Regression Notes from slides

$$\Sigma (y_i - x_i) = \Sigma y_i - \Sigma x_i$$

$$\Sigma (x_i - a) = \Sigma x_i - na$$

$$\Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - n\bar{x}^2$$

$$\bar{x} = \Sigma x_i / n$$

$$\sigma_{\bar{x}}^2 = \left(\frac{\sigma}{\sqrt{n}}\right)^2 \qquad \bar{x} \sim (\mu, \sigma/\sqrt{n})$$

t-test statistic $\quad t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}} \quad$ n-1 d.f.

CI: Sample mean $\pm$ t_{crit} $s/\sqrt{n}$

Covariance: $E[(x - E(x))(y - E(y))]$

Sample cov: $\dfrac{1}{N-1} \Sigma (x_i - \bar{x})(y_i - \bar{y})$

Correlation $= \dfrac{Cov(x,y)}{sd(x)\,sd(y)}$

$$r = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2} \sqrt{\Sigma (y_i - \bar{y})^2}}$$

Measures direction but not slope
Does NOT imply independence

β_1 is the slope of the regression line and indicates the change in the mean of the distribution of y for one unit increase in x

β_0 is the intercept of the regression line. If it is sensible to think of a value of x=0 for a particular application, then β_0 gives the mean distribution of y at x=0

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1 \ldots n$$

Assumptions:

1. x_i is the i^{th} value of the predictor variable, which is a known constant for all i
2. The observations y_i (or ε_i) are independent
3. At any given x_i, y_i (or ε_i) is normally distributed.
4. The observations y_i (or ε_i) have constant standard deviation.
5. The means of y_i can be joined by a straight line given as:

$$E[y_i] = \beta_0 + \beta_1 x_i$$

where β_0 and β_1 are unknown parameters, such that:

B_1 is the slope of the regression line and indicates the change in the mean distribution of y per one unit increase in x.

B_0 is the intercept of the regression model. If it is sensible to think of a value of $x=0$ for a particular application, then β_0 gives the mean of the distribution of y at $x=0$. So it is not always possible to have a physical explanation of this parameter.

Least Squares:

Diff wrt to β_0 and β_1 $\quad \sum(y_i - \beta_0 - \beta_1 x_i)^2$

$$b_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$S_{xx} = \sum(x_i-\bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$S_{yy} = \sum(y_i-\bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

$$S_{xy} = \sum(x_i-\bar{x})(y_i-\bar{y}) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

Estimating σ^2

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$y_i - \hat{y}_i = e_i$$

$$SSE = \sum(y_i-\hat{y}_i)^2 = \sum(y_i - \beta_0 - \beta_1 x_i)^2 = \sum e_i^2 \text{ AKA RSE}$$

$$\hat{\sigma}^2 = MSE = \frac{SSE}{N-2} = \frac{\sum(y_i-\hat{y}_i)^2}{n-2}$$

Regression Notes from Slides

$E(MSE) = \sigma^2$

Inference about slope β_1:

$E(b_1) = \beta_1$

$Var(b_1) = \frac{\sigma^2}{\Sigma(x-\bar{x})^2} = \frac{\sigma^2}{Sxx}$

$b_1 \sim N(\beta_1, \frac{\sigma^2}{Sxx})$

$Var\hat{}(b_1) = S_{b_1}^2 = \frac{MSE}{Sxx}$ $\quad Se(b_1) = \sqrt{\frac{MSE}{Sxx}}$

T-test $t_{calc} = \frac{\text{Estimator} - \text{value from } H_0}{Se(\text{estimator})}$ $\quad (n-2)$ d.F.

CI: (estimator) $\pm$ $t_{critical}$ Se (Estimator)

Important, want to test $\beta_1 = 0$, if it equals zero, no relationship exists

Inference About intercept β_0

$\beta_0 \Rightarrow b_0 = \bar{y} - \bar{x} b_1$

$E(b_0) = \beta_0$

$Var(b_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{Sxx}\right]$

$b_0 \sim N(\beta_0, \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{Sxx}])$

$Se(b_0) = \sqrt{\sigma[\frac{1}{n} + \frac{\bar{x}^2}{Sxx}]}$

Sampling Distribution of $\frac{b_0 - \beta_0}{Se(b_0)}$ is t-dist with $n-2$ d.F.

Inference for $E[y]$ at $X = X_{\text{some value}}$

Make inference about mean distribution of y ie $E[y]$ at x = some x.
- Want to construct point estimate and CI for $E(y|x=x)$.

$$E[y|x'] = \beta_0 + \beta_1 x'$$
$$\hat{y}' = b_0 + b_1 x'$$

$$Var[\hat{y}'] = \sigma^2\left[\frac{1}{n} + \frac{(x'-\bar{x})^2}{Sxx}\right] \leftarrow$$ se $(\hat{y}')$ is σ of error

$$\hat{y}' \sim N\left((\beta_0+\beta_1 x'),\ \sigma^2\left[\frac{1}{n} + \frac{(x'-\bar{x})^2}{Sxx}\right]\right)$$

Hypothesis testing for $\hat{y}$.

$H_0 : \hat{y}_i = y'$ vs $H_1 : \hat{y}' \neq y'$

Test Statistic is $\dfrac{\hat{y}' - y'}{se(\hat{y}')}$

CI $\hat{y}' \pm se(\hat{y}')t_{critical}$

Prediction of a new observation

In first case we refer to the mean distribution of y for a particular value of X, whereas in PI, we predict an individual outcome drawn from the distribution of y for a given x

$\hat{y}_{new} = b_0 + b_1 x'$ $\quad E[\hat{y}'_{new}] = y'_{new}$

$Var(\hat{y}_{new}) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x'-\bar{x})^2}{Sxx}\right]$

Variance has 2 components:

1. Variance of the sampling distribution of fitted value $\hat{y}_i$
2. Variance of distribution of y at some x

Prediction interval $\quad \hat{y}'_{new} \pm t_{critical}\sqrt{MSE\left[1 + \frac{1}{n} + \frac{(x'-\bar{x})^2}{Sxx}\right]}$

Regression Notes (slides)

$y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i$ or $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$

ANOVA

$SSTO = \sum(y_i - \bar{y})^2$ Total uncorr $\sum y_i^2$

If SSTO all responses are equal

SSE $\sum(y_i - \hat{y}_i)^2$ variability around fitted line

$SSTO = SSE + SSR$

$SSR = \sum(\hat{y}_i - \bar{y})$ variability associated with the regression line

$y_i - \bar{y} = \hat{y}_i - \bar{y} + y_i - \hat{y}_i$

SSTO SSE SSR

$\sum y = \sum \hat{y}_i$

Source of variation	SS	DF	MS	F
Regression	SSR $\sum(\hat{y}_i - \bar{y})^2$	1	MSR $SSR/1$	$F_{calc} = \frac{MSR}{MSE}$
Error	SSE $\sum(y_i - \hat{y}_i)^2$	n-2	MSE $SSE/n-2$	
Total	$\sum(y_i - \bar{y})^2$	n-1		

F_{calc} can be used to test $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$ follow $F_{1,n-2}$ dist.

If $F_{calc} \leq F_{(1-\alpha, 1, n-2)}$ do not reject H_0

$F = (t)^2$

$$\Sigma (y_i - b_0 - b_1 x_i)^2$$

db_0 $\boxed{b_0}$

$$-2\,\Sigma (y_i - b_0 - b_1 x_i) = 0$$

$$\Sigma (y_i - b_0 - b_1 x_i) = 0$$

$$\Sigma y_i - \Sigma b_0 - b_1 \Sigma x_i = 0 \qquad \frac{\Sigma x_i}{n} = \bar{x}$$

$$n b_0 = \Sigma y_i - b_1 \Sigma x_i$$

$$b_0 = \frac{\Sigma y_i}{n} - b_1 \frac{\Sigma x_i}{n}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

db_1 $\boxed{b_1}$

$$= -2\,\Sigma x_i (y_i - b_0 - b_1 x_i)$$

$$-2\,\Sigma (y_i x_i - b_0 x_i - b_1 x_i^2) = 0$$

$$\Sigma y_i x_i - b_0 \Sigma x_i - b_1 \Sigma x_i^2 = 0 \qquad b_0 =$$

$$b_1 \Sigma x_i^2 = \Sigma y_i x_i - b_0 \Sigma x_i$$

$$b_1 \Sigma x_i^2 = \Sigma y_i x_i - \left[\frac{\Sigma y_i}{n} - b_1 \frac{\Sigma x_i}{n}\right] \Sigma x_i$$

$$b_1 \Sigma x_i^2 = \Sigma y_i x_i - \frac{\Sigma y_i x_i}{n} + \frac{b_1 \Sigma x_i^2}{n}$$

$$b_1 \Sigma x_i^2 - b_1 \frac{\Sigma x_i^2}{n} = \Sigma y_i x_i - \frac{\Sigma y_i x_i}{n}$$

$$b_1 \left[\Sigma x_i^2 - \frac{\Sigma x_i^2}{n}\right] = \Sigma y_i x_i - \frac{\Sigma y_i x_i}{n}$$

$$b_1 = \frac{\Sigma y_i x_i - \frac{\Sigma y_i x_i}{n}}{\Sigma x_i^2 - \frac{\Sigma x_i^2}{n}}$$

$\bullet\, y_i$
$\bullet\, \hat{y}_i$
$\bullet\, \bar{y}$

$$\Sigma(y_i - \bar{y})^2 = \Sigma(y_i - \hat{y})^2 + \Sigma(\hat{y} - \bar{y})^2$$

$$\Sigma[(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})]^2$$

$$= \Sigma(y_i - \hat{y})^2 + \Sigma(\hat{y}_i - \bar{y})^2 + 2\,\Sigma(y_i - \hat{y}_i)(\hat{y} - \bar{y})$$

$$2\,\Sigma(y_i - \hat{y}_i)(\hat{y} - \bar{y})$$

$$\Sigma[\,\hat{y}(y_i - \hat{y}_i) - \bar{y}(y_i - \hat{y}_i)$$

$$\Sigma[(\bar{y} + b_1(x_i - \bar{x}))(y_i - \hat{y}_i) - \bar{y}(y_i - \hat{y}_i)$$

$$\Sigma\bar{y}(y_i - \hat{y}) + \Sigma b_1(x_i - \bar{x})(y_i - \hat{y}) \cancel{- \Sigma\bar{y}\, y_i - \hat{y}_i}$$

canc

$$\hat{y}_i = b_0 + b_1 x_i$$

$$b_0 = \bar{y} - b_1 x_i$$

$$\hat{y}_i = \bar{y} - b_1\bar{x}_i + b_1 x_i$$

$$\hat{y}_i = \bar{y} + b_1(x_i - \bar{x})$$

$$\Sigma b_1(x_i - \bar{x})(y_i - \hat{y}).$$

$$b_1\Sigma \underbrace{x_i(y_i - \hat{y}_i)}_{=0} - \Sigma\bar{x}(y_i - \hat{y})$$

$$\underbrace{-\Sigma\bar{x}y_i + \Sigma\bar{x}\hat{y}_i}_{=0} \quad \Sigma\hat{y}_i = \hat{y}_i$$

$$\Sigma y_i = \Sigma\bar{y} + b_1(x_i - \bar{x})$$

$$\Sigma\bar{y} + \underbrace{\Sigma b_1(x - \bar{x})}_{=0}$$

$$\Sigma\hat{y} = \Sigma\bar{y}$$

$$= n\bar{y} =$$

$$\Sigma\hat{y}_i = \Sigma y_i$$

$$b_1\Sigma x_i(y_i - \hat{y}_i)$$

$$b_1\,\Sigma x_i(y_i - [\bar{y} + b_1(x_i - \bar{x})])$$

$$b_1\,\Sigma x_i(y_i - \bar{y} - b_1(x_i - \bar{x}))$$

$$b_1\,\Sigma x_i(y_i - \bar{y}) - b_1\Sigma x_i(x_i - \bar{x})$$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

$$\Sigma(y_i - \bar{y})(x_i - \bar{x}) - b_1\Sigma(x_i - \bar{x})^2$$

$$\Sigma(y_i - \bar{y})(x_i - \bar{x}) - \Sigma(y_i - \bar{y})(x_i - \bar{x}) = 0$$