MA 575: HW4

Benjamin Draves

October 17, 2017

Exercise 5.1

(a)

```r
#Read in data
dat = read.table("~/Desktop/Courses/MA 575/book_data/Latour.txt", header = TRUE)

#take a peak
str(dat)
```

```
## 'data.frame':    44 obs. of  4 variables:
##  $ Vintage     : int  1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 ...
##  $ Quality     : num  5 4 1 3 1 4 3 2 2 4 ...
##  $ EndofHarvest: int  28 50 53 38 46 40 35 38 45 47 ...
##  $ Rain        : int  0 0 1 0 1 0 1 1 1 0 ...
```

```r
#build model
interaction_model = lm(Quality ~ EndofHarvest + factor(Rain) + EndofHarvest:factor(Rain), data = dat)
no_interaction_model = lm(Quality ~EndofHarvest + factor(Rain), data = dat)

#look at summary statistics
summary(interaction_model)
```

```
##
## Call:
## lm(formula = Quality ~ EndofHarvest + factor(Rain) + EndofHarvest:factor(Rain),
##     data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6833 -0.5703  0.1265  0.4385  1.6354
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 5.16122    0.68917   7.489 3.95e-09 ***
## EndofHarvest               -0.03145    0.01760  -1.787   0.0816 .
## factor(Rain)1               1.78670    1.31740   1.356   0.1826
## EndofHarvest:factor(Rain)1 -0.08314    0.03160  -2.631   0.0120 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7578 on 40 degrees of freedom
## Multiple R-squared:  0.6848, Adjusted R-squared:  0.6612
## F-statistic: 28.97 on 3 and 40 DF,  p-value: 4.017e-10
```

```r
summary(no_interaction_model)
```

```
##
## Call:
```

```
## lm(formula = Quality ~ EndofHarvest + factor(Rain), data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4563 -0.7366  0.1430  0.6413  1.7652
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.14633    0.61896   9.930  1.8e-12 ***
## EndofHarvest  -0.05723    0.01564  -3.660 0.000713 ***
## factor(Rain)1 -1.62219    0.25478  -6.367  1.3e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8107 on 41 degrees of freedom
## Multiple R-squared:  0.6303, Adjusted R-squared:  0.6123
## F-statistic: 34.95 on 2 and 41 DF,  p-value: 1.383e-09
```

We look to show that the interaction term (EndofHarvest:Rain) is statistical significant. We will use a *parital F test* comparing the two models above to test for the significance of the interaction term. Specifical

$$F = \frac{(RSS(reduced) - RSS(full))\,/(df(reduced) - df(full))}{RSS(full)/df(full)}$$

Here $df(full) = n - (p+1) = 44 - (3+1) = 40$, $df(reduced) = 44 - (2+1) = 41$. Moreover we can find the residual sum of squares (RSS) via the residual standard error. $RSE = \sqrt{RSS/df} \implies RSS = df * RSE^2$. Thus $RSS(full) = 40 * 0.7578^2 = 22.97043$ and $RSS(reduced) = 41 * 0.8107^2 = 26.94661$. Thus F-statistic is given by

$$F = \frac{(26.94661 - 22.9704)\,/(41-40)}{22.9704/40} = 6.925761$$

Under the null hypothesis that the interaction term has no effect on the model (i.e. regression coefficent is zero) this staistic with have degrees of freedom (1, 40). Using this we can find the rejection region given by (R, ∞) where $R = F_{(1,40),\alpha/2}$. Using R we find

```
R = qf(1 - .05, 1, 40)
```

$R = 4.084746$ so our F statistic in in the rejection region and we reject our null hypothsis that the interaction term as no effect on the model.

(b)

Using the full model fitted model, in the case there is no unwanted rain (corresponding to $Rain = 0$) our model reduces to $Quality = \beta_0 + \beta_1 EndofHarvest + e$. Thus decreasing $Quality$ by a full point corresponds $\beta_1 EndofHarvest = -1$. Using our estimate of β_1 we solve for $EndofHarvest = \frac{-1}{-0.03145} = 31.7965$. So we expect to have to wait about 32 days to decrease the quality a full point.

Now if we have unwanted rain, $Rain = 1$ our model is given by $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) EndofHarvest$. Again our problem corresponds to $(\beta_1 + \beta_3) EndofHarvest = -1$. Using our estimates, we find $EndofHarvest = \frac{-1}{-0.11459} = 8.726765$. So we expect to have to wait only 9 days until the quality drops by a full point.

Exercise 5.2

$$Var(\hat{Y}|X) = Var(X(X^TX)^{-1}X^TY|X) = X(X^TX)^{-1}X^T Var(Y|X)\big(X(X^TX)^{-1}X^T\big)^T$$
$$= X(X^TX)^{-1}X^T\sigma^2 I X\big((X^TX)^{-1}\big)^T X^T = \sigma^2 X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T$$

$$= \sigma^2 X(X^TX)^{-1}\left[X^TX(X^TX)^{-1}\right]X^T = \sigma^2 X(X^TX)^{-1}X^T = \sigma^2 H$$

Exercise 5.3

(a)

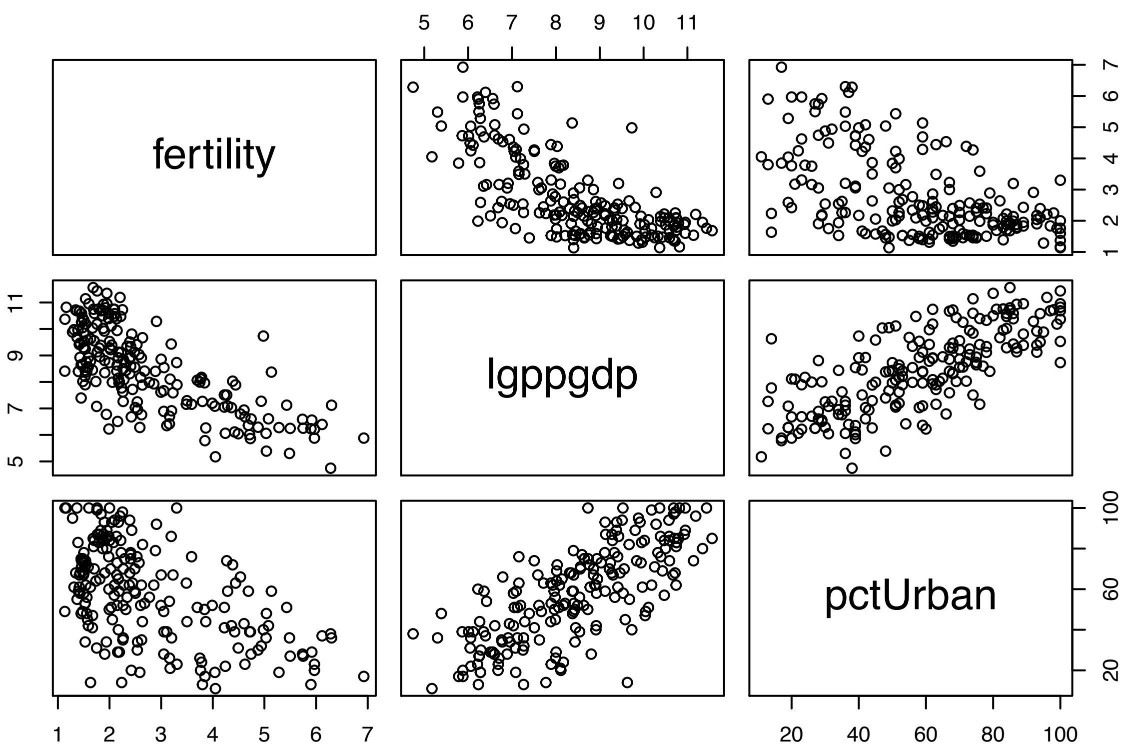
```
#read in data
dat =read.csv("~/Desktop/Courses/MA 575/book_data/UN11.csv")

#take a peak
head(dat)
```

```
##             X     region  group fertility   ppgdp lifeExpF pctUrban
## 1 Afghanistan       Asia  other     5.968   499.0    49.49       23
## 2     Albania     Europe  other     1.525  3677.2    80.40       53
## 3     Algeria     Africa africa     2.142  4473.0    75.00       67
## 4      Angola     Africa africa     5.135  4321.9    53.17       59
## 5    Anguilla  Caribbean  other     2.000 13750.1    81.10      100
## 6   Argentina Latin Amer  other     2.172  9162.1    79.89       93
```

```
#make model data
mdat = data.frame(fertility = dat$fertility,lgppgdp = log(dat$ppgdp), pctUrban = dat$pctUrban)

#take a look at the pairwise scatterplots
pairs(mdat)
```



It appears that log(ppGDP) and percent urban are strongly, postitively correlated. The pctUrban vs fertility plot appears that there is a negative correlation with decreasing variance where the trend looks exponetialy decreasing. The log(ppGPD) and fertility are negatively correlated with a similar exponetial decay trend and nonconstant variance.

(b)

```r
#build log(ppGDP) model
m1 = lm(fertility~lgppgdp, data = mdat)
summary(m1)
```

```
##
## Call:
## lm(formula = fertility ~ lgppgdp, data = mdat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16313 -0.64507 -0.06586  0.62479  3.00517
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.00967    0.36529   21.93   <2e-16 ***
## lgppgdp     -0.62009    0.04245  -14.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9305 on 197 degrees of freedom
## Multiple R-squared:   0.52,  Adjusted R-squared:  0.5175
## F-statistic: 213.4 on 1 and 197 DF,  p-value: < 2.2e-16
```

```r
#build pctUrban model
m2 = lm(fertility~pctUrban, data = mdat)
summary(m2)
```

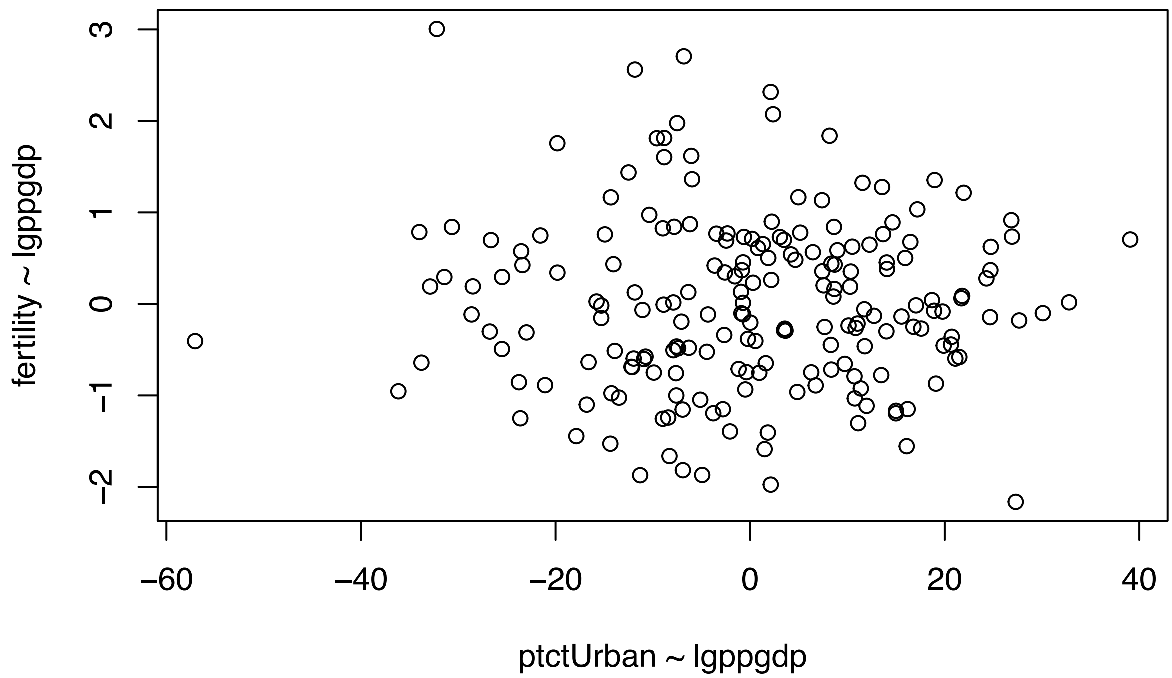
```
##
## Call:
## lm(formula = fertility ~ pctUrban, data = mdat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4932 -0.7795 -0.1475  0.6517  2.9029
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.559823   0.213681  21.339   <2e-16 ***
## pctUrban    -0.031045   0.003421  -9.076   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.128 on 197 degrees of freedom
## Multiple R-squared:  0.2948, Adjusted R-squared:  0.2913
## F-statistic: 82.37 on 1 and 197 DF,  p-value: < 2.2e-16
```

Thus we see that both β_1 coefficients are significantly different than zero.

(c)

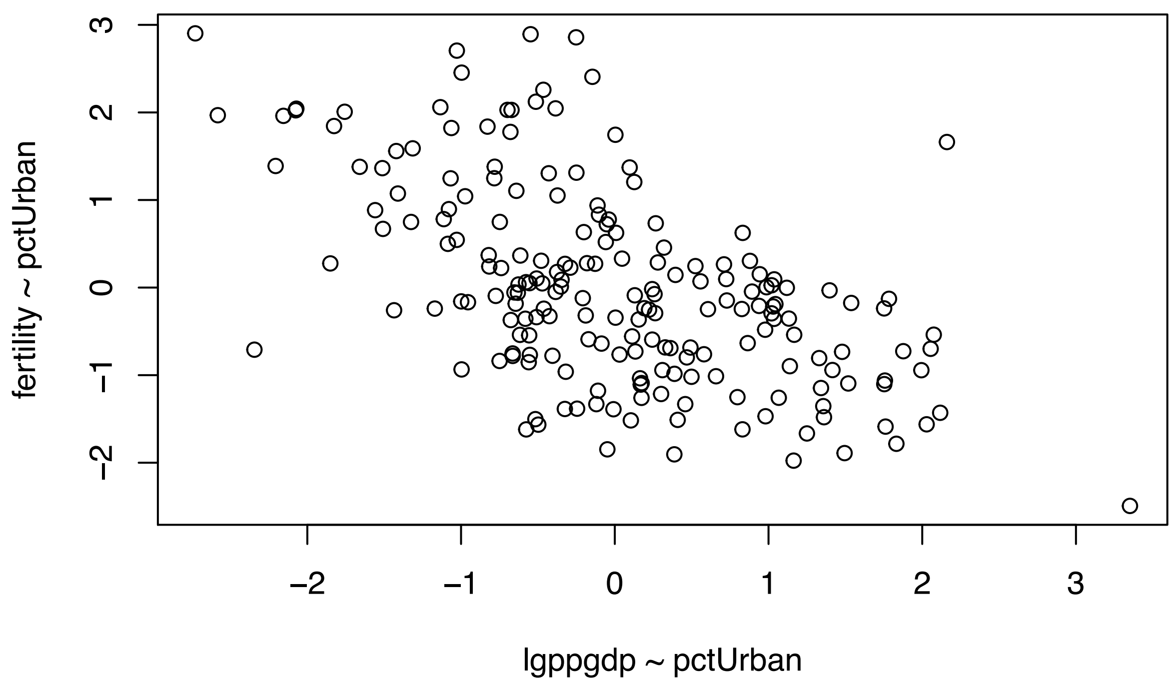
```r
#added variable plot - pctUrban
residY = lm(fertility~lgppgdp, data = mdat)$resid
residX = lm(pctUrban~lgppgdp, data = mdat)$resid
```

```r
plot(residX, residY, xlab = "ptctUrban ~ lgppgdp", ylab = "fertility ~ lgppgdp")
```




```r
#added variable plot - log(ppGDP)
residY = lm(fertility~pctUrban, data = mdat)$resid
residX = lm(lgppgdp~pctUrban, data = mdat)$resid

plot(residX, residY, xlab = "lgppgdp ~ pctUrban", ylab = "fertility ~ pctUrban")
```



There is a clear trend in the residuals when we consider the effect of log(ppGDP) on the response after removing the effect of pctUrban. There is a clear linearly decreasing trend. On the other hand, there is little affect of pctUrban on the response variable after the effect of log(ppGDP) is removed.

```r
full_model = lm(fertility ~ lgppgdp + pctUrban, data = mdat)
summary(full_model)
```

```
##
## Call:
## lm(formula = fertility ~ lgppgdp + pctUrban, data = mdat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15114 -0.64929 -0.06604  0.63253  2.99102
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.9932699  0.3993367  20.016   <2e-16 ***
## lgppgdp     -0.6151425  0.0641565  -9.588   <2e-16 ***
## pctUrban    -0.0004393  0.0042656  -0.103    0.918
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9328 on 196 degrees of freedom
## Multiple R-squared:   0.52,  Adjusted R-squared:  0.5151
## F-statistic: 106.2 on 2 and 196 DF,  p-value: < 2.2e-16
```

There results of the multiple linear regression model confirm our findings in the added variables plot. log(ppGDP) is strongly significant while pctUrban is highly insignificant ($p = .918$).

(d)

In the model above, the estimated coefficent of log(ppGDP) is given by -0.6151425. For the added variable plot, we have

```r
#get added variable plot data
residY = lm(fertility~pctUrban, data = mdat)$resid
residX = lm(lgppgdp~pctUrban, data = mdat)$resid

#build regression model
model = lm(residY~residX)
summary(model)
```

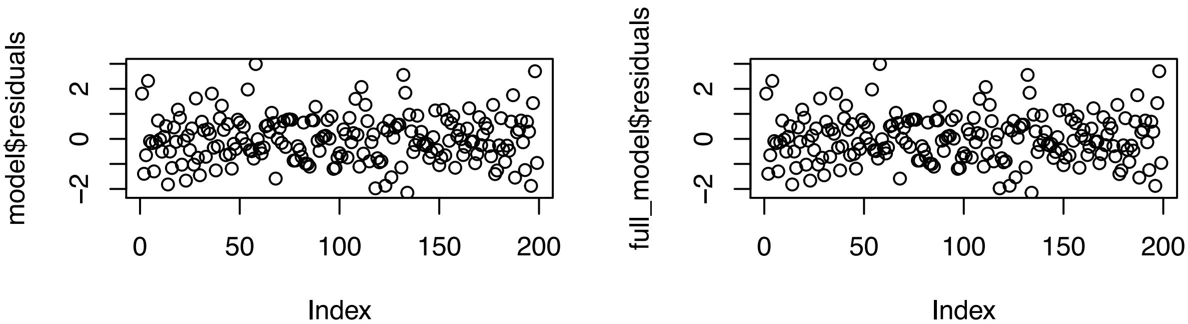
```
##
## Call:
## lm(formula = residY ~ residX)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15114 -0.64929 -0.06604  0.63253  2.99102
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.696e-17  6.596e-02   0.000        1
## residX      -6.151e-01  6.399e-02  -9.613   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9305 on 197 degrees of freedom
```

```
## Multiple R-squared:  0.3193, Adjusted R-squared:  0.3158
## F-statistic:  92.4 on 1 and 197 DF,  p-value: < 2.2e-16
```

Here we see that the estimated coefficent is again $\hat{\beta} = -0.615$. Thus, when we "remove" the effect of the other regressors, we get the MLR coefficent. This implies, that the MLR esimate of β is not independent for each regressor.

(e)

```
par(mfrow = c(2,2))
plot(model$residuals)
plot(full_model$residuals)
```



Thus the residual plots are the same in both cases.

(f)

R gives the t-value for the joint model as $t = -9.588$ and the added variable t statistic is given below

```
#build regression for added variable plot - log(GDP)
residY = lm(fertility~pctUrban, data = mdat)$resid
residX = lm(lgppgdp~pctUrban, data = mdat)$resid

summary(lm(residY~residX))
```

```
##
## Call:
## lm(formula = residY ~ residX)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15114 -0.64929 -0.06604  0.63253  2.99102
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.696e-17  6.596e-02   0.000        1
## residX      -6.151e-01  6.399e-02  -9.613   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9305 on 197 degrees of freedom
## Multiple R-squared:  0.3193, Adjusted R-squared:  0.3158
## F-statistic:  92.4 on 1 and 197 DF,  p-value: < 2.2e-16
```

Here we see that $t = -9.613$. Note that the estimates are the same in both case but the standard error changes. This is due the the degrees of freedom in both model. In the joint model, we have df = n - 2 while in the added variable model we have df = n - 1. Thus this minute change affects the t-statistics in both cases.

Exercise 5.4

Let $X = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{p1} \\ 1 & x_{12} & x_{22} & \dots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{pn} \end{bmatrix}$ Then we have

$$
\begin{aligned}
X^TX &= \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & x_{p3} & \dots & x_{pn} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{p1} \\ 1 & x_{12} & x_{22} & \dots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{pn} \end{bmatrix} \\
&= \begin{bmatrix} n & \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{2i} & \dots & \sum_{i=1}^n x_{pi} \\ \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} & \dots & \sum_{i=1}^n x_{1i}x_{pi} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{pi} & \sum_{i=1}^n x_{pi}x_{1i} & \sum_{i=1}^n x_{pi}x_{2i} & \dots & \sum_{i=1}^n x_{pi}^2 \end{bmatrix}
\end{aligned}
$$

From this, let $A_{11} = [n]$ be the 1×1 matrix, $A_{12} = \begin{bmatrix} \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{2i} & \dots & \sum_{i=1}^n x_{pi} \end{bmatrix}$, and let $A_{22} = (\sum_{i=1}^n x_{ki}x_{ji})_{1\leq k,j\leq p}$.

First note that

$$
\begin{aligned}
\mathcal{X}^T\mathcal{X} &= \begin{bmatrix} (x_{11}-\overline{x}_1) & (x_{21}-\overline{x}_1) & \dots & (x_{n1}-\overline{x}_1) \\ (x_{12}-\overline{x}_2) & (x_{22}-\overline{x}_2) & \dots & (x_{n2}-\overline{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ (x_{1p}-\overline{x}_p) & (x_{2p}-\overline{x}_p) & \dots & (x_{np}-\overline{x}_p) \end{bmatrix} \begin{bmatrix} (x_{11}-\overline{x}_1) & (x_{12}-\overline{x}_2) & \dots & (x_{1p}-\overline{x}_1) \\ (x_{21}-\overline{x}_1) & (x_{22}-\overline{x}_2) & \dots & (x_{2p}-\overline{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ (x_{n1}-\overline{x}_1) & (x_{n2}-\overline{x}_2) & \dots & (x_{np}-\overline{x}_p) \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^n(x_{i1}-\overline{x}_1)^2 & \sum_{i=1}^n(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2) & \dots & \sum_{i=1}^n(x_{i1}-\overline{x}_1)(x_{ip}-\overline{x}_p) \\ \sum_{i=1}^n(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2) & \sum_{i=1}^n(x_{i2}-\overline{x}_2)^2 & \dots & \sum_{i=1}^n(x_{i2}-\overline{x}_2)(x_{ip}-\overline{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n(x_{i1}-\overline{x}_1)(x_{ip}-\overline{x}_p) & \sum_{i=1}^n(x_{i2}-\overline{x}_2)(x_{ip}-\overline{x}_p) & \dots & \sum_{i=1}^n(x_{ip}-\overline{x}_p)^2 \end{bmatrix}
\end{aligned}
$$

Now, for *any* $1 \leq j \leq p$ and $1 \leq k \leq p$

$$
\begin{aligned}
\sum_{i=1}^n (x_{ik}-\overline{x}_k)(x_{ij}-\overline{x}_j) &= \sum_{i=1}^n x_{ik}x_{ij} - \overline{x}_j\sum_{i=1}^n x_{ik} - \overline{x}_k\sum_{i=1}^n x_{ij} + n\overline{x}_j\overline{x}_k \\
&= \sum_{i=1}^n x_{ik}x_{ij} - 2n\overline{x}_j\overline{x}_k + n\overline{x}_j\overline{x}_k \\
&= \sum_{i=1}^n x_{ik}x_{ik} - n\overline{x}_j\overline{x}_k
\end{aligned}
$$

We will now show that our $A_{22} - A_{12}^T A_{11}^{-1} A_{12} = \mathcal{X}^T\mathcal{X}$. First note that $A_{12}^T A_{11}^{-1} A_{12} = \frac{1}{n} A_{12}^T A_{12}$. Moreover, the entries of $A_{12}^T A_{12}$ are given by

$$A_{12}^T A_{12} = \big(\sum_{i=1}^n x_{ki} \sum_{i=1}^n x_{ji}\big)_{1\leq k,j\leq p} = \big(n^2 \overline{x}_k \overline{x}_k\big)_{1\leq j,k\leq p}$$

Therefore we see

$$A_{12}^T A_{11}^{-1} A_{12} = (n\overline{x}_k\overline{x}_j)_{1\leq k,j\leq p}$$

Combining this result with the definition of $A_{22} = (\sum_{i=1}^n x_{ki}x_{ji})_{1\leq k,j\leq p}$. Thus we see

$$A_{22} - A_{12}^T A_{11}^{-1} A_{12} = \Big(\sum_{i=1}^n x_{ki}x_{ji} - n\overline{x}_k\overline{x}_k\Big)_{1\leq k,j\leq p} = \mathcal{X}^T\mathcal{X}$$

Now, notice that $A_{11}^{-1}A_{12} = (\frac{1}{n}\sum_{i=1}^n x_{ik})1 \leq k \leq p = \overline{\mathbf{x}}^T$ and $A_{11}^{-1}A_{12}^T = (\frac{1}{n}\sum_{i=1}^n x_{ik})1 \leq k \leq p = \overline{\mathbf{x}}$. Having shown these relationshops hold we have

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{n} + (\overline{\mathbf{x}}^T)(\mathcal{X}^T\mathcal{X})^{-1}\overline{\mathbf{x}} & -\overline{\mathbf{x}}(\mathcal{X}^T\mathcal{X})^{-1}\overline{\mathbf{x}} \\ -(\mathcal{X}^T\mathcal{X})^{-1}\overline{\mathbf{x}} & (\mathcal{X}^T\mathcal{X})^{-1} \end{bmatrix}$$