1 Estimation

Estimator: Some function of the sample RVs, $T = g(X_1, ..., X_n)$, used to estimate pop param θ. Estimator is transformation of n RVs so is also an RV denoted by T.

Estimate: Value of estimator for sample by applying function to data, $t = g(x_i, .., x_n)$.

Unbiased: $\mathbb{E}[T_n] = \theta$ at any sample size n. Asymptotically unbiased: $\lim_{n\to\infty} \mathbb{E}[T_n] = \theta$

Consistent: Estimator is asymptotically unbiased, $\mathbb{E}[T_n] \to \theta, n \to \infty$ and $\text{Var}(T_n) \to 0, n \to \infty$.

Sample Mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$. $\overline{X}$ unbiased, consistent for μ, S^2 unbiased for σ^2.

Moments: Population RV X w pdf / mass func $f_X(x \mid \theta)$. The k^{th} pop moment is defined as expectation:

$$\mu_k := \mathbb{E}[X^k] = \begin{cases} \sum_x x^k f_X(x \mid \theta) \text{ Discrete } X \\ \int_{-\infty}^{\infty} x^k f_X(x \mid \theta) dx \text{ Continuous } X \end{cases}$$

For sample $X_1, ..., X_n$, the k^{th} sample moment estimator is $\overline{X^k} = \frac{1}{n}\sum_{i=1}^n X_i^k$

Method of Moments $X_1, ..., X_n$ random sample depending on pop params $\theta_1, ..., \theta_m$ (unknown). Suppose $(\theta_1, ..., \theta_m) = h(\mu_1, ..., \mu_2)$. MoM estimators for $\theta_1, ..., \theta_m$ are $(T_1, ..., T_m) = h\left(\overline{X}, ..., \overline{X^m}\right)$.

E.g. $X_1, ..., X_n$ iid RVs w pdf $f_X(x \mid \theta) = \frac{1}{\theta+1}$. First pop moment $\mathbb{E}[X] = \int x f_X(x \mid \theta) dx = \frac{\theta+1}{2}$. Rearrange: $\theta = 2\mathbb{E}[X] - 1$. Estimator T for θ : $T = 2\overline{X} - 1$.

Maximum likelihood: Assuming data are indep, joint pdf / mass of observations $\boldsymbol{x} = (x_1, ..., x_n)$ denoted $f(x_1, ..., x_n \mid \theta)$ is $\prod_{i=1}^n f(x_i \mid \theta)$.

Likelihood function: $L(\theta \mid \boldsymbol{x}) = \prod_{i=1}^n f(x_i; \theta)$ log-likelihood: $l(\theta \mid \boldsymbol{x}) = \log(L(\theta \mid \boldsymbol{x})) = \sum_{i=1}^n \log(f_X(x_i \mid \theta))$

Maximum Likelihood Estimate: Function of sample $x_1, ..., x_n$ that maximises $L(\theta \mid \boldsymbol{x})$. Basically find expression for estimator $\hat{\theta}$ at turning point on log-likelihood.

Confidence Interval: Let $\boldsymbol{X} = (X_1, ..., X_n)$ rep vector of sample RVs from some pop w unknown param θ. A $100(1-\alpha)\%$ confidence interval for θ w lower & upper bound estimators $l(\boldsymbol{X}), u(\boldsymbol{X})$ st: $\mathbb{P}(l(\boldsymbol{X}) < \theta < u(\boldsymbol{X})) = 1 - \alpha$.

CLT $\frac{\overline{X}-\mu}{\sqrt{\text{Var}(\overline{X})}} \sim N(0,1) \;\; n \to \infty$. Confidence interval $100(1-\alpha)\% = \left(\overline{X} - z_{\frac{\alpha}{2}}\widehat{S.E.}(\overline{X}), \overline{X} - z_{\frac{\alpha}{2}}\widehat{S.E.}(\overline{X})\right)$ where $\widehat{S.E.}(\overline{X}) = \sqrt{\text{Var}(\overline{X})}$ is standard error in estimator $\overline{X}$.

Chi squared Distribution: Let $Z_1, ..., Z_n$ be indep standard normal RVs. Define $X = \sum_{i=1}^n Z_i^2$, then RV X has a chi-squared distribution w n degrees of freedom: $X \sim \chi_n^2$

t-Distribution: Let $Z \sim N(0,1), Y \sim \chi_m^2$ indep RVs. The t-dist on m dof is $\frac{Z}{\sqrt{\frac{Y}{m}}} \sim t_m$. pdf is symm at zero, tails have more mass than standard normal, converges to standard normal as $m \to \infty$.

Distribution $X_1, ..., X_n$ indep norm RVs w unknown μ, σ^2. Then $\overline{X}, S^2$ are indep, $\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$, $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$, $T = \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$

Confidence interval for sample var: CI for $\sigma^2 = \left(\frac{(n-1)S^2}{\chi^2_{n-1;\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{n-1;1-\frac{\alpha}{2}}}\right)$. For given dataset w sample var s^2, CI for $\sigma^2 = \left(\frac{(n-1)s^2}{\chi^2_{n-1;\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{n-1;1-\frac{\alpha}{2}}}\right)$

2 Hypothesis Testing

Concluding Statements: Either reject the null hypothesis in favour of some alternative, or fail to reject the null hypothesis.

Type I error: (false positive) occurs when the null hypothesis is incorrectly rejected when it is actually true. Probability of incurring a Type I error, $\mathbb{P}(\text{Reject } H_0 \mid H_0 \text{is true})$ is the significance level of the hypothesis rest (α).

Type II error: (false negative) occurs when incorrectly fail to reject the null hypothesis when it is incorrect. $\mathbb{P}(\text{Fail to reject} H_0 \mid H_0 \text{is false})$ is β. $1 - \beta$ is called the power of the test.

p-value: Prob of obtaining a value for test stat RV, T, that is as or more extreme than observed test stat, t.

Z-test: Let $X_1, ..., X_n$ be indep $N(\mu, \sigma^2)$ where σ^2 is known. Investigate if μ is some specified μ_0.

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$$

$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, under H_0, $Z = \frac{\overline{X}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$. Critical Region: $\{z : \mid z \mid \geq z_0\}, \mathbb{P}(Z \geq z_0) = \frac{\alpha}{2}$.

Power Function: Prob that we correctly reject H_0.

$$\mathbb{P}(Z \in C \mid \mu = \mu^*) = \Phi\left(-z_{\frac{\alpha}{2}} - \frac{\mu^* - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}\right) + 1 - \Phi\left(z_{\frac{\alpha}{2}} - \frac{\mu^* - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}\right)$$

Impacted by sample size (larger $n \to$ smaller sample mean standard error, bigger n increases power), pop var (higher power if smaller pop var), significance level (higher α increases power).

One sample t-test: Key assumption for z-test is that pop var is known (usually isn't). Let $X_1, ..., X_n$ be indep $N(\mu, \sigma^2)$ RVs where σ^2 is unknown. Testing:

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$$

Test stat: $T = \frac{\overline{X}-\mu_0}{\sqrt{\frac{S^2}{n}}}$ where $\overline{X}$ and S^2 are unbiased sample mean and variance estimators respectively. Know that $T \sim t_{n-1}$, and critical value of $100\alpha\%$ is t_0 s.t.

$$\mathbb{P}(\text{Reject } H_0 \mid H_0 \text{ true}) = \mathbb{P}(\mid T \mid \geq t_0 \mid T \sim t_{n-1}) = \alpha \Rightarrow \mathbb{P}(T \geq t_0 \mid T \sim t_{n-1}) = \frac{\alpha}{2}$$

Paired t-test:

Case with two sets of samples: $X_i, ..., X_n$ from some population with expectation μ_X and $Y_1, ..., Y_m$ from second population with expectation μ_Y. Find out whether the expectations are equal or differ by some specified amount. Paired data is $(X_1, Y_1), ..., (X_n, Y_n)$. Define $D_i = X_i - Y_i$. Assume differences are independent & $D_i \sim N(\mu_D, \sigma_D^2)$ where $\mu_D = \mu_X - \mu_Y$.

$$H_0 : \mu_D = 0 \qquad H_1 : \mu_D \neq 0$$

$$T = \frac{\overline{D} - 0}{\sqrt{\frac{S_D^2}{n}}} \sim t_{n-1} \text{ under } H_0$$

C.R. is at $t_{n-1;\frac{\alpha}{2}}$.

Two sample t-test:

Next we consider case where data is not paired. Take $X_1, ..., X_n$ and $Y_1, ..., Y_m$, with μ_X, μ_Y, looking is μ_X is larger than μ_Y by some amount Δ_0.

$$H_0 : \mu_X - \mu_Y = \Delta_0 \qquad H_1 : \mu_X - \mu_Y \neq \Delta_0$$

Assume all samples are independent normal RVs:

$$X_i \sim N(\mu_X, \sigma^2), i = 1, ..., n$$
$$Y_j \sim N(\mu_Y, \sigma^2), j = 1, ..., m$$

Assume pop var is the same for both samples.

$$\overline{X} \sim N\left(\mu_X, \frac{\sigma^2}{n}\right) \qquad \overline{Y} \sim N\left(\mu_Y, \frac{\sigma^2}{m}\right)$$
$$\overline{X} - \overline{Y} \sim N\left(\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

Standardise and apply null hypothesis:

$$Z = \frac{\left(\overline{X} - \overline{Y}\right) - \Delta_0}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim N(0,1)$$

If we knew σ^2 then we could use above definition and do a two sample z-test, but typically σ^2 needs estimating:

Pooled sample variance:

Two samples $X_1, ..., X_n$ and $Y_1, ..., Y_m$. Sample variance estimator for each sample are S_X^2, S_Y^2. Pooled sample variance:

$$S_P^2 = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2 + \sum_{j=1}^m \left(Y_j - \overline{Y}\right)^2}{m+n-2} = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{m+n-2}$$

It is an unbiased estimator for common pop var.

$$T = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{S_P^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{m+n-2}$$

Test stat under H_0 for two sample t-test.

F-test for equal variances

Incase the assumption that the variances for the samples are the same is incorrect. Let σ_X^2, σ_Y^2 be pop vars for $X_1, ..., X_n$ and $T_1, ..., Y_m$. Testing

$$H_0 : \sigma_X^2 = \sigma_Y^2 \qquad H_1 : \sigma_X^2 \neq \sigma_Y^2$$

Know that estimators S_X^2, S_Y^2 are two indep chi-squared RVs.

F-distribution:

Let U, V be indep RVs s.t. $U \sim \chi_m^2, V \sim \chi_n^2$. Define F to be scaled ratio between these RVs, then new RV F follows F-dist with m and n degrees of freedom.

$$F = \frac{\frac{U}{m}}{\frac{V}{n}} \sim F_{m,n}$$

F-test stat: $F = \frac{\frac{S_X^2}{S_Y^2}\frac{\sigma_Y^2}{\sigma_X^2}}{\frac{S_X^2}{S_Y^2}} \sim F_{n-1,m-1}$ because under H_0, $\frac{\sigma_Y^2}{\sigma_X^2} = 1$

Critical regions are $F_{n,m;1-\frac{\alpha}{2}}, F_{n,m;\frac{\alpha}{2}}$. Easily found in R with qf(p probability, df1, df2)

Assessing Normality: All above tests assume that the sample RVs are normally distributed, but this should be checked.

Empirical Cumulative distribution function: $\hat{F}_X(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}[x_i \leq x]$, where $\mathbb{I}$ returns 1 if statement is true, 0 if false.

QQ plot should have an S shape?

3 Linear Regression

Simple linear model: explanatory variable x and output random response variable Y: $\mathbb{E}[Y] = \alpha + \beta x$ OR $Y = \alpha + \beta x + \varepsilon$. The intercept α and slope β are the regression parameters (unknown, to be estimated.

Errors: Assume $\mathbb{E}[\varepsilon] = 0$. Represents error in definition; errors are assumed indep.

Method of Least Squares: Consider linear model with single explanatory variable: $\mathbb{E}[Y] = \alpha + \beta x$, and suppose we have dataset $(x_1, x_2), ..., (x_n, y_n)$. The least squares estimate $\hat{\alpha}, \hat{\beta}$ that minimise sum of squares formula:

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$
$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

Both estimators are unbiased and consistent in estimating regression parameters.

$$\text{Var}(\hat{\alpha}) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}}{(n-1)s_x^2}\right) \qquad \text{Var}\left(\hat{\beta}\right) = \frac{\sigma^2}{(n-1)s_x^2}$$
$$S.E.(\hat{\alpha}) = \sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}} \qquad S.E.\left(\hat{\beta}\right) = \frac{\sigma}{\sqrt{(n-1)s_x^2}}$$

Fitted vals & Residuals: For model $Y_i = \alpha + \beta x_i + \varepsilon_i$ where $\mathbb{E}[\varepsilon_i] = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$. Compute least squares estimates $\hat{\alpha}, \hat{\beta}$ for dataset $(x_1, y_1), ..., (x_n, y_n)$. The fitted / predicted values $\hat{y}_1, ..., \hat{y}_n$ are response variables from $y_i = \hat{\alpha} + \hat{\beta}x_i$. Residuals are difference between observed and predicted response values: $e_i = y_i - \hat{y}_i$.

Residual variance: Residual sum of squares is sum of squared errors between the observed responses y_i and predicted $\hat{y}_i$ from simple linear model. Define the least squares estimate of σ^2 (residual variance):

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

It follows that $s_e = \sqrt{s_e^2}$ defines the residual standard deviation estimate for σ. The residual variance estimator $S_e^2 = \frac{1}{n-2}\sum_{i=1}^n \left(Y_i - \hat{Y}_i\right)^2$ is unbiased in estimating σ^2.

$$\widehat{S.E.}(\hat{\alpha}) = s_e\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}} \qquad \widehat{S.E.}\left(\hat{\beta}\right) = \frac{s_e}{\sqrt{(n-1)s_x^2}}$$

Normal linear regression model: Often need to test hypotheses about regression parameters / construct associated confidence intervals. So, need to determine distribution of least-squares estimators. To do this, apply additional assumption on responses by supposing the RVs follow indep normal distributions. The normal linear regression model:

$$Y_i \overset{\text{indep}}{\sim} N(\alpha + \beta x_i, \sigma^2)$$

Equivalently write $Y_i = \alpha + \beta x_i + \varepsilon_i$ where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

Maximum Likelihood estimates for α, β: The ML estimates for the regression parameters α, β in the normal linear regression model are the least squares estimates $\hat{\alpha}, \hat{\beta}$.

pdf for y_i at x_i:

$$f(y_i \mid x_i, \alpha, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}}$$

$$\Rightarrow L(\alpha, \beta, \sigma^2 \mid \boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^n f(y_i \mid x_i, \alpha, \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2}$$

$$\Rightarrow l(\alpha, \beta, \sigma^2 \mid \boldsymbol{x}, \boldsymbol{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}Q(\alpha, \beta)$$

Here, $Q(\alpha, \beta)$ is the least square estimate. It follows that maximising the log-likelihood function for α and β is equivalent to minimising $Q(\alpha, \beta)$.

Distribution of $\hat{\alpha}$ and $\hat{\beta}$: For the normal linear regression model:

$$\hat{\alpha} \sim N\left(\alpha, \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right]\right) \qquad \hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{(n-1)s_x^2}\right)$$

The distribution of the residual variance estimator is:

$$\frac{(n-2)S_e^2}{\sigma^2} \sim \chi_{n-2}^2$$

Both estimators $\hat{\alpha}, \hat{\beta}$ are independent of S_e^2 and so their marginal distributions are:

$$\frac{\hat{\alpha} - \alpha}{\sqrt{S_e^2\left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right)}} \sim t_{n-2} \qquad \frac{\hat{\beta} - \beta}{\sqrt{\frac{S_e^2}{(n-1)s_x^2}}} \sim t_{n-2}$$

Although the least squares estimators are independent of S_e^2, $\hat{\alpha}$ and $\hat{\beta}$ are not independent of each other. Above results enable calculation of confidence intervals, e.g. 95% CI for α and β:

$$\hat{\alpha} \pm t_{n-2;0.025}\sqrt{s_e^2\left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right)} \qquad \hat{\beta} \pm t_{n-2;0.025}\sqrt{\frac{s_e^2}{(n-1)s_x^2}}$$

Regression in R:

Example:

```
osmosis_data <- data.frame(
  concentration = c(0, 1, 2, 3, 4, 5), # explanatory variable, x
  weight_gained = c(130, 80, -70, -140, -170, -190) # response variable, y
)
osmosis_model <- lm(formula = weight_gained ~ concentration, data = osmosis_data)
osmosis_model
```

This outputs:

```
Call:
lm(formula = weight_gained ~ concentration, data = osmosis_data)
Coefficients:
(Intercept) concentration
    112.86       -69.14
```

For more info:

```
summary(osmosis_model)
```

Outputs

```
Call:
lm(formula = weight_gained ~ concentration, data = osmosis_data)
Residuals:
   1       2       3       4       5       6
17.143 36.286 -44.571 -45.429 -6.286 42.857
Coefficients:
             Estimate  Std.Error  t value  Pr(>|t|)
(Intercept)   112.86     31.42     3.592  0.02292 *
concentration -69.14     10.38     -6.663 0.00264 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 43.41 on 4 degrees of freedom
Multiple R-squared: 0.9174, Adjusted R-squared: 0.8967
F-statistic: 44.4 on 1 and 4 DF, p-value: 0.002635
```

Interpreting summary output:

Residuals: Residuals $e_1, ..., e_n$. Useful for assessing reasonableness of model assumptions.

Coefficients: Gives least squares estimate for each regression parameter and corresponding Std Error. May ask if under normal linear model the intercept or explanatory variables are important components to the model. Can independently assess this:

$$H_0 : \alpha = 0 \qquad H_1 : \alpha \neq 0$$
$$H_0 : \beta = 0 \qquad H_1 : \beta \neq 0$$

Recall that $\hat{\alpha}, \hat{\beta}$ follow the t-distribution after standardisation on $n-2$ dof. The third column in the coefficients table (t value) is the observed statistics under the null hypothesis:

$$t_\alpha = \frac{\hat{\alpha}}{\widehat{S.E.}(\hat{\alpha})} \qquad t_\beta = \frac{\hat{\alpha}}{\widehat{S.E.}\left(\hat{\beta}\right)}$$

The last column, $\Pr(> \mid t \mid)$ gives corresponding p value of the two sided alternative t-test. Next to these values is a star rating, corresponding to significance codes underneath the coefficients table. In example, p-values for each test is less than 5% so sufficient to reject the null hypothesis.

Residual Standard Error: Estimated residual standard error, s_e, and corresponding dof, $n-2$.

Mulitple R-squared: Provides metric on how good a statistical model is in predicting response data.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

For linear regression model with single explanatory variable, the R^2 metric corresponds to the squared correlation between the response and explanatory data: $R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}$. Range is $0 \leq r^2 \leq 1$, where $r^2 = 0$ means model finds no relationship between explanatory and response variables.

F-statistic: Gives test statistic and dofs of the corresponding F -distribution for assessing the null hypothesis that all explanatory variables provide no descriptive potential in describing the variability in the response variable. Here, only one explanatory variable so F test is equiv to t test for parameter β.

Confidence and Prediction Intervals: Best point estimate for Y_0 based on a fitted linear regression model would be $\hat{\alpha} + \hat{\beta}x_0$. This can be interpreted as the best estimate of the line of best fit at x_0, or the best prediction for the response variable of a new data pair when the explanatory variable is x_0. Both point estimates are identical, but associated description of uncertainty is different. Two intervals of interest:

(i) CI for line of best fit, $\mathbb{E}[Y_0]$
(ii) Prediction interval for response RV Y_0

CI for $\mathbb{E}[Y_0]$: $\mathbb{E}[Y_0] = \alpha + \beta x_0, \hat{\mathbb{E}}[Y_0] = \hat{\alpha} + \hat{\beta}x_0 = \hat{Y} + \hat{\beta}(x_0 - \overline{x})$. $\text{Var}\left(\hat{\mathbb{E}}[Y_0]\right) = \sigma^2\left[\frac{1}{n} + \frac{(x_0-\overline{x})^2}{(n-1)s_x^2}\right]$. Estimating σ^2 by S_e^2 and standardizing obtain the following:

$$\frac{\hat{\mathbb{E}}[Y_0] - \mathbb{E}[Y_0]}{\sqrt{S_e^2\left(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{(n-1)s_x^2}\right)}} \sim t_{n-2}$$

So the 95% CI for $\mathbb{E}[Y_0]$, the value of the line of best fit at x_0 is:

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2;0.025}\sqrt{s_e^2\left(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{(n-1)s_x^2}\right)}$$

Prediction Interval for Y_0 : $Y_0 = \alpha + \beta x_0 + \varepsilon_0$ with indep error RV $\varepsilon_0 \sim N(0, \sigma^2)$. Best method to construct prediction RV $\hat{Y}_0$:

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta}x_0 + \varepsilon_0 = \hat{\mathbb{E}}[Y_0] + \varepsilon_0$$

Assume that $\mathbb{E}[\varepsilon_0] = 0$, it follows that the point estimate for prediction is $\mathbb{E}\left[\hat{Y}_0\right] = \hat{\alpha} + \hat{\beta}x_0 = \hat{\mathbb{E}}[Y_0]$.

$$\text{Var}\left(\hat{Y}_0\right) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{(n-1)s_x^2}\right)$$

So the 95% prediction interval for Y_0 is:

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2;0.025}\sqrt{s_e^2\left(1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{(n-1)s_x^2}\right)}$$

Intervals in R:

```
Usage: predict(object, newdata, interval = c("none", "confidence", "prediction"),
level = 0.95)
Inputs: object - the linear regression object from lm()
newdata - a data frame with the new explanatory variables x_0
interval - type of interval to compute
level - confidence level of the interval
```

Example

```
cars_model <- lm(formula = dist ~ speed, data = cars)
summary(cars_model)

Call:
lm(formula = dist ~ speed, data = cars)
Residuals:
Min       1Q     Median   3Q    Max
-29.069 -9.525   -2.272 9.215 43.201
Coefficients:
Estimate       Std.   Error  t value Pr(>|t|)
(Intercept) -17.5791  6.7584 -2.601  0.0123 *
speed         3.9324  0.4155  9.464  1.49e-12 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438
F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12
```

Checking Assumptions: The normal linear regression model specifies that $Y \overset{\text{iid}}{\sim} N(\alpha + \beta x_i, \sigma^2)$. The assumptions made in defining this statistical model are:

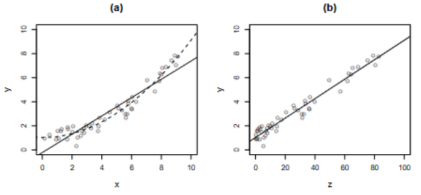
i) Independence between response variables

ii) Linearity of expectation, $\mathbb{E}[Y_i] = \alpha + \beta x_i$

iii) Homoscedasticity (i.e. they have the same variance)

iv) Normally distributed.

Before doing test, makes sense to plot data to check that a linear relationship exists at all - if not, consider simple transformation of x to form new explanatory variable:



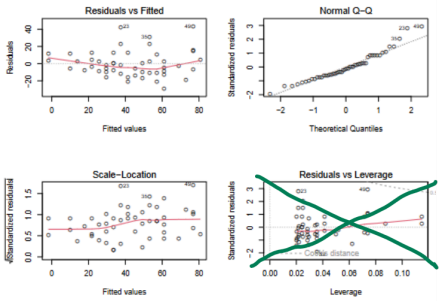
Left is y vs x with line of best fit $\mathbb{E}[Y] = \hat{\alpha}_a + \hat{\beta}_a x$. Plot y vs z where $z = x^2$ has line of best fit $\mathbb{E}[Y] = \hat{\alpha}_b + \hat{\beta}_b z$. Next need to check assumptions are reasonable; key lies in residuals, $e_i = y_i - \hat{y}_i$.

Residuals vs Fitted: Scatter of $e_1, ..., e_n$ against fitted values $\hat{y}_1, ..., \hat{y}_n$. If linear assumption $\mathbb{E}[Y] = \alpha + \beta x$ is true then residuals should be scattered randomly about zero without discernable pattern. Can also check homoscedasticity by seeing if range of scatter values is consistent.

Scale-Location: Scatter of transformation of standardized residuals $\sqrt{\mid e_1^* \mid}, ..., \sqrt{\mid e_n^* \mid}$ against fitted values $\hat{y}_1, ..., \hat{y}_n$. If homoscedastic assumption is true then points should be randomly scattered around 0.82.

Normal QQ-plot: Normal qq plot of standardised residuals $e_1^*, ..., e_n^*$. If assumption that response RVs are independently and normally distributed then can consider standardised residuals as a sample of n values from the standard normal distribution. HOw TO INTERPRET QQ?

```
par(mfrow = c(2, 2)) # Set-up 2x2 layout in plotting window
plot(cars_model)
```



4 Multiple Regression & Analysis of Variance

Extending Least Squares Estimation: Begin by extending linear regression model to incorporate more explanatory variables. Let $Y_1, ..., Y_n$ denote random response variables where each response Y_i may linearly depend on p different explanatory variables $x_{i,1}, ..., x_{i,p}$:

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_p x_{i,p} + \varepsilon_i$$

Coefficients $\beta_1, ..., \beta_p$ are regression parameters and ε_i are indep random errors satisfying conditions $\mathbb{E}[\varepsilon_i] = 0$, $\text{Var}(\varepsilon_i) = \sigma^2$.

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ x_{2,1} & \dots & x_{2,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \Rightarrow \underline{Y} = \boldsymbol{X}\underline{\beta} + \underline{\varepsilon}$$

Here $\underline{Y}$ is the response vector RV with observed values $\underline{y} = (y_1, ..., y_n)^T$, $\underline{\varepsilon}$ is the random error vector, the $n \times p$ matrix $\boldsymbol{X}$ is the *design matrix*, and $\underline{\beta}$ the regression parameter vector.

Least Squares Estimates: The least-squares vector estimate $\underline{\hat{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$ that minimises the formula $Q(\underline{\beta}) = (\underline{y} - \boldsymbol{X}\underline{\beta})^T (\underline{y} - \boldsymbol{X}\underline{\beta})$ is:

$$\underline{\hat{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\underline{y}$$

```
osmosis_data <- data.frame(
concentration = c(0, 1, 2, 3, 4, 5), # explanatory variable 1
surface_area = c(22, 21, 19, 19, 20, 22), # explanatory variable 2
weight_gained = c(130, 80, -70, -140, -170, -190) # response
variable, y
)
osmosis_model_new <- lm(formula = weight_gained ~ concentration +
surface_area,
data = osmosis_data)
osmosis_model_new
Call:
lm(formula = weight_gained ~ concentration + surface_area, data =
osmosis_data)
Coefficients:
(Intercept) concentration surface_area
-431.52 -66.89 26.28
```

The fitted equation is then $\mathbb{E}[Y_i] = -431.52 - 66.89x_{i,1} + 26.28x_{i,2}$ (to 2 d.p.). This equation describes the 'plane' of best fit.

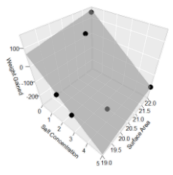


Figure 5.1: A 3D scatter plot of the osmosis data with the 'plane' of best fit.

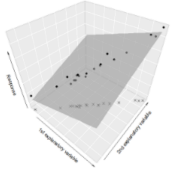


Figure 5.2: A 3D scatter plot with 'plane' of best fit to illustrate multicollinearity. Solid points represent the 3D co-ordinate of the data $(x_{i,1}, x_{i,2}, y_i)$, with the cross points presenting the projection of the data into the plane of the explanatory variables.

In choosing the explanatory variables for use, we need to ensure that the columns of the resulting design matrix are linearly independent. Ideally, if we can ensure that the explanatory variables are linearly independent, we want them to not be highly statistically correlated because the fitted regression line / plane will be highly unstable (even though we can derive the least squares estimation). This problem related to high correlation within the explanatory variables is called multicollinearity. To reduce the risk of incurring multicollinearity issues and to produce a more stable fitted regression line, ideally we would only like to use explanatory variables in the regression model that are statistically independent (i.e. have low sample correlation between the variables).

Confidence Interval: May want to quantify uncertainty in regression parameters. First, need to apply additional assumption of the errors and assume that they are independent normal RVs. So, we can express the multiple regression model as:

$$Y_i \overset{\text{iid}}{\sim} N(\beta_i x_{i,1}, \beta_2 x_{i,2}, ..., \beta_p x_{i,p}, \sigma^2)$$

Corresponding MLE for $\underline{\beta}$ are the least squares estimates $\underline{\hat{\beta}}$. An unbiased estimator for the error variance σ^2 is defined by the residual variance estimator:

$$S_e^2 = \frac{1}{n-p}\sum_{i=1}^{n}\left[Y_i - \left(\hat{\beta}_i x_{i,1}, \hat{\beta}_2 x_{i,2}, ..., \hat{\beta}_p x_{i,p}\right)\right]^2$$

$$= \frac{1}{n-p}\left(\underline{y} - \boldsymbol{X}\underline{\hat{\beta}}\right)^T\left(\underline{y} - \boldsymbol{X}\underline{\hat{\beta}}\right)$$

The marginal distribution for the least squares estimator $\hat{\beta}_j$ is:

$$\frac{\hat{\beta}_j - \beta_j}{S.E.\left(\hat{\beta}_j\right)} \sim t_{n-p}$$

Here, $S.E.\left(\hat{\beta}_j\right)$ is the standard error in the least_squares estimator for β_j. The 95% confidence interval is:

$$\hat{\beta}_j \pm t_{n-p;0.025}\widehat{S.E.}\left(\hat{\beta}_j\right)$$

```
summary(osmosis_model_new)
Call:
lm(formula = weight_gained ~ concentration + surface_area, data =
osmosis_data)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -431.524 128.197 -3.366 0.043536 *
concentration -66.890 4.534 -14.753 0.000676 ***
surface_area 26.280 6.154 4.271 0.023565 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 18.84 on 3 degrees of freedom
Multiple R-squared: 0.9883, Adjusted R-squared: 0.9805
F-statistic: 127 on 2 and 3 DF, p-value: 0.001261
```

Comparing the the results in 'Regression in R', residual standard error is lower ($s_e = 18.84$ compared to 43.41). The R^2 metric is larger (0.9883 vs 0.9174) without the surface area explanatory variable. A better measure for comparing models is the adjusted R^2 metric, because it scales to account for the number of explanatory variables:

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p}$$

ANOVA F-test for model comparison: In multiple regression it is common to test whether specified regression coefficients are zero, representing the intuitive idea of whether a given explanatory variable 'has no effect' on the response. Define the full model with p_1 explanatory variables as

$$Y_i \overset{\text{iid}}{\sim} N\left(\beta_1 x_{i,1} + ... + \beta_{p_0} x_{i,p_0} + \beta_{p_0+1} x_{i,p_0+1} + ... + \beta_{p_1} x_{i,p_1}, \sigma^2\right)$$

Interested in assessing if simplified *sub-model* with just for p_0 explanatory variables is as good at explaining variation in observed data.

$$H_0 : \beta_{p_0+1} = ... = \beta_{p_1} = 0 \qquad H_1 : \text{No restriction on regression coeffs}$$

Basically want to see if full model gives significantly better fit to data than the sub-model. Measure the goodness of fit via *residual sum of squares* or RSS. For the full model with least-squares vector $\underline{\hat{\beta}_1} = \left(\hat{\beta}_1, ..., \hat{\beta}_{p_1}\right)^T$, RSS is:

$$\text{RSS}_1 = \sum_{i=1}^{n}\left(Y_i - \left[\hat{\beta}_1 x_{i,1} + ... + \hat{\beta}_{p_1} x_{i,p_1}\right]\right)^2 = (n-p_1)S_1^2$$

Here, S_1^2 denotes the residual variance estimator for the full model. RSS for sub model under null hypothesis is defined as $\text{RSS}_0 = (n-p_0)S_0^2$ where S_0^2 corresponds to sub-model's residual variance estimator.

$$\frac{\text{RSS}_0}{\sigma^2} = \frac{(n-p_0)S_0^2}{\sigma^2} \sim \chi^2_{n-p_0} \qquad \frac{\text{RSS}_1}{\sigma^2} = \frac{(n-p_1)S_1^2}{\sigma^2} \sim \chi^2_{n-p_1}$$

If the null hypothesis is true then the following holds:

$$\frac{\text{RSS}_0}{\sigma^2} \sim \chi^2_{n-p_0}; \qquad \frac{\text{RSS}_0 - \text{RSS}_1}{\sigma^2} \sim \chi^2_{p_1-p_0}$$

And $\text{RSS}_0 - \text{RSS}_1$ is independent of RSS_1. Hence from the F-dsitribution:

$$F = \frac{(\text{RSS}_0 - \text{RSS}_1)/(p_1 - p_0)}{\text{RSS}_1/(n-p_1)} \sim F_{p_1-p_0, n-p_1}$$

Under H_0.

ANOVA F-Test: Consider the linear regression model $\underline{Y} = \boldsymbol{X}\underline{\beta} + \underline{\varepsilon}$ where $\varepsilon_1 \overset{\text{iid}}{\sim} N(0, \sigma^2)$ and $\beta = \left(\beta_1, ..., \beta_{p_0}, \beta_{p_0+1}, ..., \beta_{p_1}\right)$. We wish to assess whether some explanatory variables have negligible importance in describing the the variability in the response variable. WLOG, suppose that these explanatory variables correspond to the last $p_1 - p_0$ columns of the design matrix $\boldsymbol{X}$, resulting in the following null and alternative hypotheses:

$$H_0 : \beta_{p_0+1} = ... = \beta_{p_1} = 0 \quad H_1 : \exists j \in \{p_0+1, ..., p_1\} \mid \beta_j \neq 0$$

We consider the F -test statistic where, under H_0:

$$F = \frac{(\text{RSS}_0 - \text{RSS}_1)/(p_1 - p_0)}{\text{RSS}_1/(n-p_1)} \sim F_{p_1-p_0, n-p_1}$$

with $\text{RSS}_1 = (n-p_1)S_1^2$ and $\text{RSS}_0 = (n-p_0)S_0^2$ are the residual sum of squares under the full and sub models. The above hypotheses can be expresed in terms of a one-sided alternative:

$$H_0 : F = 0 \text{ vs } H_1 : F > 0.$$

For a α% sf, the CR for rejecting H_0 is then:

$$C = \left[F_{p_1-p_0, n-p_1; \alpha}, \infty\right)$$

where the critical value can be computed in R by:

```
qf(1 - alpha, df1 = p1 - p0, df2 = n - p1)
```

Alternatively, the p-value for the test given the observed test statistic f is $\mathbb{P}\left(F \geq f \mid F \sim F_{p_1-p_0, n-p_1}\right)$ and may be computed in R by:

```
1 - pf(f, df1 = p1-p0, df2 = n-p1)
```

EXAMPLE!!

Categorical Variables: Need to convert categorical data into something numerical in order to incorporate it into a linear regression model. To do this, introduce a sequence of binary explanatory variables.

Consider a categorical variable with k categories, then we construct a set of binary explanatory variables $\underline{a}_j = (a_{1,j}, ..., a_{n,j})^T$ for $j = 1, ..., k$ where for given sample i and category j:

$$a_{i,j} = \begin{cases} 1 \text{ if sample i belongs to category j} \\ 0 \text{ otherwise} \end{cases}$$

In R the set of binary vectors $\underline{a}_1, ..., \underline{a}_k$ are collectively referred to as a *factor variable* and is constructed using the factor command:

```
Usage: factor(x, levels)
Inputs: x - Character vector containing the categorical data.
levels - Character vector of the category names (if not specified,
then
names are taken from x and are placed in alphabetical order).
```

The factor variable vectors then form additional columns in the design matrix along with any other explanatory variables, e.g. $\boldsymbol{X} = \left(\underline{a}_1, ..., \underline{a}_k, \underline{x}_1, ..., \underline{x}_2\right)$. The corresponding regression parameter for the binary vector for category j (the coefficient vector $\underline{a}_j$) would then refer to the expected response when the value of all other numerical explanatory variables at zero, the category specific intercept.

$$\mathbb{E}[Y_i] = \alpha_1 a_{i,1} + ... + \alpha_k a_{i,k} + \beta_1 x_{i,1} + \beta_2 x_{2,i}$$

$$= \alpha_j + \beta_1 x_{i,1} + \beta_2 x_{2,i} \text{ if sample i belongs to category j}$$

```
blood_type_char <- c("A", "AB", "O", "B", "A", "A", "O", "O", "O",
"O", "A", "O", "A", "O", "O", "O", "O", "A", "O", "B")
blood_type = factor(blood_type_char, levels = c("O", "A", "B",
"AB"))
blood_type
[1] A AB O B A A O O O O A O A O O O O A O B
Levels: O A B A
```

To view the associated columns that will form part of the design matrix when perform a linear regression model, we can use the model.matrix() command as follows:

```
model.matrix( ~ blood_type)
  (Intercept) blood_typeA blood_typeB blood_typeAB
1           1           0           0            0
2           1           0           0            1
3           1           0           0            0
...
20          1           0           1            0
```

Although there are four columns, the vector corresponding to blood type 'O' is missing but is instead replaced with (Intercept) where all of the entries are 1. To view all four anticipated binary vectors associated with each blood type categories we need to include -1 into the formula:

```
model.matrix( ~ -1 + blood_type)
  blood_type O blood_typeA blood_typeB blood_typeAB
1           0           1           0            0
2           0           0           0            1
3           1           0           0            0
...
20          0           0           1            0
```

An important observation about the second matrix (and for all factor variables) is that $\underline{a}_1 + ... + \underline{a}_4 = 1$. This can be problematic when forming a design matrix with other categorical variables as the design matrix may have linearly dependent columns, preventing a unique solution to estimating the least-squares estimates.

Strategy to address is to ensure that the design matrix has a ones-vector, i.e. the 'intercept' column where where all entries are 1, and include all but one of the binary vectors for each categorical variable in the model, as demonstrated in the first case above. This strategy ensures columns of $\mathbf{X}$ being linearly independent.

If we do this then the columns do not sum to the ones vector, so we can add additional explanatory variables without concern for columns of $\boldsymbol{X}$ being linearly independent. May then add further numerical explanatory variables to define following regression model:

$$\mathbb{E}[Y_i] = \alpha + \beta_1 a_{i,A} + \beta_2 a_{i,B} + \beta_3 a_{i,AB} + \beta_4 x_i$$

If participant's blood type is O then model simplifies to $\mathbb{E}[Y_i] = \alpha + \beta_4 x_i$ whereas if it's AB then the model is $\mathbb{E}[Y_i] = (\alpha + \beta_3) + \beta_4 x_i$. Equations describe parallel regression lines with diff intercepts. Parameter α is intercept term for category 'O', parameter β_4 is diff in intercept from category 'O' to category 'AB'. As the intercepts asso- ciated with blood types 'A', 'B' and 'AB' depends on the intercept for blood type 'O', refer to first category in the factor variable (i.e. the binary vector that was ignored) as the *baseline group*.

One-way ANOVA (single categorical explanatory variable): Common task is to compare several population means. Usual way is *one-way analysis of variance* which is a special case of ANOVA. Consider k populations with expectations $\mu_1, ..., \mu_k$. Testing hypotheses:

$$H_0 : \mu_1 = ... = \mu_k [= \mu] \quad H_1 : \mu_1, ..., \mu_k \text{ not all equal}$$

Purpose is to generalise the multi-sample t-test. Suppose random samples of sizes $n_1, ..., n_k$ for the k categories with overall sample size $\sum_{j=1}^{k} n_j$. Let $y_{i,j}$ denote ith observation belonging to jth category. Suppose $y_{i,j}$ is a realised value of RV $Y_{i,j}$ where we assume that:

$$Y_{i,j} \overset{\text{iid}}{\sim} N(\mu_j, \sigma^2)$$

With assumption of independent normal random variables with equal variance. Above model is special case of multiple regression model with single categorical variable. To represent the above defi- nition as a multiple linear regression model requires the formation of a set of binary vectors that converts the categorical variable into a factor variable.

Stack all of the response RVs into a single vector $\underline{Y} = \left(Y_{1,1}, ..., Y_{n_1,1}, ..., Y_{1,k}, ..., Y_{n_k,k}\right)^T$ and then construct binary vectors $\underline{x}$ with elements given as follows based on corresponding entry $\underline{Y}$:

$$x_{i,j;r} = \begin{cases} 1 \text{ if RV } Y_{i,j} \text{ refers to category } r \\ 0 \text{ otherwise} \end{cases}$$

Constructing the design matrix as $\boldsymbol{X} = (\underline{x}_1, ..., \underline{x}_k)$ with parameter vector $\underline{\mu} = (\mu_1, ..., \mu_k)^T$ obtains:

$$\mathbb{E}[\underline{Y}] = \boldsymbol{X}\underline{\mu} \Leftrightarrow \mathbb{E}\left[Y_{i,j}\right] = \mu_i x_{i,j;1} + ... + \mu_k x_{i,j;k}$$

This describes a mult regression model that has no intercept term and regression coeffient denotes the population expectation for the respective group. The least squares estimates for these parameters equate to the group sample means: $\hat{\mu}_j = \overline{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{i,j}$.

The residual variance estimate is then

$$s_e^2 = \frac{1}{n-k}\sum_{j=1}^{k}\sum_{j_{i=1}}^{n}\left(y_{i,j} - \overline{y}_j\right)^2$$

Note: least squares estimate from R doesn't correspond to model above, instead considers alternative parameterisation that includes an intercept term corresponding to baseline case: achieved by replacing first column of design matrix with ones-vector to give $\overline{\boldsymbol{X}} = (\underline{1}, \underline{x}_1, ..., \underline{x}_2)$ and notate parameter vector as $\underline{\beta} = (\mu_1, \beta_2, ..., \beta_k)^T$. Resulting regression model:

$$\mathbb{E}[\underline{Y}] = \overline{\boldsymbol{X}} \Leftrightarrow \mathbb{E}\left[Y_{i,j}\right] = \mu_1 + \beta_2 x_{i,j;2} + ... + \beta_k x_{i,j;k}$$

Here $\beta_j = \mu_j - \mu_1$. Intercept parameter μ_1 is pop expectation for first group or *baseline category* with β_j representing difference in expectations between considered and baseline categories. Least squares estimates are then $\hat{\mu}_1 = \hat{y}_1, \hat{\beta}_j = \hat{\mu}_j - \hat{\mu}_1 = \hat{y}_j - \hat{y}_1$.

Conducting test: Null hypothesis of equal group poplation expectations $\mu_1 = ... = \mu_k$. Full model has $p_1 = k$ parameters, sub-model has p_0 parameters (overall pop expectation μ). Need to find residual sum of squares to perform F-test.

$$\sum_{j=1}^{k}\sum_{j_{i=1}}^{n}\left(Y_{i,j} - \overline{Y}\right)^2 = \sum_{j=1}^{k} n_j\left(\overline{Y}_j - \overline{Y}\right)^2 + \sum_{j=1}^{k}\sum_{j_{i=1}}^{n}\left(y_{i,j} - \overline{y}_j\right)^2$$

$$\text{RSS}_y = \text{RSS}_b + \text{RSS}_e$$

$$(n-1)S_y^2 = (k-1)S_b^2 + (n-k)S_e^2$$

- RSS_e: error residual sum of squares. Describes variation within groups around respective group sample means.
- RSS_b: residual sum of squares between group means.

ANOVA test stat:

$$F = \frac{\text{RSS}_b/(k-1)}{\text{RSS}_e/(n-k)} = \frac{S_b^2}{S_e^2} \sim F_{k-1;n-k}$$

Under H_0. Basically ratio between unbiased sample variance estimators for the explained, S_b^2, vs unexplained, S_e^2, by the regression model. If H_0 is false then expect sample group means to be different to overall sample mean, so large RSS_b in comparison to RSS_e.

In R:

```
Usage: anova(object)
Inputs: object - The object returned by the lm() command.
```

This command prints:

	Deg. free.	Res. Sum Sq.	Mean Sum Sq.	F-value	p-value
Categorical variable	$k-1$	$\text{RSS}_b = \text{RSS}_y - \text{RSS}_e$	$S_b^2 = \text{RSS}_b/(k-1)$	$F = S_b^2/S_e^2$	p
Residuals	$n-k$	RSS_e	$S_e^2 = \text{RSS}_e/(n-k)$		
Total	$n-1$	RSS_y	$S_y^2 = \text{RSS}_y/(n-1)$		

One-way ANOVA is then concluded by comparing F-test stat against appropriate critical value / comparing p-value against specified critical value.

Example:

```
Lab1 2.9 3.1 3.1 3.7 3.1 4.2 3.7 3.9 3.1 3.0 2.9
Lab2 2.7 3.4 3.6 3.2 4.0 4.1 3.8 3.8 4.3 3.4 3.3
Lab3 3.3 3.3 3.5 3.5 2.8 2.8 3.2 2.8 3.8 3.5 3.8
Lab4 3.3 3.2 3.4 2.7 2.7 3.3 2.9 3.2 2.9 2.6 2.8
Lab5 4.1 4.1 3.7 4.2 3.1 3.5 2.8 3.5 3.7 3.5 3.9

flammable_data <- data.frame(
burn_length = c(2.9, 3.1, 3.1, 3.7, 3.1, 4.2, 3.7, 3.9, 3.1, 3.0,
2.9,
2.7, 3.4, 3.6, 3.2, 4.0, 4.1, 3.8, 3.8, 4.3, 3.4, 3.3,
3.3, 3.3, 3.5, 3.5, 2.8, 2.8, 3.2, 2.8, 3.8, 3.5, 3.8,
3.3, 3.2, 3.4, 2.7, 2.7, 3.3, 2.9, 3.2, 2.9, 2.6, 2.8,
4.1, 4.1, 3.7, 4.2, 3.1, 3.5, 2.8, 3.5, 3.7, 3.5, 3.9),
lab = factor(rep(c("Lab1", "Lab2", "Lab3", "Lab4", "Lab5"), each =
11)))
boxplot(burn_length ~ lab, data = flammable_data)
mean(flammable_data$burn_length)
[1] 3.376364
tapply(flammable_data$burn_length, flammable_data$lab, mean)
  Lab1      Lab2      Lab3      Lab4      Lab5
3.336364 3.600000 3.300000 3.000000 3.645455
```

From output, overall sample mean burn length cross all labs is 3.38cm, but lab 4 has shortest ave and lab 5 the longest. Let $\mu_1, ..., \mu_5$ denote expected burn lengths from labs:

$$H_0 : \mu_1 = ... = \mu_5 \quad H_1 : \text{At least one is different}$$

```
flammable_model <- lm(burn_length ~ lab, data = data)
anova(flammable_model)
Analysis of Variance Table
Response: burn_length
          Df  Sum Sq  Mean Sq  F value   Pr(>F)
lab        4  2.9865  0.74664  4.5346   0.003337 **
Residuals 50  8.2327  0.16465
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> qf(1 - 0.05, df1 = 4, df2 = 50)
[1] 2.557179
```

See that observed test stat is $f = 4.54$. Corresponding p-value under null hypothesis is 0.0033. From table see that the 2 dofs for the F distribution are 4 and 50 respectively, from which can evaluate critical value at 5% sl to be 2.55. As observed value is greater than critical value, and equivalently p-value is less than 5%, there is enough evidence to reject null hypothesis of equal group expectations.

Two-way ANOVA (two categorical explanatory variable:

Prev supposes each observation belongs to one group from a single set of categories, but observations are often cross-classified. Consider the case where each observation is classified according to two categorical variables and perform a two-way ANOVA.

- label two categorical variables as *Block* and *Treatment* variable, & notate that there are b block groups and k treatments groups.
- $Y_{i,j}$ is response RV corresponding to block i, treatment j.
- Consider case where there is only one observation per block & treatment combo, so saample size is bk.
- Assume the RVs follow indep normal distributions with unique expectation $\mu_{i,j}$ but common variance σ^2 : $Y_{i,j} \overset{\text{iid}}{\sim} N(\mu_{i,j}, \sigma^2)$.
- Model for $\mathbb{E}\left[Y_{i,j}\right] = \mu_{i,j}$: $\mu_{i,j} = \mu + \alpha_i + \beta_j$, where μ is overall pop expectation across groups, α_i is teh i^{th} block group effect and β_j the j^{th} treatment group effect
- The group effect parameters describes how a particular block/treatment group's expectation differs from the overall expectation μ.

Need to look at this model carefully for some constant $c \in \mathbb{R}$:

$$\mu_{i,j} = \mu + \alpha_i + \beta_j = \mu + (\alpha_i - c) + (\beta_j + c) = \mu + \alpha_i^* + \beta_j^*$$

So decomposition of expectation in terms of block / treatment effects is *not unique*: $\mu_{i,j}$ can equally be expresed in terms of α_i, β_j or the shifted version α_i^*, β_j^*.

Issue must be resolved by placing additional constraints on the parameters:

$$\sum_{i=1}^{b}\alpha_i = 0 \quad \sum_{j=1}^{k}\beta_j = 0$$

Under these constraints, the least squares estimates are:

$$\hat{\mu} = \hat{y} \quad \hat{\alpha}_i = \overline{y}_{i,\bullet} - \overline{y} \quad \hat{\beta}_i = \overline{y}_{\bullet,j} - \overline{y}$$

where $\overline{y}$ is the overall sample mean, $\overline{y}_{i,\bullet}$ is the sample mean within the i^{th} block group and $\overline{y}_{\bullet,j}$ is the sample mean within the j^{th} treatment group:

$$\overline{y}_{i,\bullet} = \frac{1}{k}\sum_{j=1}^{k} y_{i,j} \quad \overline{y}_{\bullet,j} = \frac{1}{b}\sum_{i=1}^{b} y_{i,j} \quad \overline{y} = \frac{1}{k}\sum_{j=1}^{k}\overline{y}_{\bullet,j}$$

Observe the need for additional constraints on the parameters, where $b = 3, k = 2$:

$$\mathbb{E}[\underline{Y}] = \begin{pmatrix} \mu_{1,1} \\ \mu_{2,1} \\ \mu_{3,1} \\ \mu_{1,2} \\ \mu_{2,2} \\ \mu_{3,2} \end{pmatrix} = \begin{pmatrix} \mu + \alpha_1 + \beta_1 \\ \mu + \alpha_2 + \beta_1 \\ \mu + \alpha_3 + \beta_1 \\ \mu + \alpha_1 + \beta_2 \\ \mu + \alpha_2 + \beta_2 \\ \mu + \alpha_3 + \beta_2 \end{pmatrix} = \left(\begin{array}{c|cc|c} 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{array}\right) \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

The design matrix is linearly independent (as sum of column vectors within each dashed partition equate to the ones-vector)

Test: There are two possible tests to consider:

1. Is there any evidence of a difference in expected response between the block groups? Suppose that the expected response only depends on which treatment group it belongs to, so propose following:

$$H_0 : \alpha_1 = ... = \alpha_b = 0 \quad H_1 : \text{At least one } \alpha_i \text{ does not equal } 0$$

2. Is there any evidence of a difference in expected response between the treatment groups?

$$H_0 : \beta_1 = ... = \beta_b = 0 \quad H_1 : \text{At least one } \beta_i \text{ does not equal } 0$$

want to partition the overall variability in the data (RSS_y) ito three parts: that which is explained by the block group sample means, RSS_b, that explained by the treatment group sample means, RSS_t, and the remaining unexplained variability, RSS_e:

$$\sum_{i=1}^{b}\sum_{j=1}^{k}\left(Y_{i,j}-\overline{Y}\right)^2$$
$$= k\sum_{i=1}^{b}\left(\overline{Y}_{i,\bullet}-\overline{Y}\right)^2 + b\sum_{j=1}^{k}\left(\overline{Y}_{\bullet,j}-\overline{Y}\right)^2 + \sum_{i=1}^{b}\sum_{j=1}^{k}\left(Y_{i,j}-\overline{Y}_{i,\bullet}-\overline{Y}_{\bullet,j}+\overline{Y}\right)^2$$
$$\Rightarrow \text{RSS}_y = \text{RSS}_b + \text{RSS}_t + \text{RSS}_e$$
$$\Rightarrow (n-1)S_y^2 = (b-1)S_b^2 + (k-1)S_t^2 + (n-b-k+1)S_e^2$$

From this partitioning, two ANOVA F-tests can be performed where under the respective null-hypotheses:

$$F_b = \frac{\text{RSS}_b/(b-1)}{\text{RSS}_e/(n-b-k+1)} = \frac{S_b^2}{S_e^2} \sim F_{b-1;n-k-b+1}$$

$$F_t = \frac{\text{RSS}_t/(b-1)}{\text{RSS}_e/(n-b-k+1)} = \frac{S_t^2}{S_e^2} \sim F_{k-1;n-k-b+1}$$

In R: Similar to one-way:

```
lm(response ~ block_factor + treatment_factor, data = study_data)
```

	Deg. free.	Res. Sum Sq.	Mean Sum Sq.	F-value	p-value
Block	$b-1$	RSS_b	$S_b^2 = RSS_b/(b-1)$	$F_b = S_b^2/S_e^2$	p_b
Treatment	$k-1$	RSS_t	$S_t^2 = RSS_t/(k-1)$	$F_t = S_t^2/S_e^2$	p_t
Residuals	$n-k-b+1$	RSS_e	$S_e^2 = RSS_e/(n-k-b+1)$		
Total	$n-1$	RSS_y	$S_y^2 = RSS_y/(n-1)$		

The hypothesis tests for investigating evidence of block/treatment group effect can then be independently concluded by either comparing the respective f-test statistic against the corresponding one-sided critical value or evaluating significance of the stated p-values.

Example:

Reading test:

	age 9	age 10	age 11	Avg.
School *A*	68	89	86	81
School *B*	50	57	91	64
School *C*	51	65	73	63
School *D*	67	81	86	78
Avg.	59	73	84	72

```
reading_data <- data.frame(
score = c(68, 89, 86, 50, 57, 91, 51, 65, 73, 67, 81, 86),
school = factor(rep(c("SchA", "SchB", "SchC", "SchD"), each = 3),
levels = c("SchA", "SchB", "SchC", "SchD")),
age = factor(rep(c("Age9", "Age10", "Age11"), times = 4),
levels = c("Age9", "Age10", "Age11")))
 reading_model <- lm(formula = score ~ school + age, data =
reading_data)
> summary(reading_model)
Call:
lm(formula = score ~ school + age, data = reading_data)
Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  68.000     5.817      11.691   2.36e-05 ***
schoolSchB  -15.000     6.716      -2.233   0.0670 .
schoolSchC  -18.000     6.716      -2.680   0.0365 *
schoolSchD   -3.000     6.716      -0.447   0.6708
ageAge10     14.000     5.817       2.407   0.0528 .
ageAge11     25.000     5.817       4.298   0.0051 **
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 8.226 on 6 degrees of freedom
Multiple R-squared: 0.8283, Adjusted R-squared: 0.6851
F-statistic: 5.787 on 5 and 6 DF, p-value: 0.02705
```

Categories for school A and age group 9 are missing from output: form intercept baseline case whereby estimated expected reading score for age 9 pupils at school A is $\hat{\mu}_{A,9} = 68$. Estimated score for all other group combinations are then derived from this baseline, e.g. $\hat{\mu}_{A,11} = 68 + 25 = 93$.

F-test summary above is significant (p-value of $0.027 < 5\%$), this test compares the fitted model against the only sub-model (i.e. $H_0 : \alpha_1 = ... = \alpha_b = \beta_1 = ... = \beta_k = 0$). The info needed to assess each categorical variable individually is obtained from applying the anova() command:

```
anova(reading_model)
Analysis of Variance Table
Response: score
          Df  Sum Sq  Mean Sq  F value   Pr(>F)
school     3   702     234.00    3.4581  0.09154 .
age        2  1256     628.00    9.2808  0.01458 *
Residuals  6   406      67.67
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For both the school and age variables the p-value for each ANOVA F-test is below the 5% sig lvl. Concluding each test individually:

- Fail to reject the null hypothesis of equal mean effects from the four schools at the 5% significance level, suggesting the expected reading scores don't significantly depend on choice of school.
- There is sufficient evidence at the 5% level to reject the null hypothesis that states there is no differences in expected reading scores between the three age groups.

5 Useful Distributions

DRVS

Distribution	PMF	$\mathbb{E}[X]$, $\text{Var}(X)$
Bernoulli $X \sim \text{Bern}(p)$	$f_{X(x)} = p^x(1-p)^{1-x}$ $x \in [0,1]$	$\mathbb{E}[X] = p$ $\text{Var}(X) = p(1-p)$
Binomial $X \sim \text{Binom}(n,p)$	$f_{X(x)} = \binom{x}{n}p^x(1-p)^{n-x}$ $x \in \{0,1,...,n\}$	$\mathbb{E}[X] = np$ $\text{Var}(X) = np(1-p)$
Geometric $X \sim \text{Geom}(p)$	$f_{X(x)} = p(1-p)^{x-1}$ x in {1, 2, ...}	$\mathbb{E}[X] = \frac{1}{p}$ $\text{Var}(X) = \frac{1-p}{p}$
Poisson $X \sim \text{Pois}(\lambda)$	$f_X(x) = \lambda^x e^{-\frac{\lambda}{x!}}$ $x \in \{0,1,2,...\}$	$\mathbb{E}[X] = \lambda$ $\text{Var}(X) = \lambda$

CRVS

Distribution	PMF	$\mathbb{E}[X]$, $\text{Var}(X)$
Uniform $X \sim \text{Unif}(a,b)$	$f_{X(x)} = \frac{1}{b-a}$ $x \in [a,b]$	$\mathbb{E}[X] = \frac{a+b}{2}$ $\text{Var}(X) = \frac{(b-a)^2}{12}$
Exponential $X \sim \text{Exp}(\lambda)$	$f_{X(x)} = \lambda e^{-\lambda x}$ $x \in [0,\infty)$	$\mathbb{E}[X] = np$ $\text{Var}(X) = np(1-p)$
Normal $X \sim \text{N}(\mu,\sigma^2)$	$f_{X(x)} = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$ $x \in \mathbb{R}$	$\mathbb{E}[X] = \mu$ $\text{Var}(X) = \sigma^2$
Chi-squared $X \sim \chi^2_v$	$f_X(x) = \frac{1}{2^{v/2}\Gamma(\frac{v}{2})}x^{v/2-1}$ $x \in [0,\infty)$	$\mathbb{E}[X] = v$ $\text{Var}(X) = 2v$
Student's t $X \sim t_v$	$f_X(x) =$ $\frac{1}{\sqrt{v}B(\frac{v}{2},\frac{1}{2})}\left(1+\frac{x^2}{v}\right)^{-\frac{v+1}{2}}$ $x \in \mathbb{R}$	EE[X] = 0 $\text{Var}(X) = \frac{v}{v-2}(v>2)$
F $X \sim F(\lambda,v)$	$f_X(x) =$ $\frac{\lambda^{\frac{\lambda}{2}}v^{\frac{v}{2}}}{B(\frac{\lambda}{2},\frac{v}{2})}x^{\lambda/2-1}(\lambda x+v)^{-\frac{\lambda+v}{2}}$ $(v>4)$	$\mathbb{E}[X] = \frac{v}{v-2}(v>2)$ $\text{Var}(X) = \frac{2v^2(\lambda+v-2)}{\lambda(v-2)^2(v-4)}$
Gamma $X \sim \Gamma(\alpha,\beta)$	$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x)$ $x \in [0,\infty)$	$\mathbb{E}[X] = \frac{\alpha}{\beta}$ $\text{Var}(X) = \frac{\alpha}{\beta^2}$
Beta $X \sim \text{Beta}(\alpha,\beta)$	$f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ x in [0,1]	$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$ $\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

6 Useful Functions

The Gamma function, $\Gamma(\alpha)$, is defined by the integral:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\exp(-x)dx$$

Function has recurrence relationship where $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ and so is considered the continuous analogue of the factorials since $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$.

The Beta function $B(\alpha,\beta)$ is the solution to the integral:

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$$

The Beta function is symmetric in that $B(\alpha,\beta) = B(\beta,\alpha)$ and can be expressed in terms of gamma functions:

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

The standard normal cumulative distribution function $\Phi(z)$ at some $z \in \mathbb{R}$ is defined by the integral:

$$\Phi(z) = \int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)dt$$

Function is rotationally symmetric: $\Phi(z) = 1 - \Phi(-z)$. $\Phi(z)$ has no closed-form expressions except for $\Phi(0) = \frac{1}{2}$ so values need to be numerically calculated. Can use tables or R, e.g. $\Phi(0.123)$ can be found with `pnorm(0.123)` and inverses like $\Phi^{-1}(0.95)$ by `qnorm(0.95)`.

7 Miscellaneous

Relationships between Standard Randome Variables

- If $X_1, X_2, ..., X_n$ are independent and identical Bernoulli RVs with success probability p, then $\sum_{i=1}^n X_i \sim \text{Binom}(n,p)$.
- Consider a sequence (potentially infinite) of iid Bernoulli RVs with success probability p. Let X denote number of bernoulli RVs in the sequence until first success observed: $X \sim \text{Geom}(p)$.
- Suppose $X \sim \text{Binom}(n,p)$. For reasonably large sample size n the shape of the pmf is fairly symmetric and close to the bell curve of the normal distribution. Hence a reasonable approximation $X \sim N(np, np(1-p))$ can be made.
- If $X \sim \text{Binom}(n,p)$ then the limit as $n \to \infty, p \to 0$ where $np \to \lambda$ the rv X converges in distribution to the poisson distribution with rate parameter λ.

Transformation of Normal Random Variables

8 workshop 2

Prefix	Description
d-	Probability mass / density function
p-	Cummulative distribution function
q-	Quantile value from inverse cdf
r-	Generates random val from distribution

Postfix	Distribution (disc)	Postfix	Distribution (cont)
-binom	$X \sim \text{Binom}(n,p)$	-unif	$X \sim \text{Unif}(a,b)$
-pois	$X \sim \text{Pois}(\lambda)$	-exp	$X \sim \text{Exp}(\lambda)$
-geom	$X \sim \text{Geom}(p)$	-norm	$X \sim \text{N}(\mu,\sigma^2)$
		-t	$X \sim t_n$

Examples:

1. Probability $\mathbb{P}(X = 4)$ where $X \sim$Geom(0.5)
   `dgeom(x = 4, prob = 0.5)`
2. Probability $\mathbb{P}(X \leq 3)$ where $X \sim$ Exp(2)
   `pexp(q = 3, rate = 2)`
3. The 75% quantile i.e. upper quartile of $X \sim$ Unif$(-2,7)$
   `qunif(p = 0.75, min = -2, max = 7)`
4. The probability $\mathbb{P}(X = 6)$ where $X \sim$ Pois(4)
   `dpois(x = 6, lambda = 4)`
5. The probability $\mathbb{P}(X \geq 6)$ where $X \sim N(5,9)$
   `1 - pnorm(q = 6, mean = 5, sd = 3)`
6. The 97.5% quantile value of $X \sim t_7$
   `qt(p = 0.975, df = 7)`

Checking Estimators:

```
n=100
num_repeats <- 1000
samp_size <- 100
statistics <- rep(0, num_repeats)
for(i in seq_along(statistics)){
  x <- rnorm(n = samp_size, mean = 40, sd = 2)
  statistics[i] <- mean(x)
  statistics[i] <- var(x) # evaluate S^2
  statistics[i] <- sd(x) # evaluate S
}
c(mean(statistics), var(statistics))
```

9 workshop 3

1. Let $X_1, ..., X_n$ be iid $N(\mu,\sigma^2)$ random variables where μ is unknown. Observe the following data values:

$$13.9, 8.1, 11.6, 11.9, 13.1, 12.2$$

a) Provide a 99% interval for μ, assuming we know that $\sigma^2 = 4$.

The distribution of the sample mean estimator is $\overline{X} \sim N(\mu, \frac{4}{n})$ or equivalently $\frac{\overline{X}-\mu}{\sqrt{4/n}} \sim N(0,1)$. Then we have

$$0.99 = \mathbb{P}\left(-z_{0.005} \leq \frac{\overline{X}-\mu}{\sqrt{4/n}} \leq z_{0.005}\right)$$
$$= \mathbb{P}\left(\overline{X} - z_{0.005}\sqrt{\frac{4}{n}} \leq \mu \leq \overline{X} + z_{0.005}\sqrt{\frac{4}{n}}\right)$$

From R, the quantile $z_{0.005}$ is `qnorm(0.995)` is 2.576, so the CI is

$$11.8 \pm 2.576\sqrt{\frac{4}{6}} = (9.7, 13.1)$$

b) Provide a 99% confidence interval for μ, where σ^2 is unknown.

As σ^2 is unknown, $\frac{\overline{X}-\mu}{\sqrt{S^2/n}} \sim t_{n-1}$ where $\overline{X}$ is the sample mean estimator and S^2 is the unbiased sample variance estimator.

$$0.99 = \mathbb{P}\left(-t_{n-1;0.005} \leq \frac{\overline{X}-\mu}{\sqrt{S^2/n}} \leq t_{n-1;0.005}\right)$$
$$= \mathbb{P}\left(\overline{X} - t_{n-1;0.005}\sqrt{\frac{S^2}{n}} \leq \mu \leq \overline{X} + t_{n-1;0.005}\sqrt{\frac{S^2}{n}}\right)$$

From R, the quantile $t_{6-1;0.005}$ is `qt(0.995, df = 6-1)` is 4.032. So

$$11.8 \pm 4.032\sqrt{\frac{4}{6}} = (8.5, 15.1)$$

2. Hypothesis test for $\mu = 10$:

Hypothesis Statement: $H_0 : \mu = 10 \quad H_1 : \mu \neq 10$
Theory: If $\sigma^2 = 4$ then under the null hypothesis $\overline{X} \sim N(10, \frac{4}{n})$. From standardizing, we define the test statistic $Z = \frac{\overline{X}-10}{2/\sqrt{n}} \sim N(0,1)$. At the 5% s.l. the critical region for rejecting the null hypothesis is $C = (-\infty, -1.96] \cup [1.96, \infty)$.
Apply: From the sample of $n = 6$, the sample mean is $\overline{x} = 11.8$. Thus the observed test is $z = 2.205$.
Conclusion: As $z \in C$ then there is sufficient evidence to reject the null hypothesis of $\mu = 10$ at the 5% significance level for some other value.
If σ^2 is unknown, then we perform a one-sample t-test with test statistic

$$T = \frac{\overline{X}-10}{\sqrt{S^2/n}} \sim t_{n-1}$$

Under H_0, where S^2 is the sample variance estimator. Consequently, the critical region for the test at the 5% significance level becomes $C = (-\infty, -2.576] \cup [2.576, \infty)$, here the critical value was calculated by R in (a). As $s^2 = 4$, the observed test statistic is $t = 2.205$, leading to a different conclusion by failing to reject H_0 since $z \notin C$.

3. Let $X_1, ..., X_n$ be iid by $N(\mu,\sigma^2)$ random variables, where μ is known and σ^2 is unknown.

a) State the distribution of $\sum_{i=1}^n X_i$.
$N(n\mu, n\sigma^2)$. A linear combination of normally distributed RVs is also normally distributed, $\mathbb{E}\left(\sum_{i=1}^n X_i\right) = n\mu$, and $\text{Var}\left(\sum_{i=1}^n X_i\right) = n\sigma^2$ since the X_is are independent.
b) Evaluate $\mathbb{P}\left(\sum_{i=1}^n X_i > n\mu\right)$
0.5 since distributed symmetrically about mean
c) State the distribution of $\frac{X_i-\mu}{\sigma}$ and $\sum_{i=1}^n \left(\frac{X_i-\mu}{\sigma}\right)^2$.
$\frac{X_i-\mu}{\sigma} \sim N(0,1)$ independently for each i, so that $\sum_{i=1}^n \left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_n$

10 workshop 4

Blood glucose levels data:

```
control_before <- c( 5.1, 4.9, 4.5, 5.0, 5.3, 5.0, 5.0, 5.0, 5.0,
4.1, 5.2, 5.2, 5.2, 5.3, 5.4, 5.0, 5.1)
control_after <- c(10.7, 11.5, 9.6, 8.7, 9.1, 10.6, 10.3, 10.0,
7.5,9.4, 10.5, 10.2, 11.2, 9.4, 9.4, 11.6, 11.2)
test_before <- c( 4.8, 5.1, 4.8, 5.8, 4.9, 4.3, 5.0, 4.9, 4.9,
5.0, 4.8, 4.8, 5.0, 4.9, 4.8, 5.0, 4.9, 4.7)
test_after <- c( 9.1, 10.1, 9.6, 8.6, 10.3, 9.4, 8.4, 9.6, 8.2,
10.7, 9.9, 10.2, 8.9, 10.0, 10.0, 9.2, 9.7, 9.6)
```

Meal	Time	Mean, $\bar{x}$	Standard Dev., s
Control (n=17)	Before	5.02	0.31
	After	10.05	1.09
	Change	5.04	1.10
Test (n=18)	Before	4.91	0.28
	After	9.53	0.68
	Change	4.62	0.81

One-sample t-test: Verify participants baselines are consistent to what we expect. Let $X_1, ..., X_n$ denote blood glucose levels before eating and suppose they are independent samples from $X \sim N(\mu, \sigma^2)$ (both unknown). Healthy is 5mmol/L.

$$H_0 : \mu = 5 \qquad H_1 : \mu \neq 5$$

```
t.test(x = control_before, mu = 5, alternative = "two.sided")
## One Sample t-test
##
## data: control_before
## t = 0.23417, df = 16, p-value = 0.8178
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
## 4.857892 5.177402
## sample estimates:
## mean of x
## 5.017647
```

Observed test stat is $t = 0.234$, calculated by `(mean(control_before) - 5) / sqrt(var(control_before)/length(control_before))`. The p-val is 0.8178, hihgher than 5% sig lvl, so we fail to reject the null hypothesis.

Paired t-test: Let $X_1, ..., X_n$ denote the blood glucose levels after eating, and $Y_1, ..., Y_n$ the levels before. Define the differences $D_i = X_i - Y_i$, and suppose they are independent and normally distributedwith expectation $\mu_D = \mu_X - \mu_Y$. If $\mu_Y = 5$mmol/L and $\mu_X = 10$mmol/L from existing medical research, then $\mu_D = 5$mmol/L. The hypothesis statements are:

$$H_0 : \mu_D = 5 \qquad H_1 : \mu_D \neq 5$$

In this case we perform a paired t-test because two blood glucose measurements are taken from each participant, meaning that data is paired where the sampling unit are the participants. The test is implemented in R by:

```
t.test(x = control_after, y = control_before, mu = 5, paired =
TRUE, alternative = "two.sided")
## Paired t-test
##
## data: control_after and control_before
## t = 0.13182, df = 16, p-value = 0.8968
## alternative hypothesis: true mean difference is not equal to 5
## 95 percent confidence interval:
## 4.467702 5.602886
## sample estimates:
## mean difference
## 5.035294
```

The observed test stat is $t = 0.132$, calculated by

```
diff_control <- control_after - control_before
(mean(diff_control) - 5) / sqrt(var(diff_control)/
length(diff_control))
```

The p-value for the test is $p = 0.8968$ so we may fail to reject the null hypothesis at the 5% level.

Two-sample t-test: Let $X_1, ..., X_n$ denote blood glucose levels after eating the control meal where each represents iid samples from $X \sim N(\mu_X, \sigma^2)$ and similarly define $Y_1, ..., Y_m$ to be the measurements after eating the test meal representing iid samples from $Y \sim N(\mu_Y, \sigma^2)$. Consider the null hypothesis that the expected blood glucose lvl of the test group is no smaller than that for the control group:

$$H_0 : \mu_X - \mu_Y = 0 \qquad H_1 : \mu_X - \mu_Y > 0$$

For this we propose a two sample t-test (not paired as each sample relates to different individuals) under the assumption of equal population variances (requiring var.equal = TRUE in R). Alternative hyp is one-sided in positive direction, requiring input alternative = 'greater':

```
t.test(x = control_after, y = test_after, mu = 0, var.equal = TRUE,
alternative = "greater")
## Two Sample t-test
##
## data: control_after and test_after
## t = 1.7219, df = 33, p-value = 0.04723
## alternative hypothesis: true difference in means is greater than
0
## 95 percent confidence interval:
## 0.008999057 Inf
## sample estimates:
## mean of x mean of y
## 10.052941 9.527778
```

The observed test stat is $t = 1.722$, computed as follows (first requires evaulating pooled variance)

```
null_diff <- 0
xbar <- mean(control_after)
s2x <- var(control_after)
nx <- length(control_after)
ybar <- mean(test_after)
s2y <- var(test_after)
3
ny <- length(test_after)
s2p <- ( (nx-1)*s2x + (ny-1)*s2y )/(nx+ny-2)
s2p
## [1] 0.8132868
(xbar-ybar - null_diff) / sqrt(s2p*(1/nx + 1/ny))
## [1] 1.721866
```

p-val is 0.0472 which is smaller than 5% sig lvl so we reject the null hypothesis of equal expectations in favour of some other values where expectation for test group is smaller than control's.

Validating Assumptions: Let σ^2_X, σ^2_Y denote the population variances for the control and test groups respectively and consider:

$$H_0 : \sigma^2_X = \sigma^2_Y \qquad H_1 : \sigma^2_X \neq \sigma^2_Y$$

Run the equal variance F-test:

```
var.test(x = control_after, y = test_after, ratio = 1, alternative
= "two.sided")
##
## F test to compare two variances
##
## data: control_after and test_after
## F = 2.5127, num df = 16, denom df = 17, p-value = 0.06804
## alternative hypothesis: true ratio of variances is not equal to
1
## 95 percent confidence interval:
## 0.9317305 6.8796564
## sample estimates:
## ratio of variances
## 2.512659
```

Observed test stat is the ratio of sample variances:

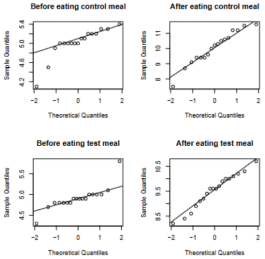
```
var(control_after) / var(test_after)
## [1] 2.512659
```

p-val is $p = 0.0680$ which is greater than the 5% sig lvl, so we fail to reject the equal variance assumption. So the stated assumption for the above two sample t-test appears reasonable, within limitations of the F-test.

Next, investigate assumption that population follows a normal distribution by generating quantile-quantile plots:

```
par(mfrow = c(2, 2)) # Creates a 2-by-2 panel to view all images
qqnorm(control_before, main = "Before eating control meal")
```

```
qqline(control_before)
qqnorm(control_after, main = "After eating control meal")
qqline(control_after)
qqnorm(test_before, main = "Before eating test meal")
qqline(test_before)
qqnorm(test_after, main = "After eating test meal")
qqline(test_after)
```



After eating control or test meal, plots roughly follow the reference line that assumes underlying population follows a normal distribution so little evidence to question validity of two-sample t-test. In the before plots, there is some deviation, so there might be some concern over underlying assumptions of the one sample t-test, so raises questions about the reliability of their conclusions.

11 workshop 10

The mean life of a sample of 9 light bulbs from batch A was observed to be 1309 hours with sample standard deviation of 420 hours. A second sample of 16 bulbs from a different batch B showed a mean life of 1205 hours with a sample standard deviation of 390 hours. Assume that the lifetime distributions associated with the two batches are $N(\mu_A, \sigma_A^2)$ and $N(\mu_B, \sigma_B^2)$.

(a) State the null and alternative hypothesis for testing if the two batches differ with respect to their mean lifetime.

$$H_0 : \mu_A - \mu_B = 0 \qquad H_1 : \mu_A - \mu_B \neq 0$$

(b) State the name of the test to be conducted.

Two sample t-test

(c) Assuming equality of variances, conduct the test at the 5% significance level.

The test statistic is

$$T = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{S_p^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

where $\overline{X}_A, n_A$ and $\overline{X}_B, n_B$ denote the sample mean estimator and sample size for batches A and B, and S_p^2 is the pooled sample variance estimator:

$$S_p^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}$$

From given data:

$$s_p^2 = \frac{8 \times 420^2 + 15 \times 390^2}{23} = 160552.17$$

$$t = \frac{1309 - 1205}{\sqrt{160552.17\left(\frac{1}{9} + \frac{1}{16}\right)}} = 0.6229$$

Under the null hypothesis, $T \sim t_{n_A+n_B-2}$, so the critical region for the two-sided alternative at the 5% lvl is $C = \{t : |t| \geq t_{23,0.025}\}$ where $t_{23,0.025} = 2.069$. Hence $t \notin C$ so we do not reject the null hypothesis. There is no evidence that the batches differ with respect to avg lifetimes.

(d) What is the purpose of performing an additional F -test in relation to the previous hypothesis test?

Assume that the variances of the two populations where our observations came from are equal. Can formally perform a significance test to examine if there is evidence against this assumption.

(e) Perform the F -test at the 5% significance level. What does your results suggest about the earlier hypothesis test on the mean light bulb lifetimes?

The following R commands may be of use:

```
> qnorm(p = 0.975, mean = 0, sd = 1)
[1] 1.959964
> qt(p = 0.975, df = c(23, 24, 25))
[1] 2.068658 2.063899 2.059539
> qf(p = c(0.025, 0.975), df1 = 8, df2 = 15)
[1] 0.2438303 3.1987381
> qf(p = c(0.025, 0.975), df1 = 9, df2 = 16)
[1] 0.2670871 3.0487535
```

$H_0 : \sigma_A^2 = \sigma_B^2 \qquad H_1 : \sigma_A^2 \neq \sigma_B^2$. For the test statistic $F = \frac{S_A^2}{S_B^2}$ where under H_0, $F \sim F_{n_A-1,n_B-1}$. At the 5% significance level, the critical region for the test is $C = \left(0, F_{8,15;0.975}\right] \cup \left[F_{8,15;0.025}, \infty\right)$, where the critical values are $F_{8,15;0.975} = 0.24$ and $F_{8,15;0.025} = 3.20$. The observed test statistic is $f = \frac{4202}{3902} = 1.16$. The test statistic is outside the critical region, so there is no evidence to reject the hypothesis of equal variances at the 5% significance level. Therefore, the equal variance assumption made made for the two-sample t-test appears to be reasonable.

12 workshop 6

1. State all assumptions that define the normal linear regression model and discuss the importance of each assumption.

- **Linearity** - this is the assume relationship to be investigated.
- **Independence** - Required when deriving estimator variance. without, covariance between data values needs to be taken into consideration.
- **Constant Variance** - Var(Y) = σ^2. Required when deriving estimator variance. Otherwise would need to account for the fact that the variance for eahc datum may be different.
- **Ranomd errors are normally distributed** - $\varepsilon \sim N(0, \sigma^2)$. Required for deriving distribution for least-squared estimators for computation of confidence intervals and performing hypothesis tests. Otherwise, exact def of estimator distributions are not easily defined.

2. When taking the maximum likelihood method to derive estimates the normal linear regression parameters, explain why the maximum likelihood estimators $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators for the regression parameters, but the estimator $\widehat{\sigma^2}$ for the common variance is biased at finite sample sizes.

Log-likelihood func for normal linear regression model:

$$l(\alpha, \beta, \sigma^2 \mid x, y) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}Q(\alpha, \beta)$$

where the sum of squares formula is $Q(\alpha, \beta) = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$. Partials w.r.t α, β:

$$\frac{\partial l}{\partial \alpha} = -\frac{1}{2\sigma^2}\frac{\partial}{\partial \alpha}Q(\alpha, \beta) \qquad \frac{\partial l}{\partial \beta} = -\frac{1}{2\sigma^2}\frac{\partial}{\partial \beta}Q(\alpha, \beta)$$

Finding coord that maximises log-likelihood is equiv to finding coord that minimises sum of squares formula. So max likelihood estimates $\hat{\alpha}, \hat{\beta}$ are least-squares estimators, which are unbiased in estimating regression params. Partial w.r.t σ^2:

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}Q(\alpha, \beta) \Rightarrow \widehat{\sigma^2} = \frac{1}{n}Q\left(\hat{\alpha}, \hat{\beta}\right)$$

Note ML estimator :

$$\widehat{\sigma^2} = \frac{n-2}{n}\frac{1}{n-2}\sum_{i=1}^n \left(Y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2 = \frac{n-2}{n}S_e^2$$

As S_e^2 is an unbiased estimator for σ^2, $\mathbb{E}\left[\widehat{\sigma^2}\right] = \frac{n-2}{n}\sigma^2$, so at finite sample size the max likelihood estimator is biased in estimating σ^2.

3. Consider the following summarised data for a simple linear regression model:

$$n = 38, \overline{x} = 20, \overline{y} = 19, s_x = 4.4, s_y = 5.5, r_{x,y} = -0.28$$

(a) Estimate the least-squares estimates $\hat{\alpha}$ and $\hat{\beta}$.

$$\hat{\beta} = r_{x,y}\frac{s_y}{s_x} = -0.28\frac{5.5}{4.4} = -0.35$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} = 19 - 20(-0.35) = 26$$

(b) Given that residual standard error is $s_e = 5.353$, evaluate standard error estimate for both least-squares estimators.

$$\widehat{SE}(\hat{\alpha}) = s_e\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}} = 5.353\sqrt{\frac{1}{38} + \frac{20^2}{37 \times 4.4^2}} = 4.093$$

$$\widehat{SE}\left(\hat{\beta}\right) = s_e\sqrt{\frac{1}{(n-1)s_x^2}} = 5.353\sqrt{\frac{1}{37 \times 4.4^2}} = 0.200$$

(c) Compute the 95% confidence interval for α.

$$\hat{\alpha} \pm t_{36;0.025}\widehat{SE}(\hat{\alpha}) = 26 \pm 2.028 \times 4.093 = (17.70, 34.30)$$

(d) Perform a hypothesis test at 5% significance level to investigate:

$$H_0 : \beta = 0 \qquad H_1 : \beta \neq 0$$

```
> qt(0.975, df = c(36, 37, 38) )
[1] 2.028094 2.026192 2.024394
```

$Y = \alpha + \beta x + \varepsilon, \varepsilon \sim N(0, \sigma^2)$. Consider tests stat under null:

$$T = \frac{\hat{\beta} - 0}{SE\left(\hat{\beta}\right)} \sim t_{36}$$

At 5% level, critical region for the test is $C = \{t : |t| \geq t_{36;0.025}\}$, where $t_{36;0.025} = 2.028$ (from R output). Given the estimates in earlier parts of the Q.

$$t = -\frac{0.35}{0.200} = -1.75$$

So $t \notin C$, so we fail to reject the null hypothesis indicating that the explanatory variabele and response variables are not linearly related.

4. For $\mathbb{E}[Y] = \alpha + \beta x$, show that

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{n-1}{n-2}\left(s_y^2 - \hat{\beta}^2 s_x^2\right)$$

where s_x^2, s_y^2 are the sample variances in the explanatory and response variables respectively, and $\hat{\beta}$ the least squares estimate of slope parameter.

Note $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}, \hat{\beta} = \frac{s_{xy}}{s_x^2}$. So $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i = \overline{y} - \hat{\beta}(x_i - \overline{x})$.

$$\begin{aligned}
s_e^2 &= \frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{y}_i)^2 \\
&= \frac{1}{n-2}\sum_{i=1}^n \left((y_i - y) - \hat{\beta}(x_i - \overline{x})\right)^2 \\
&= \frac{1}{n-2}\sum_{i=1}^n \left\{(y_i - \overline{y})^2 - 2\hat{\beta}(x_i - \overline{x})(y_i - \overline{y}) + \hat{\beta}^2(x_i - \overline{x})^2\right\} \\
&= \frac{1}{n-2}\left((n-1)s_y^2 - 2\hat{\beta}(n-1)s_{x,y} + \hat{\beta}^2(n-1)s_x^2\right) \\
&= \frac{n-1}{n-2}\left(s_y^2 - 2\hat{\beta}\frac{s_{x,y}}{s_x^2}s_x^2 + \hat{\beta}^2 s_x^2\right) \\
&= \frac{n-1}{n-2}\left(s_y^2 - 2\hat{\beta}^2 s_x^2 + \hat{\beta}^2 s_x^2\right) \\
&= \frac{n-1}{n-2}\left(s_y^2 - \hat{\beta}^2 s_x^2\right)
\end{aligned}$$

13 workshop 7

```
diamonds <- data.frame(
  value = c(176.1, 174.8, 189.5, 183.0, 188.0, 180.1, 182.1, 182.3,
192.4, 186.7, 187.8, 193.7, 190.0, 188.7, 200.8, 172.4, 186.3,
190.2, 194.8, 198.2),
  weight = c( 8.28, 10.11, 12.50, 7.88, 9.66, 10.03, 11.00, 9.39,
13.17, 9.56, 10.51, 11.47, 9.13, 8.03, 12.83, 5.87, 11.29, 9.86,
11.52, 10.53),
  clarity = c(0.97, 0.24, 0.86, 0.94, 0.51, 0.23, 0.87, 0.64, 0.94,
0.83, 1.19, 1.10, 1.27, 0.97, 0.86, 1.16, 0.91, 1.09, 1.17, 1.24),
  carat = c(1.33, 1.36, 1.30, 1.31, 1.40, 1.23, 1.27, 1.28, 1.10,
1.30, 1.10, 0.99, 1.03, 0.99, 0.96, 0.78, 0.84, 0.90, 0.74, 0.78)
)
```

Predict diamond's value base on its weight:

```
model1 <- lm(formula = value ~ weight, data = diamonds)
summary(model1)
```

Adding new data points:

```
newdiamonds <- data.frame(weight = c(6.04, 10.13))
```

i.e. two new diamonds with weights 6.04 and 10.13. To predict values from the fitted linear regression, and to construct an interval:

```
predict(model1, newdata = newdiamonds)
predict(model1, newdata = newdiamonds, interval = "confidence",
level = 0.90)
```

Multiple Linear Regression Can use more explanatory variables for better fit. E.g. $\mathbb{E}[Y_i] = \beta_1 + \beta_2 x_{weight} + \beta_3 x_{clarity}$:

```
model2 <- lm(formula = value ~ weight + clarity, data = diamonds)
summary(model2)
## Call:
## lm(formula = value ~ weight + clarity, data = diamonds)
##
## Residuals:
## Min 1Q Median 3Q Max
## -6.9133 -3.7347 0.7496 2.0655 6.9734
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 147.7248 7.2930 20.256 2.43e-13 ***
## weight 2.8328 0.6122 4.627 0.000241 ***
## clarity 11.6410 3.7048 3.142 0.005942 **
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.793 on 17 degrees of freedom
## Multiple R-squared: 0.6356, Adjusted R-squared: 0.5928
## F-statistic: 14.83 on 2 and 17 DF, p-value: 0.0001876
```

The p-val corresponding to the clarity coefficient is 0.0059 (smaller than 5% sig lvl). So we may reject the null hypothesis that the clarity coeff is zero.

Multiple & Adjusted R^2

```
model3 <- lm(formula = value ~ weight + clarity + carat, data =
diamonds)
summary(model3)
```

Determining the most suitable model:

Formula	R^2	Adj R^2
value ~ weight	42.4%	39.2%
value ~ clarity	17.7%	13.1%
value ~ carat	19.1%	14.6%
value ~ weight + carat	58.5%	53.6%
value ~ weight + clarity	63.6%	59.3%
value ~ clarity + carat	22.3%	13.1%
value ~ weight + clarity + carat	65.3%	58.7%

Highest R^2 contains all 3 variables, highest adjusted has weight & clarity as explanatory variables, so model might be overfitting the data. From summary output of model3 (ommitted) the p-val for clarity is greater than 5% sig lvl, supporting over-fitting observation.

ANOVA F-test

p-vals for both carat and clarity are high / above the 5% lvl, so the possibility that $\beta_3 = 0$ or $\beta_4 = 0$ cannot be rejected. $\mathbb{E}[Y] = \beta_1 + \beta_2 x_{weight} + \beta_3 x_{clarity} + \beta_4 x_{carat}$. The tests on the coefficients are independent, so:

$$H_0 : \beta_3 = \beta_4 = 0 \qquad H_1 : \beta_3 \neq 0 \text{ and/or } \beta_4 \neq 0$$

Model under null: $\mathbb{E}[Y] = \beta_1 + \beta_2 x_{weight}$. Test stat:

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(n - p_1)} \sim F_{p_1-p_0,n-p_1}$$

```
n <- nrow(diamonds) # sample size = 20
rss_null <- sum(residuals(model1)^2) # residual sum of squares
under the null, RSS_0
rss_full <- sum(residuals(model3)^2) # residual sum of squares of
the full model, RSS_1
p_null <- length(coef(model1)) # number of parameters under the
null model, p_0 = 2
p_full <- length(coef(model3)) # number of parameters in the full
model, p_1 = 4
f <- ((rss_null - rss_full)/(p_full-p_null)) / (rss_full/(n-
p_full))
critval <- qf(0.95, df1 = p_full-p_null, df2 = n - p_full)
```

Observed test stat os $f = 5.26$ and critical region at 5% level is $C = [3.63, \infty)]$. As $f \in C$, there is sufficient evidence to reject null hypothesis in case of some alternative where β_3 or β_4 or both are non-zero. So there exists some linear combo of them that explains some of the variability beyond what weight can explain.

```
anova(model1, model3)
## Analysis of Variance Table
##
## Model 1: value ~ weight
## Model 2: value ~ weight + clarity + carat
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1 18 617.27
## 2 16 372.33 2 244.94 5.2628 0.01752 *
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Define the full linear model as $\mathbb{E}[Y] = \beta_1 + \beta_3 x_{clarity} + \beta_4 x_{carat}$ and consider ANOVA statements:

$$H_0 : \beta_3 = \beta_4 = 0 \qquad H_1 : \beta_3 \neq 0 \text{ and/or } \beta_4 \neq 0$$

To fit the null i.e. $\mathbb{E}[Y] = \beta_1$ the full models are:

```
M_null <- lm(formula = value ~ 1, data = diamonds)
M_full <- lm(formula = value ~ clarity + carat, data = diamonds)
```

Results:

```
anova(M_null, M_full)
## Analysis of Variance Table
##
## Model 1: value ~ 1
## Model 2: value ~ clarity + carat
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1 19 1071.67
## 2 17 832.97 2 238.7 2.4358 0.1174
```

p-val is greater than 5% lvl, so fail to reject null. In other words, there is not enough evidence to distinguish between the full model and the intercept only model. The summary on the full model is:

```
summary(M_full)
##
## Call:
## lm(formula = value ~ clarity + carat, data = diamonds)
##
## Residuals:
## Min 1Q Median 3Q Max
## -19.1609 -2.3000 0.2256 4.6078 12.7784
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 192.280 15.741 12.216 7.66e-10 ***
## clarity 5.942 7.135 0.833 0.417
## carat -9.759 9.728 -1.003 0.330
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7 on 17 degrees of freedom
## Multiple R-squared: 0.2227, Adjusted R-squared: 0.1313
## F-statistic: 2.436 on 2 and 17 DF, p-value: 0.1174
```

14 workshop 8

1. How to compare below models?

- $\mathbb{E}[Y] = \beta_0 + \beta_1 a + \beta_2 b + \beta_3 c$
- $\mathbb{E}[Y] = \beta_0 + \beta_2 b$
- $\mathbb{E}[Y] = \beta_0 + \beta_4 d$

Compare 1 and 2 by ANOVA F-test as 2 is a simplified 1. Could also use adj R^2. For 1 and 3, not possible to set some parameters in one model to zero to obtain other so can't do ANOVA so best would be adj R^2. Can't do ANOVA for 2 and 3 either, but same number of params so either multiple or adj R^2.

2. Dataset contains 15 rows and 3 columns with names , t, u:

```
> model <- lm(w ~ t + u, data = dataset)
> summary(model)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.9939 4.3774 -4.796 0.000436 ***
t 9.3295 1.0127 9.213 8.62e-07 ***
u 7.8727 0.8764 8.983 1.13e-06 ***
> qt(0.95, df = c(11, 12, 13, 14, 15))
[1] 1.795885 1.782288 1.770933 1.761310 1.753050
> qt(0.975, df = c(11, 12, 13, 14, 15))
[1] 2.200985 2.178813 2.160369 2.144787 2.131450
```

a) Describe the model being fitted to the data. A multiple linear regression model with two explanatory variables. Let W_i denote response variable with t_i and u_i the two explanatory variables.

$$W_i = \beta_0 + \beta_1 t_i + \beta_2 u_i + \varepsilon_i$$

where the independent random error ε_i follows a normal distribution $N(0, \sigma^2)$.

b) Expected value of w when $t = 3, u = -0.5$?

$$\hat{\mathbb{E}}[W] = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 u = 3.06$$

c) Compute 95% confidence interval for regression coeff of t.

$$\hat{\beta}_1 \pm t_{n-p;0.025}\widehat{SE}\left(\hat{\beta}_1\right) = (7.12, 11.54)$$

d) Perform hypothesis test at the 5% sig lvl to investigate the null hypothesis that the regression coeff of u is 9 or larger.

- $H_0 : \beta_2 \geq 9 \quad H_1 : \beta_2 < 9$.
- Test stat under null: $T = \frac{\hat{\beta}_2 - 9}{\widehat{SE}(\beta)_2} \sim t_{n-p}$ where $n = 15, p = 3$.
- Critical region at 5%: Alt hyp is one-sided in negative direction, so $C = \left(-\infty, -t_{n-p;0.025}\right] = (-\infty, 1.782]$.
- $t = \frac{7.8727 - 9}{0.8764} = -1.286$.
- Observed test stat does not lie in the critical region so fail to reject null, so cannot discard posibility that coeff of u is 9 or larger.

4. Consider $\mathbb{E}[Y] = \beta_0 + \sum_{i=1}^4 \beta_i x_i$. Fitting the data with $n = 35$ obtains residual standard error of $s_1 = 2.1$. Simpler sub-model $\mathbb{E}[Y] = \beta_0 + \beta_1 x_1$ is proposed, resulting in residual standard error of 2.3. Perform ANOVA F-test to compare the two at the 5% lvl.

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \quad H_1 : \min \text{ one of } \beta_2, \beta_3, \beta_4 \neq 0$$

ANOVA F-test statistic:

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(n - p_1)} \sim F_{p_1-p_0,n-p_1}$$

where residual sum of squares for full model is $RSS_1 = (n - p_1)S_1^2$ with sample size n, number of parameters p_1 and unbiased residual variance estimator S_1^2. Terms RSS_0, p_0, S_0^2 similarly defined for null hyp sub-model.

Critical region:

```
qf(p = 0.95, df1 = 5-2, df2 = 35-5)
[1] 2.922277
```

$f = \frac{(174.57 - 132.3)/(5-2)}{132.3/(35-5)} = 3.195$

Observed test stat is greater than the critica value, so sufficient evidence to reject the null hyp at the 5% lvl. This suggests that there is some linear combo of variables x_2, x_3, x_4 that can explain some of the variability in the response beyond what x_1 can describe.

5. Complete the following one-way ANOVA for a categorical variable with $k = 5$ groups:

	Deg.Free.	Sum Sq	Mean Sq	F value
Group	4	24	6	1.2
	k - 1	$rss_b = rss_y - rss_e$	$s_b^2 = \frac{rss_b}{k-1}$	$\frac{s_b^2}{s_e^2}$
Residuals	10	50	**5**	-
	$n - k$	$rss_e = (n-k)s_e^2$	s_e^2	-
Total	**14**	**74**	-	-
	n-1	rss_y	-	-