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
$\epsilon \sim N(0, \sigma^2)$
The expected values of each of these quantities are:
$E[y \mid x] = \mu_{y|x} = E[\beta_0 + \beta_1 x + \epsilon] = E[\beta_0] + E[\beta_1 x] + E[\epsilon] = \beta_0 + \beta_1 x + 0$
$V[y \mid x] = \sigma^2_{y|x} = V[\beta_0 + \beta_1 x + \epsilon] = V[\beta_0] + V[\beta_1 x] + V[\epsilon] = 0 + 0 + \sigma^2 = \sigma^2$

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:

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

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.

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

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:

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

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

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** ($\mathbf{SS}_{res}$), 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. $\epsilon_i \sim N(0, \sigma^2)$

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: $\beta_1 = 0$, 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_{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:
$MS_R = \frac{SS_R}{1}$
$MS_{Res} = \frac{SS_{Res}}{n-2}$
10/8/17
for
$y = x\beta + \epsilon$,
$\epsilon \sim N(0, \sigma^2 I)$
Mid-term: Oct 19, Thursday Covers up to ch 4, some parts of ch 5 (depending)
- linear - constant variance - independent - Normal - $E(\epsilon) = 0$
4.4 Outliers:
- An outlier si an observation considerably different from the others - Formal tests for outliers - Points with large residuals may be outliers - Impact can be assessed by removing the points and refitting - How should they be treated?
4.5 Lack of Fit of the Regression Model
A formal test for lack of fit:
Assumes - Normality, independence, constant variance assumptions have been met - Only the first-order or straight line model is in doubt
Requires - replication of y for at least one level of x
With replication, we can obtain a "model-independent" estimate of σ^2
Say there are n_i observations of the response at the ith level of the regressor x_i, $i = 1, 2, ... m$
In other words, for a given x, you might have i occurences of it.
y_{ij} denotes the jth observation on the response at x_i, $j = 1, 2, ... n_i$
Total number of observations is $n = \sum_{i=1}^{m} n_i$
i.e. if you have 3 occurances of $x = 3.3$, and 3.3 is the nth x, then n_3 is 3.
SS_{total} (n - 1) = SS_{regr} (1) + SS_{resid} (n - 2)
$SS_{resid} = SS_{pE} + SS_{LOF}$
$SS_{pE} = (y_{ij} - \overline{y}_i)^2$
$SS_{LOF} = (\overline{y}_i - \hat{y}_{ij})^2$
$SS_{res} = (y_{ij} - \hat{y}_{ij})^2$

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\overline{y}_i - \hat{y}_{ij})^2$$

$$(n-2) = (n-m) + (m-2)$$

$$SS_{res} = SS_{PE} + SS_{LOF}$$

How small is small?
Residual plots:
e_i vs $\hat{y}_i$
x_i vs $\hat{e}_i$
something vs normal
If you have 10 distinct data points, then m is 10. If you have 17 data points total (not necessarily distinct) then n is 17.
In R, you isntall the package alr3, and do ?pureErrorAnova

10/12/17
Midterm will be Ch 0 - 5.

Ch 0 - Review (The Review HW and other basic stats. Specifically: estimating/hypothesizing/CI'ing u and σ^2)
Ch 1 - Intro
Ch 2 - Simple Linear Regression
Ch 3 - Multiple Linear Regression
Ch 4 - Model Adequacy Checking / Diagnosis
Ch 5 - Transformation and Weighting to Correct Model Inadequacies
As far as Ch 5 goes
$y = x\beta + \epsilon$
Key assumptions:
Independence -
Constant Variance -
Normally distributed -
$E[\epsilon] = 0$
Linearity -
5.1 introduction
Data Transformation (fix the problem of assumption violation e.g. non-constant variance)
Generalized and Weighted Least Squares (fix the problem of assumption violation, e.g. correlated errors or non-constant variance)
Subject-Matter Knowledge (Ideally, the choice of metric should be made by the engineer or scientist with subject-matter knowledge)
5.2 Variance stabilizing Transformations
Constant variance assumption: - Often violated when the variance is **functionally related** to the mean
i.e. $\sigma^2 \alpha$ E[y] or σ^2 = f(E[y])
- Transformation on the response may eliminate the problem
-
—-
An electric utility is interested in developing a model relating peak-hour demand (y) to total energy usage during the month (x).
This is an important planning problem because while most customers pay directly for energy usage (in kilowatt-hours), the generation system must be large enough to meet the maximum demand imposed.
Nonlinearity may be detected via the lack-of-fit test of Section 4.5
If a transformation of a nonlinear function can result in a linear function - we say it is intrinsically or transformably linear.
Ex:
$y = \beta_0 e^{\beta_1 x} \epsilon$
$\ln y = \ln \beta_0 + \beta_1 x + \ln \epsilon$