Name: _____ **Class:**

Class #: _____ **Section #:** _____

Instructor: Ralf Becker **Assignment:** Regression Inference

Question 1: (0 points)

Simple Regression: Recap

In Correlation and Regression section (part of Descriptive Statistics), we saw how the relationship between two variables can be described by using scatter plots to provide a picture of the relationship and correlation coefficients to provide a numerical measure. In many economic and business problems, a specific functional relationship is needed. For instance,

- A manager would like to know what mean level of sales can be expected if the price is set at \pounds 15 per unit.
- If 300 workers are employed in a factory, how many units can be produced during an average day?

In many cases, we can adequately approximate the desired functional relationships by a linear equation as explained in the following sections. There are three reasons why we may wish to formalise a (linear) relationship

- We may want to quantify the relationship between variables.
- We may then want to use such relationships for predicting outcomes (see the above two examples).
- We may want use such models to decide whether there are causal relationships between variables.

The last of these is actually very difficult to achieve and we will not be able to touch on this in the context of this unit. It remains true, that correlation measures and regression models are, in the first place, merely descriptive statistics. Any correlations described in such models do not automatically reflect causal relationships.

Question 2: (1 point)

Linear Regression Model

We first look at what we have studied before. We used the relationship $res_i = y_i - a - bx_i$ to calculate a residual. Re-arranging this delivers the equation which we will typically use to describe a regression relationship:

$$y_i = a + bx_i + res_i$$

where a and b represented the intercept and slope coefficients for a particular line of best fit arising from a particular sample (which is why we call a and b sample estimates - but more on this in the inference section of the course). If we want to write this relationship in a general way, i.e. not specialised on a particular sample, but for the population of our data, then we write:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

Here we have replaced the a and b with α and β and the residuals res_i with ϵ_i. α and β represent the unknown values of the intercept and slope parameters that describe a liner relationship between y_i and x_i in the population. The error terms ϵ_i now represent error terms acknowledging that, even in the population, the linear relationship will not precisely represent the data.

Only once we have a sample of data we are able to find a line of best fit (described by a and b). We should note that this line (and hence the a and b values) is unique to the particular sample. A slightly different sample would have delivered different values for a and b. (Note: This is not unique to a regression relationship. Assume you have a population of values for some random variable m_i with an unknown mean μ_m. Then you take a sample of values from this population and obtain a sample mean, $\bar{m}$. Had you taken a different sample, this $\bar{m}$ would also be different.)

When undertaking a regression analysis this is not the case. The variables on the left hand side and the right hand side have different functions and therefore we call them by different names, such as dependent variable (on the left) and explanatory variable (on the right).

$$\underset{\substack{\text{dependent variable}\\ \text{explained variable}\\ \text{outcome variable}}}{y_i} = \alpha + \beta \underbrace{x_i}_{\substack{\text{independent variable}\\ \text{explanatory variable}}} + \epsilon_i$$

We know that we should use our economic knowledge to decide which variable is dependent/explained and which is independent/explanatory.

Let's continue thinking about the relationship between height and weight. This height-weight example is fairly obvious, the weight of a person, Y, can be modeled as a linear function of the height, X. Consider a person of specific height, x_i, then that person's weight y_i can be seen as a function of that height as long as you recognise that there is an error term for individual variation. In fact, when thinking about regression it is more useful to think about groups of people rather than individuals. Consider all people with a specific height (e.g. $x = 179cm$), then the average weight of these people can be seen as a function of that height.

In the real world we know there are other factors that influence the height. These include identifiable factors, such as the age, gender and weight of parents. There are also behavioural factors such as nutrition and exercise regime. In addition, there are other unknown factors that can influence the weight.

In a simple linear equation we model the effect of all factors, other than the X variable, in this example height, are assumed to be part of the random error term, labeled as ϵ_i. This random error term is a random variable we assume to have mean 0. This is a rather crucial assumption. The error terms are unobserved and hence this is an assumption! In further econometrics units you will talk a lot about this assumption and discuss the (many) reasons why this assumption may fail.

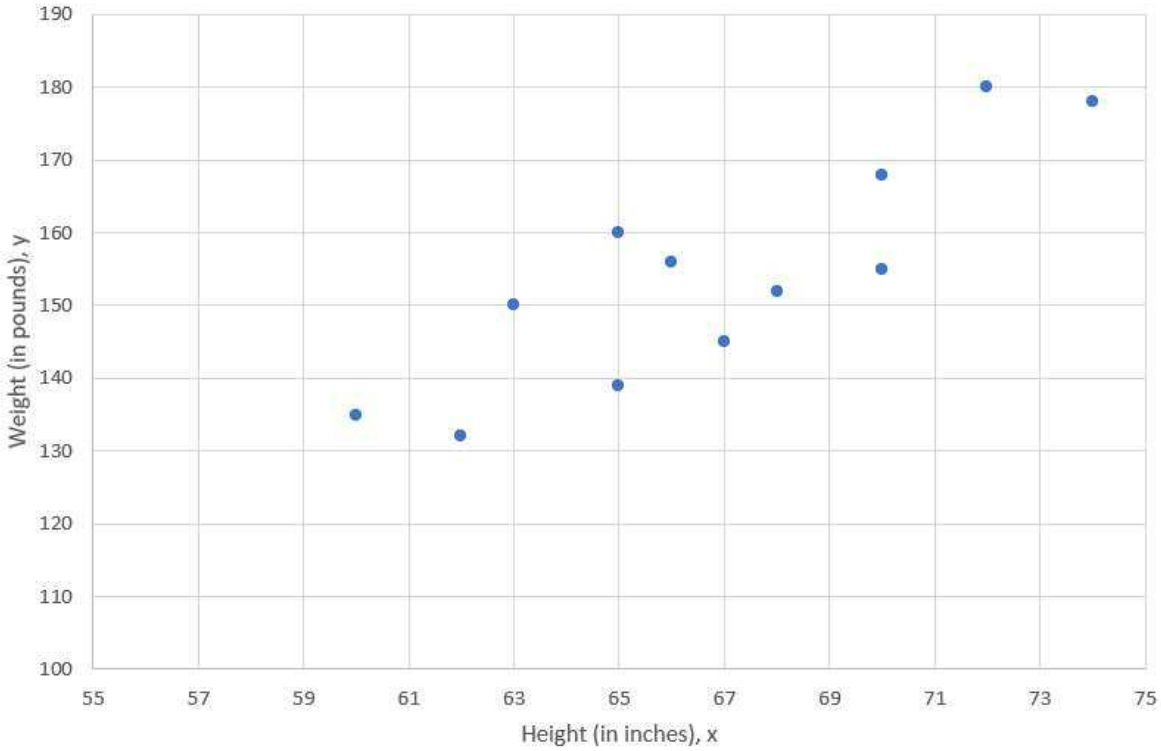
In fact, for some of the statistical inference to work below we may have to make further assumptions regarding the error terms. For instance that they are normally distributed and have constant variance. These assumptions tend to be more benign and we can deal with situations in which they are not met. Again, details here are beyond this particular course.

Thus, the model is as follows:

$$\underset{\text{Weight}}{y_i} = \alpha + \beta \underset{\text{Height}}{x_i} + \underset{\text{Other Factors}}{\epsilon_i}$$

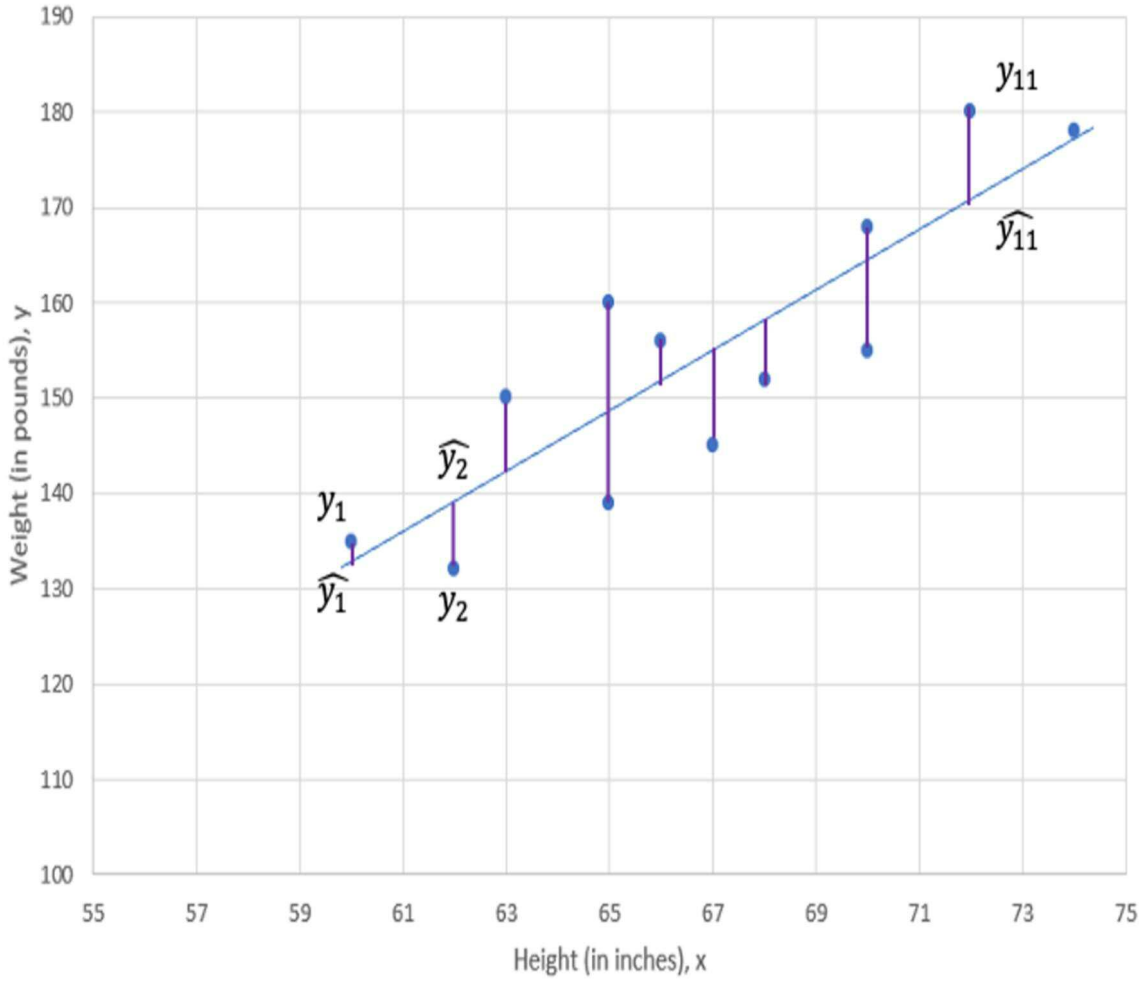
Estimation

The population regression line is a useful theoretical construct, but for applications we need to determine an estimate of the model using available data. Regression analysis is the statistical technique that finds the optimal values for α and β given a particular sample. We return to the example of 12 observations with height (taking the role of x) and weight (in this example representing y) data.



We would like to find the straight line that best fits these points, which is called line of best fit. Let's now calculate the actual sample estimates a and b. Do not forget these values will be specific to that particular sample of 12 observations. We will continue to not know what the true values α and β are and, if we had a different sample we would get somewhat different values for a and b.

Before we do so we want to point out what optimal means in this context. In fact it implies that we want to minimise (in a particular sense!) the values for res_i for all $i = 1, \ldots, n$ observations in the sample.



In fact what we want to minimise is the \emph{sum of squared residuals} and this is called least squares procedure.

As shown in Figure 2, there is a deviation between the observed value, y_i, and the predicted value, $\hat{y}_i$ (which is the value on the regression line), in the estimated regression equation for each value of X, where $y_i - \hat{y}_i$. The term we want to minimise is ($(y_i - \hat{y}_i)^2$). As we have $i = 1, \ldots, n$ of these terms, what we really wish to minimise is the sum of these squared terms:

$$SSE = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots + (y_{12} - \hat{y}_{12})^2 = \sum_{i=1}^{n} res_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

The estimates a and b are those values in

$$\hat{y}_i = a + bx_i$$

that minimise the SSE. This is where the name least squares comes from. Often you will find the term Ordinary Least Squares (OLS) where the ordinary comes from the fact that we are fitting a linear model.

This is equivalent to saying that we want to minimise the variation of our sample observations around the regression line ($a + bx$). Or better, we want to place the line of best fit such that the resulting variation is minimised.

We use differential calculus to obtain the coefficient estimators that minimise SSE (you have a function, you have two coefficients which you can vary - a and b - you know how to do that). The resulting coefficient estimator is as follows:

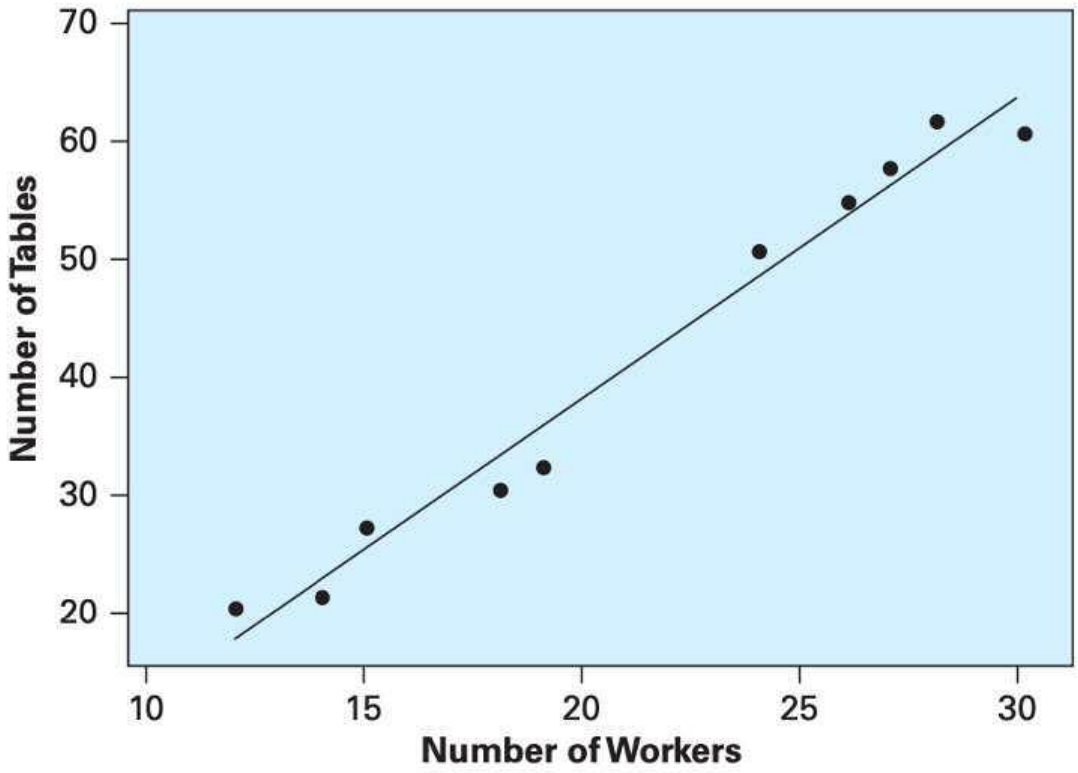
$$
\begin{aligned}
b &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{Cov(x,y)}{Var(x)} \\
&= r\frac{sy}{sx}
\end{aligned}
$$

Note that the numerator of the estimator is the sample covariance of X and Y and the denominator is the sample variance of X. The constant or intercept estimator is as follows:

$$
a = \bar{y} - b\bar{x}
$$

So far this was all a repetition of what was done in the descriptive statistics section.

The Rising Hills Manufacturing Company in Redwood Falls regularly collects data to monitor its operations. These data are stored in the data file Rising Hills. The number of workers, X, and the number of tables, Y, produced per hour for a sample of 10 workers is shown in Figure 3. If management decides to employ 25 workers, estimate the expected number of tables that are likely to be produced.



Using the data, we computed the descriptive statistics:
Cov(x,y)=106.93,
$s_x{}^2$=42.01, $\bar{y}$=41.2, and $\bar{x}$=21.3

From the covariance we see that the direction of the relationship is positive.

You specify the following linear regression model

$$
y = \alpha + \beta x + residuals
$$

Use the descriptive statistics, to compute the sample estimates for the regression coefficients:

$b =$ _____

$a =$ _____

From this, the sample regression line is as follows:

_____ = _____ + _____ $\cdot x$

For 25 employees we would expect to produce how many tables

$E(y|x = 25) =$ _____

For which of the following values of x would you feel comfortable using this estimated model to predict the output?

(a) $x = 0$

(b) $x = 10$

(c) $x = 20$

(d) $x = 30$

(e) $x = 40$

(f) $x = 50$

(g) $x = 100$

(h) $x = 500$

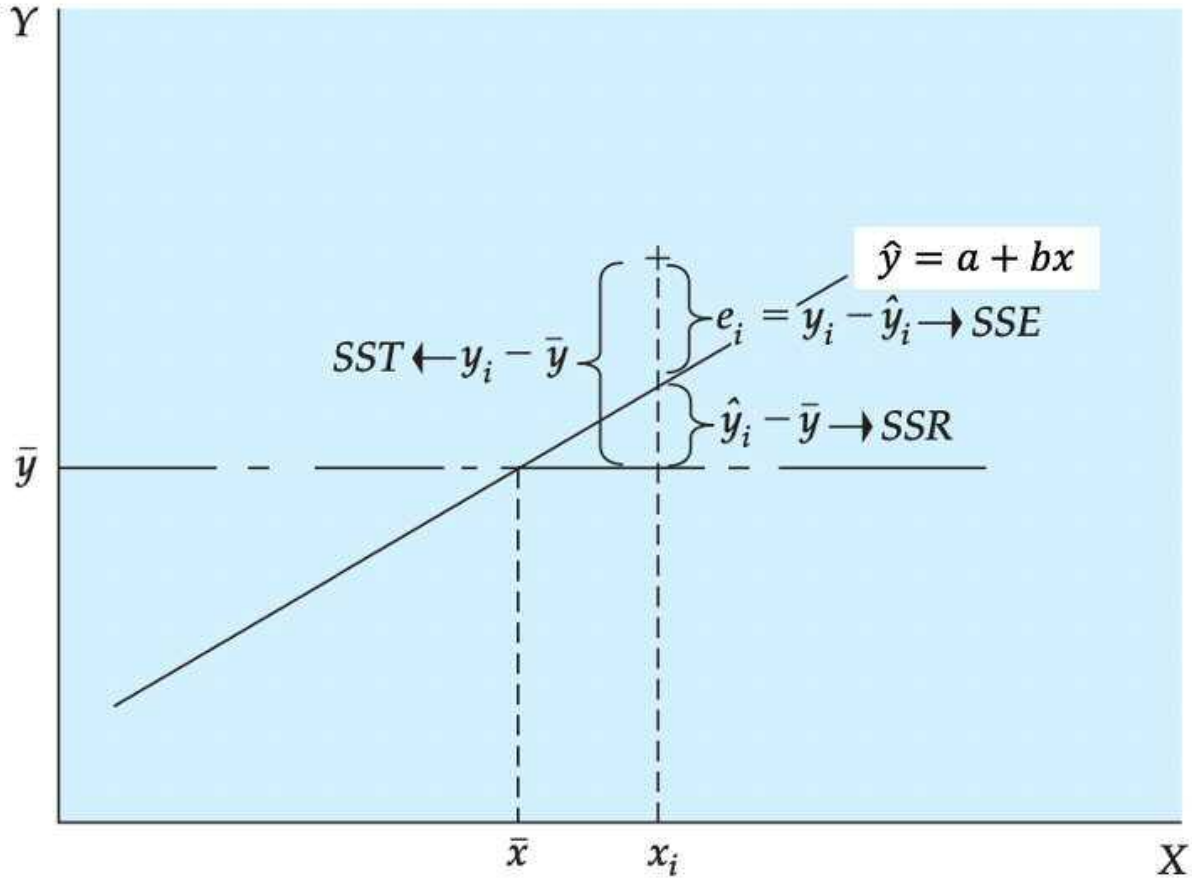
Question 3: (0 points)

Goodness of Fit

Now we are ready to develop measures that indicate how effectively the variable X explains the behaviour of Y. In our height-weight example shown in Figure 2, Weight, Y, tend to increase with Height, X, and, thus, Height explains some of the differences in Weight. The points, however, are not all on the line, so the explanation is not perfect. Here, we develop measures based on the partitioning of variability that measure the capability of X to explain Y in a specific regression application.

The analysis of variance, ANOVA, for least squares regression is developed by partitioning the total variability of Y into an explained component and an error component. In Figure 4 we show that the deviation of an individual Y value from its mean can be partitioned into the deviation of the predicted value from the mean and the deviation of the observed value from the predicted value:

$$y_i - \bar{y} = (\widehat{y_i} - \bar{y}) + (y_i - \widehat{y_i})$$



We square each side of the equation because the sum of deviations about the mean is equal to 0 and sum the result over all n points:

$$\sum_{i=1}^{12}(y_i - \bar{y})^2 = \sum_{i=1}^{12}(\widehat{y_i} - \bar{y})^2 + \sum_{i=1}^{12}(y_i - \widehat{y_i})^2$$

Some of you may note the squaring of the right-hand side should include the cross product of the two terms in addition to their squared quantities. It can be shown that the cross- product term goes to 0. This equation is expressed as follows:

$$SST = SSR + SSE$$

Here, we see that the total variability SST can be partitioned into a component SSR that represents variability that is explained by the slope of the regression equation. (The mean of Y is different at different levels of X.) The second component SSE results from the random or unexplained deviation of points from the regression line. This variability provides an indication of the uncertainty that is

associated with the regression model. In short, the total variability in a regression analysis, SST, can be partitioned, analysis of variance, into a component explained by the regression, SSR, and a component due to unexplained error, SSE. The components defined as follows:

$$\text{Sum of Squares Total:} SST = \sum_{i=1}^{12} (y_i - \bar{y})^2$$
$$\text{Sum of Squares Regression:} SSR = \sum_{i=1}^{12} (\widehat{y_i} - \bar{y})^2$$
$$\text{Sum of Squares Error:} SSE = \sum_{i=1}^{n} res_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

The ratio of the sum of squares regression, SSR, divided by the total sum of squares, SST, provides a descriptive measure of the proportion, or percent, of the total variability that is explained by the regression model. This measure is called the coefficient of determination—or, more generally, R^2:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

The coefficient of determination is often interpreted as the percent of variability in Y that is explained by the regression equation. This quantity varies from 0 to 1, and higher values indicate that the variation in the explanatory variables explain a larger proportion of the variation in the explained variable. Caution should be used in making general interpretations of R^2 because a high value can result from either a small SSE, a large SST, or both.

The coefficient of determination, R^2, only for simple regression is equal to the simple correlation squared:

$$R^2 = r^2$$

This provides an important link between correlation and R^2, the regression model.

Question 4: (0 points)

Excel application

With this background let us return to our height-weight example with data and look at how we use the partitioned variability to determine how well our model explains the process being studied. In most situations we use a spreadsheet such as Excel to obtain the regression coefficients to reduce the work load and improve computational accuracy. Figure 5 shows the EXCEL output of the height-weight example.

Video walkthrough of how to perform a simple regression estimation in EXCEL

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.863234
R Square	0.745174
Adjusted R Sc	0.719691
Standard Erro	8.232576
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1981.914	1981.914	29.24242	0.000298182
Residual	10	677.753	67.7753		
Total	11	2659.667			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-60.7461	39.81353	-1.52576	0.158053	-149.4561561	27.96398	-149.456	27.96398
X Variable 1	3.215652	0.594651	5.407626	0.000298	1.890686294	4.540618	1.890686	4.540618

From this EXCEL regression output, we could get some key information:

- R square statistic is 0.745174 so $R^2 = 0.745174$, which we can see that here 74.5\% of the variation is explained by that linear relationship.

- The sample size (Observations) of 12.
- Sum of Squared Regression (SSR) is 1981.914, which is $\sum_{i=1}^{12}(\widehat{y_i}-\bar{y})^2$.
  It shows The amount of variability explained by the regression equation.
- Sum of Squared Residuals (SSE) is 677.753, which is $\sum_{i=1}^{12}(y_i-\widehat{y_i})^2$. \\It indicates the smallest possible sum of squared residuals is 677.753.
- Sum of Squares Total (SST) is 2659.667, which is $\sum_{i=1}^{12}(y_i-\bar{y})^2$. \\It measures the amount of variation we can see in the dependent variable.
- Note that $SST = SSR + SSE$
- The intercept estimator a is -60.7461
- The coefficient estimator b is 3.2157
- The regression line (line of best fit) is therefore defined as: $\hat{y} = -60.7461 + 3.2157 \cdot x$

Question 5: (0 points)

Interpretation of Simple Regression Equation

We return to the interpretations of simple regression equation with some key points.

First, in most cases the data available to run a regression will be sample data. Therefore, the values of a and b that describe the line of best fit, are sample estimates of some unknown population parameters (usually labeled, α and β).
Second, note that b is the slope of the fitted line, $\hat{y} = a + bx$; i.e., the derivative of $\hat{y}$ with respect to x:

$$b = d\hat{y}/dx$$

and measures the increase in $\hat{y}$ for a unit increase in x. For the height-weight example, the regression line is defined as

$$\hat{y} = -60.7461 + 3.2157\, x$$

We know that when we interpret regression results, we should be aware of the units in which the dependent and explanatory variables are measured.

In the Height-Weight example, the explanatory variable (Height, x) is measured in inches (1 inch = 2.54 cm) and the dependent variable (Weight, y) is measured in pounds (1 pound = 1 lbs = 0.454 kg).

$$\underset{\text{Weight, in pounds}}{\hat{y}} = -60.7461 + 3.2157 \underset{\text{Height, in inches}}{x}$$

With this knowledge, the interpretations of $b = 3.2157$ are

- "The expected weight ($\hat{y}$) increases by 3.2157 pounds for every height increase of 1 inch."
- "On average weight ($\hat{y}$) increases by 3.2157 pounds for every height increase of 1 inch."

Interpreting the intercept only makes sense if the value of $x = 0$ is a sensible value and inside the sample of values of x. In this Height-Weight example, $a = -60.7491$ and it means that if someone has a height of 0 inches then we would expect the person to have a weight of -60.7491. This does not make any sense here.

In addition, we know that transformations of data can affect the above summary measures. Originally, height is measured in inches (1 inch = 2.54 cm) and our original model is

$$Weight[lbs]_i = \alpha + \beta\, Height[in]_i + \epsilon_i$$

If the height is measured in centimeters then the centimeter model is

$$\begin{aligned} Weight[lbs]_i &= \gamma + \delta Height[cm]_i + v_i \\ Weight[lbs]_i &= \gamma + \delta \cdot 2.54 \cdot Height[in]_i + v_i \end{aligned}$$

where $\beta = 2.54 \cdot \delta$. And indeed, the original coefficient estimated (in the inches model) is 2.54 times larger than that estimated in the centimeter model.

Question 6: (1 point)

Statistical Inference: Hypothesis Tests and Confidence Intervals

Now that we have developed the coefficient estimators, we are ready to perform inference on the population model parameters. The basic approach follows that developed in hypothesis test and confidence interval. We use the estimated parameters to test hypotheses on the unknown population parameters or alternatively we calculate confidence intervals. We do that as we are typically interested in the unknown population parameters and not the particular sample parameter estimates.

Hypothesis Tests

The true population regression line is

$$y_i = \alpha + \beta x_i + \epsilon_i$$

We obtain sample estimates for the unknown population parameters α and β. The sample estimates are a and b which then define the sample regression line

$$\hat{y}_i = a + b x_i$$

As usual we recognise that a different sample would have given us different sample estimates. This is why we need to use statistical inference techniques.

In applied regression analysis, we first usually wish to know if there is a relationship. In the regression model we see that if β is 0, then there is no linear relationship between X and Y. To determine if there is a linear relationship, we can test the hypothesis

$$H_0 : \beta = 0$$

versus

$$H_0 : \beta \neq 0$$

It turns out that inference on regression coefficients works very much like inference on a population mean with unknown population variance. In that case we used a T test

$$t = \frac{\bar{x} - \mu}{SE(\bar{x})} \sim t_{n-1}$$

this test statistic was t distributed with $n-1$ degrees of freedom if the population distribution of the error terms was normal. If it was not normal then we could approximate the distribution of T with a standard normal distribution if the sample size was big enough to invoke a CLT.

In the context of this simple regression the test statistic is defined as

$$t = \frac{b - \beta}{SE\,(b)} \sim t_{n-2}$$

where the test statistic is distributed according to a Student's t distribution with $(n-2)$ degrees of freedom. This is a bit different from what you have studied in hypothesis test section. Degree of freedom is $(n-2)$ instead of $(n-1)$ because the simple regression model uses two estimated parameters, a and b, instead of one ($\bar{x}$ when we are testing for μ). $SE\,(b)$ represents the standard error of slope coefficient b. Here we will not provide any detail on how to calculate this, we will rely on EXCEL (or any other software) to calculate this. In the image below you can see where to find the standard errors $SE(a) = 39.8135$ and $SE(b) = 0.5947$.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.86323443
R Square	0.74517369
Adjusted R Sc	0.71969106
Standard Erro	8.2325758
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1981.913623	1981.91362	29.2424157	0.00029818
Residual	10	677.7530435	67.7753043		
Total	11	2659.666667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-60.746087	39.81352824	-1.52576498	0.15805315	-149.456156	27.9639822	-149.456156	27.9639822
Height(inches	3.21565217	0.594651397	5.4076257	0.00029818	1.89068629	4.54061805	1.89068629	4.54061805

Above we indicated that we want to find if there is a linear relationship, say if β is 0. In general, we can test for any specific values β_0.

- To test the null hypothesis $H_0 : \beta \leq \beta_0$ against the alternative $H_A : \beta > \beta_0$ the decision rule is as follows:
  Reject H_0 if $t = \dfrac{b - \beta}{SE\,(b)} > t_{n-2,\alpha}$
- To test either null hypothesis $H_0 : \beta \geq \beta_0$ against the alternative $H_A : \beta < \beta_0$ the decision rule is as follows:
  Reject H_0 if $t = \dfrac{b - \beta}{SE\,(b)} \leq t_{n-2,\alpha}$
- To test the null hypothesis $H_0 : \beta = \beta_0$ against the alternative $H_A : \beta \neq \beta_0$ the decision rule is as follows:
  Reject H_0 if $t = \dfrac{b - \beta}{SE\,[b]} > t_{n-2,\alpha/2}$ or $t = \dfrac{b - \beta}{SE\,[b]} < t_{n-2,\alpha/2}$

Hypothesis tests could also be performed on the equation constant, α follow the same logic and we would use the following t-test

$$t = \frac{a - \alpha_0}{SE\,(a)} \sim t_{n-2}$$

To facilitate the application of inference to regression results, these are often presented as follows:

$$\underset{}{\hat{y}} = \underset{(39.8135)}{-60.7461} + \underset{(0.5946)}{3.2157} \cdot x$$

The standard errors of the coefficient estimates are shown in parenthesis underneath the respective sample estimate.

Let's test the hypothesis that, on average the, an additional inch of height increases weight by not more than 3 pounds at $\alpha = 0.05$.

$$H_0 : \beta \leq 3$$
$$H_A : \beta > 3$$

The test statistic is $T = \dfrac{b - \beta_0}{SEb} \sim t_{n-2}$. The decision rule is to reject H_0 if the sample test statistic is greater than 1.812 (from the t-table with 10 degrees of freedom).

$$t = \frac{3.2157 - 3}{0.5946} = 0.3627$$

As the test statistic is not larger than the critical value the null hypothesis is not rejected. The sample does not provide sufficient evidence to reject H_0.

Test the hypothesis that, on average the, an additional inch of height increases weight by not more than 2 pounds at $\alpha = 0.01$.

$H_0 : \beta$ _____ _____
$H_A : \beta$ _____ _____

The test statistic is:

(a) $T = \frac{b - \beta_0}{SE(b)} \sim t_{n-1}$

(b) $T = \frac{b - \beta_0}{SE(b)} \sim t_{n-2}$

(c) $T = \frac{\beta_0 - b}{SE(b)^2} \sim t_{n-2}$

(d) $T = \frac{b - \beta_0}{SE(b)} \sim N(0,1)$

(e) $T = \frac{b - \beta_0}{SE(b)^2} \sim t_{n-2}$

Decision Rule:

Reject _____ if _____ is _____ than _____ .

The sample test statistic is

$t =$ _____

As the test statistic is not larger than the critical value the null hypothesis is not rejected. The sample does not provide sufficient evidence to reject H_0 at $\alpha = 0.01$.

Question 7: (1 point)

Confidence Interval

You may not be interested in actually testing a particular hypothesis but merely to communicate that your sample estimate will not tell you exactly what the unknown population parameter is. The tool of choice is to present a confidence interval, say, for the slope coefficient β of the population regression line. A $100(1-\alpha)\%$ confidence interval for the population regression slope β is given by:

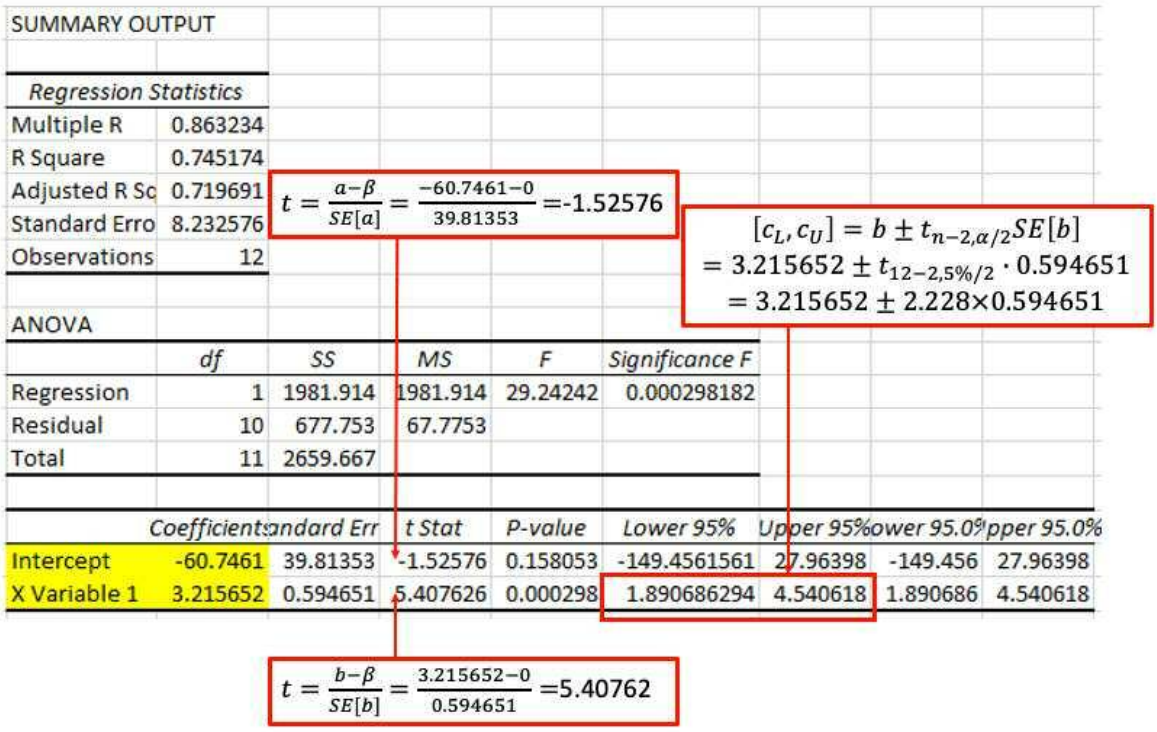
$$[c_L, c_U] = b \pm t_{n-2,\alpha/2} SE[b]$$

where $t_{n-2,\alpha/2}$ is the number for which

$$P(t_{n-2} > t_{n-2,\alpha/2}) = \alpha/2$$

and the random variable t_{n-2} follows a Student's t distribution with $(n-2)$ degrees of freedom. As in the case of the hypothesis test for a regression coefficient, the degrees of freedom is determined by the number of observations minus the number of estimated coefficients, here 2.

Excel's regression output for the Height-Weight example, is replicated below with annotations.



From that regression output you get all the ingredients you need to calculate hypothesis tests on and confidence intervals for the unknown population parameters. In fact, by default the output does provide a particular hypothesis test and a particular confidence interval. The hypothesis test provided is that of $H_0 : \beta = 0$ against $H_A : \beta \neq 0$ (or the equivalent for α) and a 95% confidence interval.

Note that the t-statistics and the p-values are specific to the $H_0 : \beta = 0$ against $H_A : \beta \neq 0$ test, but you also have all the ingredients you would need to calculate a hypothesis test with a different null hypothesis or a different confidence interval.

Let's calculate a 99% confidence interval for β.

$$\begin{aligned} b \pm & \ t_{n-2,\alpha/2} SE[b] \\ 3.2156 \pm & \ 3.169 \cdot 0.5947 = 3.2156 \pm 1.8846 \\ & \Rightarrow [1.3311, 5.1003] \end{aligned}$$

Calculate a 80% confidence interval for α.

_____ _____ _____ _____

$\Rightarrow$[_____ , _____]

Note that the 80% confidence interval does not contain 0. This is equivalent to the the p-value being smaller than 20% and hence we would reject a two-sided hypothesis test with null hypothesis $H_0 : \alpha = 0$ at $\alpha = 0.2$.

Question 8: (0 points)

Multiple Regression

So far, we developed simple regression as a procedure for obtaining a linear equation that predicts a dependent or endogenous variable as a function of a single independent or exogenous variable for example, weight as a function of height. However, in many situations, you would want several explanatory variables to jointly influence a dependent variable. Multiple regression enables us to determine the simultaneous effect of several independent variables on a dependent variable using the least squares principle.

Many important applications of multiple regression occur in business and economics. These applications include the following:

- The quantity of goods sold is a function of price, income, advertising, price of substitute goods, and other variables.
- Salary is a function of experience, education, age, and job rank.
- Capital investment occurs when a business person believes that a profit can be made. Thus, capital investment is a function of variables related to the potential for profit, including interest rate, gross domestic product, consumer expectations, disposable income, and technological level.

Business and economic analysis has some unique characteristics compared to analysis in other disciplines. Some natural scientists work in a laboratory, where many—but not all—variables can be controlled. If you wanted to figure out what the impact of one variable is you could vary that one (and keep all others constant) and then observe the impact these changes have on an outcome variable. In contrast, the economist's and manager's laboratory is the world, and conditions cannot be controlled. Thus, we need tools such as multiple regression to estimate the simultaneous effect of several variables. Multiple regression as a "lab tool" is very important for the work of managers and economists.

The "miracle" of multiple regression is going to be that it will be able to "partial" out effects of individual explanatory variables.

Model Specification

Model specification includes selection of the explanatory variables and the functional form of the model. We return to the Height-Weight example but we will discuss more than one factor this time. We previously specified:

$$\underset{\text{Weight}}{y_i} = \alpha + \beta \underset{\text{Height}}{x_i} + \underset{\text{Other Factors}}{\epsilon_i}$$

In a simple linear equation we model the effect of all other factors to be part of the random error term, labeled as ϵ_i. However, some factors in error term could be measured as well so we can move them out from the error term and include them as explanatory variables. As we have discussed, weight could be affected by gender.

Linear regression assumes that the numerical amounts in all independent, or explanatory, variables are meaningful data points. However, sex is a binary categorical variable (this is mainly true, sex referring to biological attributes. Gender identification used to be modelled as a binary variable as well, but today there is a better understanding that the gender people identify as should not be modelled as a binary variable only). Regression analysis requires numerical values to work with and therefore sex has to be coded. Categorical variables with two categories is coded as a dummy variable. A dummy variable is a variable created to assign a numerical value to the levels of the categorical variable. Here we will define such a variable, called "male" which takes the value 1 if the individual is a male and 0 if it is not. Here we have a categorical variable which takes one of two categories and we needed one variable to code this up. In general, if you have a categorical variable with k categories then you need $k-1$ categories to code these up.

This would lead us to specify a multiple regression model:

$$\underset{\text{Weight}}{y_i} = \alpha + \beta_1 \underset{\text{Height}}{x_{1i}} + \beta_2 \underset{\text{Male}}{x_{2i}} + \underset{\text{Other Factors}}{\epsilon_i}$$

Now we have an intercept coefficient (α) and two slope coefficients, β_1 and β_2 which relate to the Height and Male variables respectively. These are all unknown population coefficients.

The following table shows the data with the sex and the derived male variable.

Weight(Y, pounds)	Height(inches)	Sex	male
155	70	Female	0
150	63	Female	0
180	72	Male	1
135	60	Female	0
156	66	Female	0
168	70	Male	1
178	74	Male	1
160	65	Female	0
132	62	Female	0
145	67	Female	0
139	65	Female	0
152	68	Female	0

Question 9: (0 points)

Estimation

In general a researcher is interested in the unknown intercept coefficient (α) and two the two slope coefficients, β_1 and β_2. We will use sample data to obtain sample estimates for these. Multiple regression coefficients are computed using estimators obtained by the least squares procedure. This least squares procedure is similar to that presented in simple regression. However, the estimators are complicated by the relationships between the explanatory X_i variables.

The least squares procedure for multiple regression computes the estimated coefficients so as to minimise the sum of the residuals squared. In this case, the sample estimated regression equation is

$$\hat{y}_i = a + b_1 x_{1i} + b_2 x_{2i}$$

The residual is:

$$res_i = y_i - \hat{y}_i = y_i - (a + b_1 x_{1i} + b_2 x_{2i}) = y_i - a - b_1 x_{1i} - b_2 x_{2i}.$$

Formally, we minimise SSE:

$$SSE = \sum_{i=1}^{n} res_i^2 = \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - b_2 x_{2i})^2$$

To carry out the process formally, we use partial derivatives to develop a set of simultaneous equations that are then solved to obtain the coefficient estimators. Fortunately, the complex computations are always performed using a statistical software such as EXCEL. Our objective here is to understand how to interpret the regression results and use them to solve problems. You may learn more about the actual magic that is going on in later units.

Question 10: (0 points)

Goodness of Fit

Similar to the simple regression model, we can develop a measure of the proportion of the variability in the dependent variable that can be explained by the multiple regression model. The model variability can be partitioned into the components

$$SST = SSR + SSE$$

where these components are defined as follows:

$$\text{Sum of Squares Total:} SST = \sum_{i=1}^{12} (y_i - \bar{y})^2$$

$$\text{Sum of Squares Regression:} SSR = \sum_{i=1}^{12} (\widehat{y_i} - \bar{y})^2$$

$$\text{Sum of Squares Error:} SSE = \sum_{i=1}^{n} res_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - b_2 x_{2i})^2$$

This decomposition can be interpreted as follows:

total sample variability = explained variability + unexplained variability

The coefficient of determination, R^2, of the fitted regression is defined as the proportion of the total sample variability explained by the regression

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

and it follows that

$$0 \leq R^2 \leq 1$$

Question 11: (0 points)

Interpretation of Multiple Regression Equation

Note that the multiple regression coefficients typically change as you include additional explanatory variables. For instance, when we estimated the simple regression with only height as an explanatory variable, the estimated coefficient for the height variable was 3.2156. In the multiple regression model, after including the male dummy variable, the estimated coefficient came out as 2.1073, so a reduction of approximately 50% in this case. This will always happen if the two explanatory variables (as is the case here) are correlated.

For multiple regression, we would interpret the coefficients as being the partial derivatives.

$$b_1 = \partial \hat{y} / \partial x_1$$

This is consistent with our usual idea that, as we increase x_1 by one unit $\hat{y}$ changes by b_1. But importantly, now we also have to specify that this is true if we leave x_2 unchanged.

$$b_2 = \partial \hat{y} / \partial x_2$$

When we increase x_2 by one unit and leave x_1 unchanged, $\hat{y}$ changes by b_2.

We return to the height-weight example, the estimated coefficients are identified in the EXCEL output. The regression line is defined as

$$\hat{y} = 9.9005 + 2.1073 x_1 + 13.7051 x_2$$

Make use of the information of the units in which the dependent and explanatory variables are measured.
From above we know, Height (x_1) is measured in inches (1 inch = 2.54 cm) and the Weight (y) is measured in pounds (1 pound = 1 lbs = 0.454 kg).

We know that when we interpret regression results, we should be aware of the units in which the dependent and explanatory variables are measured. With this knowledge, the interpretations of $b_1 = 2.1073$ are

- "The expected weight ($\hat{y}$) increases by 2.1073 pounds for every height increase of 1 inch, if the other variable does not change"
- "On average weight ($\hat{y}$) increases by 2.1073 pounds for every height increase of 1 inch, if the other variable does not change"

You will note here we add the sentence "if the other variable does not change" for multiple regression. Sometimes you see the term ceteris paribus ("other things being equal") to express this. This can be abbreviated as "c.p.".

Again, note that the coefficient for height is 3.2157, which is clearly different from 2.1073 from the simple regression. This is because we added the male dummy variable into the model and as height and male are correlated (males tend to be taller) the inclusion of the male dummy variable will change the coefficient of height.

When interpreting the coefficients to dummy variables we need to be less mechanical. Before interpreting b_2, let's look at the following argument.

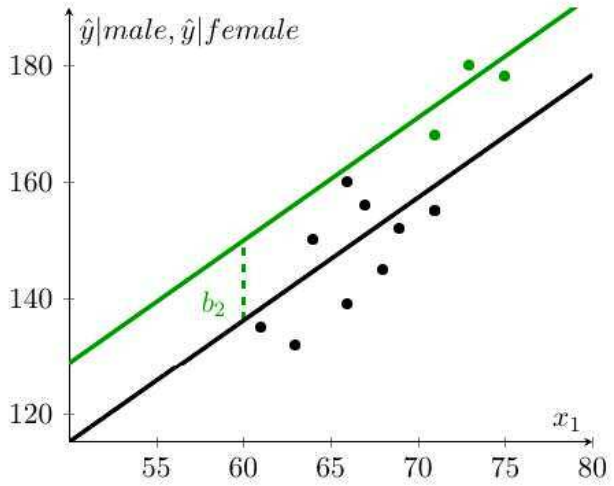
When $x_{2i} = 0$ (Female respondents), the regression specification simplifies to

$$\hat{y}_i = a + b_1 x_{1i}$$

but when $x_{2i} = 1$ (Male respondents),

$$\hat{y}_i = a + b_1 x_{1i} + b_2 \times 1 = \underbrace{a + b_2}_{\text{constant}} + b_1 x_{1i}$$

the constant is $a + b_2$. We can illustrate the two regression lines in the following plot where the data are visually separated by sex (Green = males, black = females). The fact that the two lines are parallel is enforced by the specification as we are estimating the same slope coefficient (for height, x_1) for males and females.The vertical difference between the two lines is equivalent to b_2.



Regression lines of best fit for Males and Females

We see that the dummy variable shifts the linear relationship between y_i and x_{1i} by the value of the coefficient b_2. In this way we can represent the effect of shifts in our regression equation. With this knowledge, the interpretation of $b_2 = 13.7051$ is the Male weights are, on average, 13.7051 higher than Female. When dealing with dummy variables this is the correct way to interpret regression coefficients.

.

Question 12: (1 point)

Statistical Inference: Hypothesis Tests and Confidence Intervals

We shall demonstrate how to use regression results to perform hypothesis tests and calculate confidence intervals.

Hypothesis Tests

Once you estimated a multiple regression model and obtained its regression output, performing inference on regression coefficients is really not much different to performing inference on regression coefficients in a simple regression. Let's start by re-stating the population regression model

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

We obtain sample estimates for the unknown population parameters α, β_1 and β_2. The sample estimates are a, b_1 and b_2 which then define the sample regression line

$$\hat{y} = a + b_1 x_{1i} + b_2 x_{2i} + res_i$$

As usual we recognise that a different sample would have given us different sample estimates. This is why we need to use statistical inference techniques.

Let's say we wish to test whether the respondent's sex matters for explaining their weight. We would then want to test the hypothesis

$$H_0 : \beta_2 = 0$$

versus

$$H_0 : \beta_2 \neq 0$$

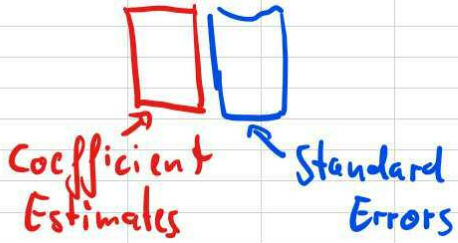
It turns out that inference on regression coefficients works very much like inference on a population mean with unknown population variance. We use a T test

$$t = \frac{b_2 - \beta_2}{SE\left(b_2\right)} \sim t_{n-3}$$

where the test statistic is distributed according to a Student's t distribution with $(n-3)$ degrees of freedom. The degree of freedom parameter is $(n-3)$ as our regression model uses three estimated parameters, a, b_1 and b_2. $SE\left(b_2\right)$ represents the standard error of slope coefficient b_2. In general the degrees of freedom are calculated from $n-k$, where k is the number of estimated coefficients. Recall that in this unit we do not cover how to calculate that standard error, we shall rely on EXCEL doing that work. Here we show again the EXCEL regression output.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.90307762
R Square	0.81554919
Adjusted R Sc	0.77456013
Standard Erro	7.38299594
Observations	12



ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	2169.08901	1084.5445	19.8967489	0.00049712
Residual	9	490.577661	54.508629		
Total	11	2659.66667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	9.90052356	52.2329906	0.18954541	0.85387124	-108.25871	128.059757	-108.25871	128.059757
Height(inches	2.10732984	0.80132226	2.62981569	0.02736692	0.29461296	3.92004672	0.29461296	3.92004672
male	13.7050611	7.39586953	1.85306961	0.09687655	-3.02555814	30.4356803	-3.02555814	30.4356803

Let us test the hypothesis that sex does not matter when explaining weight. $\alpha = 0.1$.

$$H_0 : \beta_2 = 0$$
$$H_A : \beta_2 \neq 0$$

The test statistic is $T = \frac{b_2 - \beta_{20}}{SEb_2} \sim t_{n-3}$, where β_{20} is the hypothesised value for β_2 in the null hypothesis. The decision rule is to reject H_0 if the absolute value of the sample test statistic is greater than 1.833 (from the t-table with 9 degrees of freedom, two tailed test!).

$$t = \frac{13.7051 - 0}{7.3959} = 1.8531$$

You can also find that test statistic from the regression output as EXCEL (t Stat column), by default does test exactly the hypothesis we are testing here). As the absolute value of the test statistic is larger than the critical value the null hypothesis is rejected at $\alpha = 0.1$.

Let's also calculate the p-value by referring to the t-distribution Table:

	Significance Level				
1-Tailed:	0.1	0.05	0.025	0.01	0.005
2-Tailed	0.2	0.1	0.05	0.02	0.01
k 1	3.078	6.314	12.706	31.821	63.657
9	1.383	1.833	2.262	2.821	3.250

Our test statistic is between the values of 1.833 and 2.262 which relate to the (two tailed) p-values of 0.1 and 0.05. Therefore we conclude that the p-value is between 0.05 and 0.1. If you were to use EXCEL to calculate this p-value you would use the following formula "=2*(1-T.DIST(1.8531,9,TRUE))" which delivers a precise p-value of 0.0969.

At an $\alpha = 0.1$ we conclude that the sample delivers evidence to reject the null hypothesis. Sex matters for explaining variation in weight.

When comparing this procedure to the hypothesis test we performed for a simple regression you will realise that the only difference in procedure was the degrees of freedom for the t-distribution. Otherwise everything was identical. Also, to conclude this section recall that, as this example used a small sample, we have to assume that the error terms are normally distributed in order to be able to use the t-distribution. When you use large sample sizes (justifying the application of a CLT) then we do not require this assumption.

Let us test the hypothesis that an additional inch of height, on average, adds two pounds of weight against the alternative that it adds more than two pounds. Use $\alpha = 0.01$.

$H_0 : \beta_1$ _____ _____
$H_A : \beta_1$ _____ _____

The test statistic is $T = \frac{b_1 - \beta_{10}}{SEb_1} \sim t_{dof}$, where β_{10} is the hypothesised value for β_1 in the null hypothesis.

What are the degrees of freedom for the t-distribution for this test?

$dof =$ _____

The decision rule is to reject H_0 if the value of the sample test statistic is _____ than 2.821.

The test statistic is

$t =$ _____

What is the test's p-value?

(a) p-value > 0.1

(b) $0.1 \geq$ p-value < 0.05

(c) $0.05 \geq$ p-value < 0.01

(d) p-value ≤ 0.01

Question 13: (1 point)

Confidence Interval

Presenting uncertainty about estimated coefficients we turn to confidence intervals. When we revisited hypothesis testing we realised that the only difference to hypothesis testing from a simple regression was the degrees of freedom. In fact the simple regression was just a special case of the multiple regression case. If you calculate the degrees of freedom as n minus the number of estimated coefficients you will get it right in any case.

The same applies to the calculation of confidence intervals. As long as you remember how to get the degrees of freedom this works exactly as in the case of simple regression. Let us calculate a 99% confidence interval for β_2 the coefficient for the male variable.

$$[c_L, c_U] = b_2 \pm t_{n-3,\alpha/2} SE[b_2]$$

Substituting what we get from the regression output we obtain

$$13.7051 \pm 3.250 \cdot 7.3959 = 13.7051 \pm 24.0366 \Rightarrow [-10.3315, 37.7416]$$

Calculate a 95% confidence interval for β_1 the coefficient for the height variable.

$[c_L, c_U] =$ _____ $\pm$ _____ _____

The sample confidence interval is: [_____ , _____]

Question 14: (0 points)

Worked Example and Outlook

Regression analysis is the workhorse technique of empirical economics. It is the swiss-army knife of statistical techniques. While regression analysis as it is presented here is fairly straightforward, many more complicated statistical techniques are either variations on linear regression models or use linear regression as an ingredient. So knowing regression analysis will carry you far.

We conclude this with a worked example of multiple regression analysis. This video provides a work-through a multiple regression analysis. It uses this dataset (https://manchester.mobius.cloud/web/Econ1007011/Public_Html/CountryIndicators_2019.csv (/web/Econ1007011/Public_Html/CountryIndicators_2019.csv)) of country indicators from Gapminder (https://www.gapminder.org/).