回归分析期末试卷 (2023.2.24)　　**姓名:**　　**学号:**

1.（20 分）**二选一**

(a) 右图中_____ 是回归直线 (A: 实线　B: 虚线)。

(b) 两个以上自变量的回归模型中回归系数与 _____ 成正比（A: 相关系数 B: 偏相关系数）。

(c) 线性回归模型中，若某一个自变量与其它变量高度相关，则该变量的回归系数 LS 估计的方差_____ （A: 很大　B: 很大）。

(d) 简单线性回归中，若自变量的样本方差越大，则斜率项 (自变量的回归系数）的最小二乘估计的方差_____ （A: 越大　B: 越小）。

(e) 若样本点 i 的 h_{ii} 接近于 1，则该点的_____ 对回归分析影响较大（A: 自变量　B: 响应变量）。

(f) 回归分析中删除一个数据点会导致回归平方和不变或_____（A: 增大　B: 减少）。

(g) 回归分析中增加一个变量会导致决定系数不变或 _____ （A: 增大　B: 减少）。

(h) 单因素方差分析模型假定各组的_____ 相同（A: 均值　B: 方差）。

(i) 两因素方差分析模型同时考虑处理和区组是因为它们可能 _____ （A: 相关　B: 不相关）。

(j) 假设随机选取若干长宽独立的长方形，测量其中一条边的长度 L，并以某种方式近似测得面积 S。拟合回归模型 $\log(S) = a + b\log(L) + \epsilon$, 则 LS 估计 $\widehat{b}$ 的期望等于_____（A: 1　B: 2）。

2.（15 分）**简要回答**

(a) 假设随机向量 $\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$ 的方差-协方差矩阵为 $\begin{pmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}} \end{pmatrix}$, 写出 $\mathbf{y}$ 关于 $\mathbf{x}$ 的去相关化及其方差。

(b) 应用简单线性回归模型分析 2001 年家庭人口调查数据，拟合得到妻子教育水平（WifeEdLevel, 上学的年数）与丈夫教育水平 (HusbandEdLevel) 的关系如下

$$\text{WifeEdLevel} = 5.60 + 0.57 \times \text{HusbandEdLevel} + \text{Residual},$$

如果公司今年计划送王先生到大学在职培训一年，你是否预期王太太的教育水平会相应地提高 0.57(年)? 若不是，解释 0.57 的含义。

(c) 你打算应用简单回归模型 $y = a + bx + \epsilon$ 分析体重 y 与身高 x 的关系，但你的朋友认为需要控制性别 z，分析你的朋友为什么要考虑控制 z? 如何在模型中实现对 z 的控制?

(d) 线性回归分析中，回归平方和度量了所有拟合值关于其中心的分散程度：$SS_{\text{回}} = \|\widehat{\mathbf{y}} - \mathbb{1}\bar{y}\|^2 = \sum_{i=1}^n (\widehat{y}_i - \bar{y})^2$, 其中 $\widehat{\mathbf{y}} = (\widehat{y}_1, ..., \widehat{y}_n)^\top$ 是响应变量向量 $\mathbf{y} = (y_1, ..., y_n)^\top$ 的 LS 拟合值，$\bar{y}$ 是 $y_1, ..., y_n$ 的样本均值。解释为什么所有拟合值的中心取为 $\bar{y}$。

(e) 假设线性模型

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{x}_k\beta_k + X_{-k}\boldsymbol{\beta}_{-k} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim (\mathbf{0}, \sigma^2 I_n),$$

其中设计阵 X 的第 k 列记为 $\mathbf{x}_k$ ($\|\mathbf{x}_k\| = 1$), 对应的回归系数为 β_k，其它各列组成的矩阵记为 X_{-k}, 对应的回归系数为 $\boldsymbol{\beta}_{-k}$。记 $\mathbf{x}_k^\perp = \mathbf{x}_k - P_{X_{-k}}\mathbf{x}_k$，其中 $P_{X_{-k}}$ 为 X_{-k} 张成的子空间对应的投影矩阵。写出 β_k 的 LS 估计及其方差，该方差何时最小？给出最小值。

3.（20 分）数据 *brains* 给出了 62 种哺乳动物的平均脑重量 (BrainWt, 单位：g) 和平均体重 (BodyWt, 单位：kg). R 分析该数据得到如下回归分析结果

Call: lm(formula = log(BrainWt) ~ log(BodyWt), data = brains)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.14	0.10	①	9e-30
log(BodyWt)	②	0.03	25	2e-34

Residual standard error: 0.69 on ③ degrees of freedom

Multiple R-squared: ④

F-statistic: ⑤ on ⑥ and ⑦ DF, p-value: ⑧

(a) 填上 ① - ⑧ 处的数字 (其中⑤是回归方程显著性的 F 检验统计量，⑥和⑦为其自由度).

(b) log(BrainWt) 与 log(BodyWt) 是正相关还是负相关，其样本相关系数等于多少？log(BrainWt) 与其拟合值的样本相关系数等于多少?

4.（15 分）设有数据 $(x_i, y_i), i = 1, 2, ..., 50$, 其样本相关系数为 $r = 0.5$, 样本均值分别为 $\overline{x} = 1.0, \overline{y} = 2.0$, 样本标准差分别为 $s_x = 1.0, s_y = 0.8$。现假设数据满足线性回归模型

$$y_i = a + bx_i + \epsilon_i,\ \epsilon_i,\ \text{iid}\ \sim N(0, \sigma^2)\ \text{且与}\ x_i\ \text{独立}, i = 1, ..., 50.$$

(a) 求 a, b, σ^2 的 LS 估计 $\widehat{a}, \widehat{b}, \widehat{\sigma}^2$。

(b) 计算 $H_0 : b = 0$ 的 t 检验统计量的值，并说明该检验在原假设下服从什么分布。

5.（15 分）假设线性模型 $\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1} = \mathbb{1}\beta_0 + Z\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, 其中 $\boldsymbol{\epsilon} \sim (\mathbf{0}, \sigma^2 I_n)$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \boldsymbol{\gamma} \end{pmatrix}$, $X = (\mathbb{1}, Z)$ 列满秩, $\mathbb{1}$ 是分量全为 1 的向量，Z 为 $n \times (p-1)$ 矩阵，其中心化矩阵记作 Z_c。记 $\widehat{\boldsymbol{\beta}}$ 为 $\boldsymbol{\beta}$ 的 LS 估计, $\widehat{\mathbf{y}} = X\widehat{\boldsymbol{\beta}}$ 为 LS 拟合值。定义 $\widetilde{\mathbf{y}} = \mathbb{1}\overline{y}$，均方误差 $m(\widetilde{\mathbf{y}}) = E\|\widetilde{\mathbf{y}} - X\boldsymbol{\beta}\|^2$，$m(\widehat{\mathbf{y}}) = E\|\widehat{\mathbf{y}} - X\boldsymbol{\beta}\|^2$。证明当 $\|Z_c\boldsymbol{\gamma}\|^2 \le (p-1)\sigma^2$ 时, $m(\widehat{\mathbf{y}}) \le m(\widetilde{\mathbf{y}})$。

6.（15 分）假设 n 个样本点 $(y_i, \mathbf{x}_i), i = 1, 2, ..., n$ （$\mathbf{x}_i$ 的第一分量为 1）满足线性回归模型

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i,\ \ \epsilon_i \sim (0, \sigma^2)\text{与}\mathbf{x}_i\text{独立},\ i = 1, ..., n,$$

得到残差 $e_i, i = 1, 2, ..., n$, 拟合值 $\widehat{y}_i, i = 1, 2, ..., n$, 残差平方和 $RSS = \sum_{i=1}^n e_i^2$。

(a) 证明如下结论：如果第 k（$1 \le k \le n$）个数据点的自变量 $\mathbf{x}_k = \overline{\mathbf{x}}$, 则 $\widehat{y}_k = \overline{y}$。进一步，若 $y_k = \overline{y}$, 则删除该点不影响回归系数估计和残差平方和。

(b) 现删除数据点 $(y_k, \mathbf{x}_k)$, $1 \le k \le n$, 重新拟合前述模型得到残差平方和 RSS_{-k}, 证明 $RSS_{-k} \le RSS - e_k^2$。

1.（20 分）

(a) A (实线)

(b) B （偏相关系数）

(c) A/B （很大）

(d) B （越小）

(e) A （自变量）

(f) B （减小）

(g) A （变量越多，残差平和越小，决定系数越大）

(h) B （方差分析模型假定各组的方差相同，即方差齐性）

(i) A (相关）- 该题含义不清，忽略

(j) A (1, 这是因为真实面积公式 $S_{true}=$ 长 $\times$ 宽 $=L\times W, \log(S_{true})=\log(L)+\log(W)$，而我们假设了 WL, 故未测量的 W 可以放到误差项 ϵ 中，所以模型 $\log(S)=a+b\log(L)+\epsilon$ 中误差项 ϵ 与自变量 $\log(L)$ 独立，且 b 的真值为 1，则 b 的 LS 估计是无偏的，其期望等于 $b=1$）

2.（15 分）简要回答

(a) $\mathbf{y}^{\perp}=\mathbf{y}-\Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\mathbf{x}$，$var(\mathbf{y}^{\perp})=\Sigma_{\mathbf{yy\bullet x}}$

(b) 不是，0.57 的含义是：若研究对象中某人比另外一人多上一年学，则其妻子期望比另外一人的妻子多上 0.57 年学。

(c) z 可能是干扰因素（既与响应 y 有关，也与自变量 x）有关。在线性模型中添加变量 z: $y=a+bx+cz+\epsilon$ 即控制了 z 的干扰。

(d) $\overline{y}$ 也是所有拟合值的平均值，这是因为 LS 方法的正则方程 $\sum(y_i-\hat{y}_i)=0$.

(e) $\hat{\beta}_k=\mathbf{x}_k^{\perp'}\mathbf{y}/\|\mathbf{x}_k^{\perp}\|^2$, $var(\hat{\beta}_k|X)=\sigma^2/\|\mathbf{x}_k^{\perp}\|^2$.
因为 $\|\mathbf{x}_k^{\perp}\|^2\le\|\mathbf{x}_k\|^2=1$, 当 $\mathbf{x}_k^{\perp}=\mathbf{x}_k$ 即 $\mathbf{x}_k$ 与其它各列正交时，$\hat{\beta}_k$ 的方差最小（最小值为 σ^2。（参见 1(c)).

3.（20 分）

①：21.4 ②：0.75 ③：60

④：0.91 ($F=t^2=25^2=615=(n-p)\times R^2/(1-R^2)=(60R^2/1-R^2)\rightarrow R^2=625/685=0.91$)) ⑤：625 ($25^2$)) ⑥：$60=n-p$ ⑦：$1=p-1$

⑧ $2e-34$ (与斜率显著性检验的 p 值相同)

正相关（因为 log(BodyWt) 的系数估计 0.75, 所以正相关）。相关系数都是 $\sqrt{R^2}=\sqrt{0.91}=0.95$

4.（15 分）$\hat{b}=s_{xy}/s_{xx}=rs_xs_y/s_x^2=0.5\times0.8/1=0.4$.

$\hat{a}=\overline{y}-\overline{x}\hat{b}=2-0.4=1.6$.

$\hat{\sigma}^2=(n-1)/(n-2)(1-r^2)s_y^2=49/48\times(1-0.25)\times0.8^2=0.49$ 或略去 $(n-1)/(n-2), \hat{\sigma}^2=0.48$

$t=\sqrt{s_{xx}}\hat{b}/\hat{\sigma}=\sqrt{(n-1)s_x^2}\,\widehat{b}/\hat{\sigma}=\sqrt{50-1}\times0.4/\sqrt{0.49}=4$. 原假设下 $t\sim t_{n-2}=t_{48}$.

5.（15 分）

证明：（MSE 的分解）

首先注意到 LS 拟合 $\widehat{\mathbf{y}}$ 的 MSE 等于其方差：

$$M(\widehat{\mathbf{y}}) = var(\widehat{\mathbf{y}}) = P_X\sigma^2$$

所以 $m(\widehat{\mathbf{y}}) = tr(M) = p\sigma^2$。

另一方面，$\widetilde{\mathbf{y}} = P_{\mathbb{1}}\mathbf{y}$, $M(\widetilde{\mathbf{y}}) = var(\widetilde{\mathbf{y}}) + \mathbf{b}\mathbf{b}^\top$, 其中方差：

$$var(\widetilde{\mathbf{y}}) = P_{\mathbb{1}}\sigma^2$$

偏差：

$$\mathbf{b} = E\widetilde{\mathbf{y}} - X\boldsymbol{\beta} = -(I_n - P_{\mathbb{1}})X\boldsymbol{\beta} = -Z_c\boldsymbol{\gamma},$$

其中 $(I_n - P_{\mathbb{1}})X\boldsymbol{\beta} = (I_n - P_{\mathbb{1}})(\mathbb{1}\beta_0 + Z\boldsymbol{\gamma}) = (I_n - P_{\mathbb{1}})Z\boldsymbol{\gamma}) = Z_c\boldsymbol{\gamma})$。所以

$$M(\widetilde{\mathbf{y}}) = P_{\mathbb{1}}\sigma^2 + \boldsymbol{\gamma}^\top Z_c^\top Z_c\boldsymbol{\gamma}$$

$$m(\widetilde{\mathbf{y}}) = tr(M) = \sigma^2 + \|Z_c\boldsymbol{\gamma}\|^2.$$

所以当 $\|Z_c\boldsymbol{\gamma}\|^2 \le (p-1)\sigma^2$ 时，$m(\widetilde{\mathbf{y}}) \le m(\widehat{\mathbf{y}})$.

6.（15 分）。

(a) 证明如果第 k （$1 \le k \le n$）个数据点的自变量 $\mathbf{x}_k = \overline{\mathbf{x}}$, 则 $\hat{y}_k = \overline{y}$。

证：因为 $\mathbb{1} \in C(X)$, 所以 $P_X\mathbb{1} = \mathbb{1}$, 所以

$$\hat{y}_k = \overline{\mathbf{x}}_k^\top\hat{\boldsymbol{\beta}} = (\mathbb{1}^\top X/n)(X^\top X)^{-1}X^\top\mathbf{y} = \mathbb{1}^\top P_X\mathbf{y}/n = \mathbb{1}^\top\mathbf{y}/n = \overline{y}.$$

若 $y_k = \overline{y} = \hat{y}_k$, 则完全数据的 LS 估计 $\hat{\boldsymbol{\beta}}$ 满足正则方程

$$0 = \sum_{i=1}^{n}\mathbf{x}_i(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}) = \sum_{i\ne k}^{n}\mathbf{x}_i(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}})$$

后者说明 $\hat{\boldsymbol{\beta}}$ 也满足删除数据点 k 后的正则方程，所以 $\hat{\boldsymbol{\beta}}_{-k} = \hat{\boldsymbol{\beta}}$, 且 $RSS_{-k} = RSS$。

(b) 现删除数据点 $(y_k, \mathbf{x}_k)$, $1 \le k \le n$, 重新拟合模型得到残差平方和 RSS_{-k}, 证明 $RSS_{-k} \le RSS - e_k^2$。

证明：

$$RSS_{-k} = \min_{\boldsymbol{\beta}}\sum_{i\ne k}(y_i - \mathbf{x}_k^\top\boldsymbol{\beta})^2 \le \sum_{i\ne k}(y_i - \mathbf{x}_k^\top\hat{\boldsymbol{\beta}})^2 = RSS - e_k^2$$

(a),(b) 也可直接用课件中的如下结论（但我们不要求记忆这类复杂公式）：

(1) $\widehat{\boldsymbol{\beta}}_k = \widehat{\boldsymbol{\beta}} - (X^\top X)^{-1}\mathbf{x}_k e_k/(1-h_{kk})$

(2) $RSS_{-k} = RSS - e_k^2/(1-h_{kk})$.