STATS 413 Midterm Cheatsheet - Vince Wei

z-score Regression

$$x_i = \overline{x} + z(x_i)\text{stdev}(x) \tag{1}$$
$$r(x,y) = \frac{1}{n-1}\sum_{i=1}^{n} z(x_i)z(y_i) \tag{2}$$
$$z(\hat{y}) = r \cdot z(x) \tag{3}$$
$$\frac{\hat{y}-\overline{y}}{\text{stdev}(y)} = r \cdot \frac{x-\overline{x}}{\text{stdev}(x)} \tag{4}$$
$$\hat{y} = r\frac{\text{stdev}(y)}{\text{stdev}(x)}x + \overline{y} - r\frac{\text{stdev}(y)}{\text{stdev}(x)}\overline{x} \tag{5}$$
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{6}$$

Where $\hat{\beta}_1 = r\frac{\text{stdev}(y)}{\text{stdev}(x)}$ and $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$.

OLS Estimators

Weaker linear model assumptions:

1. **Linearity**: $E(y|X) = X\beta$, $E(\varepsilon) = \mathbf{0}$
2. **Homoskedasticity**: $Var(\varepsilon_i|x_i) = \sigma_\varepsilon^2$
3. **Uncorrelated noise**: $Cov(\varepsilon_i, \varepsilon_j|x_i, x_j) = 0$ for $i \neq j$

OLS aims to optimize for:

$$\min_{\tilde{\beta}} \sum_{i=1}^{n} e_i^2 = \min_{\tilde{\beta}} e^T e \tag{7}$$
$$= \min_{\tilde{\beta}} (y - X\tilde{\beta})^T(y - X\tilde{\beta}) \tag{8}$$

The linear regression estimator $\hat{\beta}$ must satisfy the normal equation:

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

As long as $X^T X$ is invertible, i.e. $\text{rank}(X) = p+1$, we will have a unique solution.

If $X^T X$ is invertible, the unique solution is

$$\hat{\beta}_{(p+1)\times 1} = (X^T X)^{-1}X^T y$$

Expectations and Variances of y:

$$E(y) = E(X\beta) + \epsilon = E(X\beta) = X\beta \tag{9}$$
$$Var(y) = Var(X\beta) + Var(\epsilon) = \sigma_\epsilon^2 I_{n\times n} \tag{10}$$

Hat Matrix:

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1}X^T y = Hy \tag{11}$$
$$H = X(X^T X)^{-1}X^T \tag{12}$$
$$\hat{y} = Hy \tag{13}$$
$$e = y - \hat{y} = (I - H)y \tag{14}$$

The hat matrix H has the following properties:

$$H^T = H \tag{15}$$
$$HH = H \tag{16}$$

For any vector $a \in R^n$ in the column space of X (i.e. $a = Xb$ for some vector $b \in R^{p+1}$)m we have

$$a^T e = 0$$

Expectations and Variances of β:

$$E(\hat{\beta}) = \beta \tag{17}$$
$$Var(\hat{\beta}) = \sigma_\epsilon^2(X^T X)^{-1} \tag{18}$$

Residual Sum of Squares (RSS):

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y-\hat{y})^2 = (y - X\hat{\beta})(y - X\hat{\beta})^T = e^T e$$

Total Sum of Squares (TSS):

$$TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2 = (n-1)\text{Var}(y)$$

R^2:

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

Additionally,

$$R^2 = Cor(\hat{y}, y)$$

in the case of simple regression ($p = 1$), $R^2 = Cor(x, y)$.

Hypothesis Tests and Confidence Intervals

A Type I Error occurs if we reject the null hypothesis when it was actually true ($P(\text{Type 1}|H_0) \leq \alpha$).
A Type II Error occurs if we fail to reject the null hypothesis when it was actually false (power $= 1 - P(\text{Type II Error})$).
Stronger Linear Model:

1. **Linearity**: $E(y_i) = x_i^T\beta$
2. **Homoskedasticity**:$Var(y_i) = \sigma_\epsilon^2$
3. **Normality**: $\epsilon_i \sim N(0, \sigma_\epsilon^2)$

Distributions, Expectations, Variances of $\hat{\beta}$:

$$\hat{\beta} \sim MVN(\beta, \sigma_\epsilon^2(X^T X)^{-1})$$

RMSE:

$$\hat{\sigma_\epsilon} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-p-1}} = \frac{RSS}{n-p-1}$$

where $n-p-1$ are referred to as the degrees of freedom for the residual vector e.
With an estimation for σ_ϵ , we can now estimate $\text{stdev}(\hat{\beta}_j)$ using $\text{se}(\hat{\beta}_j)$, the standard error for the j th coefficient:

$$\text{se}(\hat{\beta}_j) = \hat{\sigma_\epsilon}\sqrt{(X^T X)^{-1}_{(j+1),(j+1)}}$$

With this, we'll consider a new test statistic which replaces $\text{stdev}(\hat{\beta}_j)$ with its sample analogue $\text{se}(\hat{\beta}_j)$

$$t_{\text{stat}} = \frac{\hat{\beta}_j^{\text{obs}} - \gamma_0}{\text{se}(\hat{\beta}_j)}$$

Confidence Interval (Two-tailed):

$$\hat{\beta}_j \pm t_{1-\frac{\alpha}{2},n-p-1}\text{se}(\hat{\beta}_j)$$

The corresponding command for $t_{1-\frac{\alpha}{2},n-p-1}$ is `qt(1-a/2, n-p-1)`.
If our alternative is two-sided and we conduct a test with significance level α, we can reject the null γ iff γ DOES NOT fall within the $100(1-\alpha)\%$ confidence interval.

$\mathcal{F}$-Tests

We are interested in the hypothesis tests:

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$

and

$$H_a = \text{ at least one } \beta_j \neq 0$$

The statistic takes on the form:

$$F_{\text{stat}} = \frac{R^2}{1-R^2}\left(\frac{n-p-1}{p}\right)$$

Formally, F_{stat} follows a $\mathcal{F}_{p,n-p-1}$ distribution, where p is the number of predictor variables and n is the sample size. We calculate the p -value as $P(\mathcal{F}_{p,n-p-1} \geq F_{\text{stat}})$: here the extremes are always in the right tail.
Alternatively, we may write

$$F_{\text{stat}} = \frac{\frac{TSS-RSS}{p}}{\frac{RSS}{n-p-1}} = \frac{\frac{TSS-RSS}{p}}{\hat{\sigma}_\epsilon^2} = \left(\frac{R^2}{1-R^2}\right)\left(\frac{n-p-1}{p}\right)$$

Interactions

The interaction term between a continuous and categorical covariate takes on the following form:

$$X_i\mathbf{1}\{X_j = c\} = \begin{cases} 0 & X_j \neq c \\ X_i & X_j = c \end{cases}$$

We are interested in the null hypothesis that some subset of slopes $\mathcal{I} \subseteq \{1, \ldots, p\}$ are zero:

$$H_0 : \beta_i = 0 \text{ for some subset of covariates } \mathcal{I}$$

The alternative hypothesis is then

$$H_a : \text{at least one slope in subset } \mathcal{I} \text{ is non zero}$$

The test statistic may be expressed as:

$$F_{\text{stat}} = \frac{(RSS_{\text{red}} - RSS_{\text{full}})/(\text{df}_{\text{red}} - \text{df}_{\text{full}})}{RSS_{\text{full}}/\text{df}_{\text{full}}}$$

Under the null: $F_{\text{stat}} \sim \mathcal{F}_{(\text{df}_{\text{red}}-\text{df}_{\text{full}}),\text{df}_{\text{full}}}$.

Confidence Intervals for Expected Response

We have that

$$a^T\hat{\beta} = \sum_{j=0}^{p} a_{j+1}\hat{\beta}_j$$

$$\begin{aligned}
\text{Var}(\hat{\beta}_{\text{Advert, South}}) &= \text{Var}(a^T\hat{\beta}) && (19)\\
&= \sigma_\epsilon^2 a^T(X^TX)^{-1}a && (20)\\
&= \text{Var}(\hat{\beta})_{(5,5)} + \text{Var}(\hat{\beta})_{(8,8)} + 2\text{Var}(\hat{\beta})_{(5,8)} && (21)\\
&= \text{Var}(\hat{\beta}_{\text{Advert}}) + \text{Var}(\hat{\beta}_{\text{South}}) + \text{Cov}(\hat{\beta}_{\text{A}}, \hat{\beta}_{\text{S}}) && (22)\\
\implies \text{stdev}(\hat{\beta}_{\text{Advert, South}}) &= \sigma_\epsilon\sqrt{a^T(X^TX)^{-1}a} && (23)\\
\implies \text{se}(\hat{\beta}_{\text{Advert, South}}) &= \hat{\sigma}_\epsilon\sqrt{a^T(X^TX)^{-1}a} && (24)
\end{aligned}$$

Distributions, Expectations, Variances of Inferences:

$$E(\hat{\mu}_{y|\tilde{x}}) = \mu_{y|\tilde{x}}$$

$$\text{Var}(\hat{\mu}_{y|\tilde{x}}) = \sigma_\epsilon^2\tilde{x}^T(X^TX)^{-1}\tilde{x}$$

$$T = \frac{\hat{\mu}_{y|\tilde{x}}}{\hat{\sigma}_\epsilon\sqrt{\tilde{x}^T(X^TX)^{-1}\tilde{x}}} \sim t_{n-p-1}$$

$$\text{se}(\hat{\mu}_{y|\tilde{x}}) = \hat{\sigma}_\epsilon\sqrt{\tilde{x}^T(X^TX)^{-1}\tilde{x}}$$

We may construct a $100(1-\alpha)\%$ t -based confidence interval for $\mu_{y|\tilde{x}}$:

$$\hat{\mu}_{y|\tilde{x}} \pm t_{1-\alpha/2,n-p-1}\text{se}(\hat{\mu}_{y|\tilde{x}})$$

$$t_{\text{stat}} = \frac{\hat{\mu}_{y|\tilde{x}} - \mu_0}{\text{se}(\hat{\mu}_{y|\tilde{x}})}$$

Prediction Intervals

Distributions, Expectations, Variances of New Observations:

$$E(y^\star - \hat{\mu}_{y|\tilde{x}}) = 0$$

$$\begin{aligned}
\text{Var}(y^\star - \hat{\mu}_{y|\tilde{x}}) &= \text{Var}(y^\star) + \text{Var}(\hat{\mu}_{y|\tilde{x}}) && (25)\\
&= \sigma_\epsilon^2 + \sigma_\epsilon^2\tilde{x}^T(X^TX)^{-1}\tilde{x} && (26)\\
&= \sigma_\epsilon^2(1 + \tilde{x}^T(X^TX)^{-1}\tilde{x}) && (27)
\end{aligned}$$

$$T = \frac{y^\star - \hat{\mu}_{y|\tilde{x}}}{\hat{\sigma}_\epsilon\sqrt{(1 + \tilde{x}^T(X^TX)^{-1}\tilde{x})}} \sim t_{n-p-1}$$

Under the stronger linear model. Based on this statistic, an $100(1-\alpha)\%$ prediction interval for a covariate $\tilde{x}$ is

$$\hat{\mu}_{y|\tilde{x}} \pm t_{1-\alpha/2,n-p-1}\hat{\sigma}_\epsilon\sqrt{(1 + \tilde{x}^T(X^TX)^{-1}\tilde{x})}$$

Regression Diagongstics

$$e = (I - H)y$$

Let $\text{cov}(\cdot,\cdot)$ be the sample covariance operator. Then,

$$\text{cov}(e, x_j) = 0$$

$$\text{cov}(e, \hat{y}) = 0$$

And

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

Under the weaker linear model

$$E(e) = E((I-H)y) = \mathbf{0}$$

using $E(y - Hy) = E(y - X\hat{\beta}) = 0$ and $E(\hat{\beta}) = \beta$, $E(y - X\beta) = 0$ under assumptions; and

$$E(\hat{y}) = E(Hy) = X\beta$$

Under the weaker linear model

$$\text{Var}(e) = \sigma_\epsilon^2(I - H)$$

using $\text{Var}[(I-H)y] = (I-H)\text{Var}(y)(I-H)^T$ $\text{Var}(y) = \sigma_\epsilon^2$ under assumptions; and

$$\text{Var}(\hat{y}) = \sigma_\epsilon^2 H$$

using $\text{Var}(\hat{y}) = \text{Var}(Hy) = H\text{Var}(y)H^T$ and $\text{Var}(y) = \sigma_\epsilon^2$ under assumptions. When performing regression with an intercept and p predictors, the **diagonal entries** of the hat matrix H, h_{ii} , satisfy the following

$$\frac{1}{n} \le h_{ii} \le 1$$

for $i = 1, \ldots, n$. And

$$\sum_{i=1}^{n} h_{ii} = p + 1$$

This implies that

$$\text{Var}(e_i) = \sigma_\epsilon^2(1 - h_{ii}) \le \sigma_\epsilon^2 = \text{Var}(\epsilon_i)$$

In addition, for any $i, j = 1, \ldots, n$ we have

$$\text{Cov}(e_i, e_j) = -\sigma_\epsilon^2 h_{ij}$$

$$\text{Cov}(\hat{y}_i, \hat{y}_j) = \sigma_\epsilon^2 h_{ij}$$

Covariance of Residuals and Fitted Values:

$$\begin{aligned}
\text{Cov}(e, \hat{y}) =& \text{Cov}((I-H)y, Hy) && (28)\\
&= (I-H)\text{Cov}(y,y)H^T && (29)\\
&= \text{Cov}(y,y)(I-H)H^T && (30)\\
&= \text{Cov}(y,y)0 \qquad (H^T = H, HH = H) && (31)\\
&= \mathbf{0} && (32)
\end{aligned}$$

Distribution of Residuals and Fitted Values:

$$\begin{bmatrix} e \\ \hat{y} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mathbf{0} \\ X\beta \end{bmatrix}, \begin{bmatrix} (I-H)\sigma_\epsilon^2 & \mathbf{0} \\ \mathbf{0} & H\sigma_\epsilon^2 \end{bmatrix}\right) \qquad (33)$$

and e and $\hat{y}$ are independent of one another (relies crucially on multivariate normality).

Detecting non-linearity: Residual Plots (e on y-axis and one predictor x or fitted values $\hat{y}$ on the x-axis)

Standardized / Internally Studentized Residuals:

$$r_i = \frac{e_i}{\hat{\sigma}_\epsilon\sqrt{1 - h_{ii}}}$$

Detecting Heteroskedasticity

1. Plot r_i against x_{ij} for $j = 1, \ldots, p$
2. Plot r_i against $\hat{y}_i$
3. Plot $\sqrt{|r_i|}$ against x_{ij} for $j = 1, \ldots, p$
4. Plot $\sqrt{|r_i|}$ against $\hat{y}_i$

With any of these approaches, want to see if the magnitude of the residuals is varying as a function of x_{ij} or $\hat{y}_i$.