2 Linear Regression

Exercise 11 — Assume we have n observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ and we consider a linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. We estimate parameters β_0 and β_1 by minimizing mean squared error:

$$MSE(\hat{\beta}_0, \hat{\beta}_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 + \hat{\beta}_1 x))^2 \ .$$

Show that in such a case

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

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

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are sample means. Argue that the obtained line always passes through the point $(\bar{x}, \bar{y})$. (2p)

Exercise 12 — Derive the bias, variance and standard error for estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. We assume that $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and all ε_i for $i \in \{1, \ldots, n\}$ are independent. (2p)

Exercise 13 — Recall how to prove that the sum of two independent normally distributed random variables is normally distributed. (2p)

Exercise 14 — Explain why there is approximately a 95% chance that the interval

$$\hat{\beta}_1 \pm 2\sqrt{\mathrm{Var}\left(\hat{\beta}_1\right)}$$

contains the true value of β_1. (2p)

Exercise 15 — Recall that for $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $TSS = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ we define R^2 as

$$R^2 = 1 - \frac{RSS}{TSS} \ .$$

What's the interpretation for R^2? Show that if we consider a model $Y = \beta_0 + \beta_1 X + \varepsilon$ we have

$$R^2 = Corr(X, Y)^2 \ ,$$

where $Corr(X, Y)$ is correlation coefficient. (3p)

Exercise 16 — Recall what's t-statistic and how we can use it in the context of linear regression. What's p-value? (3p)

Exercise 17 — Show that for a linear regression model with $k+1$ parameters we can obtain estimations of

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$$

as

$$\hat{\beta} = (XX^T)^{-1}X^T\vec{y} \ ,$$

where X is data matrix and $\vec{y}$ is vector of responses (see e.g. here). (3p)

Exercise 11

We can recall that those presented parameters gives minimum error variance, in case where:

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

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

For those parameters by minimizing the MSE:

$$MSE(\hat{\beta}_0, \hat{\beta}_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 + \hat{\beta}_1 x\right)^2$$

We can prof that the obtained line allways go through the point $(\hat{x}, \hat{y})$, becouse we can note how the $\hat{y}_i$ parameters are estimated by $\hat{x}_i$:

$$\hat{y} = \frac{1}{n}\sum y_i = \frac{1}{n}\sum(\hat{y}_i + \hat{\epsilon}_i)$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{i1} + \cdots + \hat{\beta}_n \hat{x}_{in} + \hat{\epsilon}_i$$

We get then:

$$\hat{y} = \frac{1}{n}\sum\left(\hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{i1} + \cdots + \hat{\beta}_n \hat{x}_{in} + 0\right)$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{i1} + \cdots + \hat{\beta}_n \hat{x}_{in}$$

From that we see the prof, the regression line will go through the point $(\hat{x}, \hat{y})$.

Exercise 12

We assume that $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma^2)$ and assume that all ϵ_i for $i \in \{1, \ldots, n\}$ are independent.

So the $Bias$ for estimators β_0 and β_1 will look like this:

$$Bias(\beta_0) = E(\beta_0) - \beta = E(\bar{X}) - \beta = E\left(\frac{1}{n}\sum x_i\right) - \beta = \beta - \beta = 0$$

$$Bias(\beta_1) = E(\beta_1) - \beta_0 = (\beta^* - \beta) = 0 \; as \; \beta^* = \beta$$

When the value is 0, the estimator is unbiased.

Variance:

$$Var(\beta_0) = Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2} n Var(x_i) = \frac{1}{n} Var(x_i) = \sigma^2$$

$$Var(\beta_1) = \frac{\sigma^2}{n}$$

Standard error of estimation:

$$S_{est} = \sqrt{\frac{\sum(\beta_0 - \beta_1)^2}{n}}$$

Exercise 13 — Recall how to prove that the sum of two independent normally distributed random variables is normally distributed.

In this case we should be able to use characterisitc function, for example:

$$\varphi_{X+Y}(t) = E(e^{it(X+Y)})$$

In this case we know that the sum of those given variables X and Y is just a product of two separated characteristic functions:

$$\varphi_X(t) = E(e^{it(X)})$$

$$\varphi_Y(t) = E(e^{it(Y)})$$

We know that the expected value μ and the variance θ^2 in the characteristic function of the normal distribution is:

$$\varphi(t) = exp(\mathrm{it} * \mu - \frac{\theta^2 * t^2}{2})$$

We can then conclude that:

$$\varphi_{X+Y}(t) = \varphi_X(t) * \varphi_Y(t) = exp\left(it * \mu_X - \frac{{\theta_X}^2 * t^2}{2}\right) exp\left(it * \mu_Y - \frac{{\theta_Y}^2 * t^2}{2}\right)$$

$$\varphi_{X+Y}(t) = exp\left(it(\mu_X + \mu_Y) - \frac{({\theta_X}^2 + {\theta_Y}^2) * t^2}{2}\right)$$

We know that we have there characteristic function oft he normal distribution with expected value $\mu_X + \mu_Y$ and the variance ${\theta_X}^2 + {\theta_Y}^2$. We can then conclude that there are not wo distinct distributions that can both have the same characteristic function. Then the destribution of the $X + Y$ values must be just this normal distribution.

Exercise 14

We need to explain why there is approximatly 95% chance that $\hat{\beta}_1$ value in this interval is true value.

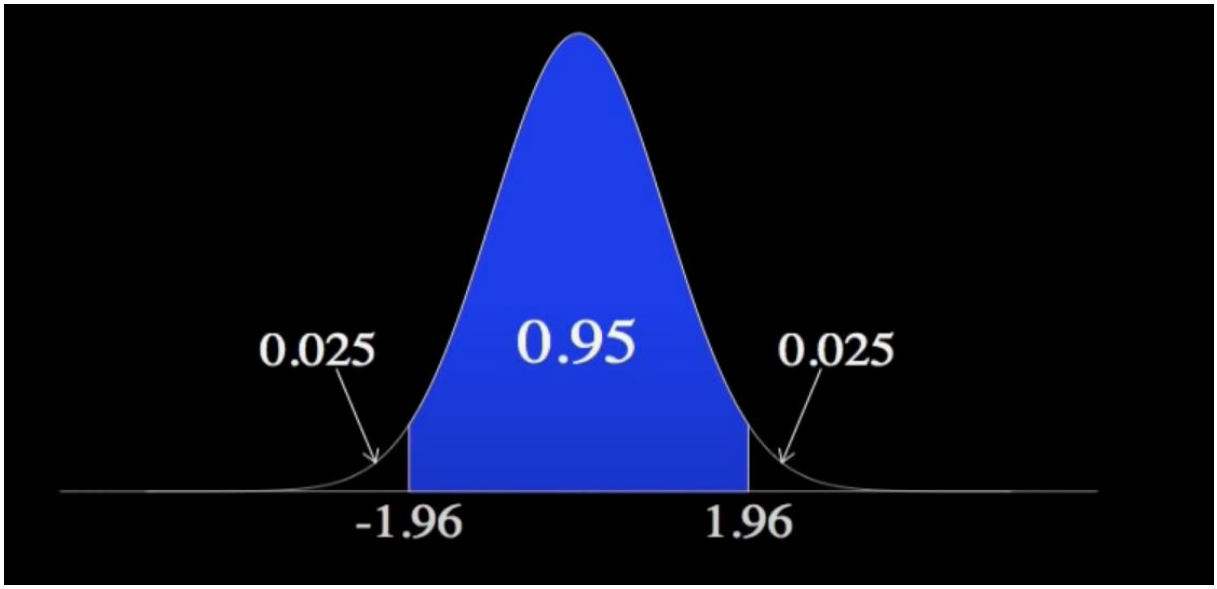
$$\hat{\beta}_1 \pm 2\sqrt{Var(\widehat{\beta_1})}$$

We need to recall into confidence intervals in this case, a 95% confidence interval means that if we were to take 100 different samples and compute a 95% confidence interval for each sample, then approximately 95 of the 100 confidence intervals will contain the true mean value.

Then:

For the standard normal distribution $P(-1.96 < \widehat{\beta_1} < 1.96) = 0.95$

$$\hat{\beta}_1 = \frac{\hat{X} - \mu\hat{x}}{\sigma\hat{x}} = \frac{\hat{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$



This comes with a conclusion $(1-\alpha) * 100\% = 95\%$

$$P\left(-1.96 < \frac{\hat{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96\right) = 0.95$$

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} < \mu < \hat{X} < \mu < 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} < \mu < \hat{X} < \mu < 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$95\%\ CI = \hat{\beta}_1 \pm 2\sqrt{Var(\widehat{\beta_1})}$$

$$95\%\ CI = \left(-2\sqrt{Var(\hat{\beta}_1)} < \hat{\beta}_1 < 2\sqrt{Var(\hat{\beta}_1)}\right)$$

Exercise 15

We need then to show that the R^2 statistic is equal to the square of correlation between X and Y. For simplicity in calculations we can assume that $\bar{x} = \bar{y} = 0$.

RSS – Residual Sum of Squares, TSS = Total Sum of Squares

So, we have the following equalities:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i (y - \hat{y}_i)^2}{\sum_j y_j^2}$$

With $\hat{y}_i = \hat{\beta}_1 x_i$, we can write it as:

$$R^2 = 1 - \frac{\sum_i \left(y_i - \frac{\sum_j x_j y_j}{\sum_j x_j^2 x_i} \right)^2}{\sum_j y_j^2} = \frac{\sum_j y_j^2 - \left(\sum_i y_i^2 - 2 \sum_i y_i \left(\frac{\sum_j x_j y_j}{\sum_j x_j^2} \right) x_i + \sum_i \left(\frac{\sum_j x_j y_j}{\sum_j x_j^2} \right)^2 x_i^2 \right)}{\sum_j y_j^2}$$

From this point we can show the $Cor(X, Y)^2$:

$$R^2 = \frac{\frac{2(\sum_i x_i y_i)^2}{\sum_j x_j^2} - \frac{(\sum_i x_i y_i)^2}{\sum_j x_j^2}}{\sum_j y_j^2} = \frac{(\sum_i x_i y_i)^2}{\sum_j x_j^2 \sum_j y_j^2} = Cor(X, Y)^2$$

Exercise 16

t-statistic is short from word test statistic. The test statistic shows how closely the result of the test data distribution is predicted with respect to the range of the null hypothesis used in this test statistic.

The distribution of these data determines exactly what the frequency of observations is. It can be determined by the central tendency and the variation around the total route of this central tendency. Due to the fact that each type of statistical test provides for different possibilities of distributions, it is necessary to fit the appropriate test for your case.

The t-statistic summarizes the totality of the observed data as a single number using the central tendency, their variability, sample size, and the total number of predictor variables in a given statistical model.

In short, t-statistic is computed as the correlation between the variables divided by the variance of the data (i.e., standard deviation).

In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the

null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

Let's go with explanation, for example we calculacted the t-value of 2.39 that is farm from expected range of t-values under the null hypothesis and the p-value is less than 0.05. In this situation we would expect to see a t-value as large or larger than 2.39 value in less than 5% of the time.

Real life implementation:

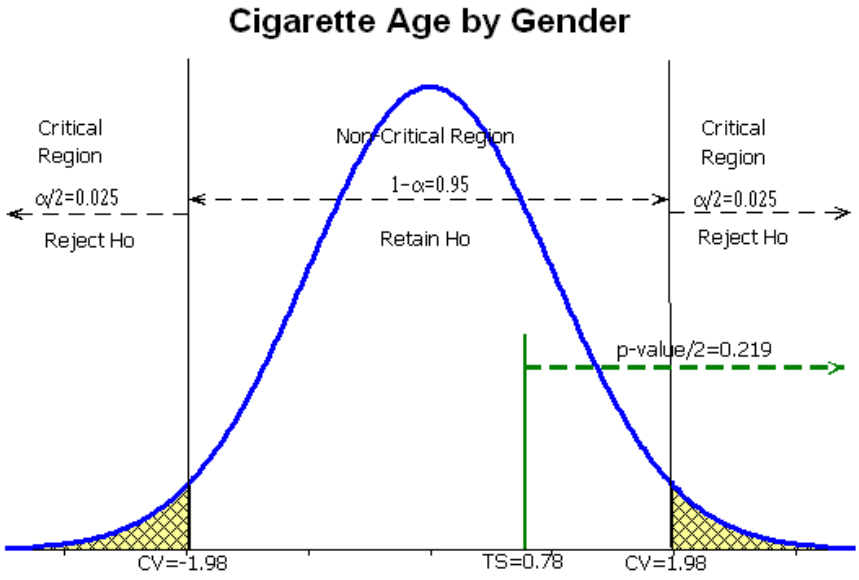
In this case we try to calculate if the men and women try cigarettes at the same age.

Two-sample T for cigarette

```
gender   N   Mean  StDev  SE Mean
Female  78  15.14   2.80     0.32
Male    72  14.76   3.11     0.37

Difference = mu (Female) - mu (Male)
Estimate for difference:  0.377
95% CI for difference:  (-0.581, 1.335)
T-Test of difference = 0 (vs not =): T-Value = 0.78  P-Value = 0.438  DF = 143
```

Looking ad the data we have two tail test and the value of DF is 143, in this case then the critical values for α = 0.05 significance level are t = ±1.976692.



Looking at the pictuer the T-statistic of t = 0.78 falls in the non-critical region, we can then retain the null hypothesis. Then the p-value of 0.438 is greater than the visible significance level of 0.05, we again retain the null hypothesis. The overall claimed difference of 0 falls in the confidence interval of $-0.581 < \mu_f - \mu_m < 1.335$, once more then we retain the null hypothesis.

Then in all three approaches to hypothesis testing we retain the null hypothesis. We can conclude that then there is not enough evidence to reject the claim that the age at which men and women

first try cigarettes is the same. There is not enough evidence to support the claim that the age at which men and women first try cigarettes is different.

Exercise 17

We know that we have here multiple linear regression model according to the information given, we then see this as:

$$h_{\hat{\beta}}(x) = \hat{\beta}_0 x_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

We want then to minimize the least square cost function:

$$J(\hat{\beta}_0 \dots k) = \frac{1}{2l} \sum \left[h_{\hat{\beta}}(x^{(i)}) - y^{(i)})^2 \right]$$

In which we use l as a number of training samples, x is input variable and y is output variable for every next i sample

Now, that the regression coefficents $\hat{\beta}$ are essentialy a vector, and each of the l input samples are also a vector with $k + 1$ dimensions (for convenience we can assume that $x_0 = 1$). We get then in matrix notation that:

$$H_{\hat{\beta}}(x) = \hat{\beta}^{\,T} x$$

We can then rewrite the least squares cost function and use matrix multiplication to get:

$$J(\hat{\beta}) = \frac{1}{2l} (X\hat{\beta} - y)^T (X\hat{\beta} - y)$$

In this case we can ignore $\frac{1}{2l}$ and apply some matrix transpose identities, since we will be evaluating derivative later in the calculation. The final form will be simplified equation for $J(\hat{\beta})$:

$$J(\hat{\beta}) = ((X\hat{\beta})^T - y^T)(X\hat{\beta} - y)$$

$$J(\hat{\beta}) = ((X\hat{\beta})^T X\hat{\beta} - (X\hat{\beta})^T y - y^T (X\hat{\beta}) + y^T y$$

$$J(\hat{\beta}) = \hat{\beta}^T X^T X\hat{\beta} - 2(X\hat{\beta})^T y + y^T y$$

Right now we need to find minimum of the above function, to do that we need to find the derivative wrt $\hat{\beta}$, and also equate to 0:

$$\frac{\partial J}{\partial \hat{\beta}} = 2X^T X\hat{\beta} - 2X^T y = 0$$

Now it looks like it should:

$$X^T X\hat{\beta} = X^T y$$

At the end we assume that given matrix $(X^T X)$ is invertible. We multiply both sides by $(X^T X)^{-1}$, and finally we are left with normal equation:

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