# ST5202: Applied Regression Analysis

Department of Statistics and Applied Probability
National University of Singapore

19-Feb-2018
Lecture 6

# Announcement

- Assignment #3 released. Due by 5th of March.
- Midterm will cover from lecture 1 to lecture 6.
- Midterm scheduled on 12th of March.
- Last day to make a request for a make-up midterm is 26th of February (official document needed).

# Lecture 6

## Mid review &
## Multiple regression II (Chapter 7)

# Outline

- Mid review
- Multiple regression II
  - Model diagnostics
  - (Partial) F-tests, extra sum of squares (coefficients of partial determination)

## Testing $\beta_1 = 0$ in SLR: three approaches

- Approach 1:
  t-test: sampling distribution approach
- Approach 2:
  F-test: Analysis of Variance (ANOVA) approach
- Approach 3:
  General linear test approach

## Test $\beta$'s in SLR using $t$ test

- Test $\beta_0 = 0$ or $\beta_1 = 0$ can be derived from respectively
$$\frac{b_0 - \beta_0}{s\{b_0\}} \sim t(n-2), \quad \frac{b_1 - \beta_1}{s\{b_1\}} \sim t(n-2)$$
- Under each $H_0$, we have $\frac{b_0}{s\{b_0\}} \sim t(n-2)$, and $\frac{b_1}{s\{b_1\}} \sim t(n-2)$ respectively. Here, both $s\{b_0\}$ and $s\{b_1\}$ should be calculated from sample.
- Test statistics and the decision rules are as follows:
$$\text{For } T^* = \frac{b_0}{s\{b_0\}}, \text{ or } T^* = \frac{b_1}{s\{b_1\}}$$
$$\begin{aligned} |T^*| &\leq t(1-\alpha/2; n-2), \text{ accept } H_0 \\ |T^*| &> t(1-\alpha/2; n-2), \text{ reject } H_0 \end{aligned}$$
- Absolute values make the test two-sided

# Test $\beta$'s in SLR using $t$ test–R code

```
> # fit a linear model with the lm command:
> mod = lm(GPA ~ ACT)
> summary(mod)
Call:
lm(formula = GPA ~ ACT)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.11405 0.32089   6.588 1.30e-09 ***
ACT            0.03883 0.01277   3.040 0.00292 **
---

Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262,   Adjusted R-squared:  0.06476
F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

# ANOVA table

- We collect the above ANOVA analysis as a table as follows

| Source | SS | df | MS | F | p-value(s) |
|---|---|---|---|---|---|
| Regression | SSR | 1 | MSR | $F^*$ | $P(F(1, n-2) \geq f^*)$ |
| Error | SSE | n-2 | MSE | | |
| Total | SSTO | n-1 | | | |

  where $f^*$ denotes the computed value of $F^*$ from the sample

- One of the important role of the above ANOVA table is to test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$

## Analysis of Variance (ANOVA) approach

- Total sum of squares (SSTO): $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
  - independent of $X_i$: lose 1 df, so that df is $n-1$
- Error sum of squares (SSE):

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$$

  - lose 2 df, so that the df is $n-2$
- Regression sums of squares (SSR):

$$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = b_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$$

  - df is 1
- SSTO = SSR + SSE

## Coefficient of Determinant

- SSTO: a measure of uncertainty of $Y$ when $X$ is not taken into account
- SSE: a measure of uncertainty of $Y$ when $X$ is taken into account
- coefficient of determination $R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$: reduction of uncertainty of $Y$ due to considering $X$
- $0 \leq R^2 \leq 1$

# Analysis of Variance (ANOVA) approach–R code

```
> anova(mod)
Analysis of Variance Table

Response: GPA
           Df Sum Sq Mean Sq F value   Pr(>F)
ACT         1  3.588  3.5878  9.2402 0.002917 **
Residuals 118 45.818  0.3883
```

| Source | SS | df | MS | F | p-value(s) |
|---|---|---|---|---|---|
| Regression | SSR | 2-1 | MSR | $F^* = \frac{MSR}{MSE}$ | $P(F(1, n-2) \geq 9.2402)$ |
| Error | SSE | $120-2$ | MSE | | |
| Total | SSTO | $120-1$ | | | |

## t test is equivalent to $F$ test

- $t(m)^2$, and $F(1, m)$ have the same distribution.
- $t^* = \frac{b_1}{s\{b_1\}} = \frac{b_1}{\sqrt{MSE}/\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$
- $t^{*2} = f^*$, the computed value of $F^*$
- $F^*$ is the ratio MSR/MSE in anova (Cochran's theorem)

## t test is equivalent to $F$ test–R code

```
mod <- lm(GPA~ACT, data=gpa)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
ACT          0.03883    0.01277   3.040  0.00292 **
---
Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262,   Adjusted R-squared:  0.06476
F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917

Analysis of Variance Table
Response: GPA
           Df Sum Sq Mean Sq F value   Pr(>F)
ACT         1  3.588  3.5878  9.2402 0.002917 **
Residuals 118 45.818  0.3883
```

## General linear test approach

- Full model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, $i = 1, \cdots n$
  - SSE(F): $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$
- Reduced model (under $H_0 : \beta_1 = 0$): $Y_i = \beta_0 + \epsilon_i$, $i = 1, \cdots, n$
  - SSE(R): $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- SSE(R) $\geq$ SSE(F)
- The idea: if the full model is better than reduced model, then $\frac{SSE(R) - SSE(F)}{SSE(F)}$ tends to be *significantly* large $\rightarrow$ another $F$ test

# General linear test approach–R code

```
> mod1 <- lm(GPA~ACT, data=gpa)
> mod2 <- lm(GPA~-ACT, data=gpa)
> anova(mod1)
> anova(mod2)

Analysis of Variance Table

Response: GPA
           Df Sum Sq Mean Sq F value   Pr(>F)
ACT         1  3.588  3.5878  9.2402 0.002917 **
Residuals 118 45.818  0.3883

Analysis of Variance Table

Response: GPA
           Df Sum Sq Mean Sq F value Pr(>F)
Residuals 119 49.405 0.41517
```

# General linear test approach–R code

```
> anova(mod2,mod1)
Analysis of Variance Table

Model 1: GPA ~ -ACT
Model 2: GPA ~ ACT
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    119 49.405
2    118 45.818  1    3.5878 9.2402 0.002917 **
---
```

- $F^* = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F} = \frac{\frac{49.818 - 49.405}{119-118}}{\frac{45.818}{118}} = 9.24$
- **Note:** for the case of SLR, when testing $\beta_1 = 0$, it happens that $SSE(R) = SSTO$. $F^* = \frac{(SSTO - SSE)/(df_R - df_F)}{SSE/df_F} = \frac{MSR}{MSE} = F^*$ identical to $F$ statistic in 'original' anova

ANOVA table for multiple linear model

| Source | SS | df | MS | F | p-value(s) |
|---|---|---|---|---|---|
| Regression | SSR= $\mathbf{Y}'\left(\mathbf{H}-\frac{1}{n}\mathbf{J}\right)\mathbf{Y}$ | p-1 | MSR= $\frac{SSR}{p-1}$ | $F^* = \frac{MSR}{MSE}$ | $P(F(p-1, n-p) \geq f^*)$ |
| Error | SSE= $\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}$ | $n-p$ | MSE= $\frac{SSE}{n-p}$ | | |
| Total | SSTO= $\mathbf{Y}'\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\mathbf{Y}$ | $n-1$ | | | |

where $f^*$ is computed value of $F^*$ from the sample.

- One of the important role of the above ANOVA table is to test $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$ versus $H_a$ : at least one $\beta_k \neq 0$ $(k = 1, \cdots, p-1)$

# Multivariate data example

| | Population | Income | Illiteracy | Life.Exp | Murder | HS.Grad | Frost | Area | Density |
|---|---|---|---|---|---|---|---|---|---|
| Alabama | 3615 | 3624 | 2.1 | 69.05 | 15.1 | 41.3 | 20 | 50708 | 71.2905261 |
| Alaska | 365 | 6315 | 1.5 | 69.31 | 11.3 | 66.7 | 152 | 566432 | 0.6443845 |
| Arizona | 2212 | 4530 | 1.8 | 70.55 | 7.8 | 58.1 | 15 | 113417 | 19.5032491 |
| Arkansas | 2110 | 3378 | 1.9 | 70.66 | 10.1 | 39.9 | 65 | 51945 | 40.6198864 |
| California | 21198 | 5114 | 1.1 | 71.71 | 10.3 | 62.6 | 20 | 156361 | 135.5708904 |
| Colorado | 2541 | 4884 | 0.7 | 72.06 | 6.8 | 63.9 | 166 | 103766 | 24.4877898 |
| Connecticut | 3100 | 5348 | 1.1 | 72.48 | 3.1 | 56.0 | 139 | 4862 | 637.5976964 |
| Delaware | 579 | 4809 | 0.9 | 70.06 | 6.2 | 54.6 | 103 | 1982 | 292.1291625 |
| Florida | 8277 | 4815 | 1.3 | 70.66 | 10.7 | 52.6 | 11 | 54090 | 153.0227399 |
| Georgia | 4931 | 4091 | 2.0 | 68.54 | 13.9 | 40.6 | 60 | 58073 | 84.9103714 |
| Hawaii | 868 | 4963 | 1.9 | 73.60 | 6.2 | 61.9 | 0 | 6425 | 135.0972763 |
| Idaho | 813 | 4119 | 0.6 | 71.87 | 5.3 | 59.5 | 126 | 82677 | 9.8334482 |
| Illinois | 11197 | 5107 | 0.9 | 70.14 | 10.3 | 52.6 | 127 | 55748 | 200.8502547 |
| Indiana | 5313 | 4458 | 0.7 | 70.88 | 7.1 | 52.9 | 122 | 36097 | 147.1867468 |
| Iowa | 2861 | 4628 | 0.5 | 72.56 | 2.3 | 59.0 | 140 | 55941 | 51.1431687 |
| Kansas | 2280 | 4669 | 0.6 | 72.58 | 4.5 | 59.9 | 114 | 81787 | 27.8772910 |
| Kentucky | 3387 | 3712 | 1.6 | 70.10 | 10.6 | 38.5 | 95 | 39650 | 85.4224464 |
| Louisiana | 3806 | 3545 | 2.8 | 68.76 | 13.2 | 42.2 | 12 | 44930 | 84.7095482 |
| Maine | 1058 | 3694 | 0.7 | 70.39 | 2.7 | 54.7 | 161 | 30920 | 34.2173351 |
| Maryland | 4122 | 5299 | 0.9 | 70.22 | 8.5 | 52.3 | 101 | 9891 | 416.7424932 |
| Massachusetts | 5814 | 4755 | 1.1 | 71.83 | 3.3 | 58.5 | 103 | 7826 | 742.9082545 |
| Michigan | 9111 | 4751 | 0.9 | 70.63 | 11.1 | 52.8 | 125 | 56817 | 160.3569354 |
| Minnesota | 3921 | 4675 | 0.6 | 72.96 | 2.3 | 57.6 | 160 | 79289 | 49.4520047 |
| Mississippi | 2341 | 3098 | 2.4 | 68.09 | 12.5 | 41.0 | 50 | 47296 | 49.4967862 |
| Missouri | 4767 | 4254 | 0.8 | 70.69 | 9.3 | 48.8 | 108 | 68995 | 69.0919632 |
| Montana | 746 | 4347 | 0.6 | 70.56 | 5.0 | 59.2 | 155 | 145587 | 5.1240839 |
| Nebraska | 1544 | 4508 | 0.6 | 72.60 | 2.9 | 59.3 | 139 | 76483 | 20.1874926 |
| Nevada | 590 | 5149 | 0.5 | 69.03 | 11.5 | 65.2 | 188 | 109889 | 5.3690542 |
| New Hampshire | 812 | 4281 | 0.7 | 71.23 | 3.3 | 57.6 | 174 | 9027 | 89.9523651 |
| New Jersey | 7333 | 5237 | 1.1 | 70.93 | 5.2 | 52.5 | 115 | 7521 | 975.0033240 |
| New Mexico | 1144 | 3601 | 2.2 | 70.32 | 9.7 | 55.2 | 120 | 121412 | 9.4224624 |
| New York | 18076 | 4903 | 1.4 | 70.55 | 10.9 | 52.7 | 82 | 47831 | 377.9139052 |
| North Carolina | 5441 | 3875 | 1.8 | 69.21 | 11.1 | 38.5 | 80 | 48798 | 111.5004713 |
| North Dakota | 637 | 5087 | 0.8 | 72.78 | 1.4 | 50.3 | 186 | 69273 | 9.1955019 |
| Ohio | 10735 | 4561 | 0.8 | 70.82 | 7.4 | 53.2 | 124 | 40975 | 261.9890177 |
| Oklahoma | 2715 | 3983 | 1.1 | 71.42 | 6.4 | 51.6 | 82 | 68782 | 39.4725364 |
| Oregon | 2284 | 4660 | 0.6 | 72.13 | 4.2 | 60.0 | 44 | 96184 | 23.7461532 |
| Pennsylvania | 11860 | 4449 | 1.0 | 70.43 | 6.1 | 50.2 | 126 | 44966 | 263.7548370 |
| Rhode Island | 931 | 4558 | 1.3 | 71.90 | 2.4 | 46.4 | 127 | 1049 | 887.5119161 |
| South Carolina | 2816 | 3635 | 2.3 | 67.96 | 11.6 | 37.8 | 65 | 30225 | 93.1679074 |
| South Dakota | 681 | 4167 | 0.5 | 72.08 | 1.7 | 53.3 | 172 | 75955 | 8.9658350 |
| Tennessee | 4173 | 3821 | 1.7 | 70.11 | 11.0 | 41.8 | 70 | 41328 | 100.9727062 |
| Texas | 12237 | 4188 | 2.2 | 70.90 | 12.2 | 47.4 | 35 | 262134 | 46.6822312 |
| Utah | 1203 | 4022 | 0.6 | 72.90 | 4.5 | 67.3 | 137 | 82096 | 14.6535763 |
| Vermont | 472 | 3907 | 0.6 | 71.64 | 5.5 | 57.1 | 168 | 9267 | 50.9334197 |
| Virginia | 4981 | 4701 | 1.4 | 70.08 | 9.5 | 47.8 | 85 | 39780 | 125.2136752 |
| Washington | 3559 | 4864 | 0.6 | 71.72 | 4.3 | 63.5 | 32 | 66570 | 53.4625207 |
| West Virginia | 1799 | 3617 | 1.4 | 69.48 | 6.7 | 41.6 | 100 | 24070 | 74.7403407 |
| Wisconsin | 4589 | 4468 | 0.7 | 72.48 | 3.0 | 54.5 | 149 | 54464 | 84.2574912 |
| Wyoming | 376 | 4566 | 0.6 | 70.29 | 6.9 | 62.9 | 173 | 97203 | 3.8681934 |

- Life expectancy does have a bivariate relationship with a lot of the other variables
- But many of those variables are also related to each other
- Multiple regression allows us to tear all of this apart and investigate the relationship in a “purer” (but not exactly pure) form.

## R code

```
st[,9] = st$Population*1000/st$Area # add a variable
colnames(st)[9] = "Density" #create and name a new column

> names(st)                        # for handy reference
[1] "Population" "Income"   "Illiteracy" "Life.Exp"   "Murder"
[6] "HS.Grad"    "Frost"    "Area"       "Density"
```

# Begin by throwing all the predictors into a linear model

```
> model1 = lm(Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad + Frost + Area
+ Density, data=st)
> summary(model1)

Call:
lm(formula = Life.Exp ~ Population + Income + Illiteracy + Murder +
    HS.Grad + Frost + Area + Density, data = st)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.995e+01  1.843e+00  37.956  < 2e-16
Population   6.480e-05  3.001e-05   2.159   0.0367
Income       2.701e-04  3.087e-04   0.875   0.3867
Illiteracy   3.029e-01  4.024e-01   0.753   0.4559
Murder      -3.286e-01  4.941e-02  -6.652 5.12e-08
HS.Grad      4.291e-02  2.332e-02   1.840   0.0730
Frost       -4.580e-03  3.189e-03  -1.436   0.1585
Area        -1.558e-06  1.914e-06  -0.814   0.4205
Density     -1.105e-03  7.312e-04  -1.511   0.1385

Residual standard error: 0.7337 on 41 degrees of freedom
Multiple R-squared: 0.7501,     Adjusted R-squared: 0.7013
F-statistic: 15.38 on 8 and 41 DF,  p-value: 3.787e-10
```

- Higher populations are related to increased life expectancy and higher murder rates are strongly related to decreased life expectancy; not seeing too much beyond that.

# ANOVA at a glance but not too helpful

```
> summary.aov(model1)
             Df  Sum Sq Mean Sq F value    Pr(>F)
Population    1  0.4089  0.4089  0.7597  0.388493
Income        1 11.5946 11.5946 21.5413 3.528e-05
Illiteracy    1 19.4207 19.4207 36.0811 4.232e-07
Murder        1 27.4288 27.4288 50.9591 1.051e-08
HS.Grad       1  4.0989  4.0989  7.6152  0.008612
Frost         1  2.0488  2.0488  3.8063  0.057916
Area          1  0.0011  0.0011  0.0020  0.964381
Density       1  1.2288  1.2288  2.2830  0.138472
Residuals    41 22.0683  0.5383
```

- This is a bit different from the ANOVA table we had (will explain later)
- Now we need to start winnowing down our model to a minimal adequate one. The least significant slope is that for "Area", so let's toss out "Area" first.

## Comparing two models

```
> model2 = update(model1, .~.-Area)
> summary(model2)

Call:
lm(formula = Life.Exp ~ Population + Income + Illiteracy + Murder +
    HS.Grad + Frost + Density, data = st)


Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.094e+01  1.378e+00  51.488  < 2e-16
Population   6.249e-05  2.976e-05   2.100   0.0418
Income       1.485e-04  2.690e-04   0.552   0.5840
Illiteracy   1.452e-01  3.512e-01   0.413   0.6814
Murder      -3.319e-01  4.904e-02  -6.768 3.12e-08
HS.Grad      3.746e-02  2.225e-02   1.684   0.0996
Frost       -5.533e-03  2.955e-03  -1.873   0.0681
Density     -7.995e-04  6.251e-04  -1.279   0.2079

Residual standard error: 0.7307 on 42 degrees of freedom
Multiple R-squared: 0.746,      Adjusted R-squared: 0.7037
F-statistic: 17.63 on 7 and 42 DF,  p-value: 1.173e-10
```

- reduce our model to a point where all the remaining predictors are significant, and we want to do this by throwing out one predictor at a time. "Area" goes out first.

# Comparing two models–continued

```
> anova(model1, model2)
Analysis of Variance Table

Model 1: Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
    Frost + Density
Model 2: Life.Exp ~ Population + Income + Illiteracy + Murder + +HS.Grad +
    Frost + Area + Density
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     42 22.425
2     41 22.068  1   0.35639 0.6621 0.4205
```

- We have seen this comparison before. Where?
- Removing "Area" had no significant effect on the model ($p = 0.4205$). Compare the p-value to that for "Area" in the first summary table above.
- Does the order in anova(model1, model2) matter here?
- Notice that removing "Area" has cost us very little in terms of R-squared, and the adjusted R-squared actually went up, due to there being fewer predictors.

## What goes out next? Illiteracy

```
> model3 = update(model2, .~.-Illiteracy)
> summary(model3)

Call:
lm(formula = Life.Exp ~ Population + Income + Murder + HS.Grad +
    Frost + Density, data = st)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.131e+01  1.042e+00  68.420  < 2e-16
Population   5.811e-05  2.753e-05   2.110   0.0407
Income       1.324e-04  2.636e-04   0.502   0.6181
Murder      -3.208e-01  4.054e-02  -7.912 6.32e-10
HS.Grad      3.499e-02  2.122e-02   1.649   0.1065
Frost       -6.191e-03  2.465e-03  -2.512   0.0158
Density     -7.324e-04  5.978e-04  -1.225   0.2272

Residual standard error: 0.7236 on 43 degrees of freedom
Multiple R-squared: 0.745,     Adjusted R-squared: 0.7094
F-statistic: 20.94 on 6 and 43 DF,  p-value: 2.632e-11
```

- Things are starting to change a bit. R-squared went down again, as it will always do when a predictor is removed, but once more adjusted R-squared increased.
- Now “Frost” becomes a significant predictor of life expectancy.

# What goes out next? Income

```
> model4 = update(model3, .~.-Income)
> summary(model4)

Call:
lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost +
    Density, data = st)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.142e+01  1.011e+00  70.665  < 2e-16
Population   6.083e-05  2.676e-05   2.273  0.02796
Murder      -3.160e-01  3.910e-02  -8.083 3.07e-10
HS.Grad      4.233e-02  1.525e-02   2.776  0.00805
Frost       -5.999e-03  2.414e-03  -2.485  0.01682
Density     -5.864e-04  5.178e-04  -1.132  0.26360

Residual standard error: 0.7174 on 44 degrees of freedom
Multiple R-squared: 0.7435,     Adjusted R-squared: 0.7144
F-statistic: 25.51 on 5 and 44 DF,  p-value: 5.524e-12
```

- R-squared went down hardly at all. Adjusted R-squared went up. “Income” will be kicked out.

# Now all the predictors are significant expect “Density”

```
> model5 = update(model4, .~.-Density)
> summary(model5)

Call:
lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,
    data = st)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47095 -0.53464 -0.03701  0.57621  1.50683

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16
Population   5.014e-05  2.512e-05   1.996  0.05201
Murder      -3.001e-01  3.661e-02  -8.199 1.77e-10
HS.Grad      4.658e-02  1.483e-02   3.142  0.00297
Frost       -5.943e-03  2.421e-03  -2.455  0.01802

Residual standard error: 0.7197 on 45 degrees of freedom
Multiple R-squared: 0.736,      Adjusted R-squared: 0.7126
F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

- Adjusted R-squared slipped a bit this time, but not significantly. How can we tell it?

# Dropping “Density”

```
> anova(model5, model4)
Analysis of Variance Table

Model 1: Life.Exp ~ Population + Murder + HS.Grad + Frost
Model 2: Life.Exp ~ Population + Murder + HS.Grad + Frost + Density
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     45 23.308
2     44 22.648  1   0.66005 1.2823 0.2636
```

- So, letting “Density” out is fine
- So far, we have (implicitly by not saying otherwise) set our alpha level at 0.05, so now population must be out.
- This could have a substantial effect on the model, as the slope for “Population” is very nearly significant.

# One way to find out

```
> model6 = update(model5, .~.-Population)
> summary(model6)

Call:
lm(formula = Life.Exp ~ Murder + HS.Grad + Frost, data = st)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.036379   0.983262  72.246  < 2e-16
Murder      -0.283065   0.036731  -7.706 8.04e-10
HS.Grad      0.049949   0.015201   3.286  0.00195
Frost       -0.006912   0.002447  -2.824  0.00699

Residual standard error: 0.7427 on 46 degrees of freedom
Multiple R-squared: 0.7127,     Adjusted R-squared: 0.6939
F-statistic: 38.03 on 3 and 46 DF,  p-value: 1.634e-12
```

- We have reached the one of the minimal adequate models (may not be unique).

# Multiple regression II

## Model diagnostics and other issues in multiple linear regressions

## Portrait studio example–R code

```
> mod = lm(Y ~ X1 + X2)
> summary(mod)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.8571    60.0170  -1.147   0.2663
X1            1.4546     0.2118   6.868    2e-06 ***
X2            9.3655     4.0640   2.305   0.0333 *
---

Residual standard error: 11.01 on 18 degrees of freedom
Multiple R-squared: 0.9167, Adjusted R-squared: 0.9075
F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10
```

- $Y$: sales in a community
  $X_1$: the number of persons aged 16 or younger in the community
  $X_2$: per capita personal income in the community

# Portrait studio example–R code

Getting the 95% confidence interval for the mean at $X_1 = 65.4, X_2 = 17.6$ (with prediction interval)

```
> xh<- data.frame(cbind(X1 = 65.4, X2 = 17.6))
> predict(mod, xh, interval="confidence", level=0.95)
       fit      lwr      upr
191.1039 185.2911 196.9168

> predict(mod, xh, interval="predict", level=0.95)
       fit      lwr     upr
 191.1039 167.2589 214.949
```

- $Y$: sales in a community
  $X_1$: the number of persons aged 16 or younger in the community
  $X_2$: per capita personal income in the community

# Portrait studio example–R code

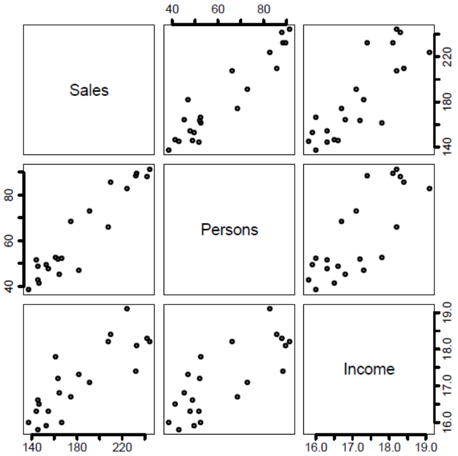
Getting the $\boxed{90\%}$ prediction intervals (Bonferroni) at both $X_1 = 65.4, X_2 = 17.6$ and $X_1 = 53.1, X_2 = 17.7$

```
xh1 <- data.frame(cbind(X1 = 65.4, X2 = 17.6))
xh2 <- data.frame(cbind(X1 = 53.1, X2 = 17.7))

xh <-rbind(xh1,xh2)
< xh
    X1   X2
1 65.4 17.6
2 53.1 17.7

> predict(mod, xh, interval="predict", level=0.95)
       fit      lwr      upr
1 191.1039 167.2589 214.9490
2 174.1494 149.0867 199.2121
```

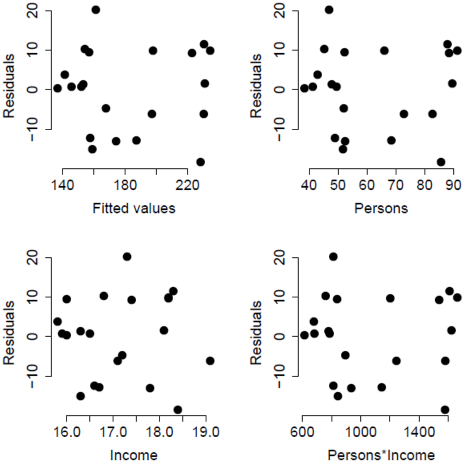
## Diagnostics and remedial measures

- Scatter plot matrix
- Residual plots:
  just as before
  - Against time or some other sequence to check error dependency
  - Against each $X$ variable for potential nonlinear relationship and nonconstancy of error variances
  - Against omitted variables (including the interaction terms). Interaction terms will be discussed in more detail in Ch. 8
- Correlation Test for Normality (same as simple linear regression)
- Brown-Forsythe Test, and Breush-Pagan test for constancy of error variance
- F test for lack of fit
  (need to have a "replicate" observation matching all $X_{i1}, \cdots, X_{i,p-1}$)
- Box-Cox transformations (same as in simple linear regression)

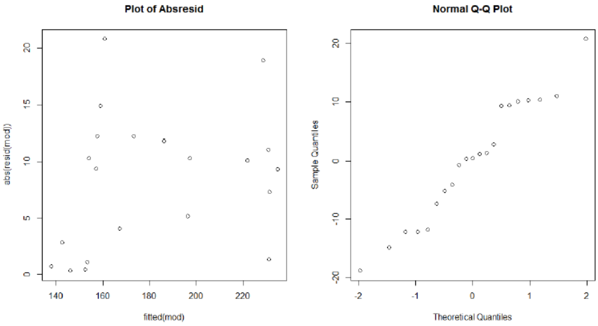
# Portrait studio example–scatter plot



# Portrait studio example–residual plot plot



# Portrait studio example–residual plot plot



## Diagnostic tests

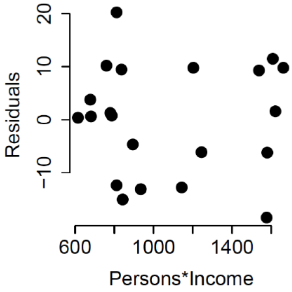
- Constant variance
  - Brown-Forsythe test, and Breusch-Pagan test
- F-test of lack of fit:
  - "Compare local means to prediction with linear model at different X-levels"
    Note: here we need repeated $Y_{ij}$'s at a combination of predictors $(X_{j1}, \cdots, X_{j,p-1})$
  - Test statistic:

$$\begin{aligned} F^* &= \frac{SSE(R) - SSE(F)}{df_R - df_F} \cdot \frac{df_F}{SSE(F)} \\ &= \frac{SSE - SSPE}{c - p} \cdot \frac{n - c}{SSPE} \\ &= \frac{MSLF}{MSPE} \sim F(c - p, n - c) \text{ under } H_0 \end{aligned}$$

## Portrait studio example–residual against the interaction

- Model: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 * X_2$
- Interaction term: $\beta_3 X_1 * X_2$
  Due to this term, the effect of $X_1$ varies depending on the level of $X_2$ (vice verso also holds)

No systematic pattern $\rightarrow$ NO interaction effects reflected by the model term $\beta_3 X_1 * X_2$ appear to be present



# Portrait studio example: checking if the interaction term is necessary ($H_0 : \beta_3 = 0$)

- Reduced model: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- Full model: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 * X_2$

```
mod1 <- lm(Y ~ X1+X2)
mod2 <- lm(Y ~ X1 + X2 + X1:X2)
anova(mod1,mod2)

Analysis of Variance Table

Model 1: Y ~ X1 + X2
Model 2: Y ~ X1 + X2 + X1:X2
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     18 2180.9
2     17 2172.5  1    8.4336 0.066 0.8003
```

Small F values mean that we do not have have enough evidence to reject the $H_0 : \beta_3 = 0$.

## Extra sum of squares and partial F-tests

- Testing whether subsets of the regression coefficients are equal to zero in multiple regression model with
  $E\{Y\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$
- Example of such a test:
  $H_0 : E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
  $H_a : E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$
  with additional predictor variables $X_3$, $X_4$, and $X_5$.
  It can be $X_3 = X_1 X_2$, $X_4 = X_1^2$, $X_5 = X_2^2$
- You can test this with a general linear test approach, with

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \cdot \frac{df_F}{SSE(F)}$$

- The $F-$test is called a partial $F-$test and the difference SSE(R)-SSE(F) is called an extra sum of squares

## Extra sum of squares, a bit more notation

- For model with $p-1$ predictor variables:

  $$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \epsilon$$

  write ANOVA as

  $$SSTO = SSR(X_1, X_2, \cdots, X_{p-1}) + SSE(X_1, X_2, \cdots, X_{p-1})$$

  to make clear which model the SSR and SSE are referring to (i.e., which variables are included in the model)
- Note that
  - SSTO does not depend on which predictor variables were included in the model!
  - SSE can never increase if more predictor variables are added to the model; e.g., $SSE(X_1, X_2) \leq SSE(X_2)$

## Extra sum of squares, a bit more notation–continued

- SSTO = SSE + SSR for each model.
  Therefore, SSR can never decrease if more predictor variables are added to the model:
  e.g., $SSR(X_1, X_2) \geq SSR(X_1)$
- We would like to decompose SSR to measure marginal reduction in error sum of squares when an extra variable is added to the model:
  e.g., $SSR(X_2|X_1)$

## Extra sum of squares, a bit more notation–continued

- For a model with two predictor variables, the *extra sum of squares* when adding $X_2$ to the model with $X_1$ in it, is defined as:

$$\begin{aligned} SSR(X_2|X_1) &= SSE(X_1) - SSE(X_1, X_2), \\ &= SSR(X_1, X_2) - SSR(X_1) \end{aligned}$$

  which is the increase (reduction) in the regression (error) sum of squares when adding $X_2$ to the model when $X_1$ is already included
    - Is $SSR(X_2|X_1) = SSR(X_1|X_2)$?
- The degrees of freedom of an extra sum of squares is
    - the difference in the degrees of freedom of its SSE's (or similarly of its SSR's)
    - the number of predictors that is added to the model

## Decomposing SSR in ANOVA

- General definition for two sets $S$ and $R$ of predictor variables:

$$SSR(X_S|X_R) = SSR(X_S, X_R) - SSR(X_R)$$

- E.g., for $S = \{2, 3\}$ and $R = \{1\}$,
  $SSR(X_2, X_3|X_1) = SSR(X_1, X_2, X_3) - SSR(X_1)$,
  for $S = \{3\}$ and $R = \{1, 2\}$,
  $SSR(X_3|X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2)$
- It follows from the definition of the extra sum of squares that (verify it!)

$$\begin{aligned} & SSR(X_1, X_2, \cdots, X_{p-1}) \\ & = SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2) + \cdots + SSR(X_{p-1}|X_1, \cdots, X_{p-2}) \end{aligned}$$

- This can be used to decompose SSR in ANOVA table

## ANOVA for portrait studio data (Ch. 6.9)

```
> anova(mod)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq  F value   Pr(>F)
X1         1 23371.8 23371.8 192.8962 4.64e-11 ***
X2         1   643.5   643.5   5.3108  0.03332 *
Residuals 18  2180.9   121.2
---
Signif. codes:  0 Ã***Ã 0.001 Ã**Ã 0.01 Ã*Ã 0.05 Ã.Ã 0.1 Ã
```

- "anova" command in R for multiple regression model gives the break-down of SSR in $SSR(X_1), SSR(X_2|X_1)$, and so on.

# Body fat example

- body fat percentage ($Y$)
- triceps skin fold thickness ($X_1$)
- thigh circumference ($X_2$)
- midarm circumference ($X_3$)

| Subject $i$ | Triceps Skinfold Thickness $X_{i1}$ | Thigh Circumference $X_{i2}$ | Midarm Circumference $X_{i3}$ | Body Fat $Y_i$ |
|---|---|---|---|---|
| 1 | 19.5 | 43.1 | 29.1 | 11.9 |
| 2 | 24.7 | 49.8 | 28.2 | 22.8 |
| 3 | 30.7 | 51.9 | 37.0 | 18.7 |
| ... | ... | ... | ... | ... |
| 18 | 30.2 | 58.6 | 24.6 | 25.4 |
| 19 | 22.7 | 48.2 | 27.1 | 14.8 |
| 20 | 25.2 | 51.0 | 27.5 | 21.1 |

# Body fat example

- Regression on $X_1$
- Regression on $X_2$

(a) Regression of $Y$ on $X_1$

$\hat{Y} = -1.496 + .8572X_1$

| Source of Variation | $SS$ | $df$ | $MS$ |
|---|---|---|---|
| Regression | 352.27 | 1 | 352.27 |
| Error | 143.12 | 18 | 7.95 |
| Total | 495.39 | 19 | |

| Variable | Estimated Regression Coefficient | Estimated Standard Deviation | $t^*$ |
|---|---|---|---|
| $X_1$ | $b_1 = .8572$ | $s\{b_1\} = .1288$ | 6.66 |

(b) Regression of $Y$ on $X_2$

$\hat{Y} = -23.634 + .8565X_2$

| Source of Variation | $SS$ | $df$ | $MS$ |
|---|---|---|---|
| Regression | 381.97 | 1 | 381.97 |
| Error | 113.42 | 18 | 6.30 |
| Total | 495.39 | 19 | |

| Variable | Estimated Regression Coefficient | Estimated Standard Deviation | $t^*$ |
|---|---|---|---|
| $X_2$ | $b_2 = .8565$ | $s\{b_2\} = .1100$ | 7.79 |

# Body fat example

- Regression on $X_1, X_2$
- Regression on $X_1, X_2, X_3$

(c) Regression of $Y$ on $X_1$ and $X_2$

$\hat{Y} = -19.174 + .2224X_1 + .6594X_2$

| Source of Variation | $SS$ | $df$ | $MS$ |
|---|---|---|---|
| Regression | 385.44 | 2 | 192.72 |
| Error | 109.95 | 17 | 6.47 |
| Total | 495.39 | 19 | |

| Variable | Estimated Regression Coefficient | Estimated Standard Deviation | $t^*$ |
|---|---|---|---|
| $X_1$ | $b_1 = .2224$ | $s\{b_1\} = .3034$ | .73 |
| $X_2$ | $b_2 = .6594$ | $s\{b_2\} = .2912$ | 2.26 |

(d) Regression of $Y$ on $X_1$, $X_2$, and $X_3$

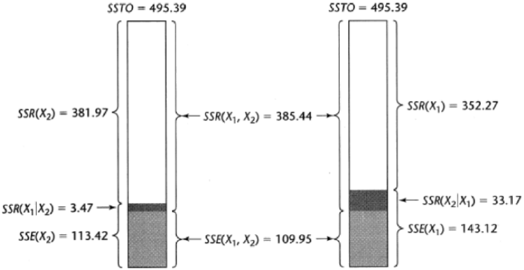
$\hat{Y} = 117.08 + 4.334X_1 - 2.857X_2 - 2.186X_3$

| Source of Variation | $SS$ | $df$ | $MS$ |
|---|---|---|---|
| Regression | 396.98 | 3 | 132.33 |
| Error | 98.41 | 16 | 6.15 |
| Total | 495.39 | 19 | |

| Variable | Estimated Regression Coefficient | Estimated Standard Deviation | $t^*$ |
|---|---|---|---|
| $X_1$ | $b_1 = 4.334$ | $s\{b_1\} = 3.016$ | 1.44 |
| $X_2$ | $b_2 = -2.857$ | $s\{b_2\} = 2.582$ | -1.11 |
| $X_3$ | $b_3 = -2.186$ | $s\{b_3\} = 1.596$ | -1.37 |

# Body fat example

- $p!$ different partitions



# ANOVA Table

| Source of Variation | SS | df | MS |
| --- | --- | --- | --- |
| Regression | $SSR(X_1, X_2, X_3)$ | 3 | $MSR(X_1, X_2, X_3)$ |
| $X_1$ | $SSR(X_1)$ | 1 | $MSR(X_1)$ |
| $X_2 \mid X_1$ | $SSR(X_2 \mid X_1)$ | 1 | $MSR(X_2 \mid X_1)$ |
| $X_3 \mid X_1, X_2$ | $SSR(X_3 \mid X_1, X_2)$ | 1 | $MSR(X_3 \mid X_1, X_2)$ |
| Error | $SSE(X_1, X_2, X_3)$ | $n-4$ | $MSE(X_1, X_2, X_3)$ |
| Total | $SSTO$ | $n-1$ | |

## Partial F-tests, for one predictor variable

- Test $\beta_k = 0$ with general linear test approach:
  Reduced model

$$\begin{aligned} E\{Y\} &= \beta_0 + \sum_{j \neq k} \beta_j X_j \\ \text{versus full model } E\{Y\} &= \beta_0 + \sum_{j} \beta_j X_j \end{aligned}$$

- Partial F-statistic:

$$\begin{aligned} F^* &= \frac{SSE(R) - SSE(F)}{1} \Big/ \frac{SSE(F)}{n-p} \\ &= \frac{SSR(X_k | X_{-k})}{SSE(X_1, \cdots, X_{p-1})/(n-p)} \\ &\sim F(1, n-p) \text{ under } H_0 \end{aligned}$$

- Comparison with t-test $\beta_k = 0$: $F^* = (t^*)^2$

## General liner test in R

- Use anova to carry out F-test from general linear model approach
- Put in the reduced model first, then the full model

```
anova(mod1, mod12)
Analysis of Variance Table

Model 1: Y ~ X1
Model 2: Y ~ X1 + X2
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     19 2824.40
2     18 2180.93  1    643.48 5.3108 0.03332 *
---
```

## Body fat example–continued

- Body fat: can $X_3$ (midarm circumference) be dropped from the model?

| Source of Variation | SS | df | MS |
|---|---|---|---|
| Regression | 396.98 | 3 | 132.33 |
| $X_1$ | 352.27 | 1 | 352.27 |
| $X_2 \vert X_1$ | 33.17 | 1 | 33.17 |
| $X_3 \vert X_1, X_2$ | 11.54 | 1 | 11.54 |
| Error | 98.41 | 16 | 6.15 |
| Total | 495.39 | 19 | |

$$F^* = \frac{SSR(X_3|X_1, X_2)}{1} \div \frac{SSE(X_1, X_2, X_3)}{n-4}$$
$$= \frac{11.54}{1} \div \frac{98.41}{16} = 1.88$$

- For $\alpha = 0.01$, we require $F(0.99; 1, 16) = 8.53$
- We observe $F^* = 1.88$, so we conclude $H_0$, $\beta_3 = 0$.

## Partial F-tests, for a subset of predictor variables

- Test if several regression coefficients are zero:
  Test $H_0 : \beta_k = 0$ for any $k \in S$,
  (with $S$ a set of indices, e.g., $S = \{3, 4, 5\}$)
  versus $H_a : \exists k \in S$, with $\beta_k \neq 0$,
- Partial F-statistic (with $\tilde{S}$ the number of elements in $S$):

$$
\begin{aligned}
F^* &= \frac{SSR(X_S|X_{-S})/\tilde{S}}{SSE(X_1, \cdots, X_{p-1})/(n-p)} \\
&\sim F(\tilde{S}, n-p) \text{ under } H_0
\end{aligned}
$$

## R-squared continued

- Coeff. of multiple determination $R^2 = SSR/SSTO$; the proportionate reduction in variation in $Y$ associated with the predictor variables $X_1, \cdots, X_{p-1}$
- Coeff of **partial** determination

$$R^2_{Y2|1} = \frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1)} = SSR(X_2|X_1)/SSE(X_1);$$

the proportionate reduction in variation in $Y$ remaining after $X_1$ was included in the model, gained by also including the predictor variable $X_2$ (relative marginal reduction)
- Generally for subsets $S$ and $R$:

$$R^2_{YS|R} = \frac{SSE(X_R) - SSE(X_S, X_R)}{SSE(X_R)} = SSR(X_S|X_R)/SSE(X_R),$$

the proportionate reduction in variation in $Y$ remaining after $X_j$, $j \in R$ were included in the model, gained by also including the predictor variables $X_j$, $j \in S$

# Portrait studio example (Ch. 6.9)

- How much information does $X_2$ (average disposable income in a city) add to estimating $E\{Y\}$ (the expected sales of the portrait studio) given $X_1$ (number of persons $< 16$) was already included in the model?

```
------------------------------------------------------
Analysis of Variance Table for Y~X1+X2
          Df  Sum Sq Mean Sq  F value    Pr(>F)
X1         1 23371.8 23371.8 192.8962 4.64e-11 ***
X2         1   643.5   643.5   5.3108  0.03332 *
Residuals 18  2180.9   121.2
------------------------------------------------------
Analysis of Variance Table for Y ~ X1
X1         1 23371.8 23371.8  157.22 1.229e-10 ***
Residuals 19  2824.4   148.7
------------------------------------------------------
```

- $SSR(X_2|X_1) = 643$, $SSE(X_1) = 2824$, thus $R^2_{Y2|1} = 0.23$:
  23% of variation in $Y$, remaining after including $X_1$ is “explained” by $X_2$

## Body fat example-cont'd

| Source of Variation | SS | df | MS |
|---|---|---|---|
| Regression | 396.98 | 3 | 132.33 |
| $X_1$ | 352.27 | 1 | 352.27 |
| $X_2 \mid X_1$ | 33.17 | 1 | 33.17 |
| $X_3 \mid X_1, X_2$ | 11.54 | 1 | 11.54 |
| Error | 98.41 | 16 | 6.15 |
| Total | 495.39 | 19 | |

- $R^2_{Y2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{33.17}{143.12} = .232$
- $R^2_{Y3|12} = \frac{SSR(X_3|X_1,X_2)}{SSE(X_1,X_2)} = \frac{11.54}{109.95} = 0.105$
- $R^2_{Y1|2} = \frac{SSR(X_1|X_2)}{SSE(X_2)} = \frac{3.47}{113.42} = 0.031$
- Adding $X_2$ to the model containing $X_1$, SSE would be reduced by 23.2%; SSE would be reduced by 10.5% if $X_3$ is added given $X_1$ and $X_2$ in the model.
- How about a model already contains $X_2$?

## Another way to get $R^2_{Y2|1}$ when $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

- Fit 3 models:
  - Model (a): $Y \sim X_1$, denote residuals by $e(Y|X_1)$
  - Model (b): $X_2 \sim X_1$, denote residuals by $e(X_2|X_1)$
  - Model (c): $e(Y|X_1) \sim e(X_2|X_1)$
- In (c) we are modeling the part of $Y$ that is not explained by $X_1$, with the part of $X_2$ that is not explained by $X_1$
- In model (c):
  - The regression coefficient for $e_i(X_2|X_1)$ is the regression coefficient of $X_2$ in model $Y \sim X_1 + X_2$
  - $SSR = SSR(X_2|X_1)$
  - $R^2$ for model (c) $= R^2_{Y2|1}$
- Plot of $e_i(Y|X_i)$ against $e_i(X_2|X_1)$ is called the added variable plot (for the effect of $X_2$ on $Y$, after controlling for $X_1$)