Prior Knowledge:

A **sample space** Ω is a set of possibilities. A **random variable** (RV) is a function $X : \Omega \to \mathbb{R}$. **Discrete** RV means Ω is countable. **Continuous** means uncountable possibilities.

For discrete RVs, $\sum_i P(X = x) = 1$, for continuous, $\int_{-\infty}^{\infty} f(x) = 1$, and $P(X = x) = 0$. We define the **expectation** as $E[X] = \sum_i xP(X = x)$ or $\int_{-\infty}^{\infty} xf(x)$, and the **variance** as $Var[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2$. The **standard deviation** is $SD(X) = \sqrt{Var[X]}$. We know $E[aX + b] = aE[X] + b$, and $Var[aX + b] = a^2 Var[X]$.

For a random variable X and event A, we define the **conditional probability** $P(X|A) = \frac{P(X \cap A)}{P(A)}$. Conditional expectations and variances are defined similarly.

RVs X and Y are **independent** if $P(X \cap Y) = P(X)P(Y)$. Random variables are **iid** means they are independent and identically distributed.

A sequence of random variables converges in probability to another RV if $\lim_{n\to\infty} P(|X_n - Y| > \epsilon) = 0$, denoted by $X_n \xrightarrow{P} Y$. For example, the (weak) **Law of Large Numbers** (LLN) states $\bar{X}_n \xrightarrow{P} E[X]$, where $\bar{X}_n$ is the mean of n samples. Almost sure convergence is a stronger result, which states that $P(\lim_{n\to\infty} X_n = Y) = 1$, or $X_n \xrightarrow{a.s} Y$. Convergence in distribution means that $\forall x \in \mathbb{R}, \lim_{n\to\infty} P(X_n < x) = P(Y < x)$. Written as $X \xrightarrow{D} Y$.

The **Central Limit Theorem** states that for any RV X, if we define $Z_n = \frac{\sum_{i=1}^{n} X_i - nE[X]}{\sqrt{nVar[X]}}$, then $Z_n \xrightarrow{D} N(0, 1)$.

A Chi-Square distribution with m degrees of freedom is the sum of m squared Standard Normal distributions. A $t_{\text{df}=m}$ is $\frac{N(0,1)}{\sqrt{\frac{\chi^2_m}{m}}}$. An $F_{\text{df}=m,n}$ distribution is $\frac{\frac{\chi^2_m}{m}}{\frac{\chi^2_n}{n}}$. Note: $F_{m,n} = \frac{1}{F_{n,m}}$, and $t^2_m = F_{1,m}$.

The **covariance** of two RVs is defined as $Cov(X, Y) = E[(X-E[X])(Y-E[Y])] = E[XY]-E[X]E[Y]$. Thus, $Var[X + Y] = Var[X] + 2Cov[X, Y] + Var[Y]$. We define their **correlation** as $Corr(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var[X]Var[Y]}}$. We can show that $-1 \leq Corr(X, Y) \leq 1$. Correlation is useful because of this 'standardization'.

A **parameter** is a true value of something, it is a underlying number which describes a population. An **estimator** is a random variable which estimates the value of a parameter. It is a function of collected data (recall $Y = h(X_1, \ldots, X_n)$ is an RV). An **estimate** is the value of an estimator for a given dataset.

The **bias** of an estimator is $Bias(\hat{\theta}) = E[\hat{\theta}] - \theta$. The **mean squared error** (MSE) is $E[(\hat{\theta} - \theta)^2] = Bias(\hat{\theta})^2 + Var[\hat{\theta}]$.

A **moment-generating function** for a distribution is $M_X(t) = E[e^{tX}]$. These are unique for different distributions, and can be used to prove distributions are equivalent.

Markov's Inequality states that if X is a non-negative RV, then for $a > 0$, we have $P(X \geq a) \leq \frac{E[X]}{a}$. **Chebyshev's Inequality** says that if X is a random variable with finite mean μ and variance σ^2, then $\forall k \in \mathbb{R}^+, P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$.

Simple Linear Regression:

Motivation:

Definitions: A **simple linear regression equation** could be $Y = a + bX$. This isn't realistic, and we usually find that this is not perfect, due to inherent randomness. Instead, we can look at $E[Y|X] = \beta_0 + \beta_1 X$, which can be written as $Y = \beta_0 + \beta + 1X + \epsilon$, for some **error term**. The values $\beta_0, \beta_1, \epsilon$ are all parameters, which we estimate.

We have to make assumptions for linear regression. We assume there is a linear relationship between X (the **predictor** variable) and Y (the **response** variable). We want $E[\epsilon|X = x] = 0$.

We assume $y_i = b_0 + b_1 x_i + e_i$. The **residual** for an observation is $e_i = y_i - \hat{y_i}$, and the **prediction** is $\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_i$. We want to minimize the **Residual Sum Squared** (RSS), which is $\sum_{i=1}^n e_i^2$. Minimizing RSS yields our estimators b_0 and b_1.

Differentiating, rearranging, and solving yields $b_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$, and $b_0 = \bar{y} - b_1 \bar{x}$. These are the **least squares estimates**. We can consider this as $b_1 = \frac{SXY}{SXX}$, where SXY (sample covariance) $= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$, and SXX (sample variance) $= \sum_{i=1}^n (x_i - \bar{x})^2$.

We can interpret a regression line equation by $\hat{\beta_0}$ is the 'baseline' for when $X = 0$, and $\hat{\beta_1}$ represents the expected change in Y per unit change in X.

Assumptions:

We make some key assumptions for linear regression:

- Linearity:
  We assume the data is linear; $E[Y|X = x]$ is linear with respect to the value of X. This means linearity in parameters, not linearity in the actual function.
- Independence:
  We assume each error term ϵ_i is independent of others, and we ideally want them to be iid.
- Homoscedasticity:
  This means that $Var[\epsilon|X = x]$ is a constant in terms of X. We want error terms with constant variance.
- Normality:
  This isn't 100% needed but we want $\epsilon \sim N(0, \sigma^2)$ to make interpretation and inferences easier.

To estimate the value of σ^2, the error variance, we use $S^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}$.

More on $\hat{\beta_1}$:

We know that $\hat{\beta_1} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$. This is equivalent to $\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i)}{\sum_{i=1}^n (x_i - \bar{x})(x_i)}$.

We can use this to show $E[\hat{\beta_1}|X = x_i] = \beta_1$. It is an unbiased estimator! Its variance is $\frac{\sigma^2}{SXX}$. Let $c_i = \frac{x_i - \bar{x}}{SXX}$. Then $\hat{\beta_1} = \sum_{i=1}^n c_i y_i$, which is Normally distributed via assumption of ϵ being normal.

We can't find σ^2 (usually), so we estimate a confidence interval using $S^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}$, and use a $t_{\text{df}=n-2}$ distribution.

More on $\hat{\beta_0}$:

Since $\hat{\beta_0} = \bar{y} - \hat{\beta_1} x$, we can find that $E[\hat{\beta_0} = \beta_0$. Also unbiased!

We can also find that $Var[\hat{\beta}_0] = \sigma^2(\frac{1}{n}+\frac{\bar{x}^2}{SXX})$. Similarly to $\hat{\beta}_1$, we know that $\hat{\beta}_0$ is Normally distributed, as it is a linear combination of Normals.

Confidence in the Regression Line:

For a given value x^*, the fitted/predicted value is $y^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$. This is an unbiased estimator, as $E[y^*|X = x^*] = E[\hat{\beta}_0 + \hat{\beta}_1 x^*|X = x^*] = \beta_0 + \beta_1 x^* = E[Y|X = x^*]$.

The variance for our regression line can be shown to be $Var[y^*|X = x^*] = \sigma^2(\frac{1}{n} + \frac{(x^*-\bar{x})^2}{SXX})$.

The prediction error for our regression line is $Var[y^* - \hat{y}^*|X = x^*] = \sigma^2(1 + \frac{1}{n} + \frac{(x^*-\bar{x})^2}{SXX})$.

ANOVA:

The **total sum of squares** (TSS or SST) is $\sum_{i=1}^n (y_i - \bar{y})^2$. We can decompose this into two terms, $RSS = \sum_{i=1}^n (y_i - \hat{y_i})^2 = (n-2)S^2$, and $SS_{Reg} = \sum_{i=1}^n (\hat{y_i} - \bar{y})^2$.

We can use an F-test to determine the goodness of fit for our line. We define $R^2 = \frac{SS_{Reg}}{SST}$ as a measure of variance explained by our regression line. Note that $R^2 = Corr(X,Y)^2$.

Additionally, we know that $\frac{SS_{Reg}}{\sigma^2} \sim \chi_1^2$, and $\frac{RSS}{\sigma^2} \sim \chi_{n-2}^2$.

Categorical Predictors:

Instead of doing 2-sample t-tests, we can simply do a linear regression with a dummy variable to check if a categorical predictor means anything. In this case, as X is discrete, the value $\hat{\beta}_1$ represents a difference in groups, not a slope.

Diagnostics

Diagnostics help us to determine whether our data is fit for linear regression analyses.

We check whether the error terms are homoscedastic, check for the linearity of our data, and check for independence and Normality by examining the residuals relative to the estimated values or observed values.

Leverages and Outliers:

Definition:

A **leverage point** is one with influence on the fitted model. Usually, it is one with $X = x_i$ far from $\bar{x}$. Bad leverage points change OLS estimates a lot, good leverage points don't.

An **outlier** is a point with $Y = y_i$ far from $\bar{y}$. We need to numerically determine these points to see how they influence our line.

Recall $\hat{\beta}_1 = \sum_{i=1}^n \frac{x_i - x}{SXX} y_i = \sum_{i=1}^n c_i y_i$. Rearranging yields $\hat{y_i} = \sum_{j=1}^n (\frac{1}{n} + \frac{(x_i-\bar{x})(x_j-\bar{x})}{SXX}) y_j = \sum_{j=1}^n h_{ij} y_j$, where h_{ij} are the elements of an $n \times n$ **hat matrix**. In this case, $h_{ij} = \frac{1}{n} + \frac{(x_i-\bar{x})(x_j-\bar{x})}{SXX}$. By this definition, $\sum_{i=1}^n h_{ii} = 2$, and $\sum_{i=1}^n h_{ij} = 1$.

High h_{ii} means the point is influential. We use the cutoff of $h_{ii} > \frac{4}{n}$, since the mean h_{ii} is $\frac{2}{n}$.

Residuals:

The residuals are defined as $e_i = y_i - \hat{y_i}$. We find that $Var(e_i) = \sigma^2(1 - h_{ii}) = \sigma^2(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{SXX})$. Residuals are not entirely independent, and they don't have equal variances!

We can standardize a residual using $r_i = \frac{e_i}{\sqrt{S^2(1-h_{ii})}}$, where $S^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}$. We should have r_is be Normally distributed.

Cook's Distance:

The Cook's Distance is a formula for the influence which a point has on a regression line. The formula is $D_i = \frac{r_i^2}{2} \frac{h_{ii}}{1-h_{ii}}$. We use the cutoff $D_i > \frac{4}{n-2}$ to determine if a point is too influential.

Variance Stabilizing:

We know that $f(Y) \approx f(E[Y]) + f'(E[Y])(Y - E[Y])$. Thus, $Var[Y] = (f'(E[Y]))^2 Var[Y]$.

We can do transformations like this on X, Y or both X and Y to make our data more linear and improve prediction error. This works but also sucks for interpreting results.

The Box-Cox transformation is a procedure which can help to make our data more linear, and almost always works for any distribution of data. It makes things almost impossible to understand, since we pick a very arbitrary power function ($Y \mapsto Y^\lambda$) which is hard to explain.

Multiple Linear Regression:

Motivation:

What if we want more than 1 predictor? We could have $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$. This produces a hyperplane, which we can't visualize but it is very useful.

We define $RSS(\hat{\beta}_0, \ldots, \hat{\beta}_p) = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \sum_{j=1}^p \hat{\beta}_j X_{i_j})^2$, and minimize with respect to betas to fit regressions.

Expectations of Vectors

If $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$, then $E[\mathbf{Y}] = (E[Y_1], E[Y_2], \ldots, E[Y_n])^T$, and $Var[\mathbf{Y}]$ is a variance-covariance matrix.

$\mathbf{Y}$ is an $n \times 1$ column vector, and $\mathbf{X}$ is an $n \times (p+1)$ matrix. The first column is all 1s.

$RSS = \epsilon^T \epsilon$, where ϵ is an $n \times 1$ column vector defined like with SLR. We find that $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, which is a $(p+1) \times 1$ column vector.

The projection matrix is $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Thus, we define the projection/hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, and $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$. Hat and residual matrix are both symmetric and idempotent.