Definitions:

Pearson's correlation coefficient:

The covariance of two variables divided by the product of their standard deviations.

For a population:

p$_{x,y}$ $= \frac{\text{Cov(X, Y)}}{\sigma_X \sigma_Y}$
where
Cov(X, Y) $=$ E[(X - E[X])(Y - E[Y])]

For a sample:

It's often referred to as the sample correlation coefficient, commonly abbreviated to just "r"

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

(Above: the sample covariance divided by the product of the sample standard deviations)
which can be manipulated to get:

$$r = r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

The correlation coefficient always takes a value between -1 and 1, with 1 or -1 indicating perfect correlation (all points would lie along a straight line in this case).

a. A positive correlation indicates a positive association between the variables (increasing values in one variable correspond to increasing values in the other variable).

b. A negative correlation indicates a negative association between the variables (increasing values is one variable correspond to decreasing values in the other variable).

c. A correlation value close to 0 indicates no association between the variables.

The square of the correlation coefficient, R^2 , is a useful value in linear regression. This value represents the fraction of the variation in one variable that may be explained by the other variable. Thus, if a correlation of r $= 0.8$ is observed between two variables (say, height and weight, for example), then a linear regression model attempting to explain either variable in terms of the other variable will account for 64% (r^2 $= 0.8^2$ $= .64$) of the variability in the data.
The correlation coefficient also relates directly to the regression line Y $=$ a + bX for any two variables, where b $=$ r$\frac{s_x}{s_y}$
I found this info here:
http://www.stat.yale.edu/Courses/1997-98/101/correl.htm
and on the wikipedia page.

Regression Analysis definition

A statistical technique for modeling and investigating the relationship between variables.
The basic model is:

$$y = \beta_0 + \beta_1 x + \epsilon$$

The **response variable**, y, is the variable you're analyzing to see how much it's influenced by the other variable(s).
The **regressor variable(s)**, x, is (are) the variable(s) you're estimating regression coefficients for in order to predict future response variables.
The **regression coefficients**, β $_0$, β $_1$, ... are the coefficients for each regressor variable (and a slope, usually) that best minimize the random error (ϵ) for the model.
The **random error** term, ϵ , is the random variable that accounts for the failure of the model to fit the data exactly. For example, for a particular (x$_i$, y$_i$), the ϵ $_i$ is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where
ϵ ~N(0, σ^2)
The expected values of each of these quantities are:
E[y | x] = $\mu_{y|x}$ = E[β $_0$ + β $_1$x + ϵ] = E[β $_0$] + E[β $_1$x] + E[ϵ] = β $_0$ + β $_1$x + 0
V[y | x] = $\sigma^2{}_{y|x}$ = V[β $_0$ + β $_1$x + ϵ] = V[β $_0$] + V[β $_1$x] + V[ϵ] = 0 + 0 + σ^2 = σ^2

What're the 3 key assumptions?

a. Uncorrelated Errors (what does this mean specifically?)

b. Constant Variance (**between what?**)

c. and one other...

Constructing a Regression model:

The β 's must all be estimated.
For a sample regression model:

$$y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_{i,k} x_{i,k} + \epsilon_i \text{ for i } = 0,\ 1,\ 2\ ...\ \text{n}$$

Least squares estimation seeks to minimize the sum of the squares of the differences between the observed responses (the y$_i$'s) and the straight line.

$$\text{S}(\beta\ _0,\ \beta\ _1,\ ...) = \sum \epsilon_i^2 = \sum [y - (\beta_0 + \beta_1 x_{i,1} + ...)]^2$$

When you take the partial derivative of each β , you get k + 1 equations. Since you have k + 1 unknowns, you can do some linear algebra to solve for each β . In the case of simple linear regression:

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

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

Put another way:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x}) y_i$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Residuals

The **residuals** of a linear regression model are the errors for each sample which will later be used to determine the adequacy of the model.

$$\epsilon_i = y_i - \hat{y}_i$$

Some properties of the Least Squares Estimators (2.2.2)

The ordinary least-squares (OLS) estimator of the slope ($\hat{\beta}_1$) is a linear combination of the observations, y$_i$:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i$$

where

$$c_i = \frac{(x_i - \overline{x})}{S_{xx}}, \qquad \sum_{i=1}^{n} c_i = 0, \qquad \sum_{i=1}^{n} c_i^2 = \frac{1}{S_{xx}}, \qquad \sum_{i=1}^{n} c_i x_i = 1$$

The last 3 are useful in showing expected value and variance properties:

$$E[\hat{\beta}_1] = \beta_1 \qquad E[\hat{\beta}_0] = \beta_0$$

$$V[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}} \qquad V[\hat{\beta}_0] = \sigma^2 (\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}})$$

The OLS Estimators are the **Best Linear Unbiased Estimators (BLUE)** by the **Gauss-Markov Theorem**, which states that:
In a linear regression model in which the errors

a. have expectation zero,

b. are uncorrelated, and

c. have equal variances,

the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator. Here, "best" means the estimator has the lowest variance as compared to other unbiased, linear estimators.
The errors do not need to be normal, nor do they need to be independent and identically distributed (only uncorrelated with mean zero and **homoscedastic** (i.e. all random variables have the same finite variance)).
More useful properties of the least squares fit:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i, \qquad \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} \epsilon_i = 0, \qquad \sum_{i=1}^{n} \epsilon_i x_i = \sum_{i=1}^{n} \epsilon_i \hat{y}_i = 0$$

The regression line also always passes through the centroid ($\overline{y}$, $\overline{x}$) of the data.

Estimation of σ^2(2.2.3)

The **residual (error) sum of squares** (**SS**$_{res}$), is defined to be:

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \epsilon_i^2$$

The **total sum of squares** is defined to be

$$SS_{total} = SS_{model} + SS_{res}$$

To estimate σ^2, we use

$$\hat{\sigma^2} = \frac{SS_{res}}{n-2} = MS_{res}$$

The quantity n - 2 is the number of **degrees of freedom** (df) for the residual sum of squares. The df = n - 2 because... ***
Since this estimate depends on the model and SS$_{res}$, any model error assumption violations could impact this estimate as well.

Hypothesis Testing on the Slope and Intercept

Three assumptions needed to apply procedures such as hypothesis testing and confidence intervals. Model errors, ϵ_i, are

a. normally distributed,

b. independently distributed, and

c. have constant variance

i.e. $\epsilon_i \sim$N(0, σ^2)
Let's say we want to test if the slope ($\hat{\beta}_1$) is **NOT** equal to some constant, c.
This means we'd want to disprove the null hypothesis, H$_0$, that $\hat{\beta}_1$ = c.
At this point, we'd need to calculate the **standard error** (aka standard deviation aka $\sqrt{V[\sigma^2]}$) of $\hat{\beta}_1$. This is defined like so:

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{MS_{res}}{S_{xx}}}$$

Our test statistic will then be:

$$t_0 = \frac{\hat{\beta}_1 - c}{\text{se}(\hat{\beta}_1)}$$

We reject H$_0$ (i.e. conclude there is sufficient evidence to believe that H$_a$ is true) if:

$$|t_0| > t_{\frac{\alpha}{2}, n-2}$$

We can also use the p-value approach here as well.
To test if the intercept ($\hat{\beta}_0$) is **NOT** equal to some constant, c, we would do the same procedure except use $\hat{\beta}_0$'s standard error:

$$\text{se}(\hat{\beta}_0) = \sqrt{MS_{res}(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}})}$$

Testing the significance of the regression (2.3.2)

H$_0$: β_1 = 0, H$_a$: $\beta_1 \neq 0$
This tests the significance of regression; that is, is there a linear relationship between the response and the regressor?
Failing to reject H$_0$, implies that there is no linear relationship between y and x.
There is also an **analysis of variance** (ANOVA) approach.
SS$_T$ (or SS$_{Total}$) = SS$_{\text{Model or Regression or R}}$ + SS$_{\text{Residual}}$, so:

$$SS_T = SS_R + SS_{Res}, \qquad \text{where SS}_R = \hat{\beta}_1 S_{xy}$$

$$df_T = df_R + df_{Res} \longrightarrow n - 1 = 1 + (n-2)$$

Mean Squares:
$\mathrm{MS}_R \;= \frac{SS_R}{1}$
$\mathrm{MS}_{Res} \;= \frac{SS_{Res}}{n-2}$

	Q1	Q2	Q3	Q4	Q5
50 Points	10	14	8	8	10

Question 1

For multiple regression

$$y = X\beta + \epsilon,\ \epsilon \sim \mathrm{N}(0,\, \sigma^2)$$

Derive or show that

a. $\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}$

b. $\mathrm{E}[\hat{\beta}] = \beta$

c. $\mathrm{V}[\hat{\beta}] = \sigma^2(\mathrm{X}'\mathrm{X})^{-1}$

d. $\mathrm{E}[\hat{Y}] = \mathrm{X}\beta$

e. $\mathrm{V}[\hat{Y}] = \sigma^2\mathrm{H}$, where H is the hat matrix and $\mathrm{H} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$

Question 2 (problems 3.1 and 3.3 on page 121)

a. Fit a multiple linear regression model relating the number of games won to the team's passing yardage (x_2), the percentage of rushing plays (x_7), and the opponents' yards rushing (x_8).

b. Construct the analysis-of-variance table and test for significance of regression.

c. Calculate t statistics for testing the hypotheses H_0: $\beta_2 = 0$, H_0: $\beta_7 = 0$, H_0: $\beta_8 = 0$. What conclusions can you draw about the roles the variables x_2, x_7, and x_8 play in the model?

d. Calculate R^2 and R^2_{adj} for this model.

e. Using the partial F test, determine the contribution of x_7 to the model. How is this partial F statistic related to the t test for β_7 calculated in part c above?

f. Find a 95% CI on β_7. (This is part a of problem 3.3, and the following one is part b of problem 3.3.)

g. Find a 95% CI on the mean number of games won by a team when $x_2 = 2300$, $x_7 = 56.0$, and $x_8 = 2100$.

Note: For c, d, f, and g, please show two versions of your results: (1) obtained using R code and (2) based on your manual calculation (please show detailed step for your manual calculation. You can use the partial output from the lm or ANOVA, e.g., the SS_{reg}, SS_{res}, the estimated value of β and its variance or standard deviation).

Question 3 (Exercise 3.4 on page 122

Reconsider the National Football League data from Problem 3.1. Fit a model to this data using only x_7 and x_8 as the regressors.

a. Test for significance of the regression.

b. Calculate R^2 and $R^2{}_{adj}$. How do these quantities compare to the values computed for the model in problem 3.1, which included an additional regressor (x^2)?

c. Calculate a 95% CI on β_7. Also, find a 95% CI on the mean number of games won by a team when $x_7 = 56.0$ and $x_8 = 2100$. Compare the lengths of these CIs to the lengths of the corresponding CIs from problem 3.3 (that is, the above part f and g in question 2)

d. What conclusions can you draw from this problem about the consequences of omitting an important regressor from a model?

Question 4 (exercise 4.2 on page 165

Consider the multiple regression model fit to the National Football League (NFL) team performance data in problem 3.1.

a. Construct a normal probability plot of the residuals. Does there seem to be any problem with the normality assumption?

b. Construct and interpret a plot of the residuals versus the predicted response.

c. Construct plots of the residuals versus each of the regressor variables. Do these plots imply that the regressor is correctly specified?

d. Construct the partial regression plots for this model. Compare the plots with the plots of residuals versus regressors from part c above. Discuss the type of information provided by these plots.

Question 5

Show that the hat matrix $H = X(X'X)^{-1}X'$and I - H (where I is the identity matrix) are symmetric and idempotent. That is, please show:

a. $H' = H$ and $HH = H$ (H'means the transpose of H, HH means H * H)

b. $(I - H)' = I - H$ and $(I - H)(I - H) = I - H$

Hint: $A = X'X$ is a symmetric matrix, and for a symmetric matrix, $(A')^{-1} = (A^{-1})'$. You can use this property directly in your proof of **(a)** and **(b)**. If you are interested in the proof of this property, you may check the following web page:
https://math.stackexchange.com/questions/325082/is-the-inverse-of-a-symmetric-matrix-also-symmetric