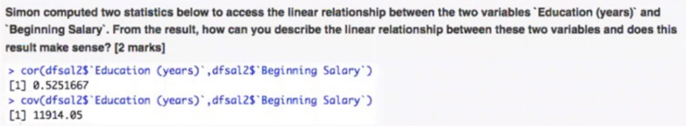
- did not indicate correct test and why in i

Example of a student's answer that got full credits

i) 2 sample t test and ANOVA test.
2 sample t test is correct. ANOVA is used to compare (means of) more than 2 samples.

ii) the p value of 2 sample t test is greater than 0.05, so we cannot reject H0. it means the mean of current salary for male versus female does not have significant difference.

Structured Qn2; Part c



Simon computed two statistics below to access the linear relationship between the two variables `Education (years)` and `Beginning Salary`. From the result, how can you describe the linear relationship between these two variables and does this result make sense? [2 marks]

```
> cor(dfsal2$`Education (years)`,dfsal2$`Beginning Salary`)
[1] 0.5251667
> cov(dfsal2$`Education (years)`,dfsal2$`Beginning Salary`)
[1] 11914.05
```

Common Mistakes:

- did not answer the second part on whether the positive linear relationship makes sense
- some mentioned that cor show small positive linear relationship but cov shows large positive linear relationship

Example of a student's answer that got full credits

The correlation coefficient is 0.5251667 which suggest a moderately strong positive linear relationship between the two variables. The positive covariance indicates a positive linear relationship between the two variables. I would think it makes sense as the more years you spent being educated, the higher qualification you attain and hence you attract a higher beginning salary.