

# Linear Models - Regression

Ananthan Ambikairajah

27 February, 2020

# Common statistical tests are linear models

Last updated: 02 April, 2019

See worked examples and more details at the accompanying notebook: <https://lindeleov.github.io/tests-as-linear>

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: $lm(y \sim 1 + x)$	<b>y is independent of x</b> P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	$lm(y \sim 1)$ $lm(\text{signed\_rank}(y) \sim 1)$	✓ for $N > 14$	One number (intercept, i.e., the mean) predicts y. - (Same, but it predicts the signed rank of y.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y1, y2, paired=TRUE) wilcox.test(y1, y2, paired=TRUE)	$lm(y_2 - y_1 \sim 1)$ $lm(\text{signed\_rank}(y_2 - y_1) \sim 1)$	✓ for $N > 14$	One intercept predicts the pairwise $y_2 - y_1$ differences. - (Same, but it predicts the signed rank of $y_2 - y_1$ .)	
	<b>y ~ continuous x</b> P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	$lm(y \sim 1 + x)$ $lm(\text{rank}(y) \sim 1 + \text{rank}(x))$	✓ for $N > 10$	One intercept plus x multiplied by a number (slope) predicts y. - (Same, but with ranked x and y)	
	<b>y ~ discrete x</b> P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y1, y2, var.equal=TRUE) t.test(y1, y2, var.equal=FALSE) wilcox.test(y1, y2)	$lm(y \sim 1 + G_2)^4$ $glm(y \sim 1 + G_2, \text{weights}=\dots)^4$ $lm(\text{signed\_rank}(y) \sim 1 + G_2)^4$	✓ for $N > 11$	An intercept for group 1 (plus a difference if group 2) predicts y. - (Same, but with one variance per group instead of one common.) - (Same, but it predicts the signed rank of y.)	
Multiple regression: $lm(y \sim 1 + x_1 + \dots + x_k + \dots)$	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	$lm(y \sim 1 + G_2 + G_3 + \dots + G_k)^4$ $lm(\text{rank}(y) \sim 1 + G_2 + G_3 + \dots + G_k)^4$	✓ for $N > 11$	An intercept for group 1 (plus a difference if group $\neq 1$ ) predicts y. - (Same, but it predicts the rank of y.)	
	P: One-way ANCOVA	aov(y ~ group + x)	$lm(y \sim 1 + G_2 + G_3 + \dots + G_k + x)^4$	✓	- (Same, but plus a slope on x.) Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.	
	P: Two-way ANOVA	aov(y ~ group * sex)	$lm(y \sim 1 + G_2 + G_3 + \dots + G_k + S_2 + S_3 + \dots + S_k + G_2 * S_2 + G_2 * S_3 + \dots + G_k * S_k)$	✓	Interaction term: changing sex changes the y - group parameters. Note: $G_{2:k}$ is an indicator (0 or 1) for each non-intercept levels of the group variable. Similarly for $S_{2:k}$ for sex. The first line (with $G_2$ ) is main effect of group, the second (with $S_2$ ) for sex and the third is the group * sex interaction. For two levels (e.g. male/female), line 2 would just be ' $S_2$ ' and line 3 would be ' $S_2$ multiplied with each $G_2$ '.	[Coming]
	<b>Counts ~ discrete x</b> N: Chi-square test	chisq.test(groupXsex_table)	<b>Equivalent log-linear model</b> $glm(y \sim 1 + G_2 + G_3 + \dots + G_k + S_2 + S_3 + \dots + S_k + G_2 * S_2 + G_2 * S_3 + \dots + G_k * S_k, \text{family}=\dots)^4$	✓	Interaction term: (Same as Two-way ANOVA.) Note: Run glm using the following arguments: <code>glm(model, family=poisson())</code> As linear-model, the Chi-square test is $\log(y) = \log(N) + \log(\alpha) + \log(\beta) + \log(\alpha\beta)$ where $\alpha$ and $\beta$ are proportions. See more info in the accompanying notebook.	Same as Two-way ANOVA
Multiple regression: $lm(y \sim 1 + x_1 + \dots + x_k + \dots)$	N: Goodness of fit	chisq.test(y)	$glm(y \sim 1 + G_2 + G_3 + \dots + G_k, \text{family}=\dots)^4$	✓	(Same as One-way ANOVA and see Chi-Square note.)	1W-ANOVA

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation  $y \sim 1 + x$  is shorthand for  $y = 1 \cdot b + a \cdot x$  which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they all are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables  $G_i$  and  $S_i$  are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when  $\Delta x = 1$  between categories the difference equals the slope. Subscripts (e.g.,  $G_2$  or  $y_2$ ) indicate different columns in data. It requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at <https://lindeleov.github.io/tests-as-linear>.

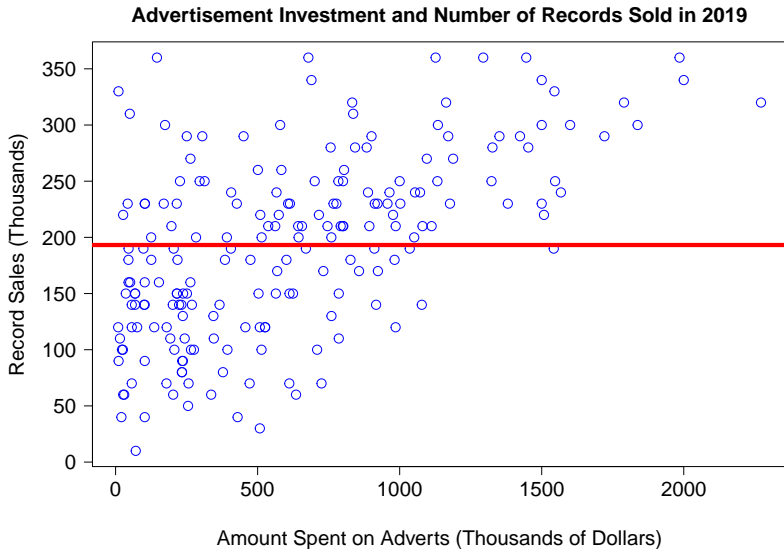
<sup>4</sup> See the note to the two-way ANOVA for explanation of the notation.

<sup>4</sup> Same model, but with one variance per group: `glm(value ~ 1 + G_2, weights = varIdent(form = ~1|group), method="ML")`.

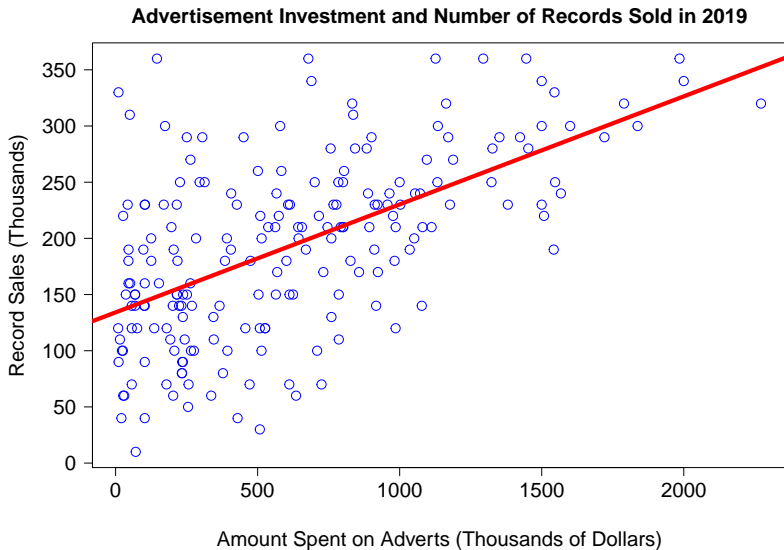


Jonas Kristoffer Lindelev  
<https://lindeleov.net>

# The Mean - A Very Simple Statistical Model

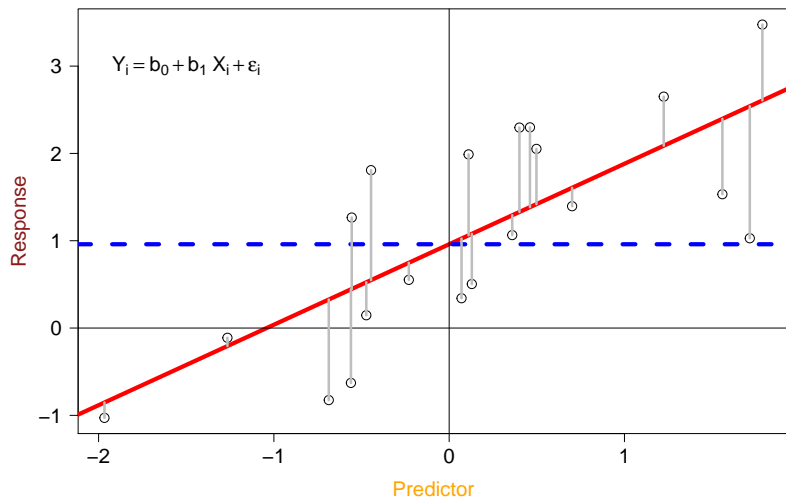


# The Method of Least Squares

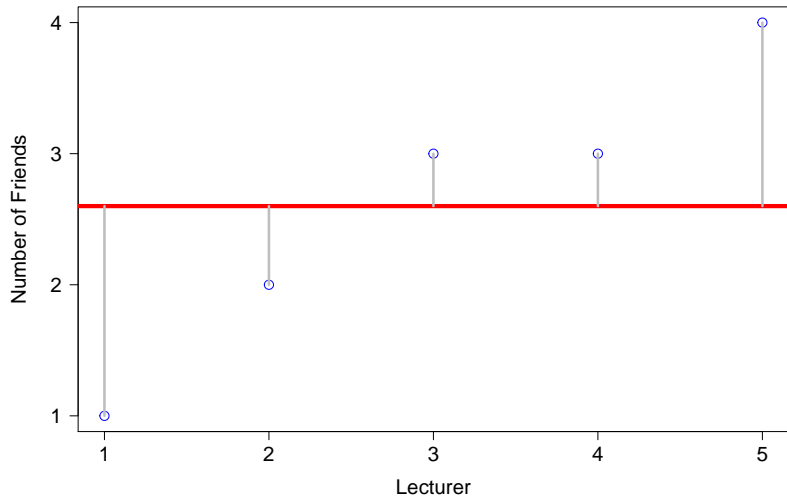


# Linear Regression

Response = Intercept + Slope x Predictor + Error

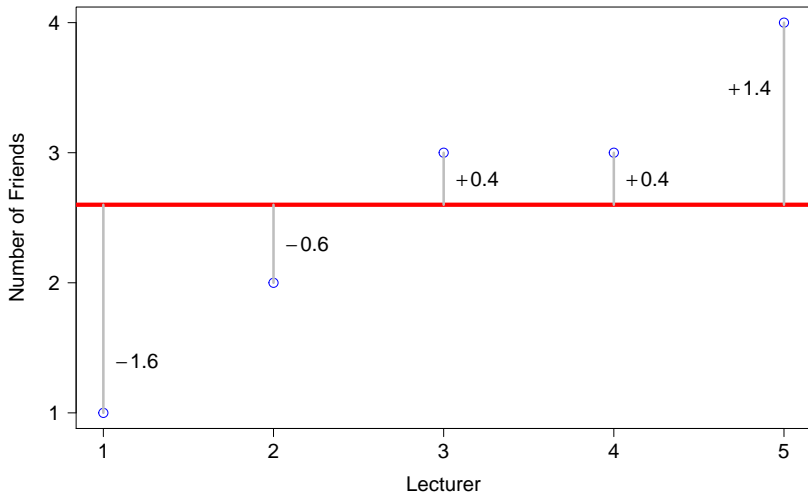


## Concept Check - Variance



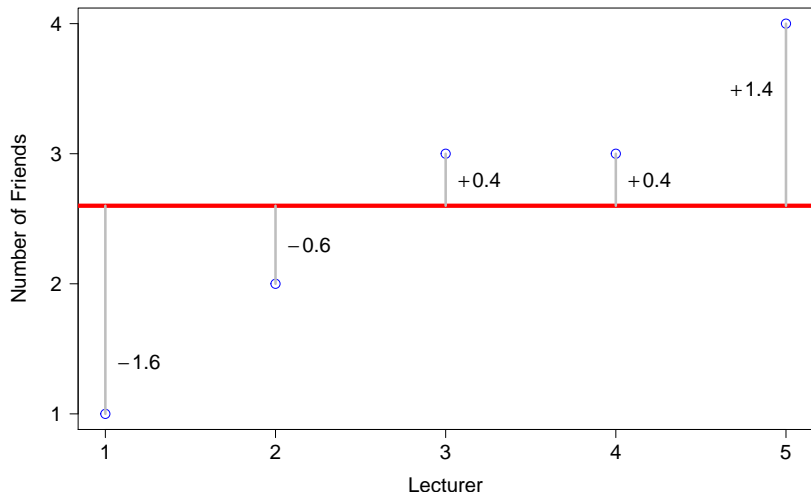
## Concept Check - Variance

- total error = sum of deviances =  $\sum (x_i - \bar{x})$



## Concept Check - Variance

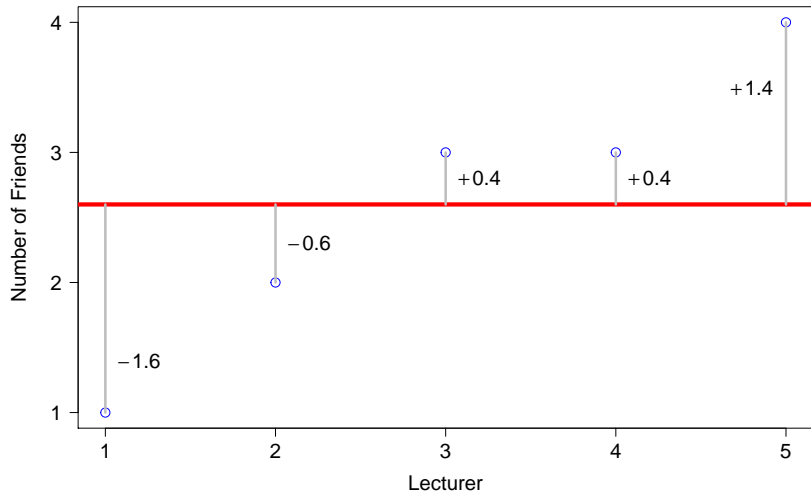
- ▶ total error = sum of deviances =  $\sum (x_i - \bar{x})$
- ▶ sum of squared errors (SS) =  $\sum (x_i - \bar{x})(x_i - \bar{x})$





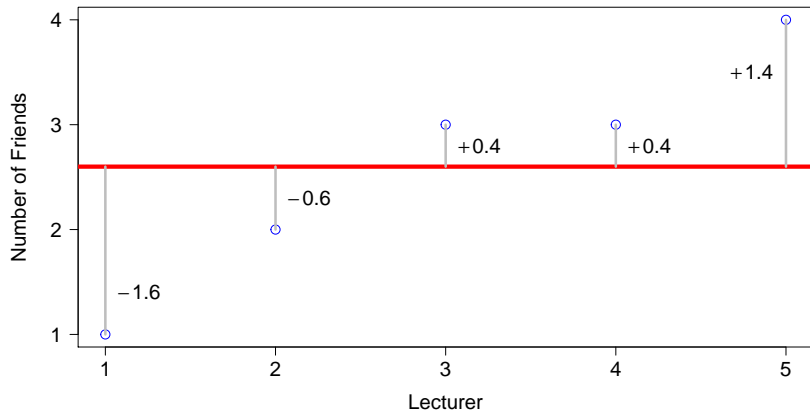
## Concept Check - Variance

- ▶ total error = sum of deviances =  $\sum (x_i - \bar{x})$
- ▶ sum of squared errors (SS) =  $\sum (x_i - \bar{x})(x_i - \bar{x})$
- ▶ variance ( $s^2$ ) =  $\frac{SS}{N-1} = \frac{\sum (x_i - \bar{x})^2}{N-1}$



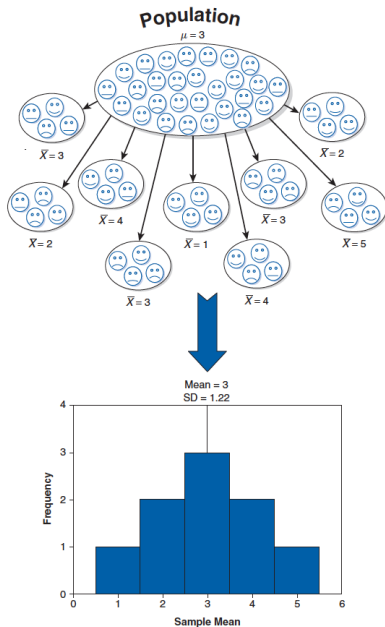
## Concept Check - Variance

- ▶ total error = sum of deviances =  $\sum (x_i - \bar{x})$
- ▶ sum of squared errors (SS) =  $\sum (x_i - \bar{x})(x_i - \bar{x})$
- ▶ variance ( $s^2$ ) =  $\frac{SS}{N-1} = \frac{\sum (x_i - \bar{x})^2}{N-1}$
- ▶ standard deviation ( $s$ ) =  $\sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}}$



# Concept Check - Standard Error

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}}$$

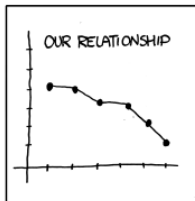


# Constructing Simple Regressions in R

```
album_data <- read.delim("Album Sales 1.dat", header = TRUE)
album_lm_1 <- lm(sales ~ 1 + adverts, data = album_data)
summary(album_lm_1)
```

```
##
## Call:
## lm(formula = sales ~ 1 + adverts, data = album_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -152.949  -43.796   -0.393   37.040  211.866
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.341e+02  7.537e+00  17.799   <2e-16 ***
## adverts      9.612e-02  9.632e-03   9.979   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.99 on 198 degrees of freedom
## Multiple R-squared:  0.3346, Adjusted R-squared:  0.3313
## F-statistic: 99.59 on 1 and 198 DF, p-value: < 2.2e-16
```

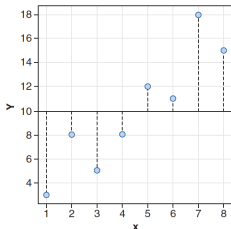
# Exercise 1 - Simple Regression



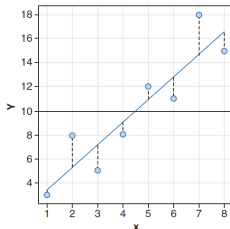
# The Theory - Regression Output

```
##
## Call:
## lm(formula = sales ~ 1 + adverts, data = album_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -152.949  -43.796   -0.393   37.040  211.866
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.341e+02  7.537e+00  17.799  <2e-16 ***
## adverts      9.612e-02  9.632e-03   9.979  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.99 on 198 degrees of freedom
## Multiple R-squared:  0.3346, Adjusted R-squared:  0.3313
## F-statistic: 99.59 on 1 and 198 DF,  p-value: < 2.2e-16
```

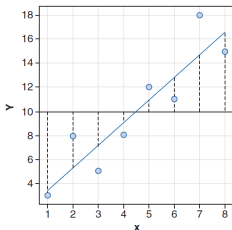
# The Model - Assessing Goodness of Fit ( $R^2$ )



$SS_T$  uses the differences between the observed data and the mean value of  $Y$



$SS_R$  uses the differences between the observed data and the regression line



$SS_M$  uses the differences between the mean value of  $Y$  and the regression line

# The Model - Assessing Goodness of Fit ( $R^2$ )

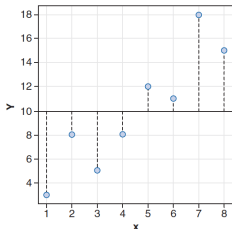
$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N - 1}}$$

$$R^2 = \frac{SS_M}{SS_T}$$

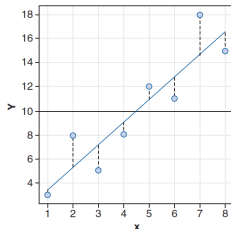
$$1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

$$F = \frac{MS_M}{MS_R}$$

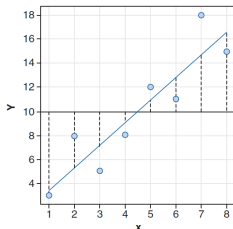
$$r = \sqrt{R^2}$$



$SS_T$  uses the differences between the observed data and the mean value of  $Y$



$SS_R$  uses the differences between the observed data and the regression line



$SS_M$  uses the differences between the mean value of  $Y$  and the regression line



# The Theory - Regression Output

```
##  
## Call:  
## lm(formula = sales ~ 1 + adverts, data = album_data)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -152.949  -43.796   -0.393   37.040  211.866   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept) 1.341e+02  7.537e+00  17.799  <2e-16 ***  
## adverts      9.612e-02  9.632e-03   9.979  <2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 65.99 on 198 degrees of freedom  
## Multiple R-squared:  0.3346, Adjusted R-squared:  0.3313   
## F-statistic: 99.59 on 1 and 198 DF,  p-value: < 2.2e-16
```

# The Theory - The Predictors

## A.2 Critical values of the $t$ -distribution

$$t = \frac{b_{\text{observed}} - b_{\text{expected}}}{SE_b}$$

$$t = \frac{b_{\text{observed}}}{SE_b}$$

$$df = N - p - 1$$

df	Two-Tailed Test		One-Tailed Test	
	0.05	0.01	0.05	0.01
1	12.71	63.66	6.31	31.82
2	4.30	9.92	2.92	6.96
3	3.18	5.84	2.35	4.54
4	2.78	4.60	2.13	3.75
5	2.57	4.03	2.02	3.36
6	2.45	3.71	1.94	3.14
7	2.36	3.50	1.89	3.00
8	2.31	3.36	1.86	2.90
9	2.26	3.25	1.83	2.82
10	2.23	3.17	1.81	2.76
11	2.20	3.11	1.80	2.72
12	2.18	3.05	1.78	2.68
13	2.16	3.01	1.77	2.65
14	2.14	2.98	1.76	2.62
15	2.13	2.95	1.75	2.60
16	2.12	2.92	1.75	2.58
17	2.11	2.90	1.74	2.57
18	2.10	2.88	1.73	2.55
19	2.09	2.86	1.73	2.54
20	2.09	2.85	1.72	2.53
21	2.08	2.83	1.72	2.52
22	2.07	2.82	1.72	2.51
23	2.07	2.81	1.71	2.50
24	2.06	2.80	1.71	2.49
25	2.06	2.79	1.71	2.49
26	2.06	2.78	1.71	2.48
27	2.05	2.77	1.70	2.47
28	2.05	2.76	1.70	2.47
29	2.05	2.76	1.70	2.46

## Exercise 2 - Derive and Interpret the Output

```
##
## Call:
## lm(formula = sales ~ 1 + adverts, data = album_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -152.949  -43.796   -0.393   37.040  211.866
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.341e+02  7.537e+00  17.799  <2e-16 ***
## adverts      9.612e-02  9.632e-03   9.979  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.99 on 198 degrees of freedom
## Multiple R-squared:  0.3346, Adjusted R-squared:  0.3313
## F-statistic: 99.59 on 1 and 198 DF,  p-value: < 2.2e-16
```

## Exercise 3 - Anscombe's Quartet



Thank You

