# STAT 331 Course Notes

Applied Linear Models

Peter Balka • Winter 2018 • University of Waterloo

Last Revision: January 18, 2018

## Table of Contents

**Abstract**

These notes are intended as a resource for myself; past, present, or future students of this course, and anyone interested in the material. The goal is to provide an end-to-end resource that covers all material discussed in the course displayed in an organized manner. If you spot any errors or would like to contribute, please contact me directly.

# 1 January 4, 2018

## 1.1 Simple linear regression review

In SLRM, there is a single explanatory variate and a response variate.
A good graphical summary for SLRM are **scatterplots**.
A good numerical summary for SLRM is the **correlation coefficient** defined as

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

where $-1 \leq r \leq 1$. If $|r| \approx 1$ then the explanatory/response variates have a strong linear relationship.

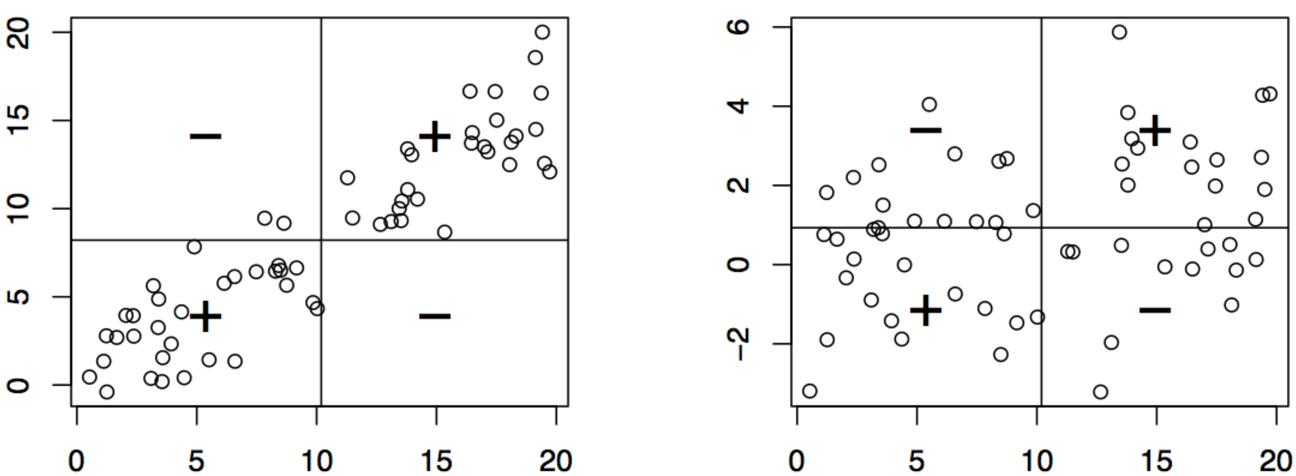
# 2 January 9, 2018

## 2.1 Correlation coefficient and covariance

Note: the measure $r$ is also the covariance divided by the standard deviations or

$$r = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

Note that the covariance $E[(X - E[X])(Y - E[Y])]$ can be graphically separated by the means $\bar{X}$ and $\bar{Y}$.



One can see that the covariance signage is determined by the sum of the magnitudes in the positive and negative quadrants.

## 2.2 Simple linear regression (SLR) model

An SLR model can be thought of as a line with covariates $x$ and $y$ where

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \ldots, n$$

where $\epsilon_i$ is some error term for each $i$.

**Example 2.1.** From the dataset

| Overhead | Office Size |
|---|---|
| 218955 | 1589 |
| 224513 | 1912 |
| ⋮ | ⋮ |

Thus we have the SLR model

$$\begin{aligned} 218955 &= \beta_0 + \beta_1(1589) + \epsilon_1 \\ 224513 &= \beta_0 + \beta_1(1912) + \epsilon_2 \end{aligned}$$

## 2.3 Methods of least squares

Find (estimate) the value of $\beta_0, \beta_1$ (denoted by $\hat{\beta}_0, \hat{\beta}_1$, respectively) that minimizes the sum of squares of the errors $\sum_{i=1}^n \epsilon_i^2$. That is: we find values of $\beta_0, \beta_1$ that minimizes the function

$$S(\beta_0, \beta_1) = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$$

We take the partial derivatives and set to 0 to find the minimum (assuming convexity)

$$\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2 \sum_{i=1}^n y_i - (\beta_0 + \beta_1 x_i) = 0 \\ \frac{\partial S}{\partial \beta_1} &= -2 \sum_{i=1}^n x_i [y_i - (\beta_0 + \beta_1 x_i)] = 0 \end{aligned}$$

which yields (the notation changes to estimates of $\beta$ assuming we can calculate those)

$$\begin{aligned} \sum_{n=1}^n y_i &= n\hat{\beta}_0 + \sum_{n=1}^n x_i \hat{\beta}_1 \\ \sum_{n=1}^n x_i y_i &= \sum_{i=1}^n x_i \hat{\beta}_0 + \sum_{n=1}^n x_i^2 \hat{\beta}_1 \end{aligned}$$

which gives us the estimates

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

where the second equation follows from substituting in the first and re-deriving for $\frac{\partial S}{\partial \beta_1}$.
The corresponding fitted line is

$$\hat{\mu}_{y|X=x} = \hat{\mu} + \hat{\beta}_0 + \hat{\beta}_1 x$$

For the example with overhead above, we'd have

$$\hat{\mu} = -27877.06 + 126.33x$$

## 2.4 Fitted residuals

These are the difference between the actual values and our fitted value (distinct from the error terms previously)

$$e_i = (y_i - \hat{\mu_i}) = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Some **key points** regarding this model

- By estimating two parameters $(\beta_0, \beta_1)$, we have imposed two constraints on our residuals (from our partial derivatives)

$$\sum e_i = 0$$
$$\sum x_i e_i = 0$$

  These reduces our number of $n$ independent measures by 2 since we can compute the remaining two residuals from $n-2$ observations. Thus we have $n-2$ **degrees of freedom** (or in general, $n-k$ dfs where $k$ is the number of estimated parameters>).

## 2.5 Interpretation of estimated parameters $\hat{\beta}_i$

$\beta_1$

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x$$
$$\hat{\mu_{x+1}} = \hat{\beta}_0 + \hat{\beta}_1(x+1)$$
$$= \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_1$$
$$= \hat{\mu} + \hat{\beta}_1$$

thus $\hat{\beta}_1$ can be interpreted as the estimated mean change in the response $(y)$ associated with one unit change of $x$.

$\beta_0$ For $x = 0$, $\hat{\mu} = \hat{\beta}_0$.

However, in the example with overhead, it's evident that when $x = 0$ overhead is negative $(-27877.06)$ which is nonsensical.

Never extrapolate results outside the range of the values of the explanatory variate(s).

# 3 January 16, 2018

## 3.1 Invariants for normal SLR models

Recall for the normal SLR model we have

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \ldots, n$$

where $\epsilon_i \sim N(0, \sigma^2)$ is some error term for each $i$.

- $\beta_0 + \beta_1 x_i$ is the **deterministic** and $\epsilon_i$ is the **random** components of the model.
- $Var(\epsilon_i) = 0$ for all $i$ (constant variance)
- $\epsilon_i, \epsilon_j$ for $i \neq j$ are independent (otherwise we'd need time series)

## 3.2 Estimate of variance

Each of our error terms follow a $N(0, \sigma^2)$ distribution. The **unbiased** estimate of $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}$$

The **residual standard error** is $\hat{\sigma}^2 = \sqrt{\frac{\sum e_i^2}{n-2}}$.

## 3.3 Unbiased estimator of $\hat{\beta}_1$

The **estimator** of $\beta_1$ is a random variable that is similar to the estimate but with r.v. $Y_i$ and $\bar{Y}$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}$$

Note: $\hat{\beta}_1$ can be expressed as a linear combination of response varaibles $Y_i$, $i = 1, 2, \ldots, n$.

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum (x_i - \bar{x}) Y_i - \bar{Y} \sum (x_i - \bar{x})}{S_{xx}} \\
&= \frac{\sum (x_i - \bar{x}) Y_i}{S_{xx}} && \sum (x_i - \bar{x}) = 0 \\
&= \sum_{i=1}^n c_i Y_i && c_i = \frac{(x_i - \bar{x})}{S_{xx}}
\end{aligned}$$

Remember that

$$\begin{aligned}
&\epsilon_i \sim N(0, \sigma^2) \text{ independent } \Rightarrow Y_i \sim N(\beta_0 + \beta_1 x, \sigma^2) \text{ independent} \\
\Rightarrow &\hat{\beta}_1 \sim \text{ Normal (sum of independent normal r.v.'s)}
\end{aligned}$$

Thus we have

$$
\begin{aligned}
E(\hat{\beta}_1) = E(\sum c_i Y_i) &= \sum c_i E(Y_i) \\
&= \sum (\frac{(x_i - \bar{x})}{S_{xx}})(\beta_0 + \beta_1 x_i) \\
&= \frac{\beta_0 \sum (x_i - \bar{x}) + \beta_1 \sum x_i (x_i - \bar{x})}{S_{xx}} \\
&= \frac{\beta_1 \sum x_i (x_i - \bar{x}) - \beta_1 \bar{x} \sum (x_i - \bar{x})}{S_{xx}} && \text{eliminate and introduce 0 term} \\
&= \frac{\beta_1 \sum (x_i - \bar{x})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \\
&= \beta_1
\end{aligned}
$$

Since $E(\hat{\beta}_1) = \beta_1$, then $\hat{\beta}_1$ is an unbiased estimator of $\beta_1$.
The variance of our estimator $\hat{\beta}_1$ is

$$
\begin{aligned}
Var(\hat{\beta}_1) &= Var(\sum c_i Y_i) \\
&= \sum c_i^2 Var(Y_i) && Y_i \text{ independent} \\
&= \sum \frac{\sigma^2 (x_i - \bar{x})^2}{S_{xx}^2} \\
&= \frac{\sigma^2}{S_{xx}} \\
&= \frac{\sigma^2}{\sum (x_i - \bar{x})^2}
\end{aligned}
$$

Since our estimator $\hat{\beta}_1$ follows (from above)

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

we have

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0, 1)$$

in terms of the sample variance (or estimate $\hat{\sigma}$ we have the **T-distribution**

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{\sigma}}{\sqrt{S_{xx}}}} \sim t_{n-2}$$

## 3.4 Central limit theorem

Recall that the distribution of the sample means follows a normal distribution

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

so we have

$$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$
$$\frac{\bar{X}-\mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim t_{n-1}$$

This follows from

$$SD(X) = \sigma$$
$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$
$$SE(\bar{X}) = \frac{\hat{\sigma}}{\sqrt{n}}$$
$$\frac{\bar{X}-\mu}{SE(\bar{X})} \sim t_{n-1}$$
$$\frac{\hat{\beta}_1-\beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

# 4 January 18, 2018

## 4.1 Inference for $\beta_1$

"Is there a relationship between overhead and office size (for the population of offices)?"
There is no (linear) relationship $\iff \beta_1 = 0$.
We can statistically check this using two methods

1. Confidence interval for $\beta_1$
2. Hypothesis test for $\beta_1$ ($H_0 : \beta_1 = 0$)

## 4.2 Confidence interval

For a given distribution with one parameter $\mu$, we can calculate the $(1-\alpha)100\%$ confidence interval for $\mu$ (note: we need only one $t$ value since the T-distribution is symmetric)

$$\hat{\mu} \pm t_{n-1,1-\frac{\alpha}{2}} \cdot SE(\hat{\mu})$$
$$\Rightarrow \bar{x} \pm t_{n-1,1-\frac{\alpha}{2}} \left(\frac{\hat{\sigma}}{\sqrt{n}}\right)$$

The $(1-\alpha)100\%$ confidence interval for $\beta_1$ (where the we have $n-2$ degrees of freedom)

$$\hat{\beta}_1 \pm t_{n-2,1-\frac{\alpha}{2}} \cdot SE(\hat{\beta}_1)$$

**Example 4.1.** The 95% C.I. for $\beta_1$ for overhead data is

$$\hat{\beta}_1 \pm t_{22,0.975} SE(\hat{\beta}_1)$$
$$=126.33 \pm 2.074(10.88)$$
$$=126.33 \pm 22.57 = (103,76,148.90)$$

where $\pm 22.57$ is the **margin of error**.

Since $\beta_1 = 0$ is not in the interval, we can conclude that there is a *significant* positive relationship between overhead and office size.

**Remark 4.1.** An X% confidence interval can be interpreted as: X% of X% confidence intervals established from repeated samples contain the true value.

### 4.3 Hypothesis testing

We form a **null hypthoesis** $H_0$ and an alternative hypothesis $H_1$, where we assume $H_0$ unless there is statistical significance rejecting $H_0$.
For simple linear regression, we suppose

$$\begin{aligned} H_0 &: \beta_1 = 0 && \text{no relationship} \\ H_1 &: \beta_1 \neq 0 && \text{two-sided alternative} \end{aligned}$$

Our **test statistic** $t$ is the distribution

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

**Example 4.2.** Under $H_0$ we have for our example

$$t = \frac{126.33 - 0}{10.88} = 11.61$$

If we look at our $t_{22}$ distribution, we find the total probability of the pdf at $P(t \leq 11.61)$ and $P(t \geq 11.61)$ (the **p-value**).
We see that $P(t_{22} > 2.819) = 0.0005 \Rightarrow P(t_{22} > 11.61) << 0.005$.
Thus the p-value is $2P(t_{22} > 11.61) << 0.01$ (in fact, it is $7.47 \times 10^{-11}$), which is lower than **0.05 (the significance level)**, so we reject the null hypothesis.

**Remark 4.2.** The p-value of a hypothesis test can be interpreted as: the probability that our sample holds (the observed, or more extreme, results) under the null hypothesis. If it is extremely low (past a certain threshold), then we may reject the null hypothesis as not possible.

### 4.4 Confidence interval vs hypothesis testing

Deciding which method to use is problem dependent: usually, hypothesis testing is simpler to interpret for many variates and a confidence interval is only relevant for single variates.
A 95% confidence interval corresponds with a hypothesis test with a 0.05 significance level i.e. we will derive an equivalent conclusion.

### 4.5 Multiple Regression Model

We want to model the following relationship

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

where $\epsilon_i = N(0, \sigma^2)$ and independent.
Note we have $p$ variates and $p+1$ parameters (the bias term) thus we have $n - (p+1)$ degrees of freedom.

In matrix form, this is represented as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

which can be written succinctly as

$$Y = X\beta + \epsilon$$

where $\epsilon = N(\vec{0}, \sigma^2 I)$ or $Var(\epsilon) = \sigma^2 I$ (the **covariance matrix**; note that the covariance between $\epsilon_i, \epsilon_j$ $i \neq j$ is 0 since they are independent).