MATH-338 Final Cheat Sheet [Linear Regression]

TERMS

Direction: Positive($+\beta_1$), Negative($-\beta_1$), no association, complex (parabolic). **Correlation:** unitless, invariant to linear transformation. Highly susceptible to outliers. Can have strong non linear association but correlation close to 0 (complex). $-\mathbf{1} \leq \mathbf{r} \leq \mathbf{1}$**:** when $r = 1$, all points on line with positive slope. When $r = -1$, all points on line with negative slope. Only interpretable for linear. **Coefficient of determination:** represents the proportion of variation y that is explained by/accounted for by the model. ANOVA tests whether this proportion is "significant". **Assumptions for Linear Regression Inference: 1:** linear model appropriate. **2:** residuals are normally distributed. **3:** residuals will have $\mu = 0$ and $\sigma =$?.

FORMULAS

- $y = \beta_o + \beta_1 x + \epsilon$ [standard formula for linear regression]
- $\epsilon \sim N(0, \sigma)$ [random variable]
- $r = \frac{1}{n-1} \times \Sigma(\frac{x-\bar{x}}{S_x})(\frac{y-\bar{y}}{S_y})$ [correlation]
- $x_i > \bar{x} \rightarrow y_i > \bar{y}$ [contribution to r is t]
- $\hat{y} = b_o + b_1 x$ [least squares regression line]
- $t = \frac{\text{Statistic - parameter}}{standarderror} = \frac{b_1 - \beta_1}{SE_{b_1}}$ $[\beta_1 = 0]$
- $F_{observed} \sim F(P, n-P-1)$
- $t_{observed} \sim t(n-p-1)$
- $b_1 = r \times \frac{s_y}{s_x}$
- $b_0 = \bar{y} - b_1 \bar{x}$
- $e_i = y_i - \hat{y}_i$ [prediction error (residual)]
- $r^2 = \frac{SSM}{SST}$ [coefficient of determination]

t-Test

- If P-value $\leq$ sig level $\implies$ reject $\mathrm{H_o}$ ($\mathrm{x_j}$ is a sig predictor of y)
- Else, $\mathrm{x_j}$ is not sig, $\therefore$ $\mathrm{x_j}$ does not have linear relation with y

ANOVA

Source	DF	Σ (squares)	μ (squares)	F	$Pr > F$
Model	p	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$\frac{SSM}{DFM} = MSM$	$F_{obs} = \frac{MSM}{MSE}$	p-value
Error (e_i)	$n-p-1$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$\frac{SSE}{DFE} = MSE$		
Total	$n-1$	$\sum_{i=1}^{n}(y_i - \bar{y}_i)^2$			

- $\mathrm{H_0}$: $\mu_1 = \mu_2 = ... = \mu_i$
- Population mean does not depend on group
- Population mean of y does not depend on x ($\mathrm{H_0}$: $\mu_{y|x} = \mu_y \implies \beta_1 = 0$)
- Reject $\mathrm{H_0}$: Our model is "significantly better" than null model at explaining changes in y $\implies$ we should use linear model
- Fail to reject $\mathrm{H_0}$: Our model is not significantly better than null model $\implies$ should use smaller model (null model)

SYMBOL CHART

β_o	y-intercept (b portion in $y = mx + b$)
β_1	slope (m portion in $y = mx + b$)
$\hat{y}$	predicted value of y
b_o	y-intercept for predicted
b_1	slope for predicted
r	coefficient of correlation
x^*	predictor for $\hat{\mu}$ and $\hat{y}$

t-Test for Slope

- $\mathrm{H_0}$: $\beta_1 = 0$ $t \sim t(n-2)$
- $\mathrm{H_a}$: $\beta \neq= 0$
- We reject $\mathrm{H_0}$ if slope is not 0, so a linear relationship exists between x and y.
- Fail to reject $\mathrm{H_0}$ and it is reasonable to believe that slope is 0 and there is no linear relationship between x and y.

Mean Response

- Model: $\mu_{y|x} = \beta_0 + \beta_1 x$
- $\uparrow$ n $\rightarrow$ $\uparrow$ CI $\downarrow$ relationship
- CI at x^* close to $\bar{x}$ $\rightarrow$ narrower
- CI at x^* far from $\bar{x}$ $\rightarrow$ wider
- $PI = \hat{y} \pm t^{**} \times SE_{\hat{y}}$
- We are 95% confident in our estimate that when x-variable is value of x^*, the population mean of y-variable for a new observation is between lower and upper bound.

Confidence Interval for Slope

- Want to find values of β_1 for which t-statistic is NOT in the critical region
- We are 95% confident in our estimate that when x-variable increases by 1 unit, the population mean of y-variable increases (1 decreases) by between lower bound and upper bound.

MULTIPLE LINEAR REGRESSION

basic model

- $\mu_{y|x} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$

least squares line

- $\hat{y} = b_0 + b_1 x_1 + ... + b_p x_p \implies y_i = \hat{y}_i + e_i$

ANOVA

- If p-value $\leq$ sig level $\implies$ reject $\mathrm{H_o}$ (our model is significantly better than null for prediction)
- Else, our model is not significantly better than the null for predicting
- Some x-variables may still be important predictors, however "more important" predictors are left out of the model (backward selection)